# Computational Measures of Language Variation in Textual Utterances

by

Krishnapriya Vishnubhotla

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Department of Computer Science
University of Toronto

Computational Measures of Language Variation
in Textual Utterances

Krishnapriya Vishnubhotla
Doctor of Philosophy

Department of Computer Science
University of Toronto
2024

# Abstract

The use of language as a tool for self-expression and social communication is marked by extensive variation in the way that language is used by different people and communities. The computational modeling of this variation studies the *social meaning* (as opposed to semantics) encoded in the choice of specific words from among synonyms, specific sentence structures, rhetorical devices, strategies of information organization, and so on. The types of textual utterances studied span various types of electronic communication (emails, social media posts), short-form written media (essays, newspaper articles), and longer forms of writing (full-length books), each of which are influenced in their style by different dimensions of social context. Research in this area is made challenging by the difficulty in exhaustively enumerating the aspects of social context that influence the style of a text, and consequently, a scarcity of labelled datasets.

We present datasets and computational frameworks to understand variation in language use as a function of speaker identity. In the domain of full-length literary novels, we use the term "speaker" to refer to the various fictional characters that interact via dialogue throughout the course of a novel. Within any fictional narrative, characters play a central role in holding the attention of their audience, and remain memorable for a wide variety of reasons: their personality, their emotional journey, their snappy retorts. We quantify some of these aspects for the first time in full-length English-language literary novels published in the period 1810-1950. Our research is enabled by the creation of a large dataset of novels that are annotated for various aspects of character and dialogue, which exceeds the size of prior datasets in this domain by a factor of 10.

In the final portion of the thesis, we also demonstrate the presence of, and regularity in, variation in the emotional features of language used on Twitter (now X). We find significant differences in these features along the two axes of time and geographic location, and in the concluding section briefly discuss the possibilities that lie at the intersection of computational measures of emotional expression and the affective sciences.

To Family

# Acknowledgements

I am grateful to the University of Toronto, the Natural Sciences and Engineering Research Council of Canada, and the Vector Institute for providing the funds and resources that made this work possible.

Thank you to Yang Xu for being on my supervisory committee and for the valuable feedback at every checkpoint, which was instrumental in shaping this thesis. Many thanks also to Suzanne Stevenson for being a part of my final committee, and to David Bamman for agreeing to be my external examiner, being so kind with his feedback, and making the defence an exciting and enjoyable one.

I am indebted to the wonderful mentors who helped me navigate the research landscape with kindness, wit, and warmth, and showed me how to keep a love for language at the center of all my projects: Graeme Hirst, Adam Hammond, Frank Rudzicz, and Saif Mohammad.

Graeme, thank you for taking a chance on me as a Masters student, and for since then showing a steady faith in me that has been my anchor through the many ups and downs of a PhD. Our weekly meetings have been the most rewarding part of my time as a graduate student. I hope I was a worthwhile distraction from what must have been a well-timed retirement.

Adam, thank you for guiding so much of my work, and for being so generous with your time, resources, and expertise. I would have run out of ideas a long time ago if not for your constant enthusiasm and insights. My thanks also to the many wonderful annotators from the English department who made our work possible.

Frank, thank you for keeping me connected to the big wide world of NLP research, giving me many opportunities to grow as a researcher, and for constantly reminding me how fun language and literature is.

Saif, I am so grateful to have found your mentorship when I needed it the most. Working with you has been an invaluable lesson in learning to ask and answer the right questions, and in valuing the journey above the destination. Thank you for all your support.

To my collaborators, Mohammad Abdalla and Daniela Teodorescu, and the wonderful team at the UNC Affective Sciences Lab, thank you for showing me how the work is done, I have learnt so much from you all. There are several others, too many to mention, who have trusted me with opportunities along the way and have made this a colorful PhD. I am deeply thankful to each and every one of them.

One of the best things about graduate school, despite the individuality of its effort, is that you get to hang out with some very smart, special people. Thank you Kawin,

for a friendship that I value most highly and cherish most dearly. Jonah, Ayoub, Siva, Nikita, I am very lucky to have found your friendships.

I am thankful for the many faces that have made up my UofT community over the years, they made the journey much more enjoyable: Akshay, Chloe, Serena, Nona, Sean, Julia, Jai, Patricia, Zining, Raeid, Bai, Mete, Alice, Akós, among others.

I am also very lucky to have found friends in Toronto who have become family, and without whom I could not call the city home: Sid, Shruti, Gokhul, Amlan, Sindhu, Aparna, Deeksha, thank you for keeping me grounded, for always keeping life interesting, and for being there when it mattered.

To the friends from past lives whom I can count on still: Nidhi, Smriti, Meet, and Ani, thank you for being sources of comfort in a sometimes scary world, for conversations that make me smile, and for the reminders that life has a way of figuring itself out.

Finally, everything in this thesis, as with every endeavour of my life, academic or otherwise, is made possible by a family that has always cheered me on, no matter the distance between us, and celebrated every achievement with unconditional love and pride. My parents instilled in me a work ethic that is simple and honest, placed in me a confidence that never wavered, and provided me with every opportunity I could ever need or want. I am so lucky to have two beautiful sisters, around whom joy, love, and laughter is a guarantee. I am inspired by all of your lives, and this thesis is dedicated to you.

# Table of Contents

# Chapter 1

# Introduction

## 1.1 Overview

Variation in language use is witnessed across speakers and communities, and is influenced in complex ways by a number of variables: personal demographics (age, gender, ethnicity, personality), the audience or participants, social setting and associated norms, the topic being discussed, the medium of communication and associated genres (a speech, a formal essay, a social media post), and additional, intentional stylistic effects (sarcasm, imitation). This variation is expressed linguistically through differences in the choices of words and the syntactic structures of sentences, the use of rhetorical devices, and strategies of information organization, among others. Oftentimes, this variation is studied along a set of higher-level lexical and social dimensions: the level of emotionality, formality, or humour, the effects of age, gender, ethnicity, the composition of one's social networks, and regional and dialectal variation.

Computationally modeling such variation is challenging because of the lack of explicit labels on the factors influencing the style of a particular utterance, as well as a general lack of organized data that demonstrates such variation along different axes in a controlled manner. Nevertheless, progress has been made by operating within controlled data settings where variation is demonstrated only along one or two axes, all others being equal, and carefully designing experimental setups and computational models that can effectively account for the effects of confounding variables.

In this thesis, I examine variation in language use primarily as an effect of speaker identity. How, and to what extent, do speakers operating in a common situational context differ in their use of language? The *types* of variation I study can be classified as stylistic, in the sense that they capture a facet of linguistic usage that varies from person to person. Computationally characterizing such variation is of interest for two reasons: it is essential to building *robust* natural language processing systems which are semantically invariant to stylistic variation in inputs, as well as systems that are *flexible* enough to modify their stylistic tone to match that of the user; and it provides a data-driven way of studying sociolinguistic variation, using the large amounts of user-generated textual content that can be found on the Web.

The three sections of this thesis examine various facets of stylistic variation in three distinct scenarios. First, I critically examine the merits and demerits of neural models that seek to learn separate, disentangled representations of meaning and form for sentences. These methods show promise in being able to capture the two broad facets of meaning (semantics) and style (the surface form realization of the meaning) in distinct embeddings. However, I show that these methods work well only in constrained data settings where a notion of semantic equivalence can be established between sentence pairs, i.e., with paraphrase datasets. I then look into ways of measuring closeness in meaning between sentence pairs, and develop a novel dataset of semantic textual relatedness between sentence pairs along with models trained to predict such a relationship.

The above research direction, though full of potential, suffers from too many shortcomings to be of utility when applied to real-world datasets demonstrating linguistic variation. The second set of works in this thesis focuses on a specific type of variation observed in the literary domain — that of characters and their utterances. This work is influenced by the literary theory of dialogism, which views novels as complex fabrics of differentiated voices speaking to and about one another, mediated by a narrator. Previous work in NLP on literary analysis has largely interpreted the novel as a single, undifferentiated blob of text attributed solely to the author. This is a step towards a more nuanced computational analysis of novels, wherein I show that character voices within a text can be distinguished from one another based solely on the content and style of their utterances, and that the emotional arcs (a temporal view of emotion states) of these characters are quite distinct from one another and also from the arc of the narration itself. As a step towards enabling more large-scale analyses of characters and their voices, I also evaluate and advance the state-of-the-art in speaker attribution in literary novels.

The third and final section is also focused on measuring temporal emotional variation, but in the more real-world domain of social media data, specifically tweets. Using the framework of Utterance Emotion Dynamics, we look at temporal and geographic variation in the emotions expressed by Twitter users from the US and Canada during the years 2015–2021, a time-period that comprises certain events of particular interest, such as the onset of the COVID-19 pandemic and the subsequent development and release of associated vaccines, and the 2020 US presidential elections.

In summary, this thesis proposes, and demonstrates the utility of, computational models to capture specific dimensions of variation in language use in an inter-disciplinary setting, where usability and interpretability of the models and results are vital. Along the way, we discover the challenges posed by the specific domains of data we work with, and formulate ways to overcome these challenges. While these are useful steps forward, this thesis also serves to reinforce the intricacies of computationally modelling language use variation, where the increasingly favoured NLP paradigm of one-model-fits-all will likely not be of much use. I expand further on these points in the concluding chapter of this thesis.

## 1.2 Background

In linguistics, the distinctions between semantics, or meaning, and style, or form, have been extensively theorized and debated. Form is generally used to refer to the surface-level symbols (words, phrases, sentences) used to represent an underlying meaning. While words can be composed in different ways to yield texts that convey drastically different meanings, one can also express the same meaning with different surface-form realizations, i.e, paraphrases (*"This is my wife."*, *"I am married to her."*). Whether form is entirely orthogonal to meaning, however, is not very clear; often, variations in the words used to convey a certain sentiment carry a *social meaning*, subtle indicators of contextual and social information that are external but complementary to the semantic information being conveyed. A change in the voice of a sentence from active to passive, for example, changes the emphasis or focus of the reader from the subject (the agent) to the object (the beneficiary) of the sentence. Even at the word-level, the choice of a particular synonym is influenced by certain connotative nuances of meaning (what influences the selection of the appropriate word between *forest, woods, jungle*?).

Linguistic variation as a function of social aspects has been studied extensively in linguistics, and by extension, in computational linguistics, under the broad umbrella of (computational) sociolinguistics. William Labov in the 1960s studied how linguistic features varied among the speakers of different dialects of English, such as AAVE (African American Vernacular English), and among social classes of people living in New York City (Weinreich et al., 1968; Entwisle and Labov, 1975). Subsequent works in computational linguistics and NLP have prominently analyzed the effects of gender and age (Labov, 1990; Eckert, 1989; Bamman et al., 2012; Voigt et al., 2018), geographic regions (Johnstone, 2002; Eisenstein et al., 2010), and social networks (Eckert, 2000; Danescu-Niculescu-Mizil and Lee, 2011; Jurgens et al., 2023).

### 1.2.1 Data Sources

Traditional ways of cataloguing and analyzing language variation involved taking surveys, or interviewing groups of people who represented the dimension of social variation that was of interest. These methods arguably involve an *observation bias*, wherein the default manner of speaking is automatically modified to suit the conversational context of a survey or an interview (referred to as the observer's paradox). The creation of digital text archives, as well as the advent of social media and its widespread usage, has brought about new opportunities to study linguistic variation in a more natural setting, across much larger sections of the population, and over larger spans of time, though the latter is also confounded by the idiosyncracies of the medium itself.

Some of the initial datasets to study stylistic variation included collections of essays (Goldstein-Stewart et al., 2008), news articles (Stamatatos, 2013), blogs (Argamon et al., 2007), and emails (Keila and Skillicorn, 2005). Authorship attribution studies, wherein stylistic features are used to identify the (unknown) author of a text, were conducted on a variety of literary sources such as plays, short stories, and full-length novels (and famously, on the Federalist papers written by some of the Founding Fathers of the United States of America) (Stamatatos, 2009; Stamatatos et al., 2000). The study of style in literature is, however, complicated by other confounders such as the genre, topic, and intra-novel variation in prose styles. In the last couple of decades, data from social media sites like Twitter and Reddit have gained popularity, allowing data-hungry methods involving neural networks to be applied to studies of style. The yearly-challenges from the PAN tasks on authorship attribution and profiling have led to the collection and release of several such datasets (Juola, 2012; Bevendorff et al., 2022). More recently, the focus on controllable text generation in NLP has spurred the creation of datasets demonstrating stylistic variation along certain axes (formality, politeness, humour) (Rao and Tetreault, 2018b; Madaan et al., 2020; Lyu et al., 2021).

### 1.2.2 Methodology

The primary computational frameworks to model linguistic variation have largely stayed the same, but the representational and predictive power of the models has greatly improved. Typically, one defines a set of linguistic features that are thought to capture the style of a text span – also called stylometric indicators. The features are computed for the set of datapoints representing the different social groups or contexts, and tested for a consistent, significant difference in value across groups. Generally, we hope that the inter-group variation in stylometric indicators overpowers any intra-group variation. Classification and regression models, such as linear or logistic regression, are used to test whether the textual features are predictive of the social group they represent.

An alternative modelling approach is to use generative latent variable models, like LDA (Latent Dirichlet Allocation) (Blei et al., 2009). These models are defined with a set of latent factors that are hypothesized to affect the distribution of words in a document, and trained to identify the level of influence of each of these latent factors that in turn maximizes the probability of the dataset being generated. These approaches allow us to model multiple stylistic factors, as well as their interdependence, in a single framework (Bamman et al., 2014a; Brooke and Hirst, 2013a).

With neural approaches to modelling language, the need to manually select relevant linguistic features is diminished. These models *learn* a numerical representation of the input text as part of the parameter optimization process for a particular predictive or generative task. While they reduce the subjectivity and human error involved in manual feature engineering, they are also consequently less interpretable in allowing us to analyze which linguistic features actually contributed to a high predictive accuracy of the group label. Neural methods have been shown to take advantage of the selection biases exhibited in datasets to rely on irrelevant features in order to maximize prediction power. These methods also depend on the availability of large amounts of data, in the range of a few thousand examples per class, for training and validation.

Generative neural models of language have grown particularly powerful over the last few years. These models are trained on large amounts of text data scraped from every possible source on the Internet, using architectures whose training can be parallelized across multiple nodes of computation (i.e, distributed training with several GPUs). While these models are now able to generate text in a wide variety of styles with astonishing fluency, we are less sure than before of the mechanisms enabling such generative flexibility, other than to say that there are billions of matrix computations involved. Qualitative studies of language variation have therefore advanced at a slower rate than quantitative models of it.

This thesis takes the side of qualitative understanding over modeling power. In the first section, we briefly study neural models of style representation, and find them quite unsatisfactory for modelling the kinds of language variation we are interested in, which occurs at a smaller scale than demanded by the former. We instead work with simpler, lexical models of style and emotion to characterize a speaker's utterances, and demonstrate the extent of their variation in multiple situations — primarily, English literature written in the period 1810-1950, and tweets from North America in the years 2015-2021.

# Part I

# Meaning and Form

# Introduction

A natural approach towards understanding stylistic variation in language use is to ask how the same information, or meaning, or content, can be conveyed using different surface forms, or styles. Texts that differ in such a manner are termed paraphrases, formally defined as "sentences or phrases that convey the same meaning using different wording". A formal way of greeting someone, for example, could be "Hello, how are you today?", whereas the informal variation would be something along the lines of"Hey, what's up?". As dicussed in the introductory chapter of this thesis, the distinction between content and style, while somewhat intuitive, can be hard to define theoretically. Style can be viewed as the aspect of language expression that is extraneous to meaning, the aspects of a text that are lost in translation or paraphrasing. Functionally, we can define a level of meaning-equivalence by, say, checking how close the definitions of two words are in a dictionary. For sentences and longer texts, however, it becomes harder to find formal definitions of equivalence. Relevant NLP research in this area includes work on formal meaning representations, the collection of paraphrase datasets, and the development of representative paraphrase typologies.

Semantic representation frameworks seek to encapsulate the meaning of a sentence in a logical form, often expressed using formalisms like lambda calculus (Carpenter, 1997; Zettlemoyer and Collins, 2005). Typically, these methods rely on first, defining the important components of meaning understanding — who did what to whom, where, when, and why — along with an ontology that categorizes each event or argument into types (FrameNet, VerbNet, and PropBank are some frameworks in NLP) (Baker et al., 1998; Schuler and Palmer, 2005; Palmer et al., 2005); second, parsing natural language sentences to extract the relevant arguments for each of these components; and then representing them using a specific formalism. Abstract Meaning Representations (AMR) and Minimal Recursion Semantics (MRS) are examples of semantic representation languages developed in NLP (Copestake et al., 2005; Banarescu et al., 2013). However, these works are still evolving and far from being a complete encapsulation of meaning; parsing algorithms used to build these representations are also limited in their ability to process complex and informal sentence types.

The collection of paraphrase datasets has helped advance an alternate way of representing meaning in NLP, one that is largely data-driven (Cohn et al., 2008; Hovy and Bhagat, 2009). Vector representations of sentences are learned by training a neural network on a proxy objective that captures meaning equivalence — training the model to predict whether two sentences are paraphrases or not, or encouraging representations of paraphrases to be closer to each other in the vector space than non-paraphrases, for example. Some of these datasets are collected from naturally occurring paired variations, such as translations of books and articles, whereas others are collected using human annotations or using automatic machine translations (Zhou and Bhat, 2021).

In the next two chapters, I describe projects that contribute further to our understanding of meaning and style. In the first, I look into models for disentangled representation learning for texts, where the two aspects of style and semantics are viewed as containing complementary, orthogonal information. I unify several learning methods proposed for this task into a single framework, and evaluate the contributions of each component systematically on a highly-structured Natural Language Generation dataset. The outcomes of this work influence the direction of the rest of my thesis, by demonstrating that vector space representations of semantics and style are hard to obtain for unstructured, unlabelled, small-scale datasets. In the following chapter, I describe my contributions to a project led by another PhD student, Mohamed Abdalla, on building a dataset of semantic relatedness between sentence pairs (STR-2022). We show the shortcomings of existing datasets and the methods used to obtain them, and demonstrate the utility of STR-2022 in training sentence representation models that are better able to capture closeness in meaning.

# Chapter 2

# Disentangling Neural Representations

This work was published in the Findings of ACL 2021 as follows, and is reproduced here with little to no modification:
Krishnapriya Vishnubhotla, Graeme Hirst, and Frank Rudzicz. 2021. An Evaluation of Disentangled Representation Learning for Texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1939–1951, Online. Association for Computational Linguistics.

**Author Contributions:** All authors conceived the idea for the project during regular discussion sessions. Krishnapriya developed the methodology and experimental setup, with supervision from Profs. Graeme and Frank. Krishnapriya wrote the code, performed the analyses, and wrote the first draft of the paper. All three authors developed the discussion sections for the final version of the paper.

## 2.1 Introduction

The similarity of texts can be assessed along multiple dimensions. They could contain the same topics, as identified by semantic similarity. They could belong to the same genre or be written by the same author, in which case we might identify stylistic similarity. Texts that present a positive sentiment may be considered similar to one another when compared to those that express a negative sentiment, even if they talk about different topics. The similarity of texts, therefore, must be defined together with a frame of reference or a pre-specified dimension of variation.

Text representations obtained by current representation learning methods combine all of these different aspects of a text into a single vector embedding (Conneau et al., 2017; Reimers and Gurevych, 2019a). This results in only a fuzzy measure of text similarity when it is calculated using methods such as the cosine distance between vector embeddings. Recently, some research in NLP has focused on learning *disentangled* representations for texts, which aim to capture the different dimensions of variation of a text in separate vector embeddings. These methods have been investigated for style transfer to obtain disentangled representations of content and style (John et al., 2019; Romanov et al., 2019; Cheng et al., 2020), and paraphrase generation for disentangling syntax and semantics (Chen et al., 2019; Balasubramanian et al., 2020). Inspired by parallel developments on style transfer and disentanglement in computer vision, many of them operate within the variational autoencoder framework, where the autoencoder is modified to now encode a text into two latent vectors: one capturing the style (the aspect of variation), and the other capturing the content. Style transfer is then achieved by combining the content vector of the input with a style vector of the target style.

Disentanglement-based models offer two main advantages when compared to end-to-end style transfer methods:

1. Sampling from the latent space of the style embeddings allows for more diverse and controlled stylistic generation.

2. Similarity of documents can now be calculated for each aspect of variation, allowing for finer-grained retrieval.

Works on style transfer in NLP operate with varying definitions of what constitutes style. Many choose to define this as a factor of variation in data that can be manipulated, including aspects such as topic and sentiment. This approach has been contested by others who maintain that the semantic content of a text should not be modified when manipulating style. The latter definition fits with what stylometric analysis and linguistics consider to be the style of a text. Thus, the output of a style transfer system should be a paraphrase of the input text.

In this work, we focus on models that aim to disentangle **content** from **form**, or meaning from style, for texts. Thus, style transfer is viewed as a form of paraphrasing, where the paraphrase demonstrates certain stylistic properties. It is important to make this distinction between what constitutes style versus meaning for a text, more so when formulating style transfer problems, in order to have measurable definitions of what information may and may not be changed by the model. Parallel paraphrase datasets, therefore, are a much-needed resource for the effective evaluation of these

| **Meaning Representation** | **name[nameVariable], food[Indian], customerRating[average]** |
|---|---|
| EXTROVERT | nameVariable is an Indian place, also nameVariable has an average rating, you know. |
| UNCONSCIENTIOUSNESS | Yeah, mmhm... I don't know. nameVariable is an Indian place with a damn average rating. |
| CONSCIENTIOUSNESS | Did you say nameVariable? I see, well it is an Indian restaurant with an average rating. |
| DISAGREEABLE | Actually, basically, everybody knows that nameVariable is an Indian restaurant, also it has an average rating. |
| AGREEABLE | Let's see what we can find on nameVariable. Well, right, it is an Indian restaurant with a quite average rating. |

Table 2.1: The same meaning representation mapped to different stylistic surface realisations in the PersonageNLG dataset.

models. However, few works on disentangled representation learning actually evaluate their models on such datasets, testing instead only on the non-parallel datasets used for training. Further, some works evaluate exclusively on metrics from the style transfer task, ignoring the retrieval aspect.

The goal of this study is to conduct a systematic and grounded evaluation of various disentangled representation learning models. We first use, as a testbed for our evaluation strategy, a highly-structured Natural Language Generation dataset, PersonageNLG (Oraby et al., 2018), which maps a meaning representation to a set of stylistically different surface realisations corresponding to five personality types (Table 2.1). This dataset provides us with textual variation and gold-standard annotations for the two dimensions of interest, content and form. The structured and somewhat synthetic nature of this dataset allows us to systematically investigate the quality of the disentangled representations for metrics of aspect-specific retrieval as well as style transfer.

We then extend our experiments to two other parallel style transfer datasets: the GYAFC formality corpus (Rao and Tetreault, 2018a), and the Bible dataset (Carlson et al., 2018). Although parallel, they are not annotated for semantic content as the PersonageNLG dataset is; however, they are arguably more representative of the kinds of data we expect to obtain in the real world. Despite testing our models with loss functions that do not require parallel data, we limit ourselves to such datasets for the ease and consistency of evaluation. Our code is publicly available at `github.com/priya22/drl-nlg-eval`.

## 2.2 Model Architectures

Disentanglement of latent spaces has been widely studied and very successful in computer vision applications, but less so in NLP. This can be attributed to the vague nature of what actually constitutes style as opposed to content for a text, and uncertainty as to whether they can actually be disentangled at all (Lample et al., 2019; Yamshchikov et al., 2019). However, by using some supervision with respect to these two dimensions, researchers have attempted to obtain representations that *for the most part* encode information relating to only style or only semantics.

The models used to achieve style transfer fall into a few broad categories. End-to-end sequence transformation models are inspired by machine translation seq-2-seq models, where the translation is done from style $A$ to style $B$. These sometimes require parallel data, but methods such as backtranslation circumvent that (Prabhumoye et al., 2018; He et al., 2020). Some others look at this as a controlled text generation problem, where the control is generally a categorical variable indicating the desired stylistic class of the output, and is passed along with the input to a text generation module such as an LSTM (Hu et al., 2017; Ficler and Goldberg, 2017).

The focus of this work is on a third class of models that first learn disentangled latent representations of style and not-style (henceforth referred to as content) for a text, and train a generator that takes both vectors as input. To transform a text $A$ into the style of text $B$, we extract the content vector of the former, the style vector of the latter, and pass them through the generator. Note that here, the style vectors of each text are not the same categorical variable, but rather a vector embedding that encodes the style-specific properties of the text. One can also obtain a single style vector representation by averaging the style vectors of all texts belonging to that class, as (Fu et al., 2018) did; however, we are more interested in disentangling information at the individual text level rather than in corpus-level indicators.

Romanov et al. (2019) first proposed obtaining separate embeddings of form and meaning of texts. Starting with an encoder-decoder setup, they added adversarial and motivational losses based on style labels that encourage the form vector to encode information relevant to the label. Their models were evaluated on non-parallel datasets with two types of stylistic variation: diachronic language shift and newspaper titles versus scientific paper titles. In parallel work, John et al. (2019) proposed a disentanglement model that appends additional content-based losses, where content is approximated by a bag-of-words representation of the text. Their approach was applied to sentiment transfer for Yelp and Amazon reviews.

Other work has looked at disentangling syntax from the semantics of a text. Chen et al. (2019) proposed a VAE-based model that used parallel paraphrase corpora; this was also the focus of Bao et al. (2019) and Balasubramanian et al. (2020).

All of these works are very similar in the base model architecture and the kinds of loss functions used to guide disentanglement. In the following sections, we consolidate and propose a broad categorization of these losses that we hope will guide future work in this area. We then evaluate these models on parallel style transfer datasets, with ablation studies on the PersonageNLG dataset.

**Note on unsupervised disentanglement:** While unsupervised approaches such as the $\beta$-VAE have been very successful at disentangling factors of variation in visual data (Higgins et al., 2017), we are still far from achieving such a clean separation of the data generating factors for text. A recent promising approach in this direction was presented by Xu et al. (2020), who use pretrained models along with a novel constraint over the latent space of a VAE to control the sentiment and topic of a text.

## 2.3 Methodology

### 2.3.1 Autoencoder Model

Following previous literature, our encoder module takes as input a text, and computes latent vector embeddings for each aspect: content and form. The decoder takes as input both vectors, and generates output text. The entire autoencoder model is trained to reconstruct the input text.

Let us denote our content and form encoders by $E_c$ and $E_f$, the decoder by $G$, and their model parameters by $\theta_{E_f}$, $\theta_{E_c}$ and $\theta_G$ respectively. Our base loss can thus be written as:

$$L_{AE} = L_{rec} + \beta L_{reg} \tag{2.1}$$

where

$$\begin{aligned} L_{rec}(\theta_{E_c}, \theta_{E_f}, \theta_G) = & \\ & \mathop{\mathbb{E}}_{x}[-\log \; p_g(x \mid E_f(x), E_c(x))] \end{aligned} \tag{2.2}$$

is the reconstruction loss of the autoencoder given input $x$, $p_g$ is the decoder distribution, and $L_{reg}$ is an additional regularization term. For a Variational Autoencoder (VAE) model, this is the Kullback-Leibler divergence between the latent posterior distributions $q$ of the encoders and the latent prior $p(z)$:

$$L_{reg}(\theta_E) = D_{KL}(q(z \mid x) \parallel p(z)) \tag{2.3}$$

An alternative regularization for text autoencoders was proposed by Shen et al.

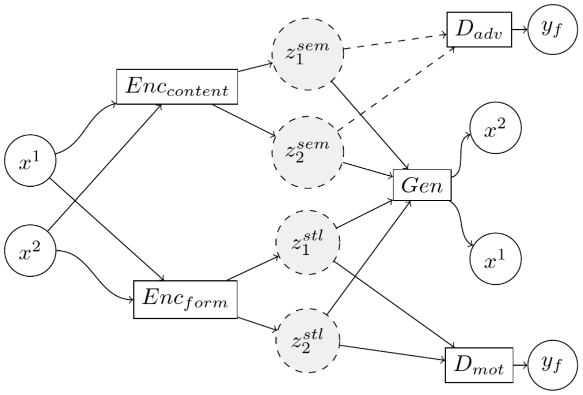


Figure 2.1: The main components of a Disentangled Representation Learning model. $z^{sem}$ and $z^{stl}$ denote the content and form vectors respectively; each is input to a motivational and an adversarial network. The generator is trained to reconstruct the original input as well as paraphrases.

(2020), where the AE loss is augmented with a denoising objective. The input text is perturbed with small amounts of "noise" in the form of word deletions or substitutions; the autoencoder is still trained to reconstruct the original text. Here,

$$L_{reg}(\theta_E, \theta_G) = \mathop{\mathbb{E}}_{(x,\tilde{x})}[-\log\ p_g(x \mid E(\tilde{x}))] \tag{2.4}$$

where $\tilde{x}$ is the noisy version of the input text $x$. These denoising autoencoders (DAEs) were shown to be more stable than VAEs for text modeling.

### 2.3.2 Losses for Disentanglement

With our base autoencoder in hand, we can now start adding losses that encourage each latent vector to encode information relevant to the corresponding aspect, i.e., content (semantics) and form (style).

### 2.3.3 Proxy-based Losses

Supervised losses are usually based on some form of proxy information present for a specific aspect.

For the form dimension, the most common proxy is class labels that indicate the style of a particular datapoint, such as formal or informal. A stronger proxy could include a list of linguistic attributes of the sentence that are highly indicative of and inform its style. These usually have to be manually defined and extracted, as by

John et al. (2019), who use high-polarity sentiment words as a proxy for the sentiment aspect.

An attribute-based proxy for content can be found by looking at the information present in, say, the meaning representation of a sentence (as provided in NLG datasets), or extracting semantics-predictive information such as semantic role labels. John et al. (2019), for example, use the bag-of-words representation of a text as a proxy for semantic information.

These additional losses are usually combined with the autoencoder objective in two ways: as a **motivational** loss, which *encourages* a latent vector to encode the proxy information, and as an **adversarial** loss, which *discourages* a latent vector from encoding the proxy information. Thus, once we define a proxy loss for, say, content, we would append a motivational loss to the content encoder and a corresponding adversarial loss to the form encoder.

Below, we use $z_c$ and $z_f$ to denote the content and form vectors of a text $x$.

### Loss Functions for Form

**Motivational:** For the datasets that we consider here, and in most real-world applications, we have the stylistic class of a text as a proxy for the form aspect. The motivational and adversarial networks are implemented as classifiers that are trained to predict this label from the corresponding latent representation. The loss function of the former is simply the cross-entropy loss of the classifier:

$$L_{mot}(\theta_D, \theta_{E_f}) = \underset{z_f}{\mathbb{E}}[-\log\ D(z_f)] \tag{2.5}$$

$D$ and $\theta_D$ represent the classifier and its parameters respectively.

**Adversarial:** We now want to ensure that the content vector does not contain any information about the form class of the text. Thus, we aim to *maximize* the entropy of the adversarial classifier. This is the approach followed by many prior works (John et al., 2019; Fu et al., 2018), which we also adopt here, as it can be nicely extended to multi-label classification, which will prove useful in the content-based losses.

Adversarial training occurs in two steps. First, the classifier is trained to predict the form label given the content representation. Then, the content encoder's parameters are updated based on the entropy loss:

$$L_{adv}(\theta_D) = \underset{z_c}{\mathbb{E}}[-\log\ D(z_c)] \tag{2.6}$$

$$L_{adv}(\theta_{E_c}) = \underset{z_c}{\mathbb{E}}[\mathbb{H}(D(z_c))] \tag{2.7}$$

where $\mathbb{H}(D(z_c))$ is the entropy calculated over the classifier-predicted label distribution.

### Loss Functions for Content

Proxy information for content is generally rare, and needs to be formulated by means of some heuristic measure. In the case of NLG datasets, we have annotated meaning representations that serve as a good proxy. However, such structured representations of meaning are difficult to obtain for general texts.

Let us assume we have a list of $k$ key-value pairs that represent content, as in the MR from Table 2.1. We represent the content proxy as a $k$-dimensional multi-hot vector $y_c$, where each dimension $y_c^i$ is a binary indicator of whether key $k_i$ is present in the MR.

**Motivational:** The motivational loss is thus defined as the multi-label cross-entropy loss over the classifier prediction, similar in form to Eq. 2.6, but now taking the content vector as input.

**Adversarial:** In turn, the adversarial content loss is found by first training a multi-label classifier that takes the form vector as input and predicts the content attribute vector, and then training the form encoder to maximize the entropy of this classifier.

## 2.3.4 Parallel Losses

These losses require as input a pair of paraphrases, say $x^1$ and $x^2$. We obtain the latent vectors for content and form for each of these: $z_c^1, z_f^1, z_c^2, z_f^2$ respectively.

**Paraphrase reconstruction loss:** Here, we swap the content vectors of the paraphrases, retain the form vectors, and attempt to reconstruct the original inputs. This was used by Chen et al. (2019) to disentangle syntax and semantics in paraphrase corpora.

$$\begin{aligned} L_{para}(\theta_{E_c}, \theta_{E_f}, \theta_G) = & \mathop{\mathbb{E}}_{x_1,x_2}[-\log\ p_g(x^1 \mid z_f^1, z_c^2)] \\ & + \mathop{\mathbb{E}}_{x_1,x_2}[-\log\ p_g(x^2 \mid z_f^2, z_c^1)] \end{aligned} \tag{2.8}$$

**Distance-based loss:** This takes the form of a max-margin loss that aims to keep the cosine similarity between the content embeddings of paraphrases higher than that between a random selection of negative example pairs. This particular loss is used by Chen et al. (2019) and Balasubramanian et al. (2020) to disentangle syntax and semantics, although they differ slightly in the criteria to select positive and negative pairs.

## 2.4 Datasets

**PersonageNLG Dataset:** The PersonageNLG corpus (Oraby et al., 2018) is a set of 88,000 pairs of meaning representations and natural language utterances, based on the E2E challenge dataset. Each utterance is associated with a unique style, which corresponds to one of five personality types: Agreeable, Disagreeable, Conscientious, Unconscientious, and Extrovert. The utterances are obtained by means of a statistical NLG system, and by varying a set of 36 predefined stylistic parameters that specify certain phrase aggregation and pragmatic markers (Table 2.1). The dataset essentially provides us with a structured and synthetic corpus of textual variation, with each utterance annotated for both content (a meaning representation) and form (the stylistic personality class). This makes it ideal for evaluating the quality of disentangled representations.

**GYAFC Dataset:** Introduced by Rao and Tetreault (2018a), the GYAFC corpus consists of 120,000 parallel sentence pairs that are paraphrased in two styles: formal and informal. See section 2.A.1 for details. GYAFC is one of the very few parallel datasets available for style transfer research in NLP.

**Bible dataset:** This dataset, compiled by Carlson et al. (2018), consists of eight verse-aligned public domain versions of the Bible; see section 2.A.2 for details. These versions are spread out across different decades, and thus belong to their own unique stylistic class. The natural parallel alignment between verses, as well as the relatively stable nature of their semantic content across time, makes this dataset ideal for studies in style transfer (although surprisingly few works on style transfer use it).

## 2.5 Evaluation

The goal of our model is to encode in separate vectors the style-specific and content-specific features of a text. The following metrics guide our similarity measures for content and form:

- **Content ($C_{sim}$):** For the PersonageNLG dataset, content similarity between two sentences is measured as fraction overlap between content labels (Section 2.3.3). For generated sentences, we use all possible slot values for each field of the Meaning Representation (Table 2.1) to approximate a bag-of-words content representation, and calculate fraction overlap of content terms in both sentences. For the other two datasets, we use the BLEU scores between the generated text and the target paraphrase as a measure of content preservation.

- **Form ($F_{class}$, $F_{sim}$):** For all three datasets, we first train a fasttext[1] classifier on their respective training sets to predict stylistic class given the input text ($F_1$ scores on the test sets are shown in Table 2.2). This classifier is then used to predict the style class of a generated text. $F_{class}$ is the $F_1$ score of the predicted labels for generated texts, using the target labels as ground truth.
  Additionally, for the NLG dataset, we use an $F_{sim}$ measure that measures the fraction overlap of non-content words of the two texts, where "non-content" is defined as all words that are not associated with content as defined above.

We divide our evaluation metrics into three groups, based on the capabilities and use-cases of learning disentangled representations.

### 2.5.1 Autoencoder Capabilities

**Reconstruction:** One of the basic functions of our model is as an autoencoder, i.e., a model that can reconstruct the input text from its latent encoding. We use the self-BLEU score between the input (reference) and the generated text to measure reconstruction quality.

### 2.5.2 Disentanglement

The quality of disentanglement of representations is assessed in two main ways.

**Classification:** The first is a classification task that aims to predict the proxy information for each text using the latent vectors. For each of our dimensions of content and form, this gives us four measures corresponding to the accuracy of a classifier trained to predict content (form) information from the content (form) vectors, and that of a classifier trained to predict form (content) information from the content (form) vectors. Ideally, we want the former numbers to be high and the latter to be close to random chance.

**Retrieval:** As stated, one of the advantages of having disentangled representations for each aspect is that we can now obtain aspect-specific similarity scores. Since all our datasets are parallel paraphrase corpora, we can measure how well the content

[1] https://fasttext.cc/

| **Dataset** | **$F_1$ score** |
|---|---|
| PersonageNLG | 0.99 |
| GYAFC | 0.87 |
| Bible | 0.72 |

Table 2.2: Performance of the external fasttext classifier on test sets.

vectors perform at retrieving paraphrases. For each sentence in our test set, we obtain the cosine similarity scores of its content vector with that of every other sentence, and look at how many of the top-$k$ matches are paraphrases of the input. We evaluate this for $k = 5$ for the GYAFC and Bible datasets, and $k = 1$ for the NLG corpus.

Similarly for form, we find the top-$k$ neighbours for the form vector of each sentence and report the precision@$k$ of retrieving texts from the same stylistic class. This metric is particularly informative for PersonageNLG, where we look at the $F_{sim}$ between the input and the closest match.

### 2.5.3 Style Transfer

Finally, we evaluate the effectiveness of our model for the task of style transfer, by testing with paraphrase pairs. Thus, for each pair of paraphrases in the test set, we obtain the content vector of the first and the form vector of the second, and pass them to the decoder module (and vice-versa). The **content preservation** and **transfer quality** of generated sentences are measured using $C_{sim}$ and $F_{class}$ respectively. We also measure the **fluency** of the generated text by measuring the perplexity of generated sentences with a trigram Kneser-Ney language model trained on the training set of each dataset.

## 2.6 Experiments

### 2.6.1 Setup

The encoder and decoder of our base model are 2-layer LSTM networks with a hidden size of 64. Both the content and form vectors are of the same size for each dataset: 16 for PersonageNLG and 32 for the others. At each decoder timestep, the concatenated latent vector $z = [z_c, z_f]$ is added to the input to obtain the next prediction. During training, teacher forcing with probability 0.4 is used; we use greedy decoding for the PersonageNLG dataset and and beam search with a beam size of 5 otherwise.

Motivational and adversarial classifiers are single-layer linear networks trained with RMSprop.

The GYAFC and NLG datasets come with predefined training and test splits. For the Bible dataset, we use a random stratified split with 65–15–20 split for training, validation, and test respectively.

### 2.6.2 Experimental Method

Our goal is to methodologically evaluate the effectiveness of each of these losses for disentangling content from form. We start with our vanilla autoencoder model ($L_{ae}$), and at each step, add additional losses based on incorporating some supervised information into our model. The terms we add are guided by some intuition on the kinds of supervision we would expect to see in the real world.

1. **Form losses $L_{form}$:** This assumes that each text is labeled with a class that indicates its stylistic category, such formal / informal, Shakespearean / modern, positive / negative, etc. This enables us to append two of our losses to the base loss: the motivational and adversarial form losses (Section 3.3.1).

2. **Motivational only $L_{mot}$:** We now add our proxy information for content. We first keep only the motivational losses and remove the adversarial losses for each aspect.

3. **Combined proxy losses $L_{proxy}$:** We add adversarial losses for form and content to the model above, giving us our full proxy-loss–based model.

4. **Paraphrase losses:** Finally, we add the parallel losses detailed in Section 3.4, taking advantage of our parallel datasets. The alignment of two paraphrases essentially acts as a proxy for the equivalence of semantic content between two texts. Accordingly, we test the following loss combinations:

    - Parallel losses only (Section 3.4) ($L_{para}$);
    - Parallel losses + form losses from point 1 above ($L_{para_f}$).

**Baseline:** We additionally compare the effectiveness of these models when compared to a categorical conditional generation model. Here, the form vector is simply an 8-dimensional encoding of the style class label, rather than derived from the input text. The model is trained using the $F_{adv}$ and $C_{mot}$ losses to ensure the content embedding doesn't encode style information, along with the reconstruction loss $L_{rec}$.

All of these loss combinations are tested on the PersonageNLG dataset, since it is annotated with proxies of both content and form.

## 2.7 Results and Analysis

We experimented with both the VAE and the DAE models for our base architecture, and found that the latter was more stable during training. Training the VAE with multiple latent vectors and additional losses often resulted in the model completely ignoring one of the latent vectors; stable modeling of such architectures is still an active area for text data and is left to future work.

### 2.7.1 Disentanglement

We first examine how well our models are able to disentangle information pertaining to form and content into the respective latent vectors. Table 2.3 reports the performances of each model for the metrics discussed in Section 2.5.2. For conciseness, we only report cross-aspect classification scores in the Classification column, where a lower number indicates better disentanglement. More detailed results with same-aspect scores are presented in Appendix 2.C.1.

In the absence of parallel data, we see that directly adding supervised losses along each dimension is the most effective strategy of disentangling information. Accordingly, the largest performance drops on cross-aspect classification are achieved with the addition of motivation losses $L_{form}$ and $L_{mot}$ for form and content. Adversarial losses do help the overall performance of the model as demonstrated by the drop in cross-aspect classification metrics, especially in the form domain. The maximal supervision afforded by the paraphrase losses $L_{para}$ demonstrates a significant improvement over the best proxy-based model here, indicating that proxy information is generally not complete enough to capture semantic content. However, the lack of similar supervision along the form dimension is reflected in the higher cross-aspect classification scores across all models.

We show t-SNE plots of the form and content vectors computed by each model in Appendix 2.B. The paraphrase model gives us neat clusters of the content vectors corresponding to the different meaning representations.

However, classification numbers alone don't present the whole picture. Our measures of retrieval quality help to isolate the effects of classifier effectiveness from the goodness of the representations alone. For the NLG dataset in particular, the re-

| | **Autoencoder** | **Disentanglement** | | | |
|---|---|---|---|---|---|
| | BLEU | Classification: $F_1 \downarrow$ | | Retrieval $\uparrow$ | |
| Target → | | Form | Content | Form | Content |
| Input → | | $z_c$ | $z_f$ | $z_f$ | $z_c$ |
| $L_{ae}$* | 43.4 | 0.96 | 0.73 | 0.57 | 0.85 |
| $L_{form}$ | −0.07 | −0.67 | −0.11 | 0.13 | 0.08 |
| $L_{mot}$ | 0.01 | −0.31 | **−0.14** | 0.08 | 0.13 |
| $L_{proxy}$ | −0.05 | −0.73 | −0.13 | **0.13** | **0.14** |
| $L_{para}$ | 0.06 | −0.68 | −0.03 | 0.11 | 0.13 |
| $L_{para_f}$ | −0.03 | **−0.75** | −0.10 | 0.12 | 0.12 |
| $L_{baseline}$ | −0.03 | −0.65 | – | – | 0.09 |

Table 2.3: Results on reconstruction and disentanglement quality for the PersonageNLG dataset. The first row reports the absolute metric for the base autoencoder model $L_{ae}$; subsequent rows report the difference from this base score. The first column reports the self-BLEU score between the reconstructed and input text. For classification, we report the cross-aspect $F_1$ scores of a classifier trained to predict the target aspect from the input. For retrieval, we report the $C_{sim}$ and $F_{sim}$ scores between the input text and its nearest neighbour in the latent space.

trieval scores tell us whether the form vector of a text actually encodes information about the linguistic features informing its style, rather than simply encoding enough to be classified in the right stylistic class. Are sentences with similar *textual* stylistic or content features closer to each other in the embedding space when compared to other sentences from the same style/content class? The relatively low delta scores when compared to classification performance indicate that this is not the case. While there are marginal improvements, proxy-based losses don't seem to be informative enough to enforce fine-grained structure in the latent space. Our experiments on style transfer in the next section reinforce this conclusion.

### 2.7.2 Style Transfer

We swap the form and content vectors of paraphrases from our test set, and evaluate the generated sentences using the metrics defined in Section 2.5.3. For the NLG dataset, as before, we use term-overlap measures of the similarity for the content and style terms between the generated text and the target paraphrase ($C_{sim}$ and $F_{sim}$); results are shown in Table 2.4. Both of these measures are far from their ideal values of 1.0.

The full proxy model $L_{proxy}$ achieves the best performance across all metrics (sample outputs are shown in Appendix 2.C.2). The paraphrase models tend to perform worse

|  | $C_{sim}$ ↑ | $F_{sim}$ ↑ | Fluency ↓ |
|---|---|---|---|
| $L_{ae}$ | 0.29 | 0.46 | 1.11 |
| $L_{form}$ | 0.28 | 0.58 | 1.08 |
| $L_{mot}$ | 0.36 | 0.48 | 1.09 |
| $L_{proxy}$ | **0.39** | **0.72** | 1.10 |
| $L_{para}$ | 0.33 | 0.45 | 1.11 |
| $L_{para_f}$ | 0.35 | 0.55 | 1.09 |
| $L_{baseline}$ | 0.30 | 0.60 | 1.06 |

Table 2.4: Evaluation of style transfer on the PersonageNLG dataset. Arrows denote desired direction of change.

|  |  | Disentanglement |  | Style Transfer |  |
|---|---|---|---|---|---|
|  |  | Clf. ↓ | Ret. ↑ | $C_{sim}$ ↑ | $F_{class}$ ↑ |
| GYAFC | $L_{base}$ | 0.43 | 0.20 | 1.5 | 0.50 |
|  | $L_{para_f}$ | 0.35 | 0.49 | 3.6 | 0.83 |
| Bible | $L_{base}$ | 0.64 | 0.25 | 1.3 | 0.11 |
|  | $L_{para_f}$ | 0.12 | 0.72 | 3.4 | 0.39 |

Table 2.5: Results on disentanglement quality and style transfer for the GYAFC and Bible datasets. The Clf. column reports the $F_1$ score of a classifier trained to predict the stylistic class label from the content vector; Ret. reports the P@5 for retrieving paraphrases using the content vectors.

than the baseline, especially on the transfer strength metric, $F_{sim}$. This points to the form vector not being informative enough, especially when no motivational losses are used. It also indicates that the adversarial losses from the proxy-based models were indeed helpful in disentanglement.

We see similar trends in both disentanglement quality and style transfer for the GYAFC and Bible datasets. The quality of text generated was significantly worse when compared to the NLG dataset, but we are still able to encode the style and content-related information in separate vectors with some success, as evidenced by the retrieval scores.

### Does Disentanglement Help?

Our comparison with the categorical baseline $L_{baseline}$ tells us whether learning disentangled representations indeed provides an advantage for the style transfer task. From Table 2.4, we see that it does quite well on the $C_{sim}$ metric, but is notably lower than $L_{proxy}$ for $F_{sim}$. This demonstrates the advantage of having a separate vector representation of the form of a text, as opposed to the stylistic class.

## 2.8 Discussion

Our experiments all demonstrate that direct supervision along each aspect is crucial for learning good aspect-specific representations. This is the case even for the synthetic PersonageNLG dataset, which is by design constrained to have two separable aspects of variation (meaning and style); this is quite rare in real-world data. Indeed, the best performing style transfer model on this dataset, from Harrison et al. (2019), is a heavily supervised one that conditions a seq-2-seq model with annotations for each type of variation in the surface realisations (i.e., the presence of certain tokens).

In the absence of parallel datasets, proxy information is widely used to encourage disentanglement. However, our results show that such supervision is not sufficient to ensure that the embeddings actually encode the linguistic properties that are characteristic of a text's stylistic class (or meaning). With the retrieval experiments on the NLG dataset, we can see that the $F_{sim}$ scores do not significantly differ between the different models. This indicates the difficulty of learning linguistic properties from class labels alone. This also explains the rather high $F_1$ scores for content classification from form embeddings.

The poor performance of these models on the style transfer task in particular indicates that the decoder, and hence the reconstruction objective itself, is somewhat lacking. This is reflected in the high classification scores of content information from form vectors, especially for the paraphrase model $L_{para}$. Additional constraints such as the backtranslation loss (Prabhumoye et al., 2018) go some way towards mitigating this issue. On the style transfer task, the baseline model $L_{baseline}$ shows performance comparable to the disentanglement models. One explanation for their poor performance is the inherent defects of variational models of text, such as the latent space vacancy issue, as demonstrated by other works (Xu et al., 2020; Shen et al., 2020).

For evaluation of such disentangled representations, traditional metrics of style transfer, such as the accuracy of an external classifier, are not the best indicators of disentanglement, nor a good demonstration of the usefulness of such embeddings. Most works on disentangled representations for style transfer do end up using a single, averaged vector embedding to inform the decoder of the desired target style. If the goal of learning disentangled representations is to perform style transfer between two classes, then a conditioned language model such as that of Ficler and Goldberg (2017) would suffice.

A more useful use-case for disentangled representations is for calculating aspect-specific similarity and retrieval between texts. However, it is not clear whether we can

| | |
|---|---|
| Formal | I'd say it is punk though. |
| Informal | However, I do believe it to be punk. |
| Informal | Gotta see both sides of the story. |
| Formal | You have to consider both sides of the story. |

Table 2.6: Sample paraphrases from the GYAFC dataset.

achieve such disentanglement with current models without fine-grained supervision along each aspect. While the NLG dataset provides us with the necessary supervision to introduce such constraints (via adversarial losses), and also evaluate them, such supervision is not available for real-world datasets.

Encoding the different factors of variation in data in separate embeddings is a desirable goal for learning robust and interpretable text representations, as well as for controllable text generation. While style transfer, and sentiment transfer in particular, has guided most of the prior research in this area, we have shown that the associated metrics and datasets are not entirely representative of the goals of learning disentangled text representations. We re-purposed an existing NLG dataset for this task instead, and performed a stronger evaluation of current models for disentangled representation learning. We have also shown that heavy supervision is needed along each aspect to obtain useful representations. Improvements in variational generative models that can overcome issues of posterior collapse and the use of decoding constraints stronger than the reconstruction loss would greatly benefit such models.

## 2.A Parallel Style Datasets

### 2.A.1 GYAFC Corpus

The *Grammarly's Yahoo Answers Formality Corpus*, or GYAFC for short, is a benchmark corpus for formality style transfer in NLP[2]. It consists of a total of 120,000 informal / formal sentence pairs, split into training, validation, and test sets.

