Demographic-Aware Natural Language Processing

by

Aparna Garimella

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2020

Doctoral Committee:

Professor Rada Mihalcea, Chair
Professor Emily M. Provost
Professor James Pennebaker
Professor Kevyn-Collins Thompson

Aparna Garimella
gaparna@umich.edu
ORCID iD: 0000-0003-3111-0686

DEDICATION

This dissertation is in honor of my father Dr. Venkateswarlu Garimella and mother Dr. Lakshmi VVS, who give education utmost importance. They have always encouraged me to pursue my passions, even when the society then looked differently at the education of girls.

ACKNOWLEDGMENTS

This thesis has been a collective endeavor of many wonderful people, without whose love and support I can not imagine finishing my PhD.

Firstly, I would like to thank my advisor Prof. Rada Mihalcea, who has been my biggest source of support through the last few years. By no means have I been an easy graduate student to advise, but she took it in her stride, and continuously pushed me when required, stepped back when needed, and most importantly, never gave up on me. In addition to learning from her about the art of doing research, I am deeply inspired by her contributions to many initiatives beyond research, while still spending quality time with her family.

A close second is Carmen Banea, who has been a constant source of strength through my PhD years. Conversations with her have challenged me to dig deeper while still being focused on what matters the most. I am thankful to her for always being by my side, and bearing with me through all our late night submissions :)

I am grateful to have had an amazing set of committee members from very diverse backgrounds – Emily Provost, Kevin Thompson, and Jamie Pennebaker. Their feedback has been invaluable, has made this dissertation much better, and gave me new research directions to pursue in the future.

I am blessed to have been able to collaborate with and learn from amazing people through my PhD years. I would like to thank all my collaborators: Dirk Hovy, James Pennebaker, Laura Chiticariu, Yunyao Li, Svitlana Volkova, and Nabil Hossain. Their invaluable insights made me approach research through an improved lens. To the whole LIT group – I could not have asked for a kinder, smarter, and more accepting lab to be a part of.

To the amazing friends I have had the pleasure and privilege of knowing – Aniket Deshmukh, Abhishek Bafna, Niket Prakash, Nikita Bhutani, thank you for putting up with me! To the lovely Sneha Agarwal, I could not have asked for a kinder and more fun person for a flatmate. Thanks a ton for dealing with my mood swings, and creating a laughter even in the most serious situations. To my best half Kedar Kulkarni, for picking me up from my deepest lows, and constantly striving to bring a smile on my face – thank you! And last but most definitely not the least, to my parents and my dear sister, words would not do justice to acknowledge your unconditional love, unwavering support and countless silent sacrifices. Everything I am today is because of you. I will do my best to excel in, and contribute to others, in everything that I do.

This material is based in part upon work supported by the Michigan Institute for Data Science, by the National Science Foundation (grants #095688 and #074251), and by the John Templeton Foundation (grants #352839 and #391616).

PREFACE

Language is not just the interaction between words, but is much more than that – it is the interaction between people. The thesis has stemmed from the shared passion of my advisor and I to have a better and more nuanced understanding and representation of people through their language use. With the increasing amount of content on digital platforms, a new direction has been created for the current day natural language processing systems. With the degree to which people share their thoughts and opinions on social media, it is only fitting that the content they read also caters to their backgrounds and personalities. I have had several inspirations in this respect in the form of interesting interactions with researchers from diverse backgrounds, only to understand that language and culture have significant effects on each other. It is my vision to have personalized systems to perform various language processing tasks for people, while also understanding and addressing the subtleties and potential issues of such developments.

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The underlying traits of our demographic group affect and shape our thoughts, and therefore surface in the way we express ourselves and employ language in our day-to-day life. Understanding and analyzing language use in people from different demographic backgrounds help uncover their demographic particularities. Conversely, leveraging these differences could lead to the development of better language representations, thus enabling further demographic-focused refinements in natural language processing (NLP) tasks. In this thesis, I employ methods rooted in computational linguistics to better understand various demographic groups through their language use. The thesis makes two main contributions.

First, it provides empirical evidence that words are indeed used differently by different demographic groups in naturally occurring text. Through experiments conducted on large datasets which display usage scenarios for hundreds of frequent words, I show that automatic classification methods can be effective in distinguishing between word usages of different demographic groups. I compare the encoding ability of the utilized features by conducting feature analyses, and shed light on how various attributes contribute to highlighting the differences.

Second, the thesis explores whether demographic differences in word usage by different groups can inform the development of more refined approaches to NLP tasks. Specifically, I start by investigating the task of word association prediction. The thesis shows that going beyond the traditional "one-size-fits-all" approach, demographic-aware models achieve better performances in predicting word associations for different demographic groups than generic ones. Next, I investigate the impact of demographic information on part-of-speech tagging and syntactic parsing, and the experiments reveal numerous part-of-speech tags and syntactic relations, whose predictions benefit from the prevalence of a specific group in the training data. Finally, I explore demographic-specific humor generation, and develop a humor generation framework to fill-in the blanks to generate funny stories, while taking into account people's demographic backgrounds.

CHAPTER 1

Introduction

> “To be a member of a group is to think and act in a certain way, in the light of particular goals, values, pictures of the world; and to think and act so is to belong to a group [1].”
>
> Richard A. Shweder

Demographics of people affect and shape their thoughts, beliefs, behaviors and actions, thus manifesting in their language use in day-to-day life. People are contributing self-authored online content that they disseminate via various social media platforms, such as Facebook, Twitter, and Google Blogger, at an increasing rate. In addition to content describing their lifestyles and interests, these writings often directly contain demographic information of the authors. Hence, these writings have inspired researchers in the fields of linguistics, sociology and psychology, to study the relationship between language and demographics to better understand people and their worldview [2, 3, 4, 5, 6].

This thesis finds inspiration in a line of research in psychology that poses that people from different demographic backgrounds and/or speaking different languages perceive the world around them differently, with these differences being reflected in their perception of time and space [7, 8], body shapes [9], or surrounding objects [10]. As an example, consider the study described in [10], which shows how the perception of objects in different languages can be affected by their genders in the respective languages. For instance, one of the words used in their study is the word ”bridge,” which is masculine in Spanish and feminine in German. When asked about the descriptive properties of a bridge, Spanish speakers described bridges as being *big*, *dangerous*, *long*, *strong*, *sturdy* and *towering*, while German speakers said they are *beautiful*, *elegant*,*fragile*, *peaceful*, *pretty* and *slender*. Similarly, the word “sun” is masculine in Spanish was described as *powerful* and *threatening*. German speakers described sun focusing on its *warming* and *nourishing* qualities.

While the above mentioned research has the benefit of careful in-lab studies that explore differences in worldview for one dimension (e.g., time, space) or word (e.g., bridge, sun) at a time, it also has limitations. First, participants in the study can be asked a only limited number of questions in a given lab session, and it is likely that they provide less natural or unrealistic answers, as they are aware that they are being studied. Second, This type of studies are expensive, with very limited in scope, and they may not generalize well across large populations. In addition, any new aspect that needs further evaluation / analysis requires setting up and conducting a new study. In contrast, such shortcomings can be readily addressed using computational linguistics, by automatically learning from large amounts of data, which in our case includes self-authored text that people contribute on the Internet. Such texts follow the speaker's natural language patterns, as the participants will not be aware that they are being studied. In addition, the speakers avail to the reader various demographic details about themselves, such as gender, age, occupation, etc. This information can be automatically augmented to the text, to enable the creation of demographic-based language representations over very large amounts of data. The additional benefit of such models, besides very low cost and high replicability, lies in the possibility of selecting and modeling demographic groups that may be extremely rare or inaccessible to researchers in a typical in-lab study The availability of such rich data, paired with advances in machine learning techniques, enables reconsidering NLP tasks through a demographic lens, while also verifying whether current models are resilient to / representative of demographics, and exploring how they could be enriched to enable modeling of both the speakers and their language at the same time.

Despite this enormous potential, there is only limited work in computational linguistics that explores demographic differences in language use. Previous research in demographic-based computational linguistics has mainly focused on studying how the author demographics affect people's language use [11, 12, 13, 14]. Further, these differences are in turn used to infer authors' demographic attributes, such as age [15], gender [16, 17, 18, 19], income level [20], location [21], political affiliation [22], personality [23], popularity [24], socio-economic status [25], and mental health conditions [26, 27, 28]. Very few attempts have been made to identify and analyze demographic differences in language [29, 30]. Even fewer attempts were made to benefit from these differences in language use, and go beyond just inferring author attributes, and develop better language models for linguistic tasks [31, 32].

In this thesis, I explore the way demographic differences are encoded in social media language, by using the methods rooted in computational linguistics to model the interaction between language and demographics, and conversely explore if the demographic differences, when they exist, could help develop better language representations for a selected set of NLP tasks, thus paving the way for demographic-aware NLP.

As an initial effort to analyze demographic differences, I investigate whether automatic classification methods developed using linguistic features can discriminate word usages of different demographic groups with an accuracy higher than chance. To achieve this, I use very large social media datasets from which I extract frequently used words. The high accuracy of classifiers serves as an indication that there exist significant differences in the usage of these words by different groups. Owing to the frequent use of the selected words for the study in their day-to-day conversations, I consider the word usage differences as empirical evidence for the existence of demographic biases in people's worldviews. Using analyses over various feature types and ablation studies, I identify the attributes of language that most contribute to the differences [33].

Next, I investigate if demographic differences in language use can be leveraged to develop better linguistic models for NLP tasks. While the current NLP methods generally deal with advanced tasks such as relation extraction [34, 35], text similarity [36, 37], at the very core, many of these tasks assume some way of drawing connections (or associations) between words. The task of word associations involves predicting a word in response to / association with a given stimulus word. For instance, one may associate "mother" with "warmth," and "fire" with "burn."

Word associations [38] start forming early in life as language is acquired, and one learns them based on the environment where concepts lie in relation to each other. Hence, they are believed to mirror the human mind [39, 40], and variations in word associations across demographic groups provide insights into people's psychologies and their worldviews [41]. Computational linguistics has traditionally taken the "one-size-fits-all" approach for predicting associations, with most models being agnostic to the backgrounds of the speakers behind the language. Most of them are based on statistics (word similarity measures) derived from large corpora [42, 38, 43]. In this thesis, I go beyond generic corpus-based methods for extracting word associations, and propose a demographic-aware association prediction model, which takes into account the demographic attributes of the language users in the study. Using comparative experiments on large datasets, this thesis brings empirical evidence in support to the hypothesis that demographic-aware models outperform generic ones [44]. I regard this as the first step towards demographic-aware NLP.

As a next step, I examine more complex linguistic tasks, namely part-of-speech (POS) tagging [45] and syntactic parsing [46]. In particular, I analyze the role played by gender in part-of-speech tagging and syntactic parsing. However, the major roadblock for this type of study is the lack of availability of a dataset that is annotated with both linguistic and demographic information. In this thesis, the Wall Street Journal (WSJ) subset of the Penn Treebank dataset is augmented with gender information, and taggers and parsers are trained on this data to analyze syntactic differences related to author gender. The results underscore the importance of accounting for gendered differences in syntactic tasks, and outline future venues for developing more accurate taggers and parsers [47].

Finally, I study the humor preferences across different demographic groups, by examining the task of generating words to fill-in the blanks to create funny stories. In this thesis, a demographic-aware benchmark for humor generation is collected, and a demographic-aware humor generation framework is developed that takes into account the demographic backgrounds of the people involved in the study. Through human evaluations, the proposed framework is shown to outperform general baselines and human fillings, and qualitative and quantitative analyses are presented to understand what makes stories generated by the proposed framework funny, and how they differ from those by humans. The thesis presents further demographic-specific observations about the humor preferences of different demographic groups.

1.1 Research Questions

The goal of the research described in this thesis is to explore, on a large scale, the demographic differences between different groups of language users, and leverage these differences to develop better language representations and tools. Specifically, the thesis seeks to find answers to the following research questions:

1. **Are there significant differences in how words are used by different demographic groups in naturally occurring social media data? If yes, can these differences and the linguistic features responsible for them be characterized? How do the word usage differences vary for a wider variety of demographic categories?** Assuming the availability of a large collection of social media data, numerous experiments are run to determine if there exist words with significant usage differences between different demographic groups. To better understand the nature of these differences, the encoding ability of various feature types is explored and evaluated via ablation studies. Similar experiments are run on social media data to characterize demographic differences in word usage for a larger set of demographic categories, and within each category, a higher number of demographic groups.

2. **Can the demographic information of the authors be used to develop a more refined demographic-aware prediction model for word associations?**

A novel data set consisting of demographic-aware word associations is collected on Amazon Mechanical Turk (AMT). The data set consists of 800 responses for each of 300 stimulus words, obtained from a demographically-diverse group of respondents. A selected set of qualitative and quantitative analyses on the responses for the stimulus words show that word associations indeed vary across demographics. Further, the thesis proposes a new demographic-aware association prediction framework, which takes into account the demographics of the people behind the language. It compares the predictions of the proposed framework with those of traditional corpus-based measures and a generic word association prediction system, that do not account for the author demographics.

3. **Do demographics play a role in complex linguistic tasks such as part-of-speech (POS) tagging and syntactic parsing?**
   Existing datasets with demographic information are either too small to train on, or lack syntactic information, while sufficiently large syntactic datasets are not labeled with demographic details of the authors. To address this problem, heuristics are devised in this thesis to augment the Wall Street Journal (WSJ) subset of the Penn Treebank (PTB) corpus with the gender information of its contributors. State-of-the-art neural taggers and parsers are trained on gender-specific WSJ data, and the performance differences are analyzed between texts authored by men and women. Further analyses are conducted to study tag-wise part-of-speech tag and syntactic differences related to author gender.

4. **Can an automatic framework be developed for demographic-aware humor generation? How do the humor preferences vary across demographic groups?**
   The task of demographic-aware humor generation for the fill-in-the-blank game of Mad Libs is studied, with the aim of generating funny stories in a demographic-aware setting. A novel dataset for demographic-aware humor generation is collected on AMT. A demographic-aware humor generation framework for filling-in Mad Lib stories is proposed, which takes into account the demographic backgrounds of the people filling the stories. Through human evaluations, the proposed approach is compared with a general baseline and human fillings. Qualitative and quantitative analyses are conducted to understand what makes the stories generated by the proposed framework humorous, and how they differ from those of the baseline and humans. Further demographic-specific observations are made about the humor preferences of the different demographic groups considered in the study.

1.2 Thesis Organization

This thesis is organized as follows. Chapter 2 reviews previous research investigating the effects of demographics on language use, including studies undertaken in the fields of sociology, psychology and linguistics, in addition to computational linguistics.

Research question 1 regarding word usage differences across various demographic categories and groups is addressed in Chapters 3 and 4. Specifically, the social media dataset used to examine the existence of word usage differences between demographic groups is introduced, and automatic classification models are developed using various linguistic features for word usage discrimination. Various feature types used and further feature analysis experiments are detailed, shedding light on the nature of these differences. This framework is further extended to study the demographic word usage differences across a wider variety of demographic categories and groups in Chapter 4.

Chapter 5 provides a detailed description of the collection of the demographic-aware word association dataset, addressing the third research question. Further, it highlights several qualitative and quantitative differences in word associations between demographic groups, and describes the development of the demographic-aware word association prediction framework.

The fourth research question is addressed in Chapter 6, which describes the annotation framework for the WSJ subset of Penn Treebank with the gender information of the WSJ authors. Experiments and analyses are conducted to study the role played by gender in part-of-speech tagging and syntactic parsing.

Chapter 7 address the fifth research question and focuses on demographic-aware humor generation, and presents studies on humor preferences across demographic groups. In particular, it studies Mad Libs, a popular fill-in-the-blank game which involves generating word completions that turn the original texts humorous. The chapter provides description of the creation of a demographic-aware humor generation benchmark based on Mad Libs. It details the development of a demographic-aware humor generation framework for Mad Libs, and presents qualitative and quantitative analyses about humor preferences across demographic groups.

Finally, Chapter 8 summarizes the findings of the thesis and revisits the research questions posed in the introduction.

CHAPTER 2

Research on Demographic-Aware NLP

2.1 Introduction

Demographics is the study of a population based on factors such as age, gender, nationality, education, economic status, occupation, among others. Most of the previous research in demographics has been undertaken in fields such as sociology, psychology, and anthropology [2, 6, 48, 49]. In this chapter, the relevant research studies on gender, geography, income, and age groups in the fields of sociology and psychology are summarized, and the past and present developments on language and demographics in linguistics and computational linguistics are surveyed.

2.2 Psychology and Demographics

Field work in psychology has had much to say about the differences between various demographic groups. The masculine is stereotyped as detached, rational, and aggressive, and the feminine as nurturing, gentle, and tactful [50]. Shweder [6] examined similarities and differences in the perceptions, emotions, and ideologies of people belonging to different geographic locations. One of the earliest attempts in experimental ethnography belongs to Cohen et al. [48], who hypothesized that people belonging to different geographies behave differently. They examined the relationship between aggression level and geographic background among males who grew up in the North and South of the United States. By subjecting the participants to insult, they found that the southerners are more likely to engage in aggressive and dominant behaviour, when compared to northerners.

Surveys have been the principal means to examine the perceptions of people belonging to different demographics. Chen et al. [51] examined the demographic differences between East Asian and North American high-school students in response styles regarding the use of rating scales through various surveys. They found that the Asian students were more likely to use the midpoint on scales, while the North American groups were more likely to use extreme values. Similar works include those by Cox et al. [52], Li and Kirkup [53] and several others.

A recent study in psychology is that by Pennebaker et al. [5], in which the emotional expressiveness was measured among the northerners and southerners in their own countries, to test Montesquieu's geography hypothesis [2], which states that residents of warmer climates are more emotionally expressive than those living in cooler ones. More recently, the findings of Boroditsky et al. [10] indicate that people's perceptions of certain inanimate objects (such as bridge, violin, key, etc.) are influenced by the grammatical genders assigned to these objects in their native languages.

2.3 Language and Demographics

Previous research studies on language and demographics which looked at online data can be differentiated with respect to their aims. (1) Studies from sociolinguistic community that aim to empirically confirm hypotheses, such as that females tend to use more pronouns, or that males tend to more negations in their daily language use. (2) Studies that utilize the computational power to develop automatic methods to solve tasks involving demographics, such as examining the correlation of demographic attributes with linguistic features, or predicting user age, gender, or nationality.