Sentences were initially sampled from the Yahoo Answers L6 corpus, and formal and informal rewrites from each were collected from workers on Amazon Mechanical Turk (Rao and Tetreault, 2018a). Table 2.6 shows example paraphrases from this corpus.

[2] https://github.com/raosudha89/GYAFC-corpus

| Version | Verse |
|---|---|
| KJV | The heart of the prudent getteth knowledge; and the ear of the wise seeketh knowledge. |
| ASV | The heart of the prudent getteth knowledge; And the ear of the wise seeketh knowledge. |
| BBE | The heart of the man of good sense gets knowledge; the ear of the wise is searching for knowledge. |
| DARBY | The heart of an intelligent getteth knowledge, and the ear of the wise seeketh knowledge. |
| DRA | A wise heart shall acquire knowledge: and the ear of the wise seeketh instruction. |
| LEB | An intelligent mind will acquire knowledge, and the ear of the wise will seek knowledge. |
| WEB | The heart of the discerning gets knowledge. The ear of the wise seeks knowledge. |
| YLT | The heart of the intelligent getteth knowledge, And the ear of the wise seeketh knowledge. |

Table 2.7: The same verse (Proverbs 18:15) paraphrased in 8 different diachronic versions of the Bible, from the Bible dataset: the King James Version (KJV, 1611), American Standard Version (ASV, 1901), Bible in Basic English (BBE, 1965), Darby Bible (DARBY, 1890), Douay-Rheims edition (DRA, 1899), Lexham English Bible (LEB, 2010), World English Bible (WEB, 2000), and Young's Literal Translation (YLT, 1862).

### 2.A.2 Bible Dataset

More than 30 English translations of the Bible have been published over the course of four centuries, the earliest being the King James Version of 1611. These versions are all highly parallel, aligned by verse, and are high-quality translations due to the importance of the source. Carlson et al. (2018) identified 8 of these versions that are in the public domain and released aligned corpora for each[3]. Table 2.7 shows a sample verse paraphrased in each of the 8 versions we consider. Each version consists of 31,096 verses, giving us close to 870,000 paraphrase pairs. We first split this into an 80–20 development–test split; the development set is further split into training and validation sets with the same ratio.

## 2.B t-SNE Visualization

t-Distributed Stochastic Neighbor Embedding is a non-linear dimensionality reduction technique useful for visualizing high-dimensional data. Figure 2.2 shows t-SNE plots of the form vectors (left column) and content vectors (right column) for sentences in the test set of the PersonageNLG dataset, for each of the loss function combinations we tested. Adding the supervised losses for form successfully groups the form vectors together into five clusters for each of the personality classes. While content vectors also show some clustering with the adversarial and motivational losses, paraphrase losses here are the most effective at grouping them into neat clusters for each of the unique meaning representations in our test set.

[3] https://github.com/keithecarlson/StyleTransferBibleData

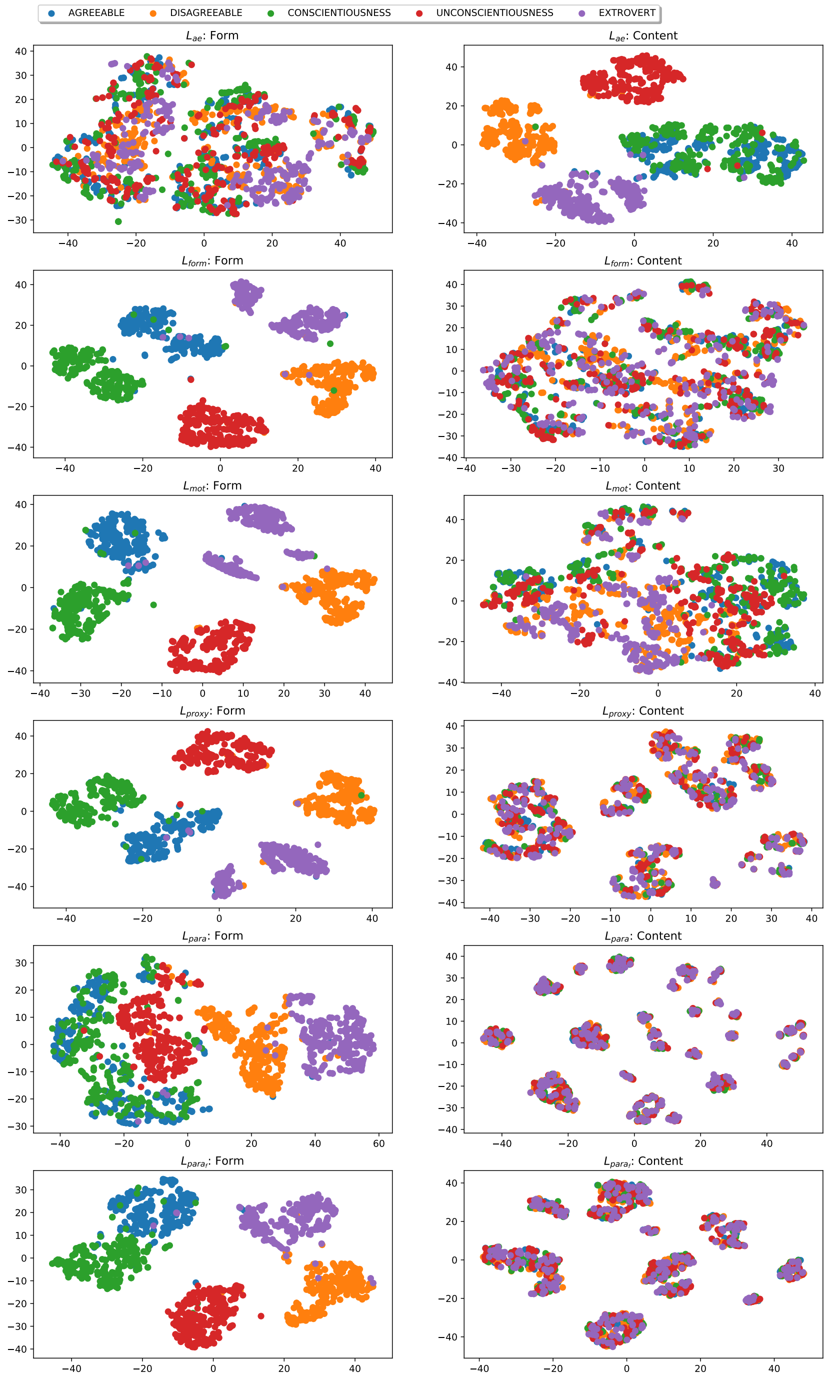


Figure 2.2: t-SNE visualization of form and content vectors for the PersonageNLG dataset, for each of our models. We see that the paraphrase losses enable a clean clustering of the meaning representations across stylistic variations. The domination of extrovert (purple) in some of the conditions is an artifact of the visualization when points fall in the same place.

| Model | Classification: $F_1$ | | | | Retrieval | | | |
|---|---|---|---|---|---|---|---|---|
| | Form | | Content | | Form: $F_{sim}$ | | Content: $C_{sim}$ | |
| | $z_f \uparrow$ | $z_c \downarrow$ | $z_c \uparrow$ | $z_f \downarrow$ | $z_f \uparrow$ | $z_c \downarrow$ | $z_c \uparrow$ | $z_f \downarrow$ |
| $L_{ae}$ | 0.73 | 0.96 | 0.58 | 0.73 | 0.57 | 0.95 | 0.85 | 0.70 |
| $L_{form}$ | 0.98 | 0.29 | 0.62 | 0.62 | 0.70 | 0.90 | 0.93 | 0.55 |
| $L_{mot}$ | 0.98 | 0.65 | 0.92 | 0.59 | 0.65 | 0.90 | 0.98 | 0.63 |
| $L_{proxy}$ | 0.98 | 0.23 | 0.92 | 0.60 | 0.70 | 0.85 | 0.99 | 0.54 |
| $L_{para}$ | 0.95 | 0.28 | 0.80 | 0.70 | 0.68 | 0.93 | 0.98 | 0.55 |
| $L_{para_f}$ | 0.98 | 0.21 | 0.75 | 0.63 | 0.69 | 0.87 | 0.97 | 0.54 |

Table 2.8: Classification and Retrieval scores that measure the quality of disentanglement of information for each of our models, evaluated on the PersonageNLG dataset

## 2.C More Results

### 2.C.1 Detailed Disentanglement Evaluation

In Table 2.8, we present a more detailed evaluation on the disentanglement metrics for our models. Here, the Classification column presents both same-aspect and cross-aspect $F_1$ scores. Higher scores for the former and lower scores for the latter indicate better disentanglement.

We notice that form information is not effectively removed from the content representations, as evidenced by the higher $F_{sim}$ scores for the content vectors $z_c$. This is a consequence of the weaker label-based proxy used for style, as opposed to the Meaning Representation-based attribute proxy for content.

### 2.C.2 Style Transfer Outputs

Table 2.9 shows sample outputs from the style transfer experiments on PersonageNLG. The model used is the best performing proxy-based model $L_{proxy}$, with motivational and adversarial losses for both style and content. Two paraphrases with different styles are first encoded into their form and content vectors. The output is generated by passing the form vector of the first sentence and the content vector of the second to the decoder. We see that the model transfers the form attributes quite well across the inputs, but content attributes are not retained perfectly.

| | |
|---|---|
| **Input (Style A)** | nameVariable is near nearVariable pal, nameVariable is a restaurant and it isn't family friendly, also the rating is average, you know! |
| **Target (Style B)** | You want to know more about nameVariable? Yeah, it isn't rather family friendly with an average rating, also it is sort of near nearVariable, also it is a restaurant, you see? |
| **Output (Style A → Style B)** | You want to know more about nameVariable? Oh it is sort of near nearVariable, also it is a restaurant, also it isn't family friendly, you see |
| **Input** | nameVariable is moderately priced, also it's in riverside. It is near nearVariable. It is a pub. it's an Italian restaurant. oh God basically, nameVariable is kid friendly. |
| **Target** | Yeah, err... I am not sure. nameVariable is an Italian place near nearVariable in riverside, damn kid friendly and moderately priced and nameVariable is a pub. |
| **Output** | Yeah, I am not sure. nameVariable is darn moderately priced in city centre near nearVariable, also it is a coffee shop, also it isn't kid friendly |

Table 2.9: Sample style transfer outputs for the best performing proxy-based model, $L_{para}$, on the PersonageNLG Dataset.

# Chapter 3

# Semantics: Understanding Closeness in Meaning

This work was published at EACL 2023, as follows:
Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What Makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

**Author Contributions:** Mohamed and Saif conceived of the initial idea for the project. Mohamed collected the data and led the annotation effort to create the final dataset (Sections 3.4, 3.5). Krishnapriya performed the experiments related to part-of-speech analysis (Section 3.6) , and experiments on the finetuning and evaluation of pretrained language models for estimating semantic relatedness (Section 3.7.1). All three authors contributed to the writing of the paper. Saif supervised the entirety of the project.

## 3.1 Introduction

The degree of semantic relatedness of two units of language has long been considered fundamental to understanding meaning. Additionally, automatically determining relatedness has many applications such as question answering and summarization. However, prior NLP work has largely focused on semantic similarity, a subset of relatedness, because of a lack of relatedness datasets. In this paper, we introduce a dataset for Semantic Textual Relatedness, *STR-2022*, that has 5,500 English sentence pairs

| | | |
|---|---|---|
| **Pair 1** | a. | *There was a lemon tree next to the house.* |
| | b. | *The boy enjoyed reading under the lemon tree.* |
| **Pair 2** | a. | *There was a lemon tree next to the house.* |
| | b. | *The boy was an excellent football player.* |

Table 3.1: Most people will agree that the sentences in pair 1 are more related than the sentences in pair 2.

manually annotated using a comparative annotation framework, resulting in fine-grained scores. We show that human intuition regarding relatedness of sentence pairs is highly reliable, with a repeat annotation correlation of 0.84. We use the dataset to explore questions on what makes sentences semantically related. We also show the utility of STR-2022 for evaluating automatic methods of sentence representation and for various downstream NLP tasks.

This work was led by Mohamed Abdalla, an alumnus of the Toronto CL group, and was carried out in collaboration with Saif Mohammad at the National Research Council of Canada. My individual contributions are in understanding the contributions of different parts-of-speech to the semantic relatedness between sentence pairs (Section 3.6, RQs 2, 3, and 4), and in training sentence representation models to better capture this phenomenon (Section 3.7.1).

Sections 3.4 and 3.5 were written by Mohamed Abdalla and are presented here in a paraphrased and compressed form. Sections 3.6, 3.7, and the appendices were a joint effort between myself and Mohamed Abdalla. The remaining sections of the paper were collaboratively written by all three authors of this work.

Our dataset, data statement, and annotation questionnaire can be found at: `https://doi.org/10.5281/zenodo.7599667`.

## 3.2 Introduction

The semantic relatedness of two units of language is the degree to which they are close in terms of their meaning (Mohammad, 2008; Mohammad and Hirst, 2012). The linguistic units can be words, phrases, sentences, etc. Though our intuition of semantic relatedness is dependent on many factors such as the context of assessment, age, and socio-economic status (Harispe et al., 2015), it is argued that a consensus can usually be reached for many pairs (Harispe et al., 2015). Consider the two sentence pairs in Table 3.1. Most speakers of English will agree that the sentences in the first pair are closer in meaning to one another than those in the second. When judging the semantic relatedness between two sentences, humans generally look for commonalities in meaning: whether they are on the same topic, express the same view, originate from the same time period, one elaborates on (or follows from) the other, etc.

The semantic relatedness of two units of language has long been considered fundamental to understanding meaning (Halliday and Hasan, 1976; Miller and Charles, 1991); given how difficult it has been to define meaning, a natural approach to get at the meaning of a unit is to determine how close it is to other units. Thus, unsurprisingly, automatically determining relatedness has many applications such as question answering, text generation, and summarization (more discussion in §3.8).

However, prior NLP work has focused on semantic similarity (a small subset of semantic relatedness), largely because of a dearth of datasets on relatedness. The few relatedness datasets that exist are only for word pairs (Rubenstein and Goodenough, 1965; Radinsky et al., 2011) or phrase pairs (Asaadi et al., 2019). Further, most existing datasets were annotated, one item at a time, using coarse rating labels such as integer values between 1 and 5 representing coarse degrees of closeness. It is well documented that such approaches suffer from inter- and intra-annotator inconsistency, scale region bias, and issues arising due to the fixed granularity (Presser and Schuman, 1996). Further, the notions of *related* and *unrelated* have fuzzy boundaries. Different people may have different intuitions of where such a boundary exists. Finally, for some tasks, it is more appropriate to train on a dataset of relatedness than similarity. (§3.3.1 discusses how relatedness and similarity are different).

In this chapter, we present the first manually annotated dataset of sentence–sentence semantic relatedness. It includes fine-grained scores of relatedness from 0 (least related) to 1 (most related) for 5,500 English sentence pairs. The sentences are taken from diverse sources and thus also have diverse sentence structures, varying amounts of lexical overlap, and varying formality.

The relatedness scores were obtained using a *comparative* annotation schema: two (or more) items are presented together and the annotator has to determine which is greater with respect to the metric of interest. Since annotators are making relative judgments, the limitations discussed earlier for rating scales are greatly mitigated. Importantly, such annotations do not rely on arbitrary boundaries between arbitrary categories such as "strongly related" and "somewhat related".

We use the relatedness dataset to explore:

1. To what extent do speakers of English intuitively agree on the relatedness of sentences? (§3.5)

2. What makes two sentences more related? (§3.6)

3. How well do existing approaches of sentence representation capture semantic relatedness (by placing related sentence pairs closer to each other in vector space)? (§3.7)

4. How can an improved annotation schema to capture relatedness benefit other NLP tasks? (§3.8)

We refer to our dataset as *STR-2022*, and the task of predicting relatedness between sentences as the *Semantic Textual Relatedness (STR)* task. Data, data statement, and annotation questionnaire are made available[1]

## 3.3 Related Work and Our Approach to Annotating for Semantic Relatedness

The three subsections below discuss key ideas from past work on annotating relatedness and similarity, existing datasets, and comparative annotation, respectively. Notably, each of these subsections also discusses how relevant past work has influenced our approach to data annotation.

### 3.3.1 Annotating Relatedness and Similarity

Semantic relatedness and semantic similarity are two concepts related to closeness of meaning. Two terms are considered semantically similar if there is a synonymy, hyponymy, or troponymy relation between them; examples include *doctor–physician* and *mammal–elephant*. Two terms are considered to be semantically related if there is any lexical semantic relation at all between them. Thus, all similar pairs are also related, but not all related pairs are similar. For example, *surgeon–scalpel* and *tree–shade* are related but not similar.

Analogous to term pairs, two sentences are considered semantically similar when they have a paraphrasal or entailment relation. Determining such an equivalence of meaning is useful in NLP tasks such as text summarization and plagiarism detection. Semantic Relatedness, however, accounts for all of the commonalities that can exist between two sentences (Halliday and Hasan, 1976; Morris and Hirst, 1991). For example, the sentences in Table 3.1 Pair 1 are highly related, but they are not paraphrases

[1] doi.org/10.5281/zenodo.7599667 or https://huggingface.co/datasets/vkpriya/str-2022 or https://github.com/Priya22/semantic-textual-relatedness.

or entailing. This expands the scope of the measure to include aspects such as the relatedness between their topics, their styles, stances, and so on.

However, because semantic relatedness involves innumerable classical and ad-hoc semantic relationships, it is markedly more complex than semantic similarity, and there are no widely agreed upon linguistic theories or guidelines for judging relatedness. This presents a challenge for gathering annotations; one can either: (i) construct their own codified instructions on how to judge semantic relatedness under various scenarios (e.g., overlapping sentence structure, relatedness of topic, etc.), at the risk of artificially over-simplifying the task or (ii) abstain from explicitly and comprehensively defining relatedness for numerous types of sentence pairs, relying instead on a simple description of relatedness, a few examples, and framing the task in relative terms.[2] In this work, we chose the latter. This allows us to: (i) determine the extent to which human intuition of relatedness is reliable and (ii) use the resulting dataset to empirically determine what makes sentences semantically related.

### 3.3.2 Existing Relatedness and Similarity Data

Existing datasets created for sentence pair similarity (e.g., STS (Agirre et al., 2012, 2013, 2014, 2015, 2016), MRPC (Dolan and Brockett, 2005), and LiSent (Li et al., 2006)) ask annotators to choose among coarse similarity labels. This leads to information loss and makes annotation difficult because distinctions between categories are often not clear; for example, the STS 2012–2016 questionnaires ask annotators to make the distinction between *2: not equivalent but share some details* and *1: not equivalent, but are on the same topic*, which is often not straightforward. Further, despite claiming to determine semantic similarity, the descriptions of categories 1 and 2 incorporate aspects of semantic relatedness — an amalgamation muddying the waters with respect to the phenomenon being annotated. Such an amalgamation is also seen in the SICK (Marelli et al., 2014) dataset which combines a labeling scheme from STS with those about entailment and contradiction. These datasets have helped make progress in the field, but there is a need for relatedness datasets obtained strictly from relatedness judgments as opposed to a hybrid involving artificially created categories for similarity and entailment.

For our annotations, we avoid fuzzy ill-defined categories, and rely instead on the intuitions of fluent English speakers to judge **relative rankings** of sentence pairs by relatedness.

[2] Recall that for Table 3.1, we were able to judge relative relatedness without explicit instruction on how to judge relatedness.

### 3.3.3 Comparative Annotations

The simplest form of comparative annotations is paired comparisons (Thurstone, 1927; David, 1963). Annotators are presented with pairs of examples and asked to choose which item is greater with respect to the property of interest (relatedness, sentiment, etc.). The choices are then used to generate an ordinal ranking of items. Paired comparison avoids a number of biases, but it requires a large number of annotations ($N^2$, where $N = \#$ items).

Best–Worst Scaling (BWS) is a comparative annotation schema that builds on pairwise comparisons and requires fewer labels (Louviere and Woodworth, 1991). Annotators are given $n$ items at a time (for our work, $n = 4$ and an *item* is a pair of sentences). They are instructed to choose the best (i.e., most related) and worst (i.e., least related) item. Annotation for each 4-tuple provides us with five pairwise inequalities. For example if $a$ is marked as most related and $d$ as least related, then we know that $a > b$, $a > c$, $a > d$, $b > d$, and $c > d$. These inequalities can be used to calculate real-valued scores, and thus an ordinal ranking of items, using a simple counting mechanism (Orme, 2009; Flynn and Marley, 2014): the fraction of times an item was chosen as the best (most related) minus the fraction of times the item was chosen as the worst (least related). Given $N$ items, reliable scores are obtainable from about $2N$ 4-tuples (Kiritchenko and Mohammad, 2016, 2017).

## 3.4 Creating STR-2022

Dataset creation included several steps: curating sentence pairs for annotation, designing the questionnaire, crowdsourcing annotations, and aggregating the annotations to obtain relatedness scores. In brief, we source our sentence pairs from already existing sentence-level datasets in NLP: either paired datasets annotated for properties like entailment or semantic similarity, or individual sentences, potentially annotated for properties like sentiment or stance. We selected sentence pairs with varying amounts of lexical overlap because randomly sampling sentence pairings would result in mostly unrelated sentences. This also allowed us to systematically study the impact of lexical overlap on semantic relatedness. Table 3.2 summarizes key details of the sentence pairs in STR-2022.

| Types of Pairs | Key Attributes | # pairs |
|---|---|---|
| 1. Formality | paraphrases, style | |
| Formality_pp | paraphrases, differ in style | 300 |
| Formality_r | random pairs | 700 |
| 2. Goodreads | reviews, informal | 1000 |
| 3. ParaNMT | automatic paraphrases | |
| ParaNMT_pp | automatic paraphrases | 450 |
| ParaNMT_r | random pairs | 300 |
| 4. SNLI | captions of images | 750 |
| 5. STS | have similarity scores | 250 |
| 6. Stance | tweet pairs with same hashtag, less grammatical | 750 |
| 7. Wikipedia | formal | |
| Wiki_pp | paraphrases, formal | 500 |
| Wiki_r | random pairs, formal | 500 |
| ALL | | 5500 |

Table 3.2: Summary of sentence pair types in STR-2022.

### 3.4.1 Annotating For Semantic Relatedness

From the list of 5,500 sentence pairs, we generated 11,000 unique 4-tuples (each 4-tuple consists of 4 distinct sentence pairs) such that each sentence pair occurs in around eight 4-tuples.[3]

In our framing of the task, we did not use detailed or technical definitions; rather, we provided brief and easy-to-follow instructions, gave examples, and encouraged annotators to rely on their intuitions of the English language to judge relative closeness in meaning of sentence pairs (similar to Asadi et al.'s (2019) work on bigrams). Annotators were asked to judge the "closeness in meaning of sentence pairs". Inspired by early work in linguistics on cohesion in text (Halliday and Hasan, 1976), we also specified that: "Often sentence pairs that are more specific in what they share tend to be more related than sentence pairs that are only loosely about the same topic" and "If a sentence has more than one interpretation, consider that meaning which is closest to the meaning of the other sentence in the pair." This is in line with application scenarios where often relatedness is to be determined between sentences from the same document.

[3] The tuples were generated using the BWS scripts provided by Kiritchenko and Mohammad (2017): http://saifmohammad.com/WebPages/BestWorst.html.

| **Statistic** | **Value** |
|---|---|
| # Sentence Pairs | 5,500 |
| # Tuples | 11,000 |
| # Annotations Per Tuple | 8 |
| # Annotations | 21,936 |
| # Annotators | 389 |
| **SHR** | 0.84 |

Table 3.3: Annotation statistics of STR-2022. SHR = split-half reliability (as measured by Spearman correlation).

### Crowdsourcing Annotations

We used Amazon Mechanical Turk (MTurk) for obtaining annotations. Each 4-tuple (also referred to as a question) in our MTurk task consists of four sentence pairs. Annotators are asked to choose the (a) most-related, and (b) least-related sentence pairs from among these four options. Each question is annotated by two MTurk workers.[4]

For quality control, the task was open only to fluent speakers of English and those MTurk workers with an approval rate higher than 98%. Further, we inserted "Gold Standard" questions at regular intervals in the task. These questions were manually annotated by all the authors, and had high agreement scores. If an annotator gets a gold question wrong, they are immediately notified and shown the correct answer. This has several benefits, including keeping the annotator alert and clearing any misunderstandings about the task. Those who scored less than ∼70% on the gold questions were stopped from answering further questions and were paid for their work. All their responses were discarded.

### Annotation Aggregation

We aggregate information from various responses by using the counting procedure discussed in §3.3.3. Since relatedness is a unipolar scale, the resulting relatedness score was linearly transformed to fit within a 0–1 scale of increasing relatedness. Appendix Table 3.8 presents sample sentence pairs from each data source.

Figure 3.1 presents a histogram of relatedness scores for STR-2022. Observe that each of the subsets covers a wide range of relatedness scores; that the lexical overlap sampling strategy has resulted in a wide spread of relatedness scores; and that supposed paraphrases are spread across much of the right half of the relatedness scale.

[4] Pilot studies showed that this results in reliable scores.

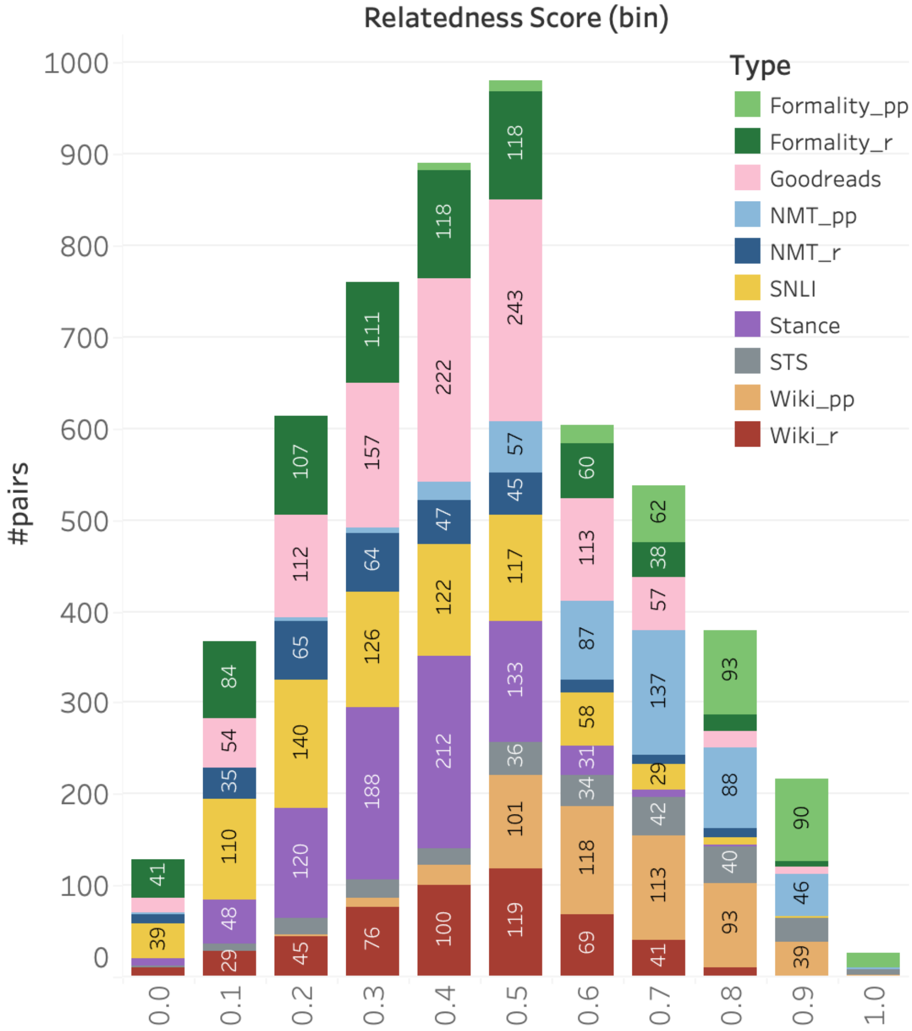


Figure 3.1: Histogram of STR-2022 relatedness scores.

## 3.5 Reliability of Annotations

For annotations producing real-valued scores, a commonly used measure of quality and reliability is *split-half reliability* (SHR) (Cronbach, 1951; Kuder and Richardson, 1937). SHR is a measure of the degree to which repeating the annotations would result in similar relative rankings of the items. To measure SHR, annotations for each 4-tuple are split into two bins. The annotations for each bin are used to produce two different independent relatedness scores. Next, the Spearman correlation between the two sets of scores is calculated — a measure of the closeness of the two rankings. If the annotations are reliable then there should be a high correlation. This process is repeated 1000 times and the correlation scores are averaged.

As shown in Table 3.3, STR-2022 has an SHR of 0.84—signifying high annotation reliability. This is a key result of this paper. Recall that our annotation guidelines did not hard code the various scenarios of sentence pair types and how they should be judged, but rather were designed to elicit how native speakers of English naturally judge relatedness. The high reliability of annotations, despite this, shows that speakers of a language are inherently consistent in their judgments of relatedness. It also validates our approach as a way to produce high-quality relatedness datasets; which, in turn, can be used to study the mechanisms underpinning relatedness (as we explore in the next section).

### 3.5.1 STR vs STS

We also conducted experiments to assess fine-grained rankings of common sentence pairs as per our relatedness scores and as per STS's similarity scores. For each of the sets of 50 sentence pairs taken from STS (with scores in (0–1], (1–2], etc.), we calculated the Spearman correlation between the rankings by similarity and rankings by relatedness. We found that the correlations are only 0.25 (weak) and 0.19 (very weak) for the bins of (1,2] and (3,4], respectively, and only about 0.49 (moderate) for the bins of (2,3] and (4,5]. Overall, this shows that the fine-grained ranking of items in the STS dataset by similarity differ considerably from that of the STR dataset.

## 3.6 What Makes Sentences More Semantically Related?

The availability of a dataset with human notions of semantic relatedness allows one to explore fundamental aspects of meaning: for example, what makes two sentences more related? In this section, we examine some basic questions. On average, to what extent is the semantic relatedness of a sentence pair impacted by presence of:

- **RQ1:** identical words (lexical overlap)?
- **RQ2:** related words?
- **RQ3:** related words of the same part of speech?
- **RQ4:** related subjects, related objects?

### 3.6.1 Method

To explore the questions above, we[5] computed relevant measures for Q2 through Q4 (lexical overlap was by the first author) for each sentence pair in our dataset. We then calculated the correlations of these scores with the gold relatedness scores.

**Lexical Overlap**. A simple measure of lexical overlap between two sentences X and Y is the Dice Coefficient (the number of unique unigrams occurring in both sentences, adjusted by their lengths):

$$\frac{2 \times |\, unigram(X) \cap unigram(Y)\,|}{|\, unigram(X)\,| + |\, unigram(Y)\,|} \tag{3.1}$$

**Related Words:** We averaged the embeddings for all the tokens in a sentence and computed the cosine between the averaged embeddings for the two sentences in a pair. This roughly captures the relatedness between the terms across the two sentences.[6] Token embeddings were taken from Google's publicly released Word2Vec embeddings trained on the Google News corpus (Mikolov et al., 2013a).

**Related Words with same POS:** The same procedure was followed as for Q2, except that only the tokens for one part of speech (POS) at a time were considered. We determined the part-of-speech of the tokens using spaCy (Honnibal et al., 2020).[7]

[5] I

[6] Other ways to estimate relatedness between sets of words across two sentences may also be used.

[7] We used the simple (coarse-grained) UPOS part-of-speech tags: https://universaldependencies.org/docs/u/pos/

| **Question** | **Spearman** | **# pairs** |
|---|---|---|
| Q1. Lexical overlap | 0.57 | 5500 |
| Q2. Related words - All | 0.61 | 5500 |
| Q3a. Related words - per POS | | |
| PROPN | 0.50 | 1907 |
| NOUN | 0.45 | 4746 |
| ADJ | 0.36 | 2236 |
| VERB | 0.31 | 3946 |
| PRON | 0.30 | 1800 |
| ADV | 0.28 | 1147 |
| AUX | 0.25 | 2069 |
| ADP | 0.23 | 2476 |
| DET | 0.20 | 3265 |
| Q3b. Related words - per POS group | | |
| Noun Group | 0.60 | 5478 |
| Verb Group | 0.32 | 4999 |
| ADJ Group | 0.29 | 4584 |
| Q4. Related Subjects and Objects | | |
| Subject | 0.29 | 1611 |
| Object | 0.43 | 1618 |

Table 3.4: Correlation between features and the relatedness of sentence pairs. A rule of thumb for interpreting the numbers: 0–0.19: very weak; 0.2–0.39: weak; 0.4–0.59: moderate; 0.6–0.79: strong; 0.8–1: very strong.

**Related Subjects and Related Objects:** For Q4, which examines the importance of different parts of a sentence, we employ the same process as Q2, except that for a given sentence: only tokens marked as subject are averaged; and only tokens marked as object are averaged. We use the packages spaCy (Honnibal et al., 2020) and Subject Verb Object Extractor (de Vocht, 2020) to determine all tokens that are the subject and object.

### 3.6.2 Results

Table 3.4 shows the results. Row Q1 shows that simple word overlap obtains a correlation of 0.57 (considered to be at the high end of weak correlation). Figure 3.2 is a scatter plot where the x-axis is the word overlap score, the y-axis is the relatedness score, and each dot is a sentence pair. Observe that a number of pairs fall along the diagonal; however, there are also a large number of pairs along the top-left side of this diagonal. This suggests that even though STR-2022 has pairs

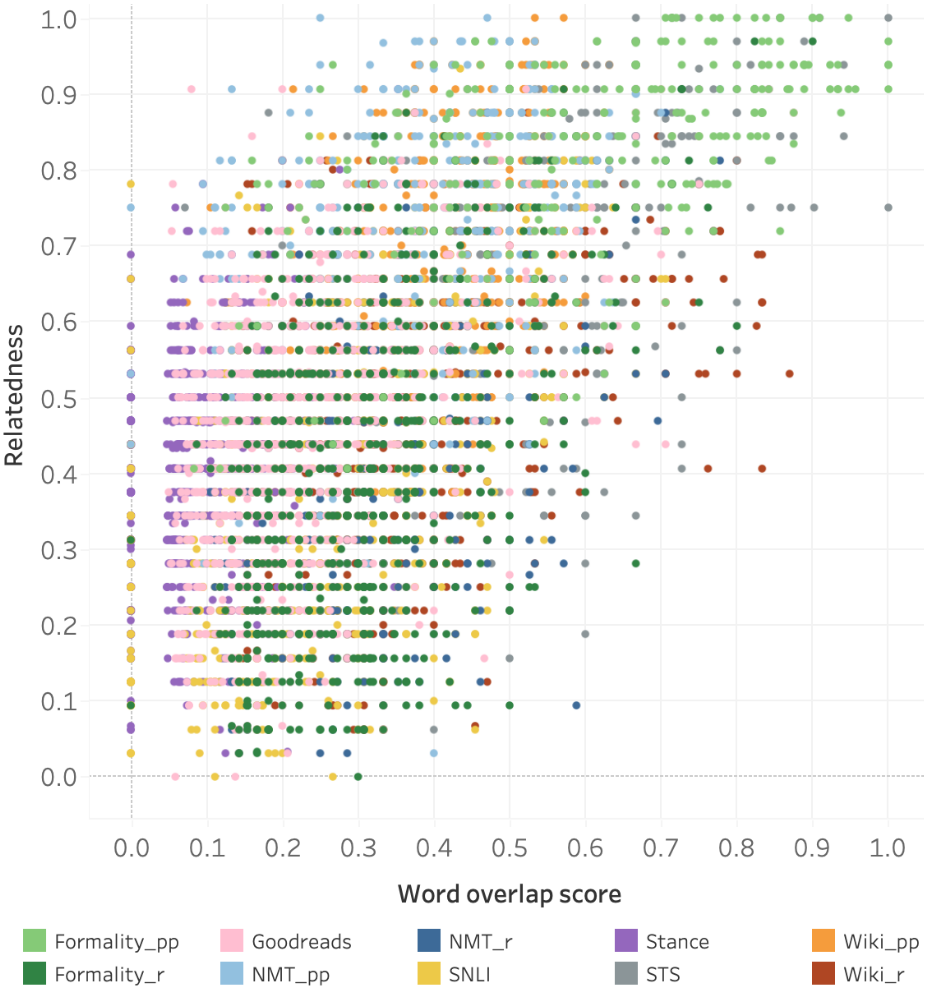


Figure 3.2: Scatter plot showing the relationship between lexical overlap and semantic relatedness of sentence pairs. Each dot in the plot is a sentence pair.

where the relatedness increases linearly with the amount of word overlap, there are also a number of pairs where a small amount of word overlap results in substantial amount of relatedness. The sparse bottom-right side of the plot indicates that it is rare for there to be substantial word overlap, and yet very low relatedness. On average, occurrence of related words across a sentence pair leads to slightly higher relatedness scores than lexical overlap (row Q2).

The Q3a rows in Table 3.4 show correlations for related tokens of a given part of speech.[8] (The rows are in order from highest to lowest correlation.) Observe that proper nouns (PROPN) and nouns have the highest numbers. It is somewhat surprising that related verbs do not contribute greatly to semantic relatedness; they have similar correlations as pronouns and adverbs, and markedly lower than adjectives and

[8] Only those POS tags that occur in both sentences of a pair in more than 10% of the pairs are considered (> 550 pairs).

nouns. Not surprisingly, determiners (DET) are at the lower end of weak correlation.

The Q3b rows show correlations of coarse POS categories: NOUN Group (NOUN, PRON, PROPN), VERB Group (VERB, AUX), and ADJ Group (ADJ, ADP, ADV). We see that presence of related nouns in a sentence pair impacts semantic relatedness much more than any other POS group.

Since related nouns were found to be especially important, we also wanted to determine what impacts overall relatedness more: the presence of related nouns in the subject position or in the object position. Q4 rows show that, on average, related objects lead to markedly higher sentence-pair relatedness than related subjects.

In order to examine whether lexical overlap and some POS are less or more relevant in low or high relatedness pairs, we repeated the experiment of Table 3.7, only for pairs with relatedness scores $< 0.5$, and separately, only for pairs with scores $\geq 0.5$. We find that for the $< 0.5$ relatedness pairs, only the existence of related proper nouns across sentence pairs has moderate correlation with the semantic relatedness of sentences; the correlation is weak for nouns, and close to 0 for all other parts of speech. The notable importance of related proper nouns and nouns is likely because they indicate a common topic, person, or object being talked about in both sentences — making the two sentence pairs related. For the $\geq 0.5$ relatedness pairs, the correlations are weak for most POS; highest for nouns; and the gap between nouns and adjectives, adverbs, and verbs is reduced. Lexical overlap in general has a much higher correlation for the $\geq 0.5$ relatedness pairs than the $< 0.5$ pairs. Detailed results are in Appendix 3.A.

## 3.7 Evaluating Sentence Representation Models using STR-2022

Since STR-2022 captures a wide range of fine-grained relations that exist between sentences, it is a valuable asset in evaluating sentence representation and embedding models. Essentially, predicting semantic relatedness is treated as a regression task, where first, using various unsupervised and supervised approaches described in the two sub-sections below, we represent each sentence as a vector. We use the cosine similarity between the vectors as a prediction of their semantic relatedness. We use the Spearman correlation between the prediction and gold relatedness scores to measure the goodness of the relatedness predictions (and in turn of the sentence representation).

The experiments below (unless otherwise specified) all involve 5-fold cross-validation

| **Model** | **Spearman** |
|---|---|
| *Baseline* | |
| 1. Lexical overlap (Dice) | 0.57 |
| *Unsupervised, Static Embeddings* | |
| 2. Word2Vec (mean, Googlenews) | 0.60 |
| 3. Word2Vec (max, Googlenews) | 0.54 |
| 4. GloVe (mean, Common Crawl) | 0.49 |
| 5. GloVe (max, Common Crawl) | 0.56 |
| 6. GloVe (mean, 200_Twitter) | 0.44 |
| 7. GloVe (max, 200_Twitter) | 0.48 |
| 8. Fasttext (mean, Common crawl) | 0.29 |
| 9. Fasttext (max, Common crawl) | 0.24 |
| *Unsupervised, Contextual Embeddings* | |
| 10. BERT-base (mean) | 0.58 |
| 11. BERT-base (max) | 0.55 |
| 12. BERT-base (cls) | 0.41 |
| 13. RoBERTa-base (mean) | 0.48 |
| 14. RoBERTa-base (max) | 0.47 |
| 15. RoBERTa-base (cls) | 0.41 |
| *Supervised (Fine-tuning on portions of STR-2022)* | |
| 16. BERT-base (mean) | 0.82 |
| 17. RoBERTa-base (mean) | 0.83 |

Table 3.5: Average correlation between human annotated relatedness of sentence pairs and the cosine distance between their embeddings across the CV runs.

(CV) on STR-2022. We report the average of the Spearman correlations across the folds. Note that even for models that do not require training (e.g., Dice score), to enable direct comparisons with trained methods, we evaluate their performance on each test fold independently and report the average of the correlations across folds.

### 3.7.1 Do Unsupervised Embeddings Capture Semantic Relatedness?

We first explore unsupervised approaches to sentence representation where the embedding of a sentence is derived from that of its constituent tokens. The token embedding can be of two types:

- **Static Word Embeddings:** We tested three popular models: Word2Vec (Mikolov et al., 2013b), GloVe Pennington et al. (2014), and Fasttext (Grave et al., 2018).
- **Contextual Word Embeddings:** We tested pretrained contextual embeddings from BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). We use the bert-base-uncased and roberta-base models from the HuggingFace library.[9]

[9] https://huggingface.co

| | **Dice** | **SBERT(RoBERTa)** | |
|---|---|---|---|
| | CV | CV | LOO CV |
| STS | 0.60 | 0.79 | 0.82 |
| SNLI | 0.53 | 0.80 | 0.77 |
| Stance | 0.20 | 0.49 | 0.39 |
| Goodreads | 0.44 | 0.73 | 0.70 |
| Wiki | 0.48 | 0.79 | 0.75 |
| Formality | 0.69 | 0.86 | 0.83 |
| ParaNMT | 0.44 | 0.80 | 0.79 |

Table 3.6: Breakdown of average test-fold correlations for each source: (a) using lexical overlap (Dice), (b) using SBERT and some in-domain data for fine-tuning (in addition to data from other domains), and (c) using SBERT and only out-of-domain data for fine-tuning (LOO CV). CV: cross-validation. LOO: leave-one-out.

We obtain sentence embeddings by both mean-pooling and max-pooling the token embeddings from the final layer. For the contextual embeddings, we also explore using the embedding of the classification token ([CLS]).

Table 3.5 shows the results. As baseline, we include how well simple lexical overlap (Dice score) predicts relatedness (row 1). Observe that mean-pooling with word2vec (row 2) obtains slightly higher correlation than the baseline, but the majority of the static embedding models fail to obtain better correlations (rows 3–9). The contextual embeddings from BERT and RoBERTa do not perform better than the word2vec embeddings (rows 10–15). Overall, the unsupervised methods leave much room for improvement.

### 3.7.2 Do Supervised Embeddings Capture Semantic Relatedness?

We[10] now evaluate the performance of BERT-based models on STR-2022 when formulated as a *supervised* regression task. We use the S-BERT cross-encoder framework of Reimers and Gurevych (2019b), and apply mean-pooling on top of the token embeddings of the final layer to obtain sentence embeddings. The model is trained using a cosine-similarity loss — the cosine between the embeddings of a sentence pair is compared to the gold semantic relatedness scores to obtain the Mean Squared Error (MSE) loss for each datapoint.

Table 3.5 rows 16 and 17 show the results: fine-tuning on STR-2022 leads to considerably better relatedness scores.

[10] I

### Impact of Domain on Fine-Tuning

The results above show that fine-tuning is critical for better sentence representation. However, it is well-documented that the domain of the data can have substantial impact on results; especially when quite different from the training data. With the inclusion of data from various domains in STR-2022 (Table 3.2), one can systematically explore performance on individual domains, as well as the extent to which performance may drop if no training data from the target domain is included for training.

Table 3.6 shows the results. The RoBERTa CV column shows a breakdown of results by source (domain). Essentially, these are results for the scenario where some portion of in-domain data is included in the training folds (along with data from other domains), and the system correlations are determined only on the test fold's target domain pairs. Observe that performance on most domains is comparable to each other.

The LOO CV column shows correlations with a leave-one-out cross-validation setup: no in-domain training data is used and system correlations are determined only for the target domain pairs. Observe that this leads to drops in scores for all domains except STS. However, the drop is small; and scores are still much higher than the lexical overlap (Dice CV) baseline. This suggests that the diversity of data in the remaining subsets is useful in overcoming a lack of in-domain training data.

## 3.8 Utility of Semantic Relatedness and STR-2022 in Downstream NLP Tasks

Semantic relatedness is central to textual coherence and narrative structure. Often, sentences in a document are not paraphrases, entailments, or similar, but rather semantically related to each other. This need for continuity of meaning has long been identified as a crucial component of language Halliday and Hasan (1976); Morris and Hirst (1991). Thus, when generating a summary or a response to a question, systems must choose sentences that are *not* paraphrases or entailments of each other, but yet suitably semantically related. Therefore, being able to judge both similarity and relatedness is crucial.

Since we made STR-2022 publicly available, it has already been used in some projects. Notable among these is Wang et al. (2022), who propose a new intrinsic evaluation method, *EvalRank*, that focuses on local neighborhoods (how well systems identify close neighbors, rather than how well they rank the full set of pairs). Using STR-2022, they are able to obtain markedly higher correlations between performance scores on the intrinsic evaluation and performance on downstream tasks (seven NLP tasks including NLI, question classification, caption retrieval, and sentiment analysis). Their ablation study demonstrates that using STS instead of STR-2022 decreases performance up to 10 points, leading them to conclude that STR-2022 is particularly useful in generating sentence embeddings for downstream tasks.

## Limitations

In our experiments, we used the most common methods for sentence representations (e.g., mean-pooling and max-pooling of traditional and contextual word embeddings). However, there may exist other embeddings which are better suited for predicting semantic relatedness (e.g., other order-aware embeddings). Expanding the set of embedding techniques tested using our dataset may yield different results and provide us a stronger understanding of the effects of different representation techniques. Furthermore, while we explored the impact of some sentence-pair features such as lexical overlap, POS, and some aspects of sentence structure (subject and object) on semantic relatedness, we did not explore the impacts of other features such as logicality and common sense reasoning on relatedness. These remain interesting directions for future work.

This paper respects existing intellectual property by making use of only publicly and freely available datasets. The crowd-sourced task was approved by our Institutional Research Ethics Board. The annotators were based in the United States of America and were paid the federal minimum wage of $7.25 per hour. Our annotation process stored no information about annotator identity and as such there is no privacy risk to them. The individual sentences selected did not have any risks to privacy either (as evaluated by manual annotation of the sentences). Models trained on this dataset may not generalize to external datasets gathered from different populations. Knowledge about language features may not generalize to other languages.

Any dataset of semantic relatedness entails several ethical considerations. We recommend careful reflection of ethical considerations relevant for the specific context of deployment when using STR-2022.

| | | **Spearman** | |
|---|---|---|---|
| **Question** | 0–1 pairs | <0.5 pairs | ≥0.5 pairs |
| Q1. Lexical overlap | 0.57 | 0.14 | 0.52 |
| Q2. Related words - All | 0.61 | 0.14 | 0.50 |
| Q3a. Related words - per POS | | | |
| PROPN | 0.50 | 0.34 | 0.26 |
| NOUN | 0.45 | 0.18 | 0.37 |
| ADJ | 0.36 | 0.04 | 0.35 |
| VERB | 0.31 | 0.03 | 0.31 |
| PRON | 0.30 | 0.01 | 0.30 |
| ADV | 0.28 | 0.04 | 0.35 |
| AUX | 0.25 | 0.03 | 0.20 |
| ADP | 0.23 | 0.07 | 0.22 |
| DET | 0.20 | 0.03 | 0.19 |
| Q3b. Related words - per POS group | | | |
| Noun Group | 0.60 | 0.34 | 0.41 |
| Verb Group | 0.32 | 0.09 | 0.29 |
| ADJ Group | 0.29 | 0.04 | 0.32 |
| Q4. Related Subjects and Objects | | | |
| Subject | 0.29 | 0.00 | 0.32 |
| Object | 0.43 | 0.14 | 0.33 |

Table 3.7: Correlation between features and the relatedness of sentence pairs in STR-2022 when considering full relatedness range (0–1), only the pairs with relatedness $< 0.5$, and only the pairs with relatedness $\geq 0.5$.
Note: The 0–1 pairs column was shown earlier in Table 4. It is repeated here for ease of comparison.

## 3.9 Conclusion

We created STR-2022, the first dataset of English sentence pairs annotated with fine-grained relatedness scores. We used a comparative annotation method that produced a split-half reliability of 0.84. Thus, we showed that speakers of a language can reliably judge semantic relatedness. We used the dataset to explore several research questions pertaining to what makes two sentences more related. Finally, we used STR-2022 to evaluate the ability of sentence representation methods to embed sentences in vector spaces such that those that are closer to each other in meaning are also closer in the vector space. The dataset is made freely available; facilitating further research in semantic relatedness and sentence representation.

## 3.A Correlation of Features in Low and High Relatedness Sentence Pairs

As discussed in Section 5.2, in order to examine whether lexical overlap and some parts of speech are less or more relevant in low or high relatedness pairs, we repeated the experiment in Table 4, only for pairs with relatedness scores less than 0.5 and also for pairs with scores greater than 0.5. Table 3.7 shows the detailed correlation scores. See Section 5.2 for a discussion of the main trends.

## 3.B Sample Sentence Pairs from STR-2022

Table 3.8 presents sample sentence pairs from different domains.

| **Source** | **Sentence Pairs** | **STR score** |
|---|---|---|
| Formality_pp | *I think Taylor is really cute, but I hate his voice.*<br>*I think Taylor is SUPER cute...but I hate his voice.* | 1.000 |
| Wiki_pp | *It is sometimes referred to as the trunk.*<br>*Some people also call it the trunk.* | 0.969 |
| Goodreads | *I loved this short story - wish it were longer!*<br>*It was a quick read and part of me wished that it would go on a little longer.* | 0.844 |
| Wiki_r | *On August 2 , a tropical storm hit Northeastern Florida .*<br>*In early October , a hurricane caused damage and erosion to northeastern Florida .* | 0.625 |
| Stance | *So unfortunate #thebriefcase @cbs. Adoption isn't always the answer.*<br>*Just remember, there is a living family out there just waiting to #adopt your aborted baby.* | 0.562 |
| SNLI | *A woman in speaking in a theater.*<br>*deleon speaking into a mic.* | 0.406 |
| ParaNMT_pp | *Are you–are you going to tell every one?*<br>*will you say it now – all] of you?* | 0.334 |
| Formality_r | *i believe in american dreams ...*<br>*You are the woman of my dreams* | 0.219 |
| STS | *A person is riding a horse.*<br>*A woman is slicing potatoes.* | 0.062 |

Table 3.8: Sample sentence pairs from different domains in the STR-2022 dataset.