2.3.1 Linguistic Theories

Previous sociolinguistic research mostly checked hypotheses formulated before the widespread availability and use of the Internet, such as that women use hedges more often [4], or men tend to use more negations [54]. Lakoff [4] found several characteristics of women's language, including words such as "lovely" and "adorable", or phrases such as "it seems to be" or "would you mind." It has been found for instance that men and women differ on private versus public speaking, on "report talk" versus "rapport talk" – these and other facets of relational dialectics are gendered and constitute the so-called "GenderLens" [55]. There is a large body of work on the connection between language and gender in the field of sociology [56].

As stated by Kramsch [57], demographics made their way into linguistics through the study of language as a window into people's social lives, ideologies, and everyday experiences. Several works in the field of psycholinguistics have examined the usage of language to understand different personality and behavior traits of people [58, 59, 60, 61, 62, 63].

Ramirez and Pennebaker [58] found that linguistic features of essays are empirically linked to health changes. For instance, use of positive emotions, increasing use of causal and other cognitive words, and shifts in pronoun use are correlated with fewer physician visits.

More recently, location has become central in sociolinguistics [64]. An early work in this direction was Kurath's *Word Geography of the Eastern United States* [65], in which he conducted interviews and mapped the occurrences of equivalent word pairs such as *stoop* and *porch*.

2.3.2 Demographics Prediction from Language

The systemic differences in language use across demographic groups can be leveraged to infer author demographic attributes, including age [15], gender [16, 17], income [20], location [21], political affiliation [22], personality [23], popularity [24], socio-economic status [25], and mental health conditions [26, 27, 28].

2.4 Demographics in Word Associations

Word associations have captured the attention of psychologists since at least the early 1900s. In [66], Kent and Rosanoff proposed the use of a set of 100 emotionally neutral words for word associations survey. The same year marked the beginning of a long term study conducted at the University of Minnesota, that was focused on capturing the primary responses of students to trigger words, collecting data in 1910, 1927, 1933, 1952 and 1960 [67]. The researchers have noticed that students' answers systematically changed across the datasets collected at different points in time, with the norms collected in a closer time frame displaying a higher correlation than those gathered between more distant survey dates. Also, primary responses that had a higher frequency exhibited a stronger resilience across datasets, while the usage of super-ordinate norms (e.g. a word whose meaning encapsulates the meaning of other words, such as blue representing the superset of teal, turquoise and navy) had steadily declined.

While this study was surveying the same age group (students) at different points in time, [41] surveyed 738 subjects with ages ranging between 18 and 87, with particular care to include people from different socioeconomic classes (from unskilled laborers to professionals). Their conclusion was that primary responses' variability increases with age, as the strength of individuals' commonality decreases. For example, older people associated "sleep" with "awake," instead of "bed" or "dream" which were the preferred answers of the younger age group.

A psycholinguistics study that looked at the impact that the nationality of respondents may have on formed word associations was carried out by Rosenzweig [68], employing the stimulus word list proposed by Kent and Rosanoff [66] manually translated into several West European languages. Based on the primary responses coming from native speakers of English, French, German and Italian, which were mapped back into English, the author concluded that the associations formed by speakers of the four languages are very similar, with "almost half the comparisons in any pair of languages yielding agreements," where the most frequent responses are encountered across pairs of languages, while rare responses do not correspond. Rosenzweig also mentions that the primary responses across genders of the same nationality have a high agreement, with French nationals split by gender agreeing 75% of the time, while American nationals agreeing 82% of the time.

Given that the primary responses were compared across languages and people with a relatively common origin (West European), the research in this thesis seeks to investigate whether similar results are encountered when looking at different locations (namely United States (US) versus India), or genders (males versus females). Furthermore, this study is conducted in English from the beginning, to eliminate a third party's subjectivity in mapping primary responses from one language to another. In addition, this thesis verifies whether the same high level of agreement holds between American men and women, and Indian men and women, while also exploring demographic strength of word association between women and men, and Indians and Americans.

There have also been attempts in computational linguistics to derive associations not based on survey results (which are static and resource intensive), but based on statistics derived from large corpora [42, 69, 38]. Research in semantic similarity can also be used to model associations based on several directions: (1) co-occurrence metrics that rely on large corpora such as PMI [38], second order PMI [37], or Dice [70]; (2) distributional similarity-based measures, that characterize a word by its surrounding context such as LSA [71], ESA [72], or SSA [73]; and (3) knowledge-based metrics that rely on resources such as lexica or thesauri [74, 75, 76, 77]. However, most of these metrics have so far been applied to model the relatedness between two words, namely given a word pair, to score how related the two words are; as such, they have not been used to predict free association norms, namely given a word, to attempt to determine the most likely word that a human would associate with that stimulus.

Large word association databases exist, such as the one collected by Deyne et al. [78], who used a set of 12,000 stimulus words and surveyed 70,000 participants. Yet to our knowledge, no concerted attempt has been made to gather word associations jointly with the demographic characteristics of the people behind them.

While not directly seeking to extract word associations but rather trying to represent language meaning through a locality lens, [79] have proposed using distributed representations to model words employed by social media users from different US states. They were able to show that the regional meaning of words can successfully be carried by word embeddings; for example the word "wicked" was most similar to the word "evil" in Kansas, while in Massachusetts, it was most similar to "super" (based on the cosine similarity of the words' vectorial representation). In contrast, the rationale in this thesis is to explore if word associations can be automatically derived from large corpora annotated with user-centered attributes such as location or gender.

2.5 Demographics in Syntax

Lexical language variation among groups has made it possible to analyze word usages across demographic groups [80, 33, 81]. Such variations can be leveraged to predict author gender and other latent attributes, such as age and geographic origin [15, 21, 16, 17]. However, language variation goes beyond the lexical level – syntax plays a crucial role in the systemic variation among groups. Within sociolinguistics, the relation between gender and syntax has a long history [82, 83]. Mondorf [11] does a systemic analysis of the usage of various types of clauses and their positions among men and women, stating that women have a higher usage of adverbial, causal, conditional and purpose clauses, while men tend to use more concessive clauses.

Several works in the NLP community use syntactic features from text, such as part-of-speech tags and dependency relations, to improve data-driven dependency parsing [84], or sentiment classification [85, 86], among other tasks. However, most of these works use social media datasets, which preserve significant differences in the writing styles of the authors, and the standard pre-trained taggers and parsers used to obtain such constructs are gender-agnostic.

There have been some works in the recent past that augment models with author demographics to develop gender-aware frameworks. [87] used continuous gender representations to augment the models for part-of-speech tagging, prepositional-phrase attachment, sentiment analysis, sarcasm detection, and stance detection, but evaluated on a dataset without gender information. They reported no significant improvement, which could be due to context effects. This thesis provides a guideline to obtain a gendered dataset with syntactic annotations, and seeks to analyze syntactic differences across genders, and explore if gender-specific settings can help taggers and parser obtain better performances.

2.6 Demographics in Humor Generation

This thesis focuses on humor generation and attempts to augment demographic information of the initiators (authors) of humor in an automatic humor generator.

2.6.1 Humor Generation

There is a long history of research in the general theories of humor [88, 89, 90, 91, 92, 93]. Among the three main theories of humor are the superiority theory, the relief theory and the incongruity theory. While the general idea behind the superiority theory is that a person laughs about either misfortunes of others as laughter expresses feelings of superiority over them, or over a former state of oneself, the relief theory maintains that laughter is a homeostatic mechanism by which psychological tension is reduced. The third and the most widely accepted one is the incongruity theory, which identifies a joke as funny when it has the surprise that fails to comply with the conventional expectations.

The Sematic Script Theory of Humor (SSTH) states that a joke emerges when it can be interpreted according to two different, generic scripts, one of which is less obvious [88]. One of the first attempts in applying the SSTH theory was by [94] to automatically generate two-line jokes. According to the Benign Violation Theory [95], the unexpected must logically follow from the set up and must not be offensive to the reader, else the reader is left confused and the joke is not funny.

In general, it is very difficult to apply humor theories directly to generate humor, because they require a high degree of common sense understanding of the world. Due to this, the most successful algorithmic work in humor generation is limited to using relatively shallow linguistic rules on specific types of jokes and certain patterns of text such as short riddles, one-liners, and acronyms. Joke Analysis and Production Engine (JAPE), as proposed in [96], used an algorithm to generate funny punning riddles, such as:

Question: What do you call a weird market?
Answer: A bizarre bazaar.

JAPE used an algorithm to generate phonologically ambiguous riddles, having noun phrase punchlines, by discovering candidate ambiguities in either spelling (e.g., cereal vs. serial) or word sense (e.g., river bank vs. savings bank). Though not as funny as human-generated jokes, JAPE riddles were found to be funnier than non-jokes, when evaluated by children.

Another attempt at computational humor generation was HAHAcronym generator

[97]. Its goal was to generate alternate, funny versions of existing acronyms, ensured by incongruity theories, as follows:

DMSO: Defense Modeling and Simulation Office.
DMSO: Defense Meat-Eating and Salivation Office.

Their algorithm retains part of the acronym, making changes in the remainder, to obtain different semantics, rhymes, extreme-sounding adverbs, and antonyms.

[98] developed an algorithm to generate jokes, such as I like my coffee like I like my war, cold, filling in the three blanks. They encoded four assumptions about what make a joke funny, using discrete probability tables learned from a large corpus of regular text data, along with part-of-speech tagging and an estimate of different possible senses of the words. 16% of their automatically generated jokes were considered funny, compared to 33% when the same type of jokes were generated by people. This is the first attempt at unsupervised joke generation with promising results.

More Recently, [99] attempted to generate humorous texts by substituting a single word in a short text, within some constraints. The resulting texts showed that taboo words made a text recognizable as humorous. [100] developed an algorithm for a computer-aided approach to help a human player fill-in the blanks of a Mad Lib story, to make it sound funny. Below is an example of their work (with the underlined text in italics is their prediction as compared to that in normal font).

As an adult, Batman wore a/an costume / *veil* (noun) to protect his identity / *jaw* (noun) while fighting crime / *acne* (noun) in Gotham.

Their technique to generate complete, funny stories involved the use of a language model to generate words, a classifier to rank and retain the top 20 funny candidates and further filter these words, and the use of humans to select possible fillings from these 20 candidates for each blank, thus creating funny Mad Lib stories in a semi-automated approach. Their analyses showed that human judges found those stories funny which were coherent. Another study in a related research line is by [101], where the task was to choose a *satisfying* ending to a story. Also, it is to be noted that while the work in [102] is called "Visual Madlibs," it essentially augments an image dataset with fill-in-the-blank questions, such as "This place is a park."

2.6.2 Effect of Demographics in Humor

In sociolinguistics, the relationship between humor and demographics is widely studied. For instance, Mundorf et al. [103], found a complex pattern in humor appreciation. Specifically, males enjoyed sexual humor more than females did, irrespective of the victim-gender. For hostile humor, each gender enjoyed humor with an opposite-gender victim more than the other gender did. In the Gender and Humor chapter of her thesis, Hay [104] surveyed previous studies that claimed women are less inclined towards humor than men. [105] statistically highlighted some interesting patterns in the humor of New Zealand men and women. Its modeling revealed that women more likely shared funny personal stories to create solidarity, while men used other strategies to achieve the same. Freud [106] claimed that women have fewer strong taboo feelings to repress, and hence do not need a sense of humor. This perception is however changing in recent times with more recent research claiming that humor is different between genders.

Humor styles also vary across nationalities, as indicated by [107], whose findings suggested that humorous communications from Korea, Germany, Thailand and United States, had variable content for humorous advertising, while sharing certain universal cognitive structures. Contingent on the ethnic background of audience members, certain humor styles can be detrimental to interpersonal relations [108]. For instance, a manager who is addressing a group of workers from a high power distance society should avoid self-defeating or affiliative humor.

CHAPTER 3

Identifying Demographic Differences in Word Usage

3.1 Introduction

In this chapter, a step-by-step framework is presented that allows identification of word usage differences between demographic groups in social media language in a data-driven bottom-up fashion using the power of large-scale computational linguistics. In particular, I hypothesize that computational models can be developed to identify differences in word usage between demographic groups in their personal writings on social media platforms. Rather than starting with predetermined hypotheses (e.g., that Spanish and German speakers would have a different way of talking about bridges), computational techniques are used to run experiments on hundreds of frequently-used content words, and consequently identify those words where usage differences exist between demographic groups. Owing to the frequent usage of these words in their personal writings, I regard the word usage differences as reflection for the demographic biases in people's worldviews. This hypothesis is explored by seeking answers to two main research questions.

First, given a word W, are there significant differences in how this word is being used by different demographic groups? Word models are developed in which classifiers are trained on several classes of linguistic features, and attempts are made to differentiate between usages of the given word W by different demographic groups. By applying them to a large number of words, these models are used to identify the words for which there exist significant usage differences between the different groups of interest.

Second, if significant differences in the usage of a word are identified, can feature analysis be used to understand the nature of these differences? Several analyses are performed: (1) Feature ablation that highlights the linguistic features contributing to these differences; (2) Topic modeling applied to the words with significant differences, used to identify the dominant topic for each demographic group and to measure the correlation between the topic distributions in different demographic groups; and (3) One-versus-all classification models, where it is attempted to isolate the idiosyncrasies in word usage for each demographic group at a time.

This chapter is organized as follows. Section 3.2 describes the datasets obtained from social media platforms for this task. The linguistic features used and classification methods developed to identify demographic usage differences are briefed in Section 3.3. Section 3.4 presents feature analysis and topic modeling experiments to analyze the nature of demographic differences, and identify the attributes of language that contribute to these differences in word usage. Finally, the findings of this this chapter are summarized in Section 3.5.

3.2 Data

The work in this chapter is based on personal writings collected from blogs, and specifically targets word usage differences between Australia and United States. These two countries are selected for two main reasons: (1) they both use English as a native language, and therefore one can avoid the noise that would otherwise be introduced by machine translation; and (2) they have a significant number of blogs contributed in recent years, which can be used to collect a large number of occurrences for the given set of words.

A large corpus of blog posts is obtained by crawling the blogger profiles and posts from Google Blogger. Google Blogger is a free online blog-publishing platform launched in August 1999, and it is used by people from multiple English speaking countries. The motivation to use the blog genre is that, blog posts are lengthier compared to microblogs originating on sites such as Twitter or Facebook, as well as richer in terms of vocabulary and topic usage. In addition, users complete their profile information using a drop-down list for countries, gender and industry, ensuring a closed vocabulary that enables better matches across users. For each profile, a maximum of 20 blogs, and for each blog, up to 500 posts are considered. Table 5.6 gives statistics of the data collected in this process. The blog posts are cleaned by removing the HTML tags, and are tagged with part-of-speech tags [109].

COUNTRY	PROFILES	BLOGS	POSTS
Australia	469	1,129	320,316
United States	374	1,267	471,257

Table 3.1: Blog statistics for the two target cultures.

Next, a pool of candidate target words is created by identifying the top $1,500$ content words based on their frequencies in the blog posts, additionally placing a constraint that they cover all open-class part-of-speech tags: of the $1,500$ words, 500 are nouns, 500 verbs, 250 adjectives, and 250 adverbs. These numbers are chosen based on the number of examples that exist for the target words; e.g., most ($>$ 490) of the 500 selected nouns have more than 300 examples; etc. All possible inflections are considered for these words; for instance for the verb *write*, the forms *writes*, *wrote*, *written*, and *writing* are considered. The possible inflections for the target words are added manually, to ensure correct handling of grammatical exceptions.

To obtain usage examples from the two demographic groups for these words, paragraphs are extracted from the blog posts that include the selected words with the given part-of-speech tags. Of these paragraphs, those are discarded that contain less than ten words. Long paragraphs are truncated so that they include a maximum of 100 words to the left and right of the target word, disregarding sentence boundaries. The contexts of the target words are then parsed to get the dependency labels of the context words in relation to the target word [110].

The data is explicitly balanced across time. Noting there could be cases where the number of blog posts published in a specific year is higher compared to that in other years due to certain events (e.g., an Olympiad, or a major weather related event), samples are drawn from the datasets from several different time periods. Specifically, for each demographic group, equal number of instances are selected from four different years (2011-2014). Table 3.2 shows the per-word average number of data instances obtained in this way for each part-of-speech tag, for each culture for the years 2011-2014.

Note that there is no attempt to balance the data across topics (domains), as potentially different topic distributions are regarded as a reflection of the differences between demographic groups (e.g., Australia may be naturally more interested in water sports than United States is). Instead, the datasets are explicitly balanced with respect to time, as described above, to avoid temporal topic peaks related to certain events.

Country	Noun	Verb	Adj	Adv
Australia	22,461	18,396	19,206	19,377
United States	15,199	12,347	12,513	12,952

Table 3.2: Average number of instances for the 1,500 target words for the years 2011-2014.

3.3 Finding Words with Demographic Bias

In this section, the first research question is addressed: given a word W, are there significant differences in how this word is being used by two different demographic groups? For this, a classification task is formulated where the goal is to identify, for the given target word W, the demographic attributes of the writer of a certain occurrence of that word. For each word *W*, a vectorial model is generated that accounts for *W*'s usage across all the groups, i.e the representations of *W* pertaining to both Australia and United States data are accounted for. Each word is represented through several lenses, by looking at potential linguistic differences, accounting for social and psycholinguistic aspects. If the accuracy of such a classifier exceeds that of random chance, this can be taken as an indication of word usage differences between the two groups. Classification experiments are run on each of the $1,500$ words described in Section 3.2, and consequently words with significant usage differences between Australia and United States are identified.

3.3.1 Features

The following four types of features are extracted from the word datasets.

Local features. These features consist of the target word itself, its part-of-speech tag, three words and their part-of-speech tags to the left and right of the target concept, nouns and verbs before and after the target concept. These features are used to capture the immediately surrounding language (e.g., descriptors, verbs) used by the writers while describing their views about the target word.

Contextual features. These features are determined from the global context, and represent the most frequently occurring open-class words in the contexts of the word W in each demographic group. At most ten such features are allowed for each group, and a threshold of minimum of five occurrences is imposed for a word to be selected as a contextual feature. Contextual features express the overall intention of the blogger while writing about the target word.