# Part II

# Literary Dialogism

# Outline

The computational analysis of stories and narratives has a rich history, and advances in language processing technologies and associated datasets have enabled large-scale quantitative studies of plot structure, characters and networks, and authorial style. The work in this section of the thesis presents new datasets and methods for a more nuanced understanding of literary texts, specifically full-length English-language novels from the period 1810-1950. Our focus is on the role that characters play in a novel — how they speak, interact, and evolve over the course of the narrative. I begin by introducing the Project Dialogism Novel Corpus, a dataset of 28 full-length novels annotated in entirety for various aspects of quotations and coreference within them, produced in collaboration with Professor Adam Hammond and his students from the Department of English, University of Toronto. PDNC allows us to perform a series of analyses on character representations, utterance styles, and emotional dynamics, which I expand on in the subsequent chapters. I also contribute towards enabling such research at a larger scale by characterizing the task of quotation attribution and training state-of-the-art models for the same. Our findings from these works indicate the feasibility and the importance of representing novels as a mix of diverse voices and perspectives expressed through it's characters, rather than a single, undifferentiated blob of text attributable to the writer.

In Chapter 5, I introduce the PDNC dataset and its annotation process. This was a hugely collaborative effort over multiple years led by Professor Hammond. I developed the software for the annotation process and handled the computational aspects of pre-processsing the texts, processing the annotations for disagreements, and post-processing them to obtain the final dataset. The selection of the texts, development of annotation guidelines (which is also a notable output of this project), and recruitment and training of the annotators was largely handled by Professor Hammond.

Chapter 6 describes our first step towards understanding differences in character voices. I begin with standard stylometric analyses of the utterances of various characters, as annotated in our dataset. We look at whether stylometric features are able to separate our utterances by speaker in straightforward classification and clustering setups. I also take advantage of the annotations for character and author gender (I clarify what gender refers to in this context in Section 4.2.1) to examine whether there are statistically significant differences in how different characters are addressed. Some of these results are encouraging, but we do not see any broad, generalizable claims being validated.

Chapter 7 tackles emotional variation more specifically. We look at emotional arcs of characters across narrative time using the framework of Utterance Emotion Dyanamics (UED), and compute measures of emotional narrative diversity for novels. We show that the emotion arcs of characters can be quite different from that of the narration, and the character arcs across novels show a higher range of similarities and dissimilarities than the arcs of narrations or the overall arcs of novels. These findings highlight the need for viewing novels as a composition of their characters.

In Chapter 8, I describe my work on the task of quotation attribution, where we seek to automatically identify the speaker of each quotation in a text. This is a vital step towards enabling research on characters and character voices at a much larger scale. I propose a modular set of sub-tasks and corresponding metrics that the quotation attribution task can be decomposed into, and evaluate a suite of tools for their performance in each sub-task. I use the insights from this study to improve the state-of-the-art pretrained attribution model, and also achieve a new state-of-the-art by fine-tuning the model on the PDNC corpus.

# Chapter 4

# Introduction

The two projects in the preceding section of the thesis looked broadly at the difficulties of defining and representing the style of, and the meaning conveyed by, texts. The focus was largely on the computational methods of doing so, and we therefore worked within the data frameworks of standard, pre-existing NLP datasets in this area. These datasets largely assume that the style of a text is well-represented and contained in the span of at most a couple of sentences.

Prior research in authorship attribution and profiling has, however, demonstrated that consistent stylistic indicators (the stylometric fingerprint) of author identity can only be reliably obtained with larger text spans (1000 tokens or more). Stylometric research has therefore been largely conducted in domains where such data can be easily obtained and attributed – essays and blog posts, emails, and literature.

The focus of this section of the thesis is on investigating a literary question of interest – the distinctiveness of character voices, or **dialogism**, in a novel – using computational techniques. We operationalize several ways of potentially answering this question with data-driven methods of text analysis, including identifying and defining various linguistic features and computational frameworks that can help us understand consistency and variation in character voice. Some of these frameworks are new, and others have been used in the past to analyze stylistic voice in other contexts, such as authorship attribution and demographic profiling. While there has been some work on individual character utterances within novels, it has been limited to data samples of at most 3–5 texts (we review related work in Section 4.3). We are among the first to conduct a (relatively) large-scale analysis of character voice and style in literary novels; as part of this effort, we also release the largest dataset to-date of characters and their utterances in this domain.

## 4.1 Computational Literary Analysis

Literature in general functions as a rich domain of experimentation for natural language processing research. It is an indisputable source of large amounts of text data, and is often explicitly associated with various socio-demographic meta-variables of interest, such as those of the author, the time-period of writing and publication, and topic and genre information. The long-form nature of the texts presents interesting technical challenges to the direct application of most NLP models, irrespective of the specifics of the task — up until very recently, text processing models were limited in their input size to at most a few hundred tokens. The very nature of the domain itself, containing a multitude of different fictional realities and worlds and perspectives within it, demands a level of reasoning and understanding far beyond what is required by, say, textbook-style informative texts.

The variety of narrative styles within literature, even if we limit ourselves to just fictional novels, makes it a particularly tricky domain to work with in NLP. Take the case of coreference resolution, for example. Characters in a novel are rarely referred to by a single name; these names can also change over the course of the narrative with the occurrence of various fictional events: the formation of new relations by marriage, a planned change in identity for purposes of deception, a situational assigned nickname, mistaken identities leading to a comedy of errors. Accurately tracking the mentions of a specific entity over an entire novel can therefore require complex reasoning abilities, and is quite tricky to achieve without manual input of some sort.

From a humanities point of view, literature is a place where one expects to find a rich set of indicators marking the social norms and manners of its time, as well as the first hints of subversions and deviations from those norms — it functions as a playing field for pushing the limits of creativity and imagination. In both computational linguistics and the digital humanities, literature therefore presents itself as a natural domain of study to understand language as a social tool.

What is the advantage to literary analysis of such computational methods? A lot has been said and debated about the dichotomy between close reading and distant reading (Hammond, 2017; Underwood, 2019) in the humanities — the former is limited by the size of the samples that can be attended to in a lifetime, but the latter reduces the study of literature to a series of mostly meaningless, and often uninteresting, numerical tasks and scores. While there are merits to each stance, it seems quite uncontroversial to state that both approaches present complementary views of literature and the worlds it represents. Perhaps there are nuances of language that cannot yet be captured by data-driven predictive and generative models, yet it allows us to create a zoomed-out, aggregate portrait of the vast amounts of writing produced throughout our history, and in the process identify macro-trends that would have been obscured at a smaller scale. Perhaps the tendency of machine learning models to reduce a phenomenon largely to the mean of its observations glosses over micro-variations, subversions of norms, outliers that are perhaps the actual phenomena of interest; however, methods of computational analysis can, and should, be customized and designed to answer the question of interest. If it is outliers that one is interested in, then we can train a model to detect outliers from trends. Distant reading is less an independent agent writing its own conclusions, and more a tool that can be finessed to test all sorts of hypotheses, both at the macro and the micro scale.

## 4.2 Centering the Character

The allure of computational literary analysis has been well legitimized by the extensive research into and progress on computational methods to quantify various aspects of the novel (see Section 4.3 for an overview of prior work). The focus of this thesis is on one such aspect, that of the character in the novel. While authorial voice and narrative voice are well-studied in both theoretical and computational literary analysis, a study of its characters — as (fictional) individuals with distinctive voices, personas, and trajectories — has been somewhat lacking in the latter. This is in part because of the pure technical difficulty of doing so, particularly in an automated manner for a large number of texts. We already touched on the challenges with coreference resolution, for example, in the introductory paragraphs of this chapter; we dive into it further in Chapter 8 by studying the problem of quotation attribution (identifying who said what in a novel).

The aspects of the novel that we investigate here are inspired by the literary theory of dialogism. In this view, characters are used by the author to represent a "plurality of voices", imparting them with distinctive manners and styles that do not necessarily reflect the practices of the author or narrator themselves. How, concretely, can this *plurality of voice* be imparted to characters within the text? They can be distinguished stylistically, though what linguistic features constitute this style is a question for our NLP models to answer. Apart from traditional features that have been explored in computational stylometry studies, such as term frequencies, we concentrate our focus on lexical features of *style*, as developed by Brooke and Hirst (2013a), and on lexical features of *emotion* (Mohammad, 2018a,b).

We then use the emotional features of character utterances to further characterize their distinctiveness, via the framework of Utterance Emotion Dynamics (UED). The UED framework computes a set of metrics that represent the temporal sequence of emotion states associated with a speaker, derived from their utterances over a period of time. Here, time is represented by the narrative flow of the novel, and each character constitutes a speaker for whom we derive UED metrics. As opposed to most prior work in this area that considers the narrative arc of a novel to be a singular arc of all of its text or narration, we again place its characters and their dialogue at the centre of our study. How representative is the narrative arc of a novel of the journeys of its characters? How diverse are the arcs of various characters within the novel, and does this measure of diversity inform us in any way about the plurality of voices within it?

The work in this section of the thesis falls somewhere in between the extremes of close reading and distant reading. From the point of view of big data analysis, we work with a laughably small number of texts (28). On the other hand, we cover a sizeable number of authors (16) and genres within our corpus, allowing us to test the generalizability of certain pre-established conclusions on character voices, authorial style, and narrative shapes. Our particular narrative feature of interest is character voices, which have in the past been limited to datasets containing between 1 and 6 novels. The set of 28 novels in our corpus provides us with utterance data for a total of 809 characters, 307 of whom can be classified as important characters contributing to at least 5% of the total dialogue in the text. Our study therefore utilizes a relatively, though perhaps not sufficiently, 'big' dataset for analysis.

### 4.2.1 Aspects of Gender and Sex

A notable portion of this section of the thesis delves into the differences in character portrayal when viewed through the lens of what I refer to as *gender* – that of the author, as well as the characters they write. Here, I clarify which aspects of individual (author) and character identities are captured by my use of the term "gender", and how it relates to prior literature in NLP as well as the humanities on gender and sex, with a specific focus on the domain of literature.

The sex of an individual refers to a set of physical characteristics or biological attributes, such as chromosomes, gene expression, and reproductive/sexual anatomy[1]. Sex is usually categorized as male or female, but variations do occur.

[1] https://cihr-irsc.gc.ca/e/48642.html

Gender, on the other hand, is a construct that is dependent on a wide variety of personal, social, and cultural factors. Unlike sex, the gender identity of an individual can be non-binary (falling on a spectrum), fluid (mutable), and multi-dimensional. As such, it defies categorization into a set of predetermined classes, or even as a variable that can be represented by a single value. Cao and Daumé (2020) describe the many aspects that encompass gender: the gender that one *experiences*, the one that is *presented* to the world, and the one that is imposed on them by social judgement or perceptions (which is typically a binary between masculine and feminine roles).

How this gender identity is realized in language further complicates efforts to quantify it. The gender categories that have a linguistic form, while constantly evolving, are severely limited when compared to the spectrum of individual gender identities ("linguistic categories of gender do not even remotely map one-to-one to social categories" (Cao and Daumé, 2020)). These linguistic categories of gender also vary widely from language to language. We can have *grammatical gender*, wherein nouns (animate and inanimate) are divided into gendered groups that determine the grammatical agreement of dependent terms in the sentence (English does not have grammatical gender). The use of gendered pronouns confers a *referential gender* upon entities, typically grouped into the male (*he/him*), female (*she/her*), and gender-indefinite categories (*everyone, someone*). These pronominal categories are continually being expanded, for example with neopronouns (*ze, em*) in English, and the recent resurgence of the use of the singular *they* as a gender-neutral pronoun. These linguistic categories of gender may loosely map to some social perceptions of the gender of the humans they refer to or represent, but this is not always the case. Additionally, *proper names* are heavily associated with a particular social gender of the referent entity in many languages, including English (*Jane* is likely to refer to a female person, *John* to a male person). We also have *lexical gender*, where terms like *mother* and *son* carry gender signals.

I now briefly describe how we operationalize gender for the two types of entities that we study in this thesis: characters within novels, and the authors of those novels.

**Character Gender:** The characters that we analyze in this thesis are fictional; there is no physical embodiment of their selves in the real world, and it therefore does not make much sense to talk about the determination of their *sex*. What aspects of their *gender*, then, are we presented with in the text? We are exposed, with varying levels of detail, to behaviours and features of the characters that we, as readers, might interpret as being associated with a particular gender; this interpreted gender, however, could vary with what each of us consider as the 'norm' for a gender, and also vary with time period, place, and culture. We are additionally likely to be presented with proper names for these characters, thereby indicating **lexical gender** by way of the names (*Mary*) and titles (*Mrs. Bennet*), and what we have defined previously as **referential gender**, by way of pronouns used by the author (and, via the author, narrators and characters) to refer to the character in the text. Referential gender is what we have focused on quantifying in this work for the characters in our novels: we label a character as Male (M) if they are consistently referred to by *he/him* pronouns, Female if they are referred to by *she/her* pronouns, Ambiguous (A) if they are referred to by a mix of male and female pronouns or with gender-neutral pronouns like *it*, and Unknown (U) if there is no pronominal information presented about the character. In some cases, when there is a lack of pronominal terms referring to a character entity, particularly for minor characters who are mentioned only in passing, we make use of lexical gender — labeling a *policeman* as Male, for example, or *Mary Jane* as Female.

**Author Gender:** With authors, we are able to talk about determining both their sex and gender, though actually doing so might not be very simple. Self-declaration is the best way to determine both of these variables for any person, but since we deal in our dataset with authors who are long dead, we rely on external resources such as Wikipedia. As before, we label authors as Male or Female if, based on the information available to us, they can be described as such both in terms of biological (sex) and cultural (gender) factors; we refer to this variable as *author gender*. In cases of apparent or possible conflict, we mark gender as "A" (ambiguous); in cases where information is lacking, we mark them as "U" (unknown). We do not encounter any of the latter two cases with the authors that we consider here, but one might encounter them when analyzing an expanded corpus of novels and authors.

## 4.3 Overview of Related Work

I now review related work from computational literary studies, spanning research in natural language processing and the digital humanities, on narrative analysis of fictional texts, with a particular focus on character analysis.

The bulk of stylometric work in computational linguistics has focused on the distinctive voices of *authors*, providing evidence that authorial style can be reliably captured with certain linguistic features. Burrows (1989) demonstrated that the major characters in Jane Austen's novels could be distinguished to some extent by their differing frequencies of usage of the most-frequently occurring words (MFW) in the corpus. Word frequencies along with eigen-decomposition (such as PCA) and visualization were used to identify distinct clusters of character voices within novels. In the same work, he also shows that male and female character dialogue clusters together, though somewhat imperfectly. This pioneering initial study might seem to contradict that of authorial style being consistent across a novel, but the two views can be reconciled by considering character voices to be a form of micro-variation observed within the overarching voice of the author. Certain works on authorial style in fictional texts do however, remove dialogue from consideration when creating feature profiles.

Clustering techniques form the most common method of demonstrating distinctive character voices. In Hoover (2017), MFW vectors are passed to a hierarchical clustering algorithm that successfully groups together most of the utterances of individual characters in Sherlock Holmes novels. One reason to prefer clustering over a predictive classification approach is the scarcity of data. Stylometric features become more "visible" and reliable across larger spans of text — previous studies use chunks of text ranging from 1000 to 6000 tokens. Obtaining several samples of this size is largely impossible for the dialogue of individual characters, rendering training a classifier an infeasible approach.

Brooke and Hirst (2013b) develop for the first time a lexicon of stylistic dimensions. Six complementary dimensions of style are identified based on prior literature: abstract vs concrete, subjective vs objective, and literary vs colloquial. The authors also propose LDA-based models to iteratively construct a large lexicon of words associated with style scores along each dimension. The style features provided by this lexicon prove to be indicative of the distinct voices presented in the free indirect discourse of Virginia Woolf's *To The Lighthouse*, and correlate with social aspects of the character in the narrative, such as gender and social class. Similar clustering techniques were applied to distinguish the various voices in T.S. Eliot's poem *The Waste Land*, with moderate success. This lexical approach has the benefit of providing a human-interpretable set of dimensions with which to quantify style (Brooke et al., 2013, 2015, 2016).

Muzny et al. (2017a) directly tackle the concept of dialogism by proposing a metric to quantify it using certain grammatical features that are representative of dialogic text as opposed to narrative text in a novel. This approach does not delve into distinguishing individual characters within the text; rather it identifies part-of-speech features that are more prominent in dialogue when compared to narration and uses this to assign a 'dialogism score' to spans of text.

### 4.3.1 Character Networks and Profiles

Bamman et al. (2014b) develop a persona model of a character within a novel that represents them as a mixture of a fixed number of latent types. The aspects that determine a character's persona are drawn from the verbs, adjectives, and nouns used to refer to them, or describe them, in the narration. The resulting model is able to cluster characters based on the similarity of their latent persona vectors, and satisfies several commonsense hypotheses that place characters within a novel to be more similar to each other than to a character from an entirely different author and novel, and so on.

Narrative references to characters have extensively been used to identify character networks and relationships, and dialogue has been used here to some extent. Elson et al. (2010) use quoted speech to find social links between characters and model relationships between them. Chaturvedi et al. (2016) focus on modeling temporal sequences of character relations, again mostly using narrative text; similar lines of work are followed by Iyyer et al. (2016). Sims and Bamman (2020) use attributed character speech to quantify information propagation within a narrative, with interesting observations on the stereotypical gender dynamics in 19th and 20th century novels that place female characters as the propagators of information, despite their rather diminished presence when compared to male characters.

### 4.3.2 Emotion Arcs of Narratives

Methods of emotion detection have evolved from a token-level determination of binary polarity to include sophisticated predictive models that take into account local and global context and deal with multiple dimensions of emotion, including discrete emotions like anger, joy, sadness, etc., and affective states like valence, arousal, and dominance. The term "sentiment analysis" is broadly used to describe a variety of related tasks, including but not limited to, detecting the overall emotion conveyed by the writer of the text; the emotion invoked in the reader of the text; and the emotions associated with one or more of the entities mentioned in the text. While sentiment and emotion analysis have long been hotbeds of NLP research, we limit our review here to work that explicitly deals with narrative analysis.

Kurt Vonnegut famously gave a lecture on what he considered to be the eight basic shapes of stories[2], where the shape is a graph representing the emotional trajectory of the protagonist through the course of the story (Vonnegut, 2009). Reagan et al.

[2] https://bigthink.com/high-culture/vonnegut-shapes/

(2016) conduct a large-scale analysis of this hypothesis using texts from the open-source repository of books, Project Gutenberg. They determine the emotion arc for a novel by dividing the text into 1000-word chunks and computing a 'happiness score' for each chunk using a word-level lexicon of scores. They cluster arcs and show that six basic shapes can indeed be derived that correspond roughly to those proposed by Vonnegut. Prior to this, Mohammad (2012) also performed a large-scale analysis of the flow of emotion-bearing words in a narrative, and showed clusters based on genre as well as time-period that correlate with real-world events in various geographic locations. Other works have since examined narrative sentiment and emotion arcs in the context of genre (Kim et al., 2017a), narrative mood (Öhman and Rossi, 2023), and reader preferences and literary quality (Moreira et al., 2023; Bizzoni et al., 2023; Ohman et al., 2024).

Nalisnick and Baird (2013) dive into emotions conveyed in character dialogue with a dataset of Shakespeare's plays; an utterance is assumed to have been uttered by the closest character mention. They use this to test various hypotheses on the relations between characters and their evolution, such as protagonist–antagonist pairs and romantic couples. Character relations, as presented in the previous section, have been primarily analyzed using mentions in the narration rather than utterances.

Hipson and Mohammad (2021) introduce for the first time the Utterance Emotion Dynamics (UED) framework, which formalizes a series of metrics drawn from the emotional arc of a sequential narrative. These metrics are inspired by works in psychology and the affective sciences that study the emotional states and dynamics of humans and their relationship to emotional, mental, and physical well-being. We largely draw from this framework for our own work on character utterances. In their work, the authors circumvent the difficulty of dealing with utterances in literary texts by focusing instead on characters in movies, whose utterances can be trivially extracted using scripts.

### 4.3.3 Automatic Quotation Attribution

In the vast research space of computational literary analysis, research into the information conveyed by character dialogue has been severely limited — justifiably limited, however, by the lack of datasets that are annotated for the necessary aspects of utterances and characters. Chapter 8 deals explicitly with automated methods to identify characters and their associated dialogue given the text of a novel; we leave a review of related work in this domain to that chapter.

# Chapter 5

# The Project Dialogism Novel Corpus

This chapter is adapted from the first four sections of the following publication: Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The Project Dialogism Novel Corpus: A Dataset for Quotation Attribution in Literary Texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848, Marseille, France. European Language Resources Association.

**Author Contributions:** Prof. Adam proposed the initial idea for the dataset creation and led the majority of the annotation effort, including recruiting and supervising annotators, and selecting the texts to annotate. Krishnapriya developed the annotation platform and wrote code to process the annotations and create the final dataset, with feedback from Profs. Graeme and Adam. Krishnapriya performed the evaluation of coreference resolution and quotation attribution models on the dataset, and wrote the initial draft of the paper. All three authors revised the final version.

## 5.1 Overview

In this chapter, I introduce a new dataset for the computational analysis of novels: the Project Dialogism Novel Corpus (PDNC). The PDNC consists of 28 novels in which all quotations are identified and annotated for speaker, addressee(s), and characters mentioned. PDNC is, by an order of magnitude, the largest corpus of its kind. Each novel is annotated manually by a pair of annotators using customized software I developed; the annotation team consisted of undergraduate and graduate students from the Department of English, and was supervised by Prof. Adam Hammond. In addition to the dataset, we also release the custom annotation software (including the source code) and the full annotation guidelines, which we hope will help standardize future work in this area. PDNC will promote a more nuanced and accurate view of novelistic discourse; whereas much research currently envisions the novel as expressing the voice of the author, the PDNC presents novels as a polyphonic fabric of characters' voices.

## 5.2 Prior Datasets

The Columbia Quoted Speech Attribution (CQSA) corpus from Elson and McKeown (2010) contains annotations for 3176 instances of quoted speech from 4 novels by each of 4 authors, and 7 short stories from 2 others; only parts of the full-length novels are annotated. Quotations are annotated at the mention-level, i.e, the speaker is chosen from a set of candidate mentions that occur in the nearby context. These mentions are then resolved to speakers by using an off-the-shelf coreference tool. He et al. (2013) annotate a dataset of three novels, *Pride and Prejudice*, *Emma*, and *The Steppe*; the latter two are also present in the CSQA corpus. Their annotation method links quotations directly to canonical characters, rather than mentions. Muzny et al. (2017b) released the QuoteLi dataset, comprising 3103 quotations annotated with both mention and speaker information. The quotations are drawn from the same three novels as those of He et al. (2013). Finally, Sims and Bamman (2020) annotate the first 2000 tokens of 100 novels from the LitBank dataset[1]. Quotations are linked to a unique speaker from a predefined list of entities. Though this dataset spans the largest number of novels (100), the restricted range of tokens considered results in only 1765 total annotations.

LitBank also contains annotations for coreference, for the same set of 2000 tokens across 100 novels. A total of 29,103 tokens are annotated, of which 24,180 refer to a person, and the rest to other named entities such as places, organizations, vehicles, etc Bamman et al. (2020). Prior to this, Vala et al. (2016) annotated coreference in *Pride and Prejudice*.

## 5.3 The Project Dialogism Novel Corpus

We draw our novels from open-source texts available on the Project Gutenberg platform. In selecting these novels, our aim has been to annotate texts in a variety of

[1] https://github.com/dbamman/litbank

| **Quotation** | **Annotations** |
|---|---|
| *“You must not be too severe upon yourself,”* replied **Elizabeth** | **Speaker:** Elizabeth Bennet<br>**Addressees:** (Mr. Bennet, Kitty)<br>**Quote type:** Explicit<br>**Referring Expression:** replied Elizabeth<br>**Mentions:** (‘you’, Mr. Bennet), (‘yourself’, Mr. Bennet) |
| With an air of indifference **he** soon afterwards added: *“How long did you say he was at Rosings?”* | **Speaker:** George Wickham<br>**Addressees:** Elizabeth Bennet<br>**Quote type:** Anaphoric<br>**Referring Expression:** he soon afterwards added<br>**Mentions:** (‘you’, Elizabeth Bennet), (‘he’, Colonel Fitzwilliam) |
| *“But not before they went to Brighton?”* | **Speaker:** Elizabeth Bennet<br>**Addressees:** Jane Bennet<br>**Quote type:** Implicit<br>**Referring Expression:**<br>**Mentions:** (‘they’, [George Wickham, Lydia]) |

Table 5.1: Annotations for three sample quotations from PDNC, one for each quotation type. The speaker in each example is highlighted in bold, and mentions within quotations are underlined.

genres (literary fiction, children's literature, detective fiction, and science fiction are represented); from the LitBank and QuoteLi corpora, to facilitate comparison and validation; and of broad interest to a variety of scholars while still relevant to our group's interest in stylistic diversity and dialogism (Hammond et al., 2020; Vishnubhotla et al., 2019). Further, we have chosen to annotate multiple novels by the same author (Jane Austen, E.M.Forster), in order to facilitate comparative analysis of a single author's oeuvre (Austen was chosen because she is included in all existing corpora).

### 5.3.1 Annotated Attributes

Each quotation in our corpus of texts is annotated with the following attributes:

1. **Speaker:** The character uttering the quotation. We limit each quotation to having a single speaker; certain special cases are highlighted in Section 5.4.4.

2. **Addressee(s):** The set of character(s) being addressed by the speaker. This includes any character that is in the vicinity of the speaker and can "hear" the uttered quotation.

3. **Quotation Type:** Following previous work, we distinguish between explicit, anaphoric, and implicit quotations. See Table 5.1 for an example of each.

4. **Referring Expressions:** For explicit and anaphoric quotations, we obtain the part of the text that indicates who the speaker is, the verb for the action of speaking, and sometimes, also the addressees.

5. **Mentions:** Finally, we also annotate all characters who are mentioned within a quotation, either explicitly by name or through a pronoun or pronominal phrase. Each mention is linked to the character or set of characters that it refers to.

In addition, each novel is also annotated with a list of characters present in the novel. Each character is associated with a "main name" (e.g., Elizabeth Bennet), as well as a set of aliases by which they are referred to in the text (e.g., Lizzy, Liz, Elizabeth). The character list includes any character who either speaks, is addressed, or is mentioned in a quotation; therefore we also have characters who are never explicitly assigned a proper name, such as "The Old Man in the Crowd".

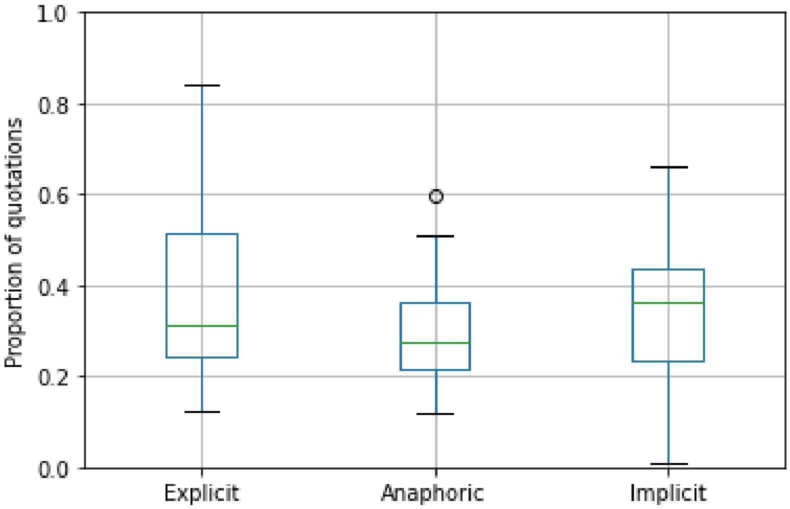


Figure 5.1: Distribution of quotation types across novels in PDNC.

### 5.3.2 Dataset Statistics

We list key characteristics of PDNC in Table 5.2. A total of 48,810 quotations are identified and annotated for the attributes listed in Section 5.3.1. On average, we have 1.79 aliases per character, and 1.82 mentions annotated per quotation. Of the 1228 characters in our character lists, 807 are speakers of a quotation; of these, 321 characters can be classified as “minor”, having 10 or fewer spoken quotations. Margaret Schlegel from *Howards End* is the most loquacious character across all novels, with 1040 quotations, followed by Jake Barnes from *The Sun Also Rises*, Katherine Hilbery from *Night and Day*, and Anne Shirley from *Anne of Green Gables*.

Figure 5.1 shows shows a box-and-whisker plot of the three quotation types annotated in our dataset. The central region (the box) indicates the “middle portion” of the data distribution, i.e, the range covered between the first quartile (the 25% mark) and the third quartile (the 75% mark), with the median (50% mark) lying at line inside the box. The whiskers, the dashes on either end of the plot, are at a distance of 1.5 times the inter-quartile length (inter-quartile length is the distance between the first and third quartiles). Points beyond the whiskers are considered outliers.

We see that Explicit quotations make up the largest percentage of annotations (36.5%), followed by Implicit (34.2%) and Anaphoric (29.2%) quotation types, though the distribution shows a large spread. *Alice in Wonderland* consists mostly of explicit quotations (84%), whereas Dostoevsky’s *The Gambler* is at only 12%.

We note that PDNC is by far the largest dataset of annotated quotations for works of English Literature. A comparison with previous datasets is presented in Table 5.3.

| **Novel** | **Author** | **# Tokens** | **# Quotations** | **# Characters** | **# Mentions** |
|---|---|---|---|---|---|
| *A Handful Of Dust* | Evelyn Waugh | 88559 | 2617 | 104 | 3198 |
| *Alice's Adventures in Wonderland* | Lewis Carroll | 34363 | 1048 | 51 | 683 |
| *Anne Of Green Gables* | Lucy Maud Montgomery | 123605 | 1779 | 113 | 5168 |
| *A Passage to India* | Edward Morgan Forster | 125104 | 2398 | 48 | 3083 |
| *A Room With A View* | Edward Morgan Forster | 84366 | 1989 | 63 | 3079 |
| *Daisy Miller* | Henry James | 26988 | 725 | 10 | 1021 |
| *Emma* | Jane Austen | 191642 | 2109 | 18 | 6310 |
| *Hard Times* | Charles Dickens | 127731 | 2414 | 38 | 4395 |
| *Howards End* | Edward Morgan Forster | 138059 | 3111 | 55 | 4344 |
| *Mansfield Park* | Jane Austen | 186364 | 1568 | 37 | 4907 |
| *Night and Day* | Virginia Woolf | 199709 | 2795 | 50 | 3568 |
| *Northanger Abbey* | Jane Austen | 91020 | 1014 | 20 | 2351 |
| *Oliver Twist* | Charles Dickens | 197490 | 4253 | 96 | 5583 |
| *Persuasion* | Jane Austen | 97860 | 682 | 35 | 2141 |
| *Pride and Prejudice* | Jane Austen | 144604 | 1708 | 74 | 4797 |
| *Sense and Sensibility* | Jane Austen | 140499 | 1543 | 24 | 4671 |
| *The Age of Innocence* | Edith Wharton | 121413 | 1592 | 55 | 2549 |
| *The Awakening* | Kate Chopin | 58902 | 728 | 22 | 978 |
| *The Gambler* | Fyodor Mikhailovich Dostoevsky (Trans. C.J.Hogarth) | 73144 | 1066 | 27 | 2056 |
| *The Invisible Man* | Herbert George Wells | 59658 | 1245 | 31 | 903 |
| *The Man Who Was Thursday* | Gilbert Keith Chesterton | 69215 | 1351 | 30 | 1695 |
| *The Mysterious Affair At Styles* | Agatha Christie | 72263 | 2212 | 30 | 3481 |
| *The Picture Of Dorian Gray* | Oscar Wilde | 95138 | 1500 | 43 | 3336 |
| *The Sign of the Four* | Arthur Conan Doyle | 51371 | 891 | 35 | 1784 |
| *The Sport of the Gods* | Paul Laurence Dunbar | 49923 | 810 | 37 | 1499 |
| *The Sun Also Rises* | Ernest Hemingway | 88361 | 3245 | 51 | 2731 |
| *Where Angels Fear to Tread* | Edward Morgan Forster | 61684 | 1236 | 18 | 1864 |
| *Winnie-The-Pooh* | Alan Alexander Milne | 29786 | 1181 | 13 | 824 |
| **Total** | | 2828821 | 48810 | 1228 | 82999 |

Table 5.2: The set of novels annotated in PDNC, with the number of annotated quotations, characters, and mentions in each.

Even though we annotate only for mentions within quotations, our count of 82,999 mention annotations is much larger than LitBank's 29,103.

PDNC also contains the largest number of tokens per document (101,033), since we annotate entire novels rather than portions of each. We believe that this is an invaluable resource for several open problems in the computational analysis of literature, allowing for tracking character mentions across larger spans of text, studying changes in character style, emotions, and character networks throughout the course of a novel, and the variation of each of these with author and genre.

| **Corpus** | **# Texts** | **# Quotations** |
|---|---|---|
| CQSA (2010) | 6 | 3176 |
| He et al. (2013) | 3 | 1901 |
| QuoteLi (2017) | 3 | 3103 |
| LitBank (2020) | 100 | 1765 |
| PDNC (2021) | 28 | 48810 |

Table 5.3: A comparison of PDNC with previous datasets for quotation attribution in literary texts.

## 5.4 PDNC: The Annotation

In this section, we describe our annotation process, from developing the guidelines to preprocessing the texts, the annotation platform, and how we resolved disagreements between annotators.

### 5.4.1 Annotation Platform

We designed our annotation platform from scratch as a web-based interface. A screenshot of the interface is shown in Figures 5.2 and 5.3. The main components include the character list, which allows the annotator to add and remove characters and associated aliases; the text box, which highlights quotations and mentions within the text (different color codes indicate the type and annotation status of the quotation or mention spans); and the annotation area, where values for the desired attributes of a quotation or mention can be set by the annotator. The platform also includes an interface that takes as input two sets of annotations of the same text and generates a file with any disagreements that occur for an annotated attribute, including mis-matches in character lists.

### 5.4.2 Annotation Process

All our annotators were university-level literature students familiar to one of the authors of this study. Each novel in our corpus was annotated separately by two annotators, and the resulting annotations were then compared to generate a list of "disagreements". Disagreements were grouped by quotation, and occur when the annotations do not match for any of the attributes listed in Section 5.3.1. The two annotators then went through a consensus exercise, where they discussed all disagreements, re-annotated the relevant quotations, and once again checked for disagreements (in practice, no more than three rounds of consensus were necessary).

### 5.4.3 Pre-processing the texts

The raw text for each novel is obtained from the Project Gutenberg platform. This is then processed using the GutenTag software[2] from Brooke et al. (2015), which outputs an initial list of characters and aliases, and also identifies quotations within the text. We also pre-identify mentions within each quotation by looking for occurrences of any character names, aliases, or words from a predefined list of pronouns.

[2] https://gutentag.sdsu.edu/

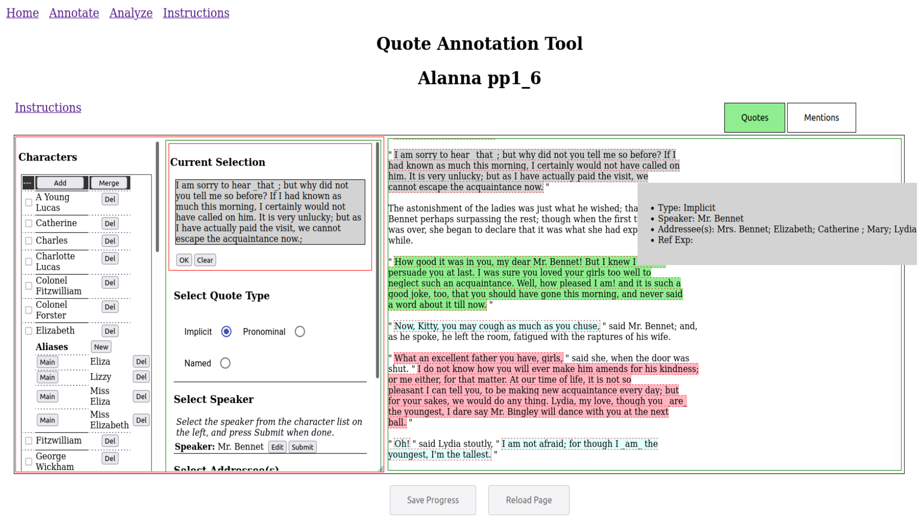
### 5.4.4 Annotation Guidelines

The complexity of narrative structure and style of literary novels means that several ambiguities can arise while determining any of the annotated attributes. We developed a comprehensive set of guidelines that attempt to cover as many as possible of the cases that we came upon in our texts. These guidelines underwent several revisions as we progressed through different novels, and were informed by feedback from our annotators as well as the authors of this work. We make the complete set of guidelines publicly available and hope it will help guide future work in this area. We highlight a few interesting cases below:

- Special aliases: Narrators of first-person narratives receive the special alias "_narr"; when more than one character speaks a quotation in unison, it is attributed to "_group"; when the identity of the speaker is unknowable in context, it is attributed to "_unknowable".

- Multiple addressees: In situations in which many characters are present, our guidelines designate an addressee as anyone "whom the speaker seems to believe can hear them."

- Locating referring expressions: Our guidelines include explicit instructions for annotating referring expressions in cases in which they are difficult to annotate, in which they introduce long or multi-part quotations, and in which multiple referring expressions are applied to single quotation.



**Quote Annotation Tool**

**Alanna pp1_6**

Instructions

**Characters**

- A Young Lucas
- Catherine
- Charles
- Charlotte Lucas
- Colonel Fitzwilliam
- Colonel Forster
- Elizabeth
  - **Aliases**: Eliza, Lizzy, Miss Eliza, Miss Elizabeth
- Fitzwilliam
- George Wickham

**Current Selection**

I am sorry to hear _that_; but why did not you tell me so before? If I had known as much this morning, I certainly would not have called on him. It is very unlucky; but as I have actually paid the visit, we cannot escape the acquaintance now.;

**Select Quote Type**

Implicit ● Pronominal ○

Named ○

**Select Speaker**

*Select the speaker from the character list on the left, and press Submit when done.*

**Speaker:** Mr. Bennet

**Select Addressee(s)**

" I am sorry to hear  that ; but why did not you tell me so before? If I had known as much this morning, I certainly would not have called on him. It is very unlucky; but as I have actually paid the visit, we cannot escape the acquaintance now. "

The astonishment of the ladies was just what he wished; tha Bennet perhaps surpassing the rest; though when the first t was over, she began to declare that it was what she had exp while.

- Type: Implicit
- Speaker: Mr. Bennet
- Addressee(s): Mrs. Bennet; Elizabeth; Catherine ; Mary; Lydia
- Ref Exp:

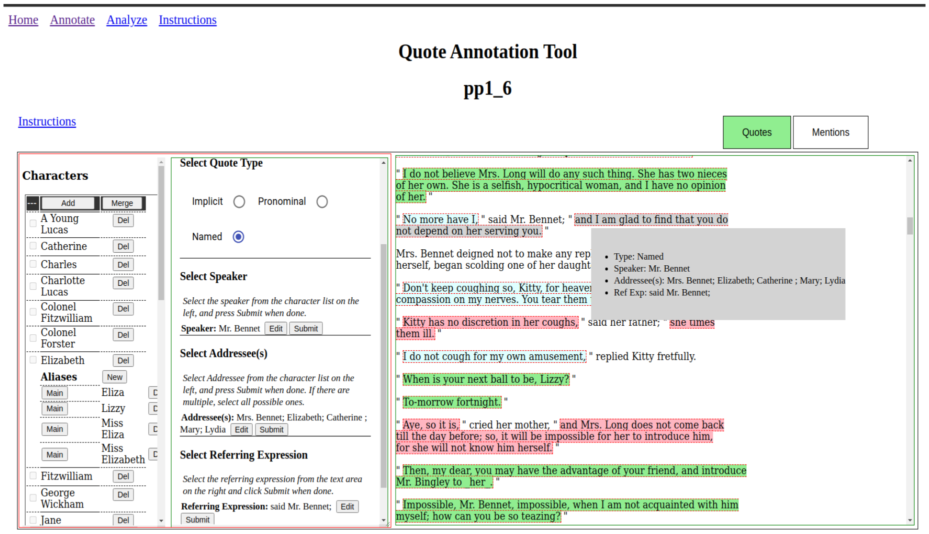
" How good it was in you, my dear Mr. Bennet! But I knew I persuade you at last. I was sure you loved your girls too well to neglect such an acquaintance. Well, how pleased I am! and it is such a good joke, too, that you should have gone this morning, and never said a word about it till now. "

" Now, Kitty, you may cough as much as you chuse, " said Mr. Bennet; and, as he spoke, he left the room, fatigued with the raptures of his wife.

" What an excellent father you have, girls, " said she, when the door was shut. " I do not know how you will ever make him amends for his kindness; or me either, for that matter. At our time of life, it is not so pleasant I can tell you, to be making new acquaintance every day; but for your sakes, we would do any thing. Lydia, my love, though you  are  the youngest, I dare say Mr. Bingley will dance with you at the next ball. "

" Oh! " said Lydia stoutly, " I am not afraid; for though I  am  the youngest, I'm the tallest. "

Figure 5.2: A screenshot of our annotation platform. The different colors indicate the type of quotation.



**Quote Annotation Tool**

**pp1_6**

Instructions

**Characters**

- A Young Lucas
- Catherine
- Charles
- Charlotte Lucas
- Colonel Fitzwilliam
- Colonel Forster
- Elizabeth
  - **Aliases**: Eliza, Lizzy, Miss Eliza, Miss Elizabeth
- Fitzwilliam
- George Wickham
- Jane

**Select Quote Type**

Implicit ○ Pronominal ○

Named ●

**Select Speaker**

*Select the speaker from the character list on the left, and press Submit when done.*

**Speaker:** Mr. Bennet

**Select Addressee(s)**

*Select Addressee from the character list on the left, and press Submit when done. If there are multiple, select all possible ones.*

**Addressee(s):** Mrs. Bennet; Elizabeth; Catherine ; Mary; Lydia

**Select Referring Expression**

*Select the referring expression from the text area on the right and click Submit when done.*

**Referring Expression:** said Mr. Bennet;

" I do not believe Mrs. Long will do any such thing. She has two nieces of her own. She is a selfish, hypocritical woman, and I have no opinion of her. "

" No more have I, " said Mr. Bennet; " and I am glad to find that you do not depend on her serving you. "

Mrs. Bennet deigned not to make any rep herself, began scolding one of her daught

- Type: Named
- Speaker: Mr. Bennet
- Addressee(s): Mrs. Bennet; Elizabeth; Catherine ; Mary; Lydia
- Ref Exp: said Mr. Bennet;

" Don't keep coughing so, Kitty, for heaven compassion on my nerves. You tear them

" Kitty has no discretion in her coughs, " said her father; " she times them ill. "

" I do not cough for my own amusement, " replied Kitty fretfully.

" When is your next ball to be, Lizzy? "

" To-morrow fortnight. "

" Aye, so it is, " cried her mother, " and Mrs. Long does not come back till the day before; so, it will be impossible for her to introduce him, for she will not know him herself. "

" Then, my dear, you may have the advantage of your friend, and introduce Mr. Bingley to  her . "

" Impossible, Mr. Bennet, impossible, when I am not acquainted with him myself; how can you be so teazing? "

Figure 5.3: A screenshot of our annotation platform showing the various attributes associated with each quotation.

# Chapter 6

# Stylometric Analysis of Character Utterances

## 6.1 Introduction

**Note:** The word *character* is used here in two different senses. For the most part, we use it to mean a character entity in a novel — a fictional person who says and does things. The other sense is as a unit of computational textual analysis, referring to the individual alphabets and digits and other special symbols that make up larger units of text like tokens and sentences. The former sense is the most common use in this section of the thesis, whereas the latter is the dominant sense across much of NLP research. In this chapter and the next, I will assume that character as a living fictional entity is the default use-case, and add a note in the appropriate places indicating when we switch to the other sense.

In this chapter and the next, we quantify the extent to which character voices in a novel are distinctive, and the variation of this measure across novels and authors. Our primary approach to determining this is to assess how well a character can be identified based solely on the linguistic characteristics of their utterances — in other words, the predictive power of the linguistic features of utterances in identifying the character that uttered them. A highly distinctive character voice[1] will make such an identification quite easy. At the novel-level, a high average accuracy of identification across characters indicates that the voices within it are easily separable, and therefore

[1] The term *voice* is used here to refer to a stylistically distinctive way of talking. *Talking*, in turn, refers to the quotations attributed to the character in the novel, rather than actual spoken-aloud acoustic signals.

distinct from one another.

A key component of this approach is the feature extractor: a function that converts textual utterances into numerical vectors of features that can then be passed to a classification model or a clustering model to quantify distinctiveness. The scope of such functions for text is wide-ranging; early computational work on textual style demonstrated the effectiveness of various lexical and syntactic features, such as frequencies of syllables, common function words, and part-of-speech tags, along with surface-level features like word and sentence lengths. These lexical features can be expanded to include various lexicons that associate words with real-valued scores representing their association with more pragmatic dimensions, such as emotion (anger, fear, positivity), style (level of abstractness, subjectivity) and various other connotative aspects (power, agency).

Compiling an exhaustive list of such features, as well as ensuring the most effective selection of features, is quite an open-ended task. An often-debated question, particularly in stylometric analyses, is the conflation of topic-related features with style. If a character in a novel talks extensively in metaphors and idioms compared to the others, then there is no question that they have a distinctive character voice. However, if a character is distinguished solely by the fact that they are a shoemaker and hence talk extensively about shoes, or by virtue of being the only king in the land they tend to use terms associated with royal activities, or if they form an isolated node in the graph of character interactions which results in certain proper nouns or names appearing with a higher frequency in their utterances, we might be less inclined to point to it as a distinctive character voice. This discrepancy touches on the style vs content debate discussed in Chapter 2, where the boundaries between the two can often be blurry. Here, we rely for the most part on feature sets that have been validated in the prior literature for various applications involving textual style, such as authorship attribution and profiling, plagiarism detection, and style change detection, which at times include content-related words.

An alternative approach to feature extraction is to use neural networks, where the feature extractor is a function with set of learnable parameters that is approximated by optimizing for an objective metric. The objective usually represents the goal of learning this representation function for the text — in our case, to maximize the accuracy of predicting the speaker. While a step-up in some ways over the manual feature engineering approach, neural networks generally require large amounts of data (approximately thousands of instances per class), and are much less interpretable in terms of identifying specific linguistic features that are most effective at the classification task.

## 6.2 Research Questions

Our research questions are primarily guided by the literary idea of dialogism, or a dialogic novel, wherein the different characters of the novel interact with one another to present multiple differing viewpoints and perspectives. We quantify this concept in a few different ways, encapsulated by the following research questions (RQs).

**RQ1: Can characters in a novel be identified by the linguistic features of their utterances?** How does this measure vary across novels and authors?
Here, we consider the accuracy of identifying speakers via their utterances to be a proxy measure of dialogism. The more separable character utterances are, or alternatively, the more distinct the clusters of utterances belonging to different characters are, the more dialogic the novel. We test this approach with various feature extractors, and models for classification as well as clustering of these features.

**RQ2: What role does the gender of the characters involved in an interaction play in determining utterance style?** Do characters speak differently based on who they are speaking to?
While RQ1 considered consistency in character voice, here we study its variation. While the former has been more widely studied in computational literary studies, we would expect that situational and social contexts affect the way characters (and people, in real life) speak; characters often evolve through the course of a novel's narrative arc as well. We consider two aspects of social context here: the gender of the speaker, and the gender of the addressee (we have defined what gender means in this context in Section 4.2.1). We model the dependency of certain lexical style features on these variables using linear mixed-effect models.

**RQ3: How do authors differ in their differentiation of character voices?** Do male and female authors differ in the portrayal of gender dynamics within character interactions?
The final research question steps out of the context of the novel to consider the gender of the author (defined in Section 4.2.1) as another aspect of social context in determining character voice. Literature oftentimes reflects the social dynamics and norms of its time, while also acting as a playground for experimentation and subversion of

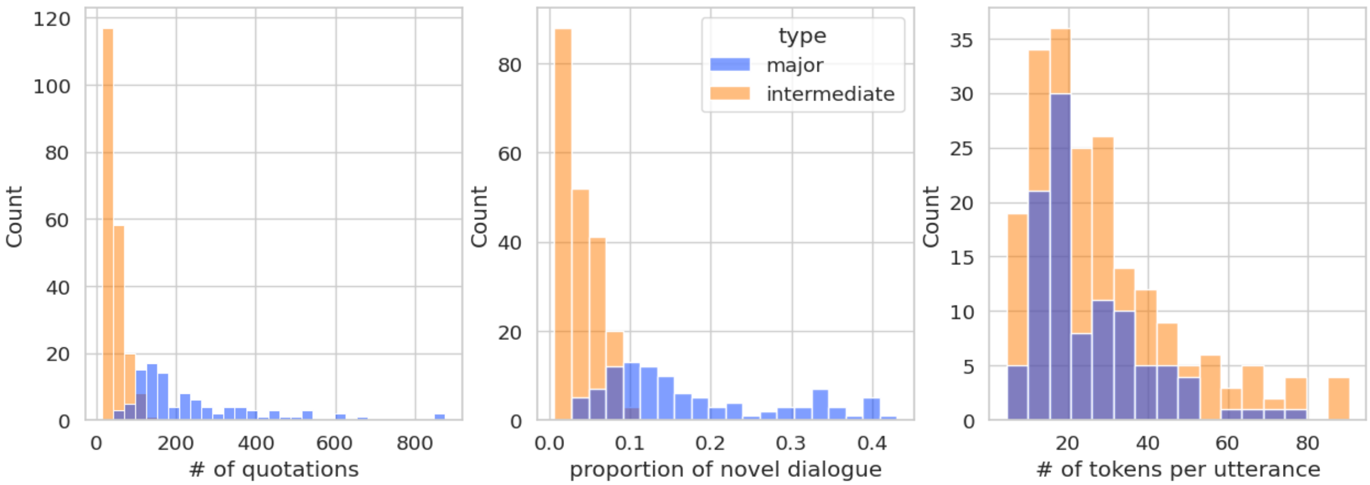


Figure 6.1: Distribution of the number of utterances by character, their proportion of the total dialogue in the novel, and average token length of their utterances.

those same norms. We study how male and female authors differ in their characterization of male and female voices and interactions in their novels using the framework of linear mixed-effect models.