Socio-linguistic features. These features include (1) fractions of words that fall under each of the 70 Linguistic Inquiry and Word Count (LIWC) categories [111]; the 2001 version of LIWC includes about 2,200 words and word stems grouped into broad categories relevant to psychological processes (e.g., emotion, cognition); (2) fractions of words belonging to each of the five fine-grained polarity classes in OpinionFinder [112], namely strongly negative, weakly negative, neutral, weakly positive, and strongly positive; (3) fractions of words belonging to each of five Morality classes [113], i.e., authority, care, fairness, ingroup, sanctity; and (4) fractions of words belonging to each of the six Wordnet Affect classes [114], namely anger, disgust, fear, joy, sadness, and surprise. These features provide social and psychological insights into the perceptions bloggers have about the words they use.

Syntactic features. These features consist of parser dependencies [115] obtained from the Stanford dependency parser [110] for the context of the target word. Among these, the following dependencies are selected for each part-of-speech tag: (1) nouns: root word of context (`root`), governor[1] if noun is nominal subject (`nsubj`), governor verb if noun is direct object (`dobj`), adjectival modifier (`amod`); (2) verbs: root, nominal subject (`nubj`), direct object (`dobj`), adjectival complement (`acomp`), adverb modifier (`advmod`); (3) adjectives: root, noun being modified (`amod`), verb being complemented (`acomp`), adverb modifier (`advmod`); (4) adverbs: root, adverb modifier (`advmod`). These features capture syntactic dependencies of the target word that are not always obtained using its context alone.

3.3.2 Demographic Word Models

The features described above are integrated into an AdaBoost classifier.[2] This classifier is selected based on its performance on a development dataset, when compared to other learning algorithms. The performance of this classifier is compared with that of a random choice baseline, which is always 50%, given the equal distribution of data between the two demographic groups. This allows identification of the words which have significant classification accuracies for the identification of the demographic group of the writers of those words, which is taken as an indication of existence of word usage differences between the two groups.

Throughout the paper, all the results reported are obtained using ten-fold cross-validation on the word datasets. When creating the folds, it is explicitly ensured that posts authored by the same blogger are not shared between the folds, which in turn ensures no overlap between bloggers in training and test sets. This is important as repeating bloggers in both the train and the test splits could potentially overfit the model to the writing styles of

[1] The convention provided in `http://nlp.stanford.edu/software/dependencies_manual.pdf` is followed.

[2] The open source machine learning framework Weka [116] is used for all the experiments. The default base classifier for AdaBoost, i.e., a DecisionStump is used.

individual bloggers, rather than learning the underlying demographic differences between them.

To summarize the cross-validation process: First, for each of the $1,500$ target words, equal number of instances containing the given target word or its inflections are collected from Australia and United States, from each of the selected years (2011-2014). The posts belonging to Australia and United States are then divided each into ten approximately equal groups, such that no two groups have bloggers in common. Finally, the corresponding groups are combined to form a total of ten bi-demographic groups that are approximately of equal size, which form our cross-validation splits. [3]

To compute the statistical significance of the results obtained, a two-sample t-test is performed over the correctness of predictions of the two systems namely, AdaBoost and random chance classifiers. Disambiguation results that are significantly better ($p < 0.05$) than the random chance baseline of 50% are marked with *.

On an average, the AdaBoost classifier leads to an accuracy of $58.36\%^*$, which represents an absolute significant improvement of 8.36% over the baseline (a random chance of 50%). Table 3.3 shows the average classification results for each part-of-speech tag, as well as the number of words for which the AdaBoost classifier leads to an accuracy significantly larger than the baseline. These results suggest that there are indeed differences in the ways in which writers from Australia and United States use the target words with respect to all the part-of-speech tags.

PART-OF-SPEECH	AVERAGE ACCURACY	WORDS WITH SIGNIFICANT DIFFERENCE
NOUNS	57.51*	393
VERBS	58.01*	395
ADJECTIVES	59.25*	207
ADVERBS	61.77*	215
OVERALL	58.36*	1,210

Table 3.3: Average ten-fold cross-validation accuracies and number of words with accuracies significantly higher than the baseline, for each part-of-speech tag, for United States vs. Australia.

[3] This condition of equal data in each group is only approximate, as there will generally not be an exact division of bloggers with equal data. The average size of a cross-validation train split for a target word is 6,247, while that for a test split is 820.

3.4 Where is the Difference?

Here, the second research question is addressed: Once significant differences in the usage of a word are identified, can feature analysis be used to understand the nature of these differences?

3.4.1 Feature Ablation

First, the role of the different linguistic features is studied when separating between word usages in Australia and United States through ablation studies. For each of the feature sets specified in Section 3.3, the word models are retrained using just that feature set type, which enables locating the features that contribute the most to the observed demographic differences.

The left side of Table 3.4 shows the ablation results averaged over the $1,500$ target words for the four sets of features. It can be seen that contextual and socio-linguistic features perform consistently well across all the part-of-speech tags, and they can alone obtain accuracies close to the all-feature performances.

POS Tag	Loc	Con	Soc	Syn	All	LIWC	OF	ML	WNA
Nouns	49.06	57.22*	56.52*	46.54	57.28*	56.62*	56.21*	54.17*	52.94
Verbs	47.97	57.65*	57.04*	47.71	57.90*	56.90*	56.28*	53.73*	53.25*
Adjectives	48.67	58.63*	57.90*	47.52	59.01*	58.03*	57.31*	55.42*	54.30*
Adverbs	50.72	61.09*	59.80*	46.85	60.81*	60.27*	59.80*	57.00*	56.57*
All	48.91	58.25*	57.47*	47.14	58.36*	57.55*	57.01*	54.70*	53.87*

Table 3.4: Feature ablation averaged over 1,500 target words. Loc: Local features, Con: Contextual features, Soc: Socio-linguistic features, Syn: Syntactic features, All: All feature types, LIWC: Linguistic Inquiry and Word Count, OF: OpinionFinder, ML: Morality Lexicon, WNA: WordNet Affect. Statistically significant values are marked with *.

Another feature ablation experiment is performed to explore the role played by various socio-linguistic features. The right side of Table 3.4 shows the classification accuracies obtained using each sociolinguistic lexicon at a time: LIWC, OpinionFinder, Morality, and WordNet Affect. Among all these resources, LIWC and OpinionFinder appear to contribute the most to the classifier; while the morality lexicon and WordNet Affect also lead to an accuracy higher than the baseline, their performance is clearly smaller.

3.4.2 Topic Modeling

Next, the analysis is focused on the top 100 words (25 words for each part-of-speech tag) that have the most significant improvements over the random chance baseline, considered to be words with demographic bias in their use. The average accuracy of the classifier obtained on this set of words is 65.45%; the accuracy for each part-of-speech tag is shown in the second column of Table 3.8.

The different usages of the words in our set of 100 words are modeled using topic modeling. Specifically, Latent Dirichlet Allocation (LDA)[4] [118] is used to find a set of topics for each word, and consequently identify the topics specific to either Australia or United States.

As typically done in topic modeling, the data is preprocessed by removing a standard list of stop words,[5] words with very high frequency ($> 0.25\times$corpus_size), and words that occur less that five times. To determine the number of topics that best describe the corpus for each of the 100 words, the average corpus likelihood over ten runs [120] is used. Specifically, the number of topics ($>= 2, <= 10$) for which the corpus likelihood is maximum is chosen.

For each data instance, a topic is said to dominate the other topics if its probability is higher than that of the remaining topics. For a given word, the *dominating topic* is identified for each demographic group as the topic that dominates the other topics in a majority of data instances. This definition of dominating topic is used in all the analyses done in this section.

Quantitative Evaluation. To get an overall measure of how the different demographic groups use the words that are found to have significant differences, the Spearman's rank correlation is computed between the topic distributions for the two demographic groups. For each topic, the number of data instances in which it dominates the other topics is obtained, in both demographic groups (Australia and United States). Subsequently, the overall Spearman correlation coefficient is measured between the dominating topic distributions for all 100 words. In other words, the distribution of topics is compared across the two demographic groups for each word. The Spearman coefficient is calculated as 0.63, which reflects a medium correlation between the usages of the words by the two demographic groups.

[4] LDA has been shown to be effective in text-related tasks, such as document classification [117].

[5] The list of stop words is taken from English WordNet [119].

Qualitative Evaluation. For qualitative evaluation, Table 3.5 and Table 3.6 show five sample words for each part-of-speech tag, along with the identified number of topics and the dominating topics for Australia and United States. Labels are associated with the hidden topics manually, after looking at the corresponding top words falling under each of them.

As seen in this table, the number of topics that best describe each word can vary widely between two topics for words such as *start* and *economic*, up to ten topics (which is the maximum allowed number of topics) for words such as *color* or *support*. The dominating topics illustrate the biases that exist in each demographic group for these words; for instance, the word *teach* is dominantly used to describe academic teaching in Australia, whereas in United States it is majorly used to talk about general life teaching. Several additional examples of differences are shown in Table 3.5.

3.4.3 One-versus-all Classification

For gaining further insight into word usage differences between United States and Australia, this study is expanded to develop word models to separate word usages in Australia (or United States) from a mix of ten different demographic groups. In other words, a one-versus-all classification is conducted using the same process as described in Section 3.3, but using Australia (United States) against a mix of other demographic groups, to examine any features specific to Australia (United States).

In order to do this, data is collected from nine additional English speaking countries, as shown in the left side of Table 3.7. As before, the data for each country is balanced over time, and it includes an equal number of instances for four different years (2011-2014). The right side of Table 3.7 shows the average number of instances per word collected for each part-of-speech tag.

In this classification, for a given target word, one half of the data is collected from Australia (United States), and the other half is collected from the remaining countries, drawing 10% from each country. These classifiers are run for all the 100 words that are previously identified as having demographic bias in their use.

The average classifier accuracy for the Australia-versus-all classification, using ten-fold cross validation, is 64.23%, as shown in the third column of Table 3.8. The same one-versus-all classification is repeated for United States, and the average accuracy is 54.89%; the results of this experiment are listed in the last column of Table 3.8.

Overall, the performance improvement over the baseline is higher for Australia-versus-others (14.23% absolute improvement) than it is for United States-versus-others (4.89% absolute improvement). From this, it can be infererd that that the performance improvement

Word	Accuracy					NT	Dominating Topic	
	All	Loc	Con	Soc	Syn		Australia	United States
Nouns								
store	67.14*	53.11*	67.15*	64.98*	50.79	10	Online (blog, card, gifts, online, sale)	Grocery (grocery, shopping, things)
support	66.05*	44.56	52.45	62.54*	51.29	10	Education (school, students, education)	Law (public, police, law, social)
color	65.56*	50.74	67.35*	65.56*	50.27	10	Background (pink, design, background)	Picture (paint, kit, picture, photo)
version	65.42*	47.75	65.47*	63.57*	47.28	4	Rendering (story, thought, school)	Album (song, music, album, classic)
phone	64.32*	53.93*	62.51*	58.19*	52.49	3	Communication (calls, email, message)	Friend (night, friend, talking)
Verbs								
go	63.93*	47.29	64.15*	65.18*	42.77	2	Time (time, day, night, love, today)	Life (life, world, work, god, children)
start	63.66*	46.80	63.46*	62.09*	54.08*	2	Daybreak (starting, day, love, morning)	School (starting, school, softball)
know	63.51*	50.18	64.20*	62.40*	47.12	2	Good Times (love, good, things, life)	Children (children, school, year, book)
sing	63.54*	51.07	63.96*	58.19*	49.80	4	Christmas (christmas, happy, kids)	Rock Singer (band, guitar, rock, singer)
teach	62.66*	52.72	61.69*	59.88*	51.18	10	Academics (teachers, education, curriculum)	Life (time, young, life, thought, work, friends)
Adjectives								
various	65.64*	54.0*	64.70*	63.18*	48.40	10	Politics (political, party, war, revolution)	Diversity (states, people, companies)
own	64.76*	56.89*	65.19*	62.71*	51.79	2	Power (life, war, political, power)	Life (love, music, work, school)
economic	61.88*	44.82	42.11	59.16*	33.07	2	Finance (economy, financial, market, tax)	Politics (political, social, war, power)
old	66.45*	42.95	67.09*	64.74*	39.89	2	Age (older, children, family, age)	Past (back, school, days, love)
human	64.37*	52.78	60.92*	59.09*	48.14	9	Rights (rights, law, freedom, civil)	Life (life, time, love, real)

Table 3.5: Five sample nouns, verbs and adjectives with significant usage difference in the Australian and American cultures. All: All the features, Loc: Local features, Con: Contextual features, Soc: Socio-linguistic features, Syn: Syntactic features, NT: Number of topics.

WORD	ACCURACY					NT	DOMINATING TOPIC	
	ALL	LOC	CON	SOC	SYN		AUSTRALIA	UNITED STATES
ADVERBS								
quite	73.56*	55.22*	70.98*	73.49*	51.50	2	EXTENT (time, back, good, love, thought)	EXTENT (time, back, good, love, thought)
else	67.35*	52.45	68.21*	64.66*	45.19	2	(make, find, world, invented, book, christ)	(time, life, night, love, work, home, god)
actually	65.62*	50.67	66.60*	65.51*	45.75	9	POLITICS (law war, people, government)	FAMILY (kids, fun, home, couple)
usually	68.37*	57.71*	70.16*	67.03*	44.35	2	SPORTS (softball, cricket, play)	ROUTINE (things, work, love, home)
certainly	64.64*	53.76*	63.25*	62.87*	43.03	2	SPORTS (game, series, softball, wrestling)	TIME (time, work, things, make)

Table 3.6: Five sample adverbs with significant usage difference in the Australian and American cultures. All: All the features, Loc: Local features, Con: Contextual features, Soc: Socio-linguistic features, Syn: Syntactic features, NT: Number of topics.

over the baseline for the Australia-versus-United States task can be majorly attributed to the different word usages in Australia from the remaining countries. In other words, United States is more aligned with the "typical" (as measured over ten different countries) usage of these words than Australia is.

3.5 Conclusions

In this chapter, the problem of identifying word usage differences between people belonging to different demographic groups is explored. Specifically, the differences between Australians and Americans are studied based on the words that are frequently in their online writings. Using a large number of examples for a set of $1,500$ words, covering different part-of-speech tags, this chapter shows that classifiers can be developed based on linguistic features that can distinguish between the word usages from the two demographic groups with an accuracy higher than chance. This is taken as an indication that there are significant differences in how these words are used in the two demographic groups, reflecting demographic bias in word use.

To better understand these differences, several analyses are performed. First, using feature ablation, the contextual and socio-linguistic features are identified as the ones playing the most important role in identifying these word usage differences. Second, focusing on the words with the most significant differences, topic modeling is used to find the main

Country	Profiles	Blogs	Posts	Word instances			
				Noun	Verb	Adj	Adv
Barbados	440	830	32,785	581	466	490	476
Canada	461	1,097	397,479	15,015	12,020	12,965	12,512
Ireland	473	978	231,240	8,936	7,161	7,919	7,809
Jamaica	451	770	41,495	1,632	1,318	1,353	1,297
New Zealand	450	1,112	226,900	8,713	7,313	7,883	8,284
Nigeria	464	908	223,772	13,719	9,710	9,796	7,631
Pakistan	458	1,404	135,473	2,861	2,130	2,243	1,847
Singapore	406	803	208,972	5,623	5,430	5,447	6,639
United Kingdom	473	934	282,740	10,887	9,432	10,021	11,066

Table 3.7: Statistics for blog data collected for additional English speaking countries.

Part-of-speech	United States vs. Australia	Australia vs. All	United States vs. All
Nouns	65.54*	63.45*	57.07*
Verbs	64.20*	63.97*	53.87*
Adjectives	65.13*	64.36*	54.48*
Adverbs	66.92*	65.13*	54.13*
Overall	65.45*	64.23*	54.89*

Table 3.8: Ten-fold cross-validation accuracies averaged over the top 100 target words for United States vs. Australia; Australia vs. a mix of ten other countries; United States vs. a mix of ten other countries.

topics for each of these words, which enabled the identification of the dominant topics for each word in each culture, pointing to several interesting word usage differences, as outlined in Table 3.5 and Table 3.6. The correlation is measured between the topic distributions for the top 100 words between the two groups, and a medium correlation of 0.63 is obtained. Finally, a one-versus-all classification is performed for these 100 words, where the word use instances drawn from either Australia or United States are compared against a mix of instances drawn from ten other groups, which suggested that United States is a more “typical” demographic group when it comes to word usage (with significantly smaller differences in these one-versus-all classifications than Australia).

In the next chapter, this work is extended to understand differences in word usages between a larger number of demographic groups.

The corresponding publication with the findings in this chapter is [33]. The demographic word datasets used in the experiments reported in this chapter are available at http://lit.eecs.umich.edu.

CHAPTER 4

Word Usage Differences Across a Larger Group of Demographics

4.1 Introduction

In this chapter, the work on identifying word usage differences is extended to address more demographics – country, gender and industry – and a larger number of countries, genders and industries in each of these demographics. Similar to the previous chapter, I investigate words that occur with preponderance in each demographic, and refine candidate lists with seemingly differing usage. For each word W in a given demographic, vectorial models are generated that account for W's usage across all demographic groups, i.e if gender is considered, the representations of W pertaining to both male and female data are accounted for. Each word is represented through several lenses, by looking at potential linguistic differences, accounting for social and psycholinguistic aspects, as done in Chapter 3. In addition, this chapter also represents the words by modeling the topical content of the context in which the word appears, owing to the large diversity in the topics observed in the previous chapter.

For the remaining of this chapter, *demographic* or *demographic category* will refer to the large demographic categories, such as gender, location, or industry, while *complementing demographics* or *slices* will refer to the sub-categories in each demographic, such as male and female for gender. Machine learning models are then trained over these word representations, aiming to predict which demographic slice each word instance belongs to. The model's performance is taken as an indicator of how much discriminating power W has within each demographic, therefore allowing to conclude whether words are used differently by people from complementing demographic slices. Throughout this chapter, the focus will be on gender, location and industry, but other demographic criteria can be explored as well.