## 6.3 Data

All of the required data for these experiments comes from the Project Dialogism Novel Corpus (PDNC), described in Chapter 5. We consider all 28 novels in PDNC, written by 19 authors, comprising 36,926 unique utterances in total from 809 speakers. For many of the subsequent analyses, we enforce a minimum threshold of 15 utterances by a character in order for that character to be considered, which results in the long tail of minor characters being eliminated. The final dataset consists of 34,273 utterances in total from 307 unique characters. Figure 6.1 plots the distribution of the number, proportion, and token length of the utterances by character type for this subset.

**Data chunking** Previous research in computational stylometry has shown that stylometric features are generally more reliable over larger spans of text (Hirst and Feiguina, 2007; Eder, 2015). Generally, word chunks of anywhere between 500 to 2000 tokens are used for both classification and clustering, with larger chunk sizes resulting in better accuracies (Burrows, 2002).

From 6.1, we note that the average number of tokens in character utterance falls somewhere between 15 and 25 tokens; we hypothesize that not many reliable stylometric indicators can be found in such short spans of text. We therefore also experiment with data chunking, wherein the set of utterances by each speaker is divided into chunks consisting of $k$ tokens each, where $k$ is varied between 500, 1000, 1500, and 2000 as a hyperparameter of interest. We also enforce a minimum number of 10 chunks per speaker, which is necessary especially when working with classification models. Given the low amount of speaker text available in most novels, testing with larger chunk sizes results in very little data to work with; this problem is exacerbated when we further delineate between speaker utterances addressed to different characters. We tend towards lower chunk sizes (500 or 1000) when larger amounts of text samples are not available for a particular experimental method.

In order to divide a speaker's utterances into chunks, we first concatenate all their temporally-ordered quotations to form one large document. Chunk boundaries are then determined at utterance boundaries, i.e, we don't break up an utterance into separate chunks (note that an utterance can comprise multiple sentences). This results in roughly equal-sized text blocks, with some deviations on either side for very large utterances.

## 6.4 Feature Extraction

We create five main sets of linguistic features to experiment with, taken from various prior works in computational stylometry and literary analysis (Stamatatos, 2009).

### Linguistic Features

**FS1: Most Frequent Words (MFW)** The most widely-used feature set in stylometry, this measures the frequency of usage of the top-$k$ most frequent words across all utterances in the comparative dataset of utterances. In experiments where we measure the accuracy of speaker identification, we take the set of 50 most-frequent words, where frequency is computed from the utterances of all the characters in that novel (since a separate classifier is trained for each novel).

**FS2: Stylometric features** The set of 89 features here, detailed in Table 6.1, are consolidated from prior work in computational stylometry and authorship attribution. Though some previous research has shown that author style subsumes any variations in the utterances of individual characters, we explore these here for posterity, and because we are working with a much larger dataset of texts. In this table, the term *character* is used to refer to the individual units that make up a token.

| **Lexical Features — Character-Level** |
|---|
| 1. Characters count (N) |
| 2. Ratio of digits to N |
| 3. Ratio of letters to N |
| 4. Ratio of uppercase letters to N |
| 5. Ratio of tabs to N |
| 6. Frequency of each alphabet (A-Z), ignoring case (26 features) |
| 7. Frequency of special characters: <>%\|{} []/\@#˜ +-*=$ˆ &_()’ (24 features). |
| **Lexical Features — Word-Level** |
| 1. Tokens count (T) |
| 2. Average sentence length (in characters) |
| 3. Average word length (in characters) |
| 4. Ratio of alphabetic characters to N |
| 5. Ratio of short words to T (a short word has a length of 3 characters or less) |
| 6. Ratio of words length to T. Example: 20% of the words are 7 characters long. (20 features) |
| 7. Ratio of word types (the vocabulary set) to T |
| **Syntactic Features** |
| 1. Frequency of Punctuation: , . ? ! : ; ’ ” (8 features) |

Table 6.1: List of stylometric features adapted from Altakrori et al. (2021).

## Lexicon-based Lexical Features

Apart from the word- and character-level features above, we incorporate lexicon-based features that capture text features along a set of higher-level, interpretable lexical dimensions, like emotionality, tone, and formality. These are extracted using word-level lexicons for each dimension, which associate words with a real-valued score representing the intensity of that word along that particular dimension (for example, the word *dejected* conveys a high level of sadness; the usage of the word *nevertheless* indicates a very formal tone). Word lexicons have been widely used in tasks involving sentiment and emotion analysis in NLP; here, we additionally use a lexicon of stylistic dimensions.

These features allow us to capture *voice* in a more literal sense: does the person (character) tend to speak with a higher-level of formality than their peers? Do they use more anger-related words than the average, or tend to be more joyful in comparison? Clearly, these scores are more informative for a qualitative understanding and characterization of voices than to say that a speaker has a higher relative usage of the word *to*.

**FS3: Lexical Affect and Emotion Scores** The third feature set is a collection of lexical scores for various dimensions of *affect*. Affect is a term that encompasses a broader range of feeling than conventional emotions like anger or joy, and incorporates dimensions that describe feelings and mood (for example, how *strong* or *weak* a particular concept is). Osgood et al. (1958) in their seminal work showed that word affect could be represented by three prominent dimensions: valence (a scale of good–bad), arousal (active–passive), and dominance (strong–weak). We use the NRC VAD lexicon for English that rates ∼20,000 words along each of these three dimensions, from a scale of 0 (low) to 1 (high).

Emotions, on the other hand, are usually described along a set of categorical dimensions: anger, joy, sadness, irritation, satisfaction, outrage, and so on. The basic emotions model in psychology posits that some emotions are more basic than others (Plutchik, 1980; Ekman, 1992); accordingly, we use here the NRC lexicons for the eight basic emotions of anger, fear, anticipation, trust, surprise, sadness, joy, and disgust, obtained from the lexicons created by Mohammad (2018b), which rank words from a scale of 0 (no emotion) to 1 (maximum intensity).

**FS4: Lexical Style scores** Analogous to works that classify words along aspects of style like formality and readability, Brooke and Hirst (2013a) create a lexicon of words associated with six aspects of lexical style. These six aspects represent the extremes of three dimensions, distilled from several linguistic theories on style: literary vs colloquial, objective vs subjective, and abstract vs concrete. Colloquial terms, as an example, consist of English slang and acronyms, like *cuz* and *lol*; the literary is represented by terms like *behold*, *amiss*, and *thine*.

## Content-based Features

**FS5: TF-IDF Counts** Finally, we vectorize the entire quotation (or chunk) text using TF-IDF counts of words, with no restriction on the tokens that can be included. This therefore potentially also includes content-indicating terms, however, as discussed previously, these word-choices cannot entirely be separated from stylistic choices. We mainly test this to quantify the jump in performance provided by including these features.

## 6.5 Speaker Identification

Let's start with our first research question (**RQ1**): are characters in a novel stylistically distinct from one another in the way they speak? We operationalize this in two ways: first, we test whether the identify of the speaker can be determined based on the features of the quotation text alone, in a classification setup; in the second method, we use clustering models to quantify at the extent to which these features cluster together by speaker. We test the performance of these methods at both the quote-level and chunk-level, where chunks are formed at different levels of token length (50, 100, 250, 500, 1000, 1500, and 2000).

**Classification:** Briefly, for each quotation (or chunk) in our dataset, we extract the set of features described in the previous section. The feature vectors are passed to a classifier that is trained to predict the the speaker in an $n$-way classification setup. Classification is done at the novel-level, i.e, we train the classifier to distinguish the speaker from the other characters within the novel, rather than at a global-level involving all possible characters from all 28 novels as candidates.

We use a Logistic Regression classifier, and use oversampling to correct from class imbalance in all novels. We report the $F_1$ score to quantify the performance of the classifier. We test multiple combinations of the 5 feature sets described in the previous section, and treat this as a hyperparameter to be optmized over.

**Clustering:** Similar to above, we extract a feature vector for each quotation/chunk in the dataset, and cluster the vectors into $K$ groups. We vary the value of $K$ from a minimum of 2 to a maximum of the number of actual characters in the data (the true number of clusters). We evaluate clustering performance using the homogeneity, completeness, and v-score measures, where the latter is the harmonic mean of the former two metrics. The v-score tells us how pure (is each character in its own cluster?) and how complete (are all the datapoints of a character in a single cluster?) the final clusters are. A high score along these metrics indicates that feature vectors cluster together well by the speaker of the associated utterances, thereby indicating consistent and distinct character voices. We experiment with both K-means and the Ward linkage clustering algorithms.

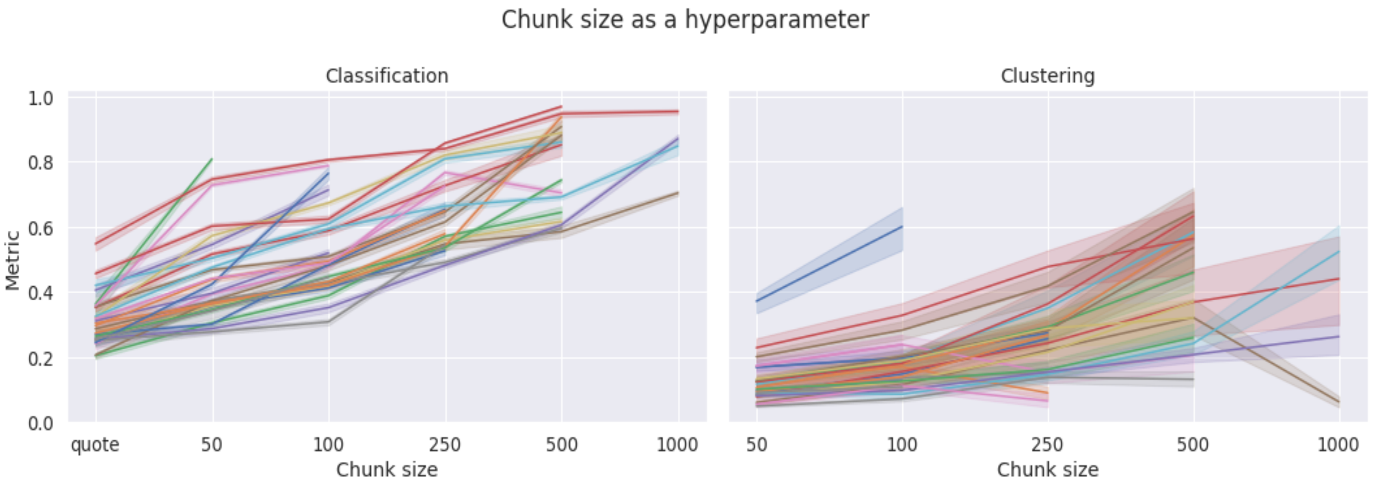


Figure 6.2: Variation in classification (micro $F_1$) and clustering (v-score) performance with different chunk sizes. Each novel is represented by a line on the plots; the shade around each line indicates the variance in the metric due to other hyperparameters (feature sets and method).

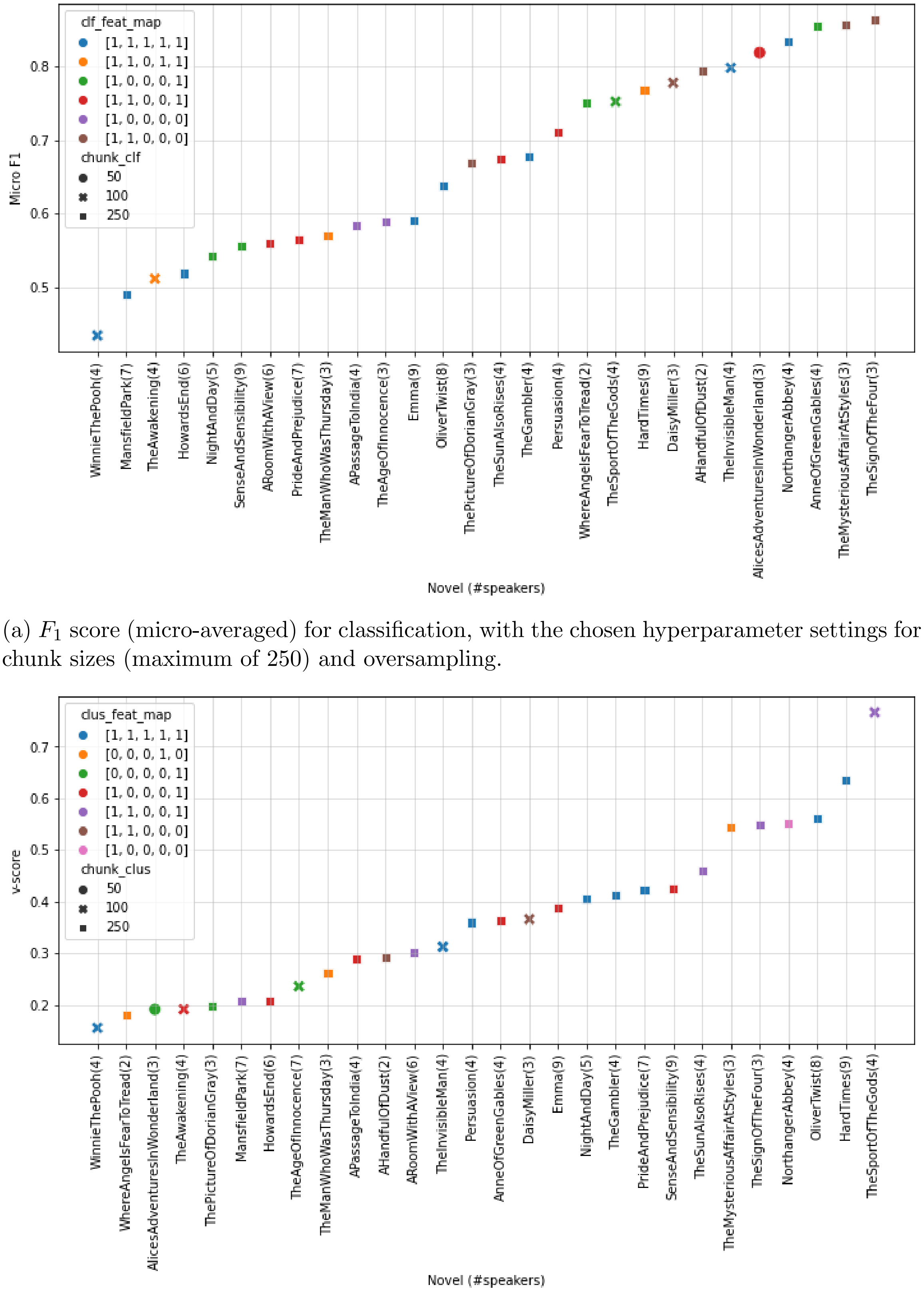
### 6.5.1 Results

Before diving into the results, we briefly examine the effect of various hyperparameters on classification performance.

#### Effect of Hyperparameters

We vary the following hyperparameters for each novel:

- **Chunk size:** As mentioned previously, we vary the input size to be of varying token lengths. We test with inputs at the individual quotation level (*quote*), and chunk sizes of length 50, 100, 250, 500, and 1000 tokens.
- **Feature sets:** We test all possible combinations of the 5 feature sets described in Section 6.4. For simplicity of reporting, we present the selected feature sets as a 5-dimensional boolean vector that is indexed by the value 1 if the feature set is used, and 0 otherwise.
- **Method:** For both classification and clustering, we test two variations in the methodology: for the former, we test classification performance with and without oversampling the minority classes, and for the latter, we vary the clustering method between the *K-means* and hierarchical *Ward* algorithms.



(a) $F_1$ score (micro-averaged) for classification, with the chosen hyperparameter settings for chunk sizes (maximum of 250) and oversampling.

(b) *v-score* for clustering, with the chosen hyperparameter settings for chunk sizes (maximum of 250) and the hierarchical Ward algorithm.

Figure 6.3: Best-performing configurations for classification and clustering of speaker identity with utterance features.

**Effect of chunk size** Figure 6.2 plots the variation in the evaluation metric – $F_1$ score for classification, *v-score* for clustering – with the token lengths of the input chunks. Overall, an increase in the chunk size corresponds to a better performance, though this trend is less pronounced for clustering. As discussed before, the danger of choosing ever-increasing chunk sizes is the reduction in the number of characters that can be analyzed. This is quite clearly represented in the two plots by the abrupt dropping off of the lines for certain novels at each token length boundary — one novel, *Alice in Wonderland*, has only one speaker (Alice) with sufficient utterances to be considered a classifiable character at chunk lengths of 50 tokens; 22 novels make it to the 250-chunks mark, 16 to the 500-word boundary, and only 4 novels have enough utterances for at least two of their characters that can be divided into a minimum of 10 chunks of 1000 words each (remember that we enforce a minimum of 10 datapoints per character for our experiments).

Clearly, there is a tradeoff that we must make here between the number of characters in the novel that we want to consider for analysis, and the effectiveness of our stylistic representations. We choose here a threshold for chunk size that allows us to keep all of the major characters, and some of the most-voluble intermediate characters, in the dataset — accordingly, we set 250 tokens as the maximum chunk length for all characters in all novels (the chunk length for a particular novel is allowed to drop below 250 (quote, 50, or 100) if those divisions result in a better performance).

**Effect of Method** With the input size set, we now look at the variation in performance with the method: whether to oversample for classification, and choosing between K-means or the Ward clustering algorithm. We find that, on average, oversampling results in an improved $F_1$ score across all novels, and the hierarchical Ward clustering algorithm performs better than K-means clustering.

**Feature Sets** Next, we select the feature combination that results in the best cross-validation score for each novel independently. We choose not to enforce a uniform feature set across all novels, as it can be an informative indicator of the stylistic idiosycracies of each novel.

In Figure 6.3, we plot the best-perfoming set of hyperparameters for each novel, for both classification and clustering. Note the wide variety of feature sets that work best for different novels; in Figure 6.3a, while TF-IDF features that incorporate content information are included in the feature sets for most novels, we are able to distinguish the top-3 speaking characters in *The Sign of The Four* (Sherlock Holmes, Dr. Watson,

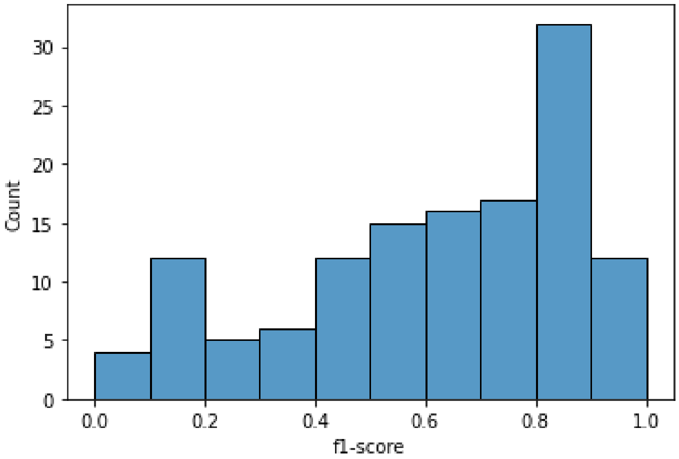


Figure 6.4: Histogram of the $F_1$ scores of classification for individual characters.

and Jonathan Small) with a high accuracy ($F_1$ score of 0.83) using just FS1 and FS2 (most frequent words, and stylometric features).

Similarly, in Figure 6.3b, for *The Mysterious Affair at Styles*, we see that just FS4 — the six lexical features of style — result in the best clustering performance (*v-score* 0.54) for the utterances of the top-3 speaking characters: Poirot, Mr. Hastings (the narrator), and John Cavendish. In contrast, for *The Sport of the Gods*, we observe a high *v-score* of 0.76 with word frequencies (MFW and TF-IDF) as well as syntactic and surface-level features from FS2.

### Qualitative Analysis

How do we perform at speaker identification at the level of individual characters? In Figure 6.4, we plot the histogram of $F_1$ scores for all eligible speakers in our dataset. The overall performance leans towards higher scores of classification, with 12 characters achieving a score above 0.9. Three characters — Stephen Blackpool and Mr. Sleary from *Hard Times*, Fannie Hamilton from *The Sport of the Gods* — have a perfect accuracy of 1.0.

Clustering, being an unsupervised algorithm, often leads to less reliable results than an optimized classifier. We check distinctiveness of character voice with clustering by computing a *cohesion score* for each character as a function of the distribution of their utterances or chunks among the formed clusters. A character whose chunks all fall perfectly into one cluster will have a perfect cohesion score of 1.0; if *most* of their utterances fall into one cluster, and a smaller proportion is scattered around the other clusters, this score falls sightly below 1. The lowest cohesion score is obtained for a character whose utterances are equally scattered across all clusters.

We see two characters achieving a perfect clustering score: Mr. Sleary from *Hard Times*, and Fannie Hamilton from *The Sport of the Gods*, agreeing with the classification performance above. They are followed by Mrs. Hilbery from *Night And Day* ($F_1$ score 0.87), and Stephen Blackpool from *Hard Times*.

On the whole, however, we don't see a very high correlation between the individual $F_1$ scores and the cohesion scores for characters (Spearman correlation: 0.52). When the 28 novels are ranked by their averaged-character scores for both methods, the correlation between rankings jumps to 0.64.

## 6.6 Modeling Effects of Gender

The results of the previous section tell us that the consistency of stylometric features across all the utterances of a character is quite inconsistent; the accuracy of classification (or clustering) varies markedly across characters and across novels. Can we explain this speaker variation in style using features other than speaker identity? One can imagine certain lexical features of utterances changing as characters interact with different entities and are placed in different situations throughout the narrative. How can we model the effects of these contexts on character voice? We can first attempt to enumerate several such factors: the set of characters being addressed by the speaker and the relationships between them, the social situation in which the interaction is taking place (a formal ball, an intimate conversation, a speech), the social background of the characters, and external events in the narrative that affect these socio-demographic features. We would need to annotate each utterance in the novel with information pertaining to these contexts, potentially categorize them into a set of nominal labels, and include them as influencing factors on utterance features along with speaker identity.

Unfortunately, obtaining such annotations is no easy task. In this section, we isolate character gender (as defined in Section 4.2.1) as a social factor of interest, and consider the effect of the genders of speakers and addressees within the novel. Do male and female characters occupy certain stereotypical roles in a story that shape the features of their utterances in consistent ways? Do female characters carry most of the emotional content of a novel? Do male characters change their way of speaking when talking to female characters, as opposed to other male characters?

Next, we step out of the fictional world of the novel to consider an external context: the influence of the author on the stylistic makeup of their characters' voices.

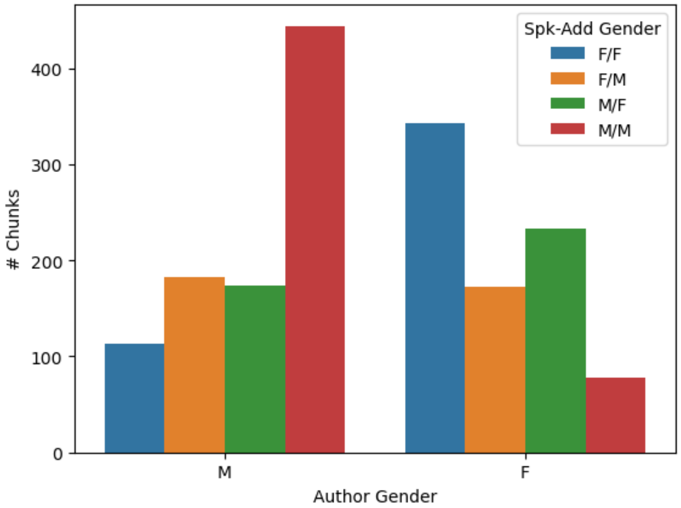


Figure 6.5: Histograms of the number of utterance chunks for each two speaker and addressee gender groups, by author gender.

Rather than model the effect of author identity, which severely limits the number of datapoints available for study (most authors have only authored one novel in our dataset), we examine if author gender (as defined in Section 4.2.1) *modulates* the effect of speaker and addressee gender on stylistic utterance features in a consistent manner, i.e, if male and female authors differ in consistent ways in how they portray male and female characters.

We create our dataset for this analysis by concatenating and chunking the utterances of all unique speaker–addressee pairs across the novels in our dataset. As before, we eliminiate the long tail of minor speakers (those with $<$ 15 utterances), and chunk into 500-token portions (we refrain from larger chunks to preserve sufficient datapoints).

Rather than using classification or clustering models with our further-diminished dataset of chunked utterances, we turn to simpler (in some senses) methods of statistically modeling dependencies between multiple variables. Specifically, we use the framework of **linear mixed-effect models**, which allow us to model differences in the mean values of a variable across several groups, and allow us to account for the non-independence of data points within a group (which, as will be explained shortly, is a key characteristic of our dataset). Our final condensed dataset consists of 1,823 dialogue chunks, uttered by 283 unique speakers (219 male, 64 female) to 281 unique addressees (217 male, 64 female).

In figure 6.5 and 6.6, we plot the distributions of the number of utterance chunks between male and female characters for both groups of authors, and their distribu-

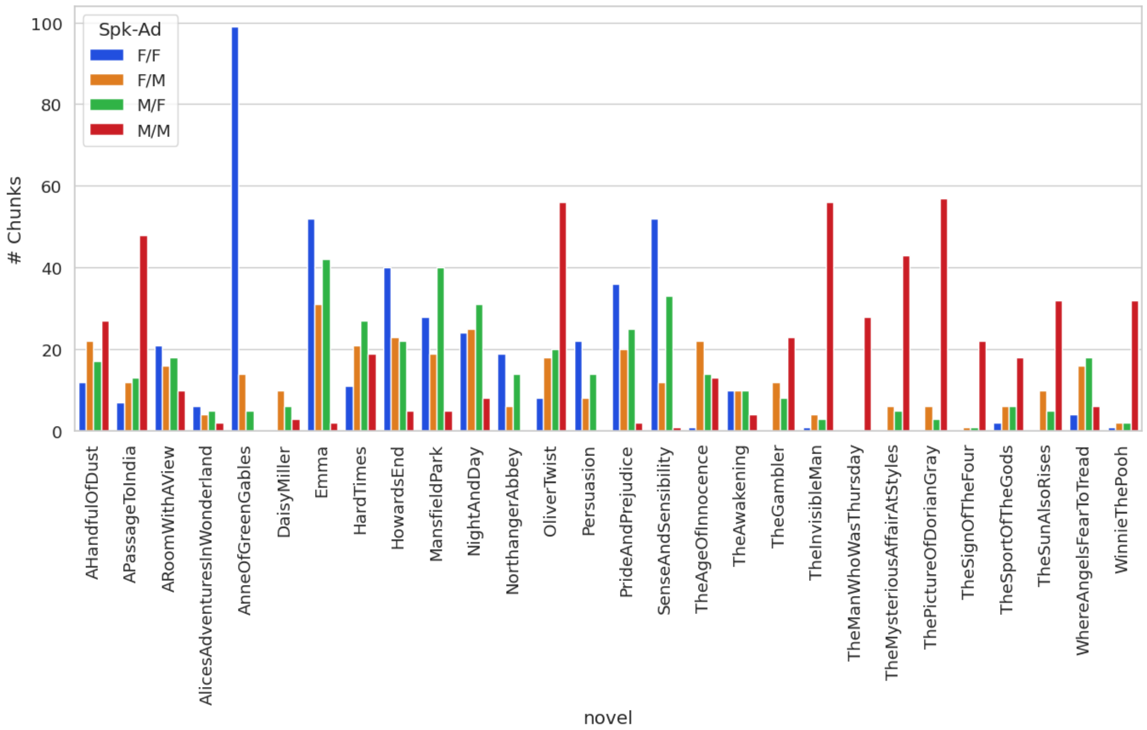


Figure 6.6: Histograms of the number of utterance chunks for each of the two speaker and addressee gender groups, by novel.

tions by novel. Clearly, there is a huge imbalance in the amount of dialogue that is exchanged by the different gender groups; some novels in the dataset have 0 interactions between female characters that constitute more than 500 words of text.

Given this imbalance, it is hard to draw a generalized conclusion on the effects of author, speaker, and addressee gender on the stylistic features of utterances. It could very well be that a higher average valence for interactions between female characters is *due to the tone of a particular novel or author*, rather than a gender-influenced effect. In other words, the distribution of the lexical scores we want to compare are affected by, and potentially correlated within, other grouping variables (the novel, and the author), external to the grouping by the gender variables.

### 6.6.1 A Brief Introduction to Linear Mixed-Effect Models

A common method for testing whether two (or more) sets of numerical scores differ from one another significantly is to use a t-test (or the ANOVA, for multiple groups). Equivalently, we are fitting a linear regression model, with the lexical score as the dependent variable and the gender group as the categorical, independent variable. From either viewpoint, a key statistical requirement is that of **independence** between the

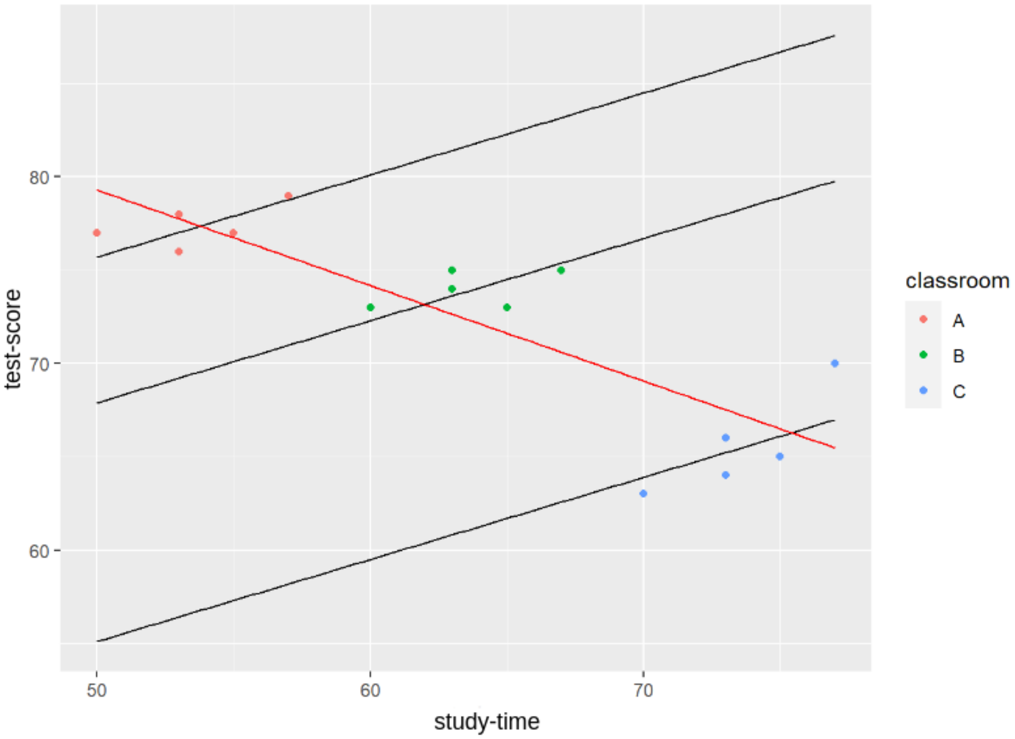


Figure 6.7: Simpson's Paradox: An illustration of the effect of correlations between groups[2].

sampled datapoints.

An intuitive example for why this is important is shown in Figure 6.7 (popularly referred to in statistical modeling as Simpson's Paradox). When modeling the relationship between `study-time` (x-axis) and `test-score` (y-axis), assuming independence of the data samples leads one to conclude that there is a negative relationship between the two (demonstrated by the red regression line); however, when considering that the students whose scores we sampled come from three distinct classrooms (an external grouping variable), the conclusion changes to a positive relationship between the variables, and an additional significant effect of classroom is apparent.

How then can we account for this correlation? A straightforward way is to add `classroom` as an additional dependent variable along with `study-time`, plus an interaction term `classroom * study-time`. This allows us to test how (a) `study-time` influences `test-score`, (b) `classroom` influences `test-score`, and (c) whether the effect of `study-time` is modulated in any way by `classroom` (say, whether the students in classrooms A and B improve on their test score with more study, but the opposite is true for students in classroom C).

The dependent variables that we explicitly compute the effects of, as above, are referred to in the statistical modeling world as **fixed effects**. Now, imagine that we measure a student's test score and study time four times in the year, once for each semester. We now have an additional grouping variable to account for: the student themselves. Different students can have their own relationship between the time spent on study, and their performance on the test. Should we include `student-id` as an additional dependent variable? We can hesitate for multiple reasons: we are not explicitly interested in the estimates of the effects for each individual student — we only want to account for this confounder when estimating the effects of `classroom` and `study-time`; the students we are analyzing here only form a (somewhat) randomly-sampled subset of the population of all students, and we are interested in making more generalized claims; the amount of data that is available to fit the regression model for each combination of dependent variables sharply drops off with each additional grouping. Extending this analogy to our dataset of character utterances, we could account for the within-novel correlation by testing for statistical significance only within the utterances of each novel, but this would allow us on average only about 10 scores to compare in the female–female interactions group.

The statistical solution to this problem is to include `student-id` as a **random effect**. By designating it a random effect, we are simply stating that we are not interested in considering it a fixed effect, and therefore not to devote too much statistical power towards computing the effects associated with each group of the random effect. Rather, the model will adjust the estimate of the fixed effects to account for the variation introduced by the random effects (say, the regression weights of the fixed effects may decrease).

Our dataset of chunked utterances and their stylistic scores follows a highly hierarchical structure: authors write multiple novels, each novel contains multiple characters, and each character speaks to several other characters. The independence assumption is therefore violated; scores can be correlated within the multiple chunks sampled between a speaker–addressee pair (though the small size of our dataset makes this quite rare), the interactions of a particular novel (this is more likely), and the interactions written by a particular author (considering that most authors have authored only one book in our dataset, this effect is largely subsumed by the one prior). We therefore model the dependence of the lexical features of utterance style on the three gender variables (speaker, addressee, and author) using the framework of **linear mixed-effect models**.

[2] Source: https://bookdown.org/anshul302/HE902-MGHIHP-Spring2020/Random.html

### 6.6.2 Constructing the Model

We construct mixed effect models for each of the six lexical style dimensions, and the three affective VAD dimensions.

**Random Effects** The first step towards building our model is to ask which external grouping variables should be included as random effects. The general recommendation is to err on the side of excess, and include all the variables which can potentially induce correlations — speaker, addressee, novel, and author identity. We test the effect of the random variables by checking the intra-class correlation (ICC) associated with each — a numerical proportion that tells us the strength of the correlation between the scores within each group, or alternatively, the proportion of variation in scores that is explained by that grouping variable.

An empirical ICC threshold for disregarding the effects of a random variable is considered to be 5%or lower. We see that author, speaker, and addressee identity must therefore be added as random effect grouping variables in any analysis of significant differences between scores based on gender.

We additionally validate the need for including random effects by comparing the log likelihood scores of a linear model that does not include random effects to a model that includes all four.

**Fixed Effects** With our random effects structure in place, we add in fixed effects for each of our dependent variables of interest — speaker gender, addressee gender, and author gender.

We follow a multi-step process to adding dependent variables, outlined below. We compare whether one model is significantly better than the other by using an ANOVA test, which compares the likelihood scores of the two.

1. **Spk model:** The categorical speaker gender variable `spk-gen` (with two levels, Male (M) and Female (F)) is the only dependent variable.

2. **Add model:** The categorical addressee gender variable `add-gen` (with two levels, Male (M) and Female (F)) is the only dependent variable.

3. **Spk–Add model:** We add both speaker and addressee gender as dependent variables. This model included the interaction term, `spk-gen * add-gen`, to test for modulating effects: do female speaker change their tone differently to male speakers, when addressing male and female characters?

| **Dimension** | `spk-id` | `add-id` | `auth-id` | `novel-id` | **Total** |
|---|---|---|---|---|---|
| `literary` | 0.132 | 0.133 | 0.337 | 0.133 | 0.618 |
| `colloquial` | 0.179 | 0.190 | 0.113 | 0.032 | 0.515 |
| `abstract` | 0.203 | 0.136 | 0.320 | 0.007 | 0.666 |
| `concrete` | 0.146 | 0.183 | 0.233 | 0.021 | 0.583 |
| `subjective` | 0.031 | 0.070 | 0.094 | 0.019 | 0.213 |
| `objective` | 0.229 | 0.165 | 0.0 | 0.225 | 0.618 |

Table 6.2: ICC scores for the individual random effects, and their combined total, for each lexical style dimension.

4. **Author model:** The categorical author gender variable `auth-gen` (with two levels, Male (M) and Female (F)) is the only dependent variable.

5. **Joint model:** We add all three gender categories as dependent variables. This model includes all possible interaction terms between the groups, effectively comparing the differences in group means of scores between all possible gender-based groupings (male-authored male characters talking to other male characters, female-authored male characters talking to other female characters, and so on).

### 6.6.3 Results: Lexical Style

**Random Effects:** The likelihood of a base model with random effects (and no fixed effects) is significantly ($p < .001$) higher than a base model with no random effects (and no fixed effects). This validates the need to include random effects in our linear model.

In Table 6.2, we list the ICC values for the four random effects in our base model, for each lexical style dimension. We note the following:

1. For all lexical style dimensions except `subjective`, the four random effects of speaker, addressee, novel, and author identity account for a major proportion (greater than 50%) of the variation seen in the scores of all the utterance chunks in our dataset. The high proportions demonstrate the non-independence of our datapoints.

2. Among the random effects, author identity contributes to a large proportion of variance observed in the literary (34.1%) and abstract (32%) dimensions. This means that the `literary` scores of utterances are quite clustered by the author

of the text. Given that most authors in our dataset are only represented by a single novel, we can say equivalently say that the scores cluster by novel. For authors that are represented by multiple novels, the differences in scores between these author-specific novels helps explains an additional 1.1% of the variation.

3. The effects of speaker identity and addressee identity also cannot be disregarded. A character in our dataset utters an average of 7 chunks, and is addressed by other characters an average of 6 times; these variables explain nearly 23% of the variation along the `objective` dimension, 13% of the variation each for the `literary` dimension, and so on.

**Fixed effects:** In Table 6.3, we report the fixed effects (including interaction terms) that emerged as having a significant effect on each lexical style dimension.

We find that once random effects are included, there is **no** consistent effect of author or addressee gender on any of these scores (in other words, none of Models 2–5 improve on the base, intercept-only random effects model).

A significant effect of speaker gender in observed for the `literary` ($p < 0.05$), `abstract` ($p < 0.007$), and `objective` ($p < 0.002$) dimensions — characters identified as Male tend to have higher scores for all three dimensions, compared to characters identified as Female (more literary, more abstract, and more objective), though the differences observed in group means are very slight.

Box-and-whisker plots of the distributions of these variables (not shown) also indicate several outliers, suggesting a skewed distribution that is not accurately captured by a normal distribution. This is turn suggests that linear regression models may not be the best way of modelling these variables; we leave an exploration of alternative methods to future work.

### 6.6.4 Results: Lexical Affect and Emotion Scores

For the eight categorical emotions, we observe that the distribution of scores is highly skewed, with a bulk of the values falling close to zero, and a rapidly-dropping tail of higher scores. Given that normality of values is usually assumed by mixed-effect models, we focus instead on the affective dimensions of valence, arousal, and dominance. Overall, all three dimensions have a normally-distributed set of values, with valence averaging at 0.61, arousal at 0.43, and dominance at 0.51.

| **Dim** | **Fixed Effects** | | **Random** |
|---|---|---|---|
| | **Gen-M** | **Estimate** | **Eff. (ICC)** |
| literary | spk | 0.011* | 0.618 |
| abstract | spk | 0.025** | 0.666 |
| objective | spk | 0.04*** | 0.619 |
| colloquial | – | | 0.514 |
| concrete | – | | 0.583 |
| subjective | – | | 0.214 |
| $^{*}p < 0.05$ | $^{**}p < 0.01$ | $^{***}p < 0.005$ | $^{****}p < 0.001$ |

Table 6.3: Effect sizes for fixed effects that are found to be significant, and the explained variance (ICC) for all random effects, for the six lexical style dimensions. Asterixes indicate the p-value threshold for significance.

| **Dim** | **Fixed Effects** | | **Random** |
|---|---|---|---|
| | **Gen-M** | **Estimate** | **Eff. (ICC)** |
| valence | auth | −0.039**** | 0.149 |
| | ad-auth | 0.0241* | |
| | spk-ad-auth | 0.029* | |
| arousal | spk | 0.005* | 0.144 |
| dominance | auth | −0.03*** | 0.228 |
| | spk-auth | 0.017* | |
| | ad-auth | 0.016* | |
| $^{*}p < 0.05$ | $^{**}p < 0.01$ | $^{***}p < 0.005$ | $^{****}p < 0.001$ |

Table 6.4: Effect sizes for fixed effects that are found to be significant, and the explained variance (ICC) for all random effects, for the three lexical affect dimensions.

Table 6.4 reports the effect sizes for significant fixed effects, and the combined explained variance (ICC scores) of the random effects, for each VAD dimension.

While including random effects improves the model fit, the effect of these random variables is notably smaller for affect dimensions when compared to stylistic dimensions. Random effects explain only 0.15, 0.14, and 0.23 of the unexplained variation in values for valence, arousal, and dominance respectively.

We observe the following significant effects of gender (of the speaker, addressee, and author) on these scores:

- Novels authored by male writers have lower average values for valence and dominance.

- Male characters have slightly higher values of arousal, irrespective of author and addressee gender.

- Interactions between male characters differ from those of the other gender groups in male-authored novels. Though male-authored novels have lower values of valence and dominance, male speakers in these novels have higher average dominance, dialogue addressed to male characters has a higher average valence and dominance compared to that addressed to female characters, and interactions between male characters have a higher valence compared to all other groups.

Overall, gender — of the authors and the characters — plays a significant role in explaining the affective features of utterances. On the other hand, there is not much consistency in affective features by speaker identity or author identity. Certainly, we expect that the emotional features of character utterances will change over the course of the narrative as the characters are placed in different situations, constituting what can be considered their *emotional arc*. The shape of such trajectories for stories has been explored in computational literary analysis; the shapes of individual character arcs within the story, and how they interact with one another, is less explored. We dive deeper into the VAD dimensions and their variation in the next chapter.

## 6.7 Conclusion

In this chapter, we took the first step towards quantifying the distinctiveness of character voices within a novel, using a large dataset of utterances annotated for speaker and addressee information. Our results demonstrate the range of stylistic variation that is seen in fictional texts, and the different ways in which this variation is expressed — our experiments with stylometric classification and clustering methods highlighted that there is no one uniform feature set that works across all novels or characters.

We also investigated the effects of gender, both of the characters and of the author, on certain lexical features of style and emotion. Our analysis with mixed-effect models showed that character identity and authorial tone (or the tone of the novel) strongly influence stylistic features of utterances, and overpower influence of gender-related effects, if any. This is further evidence of character voices being distinct from one another stylistically, both within a novel and across novels.

The same cannot be said of emotional features, where neither identity-related effects nor gender-related effects could explain the majority of the observed variation. In the next chapter, we explore emotional variation within novels in greater detail using the framework of Utterance Emotion Dynamics.

.

# Chapter 7

# Literary Emotion Dynamics

In this chapter, we continue our study of character voices and what makes them distinct in a novel. Here, we move away from style as formulated in computational stylometry, and focus more on the emotional content of a character's utterances, their temporal arc through the course of the narrative, and whether metrics characterizing this arc distinguish characters and novels.

## 7.1 Introduction

Storytelling is one of the earliest of all human traditions, predating even the invention of writing. Emotions are a powerful aspect of storytelling in any form, be it the oral tradition, or with the written word in fictional novels: a compelling story is often one that is able to evoke a strong emotional reaction from its audience. Characters (fictional named entities) are often at the center of this emotional connection that readers form with a story; we identify with their struggles, react to their pain, and celebrate their victories. The different characters within a single story usually display varying emotional trajectories through the course of the narrative – when the hero is in pain, the villain celebrates; when the villain is defeated, the rest of the cast is happy; protagonists of the story usually go through more intense emotional journeys compared to side characters. This variation is also seen at the level of the novel itself: some are tragedies, some are comedies; some have very dramatic highs and lows, others have a more even emotional tone.

Kurt Vonnegut, in his now-famous lecture[1](Vonnegut, 2009), spoke about the "shapes of stories", plotting the fluctuations of a *character's good or bad fortune*

[1] https://bigthink.com/high-culture/vonnegut-shapes/

on the y-axis as a function of narrative time, from beginning to end, on the x-axis. One can assume that he was referring to the emotional journey of the protagonist of the story in his analysis. Prior work in NLP on plotting the shapes of stories, however, has considered a novel as representing a single emotional trajectory, where the dialogue of all the different characters is merged together with the narration to create one overarching novel arc. This is somewhat of a necessary simplification because of the lack of annotated data in which characters are mapped to their utterances in novels. Automatic attribution of quotations to the characters that uttered them is still a challenging NLP problem. Here, we use the annotations from the Project Dialogism Novel Corpus (PDNC), introduced in Chapter 5, to look instead at the emotion arcs of the *individual characters* in story, distinguishing them from the arcs of the narration as well as the overall novel arc, and quantify the extent of their variation both within a single novel and across novels written by different authors.

Our goal is to quantitatively capture *longitudinal patterns* of a character's emotional states (how emotions change over time). We use the framework of Utterance Emotion Dynamics, first proposed in Hipson and Mohammad (2021), to derive metrics characterizing the emotion arcs of characters and novels. Through these metrics we examine the following research questions about emotion arcs and emotion change (at an aggregate level) in full-length English novels:

1. *The Emotion Dynamics of Novels*: How does the emotion change from the start of the novel to the end: overall, for just the narration, and for individual characters?

2. *Narration vs. Dialogue*: Do narration and dialogue have distinct emotion arcs? Prior work on the shapes of stories has largely glossed over the differences between characters, and considered either the entire novel to be a single trajectory, or worked with the text of the narration alone. We examine the validity of this approach by quantifying the differences between the emotions expressed in the narration and the dialogue of various characters.

3. *Diversity of Character Arcs*: How diverse are the emotion arcs of the characters in a story? The variety in the emotional trajectories of characters within a novel can be quite informative about the type of story it tells. A typical hero-versus-villain story, for example, might display opposing emotional arcs for the two main characters; one that follows the adventures of a group of friends might have more closely-correlated character arcs.

4. *Effects of Gender*: Are there consistent differences between the emotion arcs of gendered character groups? How does this change with different authors? Our stylometric analysis in the previous chapter indicated differences in the way male and female characters interact within novels. Here, we continue this line of research: are there consistent differences between the emotion arcs of female characters vs male characters? How does this change with the gender of the author?

To answer these questions we conduct experiments that make use of the following resources:

1. The *Project Dialogism Novel Corpus (PDNC)*, which contains manual annotations identifying the speakers of all dialogue in 28 full-length English-language novels.

2. The NRC Valence, Arousal, and Dominance Lexicon (Mohammad, 2018a), which includes ∼20,000 English words with a real-valued association scores (between 0 and 1) for the three dimensions of valence (V), arousal (A), and dominance (D) dimensions. A score of 1 indicates a maximum association or highest V/A/D, and a score of 0 indicates that the word is associated with the lowest V/A/D.

3. A simple, accurate, and interpretable way to generate emotion arcs from sequential text using an emotion lexicon (Teodorescu and Mohammad, 2022).

4. Metrics of Emotion Dynamics (Hollenstein, 2015; Kuppens and Verduyn, 2017), which quantify patterns of emotion change.

We find that novels, on average, express emotions high in valence and lower in arousal and dominance (0.65 vs 0.38 and 0.52). Most of the high valence and dominance is expressed in character dialogue rather than narration. We show that emotional arcs of characters are quite different from that of the narration, and from one another (average correlations close to 0); the extent of this variation also changes from novel to novel. We find that female authors write characters who express higher valence, lower arousal, and higher dominance; male characters written by male authors have the highest arousal for their utterances.

Our work sheds light on aspects of storytelling that have been under-explored in computational literary studies, and quantitatively demonstrates the importance of centering characters in order to gain a more nuanced understanding of novels.

## 7.2 Background: Utterance Emotion Dynamics

Emotion dynamics is a framework from psychology for measuring how an individual's emotional state changes over time (Hollenstein, 2015; Kuppens and Verduyn, 2017), and has been studied in connection with many downstream consequences to emotional, mental, and physical health. Hipson and Mohammad (2021) introduce a method to compute a sequence of emotional states (the *emotion arc*) associated with an individual over time, derived from their natural language utterances over that time period, along with a set of derived metrics characterizing this arc. They term this framework Utterance Emotion Dynamics (UED). The authors primarily characterize an emotion state as a point in the 2-dimensional valence–arousal space, and the emotion arc as a trajectory following the sequence of states in this space. In this work, we limit ourselves to analyzing emotions individually in a one-dimensional space where the temporal flow is represented on the x-axis and the emotion state values are on the y-axis.

Given a sequence of temporally ordered utterances, the emotion state at a time point is defined as the average emotion value of a small window of utterances (or words) uttered around that time point. This window is moved forward by one word at each step to obtain a sequence of temporally-ordered emotion states (the overlapping windows lead to a smoother and more continuous arc, when compared to using non-overlapping adjacent windows).

The *home base* for an individual is the space of emotion states, or values, that they are most likely to be found in, i.e., the range of most-probable values, where most-probable is usually defined as the range within which 68% of the speaker's state values are likely to fall. The home base is therefore captured by two values: the **mean**, or average, emotion state value, and therefore the center of the home base; and the **variability** (found by computing the standard deviation), which defines the bounds on either side of the centre. These bounds can be visualized as an ellipse in the 2D space; for a single dimension, it will define a range of numerical emotion state values.

Any movement outside of the range of this home base is termed a *displacement*. Displacements in turn are characterized by a series of metrics:

1. **Displacement length**: The length, in temporal steps (equivalently, number of window steps; equivalently, number of words), of the displacement — from the moment the speaker exits the home base, to when they return to the home base.

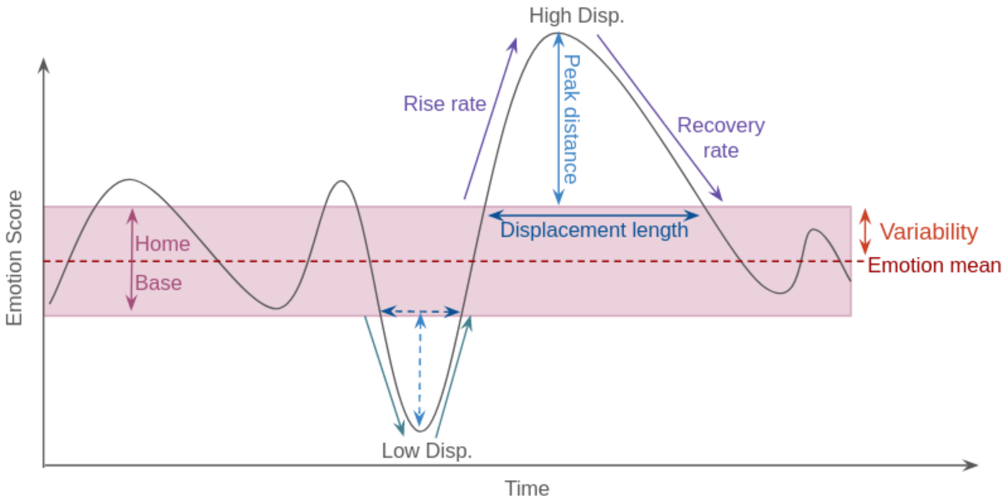


Figure 7.1: A visual representation of the various UED metrics for a sample emotion arc.

2. **Peak Distance**: The furthest that the speaker travels from the home base bounds, in emotion state value. For a 2D space, this is the maximum distance from the home base ellipse in the displacement; in the 1D space, we measure the maximum distance along the y-axis to the closest home base boundary.

3. **Rise and Recovery Rates**: The two values above allow us to compute the *rate* at which the speaker ascends to the peak emotion state, termed the rise rate, and the rate at which they climb down from the peak back to the home base (termed the recovery rate).

Figure 7.1 visualizes a sample emotion arc, and the UED metrics that are derived from this arc.

In the 1D space, a displacement can occur when the speaker's emotion state goes above the upper bound of the home base, in which case we name it a HIGH displacement, or when it goes below the lower bound of the home base, corresponding to a LOW displacement. We also break down the rise and recovery rates into separate rates corresponding to when one is moving from the home base to higher emotion values (high rise rate, Hm–Hi), from the highest value to home (high recovery rate, Hi–Hm), from the home base to lower emotion values (low rise rate, Hm–Lo), and from the lowest value to home (low recovery rate, Lo–Hm). These metrics are averaged over all the displacements of the speaker throughout their temporal arc to obtain the average displacement length, average peak distance, and average rise and recovery rates. Appendix 7.A lists all the UED metrics that characterize an emotion arc in more detail.

### 7.2.1 Emotion Scoring Method

The emotion value of a window of words can be determined in many ways: with lexicons that associate words and phrases with a numerical score along a particular emotion dimension, or with statistical and neural models that are trained to predict an emotion score given a text span as input. Lexicon-based methods are simpler and more interpretable, and do not need domain-specific training or fine-tuning of models. Teodorescu and Mohammad (2022) also showed that lexicon-based methods were able to match the true emotion arc of a text sequence with a high accuracy (correlations above 0.9) when using window sizes of a 100 words or more; Ohman et al. (2024) echo this finding for modeling literary emotion arcs in particular. We therefore follow this approach in our work.

Emotion lexicons in turn have been created for many emotion dimensions, and in multiple languages (LIWC (Tausczik and Pennebaker, 2010); WordNet-Affect (Bobicev et al., 2010), SentiWordNet (Baccianella et al., 2010), VADER (Hutto and Gilbert, 2014), and the NRC suite of emotion and affect lexicons (Mohammad, 2018b,a)). Here, with our focus on the affective dimensions of valence, arousal, and dominance, we choose the NRC VAD lexicon to construct emotion arcs; we discuss this further in Section 7.3.1. As is good practice with lexicon-based analysis, we remove high-frequency terms that are used in our corpus with a sense that is different from the sense annotated in the lexicon (like *will* and *have*).

## 7.3 Literary Emotion Dynamics

There are several emotional trajectories of interest given a novel: one can look at the emotions of entire novel text, as has been done in prior NLP work, or consider the narration and dialogue as two text streams of interest, representing the narrator and the characters (we term these text streams as being uttered by *meta-speakers*). The dialogue can further be analyzed by considering each character's utterances as an individual text stream of interest.

We categorize the characters in a novel into three groups based on the volume of their dialogue: *major characters* are those who contribute at least 10% of the total dialogue in the novel or have at least 100 attributed quotations; *minor characters* utter fewer than 35 quotations throughout; the rest are labelled *intermediate*. Table

| **Speaker Type** | | **Count** | **#tokens** |
|---|---|---:|---:|
| Meta | novel | 28 | 96973.96 |
| | narration | 28 | 57995.43 |
| Characters | major | 111 | 6547.91 |
| | intermediate | 113 | 2025.35 |
| | minor | 585 | 231.99 |

Table 7.1: The number of speakers, and average number of tokens, for each group of speakers that we consider, including meta-speakers.

7.1 presents the number of speakers and the average number of tokens for each of these speaker groups.

We order utterances for each character by their position in the text of the novel (*fabula*, the order in which events are presented), not considering the underlying chronological sequence (*syuzhet*). This is of substantial interest for literary analysis as it is the order in which the reader experiences the text uttered by a character.

The emotion arc for a speaker is computed with a rolling window size of 500 words, with the window moving forward by one word at each step until the final window subsumes the final word of their utterances – we stop as soon as the final word is included. We normalize the time for each speaker to lie between the range $[0, 1]$, i.e, each speaker starts their emotion state at time point 0, and ends at time point 1, irrespective of the volume of their dialogue. The corresponding aggregate UED metrics (average rise and recovery rates, variability, displacement lengths, etc) are computed based on this arc.

In order to compute the emotion state at a particular timepoint (i.e, for a window of words), we use the NRC-VAD lexicon Mohammad (2018a). The emotion arcs for each of the three dimensions of valence, arousal, and dominance, are individually constructed and analyzed.

### 7.3.1 Emotion Dimensions

In the previous chapter, we were briefly introduced to the three affective dimensions of valence, arousal, and dominance, collectively referred to by the acronym VAD. Here, we present a slightly expanded introduction that touches on what these dimensions capture, and what they are likely to tell us about characters and their arcs.

The VAD model is a psychological theory of emotions developed in the 1980s (Russell, 1980), which posits that core affect (emotions, feelings, moods) can be decomposed into these three dimensions. **Valence**, also called the Pleasure dimension, or alternatively the Evaluation dimension, measures how pleasant or unpleasant a particular emotional stimulus feels (the scale is also often described using the alternative labels of positive–negative, good–bad, desirable–undesirable). Emotions such as joy and excitement have a high, positive valence; fear, sadness, and disgust have a low, negative valence; and ambiguous emotions like surprise and anticipation will have a somewhat moderate valence (a surprise can be good or bad).

The *Arousal* dimension is a measure of the activity or energy evoked by an emotional state, a dimension of active–passive, stimulated–relaxed. While both rage and anger are negative valence emotions, rage has a higher arousal; similarly, excitement and serenity are both positively valenced, but the latter has a lower arousal.

The *dominance* dimension is related to feelings of control over the emotional state, linked to adjectives like powerful–powerless, dominant–submissive. Fear, for example, is a low dominance, submissive emotion, compared to confidence (valence and dominance are positively correlated). A person who feels in control in a particular situation or interaction will therefore use high dominance words.

The three VAD dimensions roughly correspond to the ones proposed in Osgood et al. (1957) as the three fundamental dimensions of word meaning, intended to the capture the subjective aspects of one's perceptions of, and reactions to, concepts. Termed the semantic differential (SD), this measurement scale is used to assess one's opinions, attitudes, and values to various objects and events; alternatively, these can be viewed as the *connotative* aspects of meaning, as opposed to the denotative (definitive) aspects. In the paper, survey participants were asked to choose where their position lies, with respect to a set of objects, words, and symbols, along several scales described with polar adjectives: sweet–bitter, fair–unfair, warm–cold, and so on. A dimensionality reduction of these ratings revealed three foundational factors, which were referred to as Evaluation (E), Activation (A), and Potency (P) – mapping (somewhat imperfectly) onto the V, A, and D dimensions (see Bakker et al. (2014) for a detailed discussion).

Even words that are not explicitly associated with emotion convey a certain affect (an aspect of their connotative meaning). Computationally, researchers have attempted to create lexicons that associate words with their implied intensity along each of three directions. Prominent efforts include ANEW from Bradley and Lang (1999), who obtained ratings for each dimensions on a 9-point rating scale for more than 1,000 words; the lexicon from Warriner et al. (2013) for 14,000 words which used a similar rating scale; and the NRC VAD lexicon for more than 20,000 words from Mohammad (2018a), which eschewed rating scales in favor of a comparative,

| Valence | | Arousal | | Dominance | |
|---|---|---|---|---|---|
| **Word** | **Score** | **Word** | **Score** | **Word** | **Score** |
| magnificent | 1.00 | aggressive | 0.971 | powerful | 0.991 |
| freedom | 0.969 | eruption | 0.913 | impress | 0.895 |
| resolute | 0.771 | vigor | 0.755 | safe | 0.759 |
| worry | 0.245 | solemn | 0.370 | gullible | 0.318 |
| bankrupt | 0.163 | patient | 0.193 | novice | 0.180 |
| toxic | 0.008 | siesta | 0.046 | frail | 0.069 |

Table 7.2: Representative terms from the polar extremes of VAD dimensions, sourced from the NRC-VAD lexicon.

| Speaker | Valence | | Arousal | | Dominance | |
|---|---|---|---|---|---|---|
| | **Mean** | **Var.** | **Mean** | **Var.** | **Mean** | **Var.** |
| Novel (meta) | 0.6472 | 0.0531 | 0.3811 | 0.0466 | 0.5192 | 0.0582 |
| Narration (meta) | 0.6273 | 0.0567 | 0.3857 | 0.0462 | 0.5076 | 0.0603 |
| Character | 0.6746 | 0.0295 | 0.3766 | 0.0282 | 0.5380 | 0.0355 |

Table 7.3: Aggregated Mean and variability (Var.) scores for the different types of speakers in our dataset, for each of the three emotion dimensions.

Best-Worst Scaling approach. Comparative annotations are generally accepted to be more reliable than normative rating scales, and we therefore choose the latter to conduct all our analyses. In Table 7.2, we list words representing the extreme ends of the VAD dimension, sourced from this lexicon[2].

## 7.4 Research Questions and Experiments

We now explore a series of research questions (RQs) on the emotional dynamics of literary novels, the diversity of emotion arcs that can be found within a novel, as well as across novels, and how they relate to one another.

### 7.4.1 The Emotion Dynamics of Novels (RQ1)

We generated emotion arcs and computed UED metrics for all the novels in PDNC.

Figure 7.2 shows the distribution of the mean and variability for the entire novel, the narration, and individual characters. Table 7.3 lists the aggregated scores for mean and variability. All other UED metrics are reported in Tables 7.4-7.6 in Appendix 7.B.

[2] http://saifmohammad.com/WebPages/nrc-vad.html

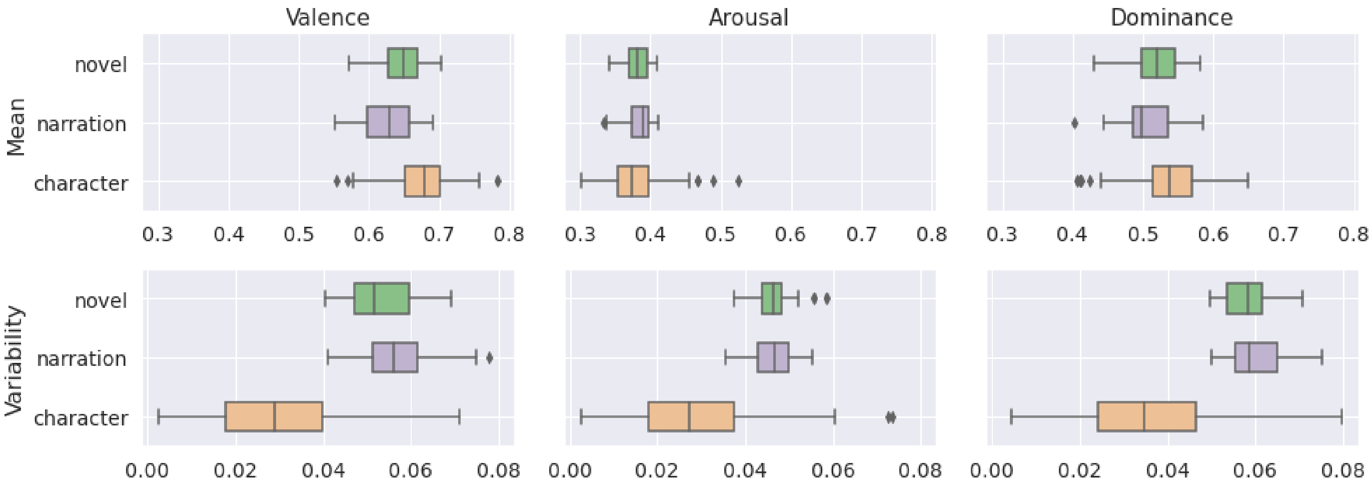


Figure 7.2: Boxplots of the distributions of mean and variability for each of the three affective dimensions and all novels in PDNC, when the entire novel and the narration are considered to be uttered by a single meta-speaker, and when each character's utterances are considered individually.

It is immediately apparent that there is much more variation in the emotion dynamics of individual characters than when the narrative and/or the entire novel are clumped together into a single voice. We quantify these distinctions in detail in the next section; here, we look at the overall distributions of these metrics for our dataset.

**Emotion mean:** On average, novels (when considered as a single speaker) have a mean value of 0.65 for valence, 0.38 for arousal, and 0.52 for dominance. Narration alone has lower mean values of valence and dominance, and about the same for arousal. Character dialogue, on the other hand, has a higher level of valence and dominance compared to narration — the third quartile for narration is close to the first quartile for dialogue — while the mean arousal is slightly lower, the spread of values exceeds the range for narration on either end of the distribution.

**Variability:** The average variability for novels (as a whole) lies around 0.053 for valence, 0.058 for dominance, and 0.046 for arousal. Narration has consistently higher average values for variability along all three dimensions when compared to character dialogue (nearly double), but the latter again covers a wider range of values.

**Peak Distance and Displacement Lengths:** On average, character dialogue has smaller peak distances when compared to narration across the three dimensions (0.016

vs 0.025 for valence). Displacements below the home base for valence have (significantly) higher peaks compared to those above the home base (for both narration and dialogue). Specifically for character dialogue, the average length of low valence displacements is much higher than all other displacements (106 words, compared to 87 for high valence, and an average of 87 and 82 words for arousal and dominance respectively). This indicates that the most common displacement in the emotional states of characters is when they enter more negative states.

**Rise and Recovery Rates:** For narration, rise rates are significantly higher when compared to recovery rates, for all three dimensions, and for both low and high displacements (all metrics are reported in the Appendix Tables 7.4-7.6). For character dialogue, we do not see significant ($p < 0.05$) differences for valence (low and high displacements), and for high dominance displacements (i.e., rise and recovery rates are quite similar).

We note several outliers on either extreme for these metrics, which are potentially interesting to researchers in literary studies, as they pinpoint characters with particularly extreme and notable personality traits. For example, the character of Dr. Watson in the Sherlock Holmes novel *The Sign of the Four* has the highest rise rate for high arousal, indicating that he is easily excited (rises quickly to states of high activity/arousal). We report all meta-speakers and characters that emerge as outliers in Table 7.7 in Appendix 7.C.

### 7.4.2 Emotion Arcs Within a Novel (RQ2)

In the previous section, we compared the aggregate UED metrics for the different text streams in a novel: the entire text, narration, and character dialogue. Here, we instead compare the shapes of the trajectories by considering a measure of arc similarity. We temporally align a pair of arcs that we wish to compare, and compute the Spearman correlation between them (details of the process are in Appendix 7.D). A higher correlation score implies that the two arcs follow similar shapes.

We continue to further quantify the diversity of emotion arcs that can be found within a single story.

**Q1:** *How similar are the emotion arcs of the narration and dialogue in a novel (regardless of which character is uttering it)?*

This question gets at the heart of how close the emotions expressed in the narration are to those of its characters' utterances. In a way, this metric captures a facet of

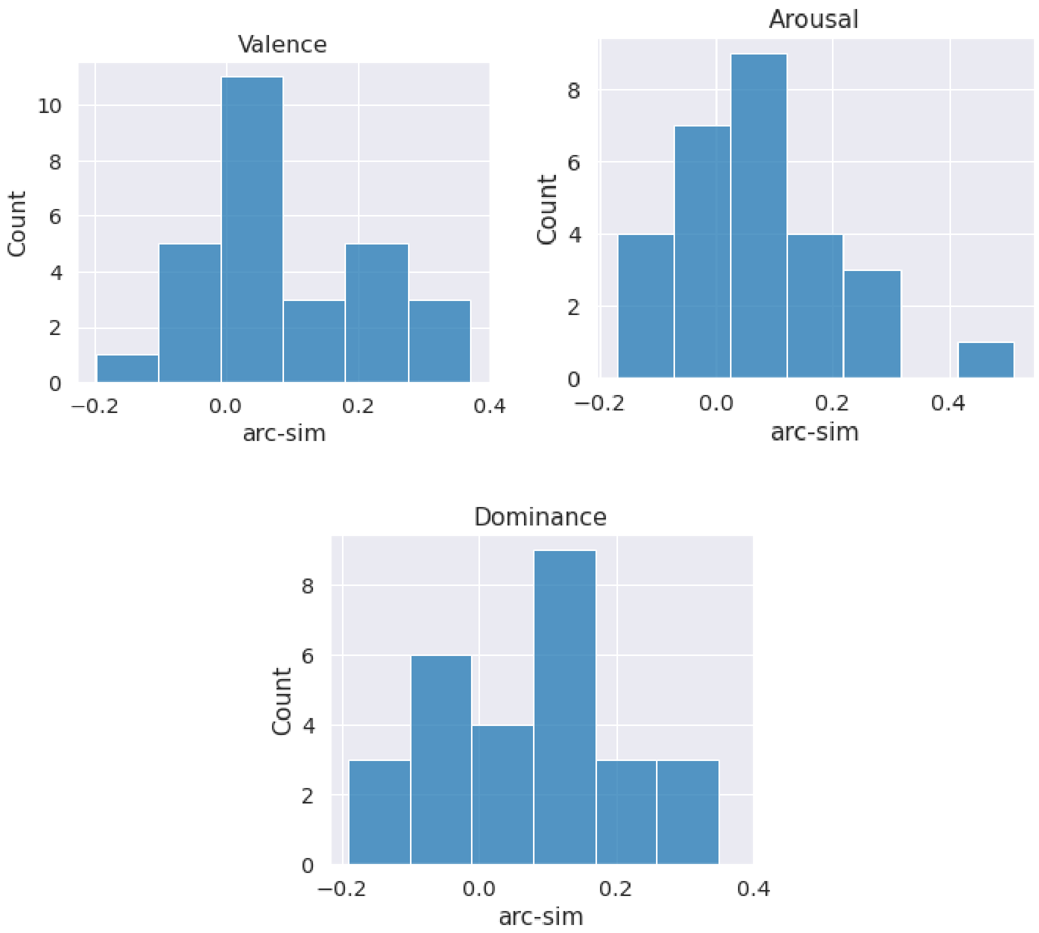


Figure 7.3: Distribution of arc correlations between narration and dialogue (irrespective of character) for all three VAD dimensions.

narrative style — is the narrator emotionally detached from what the characters are experiencing, or does the linguistic style tend to reflect their emotional journeys?

**Results:** Figure 7.3 plots the distribution of correlation scores between arcs of narration and dialogue for all 28 novels, for each of the three emotion dimensions. On average, we see near-zero correlations across all dimensions (0.06 for arousal and dominance, 0.09 for valence), with most correlation scores falling between −0.1 and 0.35 (mild positive correlation). We note the following outliers on either side of this range: for valence, *Alice's Adventures in Wonderland* has the lowest correlation (−0.2), and *The Sport of the Gods* has the highest correlation (0.37), followed by *The Age of Innocence* (0.33). Along the dominance dimension, the arcs in *The Awakening* have a mild positive correlation (0.35); for arousal, *Daisy Miller*, a first-person narrative, is a clear outlier with a correlation of 0.51. The latter outlier is more notable in terms of the absolute value of the correlation (moderate), indicating that the emotional tone of the narration is more in tune with the dialogue.

**Discussion:** The correlations between the emotional states of the narration and dialogue in novels are surprisingly low; in most cases, they are near-zero. This distinction has rarely been explored in prior work on the emotional arcs of narratives, which has largely treated the entire text of the novel as presenting a singular flow of emotions (Reagan et al., 2016; Kim et al., 2017b; Fudolig et al., 2022), and demonstrates that narration and character dialogue often represent distinct emotional arcs within a novel.

**Q2:** *How similar are the emotion arcs of the narration and each of the major characters in a novel?*

The above results tell us that the overall arcs of narration and dialogue tend to not be correlated. We now ask if, rather than being equally distant from or close to *all* the characters, the narration tends to attach itself the emotional arc of just its protagonist(s) or antagonist(s). For each novel, we measure the similarity between the arc of the narration and each of its *major* characters.

**Results:** Figure 7.4a plots the distribution of correlation scores for major characters from all 28 novels, for the valence dimension. Arcs of narration have little to no correlation with those of the major characters (average values of 0.03, 0.02, and 0.03 for valence, arousal, and dominance respectively).[3] Figure 7.8a in Appendix 7.E plots the distribution of correlations for each novel.

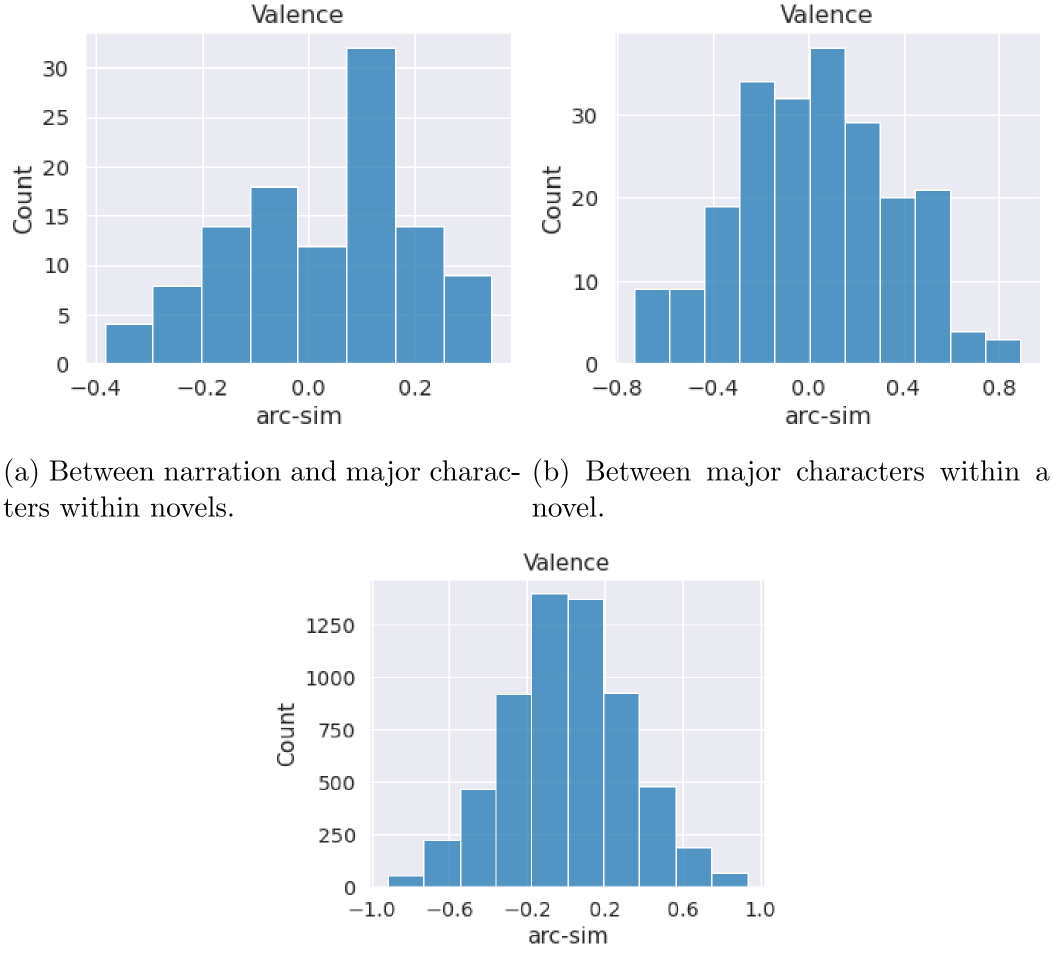
Interestingly, though 5 out of 28 novels in PDNC are written in the first-person[4], many of the correlations between the narration and the dialogue of the narrator fall in the mild ($-0.1$ to $0.2$) range for valence; for arousal, *Daisy Miller* (0.37) has one of the highest correlation scores with its narrator, whereas *The Sign of The Four* (-0.36) has among the lowest.

A close reading of these novels might provide critical literary insights into these results. With *The Sun Also Rises*, we have a traumatized and repressed narrator: what he says out loud is vastly different from what he thinks privately. For other novels, these results could be indicative of the difference between the emotions a character has in real time (dialogue) vs. those in retrospect (narration).

**Discussion:** These findings reinforce the distinctions between the narration arc and

[3] Since correlation is a bi-polar scale, we also compute average values for positive and negative correlations separately; none of them go beyond 0.18.

[4] *Daisy Miller*, *The Mysterious Affair at Styles*, *The Sun Also Rises*, *The Sign of the Four*, and *The Gambler*.



(a) Between narration and major characters within novels.

(b) Between major characters within a novel.

(c) Between major characters across all novels.

Figure 7.4: Distribution of valence arc correlations.

those of a novel's main characters — we average at near-zero correlations for all three emotion dimensions, highlighting that any computational modeling of the emotion arcs of stories should consider the distinctions between these facets of a narrative.

### 7.4.3 Diversity of Character Arcs (RQ3)

We now focus solely on the emotion arcs of character dialogue, and quantify their variation in literary novels. We look at this variation both within a single novel, and when compared across stories. These measures inform us of the diversity of emotional trajectories that a character can follow, and gets closer to the question of the "basic shapes of stories" (i.e, of character journeys) that many prior works in NLP have attempted to quantify.

**Q3:** *What is the average similarity of character arcs in a novel? Where do we see outliers?*
For each pair of major characters within a novel, we compute the similarity of their emotion arcs. Apart from looking at individual pairs of characters, we compute the mean and variance of the scores for all character pairs within a novel — which ones have the most diverse emotional journeys for their characters, and the least?
**Results:** We plot the distribution of all pairwise scores for valence in 7.4b. We can immediately see the much larger spread of the scores here (than for narration and dialogue), extending into the range of high correlation and anti-correlation (maximum of 0.89 and minimum of $-0.73$). The mean correlations, however, stay close to 0 (0.03, 0.02, 0.04 for the VAD dimensions).

We quantify the diversity of character arcs within a novel as the standard deviation of the emotion arc correlations of its major characters. *Winnie-the-Pooh* and *The Sport of the Gods* have some of the highest diversity in major character arcs for all three dimensions; *The Age of Innocence* has the lowest diversity in valence arcs, *The Gambler* and *The Invisible Man* for arousal. Recollect from the previous chapter that *Winnie-the-Pooh* also had the lowest average classification (and clustering) scores, indicating that its characters could not be distinguished based on *stylistic* utterance features. Both of these results seem to be a consequence of the genre of the novel — children's tales are often written in a simple style, and with clear-cut, extreme emotional journeys for the characters.

*The Sun Also Rises* leans the most positive with a median score of 0.44. In this novel, the characters Robert Cohn and Jake Barnes are often read as rivals (they are competing for the love of the same woman), but our analysis shows them travelling the same emotional journey (correlation of 0.89 for valence). Figure 7.8b in Appendix 7.E shows the distribution of correlations for valence by novel.
**Discussion:** The above results lend credence to the variety and diversity that can be found in novels at the level of characters — it is quite rare to find a story where the character trajectories or voices are uniform. While it might seem trivial to state that a story is not so much a singular narrative as a collection of intersecting narratives, computational analysis has largely suffered by not being able to afford this complexity in its study of stories. The higher correlations between character arcs as opposed to with the narration indicate that using a single novel-based arc is a poor representation of the novel for much literary analysis.

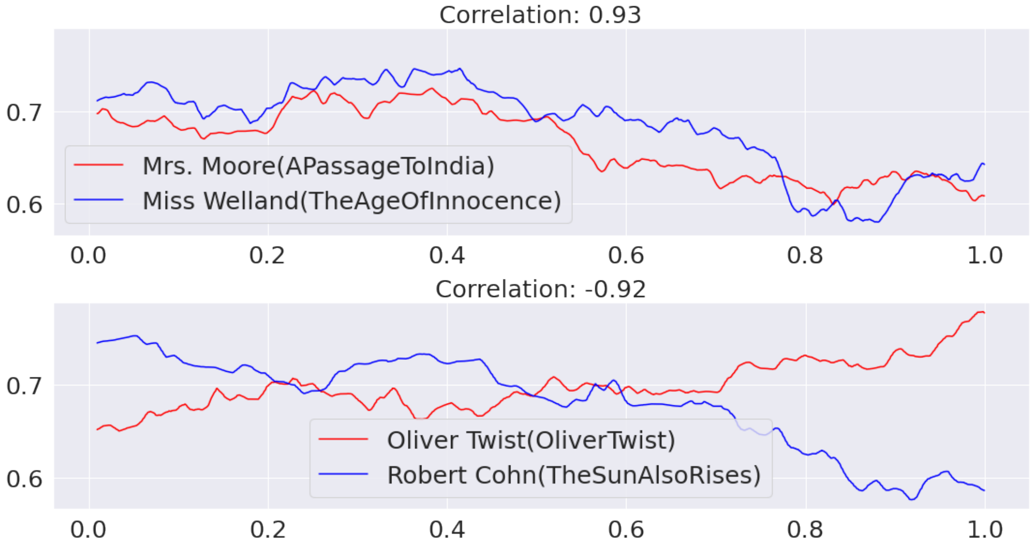


Figure 7.5: Emotion arcs of valence for character pairs with the highest and lowest correlation scores.

**Q4:** *How similar are the emotion arcs of characters across novels?*
This question explores whether we see high correlations reflecting the "shapes of stories" as described by Kurt Vonnegut, which largely describes the emotional journey of the protagonists of stories. We compute correlation scores between the arcs of all possible pairs of the 111 major characters from all 28 novels (6105 pairs).
**Results:** Figure 7.4c plots the distribution of correlation scores for valence (we see similar distributions for arousal and dominance). The average correlation between any pair of characters is close to 0, for all three dimensions. The range of correlation scores is much larger, from a minimum of $-0.92$ (Oliver Twist from *Oliver Twist* and Robert Cohn from *The Sun Also Rises*) to a maximum of 0.93 (Mrs. Moore from *A Passage To India* and Miss Welland from *The Age of Innocence*). In Figure 7.5, we plot the arcs of these two pairs of characters.
**Discussion:** The range of the correlations, extending from highly correlated to anti-correlated, demonstrate that character trajectories in different novels can indeed follow similar shapes. However, the Gaussian distribution of the correlation scores and their mean correlations of $\sim$0 indicate that these arcs do not all follow a small set of prototypical shapes, but cover a wider spectrum.

### 7.4.4 Character Groups (RQ4)

Several prior works on stories have demonstrated biases in the portrayals of characters that correspond to overgeneralized and often inaccurate stereotypes of certain

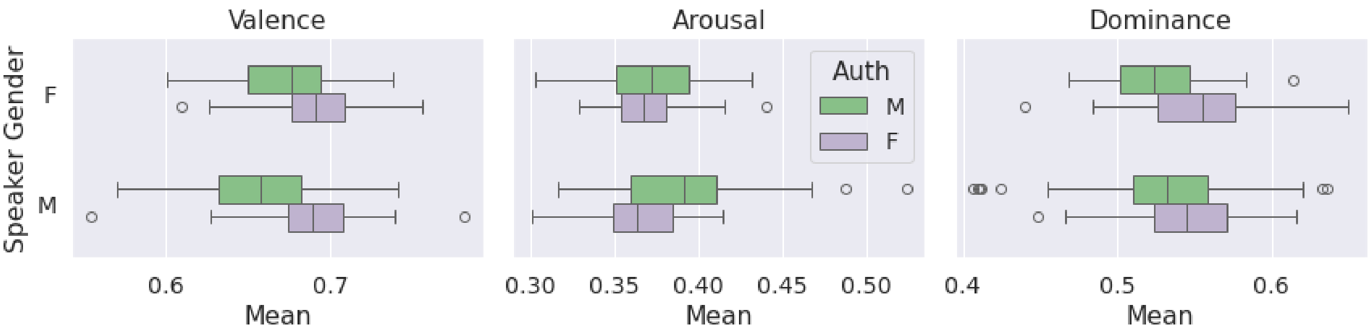


Figure 7.6: Distribution of the emotion mean for VAD, grouped by speaker (y-axis) and author (box color) gender.

demographic groups, particularly those that are marginalized (Fast et al., 2016; Sap et al., 2017). We investigate the presence of such biases for the characters in our dataset along the gender dimension, quantifying differences between the UED metrics of Male (M) and Female (F) character groups and how they are written by Male and Female authors. We note that the rather restricted range of novels and authors represented in PDNC [5] means we cannot make generalizable claims about these differences; however, our methodology is broadly applicable to any corpus of novels, and provides metrics that are useful to quantify biases in such corpora.

**Q5:** *Do the emotional dynamics of characters differ based on their (presented) gender?*

We have 89 Female and 121 Male characters in PDNC with sufficient utterances to compute UED metrics. We test for statistically significant differences in the aggregate UED metrics (mean, variability, average rise and recovery rates, etc.) for each of these groups using a two-sided independent $t$-test, and apply the Benjamini-Hochberg correction for multiple comparisons.

**Results:** We find the following significant ($p < 0.05$) effects: mean valence is higher for female characters compared to male characters (0.68 vs 0.66); mean arousal is higher for male characters (0.39 vs 0.37), and the average peak distance for arousal displacements is also higher for male characters compared to female characters (0.019 vs 0.016).

**Discussion:** This is in line with prior work on gender biases in storytelling — whether in novels, movies, or other forms of stories — that find that female characters tend to be portrayed with higher levels of positive emotions: warm, kind, caring, and more

[5] These are canonically well-regarded authors and novels, who might not be representative of the general fiction landscape of the era, and certainly not of other time-periods.

joyful (Ramakrishna et al., 2017; Xu et al., 2019) while male characters are expected to express more intense emotions relating to arousal — anger, rage, and violence.

**Q6:** *Do male authors write their characters differently from female authors?* We now compare the UED metrics of characters when grouped by author gender. There are 134 characters written by female authors (85 female, 49 male), and 183 characters written by male authors (133 male, 50 female). We test for significant differences in aggregate UED metrics for all possible groups of (M/F) characters written by (M/F) authors using the two-way ANOVA test.

**Results:** Figure 7.6 plots the distributions of mean VAD scores for all non-minor characters in our dataset, separated by speaker and author gender. Character dialogue written by female authors has a higher mean valence (0.69 vs 0.65), lower mean arousal (0.36 vs 0.38), and higher mean dominance (0.55 vs 0.52) when compared to that written by male authors in our dataset ($p < 0.001$). Additionally, we find that the peak distance for low arousal displacements is lower for female-authored characters (0.014 vs 0.019). Male-authored male characters have a particularly high mean arousal when compared to all other groups.

**Discussion:** These findings are again in line with prior work on gendered character dialogue in movie scripts and stories (Fast et al., 2016; Xu et al., 2019; Lettieri et al., 2023), which find that female writers tend to use words that are more positively valenced, and male writers, words that are marked as more arousing. We find significant trends in UED metrics that capture the intensity of emotional displacements, which have not previously been studied. The depiction of the two character gender groups for each of the author groups also reveals trends of interest — male characters have a higher average arousal in novels authored by male writers, whereas female writers tend to send their male characters into more intense low arousal states.

## 7.5 Conclusion

In this chapter, we took a closer look at the *variation* in character utterances along linguistic dimensions of affect and emotion. We first demonstrated that the emotions expressed in the narration of a story are not representative of those expressed by its characters. We then showed that the characters within a story can have widely varying emotion arcs, with correlations ranging from highly negative ($-0.8$) to strongly correlated (0.8), and averaging at near-zero correlations.

The importance of individual character arcs is also demonstrated by the high correlations we observe between the arcs of character pairs from *different* stories — contrasted, once again, with the more moderate scores observed for overall arcs of narration and dialogue.

Our analysis of the effects of author and speaker gender on the shapes of their stories echo the results from the previous chapter's experiments using mixed-effect models. While we don't find very many consistent effects of character gender alone, author gender plays an important role in how male and female characters are portrayed.

## 7.6 Discussion

The experiments in this chapter tell us that the "shapes of stories" are better represented by character journeys, and it is unclear what, if anything, a narrative arc is capturing of the stories that are told in a novel. Our results also highlight the diversity of narrative threads contained in a single novel: different characters go through wildly-varying journeys, and these cannot be summarized by a single arc — of the novel, narration, or dialogue — alone, or even with a small number of prototypical shapes. While this might seems something of a trivial statement in hindsight, we are only able to quantitatively show it with a carefully-annotated dataset of character utterances, a requirement that is not trivial to satisfy.

The works that we studied here are also set apart by an obvious *selection bias*, in that we chose novels that are quite popular and critically-acclaimed in the literary canon, written by authors who were societally positioned to be able to achieve fame and success. Will we find more evident, consistent patterns of bias conditioned on character and author gender if we expand our selection to a less-curated, and therefore more representative, dataset of texts? Perhaps, but a more pertinent question is to ask how close we are to actually being able to conduct such an analysis. Automatically extracting information about the characters in novels is computationally quite a challenging problem, as is the task of identifying and attributing the various lines of quotation within a novel to one of these characters. We expand on these challenges, and ways of overcoming them, in the subsequent chapter.

## 7.A Utterance Emotion Dynamics

The Utterance Emotion Dynamics framework derives several metrics characterizing the temporal patterns of regularity and change of emotion states derived from the textual utterances of an individual. Figure 7.1 shows a simple example emotion arc and the UED metrics corresponding to the home base and displacements below and above the home base. We briefly describe these metrics here:

- **Emotion mean (emo_mean)**: The mean of the sequence of emotion states.
- **Variability (emo_std)**: The standard deviation of the sequence of emotion states.
- **Average peak distance (emo_avg_peak_dist)**: Average of the peak emotional distance from the home base for all displacements (a measure of how emotional the speaker gets on average when they have a displacement).
- **Average displacement length (emo_avg_disp_length)**: Average of the length of a displacement, in terms of temporal steps, for all displacements (a measure of how long the speaker is outside the home base per displacement on average).
- **Average rise rate (emo_rise_rate)**: Average of the rise rates of all displacements (a measure of how quickly one reaches peak distance from home base (regardless of direction of displacement)).
- **Average recovery rate (emo_recovery_rate)**: Average of the recovery rates of all displacements (a measure of how quickly one reaches peak distance from home base (regardless of direction of displacement)).
- **Average Low peak distance (emo_low_peak_dist)**: Average of the peak emotional distance from the home base for all displacements below the home base.
- **Average Low displacement length (emo_low_disp_length)**: Average of the length of a displacement, in terms of temporal steps, for all displacements below the home base.
- **Average Home-to-Low rise rate (emo_low_rise_rate)**: Average of the rise rates of all displacements below the home base (measure of how quickly one descends to the lowest emotion state).
- **Average Low-to-Home recovery rate (emo_low_recovery_rate)**: Average of the recovery rates of all displacements below the home base (measure of how quickly one recovers from the lowest emotion state).
- **Average High peak distance (emo_high_peak_dist)**: Average of the peak emotional distance from the home base for all displacements above the home base.

| metric // speaker_type | novel | narration | character | major | intermediate | minor |
|---|---|---|---|---|---|---|
| emo_mean | 0.647 | 0.627 | 0.675 | 0.667 | 0.678 | 0.684 |
| emo_std | 0.053 | 0.057 | 0.029 | 0.038 | 0.026 | 0.018 |
| emo_avg_peak_dist | 0.023 | 0.025 | 0.016 | 0.019 | 0.015 | 0.012 |
| emo_avg_disp_length | 132.422 | 134.792 | 78.854 | 101.728 | 70.969 | 44.377 |
| emo_rise_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| emo_recovery_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| emo_low_peak_dist | 0.027 | 0.028 | 0.021 | 0.025 | 0.020 | 0.012 |
| emo_low_disp_length | 137.656 | 140.427 | 105.963 | 135.199 | 100.196 | 48.943 |
| emo_low_rise_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| emo_low_recovery_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| emo_high_peak_dist | 0.020 | 0.022 | 0.016 | 0.018 | 0.015 | 0.014 |
| emo_high_disp_length | 130.319 | 136.301 | 87.169 | 110.648 | 76.490 | 55.121 |
| emo_high_rise_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 |
| emo_high_recovery_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |

Table 7.4: Averaged UED metrics (rows) of **valence** for the different speaker types (columns) in PDNC.

- **Average High displacement length (emo_high_disp_length)**: Average of the length of a displacement, in terms of temporal steps, for all displacements above the home base.

- **Average Home-to-High rise rate (emo_high_rise_rate)**: Average of the rise rates of all displacements above the home base (measure of how quickly one ascends to the lowest emotion state).

- **Average High-to-Home recovery rate (emo_high_recovery_rate)**: Average of the recovery rates of all displacements above the home base (measure of how quickly one recovers from the highest emotion state).

## 7.B UED Metrics

We report the aggregated UED metrics for the different types of speakers – novel (meta-speaker), narration (meta-speaker), all characters, major characters, intermediate characters, and minor characters – for valence (Table 7.4), arousal (Table 7.5), and dominance (Table 7.6).

| metric // speaker_type | novel | narration | character | major | intermediate | minor |
|---|---:|---:|---:|---:|---:|---:|
| emo_mean | 0.381 | 0.384 | 0.377 | 0.381 | 0.377 | 0.366 |
| emo_std | 0.047 | 0.046 | 0.028 | 0.037 | 0.024 | 0.016 |
| emo_avg_peak_dist | 0.020 | 0.021 | 0.016 | 0.021 | 0.014 | 0.011 |
| emo_avg_disp_length | 120.911 | 121.039 | 78.358 | 106.449 | 66.519 | 39.966 |
| emo_rise_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| emo_recovery_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| emo_low_peak_dist | 0.018 | 0.018 | 0.017 | 0.020 | 0.015 | 0.012 |
| emo_low_disp_length | 126.509 | 120.441 | 89.166 | 118.320 | 75.781 | 49.623 |
| emo_low_rise_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| emo_low_recovery_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| emo_high_peak_dist | 0.023 | 0.025 | 0.019 | 0.024 | 0.016 | 0.011 |
| emo_high_disp_length | 118.200 | 125.468 | 85.401 | 117.221 | 71.227 | 40.459 |
| emo_high_rise_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| emo_high_recovery_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |

Table 7.5: Averaged UED metrics (rows) of **arousal** for the different speaker types (columns) in PDNC.

| metric // speaker_type | novel | narration | character | major | intermediate | minor |
|---|---:|---:|---:|---:|---:|---:|
| emo_mean | 0.519 | 0.508 | 0.538 | 0.532 | 0.545 | 0.536 |
| emo_std | 0.058 | 0.060 | 0.036 | 0.045 | 0.032 | 0.022 |
| emo_avg_peak_dist | 0.025 | 0.026 | 0.020 | 0.023 | 0.020 | 0.015 |
| emo_avg_disp_length | 121.263 | 124.919 | 70.747 | 91.681 | 63.142 | 38.689 |
| emo_rise_rate | 0.001 | 0.001 | 0.002 | 0.002 | 0.002 | 0.002 |
| emo_recovery_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 |
| emo_low_peak_dist | 0.026 | 0.025 | 0.023 | 0.027 | 0.022 | 0.014 |
| emo_low_disp_length | 124.514 | 126.597 | 86.892 | 116.789 | 74.472 | 39.135 |
| emo_low_rise_rate | 0.001 | 0.001 | 0.002 | 0.001 | 0.002 | 0.002 |
| emo_low_recovery_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| emo_high_peak_dist | 0.025 | 0.026 | 0.021 | 0.023 | 0.021 | 0.018 |
| emo_high_disp_length | 121.133 | 126.999 | 77.927 | 94.625 | 71.660 | 51.979 |
| emo_high_rise_rate | 0.001 | 0.001 | 0.002 | 0.002 | 0.002 | 0.002 |
| emo_high_recovery_rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 | 0.002 |

Table 7.6: Averaged UED metrics (rows) of **dominance** for the different speaker types (columns) in PDNC.

| Emotion | Metric | Type | Extreme | Speaker | Value |
|---|---|---|---|---|---|
| valence | emo_std | narration | high | narrator (ThePictureOfDorianGray) | 0.078 |
| valence | emo_mean | character | low | Monks (OliverTwist) | 0.580 |
| valence | emo_mean | character | low | Mr. Bumble (OliverTwist) | 0.599 |
| valence | emo_mean | character | low | The Invisible Man (TheInvisibleMan) | 0.578 |
| valence | emo_mean | character | low | Mike Campbell (TheSunAlsoRises) | 0.578 |
| arousal | emo_mean | narration | low | narrator (WinnieThePooh) | 0.334 |
| arousal | emo_mean | character | high | Professor De Worms (TheManWhoWasThursday) | 0.452 |
| arousal | emo_mean | character | high | Joe Hamilton (TheSportOfTheGods) | 0.488 |
| arousal | emo_std | character | low | Jock Grant-Menzies (AHandfulOfDust) | 0.008 |
| arousal | emo_std | character | low | Cassandra Otway (NightAndDay) | 0.013 |
| arousal | emo_std | character | high | Prince Charming (ThePictureOfDorianGray) | 0.061 |
| arousal | emo_std | character | high | Mike Campbell (TheSunAlsoRises) | 0.074 |
| arousal | emo_std | character | high | Piglet (WinnieThePooh) | 0.072 |
| dominance | emo_mean | narration | low | narrator (WinnieThePooh) | 0.403 |
| dominance | emo_mean | character | low | John Andrew (AHandfulOfDust) | 0.412 |
| dominance | emo_mean | character | low | Eeyore (WinnieThePooh) | 0.407 |
| dominance | emo_std | character | low | Joe Hamilton (TheSportOfTheGods) | 0.008 |
| dominance | emo_std | character | high | Mary Musgrove (Persuasion) | 0.075 |
| dominance | emo_std | character | high | Christopher Robin - Story (WinnieThePooh) | 0.080 |

Table 7.7: Outliers among characters and narration on either extreme for the mean and variability metrics, along all three dimensions.

## 7.C Outlier Characters

Speakers who emerge as outliers for the emotion mean and variability metrics, along all three VAD dimensions, are listed in Table 7.7. Speakers can include meta-speakers (like narration). Outliers are identified as points that fall outside the whiskers of a box-and-whisker plot, i.e, the scores that are below $Q_1 - 1.5 * IQR$ (low) and above $Q_3 + 1.5 * IQR$ (high), where $Q_1$ and $Q_3$ are the 25th and 75th percentiles (1st and 3rd quartiles) of the distribution, and $IQR (= Q_3 - Q_1)$ is termed the inter-quartile range. Outliers are independently identified for each speaker type (novel, narration, character).

## 7.D Aligning Emotion Arcs

Let's say we want to compare the computed emotion arcs of a set of speakers $S$:

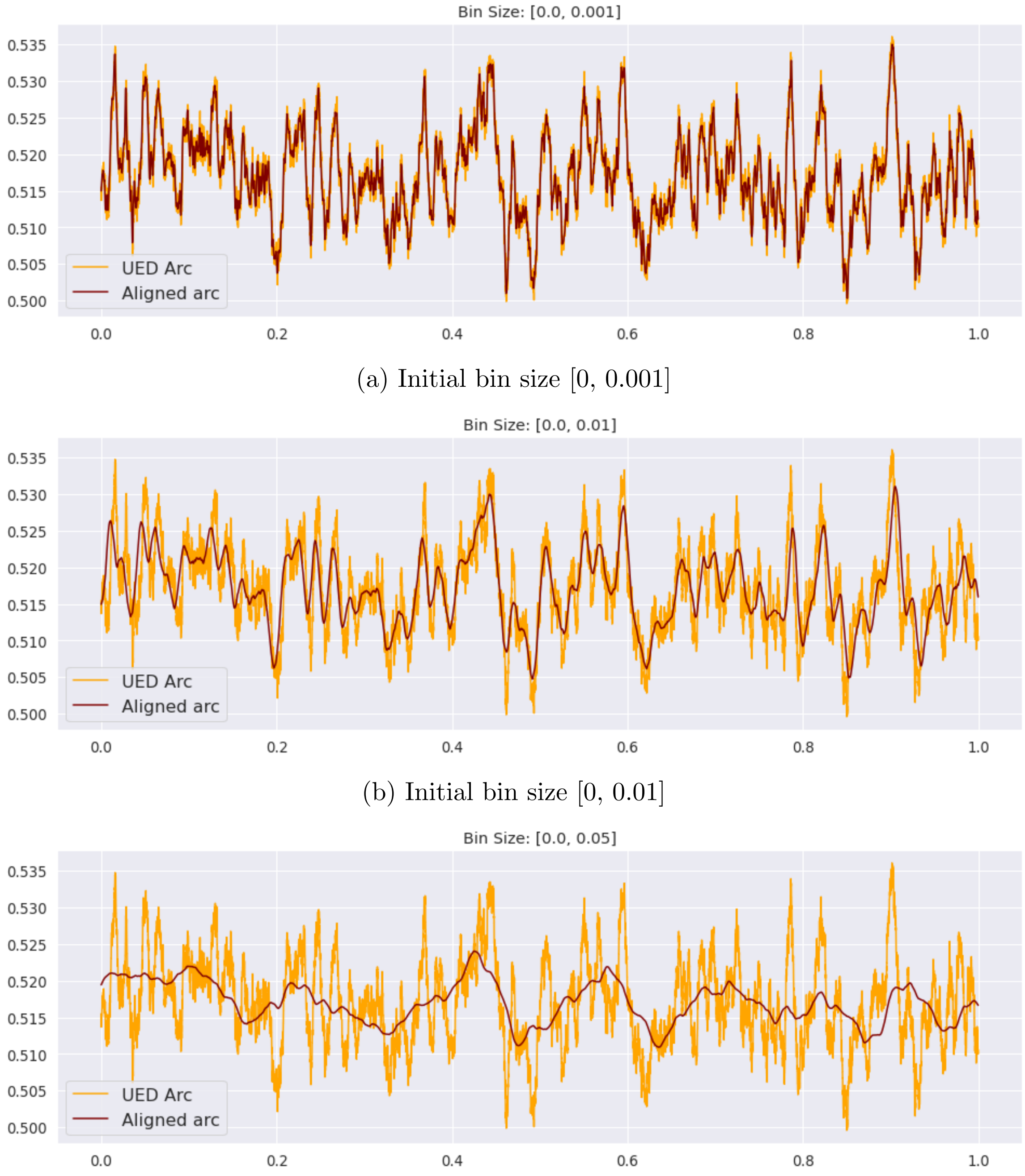
- Find the speaker $s_{sm} \in S$ with the smallest temporal length (i.e, the fewest utterances).
- Start with an initial timestep window of [0, 0.001] for this speaker (this initial window size is a hyperparameter we examine later).
- Move forward this window by one word for $s_{sm}$ – this corresponds to a new bin of timestep values for the speaker $s_{sm}$, say $[0.001, 0.011]$.
- Continue to move forward the window by one word until the end of the speaker's arc; this results in a set of $n$ time bins.
- Average the emotion state values at the timesteps contained within each the $n$ time bins to obtain $n$ comparable emotion values for each speaker in $S$.