COUNTRY	USER PROFILES	POSTS	AVG. POSTS	STD. DEV
United Kingdom	886	393,160	444	758
Australia	879	412,743	470	1,099
Canada	854	501,475	587	1,276
Nigeria	844	259,817	308	620
New Zealand	843	231,469	275	501
Ireland	831	215,702	260	612
South Africa	825	140,659	171	353
Philippines	808	243,194	301	553
United States	800	556,186	695	1,134
Singapore	748	262,006	350	556
India	719	210,456	293	797
Pakistan	707	100,755	143	688

Table 4.1: Blog data statistics for the selected geographic locations (rounded to the nearest integer).

4.2 Extracting Candidate Word Lists with Demographic Bias

Candidate word lists with demographic bias within each demographic are obtained by examining social media data collected from Google Blogger. Unlike in Chapter 3, in this chapter, the part-of-speech tags are ignored so as to closely investigate a smaller list of 500 words for each of the three demographic categories.

Country. A total of 5,527,606 blog posts are collected, associated with 20,533 user profiles from 15 countries. From these, only those countries are retained which have more than 700 profiles (in order to have sufficient user and language diversity), resulting in a set of 12 countries. Table 4.1 shows the user profile and post distribution for each country. Note that it includes countries from various continents: Africa (Nigeria, South Africa), North America (US, Canada), Asia (India, Pakistan, Philippines, Singapore), Europe (UK, Ireland) and Oceania (Australia, New Zealand), encompassing very different language usage scenarios. *Gender.* From the same data, user profiles that provide gender information are identified, resulting in approximately 9,000 profiles authoring approximately 3 million posts, in which the data is approximately equally distributed between males and females. Table 4.2 provides more detailed statistics by gender.

Industry. Out of 39 industries originally represented in the data, those that had a minimum of 100 user profiles are selected, ultimately resulting in 15 industries. These have the least number of profiles in comparison to other demographics, namely country and gender, as (a) most profiles do not disclose their industry, and (b) many of the industries do not meet

GENDER	USER PROFILES	POSTS	AVG. POSTS	STD. DEV
Female	4,683	1,256,161	385	943
Male	4,051	1,560,314	268	510

Table 4.2: Blog data statistics for the selected genders (rounded to the nearest integer).

INDUSTRY	USER PROFILES	POSTS	AVG. POSTS	STD. DEV
Arts	670	221,083	330	626
Communications	499	217,032	435	900
Technology	330	96,500	293	858
Fashion	252	46,155	183	326
Internet	197	92,609	470	1,215
Business Services	193	66,891	347	1,205
Publishing	190	93,671	493	748
Non-Profit	187	75,634	404	1,709
Engineering	176	56,514	321	803
Consulting	157	57,741	368	922
Science	138	38,734	281	498
Marketing	136	65,697	483	1,112
Religion	123	38,460	313	571
Tourism	110	28,680	261	585
Advertising	102	37,204	365	738

Table 4.3: Blog data statistics for the selected 15 industries (rounded to the nearest integer).

the 100 user profile threshold. Table 4.3 shows the data statistics for the posts associated with the 15 industries.

For each demographic, blog posts are processed to remove: (1) HTML tags, (2) email IDs and URLs, (3) repeated characters, (4) very short posts (with fewer than ten words), (5) posts with more than 25% non-English words, and (6) posts without published times or other demographic information. The content is then lemmatized using Stanford CoreNLP [121], so that a more compact word index is generated.

For each demographic, candidate target word lists are created by identifying the top 500 words satisfying the following constraints: (1) they are the most frequent words within each demographic, (2) they should not be stopwords, modal or auxiliary verbs, (3) they should not contain numerals or special characters, and (4) they should have at least three characters. As a result, three sets of 500 target words are obtained, for the three demographics location, gender, and industry, respectively. For each demographic, these words are selected based on their frequencies in all the complementing demographic slices – they should have high frequencies in *all* the demographic slices, such that the corresponding word datasets contain large number of data instances. It is to be noted there exist 319 words that are common to all the four candidate word lists. Table 4.4 shows the top 8 target words that are obtained according to the above constraints in each demographic category.

LOCATION	GENDER	INDUSTRY$_{200}$	INDUSTRY$_{100}$
one	one	one	one
time	time	look	time
use	take	time	see
know	know	day	take
day	day	new	day
work	year	take	year
people	new	think	new
first	first	year	first

Table 4.4: Sample words with the highest frequencies for each demographic category.

4.3 Encoding Word Usage for each Demographic

Similar to that in Chapter 3, for each demographic, target word datasets are constructed, featuring a single word per dataset by culling usage examples from blog posts, where the class label is the demographic slice name. These posts are truncated such that they include a maximum of 100 words to the left and right of the target word, disregarding sentence boundaries. The target word datasets are balanced with respect to author and the time when the blog was posted. Further, the following heuristics are applied on the word datasets to obtained a fairer split of blog posts with respect to time: (1) for each user profile, at most 100 posts are chosen in a round-robin fashion across the various years, and (2) the maximum number of posts collected from each year is set to 15% of the total number of posts across all years. It is to be noted that the target word datasets are not balanced across topics, as I regard potentially different topic distributions as being reflective of the word usage variations across the slices within each demographic category (e.g., India may be naturally more interested in cricket than United States).

Table 4.5 shows the per-word average number of profiles and posts,[1] and their standard deviations, retained after processing, across the three demographic categories. For industry, two scenarios are considered: (1) industries with a minimum of 200 profiles (*Industry*$_{200}$: the top four industries from Table 4.3), and (2) industries with a minimum of 100 profiles (*Industry*$_{100}$: includes all the 15 industries).

[1] For each target word, equal number of posts are selected from each country, gender or industry with each demographic category; hence, each target word dataset is *class*-balanced.

DEMOGRAPHIC	USERS	POSTS	AVG. POSTS PER CLASS
Location	6,684 ± 967	49,848 ± 24,583	4,151 ± 2,049
Gender	5,820 ± 825	146,215 ± 50,239	54,969 ± 25,119
Industry_{200}	1,145 ± 181	9,288 ± 5,541	2,322 ± 1,385
Industry_{100}	2,180 ± 348	16,713 ± 7,665	1,114 ± 511

Table 4.5: Average number of profiles for the 500 target words for the three demographics of interest (rounded to the nearest integer).

For each word in each demographic, data instances are built using four types of features derived from the corresponding target word dataset. In addition to local, contextual and socio-linguistic features that used in Chapter 3, topical features are included to capture the topic diversity in word usage across complementing demographic slices. Also, syntactic features obtained from dependency tags of syntactic parse trees are omitted, as their performances on the Australia-versus-United States binary classification are lower than random chance by 4%.

Local features (Loc). The target word itself and five context words to the left and right of the target word are considered, to capture the immediately surrounding words of the target word. The part-of-speech tags of these words, and nouns and verbs before and after the target word, are ignored based on the redundancy of these features (the part-of-speech tags did contribute to statistical improvements).

Contextual features (Con). Similar to those in Section 3.3.1 in Chapter 3, these features are extracted from the entire contexts (100 words to the left and right) of the target word.

Socio-linguistic features (Soc). Similar to those in Section 3.3.1, these features are obtained from the full contexts of each target word.

Topic features (Topic). These features are included to capture the diversity in the topics that bloggers write about when employing the target words in their articles. They consist of the topic distributions learned over a word's dataset when considering latent topics extracted using LDA [118]. The LDA implementation included with the free Python library Gensim [122] is used to obtain the distributions. As typically done in topic modeling, the datasets of the target words are preprocessed, by removing (1) a standard list of stop words,[2] (2) words with very high frequency ($> 0.25\times$corpus_size (the total number of words in the word dataset)), and (3) words with frequency less than 5.

[2] Stop word list from English WordNet [119].

4.4 Learning Word Usage for a Given Demographic

The vectorial features derived according to Section 4.3 are employed to train classifiers, and ultimately examine the word usage differences across the various demographic slices for each demographic category.

4.4.0.1 Experimental setup

The WEKA open source machine learning software toolkit[3] [116] is used for all the experiments. The word datasets are balanced across different demographic groups – for each target word, equal number of blog posts containing the given word are obtained from each demographic group. Five multi-class classifiers are considered: (1) Naive Bayes (NB), (2) Random Forest (RF), (3) Decision Tree (DT), (4) k-Nearest Neighbor (k-NN) with $k = 3$, and (5) AdaBoost (AB). Their performances are compared against the majority class baseline (BL), which assigns all instances to the class with the highest occurrence in the data. A significant increase in accuracy over the baseline for a given word suggests that the demographic slice to which a writer belongs can be automatically identified. This, in turn, is taken as an indication that there exist significant usage differences for that word among the considered demographic slices.

Similar to that in Chapter 3, all the results are obtained using ten-fold cross-validation on the word datasets, with the constraint that posts authored by same bloggers are not shared across training and test folds. All the feature types are extracted from the training and test sets separately, so as to avoid any information seen in the training data occur in the test data. Two-sample t-test is used to obtain statistical significance over the predictions made by one of our classifiers (NB, DT, RF, k-NN, AB) and the baseline classifier (BL). The results that are statistically significant with respect to the baseline ($p < 0.05$) are marked with *.