Note that, while the window is moving forward by 1 word for $s_{sm}$, it could be moving forward by $k$ words for a different speaker $s_j$ who has a longer volume of utterances. However, it ensures that we have an equal number of bins for each speaker, at approximately the same relative time points (i.e, the 1% mark, the 1.1% mark, etc).

Once a pair of emotion arcs has been aligned, we obtain equal-sized sequences of emotion states for the normalized bins, for all the speaker arcs we wish to compare (using a similarity metric like the euclidean distance or spearman correlation).

**Qualitative Assessment:** How well do the aligned emotion arcs represent the original UED arc of a novel? We qualitatively examine the aligned arcs with three different initial window sizes. The smallest novel out of the set of 28 in PDNC is *Daisy Miller*. With an initial window size of $[0, 0.01]$, we get 22,603 temporal bins; for an initial window of $[0, 0.001]$, we have 22,809 bins; $[0, 0.05]$ yields 21,690 bins. We therefore compute a new set of aligned arcs for each novel by averaging the emotion state values within each of these temporal bins, for all three temporal bin sizes.

In figure 7.7, we plot the original UED arc and the aligned arc for a longer novel, *A Room With A View*, for each of these choices. While a bin size of $[0, 0.001]$ retains many of the sharp transitions of the original arc , and the $[0, 0.05]$ blurs over too many of them, $[0, 0.01]$ seems to be a good compromise in roughly capturing the ups and downs of the emotion arc. We do note, however, that the appropriate bin size is dictated by the research question of interest; if we wanted to obtain a high-level view of the "shape of a story", we might prefer to smooth the arc even further.

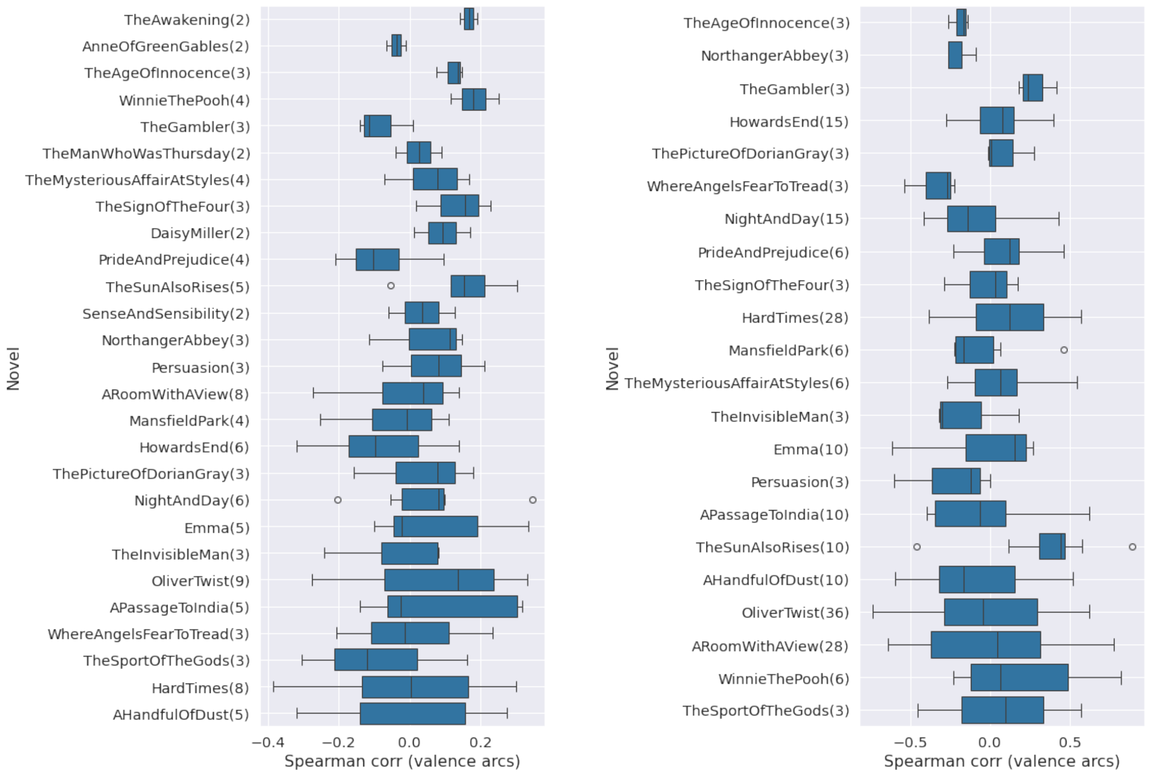


(a) Initial bin size [0, 0.001]

(b) Initial bin size [0, 0.01]

(c) Initial bin size [0, 0.05]

Figure 7.7: A visual comparison of the original UED arc with the time-aligned arc for *A Room With A View* (when aligned with the smallest novel in our dataset, *Daisy Miller*) with different initial bin sizes.



(a) Boxplots of the correlation scores of narration-major character valence arcs for each novel (ordered by variance).

(b) Boxplots of the major character valence arc correlations by novel (ordered by variance).

Figure 7.8: Within-novel correlations of **valence** arcs between narration and major character arcs. The numbers in parenthesis beside each novel indicate the number of pairwise correlations that are represented (based on the number of major characters) for each novel.

## 7.E Arc Correlations by Novel

In Figure 7.8a, we plot the distributions of the correlations between the arcs of each major character and the narration, for each novel. In Figure 7.8b, we report the distributions of the correlations between the arcs of pairs of major characters within each novel.

# Chapter 8

# Quotation Attribution in Literary Texts

The work in this chapter is derived from the following two publications:

- Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The Project Dialogism Novel Corpus: A Dataset for Quotation Attribution in Literary Texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848, Marseille, France. European Language Resources Association.

- Krishnapriya Vishnubhotla, Frank Rudzicz, Graeme Hirst, and Adam Hammond. 2023. Improving Automatic Quotation Attribution in Literary Novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 737–746, Toronto, Canada. Association for Computational Linguistics.

**Note:** The experiments in this chapter were conducted prior to the inclusion of 6 new novels in the PDNC corpus, and the presented results are therefore limited to only the first 22 novels of PDNC.

**Author Contributions:** All authors developed the idea for the project during regular discussion sessions. Krishnapriya expanded on the methodology, wrote the code and performed the experiments. Krishnapriya wrote the first draft of the paper, and all authors developed the discussion section and revised the final version of the paper.

## 8.1 Overview

In this chapter, I break from analyzing various aspects of character voice that we have so far been able to perform using the annotated data from PDNC, and focus on how such analyses can be extended to novels beyond its scope. Given the text of a novel, how can we assign dialogue to the characters that utter it, in order to then study it? This task, known as quotation attribution, is a neat but complex problem for NLP.

Current models for quotation attribution in literary novels assume varying levels of available information in their training and test data, which poses a challenge for in-the-wild inference. In this work, I present a modular formulation of the quotation attribution task as a set of four interconnected sub-tasks: character identification, coreference resolution, quotation identification, and speaker attribution. I benchmark state-of-the-art models on each of these sub-tasks independently using the Project Dialogism Novel Corpus, and use these insights and our dataset to propose a state-of-the-art quotation attribution model.

## 8.2 Introduction

The idiosyncrasies of literary text present several challenges to NLP models for named entity recognition, coreference resolution, character clustering, event detection, and speaker identification. The typical length of a text is several thousands of tokens, and the format and structure of the content vary widely depending on the genre, topic, time-period, and author of the text. Characters are referred to by various aliases, often incorporating notions of familial relations (*her father, Mr., Mrs., and Miss Bennet*) or social titles (*the baron*); mentions such as the former also can refer to different entities if used by different speakers (*my father*).

Consider, for example, the very first quotation in Jane Austen's *Pride and Prejudice*:

> *"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"*

Identifying the speaker of this quotation involves making several inferences: that the person being spoken to is Mr. Bennet, that the mention *his lady* refers to Mr. Bennet's lady, and that this is a proxy for Mr. Bennet's wife, who must be Mrs. Bennet. (The first explicit mention of Mrs. Bennet is only in Chapter 2; several other characters are introduced to us in the meantime).

Existing methods for quotation attribution fall into one of two groups: those that directly attribute the quotation to a named character entity, and those that treat it as a two-step process in which quotations are first attached to the nearest relevant *mention* of a character and mentions are then resolved to a canonical character name via a coreference resolution model. I contend that most use-cases of a quotation attribution system involve resolving the speaker mention to one among a list of character entities. Thus, the usability of these systems is very much dependent on their ability to compile such a list of character entities and to resolve each attributed mention to an entity from this list.

Here, I use the Project Dialogism Novel Corpus (introduced in Chapter 5), a large dataset of annotated coreferences and quotations in literary novels, to design and evaluate pipelines of quotation attribution. My analysis shows that state-of-the-art models are still quite poor at character identification and coreference resolution in this domain, thus hindering functional quotation attribution.

## 8.3 Prior Work

### 8.3.1 Speaker Attribution

Elson and McKeown (2010) proposed a classification approach for quotation attribution that classifies quotations into one of several types based on whether the speaker is explicitly indicated by an adjoining expression (explicit), appears without an attribution (implicit), is indicated by an anaphoric mention, is part of a dialogue chain, etc (see Table 5.1 for examples of each quotation type). A separate classifier is trained for each of these cases, taking as input a feature vector that encodes information relating to positions of mentions and quotations surrounding the target. Their model achieves an accuracy of 83% on their dataset, but uses gold labels as part of the pipeline.

O'Keefe et al. (2012) treat the task as a sequence decoding problem, where the set of speaker attributions in a document is treated as a text sequence to be predicted; i.e, the decision for the current quotation is made based on the previous $n$ attribution labels. While this method works well for news data, it fails to beat a rule-based baseline for literary texts. He et al. (2013) approach quotation attribution as a ranking problem between candidate speakers; their SVM-based ranking model selects a speaker based on a feature vector comprising contextual and topic information. The list of possible classes is taken from the set of annotated speakers.

Muzny et al. (2017b) describe a two-step process for quotation attribution, where quotations are first linked to mentions, and mentions to entities. Each step is composed of a set of deterministic sieves, designed to capture cases of increasing complexity. For example, the first sieve looks for explicit trigram patterns of Quote–Speech Verb–Mention, the next inspects dependency parses of sentences on either side of the target quotation for speech verbs with an `nsubj` relation that points to a character mention, and so on.

### 8.3.2 End-to-end Pipelines

The GutenTag package from Brooke et al. (2015) contains modules for generating character lists and identifying speakers in literary texts. It doesn't include a coreference resolution module, but instead relies on an NER system that clusters together named mentions using a bootstrapped classifier, resulting in a list of character entities and associated proper nouns (LitNER). For example, for the novel *Pride and Prejudice*, GutenTag successfully identifies *Liz, Eliza,* and *Lizzy* as aliases of the character Elizabeth Bennet; it also makes some mistakes, such as identifying *Kitty* as an alias for Lady Catherine de Bourgh, when in fact Kitty is a separate character (one of the Bennet sisters). The speaker attribution model is a simple rule-based approach that identifies the nearest named entity.

BookNLP[1] is a tool for natural language processing of literary texts (and other long documents) in English. The pipeline performs, among other things, dependency parsing, named entity recognition, coreference resolution, quotation attribution, and referential gender inference. The latest version of BookNLP is trained on LitBank's annotations of character entities (Bamman et al., 2020) and quotations (Sims and Bamman, 2020). While the exact model for quotation attribution is not described in a publication, we infer from the code that it uses a BERT-based model that takes as input the quotation text and its surrounding context, and links each quotation to a character mention. Mention-to-entity resolution is performed by a separate pipeline step that precedes quotation attribution. BookNLP combines the outputs of its named entity tagger and coreference resolution modules to generate a list of typed entities, each associated with a coreference ID; entities with the same coreference ID are assumed to be aliases for the same character entity. Here, the NER module is first used to identify a set of named entities, the coreference resolution is then applied to the text to obtain mention clusters, and named entities that co-occur in a mention cluster are taken to be aliases.

[1] https://github.com/booknlp/booknlp

## 8.4 Modularization of the Task

**Character identification:** The goal of this sub-task is to build a list of the unique character entities in a novel. Although NER models perform quite well at identifying spans of text that constitute a named entity (here, a character name), the task is complicated by the fact that characters can have multiple aliases in the text. Moreover, some characters may be introduced and referred to only by social titles (*the policeman, the Grand Inquisitor, the little old man, the bystander*).

**Coreference resolution:** The goals here are to identify text spans that refer to a character entity (which we refer to as *mentions*) and to link each mention to the correct character entity or entities to which it refers. In addition to mentions that are personal pronouns such as *he, she,* and *them*, literary texts have an abundance of pronominal phrases that reflect relationships between characters, such as *her husband* and *their father*. Such phrases can also occur within quotations uttered by a character (e.g., *my father*), requiring quotation attribution as a prerequisite for complete coreference resolution.

**Quotation identification:** Perhaps the most straightforward of our sub-tasks, here we identify all text spans in a novel that constitute dialogue, i.e., are uttered by a character entity or entities.

**Speaker attribution:** Finally, this sub-task links each identified quotation to a named character identity. While most models are designed to solve the more tractable and practical problem of linking quotations to the nearest relevant speaker mention, we subsume the mention–entity linking tasks under the coreference resolution module, equating the two tasks.

## 8.5 Models and Evaluation Metrics

We evaluate each of the modules of section 8.4 separately. In order not to confound the evaluation with cascading errors, at each step, we "correct" the outputs of the automated system from the previous step by using annotations from PDNC.

### 8.5.1 Character Identification

We evaluate the two end-to-end pipelines — GutenTag and BookNLP — on their ability to identify the set of characters in a novel, and potentially, the set of aliases for each character. In addition, we also test the NER system from the spaCy[2] module as a proxy for the state-of-the-art in NER that is not trained explicitly for the literary domain.

**Character recognition (CR):** For each novel, we compute the proportion of annotated character entities that are identified as named entities of the category 'PERSON' (Doddington et al., 2004). We use a simple string-matching approach, where we try for either a direct match, or a unique match when common prefixes such as *Mr.* and *Sir* are removed. Thus, if a particular novel has $N$ character entities annotated, the NER model outputs a list of $K$ named 'PERSON' entities, and $K'$ of these entities are in turn matched with $M$ out of the $N$ characters, the CR metric is calculated as $M/N$.

**Character clustering:** We use the clustering evaluation metrics of *homogeneity* (C.Hom), *completeness* (C.Comp), and their harmonic mean, *v-score* to evaluate named entity clusters.

These metrics evaluates how well the named entities recognized above are clustered together into character aliases. We use the annotated list of character aliases in PDNC to check the proportion of times two named entities are correctly designated as aliases of one another. Considering only the subset $K'$ of named entities identified above, we group together the names that co-occur in one or more mention clusters. We then compute the following: *homogeneity* looks at the proportion of named clusters that link to the same PDNC entity; *completeness* looks at the number of homogeneous clusters a single entity is distributed over; and their harmonic mean, the *v-measure* (note: these are slightly adapted from the standard clustering evaluation metrics of the same names, which work well for our purposes).

As an example, consider the case where we have three annotated characters for a novel: *Elizabeth Bennet*, *Mary Bennet*, and *The Queen*. The set of annotated aliases for the characters are {*Elizabeth Bennet, Eliza, Lizzie, Liz*}, {*Mary Bennet, Mary*}, and {*The Queen*}. Say model $M_1$ outputs the following entity clusters: {*Elizabeth Bennet, Eliza*}, {*Liz, Lizzie*} and {*Mary Bennet, Mary*}; model $M_2$ outputs {*Elizabeth Bennet, Mary Bennet, Eliza, Mary*}, {*Liz, Lizzie*}. Each model has recognized two out of the three characters in our list; this evaluates to a CR score of 2/3. Each of the three clusters from model $M_1$ refers solely to one character entity, resulting in a *homogeneity* score of 1.0. However, these three clusters are formed for only two unique character entities, resulting in a *completeness* score of 1.5 (*v-score* 0.6). Model $M_2$ has a homogeneity score of 0.5 and a completeness score of 1.0 (*v-score* 0.5).

[2] https://explosion.ai/blog/spacy-v3

### 8.5.2 Coreference Resolution

We consider two pipelines for coreference resolution: BookNLP (based on Ju et al. (2018)) and spaCy (based on Dobrovolskii (2021)). Given a text, these neural coreference resolution models output a set of clusters, each comprising a set of coreferent mention spans from the input.

Evaluating this module requires annotations that link each mention span in a novel to the character entity referred to. PDNC, unfortunately, contains these mention annotations only for text spans *within* quotations. We therefore evaluate coreference resolution only on a subset of the mention spans in a novel, extracted as follows: we first identify the set of mention clusters from our models that can be resolved to an annotated character entity, using the character lists from PDNC and the string-matching approach described above. We then prune this to only include those mention spans that are annotated in the PDNC dataset, i.e, mention spans that occur within quotations, and evaluate the accuracy of the resolution.

**Mention clustering (M-Clus):** We compute the fraction of mention clusters that can be matched to a *unique* (Uniq) annotated character entity rather than to multiple (Mult) or no (None) entities.

**Mention resolution (M-Res):** For those mention spans within PDNC that are identified by the model and are assigned to a cluster that can be uniquely matched to a character entity (# Eval), we compute the accuracy of the linking (Acc.).

### 8.5.3 Quotation Identification

Most models, rule-based or neural, can identify quotation marks and thus quotations. We evaluate how many of such quoted text instances actually constitute *dialogue*, in that they are uttered by one or more characters. Our gold standard is the set of quotations that have been annotated in PDNC, which includes quotations uttered by multiple characters and by unnamed characters such as *"a crowd"*.

### 8.5.4 Speaker Attribution

The speaker-attribution part of BookNLP's pipeline is a BERT-based model that uses contextual and positional information to score the BERT embedding for the quotation span against the embeddings of mention spans that occur within a 50-word context window around the quotation; the highest-scoring mention is selected as the speaker. We supplement this approach by limiting the set of candidates to resolved mention spans from the coreference resolution step, thereby directly performing quotation-to-entity linking. As we see from our results, this method, which we refer to as BookNLP+, greatly improves the performance of the speaker attribution model by eliminating spurious candidate spans.

We also evaluate a *sequential prediction model* that predicts the speaker of a quotation simply by looking at the sequence of speakers and mentions that occur in some window around the quotation. We implement this as a one-layer RNN that is fed a sequence of tokens representing the five characters mentioned most recently prior to the quotation text, one character mention that occurs right after, and, optionally, the set of characters mentioned within the quotation.

## 8.6 Improving the BookNLP Pipeline

We curate the set of mention candidates for each novel in the following manner: the mention clusters generated by BookNLP are used to extract the set of mention spans that could be successfully resolved to a character entity from the annotated PDNC character lists for each novel. We append to this set the annotated mention spans (within quotations) from PDNC, as well as explicit mention spans — that is, text spans that directly match a named alias from the character list.

Explicit matching is done with a longest-match-first approach, whereby the text span *Miss Elizabeth Bennet* would be matched before the containing spans *Elizabeth* or *Bennet*, thereby preventing overlaps as well as incorrect attributions in cases where *Bennet* is potentially annotated as an alias for a different character entity, such as *Mr. Bennet*. Overlaps between the three sets are resolved with a priority ranking, whereby PDNC annotations are considered to be more accurate than explicit name matches, which in turn take precedence over the automated coreference resolution model.

## 8.7 Experiments

The BookNLP pipeline is available to use as a Python package, as is spaCy, with pretrained models for coreference resolution and speaker attribution. For the former, these models are trained on the LitBank corpus, which is a dataset from the literary domain. We use these pretrained models to evaluate performance on the character identification and coreference resolution tasks. GutenTag can be run either via a Web interface or a command-line executable (requiring Python 2). It was designed to

| Model | CR | C.Hom | C.Comp | v-score |
|---|---|---|---|---|
| spaCy | 0.81 | 0.16 | 1.02 | 0.27 |
| GutenTag | 0.60 | 0.98 | 1.33 | 1.12 |
| BookNLP | 0.85 | 0.86 | 1.18 | 0.99 |

Table 8.1: Character identification: Average scores across all the novels in the dataset. Column headings are defined in the text. Scores for each individual novel are reported in Appendix 8.A.

interface with texts from the Project Gutenberg corpus. Some of the novels in PDNC were not found in GutenTag's predefined database of texts, so we exclude these when reporting average performance metrics.

We now describe the training setups for the two models of quotation attribution that require training: the BookNLP model and the sequential RNN model.

**Data Splits:** We experiment with two ways of dividing our dataset into training and test splits. In the **random** split, 80% of the annotated quotations from each novel are used for training and validation, and 20% as the test set. In the **leave-K-out** split, we keep 20% of the novels as the test set, with training data being obtained only from the remaining 80% of the novels, mimicking real-world applications. The metrics for each scenario are calculated in $k$-fold setup, with $k = 5$, and average performance across folds is reported.

## 8.8 Results and Discussion

From Table 8.1, we see that the neural NER models of spaCy and BookNLP are better at recognizing character names than GutenTag's heuristic system (0.81 and 0.85 vs 0.60). However, the strengths of GutenTag's simpler Brown-clustering–based NER system are evident when looking at the homogeneity; when two named entities are assigned as aliases of each other, it is almost always correct. This shows the advantage of document-level named entity clustering as opposed to local span-level mention clustering for character entity recognition. The cluster quality metric, on the other hand, tells us that GutenTag still tends to be conservative with its clustering compared to BookNLP, which nonetheless is a good strategy for the literary domain, where characters often share surnames.

Performance of these models on the coreference resolution task is significantly lower (Table 8.2). A majority of the mention clusters from both BookNLP and spaCy's coreference resolution modules end up as unresolved clusters, with no containing

| | M-Clus | | | | M-Res | |
|---|---|---|---|---|---|---|
| **Model** | **# Clus** | **Uniq** | **Mult** | **None** | **# Eval** | **Acc.** |
| spaCy | 1503.1 | 0.093 | 0.061 | 0.846 | 499.0 | 0.746 |
| BookNLP | 1662.8 | 0.043 | 0.003 | 0.953 | 1126.6 | 0.774 |

Table 8.2: Coreference resolution: All scores are averaged over the 22 novels in PDNC. Column headings are defined in the text.

| **Model** | **Quotations** | **Novels** |
|---|---|---|
| BookNLP-OG | 0.40 | 0.40 |
| BookNLP+ (LitBank) | 0.62 | 0.61 |
| Seq-RNN | 0.72 | 0.64 |
| BookNLP+ (PDNC) | 0.78 | 0.68 |

Table 8.3: Accuracy on speaker attribution for the end-to-end BookNLP model (BookNLP-OG), the restricted model with only resolved mention spans as candidates (row 2), the sequential prediction model, and the restricted model trained on PDNC, for the Quotations and the entire Novels cross-validation split.

named identifier that could be linked to a PDNC character entity. However, when we evaluate mention-to-entity linking on the subset of clusters that *can* be resolved, both systems achieve accuracy scores of close to 0.78, although spaCy is able to resolve far fewer mentions (499 vs 1127).

The importance of the character identification and coreference resolution tasks can be quantified by looking the performance of the speaker attribution models (Table 8.3). The end-to-end pretrained BookNLP pipeline, when evaluated on the set of PDNC quotations (which were identified with accuracy of 0.94), achieves an accuracy of 0.42. When we restrict the set of candidate mentions for each quotation to only those spans that can be resolved to a unique character entity, the attribution accuracy increases to 0.61. However, the RNN model still beats this performance with an accuracy of 0.72 on the random data split. When BookNLP's contextual model is trained on data from PDNC, its accuracy improves to 0.78. These scores drop to 0.63 and 0.68 for the entire-novel split, where we have the disadvantage of being restricted only to patterns of mention sequences, and not speakers.

## 8.9 Analysis

We briefly go over some qualitative analyses of the errors made by models in the different sub-tasks, which serves to highlight the challenges presented by literary text and opportunities for future research.

| | **Quotations** | | **Novels** | |
|---|---|---|---|---|
| **Model** | **Exp.** | **Rest** | **Exp.** | **Rest** |
| BookNLP-OG | 0.64 | 0.28 | 0.63 | 0.28 |
| BookNLP+ (LitBank) | 0.93 | 0.47 | 0.95 | 0.43 |
| Seq-RNN | 0.85 | 0.65 | 0.76 | 0.57 |
| BookNLP+ (PDNC) | 0.98 | 0.70 | 0.97 | 0.53 |

Table 8.4: Attribution accuracy for the speaker attribution models, broken down by quotation type, for the Quotations and Novels cross-validation splits. Column Exp. refers to explicit quotations, and column Rest refers to implicit and anaphoric quotations.

**Character Identification and Coreference Resolution:** We manually examine the mention clusters identified by our coreference resolution modules that could not be matched a unique character entity as annotated in PDNC.

We find that, by far, the most common error is conflating characters with the same surname or family name within a novel. For example, several of the women characters in these novels are often referred to by the names of their husbands or fathers, prefixed with a honorific such as *Mrs.* or *Miss*. Thus *Mrs. Archer* refers to *May Welland* in *The Age of Innocence* and *Miss Woodhouse* refers to *Emma Woodhouse* in *Emma*. However, a surname without a title, such as *Archer* or *Woodhouse*, generally refers to the corresponding male character. This results in the formation of mention clusters that take the spans *Miss Woodhouse* and *Woodhouse* to be coreferent, despite being different character entities. We see similar issues with father–son character pairs, such as *George Emerson* and *Mr. Emerson* in *A Room With A View*, and with character pairs that are siblings.

**Speaker Attribution:** We first quantify the proportion of quotations attributed to a mention cluster that cannot be resolved to a named character entity with the end-to-end application of the BookNLP pipeline.

On average, 47.7% of identified quotations are assigned to an unresolved mention cluster as the speaker. The range of this value varies from as low as 12.5% (*The Invisible Man*) to as high as 78.7% (*Northanger Abbey*). A majority of these unresolved attributions occur with implicit and anaphoric quotations (76.2%), where the speaker is not explicitly indicated by a referring expression such as *Elizabeth said*, as opposed to explicit quotations (23.8%).

In Table 8.4, we break down the performance of the speaker attribution models by quotation type. We see that even our local context–based RNN model is able to

| | | **BookNLP** | | | | | **GutenTag** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Novel** | **# Chars** | **CR** | **# Clus** | **C.Hom** | **C.Comp** | **v-score** | **CR** | **# Clus** | **C.Hom** | **C.Comp** | **v-score** |
| *A Room With A View* | 63 | 0.83 | 60 | 0.95 | 1.19 | 1.06 | 0.48 | 35 | 1.00 | 1.17 | 1.08 |
| *The Age of Innocence* | 55 | 0.84 | 48 | 0.81 | 1.26 | 0.99 | 0.64 | 49 | 1.00 | 1.40 | 1.17 |
| *Alice's Adventures in Wonderland* | 51 | 0.67 | 34 | 0.97 | 1.03 | 1.00 | 0.25 | 14 | 1.00 | 1.08 | 1.04 |
| *Anne of Green Gables* | 113 | 0.87 | 102 | 0.92 | 1.08 | 0.99 | 0.19 | 25 | 1.00 | 1.14 | 1.06 |
| *Daisy Miller* | 10 | 1.00 | 13 | 1.00 | 1.30 | 1.13 | 0.80 | 12 | 1.00 | 1.50 | 1.20 |
| *Emma* | 18 | 0.89 | 17 | 0.71 | 1.09 | 0.86 | 0.89 | 27 | 1.00 | 1.69 | 1.26 |
| *A Handful of Dust* | 104 | 0.82 | 94 | 0.89 | 1.15 | 1.01 | – | – | – | – | – |
| *Howards End* | 55 | 0.95 | 64 | 0.89 | 1.27 | 1.05 | 0.49 | 33 | 0.97 | 1.23 | 1.08 |
| *Night and Day* | 50 | 0.94 | 53 | 0.77 | 1.17 | 0.93 | 0.62 | 40 | 0.97 | 1.30 | 1.11 |
| *Northanger Abbey* | 20 | 0.90 | 12 | 0.75 | 1.00 | 0.86 | 0.85 | 23 | 0.96 | 1.29 | 1.10 |
| *Persuasion* | 35 | 0.86 | 29 | 0.79 | 1.28 | 0.98 | 0.77 | 28 | 0.96 | 1.08 | 1.02 |
| *Pride and Prejudice* | 74 | 0.81 | 62 | 0.85 | 1.10 | 0.96 | 0.35 | 30 | 0.90 | 1.35 | 1.08 |
| *Sense and Sensibility* | 24 | 0.83 | 25 | 0.56 | 1.17 | 0.76 | 0.79 | 26 | 0.96 | 1.39 | 1.14 |
| *The Sign of the Four* | 35 | 0.94 | 32 | 0.72 | 1.05 | 0.85 | 0.60 | 28 | 1.00 | 1.33 | 1.14 |
| *The Awakening* | 22 | 0.82 | 17 | 0.88 | 1.07 | 0.97 | 0.77 | 21 | 0.95 | 1.25 | 1.08 |
| *The Gambler* | 27 | 0.70 | 22 | 0.91 | 1.18 | 1.03 | 0.59 | 22 | 1.00 | 1.38 | 1.16 |
| *The Invisible Man* | 31 | 0.94 | 40 | 0.95 | 1.36 | 1.12 | 0.61 | 32 | 1.00 | 1.68 | 1.25 |
| *The Man Who Was Thursday* | 30 | 0.80 | 35 | 0.97 | 1.55 | 1.19 | 0.53 | 23 | 1.00 | 1.44 | 1.18 |
| *The Mysterious Affair at Styles* | 30 | 0.80 | 25 | 0.88 | 1.05 | 0.96 | 0.70 | 28 | 0.96 | 1.35 | 1.12 |
| *The Picture of Dorian Gray* | 43 | 0.88 | 43 | 0.98 | 1.14 | 1.05 | 0.56 | 27 | 1.00 | 1.12 | 1.06 |
| *The Sport of the Gods* | 37 | 0.81 | 34 | 0.94 | 1.23 | 1.07 | 0.54 | 28 | 0.96 | 1.50 | 1.17 |
| *The Sun Also Rises* | 51 | 0.86 | 51 | 0.96 | 1.23 | 1.08 | – | – | – | – | – |
| **Mean** | **44.5** | **0.85** | **41.45** | **0.86** | **1.18** | **0.99** | **0.60** | **27.55** | **0.98** | **1.33** | **1.12** |

Table 8.5: Results of character identification for each novel with BookNLP and GutenTag. '# Chars' is the number of characters in the novel. Other headers are the same as in Table 8.1.

identify the speaker of explicit quotations with a relatively high accuracy, and that the speaker for non-explicit quotations can also generally be modeled using the sequence of 5–6 characters mentioned in the vicinity of the quotation. The transformer-based models are of course able to use this local context more effectively by making use of linguistic cues and non-linear patterns of mentions and speakers in the surrounding text. Still, our best performing model achieves an accuracy of only 0.53 on implicit and anaphoric quotations when applied to novels unseen in the training set (the Novels split).

## 8.A Results by Novel

Tables 8.5 and 8.6 show for each novel in PDNC the per-model results for character identification that are summarized in Table 8.1.

| Novel | # Chars | CR | # Clus | C.Hom | C.Comp | v-score |
|---|---:|---:|---:|---:|---:|---:|
| *A Room With A View* | 63 | 0.78 | 64 | 0.33 | 1.24 | 0.52 |
| *The Age of Innocence* | 55 | 0.85 | 90 | 0.04 | 1.00 | 0.09 |
| *Alice's Adventures in Wonderland* | 51 | 0.80 | 44 | 0.39 | 1.00 | 0.56 |
| *Anne of Green Gables* | 113 | 0.69 | 98 | 0.24 | 1.04 | 0.40 |
| *Daisy Miller* | 10 | 0.90 | 3 | 0.00 | 0.00 | 0.00 |
| *Emma* | 18 | 0.89 | 14 | 0.07 | 1.00 | 0.13 |
| *A Handful of Dust* | 104 | 0.71 | 85 | 0.26 | 1.00 | 0.41 |
| *Howards End* | 55 | 0.84 | 72 | 0.18 | 1.08 | 0.31 |
| *Night and Day* | 50 | 0.88 | 52 | 0.15 | 1.00 | 0.27 |
| *Northanger Abbey* | 20 | 0.90 | 15 | 0.07 | 1.00 | 0.12 |
| *Persuasion* | 35 | 0.89 | 36 | 0.06 | 1.00 | 0.11 |
| *Pride and Prejudice* | 74 | 0.68 | 78 | 0.17 | 1.00 | 0.29 |
| *Sense and Sensibility* | 24 | 0.83 | 21 | 0.10 | 1.00 | 0.17 |
| *The Sign of the Four* | 35 | 0.80 | 40 | 0.05 | 1.00 | 0.10 |
| *The Awakening* | 22 | 0.86 | 24 | 0.12 | 1.00 | 0.22 |
| *The Gambler* | 27 | 0.74 | 18 | 0.22 | 1.00 | 0.36 |
| *The Invisible Man* | 31 | 0.84 | 37 | 0.22 | 1.00 | 0.36 |
| *The Man Who Was Thursday* | 30 | 0.73 | 26 | 0.19 | 1.00 | 0.32 |
| *The Mysterious Affair at Styles* | 30 | 0.87 | 29 | 0.10 | 1.00 | 0.19 |
| *The Picture of Dorian Gray* | 43 | 0.86 | 32 | 0.19 | 1.00 | 0.32 |
| *The Sport of the Gods* | 37 | 0.81 | 43 | 0.12 | 1.00 | 0.21 |
| *The Sun Also Rises* | 51 | 0.82 | 56 | 0.32 | 1.12 | 0.50 |
| **Mean** | **44.5** | **0.81** | **44.40** | **0.16** | **1.02** | **0.27** |

Table 8.6: Results of character identification for each novel with spaCy. '# Chars' is the number of characters in the novel. Other headers are the same as in Table 8.1.

# Part III

# Emotional Variation on Twitter

# Introduction

This part of the thesis switches the domain of interest from literary texts and character voices to the very real-world domain of people's utterances on the Twitter social media platform[3]. We are particularly interested here in emotional variation, and the emotion dynamics of a speaker's utterances as a characteristic indicator of their voice.

Emotions and their dynamics have been extensively studied in psychology. The dynamic sequence of person's emotional states (i.e, their emotion arc), i.e, how they experience and process the emotions triggered by various events in their lifetime, is viewed as fundamental to understanding the psychological processes of the human mind. Developing competent emotional functioning is a critical process of human growth for children and adolescents, and emotion dynamics therefore function as key indicators of mental, physical, and social health of individuals.

In Chapter 7, we introduced the Utterance Emotion Dynamics framework, which operationalizes a speaker's emotion arc as the sequence of emotion scores represented by their temporally-ordered utterances. UED metrics such as density (or mean), variability, and rise and recovery rates, correspond to emotion dynamics metrics from psychology on emotion intensity and variability, and emotion regulation.

In this work, we study the utterance emotional dynamics of people, derived from their social media posts, and their variation across two axes: geographic and temporal. In Chapter 9, I introduce our dataset TUSC (Tweets from US and Canada). It comprises millions of tweets from 46 different cities in the US and Canada collected over a two-year period from January 2020 to February 2022, as well as a subset with only country-level geotags over a longer, 7-year time-period (2015—2021), pre-processed for computational analysis. We analyze various metrics of Utterance Emotion Dynamics on this dataset, adapted here as Tweet Emotion Dynamics (TED), and their variation across cities and across months and years, uncovering interesting indications of the effects of real-world events including one-offs like the COVID-19 pandemic, and

[3] Since this work was completed, Twitter has been renamed to X, and tweets are now called posts. We will stick here with the nomenclature that was in use at the time of publication.

yearly events like the holiday season.

In the concluding section for this part, I expand on some of the follow-up work on developing computational measures of psychology-inspired metrics of emotional experience and expression. This work was carried out largely in collaboration with Dr. Saif Mohammad of the National Research Council, Canada, and the Carolina Affective Science Lab of the University of North Carolina, Chapel Hill. Our principal interest is in identifying potential correlations between various metrics of emotion extracted from textual utterances, and county-level health metrics obtained from US Census data. My main contribution to this work is in quantifying the concept of emotional granularity (the specificity of usage of different emotions) with metrics that can be extracted from textual data.

The outcomes of these projects demonstrate, quantitatively, the amount of variation in emotional expression of individual people, the systemicity of this variation across geographical locations when aggregated at the city and county level in the US and Canada, and their variation across time. The grounding of these metrics in socio-psychological theories of emotional experiences and variation, and their connection with the mental and physical health of people, makes it a compelling area of inter-disciplinary research for NLP, and also one that has to be approached with a collaborative and considerate attitude, led by subject-matter experts.

# Chapter 9

# Tweet Emotion Dynamics

This chapter was published as:
Krishnapriya Vishnubhotla and Saif M. Mohammad. 2022. Tweet Emotion Dynamics: Emotion Word Usage in Tweets from US and Canada. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4162–4176, Marseille, France. European Language Resources Association.

**Author Contributions:** Saif conceived the initial idea of the project and led the data collection effort from the social media platform Twitter (now X). Krishnapriya and Saif worked on the creation of the final dataset. Both authors developed the methodology and experimental details. Krishnapriya wrote the code and performed the experiments and analysis. Krishnapriya led the writing of the paper with Saif's supervision.

## 9.1 Overview

Over the last decade, Twitter has emerged as one of the most influential forums for social, political, and health discourse. In this paper, we introduce a massive dataset of more than 45 million geo-located tweets posted between 2015 and 2021 from US and Canada (*TUSC*), especially curated for natural language analysis. We also introduce *Tweet Emotion Dynamics (TED)* — metrics to capture patterns of emotions associated with tweets over time. We use TED and TUSC to explore the use of emotion-associated words across US and Canada; across 2019 (pre-pandemic), 2020 (the year the pandemic hit), and 2021 (the second year of the pandemic); and across individual tweeters. We show that Canadian tweets tend to have higher valence, lower arousal, and higher dominance than the US tweets. Further, we show that the COVID-19 pandemic had a marked impact on the emotional signature of tweets posted in 2020, when compared to the adjoining years. Finally, we determine metrics of TED for 170,000 tweeters to benchmark characteristics of TED metrics at an aggregate level. TUSC and the metrics for TED will enable a wide variety of research on studying how we use language to express ourselves, persuade, communicate, and influence, with particularly promising applications in public health, affective science, social science, and psychology.

## 9.2 Introduction

Over the last decade, Twitter has emerged not only as one of the most influential micro-blogging platforms, but also one of the most actively engaging (if sometimes polarizing) fronts for social, political, and even health discourse. Early work (Pak and Paroubek, 2010; Dodds et al., 2011) identified tweets as a crucial indicator of public sentiment. Since then, various samples of tweet data have been used to analyze a wide variety of phenomena, including the recent COVID-19 pandemic. However, past work largely uses topic-based keywords to obtain datasets of interest (often at the expense of geo-location information); for example, work that analyzes emotions in tweets that mention COVID-19-associated terms (Banda et al., 2020; Lwin et al., 2020). Further, very little work explores changes in patterns of emotions of individuals over time.

This paper introduces a new framework to analyze patterns of emotions associated with tweets over time, which we refer to as *Tweet Emotion Dynamics (TED).* TED builds on ideas first introduced by Hipson and Mohammad (2021), and applies metrics such as *home base, variability,* and *rise rate* to tweets. We also introduce a new dataset of geo-located English *Tweets from US and Canada (TUSC).* TUSC is not restricted to specific topics and so can be used to study tweets in general, as well as to study notable phenomena (such as a pandemic, climate change, or polarizing political events) on tweets at large (as opposed to examining tweets directly discussing those phenomena). TUSC also includes a subset, *TUSC100*, made up of tweets from 170,000 tweeters who each posted at least 100 tweets between 2020 and 2021. TUSC100 is especially well suited for longitudinal analysis. The creation of the datasets included careful post-processing to make the resource particularly suitable for textual analysis.

TUSC and TED can each be used, together or independently, to explore a wide range of research questions pertaining to tweets and emotions that may be of interest to researchers in Psychology, Affective Science, Social Science, Behavioural Science, Public Health, NLP, and Linguistics. In this paper, we use them to explore questions about how people use emotion-associated words in English tweets from US and Canada. We record the common characteristics of emotion word usage from 2015 to 2021, with a special focus on 2020 — the year that the WHO declared the Novel Coronavirus Disease (COVID-19) outbreak to be a pandemic — and its adjoining years (2019 and 2021). Finally, we benchmark individual tweeter behaviour in terms of various TED metrics. Recording this information holds considerable promise in future work; for example, for studying the emotional impact of the pandemic, for helping clinicians and patients track emotional well-being before and after health interventions, studying emotion regulation and coping strategies, etc. The data (tweet IDs), Emotion Dynamics code, and visualizations are freely available through the project homepage.[1]

## 9.3 Related Work

We group related work into two kinds: psychological and psychology-inspired research on the theory of emotions and utterance emotion dynamics; and NLP research in analyzing emotions in tweets.

### 9.3.1 Emotions

Several influential studies have shown that the three most fundamental, largely independent, dimensions of affect and connotative meaning are valence (V) (positiveness–negativeness / pleasure–displeasure), arousal (A) (active–sluggish), and dominance (D) (dominant–submissive / in control–out of control) (Osgood et al., 1957; Russell and Mehrabian, 1977; Russell, 2003). Valence and arousal specifically are commonly studied in a number of psychological and neuro-cognitive explorations of emotion.

The NRC VAD Lexicon (Mohammad, 2018a) (which we have worked with previously, in our study of dialogism) contains about twenty thousand commonly used English words (lemmas and common morphological variants) that have been scored on valence (0 = maximally unpleasant, 1 = maximally pleasant), arousal (0 = maximally calm/sluggish, 1 = maximally active/intense), and dominance (0 = maximally weak, 1 = maximally powerful).[2] As an example, the word *nice* has a valence of .93, an arousal of .44, and dominance of .65, whereas the word despair has a valence of

[1] https://github.com/Priya22/EmotionDynamics

[2] http://saifmohammad.com/WebPages/nrc-vad.html

.11, an arousal of .79, and dominance of .25. The lexicon was created using crowdsourced comparative annotations, where annotators are asked to select the terms that have the highest and lowest association with each dimension from among a set of four terms. A series of such annotations allows us to rank the entire set of lexicon terms in order of increasing association with the dimensions.

We have already been introduced to the UED framework Hipson and Mohammad (2021) in Chapter 7; we briefly re-introduce it here for continuity. The framework quantifies patterns of change of emotional states associated with utterances along a longitudinal (temporal) axis. Specifically, they proposed a series of metrics, including the following:

1. *Density or Mean*: A measure of the average utterance emotional state. This is calculated as the mean of emotion scores of the words in the utterance window.

2. *Variability*: The extent to which a speaker's utterance emotional state changes over time (measured as the standard deviation of the emotion states).

3. *Home Base:* A speaker's home base is the subspace of high-probability emotional states where they are most likely to be found. This is formulated as the range of values within one standard deviation of the average of the emotion states at each timestep.

4. *Rise and Recovery Rates:* Sometimes a speaker moves out of their home state, reaches a peak value of emotion state, before returning to the home state. The rise rate quantifies the rate at which a speaker moves towards the peak; recovery rate is the rate at which they go from the peak to the home state.

One can determine UED metrics using: 1. the utterances by a speaker, 2. the temporal information about the utterances, for example, time stamps associated with the utterances, or simply an ordering of utterances by time, and 3. features of emotional state drawn from text. The emotional state at a particular instant can be determined using lexical features (say, drawn from emotion lexicons), predictions of supervised machine learning systems, etc.

### 9.3.2 Analyzing Emotions in Tweets

Dodds et al. (2011) analyze large amounts of Twitter data to explore temporal patterns of 'societal happiness'. Larsen et al. (2015) show a correlation between patterns of emotional expression in tweets with WHO data on anxiety and suicide rates across

| Dataset | Canada | | | USA | | |
|---|---|---|---|---|---|---|
| | #tweets | # tweeters | Av.TpT | #tweets | # tweeters | Av.TpT |
| TUSC-Country | | | | | | |
| 2015 | 89,566 | 40,290 | 15.729 | 131,330 | 104,670 | 13.805 |
| 2016 | 93,280 | 40,994 | 16.164 | 133,413 | 109,110 | 14.305 |
| 2017 | 94,364 | 39,258 | 18.067 | 133,854 | 107,080 | 16.015 |
| 2018 | 95,403 | 38,866 | 21.763 | 133,066 | 105,227 | 19.394 |
| 2019 | 330,361 | 70,122 | 22.040 | 339,186 | 204,311 | 19.341 |
| 2015–2019 | 702,974 | 159,284 | 18.753 | 870,849 | 516,885 | 16.572 |
| 2020 | 321,176 | 57,465 | 22.123 | 503,976 | 250,080 | 19.698 |
| 2021 | 304,106 | 49,128 | 22.192 | 478,798 | 214,653 | 19.566 |
| 2015–2021 | 1,328,256 | 206,691 | 19.73 | 1,853,623 | 802,369 | 17.45 |
| TUSC-City | | | | | | |
| 2020 (Apr–Dec) | 15,039,503 | 716,063 | 19.275 | 23,470,855 | 2,669,081 | 17.556 |
| 2021 | 22,371,990 | 798,602 | 19.367 | 43,693,643 | 3,247,124 | 17.306 |
| 2020–2021 | 37,411,493 | 1,049,774 | 19.327 | 67,164,498 | 4,274,374 | 17.413 |

Table 9.1: Number of tweets, number of tweeters, and average number of tokens per tweet (Av.TpT) in the TUSC Datasets.

geographical location. Snefjella et al. (2018) analyze differences in language use in 40 million tweets from Canada and the USA, and find that the former tend to use more positive language, which correlates with national character stereotypes of Canadians being more agreeable and less aggressive. Twitter data has been used to study people's emotions during significant events, commonly revolving around certain tragedies and natural disasters, and significant political events. Doré et al. (2015) studied the changes in intensity of emotions of anxiety, anger, and sadness expressed on Twitter regarding the Sandy Hook Elementary School shooting. The 2016 US Presidential Election spurred several studies on the language used across geographical and political lines (Littman et al., 2016). Twitter was also used to measure the impacts of the COVID-19 pandemic on the emotional states and mental health of tweeters (Banda et al., 2020). Lwin et al. (2020), for example, looked at changes in the usage of tweets that expressed fear, anger, sadness, and joy in COVID-associated tweets from January 28 to April 9, 2020. In our work, we focus on the emotion dimensions of valence, arousal, and dominance, rather than categorical dimensions such as anger, fear, sadness, etc. We also study these patterns of emotion usage across a large time period (2015–2021), and in geo-located tweets.

## 9.4 Tweets Dataset: TUSC

### 9.4.1 Sampling Tweets

Twitter's regular API allows one to obtain a random sample of tweets from the past week. However, the search is limited to only the tweets from the past week. The Academic search API provides access to historical tweets, but with a lower rate limit. To benefit from both APIs and to confirm that our results are consistent regardless of the API and search method, we compiled two separate tweet datasets using each of the APIs:

1. Using Twitter's free API and its geo-location and random-sample switches to collect tweets from 46 prominent American and Canadian cities. Data collection began in April 1, 2020 and is ongoing. We refer to the dataset created with this method as *TUSC-City*.

2. Using Twitter's Academic API to collect tweets emanating from US and Canada from Jan 2015 to Dec 2021. The Academic API provides switches to specify the country of origin and the time span of search. However, the sample of results it provides tends to be in reverse chronological order for the specified time span. Thus, to obtain a sample of tweets from various time spans across the various years of interest, we employed the following strategy: For each year of interest, we randomly generated a date and time (using unix epoch seconds). We then specified a search interval of 8 hours starting from that date and time. We repeated this procedure thousands of times for each year. Since we were especially interested in the years of 2019, 2020, and 2021, we collected more data from these years. We refer to the resulting dataset as *TUSC-Country*.

### 9.4.2 Tweet Curation

We curated the tweet collection to make it more suitable for computational natural language analyses by applying the following steps:

- Kept one tweet per user, per day. This mitigates the impact of highly prolific tweeters and commercial accounts on the dataset.

- Kept only English language tweets (since the English set is the focus of this project). These are identified by the `iso_language` tag provided by Twitter for each tweet.

- Removed all retweets.

- Removed all tweets containing a URL and/or links to media (to focus on textual tweets). This also limits tweets by commercial organizations.

- Discarded all tweets with less than three tokens. This eliminates certain formulaic tweets such as wishes for holidays. The tweet text is tokenized using the Python implementation[3] of the Twokenizer package (Gimpel et al., 2010; Owoputi et al., 2013).

We kept quotes and replies as they include new textual information.

### 9.4.3 Key Data Statistics and Distribution

We organize the TUSC tweets as per the sampling strategy used to obtain them (see TUSC-Country and TUSC-City in §9.4.1) as well as the year of posting (2015 through 2021), and country of origin (US, Canada). Table 9.1 shows the number of tweets, number of tweeters, and average number of tokens per tweet in each of these dataset groupings. (Table 9.3 in the Appendix shows a breakdown by city for TUSC-City.) It is interesting that an average Canadian tweet has about two more tokens per tweet than a US tweet (one possible explanation is the tendency of American tweeters to use more informal and non-standard language, as found in Snefjella et al. (2018)).

TUSC-City is the larger dataset, and contains millions of tweets for many of the 46 cities for 9 months in 2020 (Apr–Dec), and all the months of 2021. It is useful for analyzing trends at the city–level, and also at the user-level, since we are more likely to have a large number of tweets from the same user.

## 9.5 Emotion Word Usage in American and Canadian Tweets

The TUSC datasets can be used to answer several important questions about emotion word usage in English tweets from US and Canada, including:

- Are there notable trends across years in the valence, arousal, and dominance of tweets? Are we tweeting with more positive words, more negative words, more high arousal words, etc. than in past years?

[3] https://github.com/myleott/ark-twokenize-py

- How has the COVID-19 pandemic impacted the emotionality of our tweets? At what point of time in the pandemic did we use the most amount of words conveying a lack of control and uncertainty? How were individual cities impacted?

- How are Canada and US different in terms of emotion word usage? Did the pandemic impact the emotionality differently in the two countries?

We will explore these, and other, questions below.

We used the NRC Valence, Arousal, and Dominance (NRC VAD) Lexicon (Mohammad, 2018a) to determine the emotion associations of the words in tweets. Specifically, we used the subset with entries for only the polar terms: i.e., only those valence entries that had scores $\leq 0.33$ (negative words) or scores $\geq 0.67$ (positive words).[4] Similarly, only those arousal and dominance entries were included that had scores $\leq 0.33$ or $\geq 0.67$. The entries with scores between 0.33 and 0.67 are considered neutral for that dimension.

**Methodological Note 1:** As is good practise in lexicon-based analysis (Mohammad, 2020), we removed lexicon entries for a small number of words that were highly ambiguous (e.g., *will, like*) or were expected to be frequently used in our tweets in a sense that is different from the usual predominant sense of the word (e.g., *trump*). The list of the 23 terms removed from the lexicon, and a description of the process of discovering them, is available in the Appendix.

**Methodological Note 2:** Similar analyses can also be performed using categorical emotions, such as joy, sadness, fear, anger, etc., using the NRC Emotion Lexicon (Mohammad and Turney, 2010, 2013).[5] See discussions on categorical and dimensional emotions in (Mohammad, 2021).

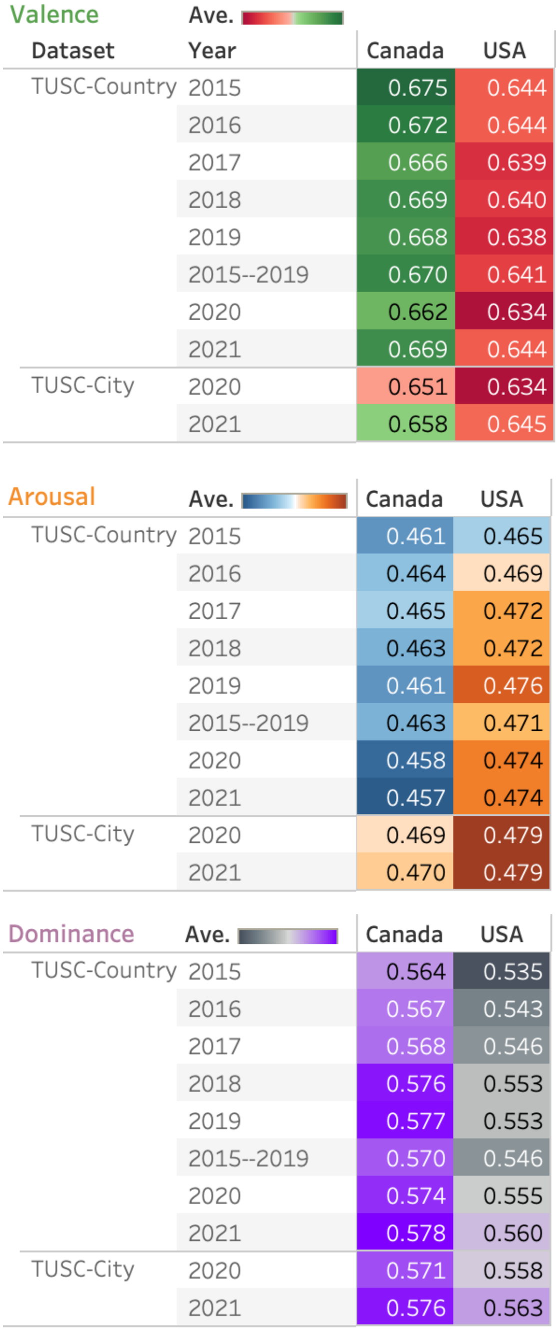
### 9.5.1 Average V, A, and D Across US–Canada

For each tweet, we take the average of the valence, arousal, and dominance values of each of the words in the tweet text. The averages are computed for TUSC-City over all tweets from each city, and for TUSC-Country at the country-level. We test whether the differences in values between countries and years are statistically significant by using the paired $t$-test, with the significance threshold for the $p$-value set to 0.001.

**Yearly Trends:** Figure 9.1 shows the average V, A, and D scores of tweets when

[4] There is no "correct" threshold to determine these classes; different thresholds simply make the positive and negative classes more or less restrictive.

[5] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm



| Valence<br>Dataset | Ave.<br>Year | Canada | USA |
|---|---|---|---|
| TUSC-Country | 2015 | 0.675 | 0.644 |
| | 2016 | 0.672 | 0.644 |
| | 2017 | 0.666 | 0.639 |
| | 2018 | 0.669 | 0.640 |
| | 2019 | 0.668 | 0.638 |
| | 2015--2019 | 0.670 | 0.641 |
| | 2020 | 0.662 | 0.634 |
| | 2021 | 0.669 | 0.644 |
| TUSC-City | 2020 | 0.651 | 0.634 |
| | 2021 | 0.658 | 0.645 |

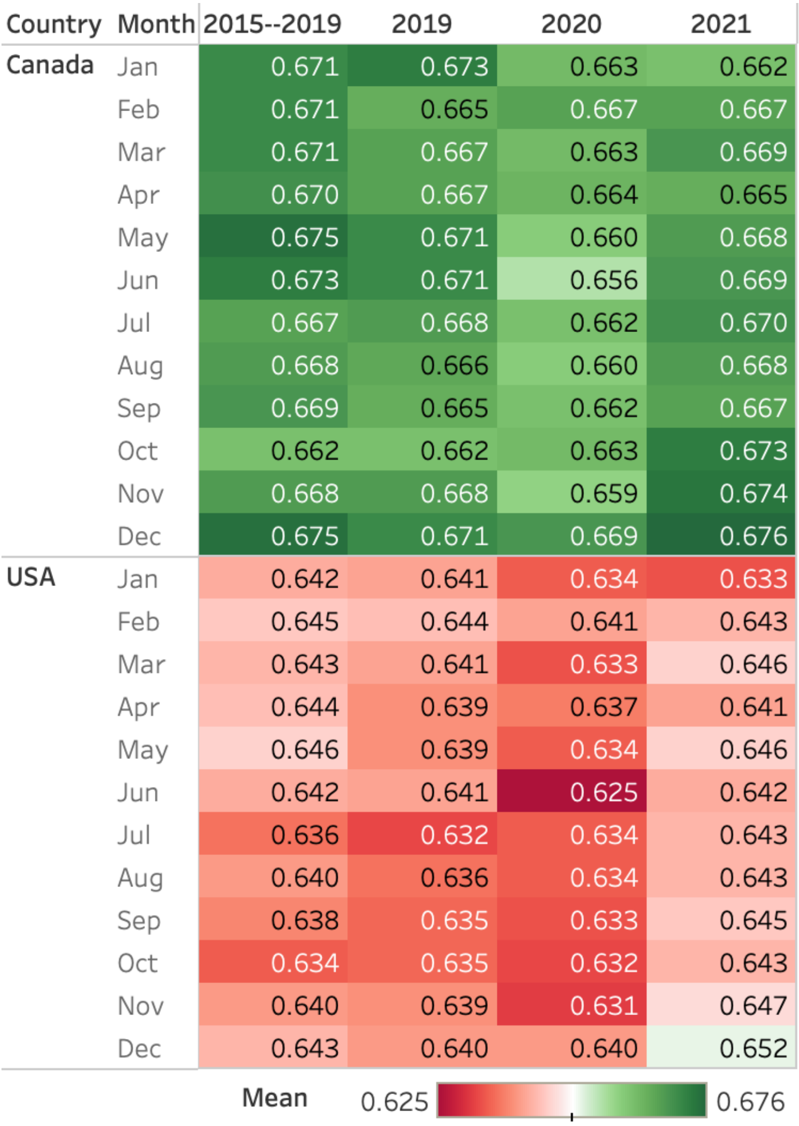
| Arousal | Ave. | Canada | USA |
|---|---|---|---|
| TUSC-Country | 2015 | 0.461 | 0.465 |
| | 2016 | 0.464 | 0.469 |
| | 2017 | 0.465 | 0.472 |
| | 2018 | 0.463 | 0.472 |
| | 2019 | 0.461 | 0.476 |
| | 2015--2019 | 0.463 | 0.471 |
| | 2020 | 0.458 | 0.474 |
| | 2021 | 0.457 | 0.474 |
| TUSC-City | 2020 | 0.469 | 0.479 |
| | 2021 | 0.470 | 0.479 |

| Dominance | Ave. | Canada | USA |
|---|---|---|---|
| TUSC-Country | 2015 | 0.564 | 0.535 |
| | 2016 | 0.567 | 0.543 |
| | 2017 | 0.568 | 0.546 |
| | 2018 | 0.576 | 0.553 |
| | 2019 | 0.577 | 0.553 |
| | 2015--2019 | 0.570 | 0.546 |
| | 2020 | 0.574 | 0.555 |
| | 2021 | 0.578 | 0.560 |
| TUSC-City | 2020 | 0.571 | 0.558 |
| | 2021 | 0.576 | 0.563 |

Figure 9.1: Average Valence, Arousal, and Dominance of words per tweet for each dataset (year).

aggregated at the country and year level, for the various data subsets. The color gradients at the top of each section show where the values lie in the spectrum from lowest to highest.

*Valence:* Observe that the average valence of Canadian tweets is consistently higher



| Country | Month | 2015--2019 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|
| Canada | Jan | 0.671 | 0.673 | 0.663 | 0.662 |
| | Feb | 0.671 | 0.665 | 0.667 | 0.667 |
| | Mar | 0.671 | 0.667 | 0.663 | 0.669 |
| | Apr | 0.670 | 0.667 | 0.664 | 0.665 |
| | May | 0.675 | 0.671 | 0.660 | 0.668 |
| | Jun | 0.673 | 0.671 | 0.656 | 0.669 |
| | Jul | 0.667 | 0.668 | 0.662 | 0.670 |
| | Aug | 0.668 | 0.666 | 0.660 | 0.668 |
| | Sep | 0.669 | 0.665 | 0.662 | 0.667 |
| | Oct | 0.662 | 0.662 | 0.663 | 0.673 |
| | Nov | 0.668 | 0.668 | 0.659 | 0.674 |
| | Dec | 0.675 | 0.671 | 0.669 | 0.676 |
| USA | Jan | 0.642 | 0.641 | 0.634 | 0.633 |
| | Feb | 0.645 | 0.644 | 0.641 | 0.643 |
| | Mar | 0.643 | 0.641 | 0.633 | 0.646 |
| | Apr | 0.644 | 0.639 | 0.637 | 0.641 |
| | May | 0.646 | 0.639 | 0.634 | 0.646 |
| | Jun | 0.642 | 0.641 | 0.625 | 0.642 |
| | Jul | 0.636 | 0.632 | 0.634 | 0.643 |
| | Aug | 0.640 | 0.636 | 0.634 | 0.643 |
| | Sep | 0.638 | 0.635 | 0.633 | 0.645 |
| | Oct | 0.634 | 0.635 | 0.632 | 0.643 |
| | Nov | 0.640 | 0.639 | 0.631 | 0.647 |
| | Dec | 0.643 | 0.640 | 0.640 | 0.652 |

Figure 9.2: Monthly trends in valence (TUSC-Country).

(more positive) than the tweets from the US (statistically significant); the difference is steady across years. There is a slight downward trend for valence in both countries from 2015 to 2019. We see the lowest values of mean valence occur in 2020 for both TUSC-Country and TUSC-City (the year the pandemic hit) for both the US and Canada. Average valence rises back up in 2021.

*Arousal:* Overall, tweets from Canada have lower average arousal (more calm, less active) than the US (statistically significant). Again, the difference in mean between the two countries remains relatively steady across years. Across years, arousal values for both countries increase from 2015 till 2017; they then drop steadily for Canada, while the USA sees a peak in 2019 followed by slight drops in subsequent years.

*Dominance:* Canada on the whole consistently has higher dominance values (greater feeling of control) than the US across the years. Both countries have the lowest dominance values in 2015, and the highest in 2021. For all three dimensions, we note that the yearly trends observed in TUSC-Country are largely also observed in the TUSC-City trends, across 2020–2021.

**Monthly Trends:** Figure 9.2 shows a breakdown of the average valence scores at the month level, across years. (Figures 9.6 and 9.7 in the Appendix show the breakdown for arousal and dominance.)

We immediately notice from the color shading that Canada consistently has higher valence (green), lower arousal (blue), and higher dominance (purple) than the US, across the months and years. June 2020 is particularly notable as it has the lowest values of valence for both USA and Canada; we hypothesize that this is an effect of both the COVID-19 pandemic (the seriousness of which was starting to become evident a couple of months earlier in March 2020) and the *Black Lives Matter* protests (which peaked after a racially-motivated shooting incident in the US around June 2020). By contrast, the final months of 2021 have the highest positivity. This could be attributed to feelings of a potential return to normalcy, and a general uptick in mood due to the holiday season (this was just before the Omicron variant of COVID-19 took root in early 2022).

The dominance numbers indicate that April and May of 2020 for Canada and the USA are marked by some of the lowest scores, suggestive of a feeling of loss of control due to the onset of the global COVID-19 pandemic.

### 9.5.2 Tweets with Emotional Terms

The experiments above showed notable differences in the average VAD scores of tweets across US and Canada. However, they also lead to further questions such as whether the higher valence in Canadian tweets is because of a greater usage of positive words or a lower usage of negative words. To explore such questions, we determine how frequently people post tweets with at least one high-valence word, how frequently people post tweets with at least one low-valence word, how frequently people post tweets with at least one high-arousal word, and so on. High and low categorization of a word is based on whether its score on the relevant dimension is $\geq 0.67$ or $\leq 0.33$, respectively. Figure 9.3 shows the results.

The darker shades of the color indicate a greater percentage of tweets had at least one of the relevant emotional words.

Observe that in both American and Canadian tweets:

- people post markedly more tweets with at least one positive word than tweets with at least one negative word (about 100% more).
- people post markedly more tweets with at least one low-arousal word than tweets with at least one high-arousal word (about 40% more).
- people post markedly more tweets with at least one high-dominance word than tweets with at least one low-dominance word (about 33% more).

In terms of their differences, we see that the tweets from Canada are marked by both a higher usage of high-valence words, as well as a lower usage of low-valence words, than the US (statistically significant). Tweets from Canada have a higher proportion of low-arousal words, whereas high-arousal word usage is similar in both countries. Canadian tweeters use about the same number of low-dominance words as those in the US, but use a greater number of high-dominance words.

Across years, low-valence words increase in usage relatively steadily until 2020, and drop in 2021. For all dimensions, the sharpest rise in usage occurs from 2016 to 2017. When comparing TUSC-Country 2020 with 2021, observe that the higher number of low-valence words used is more prominent than the lower number of high-valence words — thus, the drop in average valence in 2020 (Figure 9.1) is because people tweeted *more negative words* (and not because people tweeted *fewer positive words*).

## 9.6 Tweet Emotion Dynamics

The previous section explored emotion patterns in terms word usage frequencies, and examined city and country-level averages. In this section we apply the UED framework to quantify *individual* tweeter behaviour over time, focusing on the following research questions:

- What is the usual range and distribution of various metrics of tweet emotion dynamics (TED), such as mean and recovery rate, for American and Canadian tweeters? Establishing benchmarks for these metrics is crucial for subsequent studies that may explore, for example, the impact of a health intervention on one's TED metrics.
- Are there notable differences in the distributions of TED metrics across American and Canadian tweeters?

- Are there notable differences in the distributions of TED metrics across 2020 and 2021?

### 9.6.1 Implementation Details

As before, we compute UED metrics for individual tweeters in our dataset separately for each of the three affective dimensions of valence, arousal, and dominance. This allows us to expand the initial set of 2D-space metrics proposed in Hipson and Mohammad (2020) to include separate averages of rise and recovery rates for displacement below and above the home base (Hm-Hi, Hi-Hm, Hm-Lo, and Lo-Hm rates); we refer to this analysis henceforth as *Tweet Emotion Dynamics*, or *TED*. The timestamp of a tweet provides the temporal order. The tweets from each speaker are then concatenated together and tokenized to obtain a ordered list of tokens. Next, a rolling window of 20 tokens is considered to determine the average V, A, and D scores of the words in that window. These scores are a representation of utterance emotional state corresponding to that window. The rolling window is moved forward one word at a time to determine the subsequent averages.[6] In the rest of this section, we use the term *mean* to refer to the mean of all the rolling window averages for a speaker.

We determine TED metrics for the tweeters in the TUSC-Country dataset. Only tweeters with at least 100 tweets in a year were considered, since drawing inferences about one's tweeting behavior requires a sufficient sample size. There were about 40,000 such tweeters in the 2020 subset and about 130,000 such tweeters in the 2021 subset. We refer to their tweets (5.6 million from 2020 and 19 million from 2021) as the *TUSC100-2020* and *TUSC100-2021* datasets, respectively. The average number of tweets by a tweeter in these datasets is 153 (no tweeters had more than 365 tweets due to our earlier stated 'one tweet per user per day' pre-processing policy). See Table 9.4 in the Appendix for detailed statistics.

### 9.6.2 Results

Figure 9.4 plots the distributions of some of the metrics for the joint set of 2020 and 2021 tweeters. The plots in (a) are distributions of the mean values for the three emotion dimensions (V, A, D). The x-axis is made up of bins of size 0.005 (from 0 to 0.005, to 0.995 to 1). The y-axis indicates the number of tweeters with mean values in each of the bins. Observe that the means for V, A, and D all follow a near-normal

[6] Variations of this approach that do not use rolling windows across tweet boundaries produce similar results.

distribution. Mean valence and dominance values are more spread out compared to arousal values. For V, most fall between 0.5 and 0.8, with a median value of around 0.65, though we can see that there is a long tail of outliers. Dominance scores are spread around a median score of 0.6, and the median is even lower for arousal (0.49). Figure 9.4(b) shows distributions for rises rates and recovery rates. Observe that these have a much narrower spread, and the distributions for all three dimensions are roughly the same.

Figure 9.5 shows box and whisker plots of the same three metrics: mean, rise rate, and recovery rates. However, separate plots are shown for tweeters from US and Canada, and across 2020 and 2021. The shaded region (the box) indicates the "middle portion" of the data distribution, i.e, the range covered between the first quartile (the 25% mark) and the third quartile (the 75% mark), with the median (50% mark) lying at the border of the light and dark shaded regions. The whiskers, the lines on either end of the plot, are at a distance of 1.5 times the inter-quartile length (inter-quartile length is the distance between the first and third quartiles). Points beyond the whiskers are considered outliers. Additionally, the average value (mean) is indicated with the pink horizontal dashed line.

Observe that the mean valence is lower in 2020 than in 2021, and Canadian tweeters on average use more positive words than their US counterparts.[7] The distributions for mean arousal are quite similar across 2020 and 2021, but US tweeters have slightly higher mean arousal values. Canadian tweeters have a slightly higher median of dominance scores than US tweeters; whereas the US tweeters tend to have a wider range of dominance values. The difference in the distributions of the mean values for Canada and US is statistically significant for all three dimensions ($p$-values $< 0.001$).

The median rise rates and recovery rates do not differ markedly across countries or years. However, there is a notably large range of the third quartile (the quartile above the median) for Canadian tweeters in 2021. These are tweeters who are quicker to jump in and out of their home base. Tables 9.5, 9.6, 9.7 of the Appendix report mean scores for all of the TED metrics, averaged across all tweeters by country and year. This includes a breakdown of the rates into Hm–Hi, Hi–Hm, Hm–Lo, and Lo–Hm. Notable trends there are that the average rise and recovery rates on the high side of the home state (Hm-Hi, Hi-Hm) are lower than for the low side of the home state (Lo-Hm, Hm-Lo), for the valence and dominance dimensions. This says that tweeters are slower to rise to more positive and more dominant states, but quicker to both descend to more negative and less dominant states, and recover from them; similarly,

[7] These trends align with the trends observed in Table 2. Fix

they are slower to transit to and from states of high activity (high dominance). This difference between Hi and Lo rates is reversed for arousal. Thus, tweeters are quicker to rise to states of high arousal, and come back down from them to the home state.

We also noticed in our analyses that there exist several tweeters who have very high rise rates but normative recovery rates, and also tweeters that have very high recovery rates but normative rise rates. Identifying such characteristics and tracking them in the context of health interventions is particularly promising future work. However, it should be noted that we strongly encourage such studies, when conducted, to be led by clinicians and psychologists, with appropriate consent and ethics approvals.

### 9.6.3 City as Speaker

An interesting variation of the experiments above, is to consider each city as a 'speaker', rather than individual tweeters. Figure 9.8, in the Appendix, shows the average TED metrics for each of the 46 cities in TUSC-City. The color gradients make it easy to spot which cities have had markedly high VAD means across 2020 and 2021. Consistent with some of the earlier country-level results, we see that the Canadian cities tend to have higher valence, lower arousal, and higher dominance, than the US cities. London, Ottawa, Halifax, and Victoria have the highest valence (most positive). From the set of Canadian cities, Windsor stands out as an anomaly with valence close to many US cities. Detroit, Houston, Los Angeles, and Philadelphia have some of the lowest valence values of all cities. All cities show an increased valence from 2020 to 2021, some more drastically than others (Boston, Indianapolis, San Jose, for example). Quebec City and Windsor have the highest arousal rates in Canada; in the US, El Paso is at the top for both years. Nashville, San Francisco, San Jose, and Seattle have lower arousal rates (more in line with the average Canadian city). Washington, San Jose, and Boston also show markedly high dominance, as well as San Francisco. Among Canadian cities, Ottawa and Victoria have the highest dominance scores for 2020 and 2021, and Windsor again the lowest.

Figure 9.9, in the Appendix, shows values of the variabilities, rise rates, and recovery rates for the valence dimension. Looking at the column for variability, Windsor jumps out among the Canadian cities for having comparatively higher variability. Washington and Phoenix in 2020 have relatively high variability. Moving to the next columns, Windsor again has the highest rise and recovery rates among Canadian cities; US cities are the on the whole quicker to rise and fall.

The various metrics listed for various cities should be useful to those interested in the tweets from particular cities. Future work will drill down further into the data for individual cities to determine the factors driving the emotion word usage.

## 9.7 Conclusion

We introduced the Tweet Emotion Dynamics (TED) framework to quantify changes in emotions associated with tweets over time. We also released TUSC — a large collection of English geo-located tweets from Canada and the USA that were posted between 2015 and 2021. We studied emotion word usage in this data, using multiple metrics, for the primary dimensions of valence, arousal, and dominance. Our results showed interesting trends in the emotions expressed by tweeters from the two countries across different years, and also uncovered contrasts between Canadian and US tweeters. Future work will explore tweets from other countries and also tweets in languages other than English. In Section 9.D of the Appendix, I expand on follow-up work that looks at how the TED framework can be useful to clinicians and psychologists for measuring mental health outcomes, at an aggregate level, from social media data.

# Appendix

## 9.A Ethics Considerations

Emotions are complex, private, and central to our experience. There is also tremendous variability in how we express emotions through language. Thus several ethical considerations are relevant to textual analysis of emotions. Some that we would particularly like to highlight are listed below:

- We only release the tweet IDs for each tweet, which will need to be populated with the tweet text by users of our dataset with the Twitter API. If any of the tweets are deleted by the associated tweeter, they will no longer be accessible.
- Our work on studying emotion word usage should not be construed as detecting how people feel; rather, we draw inferences on the emotions that are conveyed by users via the language that they use. The language used in an utterance may convey information about the emotional state (or perceived emotional state) of the speaker, listener, or someone mentioned in the utterance. However, it is not sufficient for accurately determining any of their momentary emotional states. Deciphering true

momentary emotional state of an individual requires extra-linguistic context and world knowledge. Even then, one can be easily mistaken.

- The inferences we draw in this paper are based on aggregate trends across large populations. We do not draw conclusions about specific individuals or momentary emotional states.
- We do not recommend the use of TED metrics to draw inferences about individuals, unless: 1. it is exercised with extreme caution, 2. for the express benefit, and with consent, of the people whose data is used, 3. the work is led by subject-matter experts such as psychologists or clinicians, and 4. automatically drawn information is used as one source of information among many by human experts.
- Any information drawn from these metrics regarding one's language use should not be used to negatively impact the individual.

See Mohammad (2022) for a detailed discussion on the ethical considerations of automatic emotion recognition and Mohammad (2020) for practical and ethical considerations in the effective use of emotion lexicons.

## 9.B Additional Emotion Word Usage Statistics from TUSC

We present, in this section, additional tables and figures that record details of emotion word usage broken down by city (for the 46 cities considered) and by month of year. Table 9.4 shows the number of tweets and tweeters in each subset of the TUSC100 dataset. Tables 9.5, 9.6, and 9.7 tabulate the numbers corresponding to the plots in Figure 9.5 in the main paper.

## 9.C Modified NRC VAD Lexicon

When applying lexicon-based analyses to datasets from a specific domain, Mohammad (2020) recommends updating the emotion lexicons to remove terms that can be used in a sense different from the predominant word sense. Since manual examination of all the words in a large dataset is difficult, this step is recommended for at least the frequent terms.

For our analyses, we first compiled a list of all the terms from the NRC VAD lexicon that occurred in at least 0.1% of the tweets from either Canada or the USA, for any of

| | | | |
|---|---|---|---|
| have | will | one | high |
| may | way | kind | be |
| thing | things | number | seem |
| do | look | three | third |
| five | senate | say | talk |
| president | trump | like | |

Table 9.2: List of terms that were removed from the NRC VAD Lexicon.

the years in the TUSC-Country dataset (2015–2021). Both of the authors of this work examined the list and identified words that were highly ambiguous or occurred in the tweets predominantly in a sense different from what would be expected if people were shown the word out of context (as was the case of the original annotations in the NRC VAD lexicon). In all, 23 such words were identified (shown in Table 9.2). The entries for these words were removed from the lexicon before conducting the experiments described in the paper.

To examine the impact of the above lexicon update, we repeated all the experiments with the unmodified lexicon as well. We observed that while the numerical values of the UED metrics changed slightly (as would be expected), all relative trends remained the same, across countries and across years. The interested reader can find the scores for the UED metric and the complete set of experiments with the unmodified lexicon in version 2 (an older version) of the paper on ArXiv.[8]

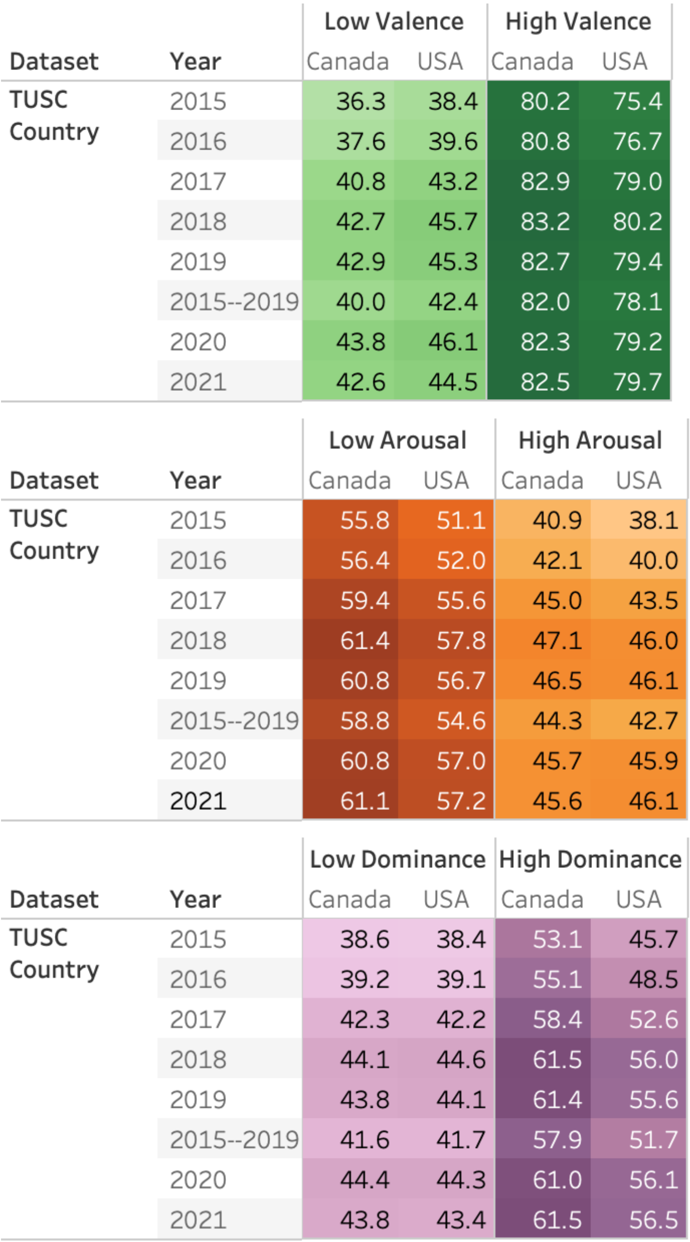
[8] https://arxiv.org/abs/2204.04862

## 9.D Further Work on Metrics of Emotional Variation

Here, I briefly describe some of my continuing work on computational metrics of emotional dynamics and variation derived from utterances that is not a part of this thesis. The purpose of this is to contextualize the standalone chapter on emotional variation on Twitter data, which is a departure from the literary focus of the majority of this thesis.

Over the last year, I have collaborated with researchers from NRC Canada and the Affective Sciences Lab at the University of North Carolina, Chapel Hill, to extend and validate the work on Tweet Emotion Dynamics. Our primary collaborators at UNC are PhD student Mallory Feldman, and the lab PI Kristin Linqdist, whose research agendas heavily revolve around the characterizing emotion dynamics in people and populations. One part of this work looks at whether the UED metrics derived for different cities in the TUSC dataset correlated with census-level data that catalogues various indicators of mental and physical health for populations. The census data is released as county-level aggregates, rather than at the city-level; we therefore collected an extended dataset of tweets with county-level co-ordinates, and computed TED metrics for the same. In addition to the 1D metrics for valence, arousal, and dominance, we also extend the framework to 2D metrics in the valence–arousal space, as is more commonly done in psychology.

A second thread of work is based on the concept of *emotion granularity*, alternatively referred to as emotion differentiation. First proposed and studied in Barrett et al. (2001), granularity quantifies one's ability to distinguish between emotional experiences with a high degree of specificity. A high emotional granularity indicates a greater ability to identify the emotions being felt in different situations, and is subsequently linked to better mental and physical health outcomes. We are operationalizing this concept for text utterances using both the emotion arcs generated with the UED framework, and with computational measures of semantic similarity of emotion-indicating utterances. These measures can be validated by studying the variation in the computed granularity metric with the mental health status of speakers; we are using several NLP datasets on mental health conditions for this purpose.

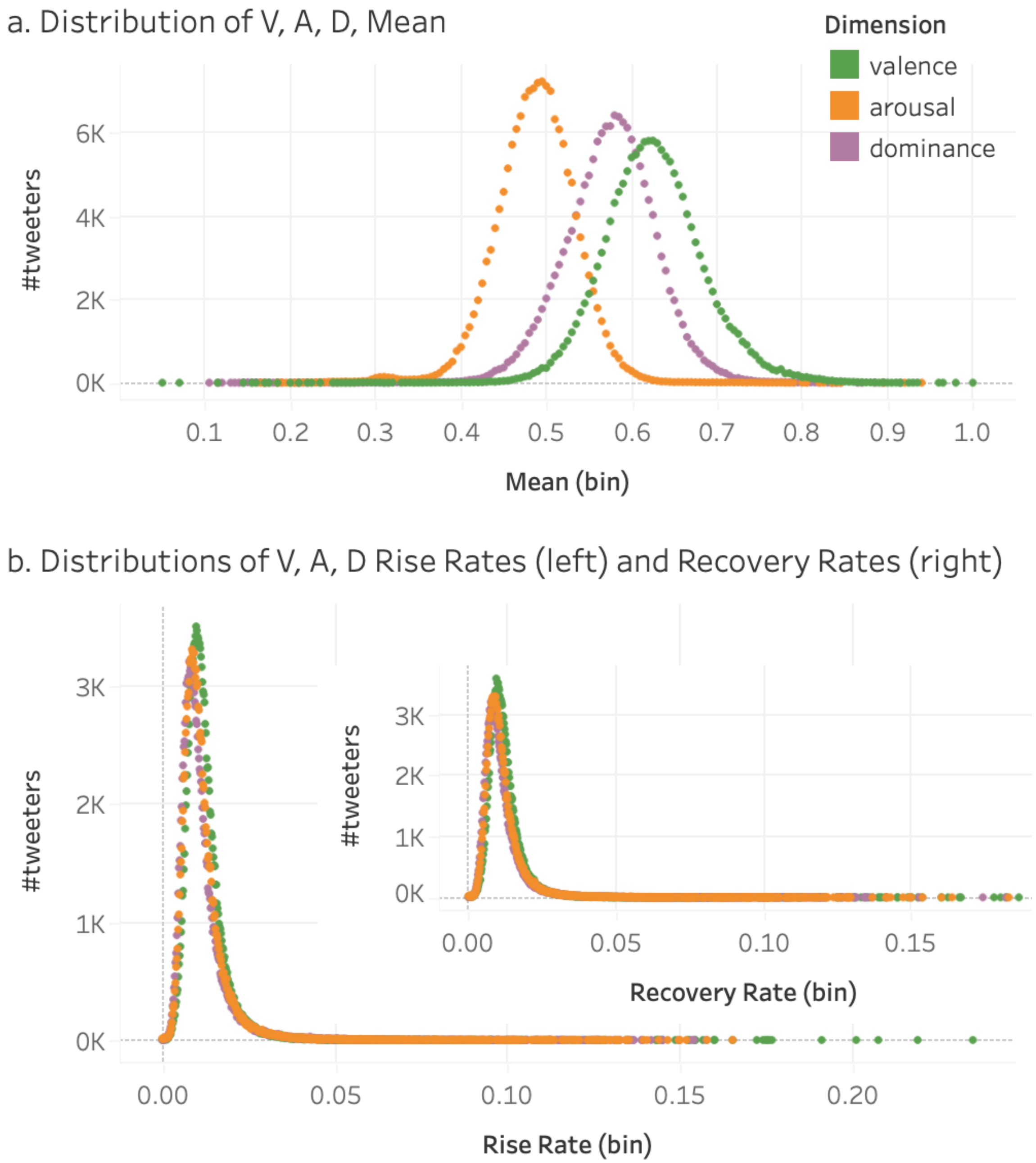


| Dataset | Year | Low Valence | | High Valence | |
|---|---|---|---|---|---|
| | | Canada | USA | Canada | USA |
| TUSC Country | 2015 | 36.3 | 38.4 | 80.2 | 75.4 |
| | 2016 | 37.6 | 39.6 | 80.8 | 76.7 |
| | 2017 | 40.8 | 43.2 | 82.9 | 79.0 |
| | 2018 | 42.7 | 45.7 | 83.2 | 80.2 |
| | 2019 | 42.9 | 45.3 | 82.7 | 79.4 |
| | 2015--2019 | 40.0 | 42.4 | 82.0 | 78.1 |
| | 2020 | 43.8 | 46.1 | 82.3 | 79.2 |
| | 2021 | 42.6 | 44.5 | 82.5 | 79.7 |

| Dataset | Year | Low Arousal | | High Arousal | |
|---|---|---|---|---|---|
| | | Canada | USA | Canada | USA |
| TUSC Country | 2015 | 55.8 | 51.1 | 40.9 | 38.1 |
| | 2016 | 56.4 | 52.0 | 42.1 | 40.0 |
| | 2017 | 59.4 | 55.6 | 45.0 | 43.5 |
| | 2018 | 61.4 | 57.8 | 47.1 | 46.0 |
| | 2019 | 60.8 | 56.7 | 46.5 | 46.1 |
| | 2015--2019 | 58.8 | 54.6 | 44.3 | 42.7 |
| | 2020 | 60.8 | 57.0 | 45.7 | 45.9 |
| | 2021 | 61.1 | 57.2 | 45.6 | 46.1 |

| Dataset | Year | Low Dominance | | High Dominance | |
|---|---|---|---|---|---|
| | | Canada | USA | Canada | USA |
| TUSC Country | 2015 | 38.6 | 38.4 | 53.1 | 45.7 |
| | 2016 | 39.2 | 39.1 | 55.1 | 48.5 |
| | 2017 | 42.3 | 42.2 | 58.4 | 52.6 |
| | 2018 | 44.1 | 44.6 | 61.5 | 56.0 |
| | 2019 | 43.8 | 44.1 | 61.4 | 55.6 |
| | 2015--2019 | 41.6 | 41.7 | 57.9 | 51.7 |
| | 2020 | 44.4 | 44.3 | 61.0 | 56.1 |
| | 2021 | 43.8 | 43.4 | 61.5 | 56.5 |

Figure 9.3: Percentage of tweets with at least one low-valence word, high-valence word, low-arousal word, high-arousal word, low-dominance word, high-dominance word — across datasets (years).


a. Distribution of V, A, D, Mean

b. Distributions of V, A, D Rise Rates (left) and Recovery Rates (right)

Figure 9.4: Distributions of Means, Rise Rates, and Recovery Rates for Valence, Arousal, and Dominance (TUSC100).

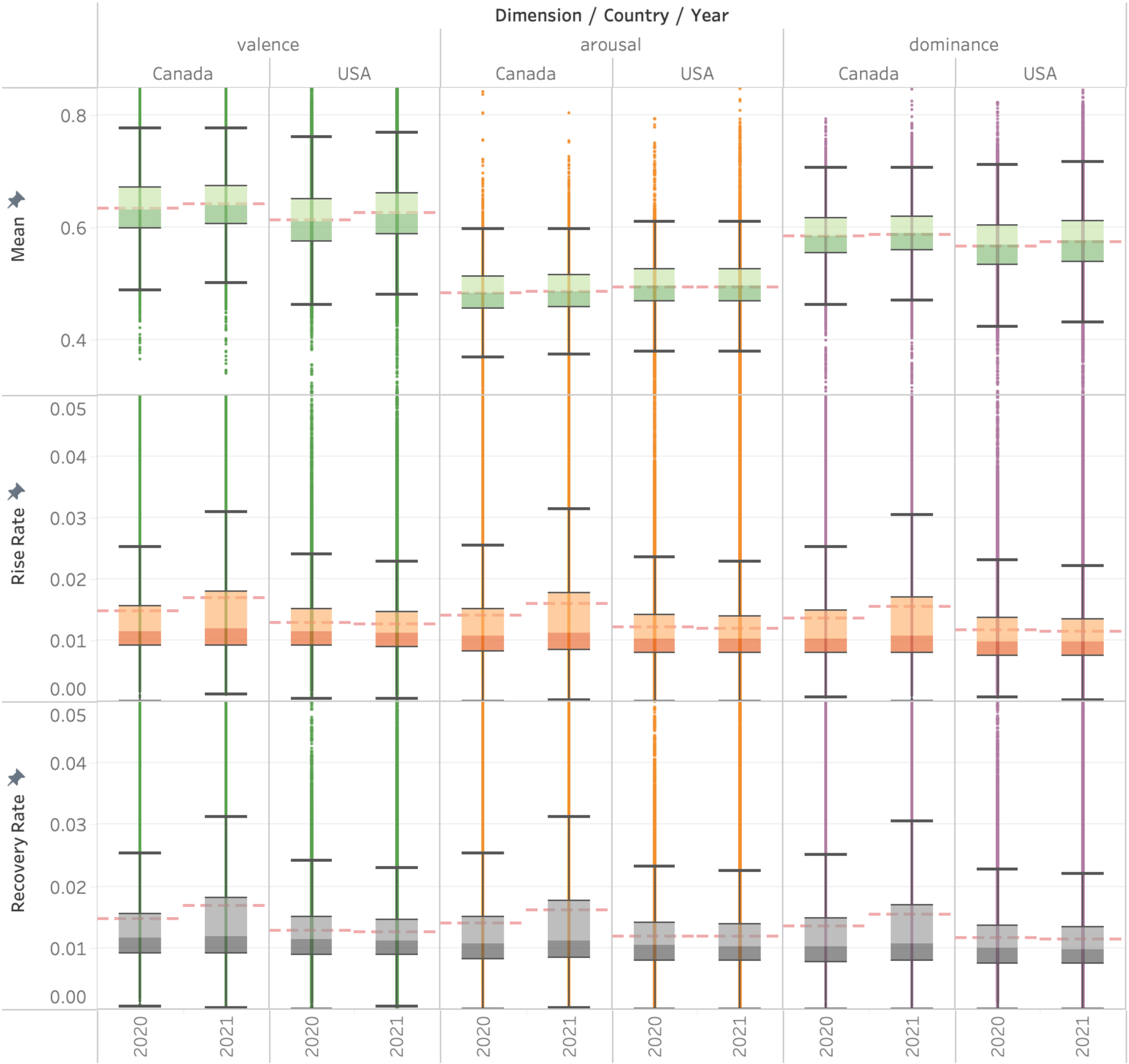
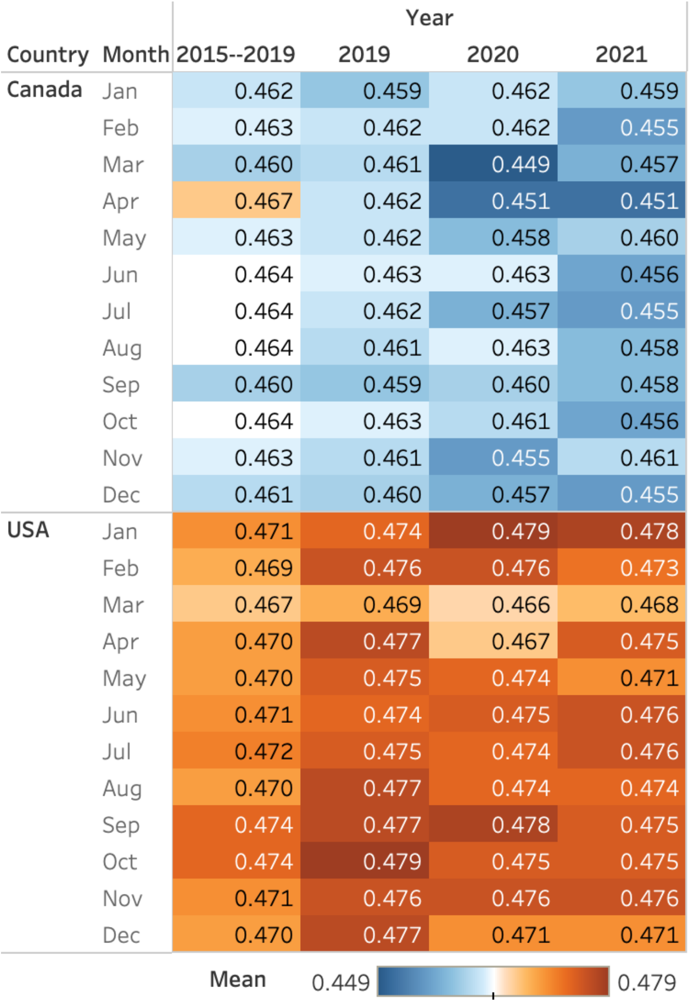


Figure 9.5: Box plots of means, rise rates, and recovery rates of Valence, Arousal, and Dominance of tweeters in 2020 and 2021 (TUSC100-2020 and TUSC100-2021).

|  | 2020 |  | 2021 |  |
|---|---|---|---|---|
| **City** | **# tweets** | **# tweeters** | **# tweets** | **# tweeters** |
| *Canada* |  |  |  |  |
| Brampton | 1,436,865 | 159,974 | 2,430,329 | 188,216 |
| Calgary | 294,911 | 31,988 | 503,173 | 39,416 |
| Edmonton | 806,116 | 43,427 | 1,319,950 | 49,058 |
| Etobicoke | 1,318,119 | 157,429 | 2,379,928 | 191,653 |
| Halifax | 572,562 | 23,733 | 678,033 | 23,541 |
| Hamilton | 446,038 | 37,023 | 702,761 | 43,537 |
| Laval | 453,670 | 48,145 | 733,844 | 58,344 |
| London | 298,615 | 16,977 | 428,929 | 18,928 |
| Mississauga | 450,835 | 97,328 | 977,517 | 142,817 |
| Montreal | 627,159 | 52,396 | 1,048,093 | 64,363 |
| North York | 1,274,462 | 148,271 | 1,685,201 | 152,105 |
| Okanagan | 30,771 | 1,814 | 37,424 | 1,813 |
| Ottawa | 1,055,035 | 55,430 | 1,332,680 | 56,621 |
| Quebec | 284,665 | 16,380 | 377,342 | 18,100 |
| Scarborough | 720,165 | 108,498 | 710,181 | 86,328 |
| Surrey | 1,115,467 | 84,001 | 1,679,642 | 94,177 |
| Toronto | 2,058,494 | 182,730 | 2,557,606 | 182,792 |
| Vancouver | 402,418 | 53,655 | 634,307 | 63,561 |
| Victoria | 340,720 | 14,787 | 436,905 | 15,565 |
| Windsor | 443,712 | 58,545 | 893,922 | 72,975 |
| Winnipeg | 608,704 | 27,954 | 824,223 | 29,365 |
| *US* |  |  |  |  |
| Austin | 1,244,776 | 102,841 | 2,242,561 | 125,526 |
| Boston | 764,257 | 100,276 | 1,641,142 | 130,145 |
| Charlotte | 997,197 | 76,528 | 1,566,062 | 86,892 |
| Chicago | 721,075 | 142,591 | 1,652,701 | 194,565 |
| Columbus | 809,160 | 69,931 | 1,445,275 | 81,135 |
| Dallas | 674,613 | 129,304 | 1,671,887 | 181,895 |
| Denver | 1,198,813 | 98,697 | 1,712,785 | 106,959 |
| Detroit | 749,506 | 77,560 | 1,418,484 | 94,202 |
| El Paso | 692,705 | 38,096 | 781,937 | 37,335 |
| Fort Worth | 1,649,842 | 188,443 | 2,794,053 | 215,169 |
| Houston | 1,557,488 | 195,548 | 2,358,286 | 209,520 |
| Indianapolis | 710,808 | 65,665 | 1,287,399 | 78,214 |
| Jacksonville | 723,513 | 48,230 | 1,000,620 | 53,257 |
| Los Angeles | 1,028,102 | 246,491 | 2,470,750 | 337,004 |
| Memphis | 876,988 | 49,835 | 1,263,728 | 54,254 |
| Nashville | 584,997 | 68,306 | 963,947 | 79,006 |
| New York | 1,079,557 | 281,109 | 2,288,500 | 361,670 |
| Philadelphia | 1,142,818 | 127,257 | 2,579,605 | 161,491 |
| Phoenix | 531,390 | 81,093 | 1,454,042 | 117,145 |
| San Antonio | 1,213,537 | 91,427 | 1,835,376 | 98,868 |
| San Diego | 1,080,361 | 101,554 | 1,940,120 | 122,278 |
| San Francisco | 1,123,356 | 132,023 | 2,179,371 | 159,727 |
| San Jose | 790,511 | 72,049 | 1,184,999 | 78,750 |
| Seattle | 940,092 | 114,848 | 1,968,319 | 141,016 |
| Washington | 585,393 | 123,576 | 1,991,694 | 189,873 |
| **All** | **38,510,358** | **3,332,189** | **66,065,633** | **3,976,481** |

Table 9.3: The number of tweets and tweeters in TUSC-City for 2020 and 2021.

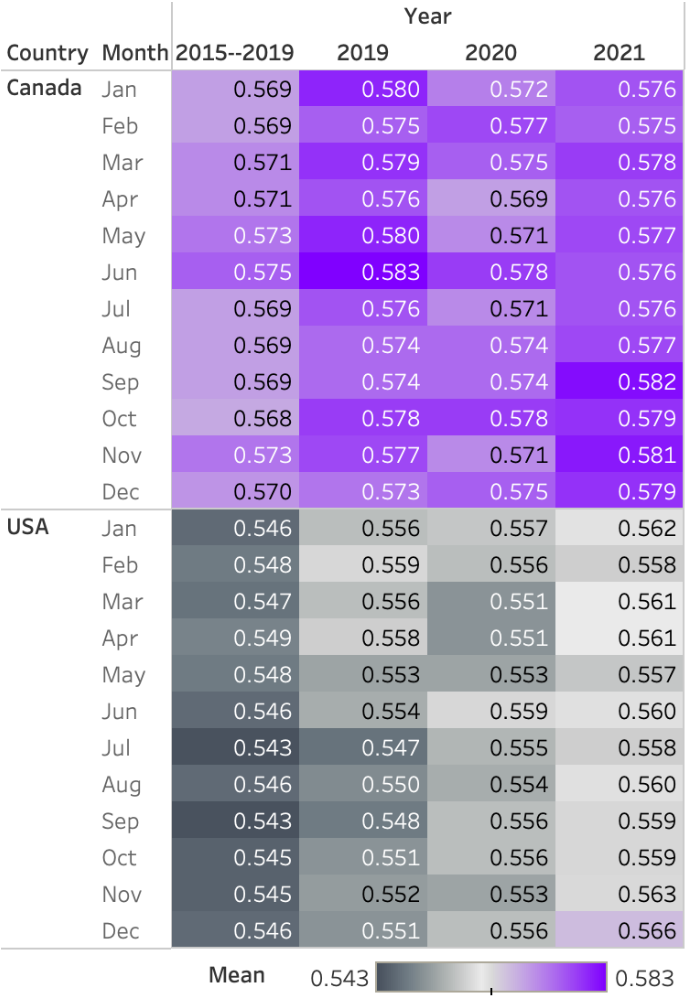


| | | Year | | | |
|---|---|---|---|---|---|
| Country | Month | 2015--2019 | 2019 | 2020 | 2021 |
| Canada | Jan | 0.462 | 0.459 | 0.462 | 0.459 |
| | Feb | 0.463 | 0.462 | 0.462 | 0.455 |
| | Mar | 0.460 | 0.461 | 0.449 | 0.457 |
| | Apr | 0.467 | 0.462 | 0.451 | 0.451 |
| | May | 0.463 | 0.462 | 0.458 | 0.460 |
| | Jun | 0.464 | 0.463 | 0.463 | 0.456 |
| | Jul | 0.464 | 0.462 | 0.457 | 0.455 |
| | Aug | 0.464 | 0.461 | 0.463 | 0.458 |
| | Sep | 0.460 | 0.459 | 0.460 | 0.458 |
| | Oct | 0.464 | 0.463 | 0.461 | 0.456 |
| | Nov | 0.463 | 0.461 | 0.455 | 0.461 |
| | Dec | 0.461 | 0.460 | 0.457 | 0.455 |
| USA | Jan | 0.471 | 0.474 | 0.479 | 0.478 |
| | Feb | 0.469 | 0.476 | 0.476 | 0.473 |
| | Mar | 0.467 | 0.469 | 0.466 | 0.468 |
| | Apr | 0.470 | 0.477 | 0.467 | 0.475 |
| | May | 0.470 | 0.475 | 0.474 | 0.471 |
| | Jun | 0.471 | 0.474 | 0.475 | 0.476 |
| | Jul | 0.472 | 0.475 | 0.474 | 0.476 |
| | Aug | 0.470 | 0.477 | 0.474 | 0.474 |
| | Sep | 0.474 | 0.477 | 0.478 | 0.475 |
| | Oct | 0.474 | 0.479 | 0.475 | 0.475 |
| | Nov | 0.471 | 0.476 | 0.476 | 0.476 |
| | Dec | 0.470 | 0.477 | 0.471 | 0.471 |

Figure 9.6: Monthly trends in **Arousal** of tweets across years (TUSC-Country).

| Dataset | | # tweets | # tweeters |
|---|---|---|---|
| **2020** | | | |
| | Canada | 3,038,530 | 20,887 |
| | USA | 2,641,694 | 19,709 |
| **2021** | | | |
| | Canada | 7,467,446 | 45,573 |
| | USA | 11,675,372 | 76,223 |

Table 9.4: Number of tweets and tweeters in the TUSC100 dataset.



|  |  | Year |  |  |  |
|---|---|---|---|---|---|
| Country | Month | 2015--2019 | 2019 | 2020 | 2021 |
| Canada | Jan | 0.569 | 0.580 | 0.572 | 0.576 |
|  | Feb | 0.569 | 0.575 | 0.577 | 0.575 |
|  | Mar | 0.571 | 0.579 | 0.575 | 0.578 |
|  | Apr | 0.571 | 0.576 | 0.569 | 0.576 |
|  | May | 0.573 | 0.580 | 0.571 | 0.577 |
|  | Jun | 0.575 | 0.583 | 0.578 | 0.576 |
|  | Jul | 0.569 | 0.576 | 0.571 | 0.576 |
|  | Aug | 0.569 | 0.574 | 0.574 | 0.577 |
|  | Sep | 0.569 | 0.574 | 0.574 | 0.582 |
|  | Oct | 0.568 | 0.578 | 0.578 | 0.579 |
|  | Nov | 0.573 | 0.577 | 0.571 | 0.581 |
|  | Dec | 0.570 | 0.573 | 0.575 | 0.579 |
| USA | Jan | 0.546 | 0.556 | 0.557 | 0.562 |
|  | Feb | 0.548 | 0.559 | 0.556 | 0.558 |
|  | Mar | 0.547 | 0.556 | 0.551 | 0.561 |
|  | Apr | 0.549 | 0.558 | 0.551 | 0.561 |
|  | May | 0.548 | 0.553 | 0.553 | 0.557 |
|  | Jun | 0.546 | 0.554 | 0.559 | 0.560 |
|  | Jul | 0.543 | 0.547 | 0.555 | 0.558 |
|  | Aug | 0.546 | 0.550 | 0.554 | 0.560 |
|  | Sep | 0.543 | 0.548 | 0.556 | 0.559 |
|  | Oct | 0.545 | 0.551 | 0.556 | 0.559 |
|  | Nov | 0.545 | 0.552 | 0.553 | 0.563 |
|  | Dec | 0.546 | 0.551 | 0.556 | 0.566 |

Figure 9.7: Monthly trends in **Dominance** of tweets across years (TUSC-Country).

| **Data** | **Year** | **Canada** | **USA** |
|---|---|---|---|
| *Mean* | 2020 | 0.6320 | 0.6132 |
| | 2021 | 0.6387 | 0.6257 |
| *Variability* | 2020 | 0.0708 | 0.0714 |
| | 2021 | 0.0700 | 0.0705 |
| *Rise Rate* | 2020 | 0.0121 | 0.0128 |
| | 2021 | 0.0117 | 0.0123 |
| *Recovery Rate* | 2020 | 0.0120 | 0.0127 |
| | 2021 | 0.0118 | 0.0123 |
| *Hm-Hi Rate* | 2020 | 0.0118 | 0.0129 |
| | 2021 | 0.0113 | 0.0121 |
| *Hi-Hm Rate* | 2020 | 0.0118 | 0.0129 |
| | 2021 | 0.0115 | 0.0122 |
| *Hm-Lo Rate* | 2020 | 0.0143 | 0.0149 |
| | 2021 | 0.0140 | 0.0145 |
| *Lo-Hm Rate* | 2020 | 0.0141 | 0.0148 |
| | 2021 | 0.0139 | 0.0144 |

Table 9.5: Tweet **Valence** dynamics metrics of tweeters in TUSC100. Averaged across all tweeters (not considering the cities they came from).

| **Data** | **Year** | **Canada** | **USA** |
|---|---|---|---|
| *Mean* | 2020 | 0.4828 | 0.4935 |
| | 2021 | 0.4854 | 0.4932 |
| *Variability* | 2020 | 0.0599 | 0.0593 |
| | 2021 | 0.0599 | 0.0595 |
| *Rise Rate* | 2020 | 0.0116 | 0.0120 |
| | 2021 | 0.0113 | 0.0117 |
| *Recovery Rate* | 2020 | 0.0115 | 0.0119 |
| | 2021 | 0.0113 | 0.0116 |
| *Hm-Hi Rate* | 2020 | 0.0129 | 0.0130 |
| | 2021 | 0.0125 | 0.0129 |
| *Hi-Hm Rate* | 2020 | 0.0127 | 0.0130 |
| | 2021 | 0.0126 | 0.0128 |
| *Hm-Lo Rate* | 2020 | 0.0121 | 0.0127 |
| | 2021 | 0.0118 | 0.0123 |
| *Lo-Hm Rate* | 2020 | 0.0120 | 0.0125 |
| | 2021 | 0.0117 | 0.0121 |

Table 9.6: **Arousal** dynamics metrics of tweeters in TUSC100. Averaged across all tweeters (not considering the cities they came from).

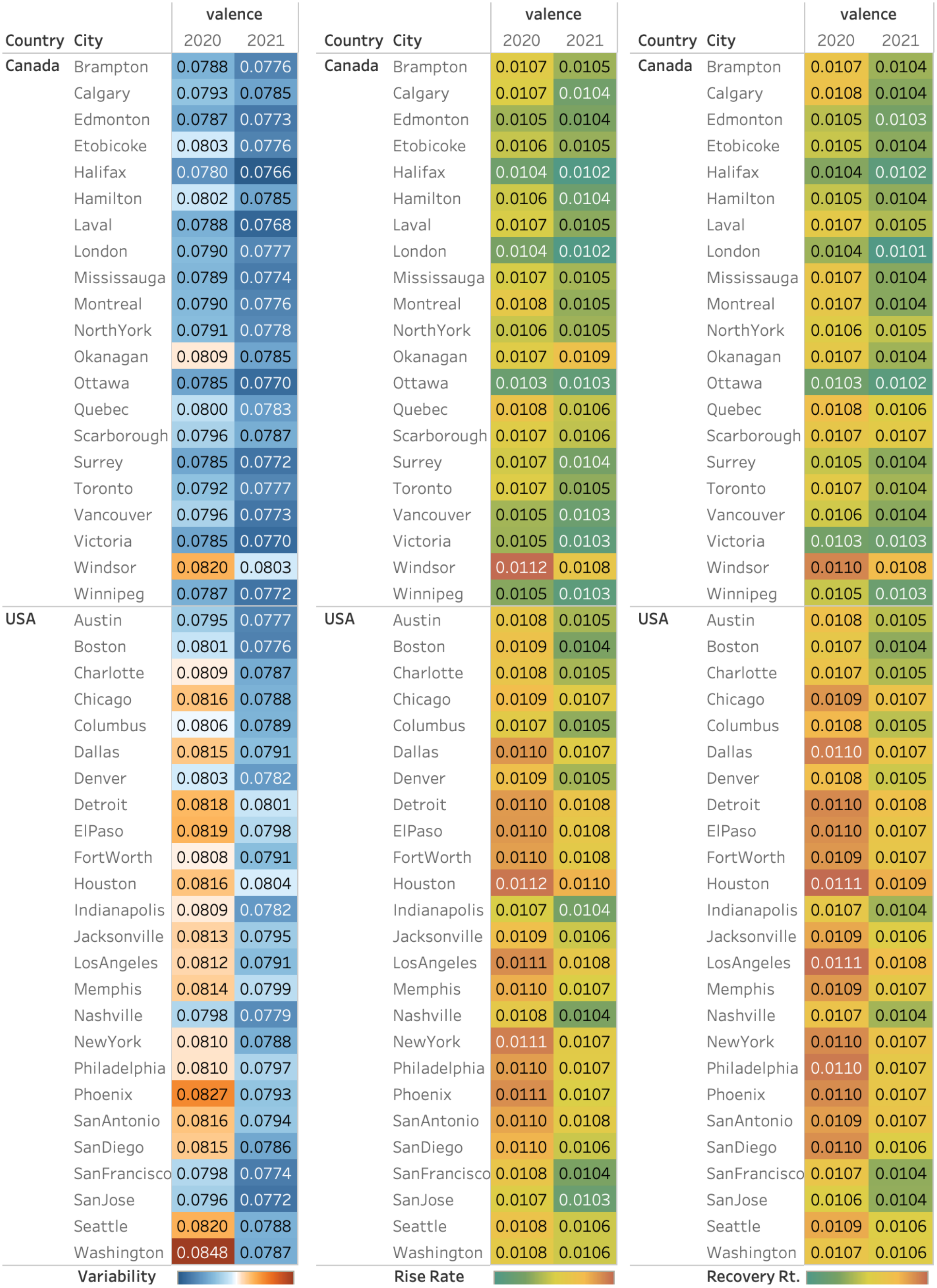
| **Data** | **Year** | **Canada** | **USA** |
|---|---|---|---|
| *Mean* | 2020 | 0.5821 | 0.5671 |
| | 2021 | 0.5873 | 0.5733 |
| *Variability* | 2020 | 0.0573 | 0.0556 |
| | 2021 | 0.0569 | 0.0557 |
| *Rise Rate* | 2020 | 0.0113 | 0.0116 |
| | 2021 | 0.0109 | 0.0113 |
| *Recovery Rate* | 2020 | 0.0112 | 0.0117 |
| | 2021 | 0.0109 | 0.0112 |
| *Hm-Hi Rate* | 2020 | 0.0114 | 0.0118 |
| | 2021 | 0.0111 | 0.0115 |
| *Hi-Hm Rate* | 2020 | 0.0114 | 0.0119 |
| | 2021 | 0.0111 | 0.0114 |
| *Hm-Lo Rate* | 2020 | 0.0127 | 0.0129 |
| | 2021 | 0.0124 | 0.0126 |
| *Lo-Hm Rate* | 2020 | 0.0126 | 0.0128 |
| | 2021 | 0.0123 | 0.0125 |

Table 9.7: Tweet **Dominance** Dynamics metrics of tweeters in TUSC100. Averaged across all tweeters (not considering the cities they came from).

| Country | City | valence |  | arousal |  | dominance |  |
|---|---|---|---|---|---|---|---|
|  |  | 2020 | 2021 | 2020 | 2021 | 2020 | 2021 |
| Canada | Brampton | 0.634 | 0.642 | 0.483 | 0.485 | 0.593 | 0.599 |
|  | Calgary | 0.636 | 0.647 | 0.483 | 0.480 | 0.599 | 0.604 |
|  | Edmonton | 0.643 | 0.649 | 0.478 | 0.480 | 0.594 | 0.598 |
|  | Etobicoke | 0.637 | 0.643 | 0.485 | 0.485 | 0.595 | 0.600 |
|  | Halifax | 0.649 | 0.655 | 0.472 | 0.473 | 0.591 | 0.595 |
|  | Hamilton | 0.644 | 0.651 | 0.482 | 0.481 | 0.599 | 0.605 |
|  | Laval | 0.638 | 0.645 | 0.488 | 0.488 | 0.593 | 0.599 |
|  | London | 0.649 | 0.656 | 0.477 | 0.477 | 0.597 | 0.604 |
|  | Mississauga | 0.635 | 0.643 | 0.485 | 0.483 | 0.594 | 0.601 |
|  | Montreal | 0.634 | 0.644 | 0.488 | 0.489 | 0.592 | 0.599 |
|  | NorthYork | 0.634 | 0.639 | 0.483 | 0.485 | 0.593 | 0.599 |
|  | Okanagan | 0.637 | 0.642 | 0.478 | 0.478 | 0.588 | 0.590 |
|  | Ottawa | 0.651 | 0.656 | 0.476 | 0.477 | 0.605 | 0.610 |
|  | Quebec | 0.638 | 0.647 | 0.494 | 0.492 | 0.592 | 0.598 |
|  | Scarborough | 0.632 | 0.633 | 0.485 | 0.487 | 0.594 | 0.599 |
|  | Surrey | 0.641 | 0.648 | 0.480 | 0.480 | 0.597 | 0.602 |
|  | Toronto | 0.634 | 0.640 | 0.484 | 0.485 | 0.594 | 0.600 |
|  | Vancouver | 0.638 | 0.645 | 0.480 | 0.480 | 0.595 | 0.601 |
|  | Victoria | 0.649 | 0.655 | 0.475 | 0.474 | 0.606 | 0.610 |
|  | Windsor | 0.613 | 0.623 | 0.497 | 0.496 | 0.576 | 0.579 |
|  | Winnipeg | 0.644 | 0.650 | 0.479 | 0.480 | 0.591 | 0.595 |
| USA | Austin | 0.627 | 0.639 | 0.489 | 0.489 | 0.587 | 0.594 |
|  | Boston | 0.625 | 0.641 | 0.488 | 0.486 | 0.592 | 0.599 |
|  | Charlotte | 0.627 | 0.638 | 0.490 | 0.493 | 0.585 | 0.590 |
|  | Chicago | 0.617 | 0.629 | 0.496 | 0.493 | 0.580 | 0.584 |
|  | Columbus | 0.625 | 0.637 | 0.492 | 0.491 | 0.583 | 0.589 |
|  | Dallas | 0.617 | 0.630 | 0.498 | 0.495 | 0.579 | 0.584 |
|  | Denver | 0.624 | 0.636 | 0.491 | 0.490 | 0.589 | 0.593 |
|  | Detroit | 0.614 | 0.626 | 0.497 | 0.495 | 0.577 | 0.581 |
|  | ElPaso | 0.629 | 0.638 | 0.502 | 0.504 | 0.570 | 0.578 |
|  | FortWorth | 0.620 | 0.630 | 0.496 | 0.496 | 0.577 | 0.583 |
|  | Houston | 0.613 | 0.620 | 0.498 | 0.499 | 0.570 | 0.573 |
|  | Indianapolis | 0.633 | 0.644 | 0.492 | 0.490 | 0.589 | 0.594 |
|  | Jacksonville | 0.626 | 0.638 | 0.492 | 0.492 | 0.590 | 0.597 |
|  | LosAngeles | 0.615 | 0.630 | 0.498 | 0.497 | 0.579 | 0.585 |
|  | Memphis | 0.621 | 0.630 | 0.493 | 0.494 | 0.578 | 0.582 |
|  | Nashville | 0.631 | 0.644 | 0.486 | 0.488 | 0.590 | 0.596 |
|  | NewYork | 0.616 | 0.630 | 0.494 | 0.493 | 0.584 | 0.591 |
|  | Philadelphia | 0.614 | 0.627 | 0.493 | 0.494 | 0.576 | 0.581 |
|  | Phoenix | 0.617 | 0.633 | 0.499 | 0.497 | 0.585 | 0.591 |
|  | SanAntonio | 0.624 | 0.634 | 0.496 | 0.499 | 0.571 | 0.579 |
|  | SanDiego | 0.620 | 0.636 | 0.496 | 0.495 | 0.584 | 0.592 |
|  | SanFrancisco | 0.626 | 0.640 | 0.484 | 0.484 | 0.593 | 0.601 |
|  | SanJose | 0.630 | 0.645 | 0.483 | 0.483 | 0.595 | 0.605 |
|  | Seattle | 0.628 | 0.639 | 0.485 | 0.483 | 0.591 | 0.596 |
|  | Washington | 0.626 | 0.635 | 0.491 | 0.487 | 0.597 | 0.603 |

Mean

Figure 9.8: TED: Tweet Valence Means (left), Arousal Means (centre), and Dominance Means (right) across American and Canadian cities in 2020 and 2021 (using tweets from TUSC-City).



| Country | City | valence | |
|---|---|---|---|
| | | 2020 | 2021 |
| Canada | Brampton | 0.0788 | 0.0776 |
| | Calgary | 0.0793 | 0.0785 |
| | Edmonton | 0.0787 | 0.0773 |
| | Etobicoke | 0.0803 | 0.0776 |
| | Halifax | 0.0780 | 0.0766 |
| | Hamilton | 0.0802 | 0.0785 |
| | Laval | 0.0788 | 0.0768 |
| | London | 0.0790 | 0.0777 |
| | Mississauga | 0.0789 | 0.0774 |
| | Montreal | 0.0790 | 0.0776 |
| | NorthYork | 0.0791 | 0.0778 |
| | Okanagan | 0.0809 | 0.0785 |
| | Ottawa | 0.0785 | 0.0770 |
| | Quebec | 0.0800 | 0.0783 |
| | Scarborough | 0.0796 | 0.0787 |
| | Surrey | 0.0785 | 0.0772 |
| | Toronto | 0.0792 | 0.0777 |
| | Vancouver | 0.0796 | 0.0773 |
| | Victoria | 0.0785 | 0.0770 |
| | Windsor | 0.0820 | 0.0803 |
| | Winnipeg | 0.0787 | 0.0772 |
| USA | Austin | 0.0795 | 0.0777 |
| | Boston | 0.0801 | 0.0776 |
| | Charlotte | 0.0809 | 0.0787 |
| | Chicago | 0.0816 | 0.0788 |
| | Columbus | 0.0806 | 0.0789 |
| | Dallas | 0.0815 | 0.0791 |
| | Denver | 0.0803 | 0.0782 |
| | Detroit | 0.0818 | 0.0801 |
| | ElPaso | 0.0819 | 0.0798 |
| | FortWorth | 0.0808 | 0.0791 |
| | Houston | 0.0816 | 0.0804 |
| | Indianapolis | 0.0809 | 0.0782 |
| | Jacksonville | 0.0813 | 0.0795 |
| | LosAngeles | 0.0812 | 0.0791 |
| | Memphis | 0.0814 | 0.0799 |
| | Nashville | 0.0798 | 0.0779 |
| | NewYork | 0.0810 | 0.0788 |
| | Philadelphia | 0.0810 | 0.0797 |
| | Phoenix | 0.0827 | 0.0793 |
| | SanAntonio | 0.0816 | 0.0794 |
| | SanDiego | 0.0815 | 0.0786 |
| | SanFrancisco | 0.0798 | 0.0774 |
| | SanJose | 0.0796 | 0.0772 |
| | Seattle | 0.0820 | 0.0788 |
| | Washington | 0.0848 | 0.0787 |

| Country | City | valence | |
|---|---|---|---|
| | | 2020 | 2021 |
| Canada | Brampton | 0.0107 | 0.0105 |
| | Calgary | 0.0107 | 0.0104 |
| | Edmonton | 0.0105 | 0.0104 |
| | Etobicoke | 0.0106 | 0.0105 |
| | Halifax | 0.0104 | 0.0102 |
| | Hamilton | 0.0106 | 0.0104 |
| | Laval | 0.0107 | 0.0105 |
| | London | 0.0104 | 0.0102 |
| | Mississauga | 0.0107 | 0.0105 |
| | Montreal | 0.0108 | 0.0105 |
| | NorthYork | 0.0106 | 0.0105 |
| | Okanagan | 0.0107 | 0.0109 |
| | Ottawa | 0.0103 | 0.0103 |
| | Quebec | 0.0108 | 0.0106 |
| | Scarborough | 0.0107 | 0.0106 |
| | Surrey | 0.0107 | 0.0104 |
| | Toronto | 0.0107 | 0.0105 |
| | Vancouver | 0.0105 | 0.0103 |
| | Victoria | 0.0105 | 0.0103 |
| | Windsor | 0.0112 | 0.0108 |
| | Winnipeg | 0.0105 | 0.0103 |
| USA | Austin | 0.0108 | 0.0105 |
| | Boston | 0.0109 | 0.0104 |
| | Charlotte | 0.0108 | 0.0105 |
| | Chicago | 0.0109 | 0.0107 |
| | Columbus | 0.0107 | 0.0105 |
| | Dallas | 0.0110 | 0.0107 |
| | Denver | 0.0109 | 0.0105 |
| | Detroit | 0.0110 | 0.0108 |
| | ElPaso | 0.0110 | 0.0108 |
| | FortWorth | 0.0110 | 0.0108 |
| | Houston | 0.0112 | 0.0110 |
| | Indianapolis | 0.0107 | 0.0104 |
| | Jacksonville | 0.0109 | 0.0106 |
| | LosAngeles | 0.0111 | 0.0108 |
| | Memphis | 0.0110 | 0.0107 |
| | Nashville | 0.0108 | 0.0104 |
| | NewYork | 0.0111 | 0.0107 |
| | Philadelphia | 0.0110 | 0.0107 |
| | Phoenix | 0.0111 | 0.0107 |
| | SanAntonio | 0.0110 | 0.0108 |
| | SanDiego | 0.0110 | 0.0106 |
| | SanFrancisco | 0.0108 | 0.0104 |
| | SanJose | 0.0107 | 0.0103 |
| | Seattle | 0.0108 | 0.0106 |
| | Washington | 0.0108 | 0.0106 |

| Country | City | valence | |
|---|---|---|---|
| | | 2020 | 2021 |
| Canada | Brampton | 0.0107 | 0.0104 |
| | Calgary | 0.0108 | 0.0104 |
| | Edmonton | 0.0105 | 0.0103 |
| | Etobicoke | 0.0105 | 0.0104 |
| | Halifax | 0.0104 | 0.0102 |
| | Hamilton | 0.0105 | 0.0104 |
| | Laval | 0.0107 | 0.0105 |
| | London | 0.0104 | 0.0101 |
| | Mississauga | 0.0107 | 0.0104 |
| | Montreal | 0.0107 | 0.0104 |
| | NorthYork | 0.0106 | 0.0105 |
| | Okanagan | 0.0107 | 0.0104 |
| | Ottawa | 0.0103 | 0.0102 |
| | Quebec | 0.0108 | 0.0106 |
| | Scarborough | 0.0107 | 0.0107 |
| | Surrey | 0.0105 | 0.0104 |
| | Toronto | 0.0107 | 0.0104 |
| | Vancouver | 0.0106 | 0.0104 |
| | Victoria | 0.0103 | 0.0103 |
| | Windsor | 0.0110 | 0.0108 |
| | Winnipeg | 0.0105 | 0.0103 |
| USA | Austin | 0.0108 | 0.0105 |
| | Boston | 0.0107 | 0.0104 |
| | Charlotte | 0.0107 | 0.0105 |
| | Chicago | 0.0109 | 0.0107 |
| | Columbus | 0.0108 | 0.0105 |
| | Dallas | 0.0110 | 0.0107 |
| | Denver | 0.0108 | 0.0105 |
| | Detroit | 0.0110 | 0.0108 |
| | ElPaso | 0.0110 | 0.0107 |
| | FortWorth | 0.0109 | 0.0107 |
| | Houston | 0.0111 | 0.0109 |
| | Indianapolis | 0.0107 | 0.0104 |
| | Jacksonville | 0.0109 | 0.0106 |
| | LosAngeles | 0.0111 | 0.0108 |
| | Memphis | 0.0109 | 0.0107 |
| | Nashville | 0.0107 | 0.0104 |
| | NewYork | 0.0110 | 0.0107 |
| | Philadelphia | 0.0110 | 0.0107 |
| | Phoenix | 0.0110 | 0.0107 |
| | SanAntonio | 0.0109 | 0.0107 |
| | SanDiego | 0.0110 | 0.0106 |
| | SanFrancisco | 0.0107 | 0.0104 |
| | SanJose | 0.0106 | 0.0104 |
| | Seattle | 0.0109 | 0.0106 |
| | Washington | 0.0107 | 0.0106 |

Figure 9.9: TED: Tweet Valence Variability (left), Rise Rate (centre), and Recovery Rate (right) across American and Canadian cities in 2020 and 2021 (using tweets from TUSC-City).

# Part IV

# Conclusion

# Chapter 10

# Conclusion

## 10.1 Summary

The work in this thesis has examined stylistic variation in language use from multiple computational perspectives. Language is a tool for self-expression, and it is no surprise that personal variations in language can offer insights into one's identity. This variation has largely been of interest to the NLP community only insofar as it is an impediment to NLP tools that rely on semantic functions — in order to be broadly useful and usable, models must be invariant to stylistic variation in the inputs. On the other hand, stylistic variation is of primary interest when studying language as a social mechanism: Why do people use the words that they use when expressing a thought? What external events can cause one's patterns of language use to change? How can we computationally capture such changes, and how can we characterize the dimensions of change at a more interpretable level, abstracting away from low-level linguistic or vectorized features? These are some of the interesting research questions that open up at the intersection of social science, linguistics, cultural analytics, and natural language processing.

In this thesis, we studied a specific subset of questions in this space. First, we investigated computational representation learning models that use neural networks and stylistically defined corpora to learn distinct representations of the *meaning* and *form* of a text. We demonstrated the difficulties of learning such representations in the absence of parallel paraphrase corpora, and their limited utility in low-data, long-text scenarios. In Chapter 3, we expanded further on what it means to say that two texts (limited again to sentences) are semantically close to one another, and we contributed a data resource that can be used to judge the closeness in meaning, or semantic relatedness, of sentence pairs.

We subsequently shifted towards studying language variation in a much more specific and defined set of domains, with a focus on quantifying diversity of linguistic features in texts uttered by different speakers. Our study of the differentiation of character voices within fictional texts in Part II is the first large-scale exploration of the stylistic diversity present within a single novel. We see that the domain of fiction is highly varied in its stylistic diversity: character voices, while they are made distinct, exhibit their distinctiveness in different ways across novels and authors. We studied lexical features of style and emotion in further detail, and showed that the Utterance Emotion Dynamics framework can be used to quantify multiple aspects of literary style, in particular the relationship of the narrative voice with that of its characters.

Part III again applied the Utterance Emotion Dynamics framework to a dataset of demographically varying speaker utterances, this time on tweets by people located in different cities in North America over the course of multiple years. This study's focal point is the socio-cultural differences that influence variation in the emotional expressivity of uttered language, and the effect of real-world events on such patterns. We observed differences in UED metrics for the populations in the USA and Canada, as well as significant inter-city variations, differences that persisted across multiple years. The impact of the COVID-19 pandemic was quantitatively shown to have lowered the positivity of people's social media outputs, along with an increase in the expression of negative feelings.

The affective intensity of words along the three primary dimensions of valence, arousal, and dominance can be particularly interesting to study as a function of the mental, physical, and emotional well-being of people and populations. An ongoing line of work extending the work from Part III is to develop frameworks and metrics to quantify emotional expression that are aligned with research in psychology and affective sciences on emotion and well-being — projects that are being carried out in collaboration with experts in these fields.

## 10.2 Future Directions

In this section, I list some concrete research directions that emerge as natural follow-ups to the work presented in this thesis.

### 10.2.1 Large-scale Literary Analysis of Character Voices

A natural follow-up to our work on character voice is its application to a much larger domain of texts. The novels represented in PDNC are critically acclaimed exemplars of their era, and thus can hardly be said to be *representative* of the bulk of the work published alongside them. Also, the socio-demographic makeup of the authors is quite restricted to a narrow range of countries, cultures, and economic classes. What would the distribution of gendered dialogue look like when considering all the published works of the 19th and 20th centuries? How distinct is the stylistic differentiation in this pool of texts? Are there consistent patterns over the decades? These are all questions of literary interest that can be answered by the application of the NLP frameworks presented in this thesis, though the technical challenges involved in processing such an uncurated corpus are many: character dialogue itself, for starters, is not always guaranteed to be marked out in neat quotation marks; character identification and coreference resolution will remain imperfect in the face of the diversity of narrative techniques and invention that mark literature. Nevertheless, in the spirit of distant reading, one can aim to cover a large proportion of the existing literature, and extract the majority of characters and attribute their utterances with a reasonable level of accuracy for analysis. (Tools like BookNLP, for example, demonstrate attribution accuracies averaging on 80%, as we have seen in Chapter 8.)

A promising step here is the ever-expanding repertoire of Large Language Models (LLMs), which are now able to ingest book-length inputs and execute functions ranging from providing general-purpose summaries to answering detailed questions about their specifics. One can imagine these models being re-purposed for novel-level document processing, handling the functionalities of character identification and speaker attribution without needing to explicitly build long-range coreference chains or disambiguate various character names. While higher-level analyses like modeling stylistic distinctiveness and mapping a character's emotional trajectories may not be processes that can be automated, LLMs can serve as a drop-in replacement for the intermediate computational models we build to extract the relevant information from these texts.

The outcomes of such large-scale studies of character voice and emotional arcs have the potential to define and refine several literary meta-variables of interest, like those of genre (a dark crime thriller or a wholesome group adventure?) and narrative style (an equidistant third-person narrator or an intimate, stream-of-consciousness character portrait?). These are categories that have no easy computational definitions yet, and often no easy theoretical boundaries either; but perhaps a data-centric grouping of the kinds of quantitative metrics we have explored in this thesis will yield a revelatory clustering of texts (along with pointing out the fluidity of their boundaries).

### 10.2.2 Developing the Character Persona

We have also not yet explored the full scope of character portrayal in novels. Stylistic voice and emotional arcs are but two aspects of the myriad ways in which characters are established as distinctive, memorable personas in fictional tales. A key aspect of Dialogism in the novel is the idea that authors use characters to represent distinct *worldviews and perspectives*, either different from or representative of their own thoughts and ideals. Such characters can also be used to chronicle a worldview that goes against the prevalent social customs of the time, a function of literature as a safe place to imagine, develop, and explain alternative human societies, non-normative lifestyles and theories, both humanistic and scientific (Johannes Kepler, a key figure of the scientific revolution of the 17th century, presented the controversial and, at the time, dangerous, Copernican theory of the heliocentric model of the solar system in a science-fiction novel that imagined the view of the Earth from the moon (Wikipedia contributors, 2023)).

Establishing these facets of characters requires reasoning beyond stylistic aspects of utterances. We want to identify their *stances* on certain *topics* — a task that involves several interesting computational challenges. What are the socially-relevant topics that are introduced or discussed in a novel? One can imagine that Jane Austen's novels, for example, prominently feature discussions of "marriage" and the expectations of gender and social class contained within it.

How can we identify a character's stance towards such topics? Yes, their utterances provide a key window into their opinions. But often, these opinions are not explicitly stated. Never does Elizabeth Bennet explicitly say "I do not believe that one's socio-economic status is a representation of one's character, nor that it determines the compatibility of two people for a marriage". We can *infer* these beliefs of hers from the things she says and the things she does in reaction to different situations — that she is bold, intelligent, independent, witty, and does not conform to societal norms and expectations. Breaking this inference down into neat little computational modules presents a novel and stimulating research direction in computational literary analysis. One can also imagine such character portraits requiring an external knowledge of the social norms of the world as well — the time period and sub-culture that the characters and the author lived in.

### 10.2.3 Emotional Spaces and Persona

The above aspects of character persona are not limited in their scope to fictional characters. With any individual, real or otherwise, we glean insights into various facets of their personality through the things they say, i.e., their utterances. Social media data is a wealth of such opinionated expression, and NLP methods have been extensively applied to these domains, characterizing individuals and communities via their linguistic preferences, stance-taking, and interaction dynamics. The work in this thesis on using emotional expression in tweets as an aspect of individual personality is a component of such sociolinguistic research that also integrates theories of psychology on how humans express and regulate their emotions, and the variation in these mechanisms across individuals, populations, and cultures.

The characterization of expressed emotion in the affective VAD space offers interesting possibilities. These three dimensions capture aspects of opinion that are key to understanding personal and societal reactions to ideas: a positive valence and positive dominance indicates that a particular concept is perceived as both friendly and competitive or effective. A negative valence, combined with positive dominance, indicates a perception of something being *dangerous*, unfriendly but also deadly, powerful — as opposed to a low dominance, indicating something is bad but also *harmless*. These dimensions of perception, formulated along the two axes of *warmth* and *competence*, have been studied under the Stereotype Content Model in psychology (Cuddy et al., 2008), and are an ideal framework to study aspects of opinion and stance with social media data. How were the COVID-19 vaccines perceived by different groups (cities, countries, demographic groups) following their release? How are immigrants perceived, or alternatively, portrayed, by members of different political groups? How do these perceptions *change over time*, i.e., what are their emotional dynamics? Did the peak of the "dangerous and deadly" sentiment towards immigrants occur during the charged 2016 presidential elections in the United States of America? The framework of Utterance Emotion Dynamics offers us a powerful set of metrics to study such social mechanisms of change. One can go even further with NLP and social network methods, perhaps looking at the most-influential terms or posts that change the course of popular perception towards a particular topic.

Scaling back down to the level of the individual, developing robust computational measures of the many metrics studied in the affective sciences and psychology as barometers of mental, physical, and emotional health is a fertile research topic in itself. I introduced *emotion granularity* as one such concept in the concluding appendix of the preceding chapter of this thesis, which postulates that the degrees of distinction made by an individual in the emotional states they experience can vary from person to person, and has downstream connections to one's well-being. There is a vast amount of theoretical and empirical research in these fields on associated metrics of emotional awareness, expression, and dynamics, as well as their effectiveness as indicators of mental and physical health. Emotion granularity itself, for example, is posited to show a non-linear relationship with age — it decreases from childhood to early adolescence, and increases towards adulthood. The onset of declining physical health towards later years, however, can again correspond to a decrease in emotional granularity. While the link between uttered language and these internal mechanics of emotions can seem tenuous, the function of social media platforms as a place for free-form expression of thoughts, and the availability of this data for computational analysis, is an opportunity for research into these mechanisms. Linguistic measures of emotional dynamics can also offer an alternative way of patient monitoring and diagnosis, alleviating some of the issues with self-reports — provided, of course, that they are conducted in a safe, confidential environment and only serve as an additional aid to human expertise.

## 10.3 Conclusion

The unifying thread behind much of the work in this thesis has been understanding personal variation in language use. A purely computational approach to this problem is to maximize the predictive accuracy of utterance attribution or the generative accuracy of language modelling. Language, however, is a social tool; one uses it to express identity in the context of personal and inter-personal relationships. It is of greater interest to therefore study language variation in the context of the social mechanisms that mould us, as much of sociolinguistics has done: the context of the interaction, other participants, and individual personalities. These qualitative aspects of studying language have somewhat receded in their importance in NLP when compared to the computational aspects; however, I believe advances in the latter can also function as more powerful tools for the former. They can serve us in creating controlled datasets, efficiently extracting attributes of interests from raw data, and making predictions (of aspects like emotion or stance) that leverage global information and contexts beyond simply that of the input. This is rightfully a very exciting time to be exploring aspects of language use that are of interest to those in the social sciences and the humanities.

# Bibliography

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. SemEval-2015 Task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-lingual Evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43.

Malik Altakrori, Jackie Chi Kit Cheung, and Benjamin CM Fung. 2021. The topic confusion task: A novel evaluation scenario for authorship attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4242–4256.

Shlomo Engelson Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A Large, Fine-Grained, Bigram Relatedness Dataset For Examining Semantic Composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 505–516.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Iris Bakker, Theo van der Voordt, Peter Vink, and Jan Boon. 2014. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33:405 – 421.

Vikash Balasubramanian, Ivan Kobyzev, Hareesh Bahuleyan, Ilya Shapiro, and Olga Vechtomova. 2020. Polarized-VAE: Proximity based disentangled representation learning for text generation. *arXiv preprint arXiv:2004.10809*.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2012. Gender identity and lexical variation in social media. *arXiv: Computation and Language*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54.

David Bamman, Ted Underwood, and Noah A Smith. 2014a. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting* *of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379.

David Bamman, Ted Underwood, and Noah A. Smith. 2014b. A bayesian mixed effects model of literary character. In *Annual Meeting of the Association for Computational Linguistics*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for Sembanking. In *LAW@ACL*.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A large-scale COVID-19 twitter chatter dataset for open scientific research - an international collaboration. *CoRR*, abs/2004.03688.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019.

Lisa Feldman Barrett, James Gross, Tamlin Conner Christensen, and Michael Benvenuto. 2001. Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition & Emotion*, 15(6):713–724.

Janek Bevendorff, Berta Chulvi, Elisabetta Fersini, Annina Heini, Mike Kestemont, Krzysztof Kredens, Maximilian Mayerl, Reynier Ortega-Bueno, Piotr Pęzik, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, and Eva Zangerle. 2022. Overview of PAN 2022: Authorship verification, profiling irony and stereotype spreaders, and style change detection. In *Conference and Labs of the Evaluation Forum*.

Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023. Sentimental matters - predicting literary quality by sentiment analysis and stylometric features. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 11–18, Toronto, Canada. Association for Computational Linguistics.

David M. Blei, A. Ng, and Michael I. Jordan. 2009. Latent Dirichlet allocation.

Victoria Bobicev, Victoria Maxim, Tatiana Prodan, Natalia Burciu, and Victoria Angheluş. 2010. Emotions in words: Developing a multilingual WordNet-Affect. In *Conference on Intelligent Text Processing and Computational Linguistics*.

Margaret M. Bradley and Peter J. Lang. 1999. Affective Norms for English Words (ANEW): Instruction manual and affective ratings.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2016. Using models of lexical style to quantify free indirect discourse in modernist fiction. *Digital Scholarship in the Humanities*, 32:234–250.

Julian Brooke and Graeme Hirst. 2013a. A multi-dimensional Bayesian approach to lexical style. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 673–679.

Julian Brooke and Graeme Hirst. 2013b. A multi-dimensional bayesian approach to lexical style. In *North American Chapter of the Association for Computational Linguistics*.

Julian Brooke, Graeme Hirst, and Adam Hammond. 2013. Clustering voices in The Waste Land. In *CLfL@NAACL-HLT*.

John F Burrows. 1989. 'an ocean where each kind...': Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23:309–321.

John F. Burrows. 2002. 'Delta': a measure of stylistic difference and a guide to likely authorship. *Lit. Linguistic Comput.*, 17:267–287.

Yang Trista Cao and Hal III Daumé. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the Bible. *Royal Society Open Science*, 5(10):171920.

Bob Carpenter. 1997. Type-logical semantics.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daumé, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *AAAI Conference on Artificial Intelligence*.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2453–2464.

Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. 2020. Improving disentangled text representation learning with information-theoretic guidance. *arXiv preprint arXiv:2006.00693*.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34:597–614.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Ann A. Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3:281–332.

Lee J Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16(3):297–334.

Amy J.C. Cuddy, Susan Tufts Fiske, and Peter Glick. 2008. *Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map*, Advances in Experimental Social Psychology, pages 61–149.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *CMCL@ACL*.

Herbert Aron David. 1963. The Method of Paired Comparisons. In *Proceedings of the Fifth Conference on the Design of Experiments in Army Research Developments and Testing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The Automatic Content Extraction (ACE) Program —– tasks, data, and evaluation. In *Language Resources and Evaluation Conference*, volume 2, pages 837–840. Lisbon.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one*, 6(12):e26752.

William B Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Bruce Doré, Leonard Ort, Ofir Braverman, and Kevin N. Ochsner. 2015. Sadness shifts to anxiety over time and distance from the national tragedy in Newtown, Connecticut. *Psychological Science*, 26(4):363–373. PMID: 25767209.

Penelope Eckert. 1989. The whole woman: Sex and gender differences in variation. *Language Variation and Change*, 1:245 – 267.

Penelope Eckert. 2000. Linguistic variation as social practice.

Maciej Eder. 2015. Does size matter? authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2):167–182.

Jacob Eisenstein, Brendan T. O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Conference on Empirical Methods in Natural Language Processing*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

David K. Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Annual Meeting of the Association for Computational Linguistics*.

David K Elson and Kathleen R McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Doris R. Entwisle and William Labov. 1975. Language in the inner city: Studies in the Black English vernacular.@@@sociolinguistic patterns. *Social Forces*, 53:658.

Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. *ArXiv*, abs/1603.08832.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104.

Terry N Flynn and Anthony AJ Marley. 2014. Best-Worst Scaling: Theory and Methods. In *Handbook of Choice Modelling*. Edward Elgar Publishing.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Mikaela Irene D. Fudolig, T. Alshaabi, Kathryn Cramer, Christopher M. Danforth, and Peter Sheridan Dodds. 2022. A decomposition of book structure through ousiometric fluctuations in cumulative word-time. *Humanities and Social Sciences Communications*, 10:1–12.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A

Smith. 2010. Part-of-speech tagging for Twitter: Annotation, features, and experiments. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Jade Goldstein-Stewart, Kerri A. Goodwin, Roberta Evans Sabin, and Ransom K. Winder. 2008. Creating and using a correlated corpus to glean communicative commonalities. In *International Conference on Language Resources and Evaluation*.

Édouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman Group.

Adam Hammond. 2017. The double bind of validation: distant reading and the digital humanities' "trough of disillusionment". *Literature Compass*, 14(8):e12402. E12402 LICO-0881.R1.

Adam Hammond, Krishnapriya Vishnubhotla, and Graeme Hirst. 2020. The words themselves: A content-based approach to quote attribution. In *Proceedings of the Digital Humanities 2020 Conference*.

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. 2015. *Semantic Similarity from Natural Language and Ontology Analysis*. Morgan & Claypool Publishers.

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn Walker. 2019. Maximizing stylistic control and semantic accuracy in NLG: Personality variation and discourse contrast. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 1–12.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017*.

Will Hipson and Saif M. Mohammad. 2020. PoKi: A large dataset of poems by children. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1578–1589, Marseille, France. European Language Resources Association.

Will E. Hipson and Saif M. Mohammad. 2021. Emotion dynamics in movie dialogues. *PLOS ONE*, 16:1–19.

Graeme Hirst and Olga Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Lit. Linguistic Comput.*, 22:405–417.

Tom Hollenstein. 2015. This time, it's real: Affective flexibility, time scales, feedback loops, and the regulation of emotion. *Emotion Review*, 7:308 – 315.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. Python Package; spaCy: Industrial-strength Natural Language Processing in Python.

David Lowell Hoover. 2017. The microanalysis of style variation. *Digit. Scholarsh. Humanit.*, 32:ii17–ii30.

Eduard H. Hovy and Rahul Bhagat. 2009. Learning paraphrases from text.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan L. Boyd-Graber, and Hal Daumé. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *North American Chapter of the Association for Computational Linguistics*.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Barbara Johnstone. 2002. Language and place.

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1446–1459, New Orleans, Louisiana. Association for Computational Linguistics.

Patrick Juola. 2012. An overview of the traditional authorship attribution subtask. In *Conference and Labs of the Evaluation Forum*.

David Jurgens, Agrima Seth, Jack E. Sargent, Athena Aghighi, and Michael Geraci. 2023. Your spouse needs professional help: Determining the contextual appropriateness of messages through modeling social relationships. In *Annual Meeting of the Association for Computational Linguistics*.

P. S. Keila and David B. Skillicorn. 2005. Structure in the Enron email dataset. *Computational & Mathematical Organization Theory*, 11:183–199.

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017a. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics.

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017b. Investigating the relationship between literary genres and emotional plot development. In *LaTeCH@ACL*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and Best–Worst scaling. In *Proceedings of*

*The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California.

G Frederic Kuder and Marion W Richardson. 1937. The Theory of the Estimation of Test Reliability. *Psychometrika*, 2(3):151–160.

Peter Kuppens and Philippe Verduyn. 2017. Emotion dynamics. *Current Opinion in Psychology*, 17:22–26. Emotion.

William Labov. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, 2:205 – 254.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations (ICLR 2019)*.

Mark E Larsen, Tjeerd W Boonstra, Philip J Batterham, Bridianne O'Dea, Cecile Paris, and Helen Christensen. 2015. We feel: mapping emotion on Twitter. *IEEE journal of biomedical and health informatics*, 19(4):1246–1252.

Giada Lettieri, Giacomo Handjaras, Erika Bucci, Pietro Pietrini, and Luca Cecchetti. 2023. How male and female literary authors write about affect across cultures and over historical periods. *Affective Science*.

Yuhua Li, David McLean, Zuhair A Bandar, James D O'shea, and Keeley Crockett. 2006. Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150.

Justin Littman, Laura Wrubel, and Daniel Kerchner. 2016. 2016 United States Presidential Election Tweet Ids.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Jordan J Louviere and George G Woodworth. 1991. Best-Worst Scaling: A Model For The Largest Difference Judgments. Technical report, Working paper.

May Oo Lwin, Jiahui Lu, Anita Sheldenkar, Peter Johannes Schulz, Wonsun Shin, Raj Gupta, and Yinping Yang. 2020. Global sentiments surrounding the COVID-19 pandemic on Twitter: Analysis of Twitter trends. *JMIR Public Health Surveill*, 6(2):e19447.

Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard H. Hovy, Barnab'as P'oczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. StylePTB: A compositional benchmark for fine-grained controllable text style transfer. *ArXiv*, abs/2104.05196.

Aman Madaan, Amrith Rajagopal Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Annual Meeting of the Association for Computational Linguistics*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2014)*, pages 216–223. Reykjavik.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

George A Miller and Walter G Charles. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1):1–28.

Saif Mohammad. 2008. *Measuring Semantic Distance Using Distributional Profiles of Concepts*. Ph.D. thesis, University of Toronto.

Saif M. Mohammad. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decis. Support Syst.*, 53:730–741.

Saif M. Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Saif M. Mohammad. 2018b. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Saif M. Mohammad. 2020. Practical and ethical considerations in the effective use of emotion and sentiment lexicons. *arXiv:2011.03492*.

Saif M. Mohammad. 2021. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In Herbert L. Meiselman, editor, *Emotion Measurement (Second Edition)*, second edition edition, pages 323–379. Woodhead Publishing.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *To Appear in Computational Linguistics*.

Saif M Mohammad and Graeme Hirst. 2012. Distributional Measures of Semantic Distance: A Survey. *arXiv preprint arXiv:1203.1858*.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, LA, California.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Pascale Moreira, Yuri Bizzoni, Kristoffer Nielbo, Ida Marie Lassen, and Mads Thomsen. 2023. Modeling readers' appreciation of literary narratives through sentiment arcs and semantic profiles. In *Proceedings of the 5th Workshop on Narrative Understanding*, pages 25–35, Toronto, Canada. Association for Computational Linguistics.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

Grace Muzny, Mark Andrew Algee-Hewitt, and Dan Jurafsky. 2017a. Dialogism in the novel: A computational model of the dialogic nature of narration and quotations. *Digit. Scholarsh. Humanit.*, 32:ii31–ii52.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017b. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the*

*European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470.

Eric T. Nalisnick and Henry S. Baird. 2013. Character-to-character sentiment analysis in Shakespeare's plays. In *Annual Meeting of the Association for Computational Linguistics*.

Emily Ohman, Yuri Bizzoni, Pascale Feldkamp Moreira, and Kristoffer Nielbo. 2024. EmotionArcs: Emotion arcs for 9,000 literary texts. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 51–66, St. Julians, Malta. Association for Computational Linguistics.

Emily Öhman and Riikka Rossi. 2023. Affect as a proxy for literary mood. *Journal of Data Mining & Digital Humanities*.

Shereen Oraby, Lena Reed, Shubhangi Tandon, TS Sharath, Stephanie Lukin, and Marilyn Walker. 2018. Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 180–190.

Bryan Orme. 2009. MaxDiff Analysis: Simple Counting, Individual-Level Logit, and Hb. *Sawtooth Software*.

Charles E Osgood, George J Suci, and Percy Tannenbaum. 1957. *The measurement of meaning*. University of Illinois Press.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. 1958. The measurement of meaning.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 380–390.

Tim O'Keefe, Silvia Pareti, James R Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Martha Palmer, Paul R. Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:71–106.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876.

Stanley Presser and Howard Schuman. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. SAGE Publications, Inc.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A Word at a Time: Computing Word Relatedness Using Temporal Semantic Analysis. In *Proceedings of the 20th International Conference on World Wide Web*, pages 337–346.

Anil Ramakrishna, Victor R. Martinez, Nikos Malandrakis, Karan Singla, and Shrikanth S. Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *Annual Meeting of the Association for Computational Linguistics*.

Sudha Rao and Joel Tetreault. 2018a. Dear Sir or Madam, May I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Sudha Rao and Joel R. Tetreault. 2018b. Dear Sir or Madam, May I Introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *North American Chapter of the Association for Computational Linguistics*.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. Adversarial decomposition of text representation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 815–825.

Herbert Rubenstein and John B Goodenough. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Maarten Sap, Marcella Cindy Prasetio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *EMNLP*.

Karin Kipper Schuler and Martha Palmer. 2005. Verbnet: a broad-coverage, comprehensive verb lexicon.

Tianxiao Shen, Jonas Mueller, Regina Barzilay, and Tommi Jaakkola. 2020. Educating text autoencoders: Latent representation guidance via denoising. In *International Conference on Machine Learning*, pages 8719–8729. PMLR.

Matthew Sims and David Bamman. 2020. Measuring information propagation in literary social networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 642–652.

Bryor Snefjella, Daniel Schmidtke, and Victor Kuperman. 2018. National character stereotypes mirror language use: A study of Canadian and American tweets. *PloS one*, 13(11):e0206188.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Assoc. Inf. Sci. Technol.*, 60:538–556.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character N -gram features. *Journal of law and policy*, 21:7.

Efstathios Stamatatos, Nikos Fakotakis, and George K. Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26:471–495.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24 – 54.

Daniel Teodorescu and Saif M. Mohammad. 2022. Frustratingly easy sentiment analysis of text streams: Generating high-quality emotion arcs using emotion lexicons. *ArXiv*, abs/2210.07381.

Louis L Thurstone. 1927. A Law of Comparative Judgment. *Psychological Review*, 34(4):273.

Ted Underwood. 2019. *Distant horizons: digital evidence and literary change*. University of Chicago Press.

Hardik Vala, Stefan Dimitrov, David Jurgens, Andrew Piper, and Derek Ruths. 2016. Annotating characters in literary corpora: A scheme, the CHARLES tool, and an annotated novel. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 184–189.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2019. Are fictional voices distinguishable? Classifying character voices in modern drama. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–34.

Peter de Vocht. 2020. Python Package: Subject Verb Object Extractor. Github.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *International Conference on Language Resources and Evaluation*.

Kurt Vonnegut. 2009. *Palm Sunday: An Autobiographical Collage*. Random House Publishing Group.

Bin Wang, C-C Jay Kuo, and Haizhou Li. 2022. Just rank: Rethinking evaluation with word and sentence similarities. *arXiv preprint arXiv:2203.02679*.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45:1191 – 1207.

Uriel Weinreich, William Labov, and Marvin Herzog. 1968. *Empirical foundations for a theory of language change*. University of Texas Press.

Wikipedia contributors. 2023. Somnium (novel) — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Somnium_(novel)&oldid=1177279891`. [Online; accessed 19-December-2023].

Huimin Xu, Zhang Zhang, Lingfei Wu, and Cheng-Jun Wang. 2019. The Cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PloS one*, 14(11):e0225385.

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. 2020. On variational learning of controllable representations for text without supervision. In *International Conference on Machine Learning*, pages 10534–10543. PMLR.

Ivan P Yamshchikov, Viacheslav Shibaev, Aleksander Nagaev, Jürgen Jost, and Alexey Tikhonov. 2019. Decomposing textual information for style transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 128–137.

Luke Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *ArXiv*, abs/1207.1420.

Jianing Zhou and S. Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Conference on Empirical Methods in Natural Language Processing*.