4.4.0.2 Results & Discussion

Table 4.6 shows the accuracies obtained with the five classifiers under consideration, averaged over 50 randomly selected words. The NB classifier performs consistently better than the others: for *location*, by 4.09% compared to the second best classifier k-NN, and for *industry*, by 2.07% for Industry$_{200}$ compared to the second best decision tree classifier and by 3.33% for Industry$_{100}$ compared to second best AdaBoost. The only drop in performance for NB is for *gender*, where some of the other classifiers exhibit a stronger predictive

[3] https://www.cs.waikato.ac.nz/ml/weka/.

ability by at most 0.94%, though this difference is not statistically significant. Since ensemble methods (RF and AB) work best for a small number of classes, it is expected that having only two classes in the case of gender caused a small uptick in performance for that category. As none of the other classifiers achieve top or second best results for more than one demographic category, the NB classifier is selected to be used for the remaining experiments in this article.

CLASSIFIER	LOCATION	GENDER	IND$_{200}$	IND$_{100}$
Avg NB	**17.18***	70.12	**40.28***	**14.06***
Avg DT	10.48	*70.69*	*38.21*	9.89
Avg RF	11.52	**71.06**	35.20	10.09
Avg k-NN	*13.08*	63.89	33.14	9.06
Avg AB	12.10	70.27	35.49	*10.71*
Diff.	4.09	-0.94	2.07	3.33
Avg BL	8.33	50.00	25.00	6.67

Table 4.6: Ten-fold cross-validation accuracies averaged over 50 words using all classifiers and all features. Row **Diff.** shows the difference in accuracy between NB and the best performing other classifiers. Overall best results are shown in boldface and second best in italics.

Figure 4.1 shows the average accuracies of the NB classifier over the set of 450 target words selected for each demographic category (excluding the words used for 50 words in Table 4.6 for model selection) using all feature types. The highest average improvement in accuracy with respect to the baseline is observed for gender (two classes) at 20.08%, while location (12 classes) experiences an average improvement of 8.66%. In the case of industry, for Industry$_{200}$ (4 classes), the improvement is 15.33%, while for Industry$_{100}$ (15 classes), the improvement is 7.13%.

Typically, while averaging over a series of results, some of the individual scores may be higher or lower compared to the baseline, yet upon averaging them, only overall improvements are noticed. In this case, each and every one of the target words exhibits a significantly higher NB accuracy over the majority class baseline. For location, these range from 11.71% for the word *Monday* to 4.1% for the word *hand*, for industry, from 9.18% for the word *create* to 3.69% for the word *movie*, and for gender, from 26.3% for the word *product* to 8.45% for the word *government*.

These results indicate that there are indeed differences in the ways bloggers from the various demographic slices employ the target words.

Based on the confusion matrices, the various slices within each demographic category can be clustered based on which locations, industries, or genders are confused with each

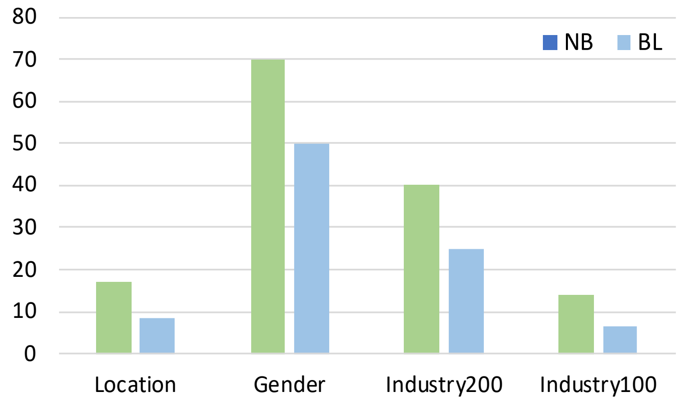


Figure 4.1: Ten-fold cross-validation accuracies averaged over 450 target words when using all features.

other by the classification model. In that respect, each demographic category has specific slices that the model often defaults to, or in other words, there are generic patterns based on the four feature types, that cause the model to get confused.

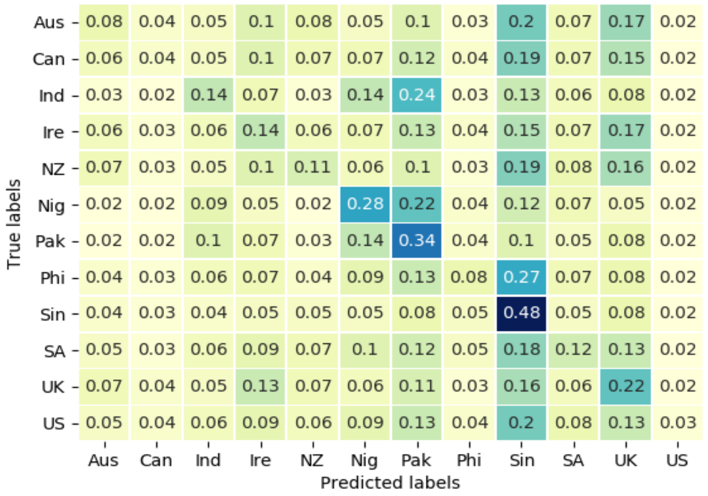


Figure 4.2: Confusion matrix for locations.

Figure 4.2 shows the confusion matrix for location, obtained by aggregating the confusion matrices of all the corresponding target words. From the figure, it can be seen that Australia, Canada, New Zealand, Philippines, South Africa and United States often get confused with Singapore, while India gets confused with Pakistan, and Ireland with United Kingdom. While Singapore is over-predicted for six slices, United States is predicted the minimum number of times. In other words, in comparison to the other countries, Singapore displays the most distinct word usage, while the United States showcases the most generic word usage.

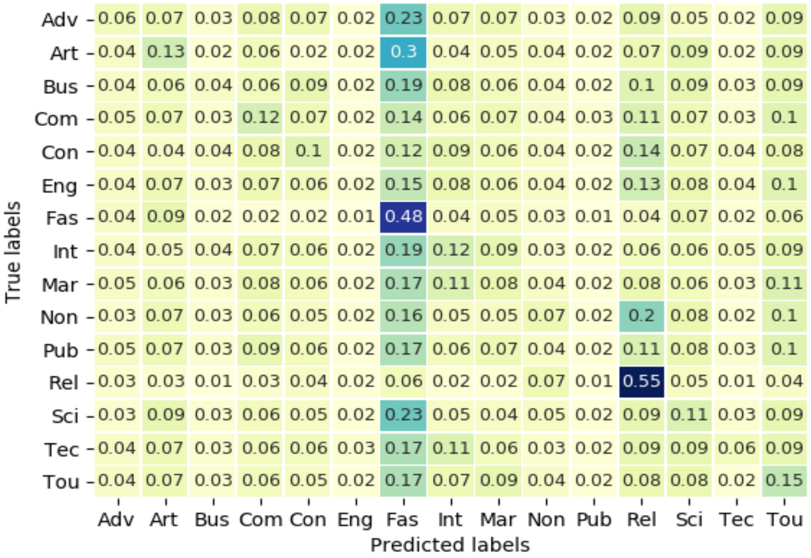


Figure 4.3: Confusion matrix for industries.

Figure 4.3 shows the confusion matrix for industry obtained by aggregating the confusion matrices of all the corresponding target words. Most industries (12 out of 15) often get confused with Fashion, with the exception of consulting, religion and non-profit. While fashion is over-predicted for most of the industries, engineering is the industry that is predicted the minimum number of times, despite not being the most sparsely represented. Hence, fashion follows the most distinct word usage, while engineering employs the most generic word usage across all industries considered.

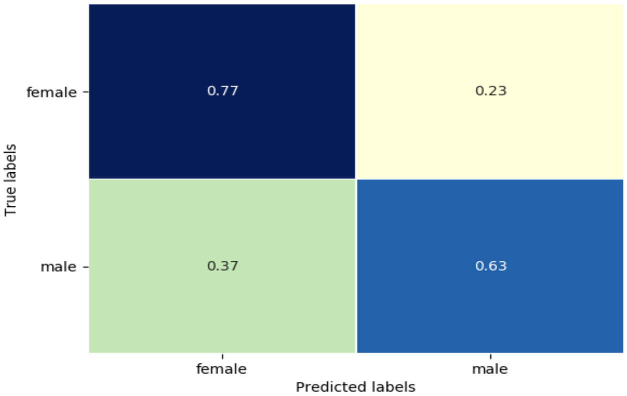


Figure 4.4: Confusion matrix for genders.

In the case of gender (Figure 4.4), since the classifier accuracies are greater than 50% and there are only two classes, the majority of instances are classified correctly. One interesting note is that female word usage offers a stronger signal compared to male word usage (that males are more likely to be confused by the classifier with females (0.37), than females with males (0.23)).

As seen from the Figures 4.1 through 4.4, words do exhibit differences in their usage across various countries, genders and industries, with some demographic slices being more similar to each other, while other slices exhibiting stronger differences. In the next section, analyses similar to those in Chapter 3 are conducted to examine which factors contribute the most to these usage differences, by focusing on the various features used, both from qualitative and quantitative perspectives.

4.5 What factors best encode usage differences across demographic categories?

A closer look at each demographic category is taken to analyze the extent to which the various linguistic features contribute to word usage differences. For each target word, the differing usage patterns are aimed to be studied based on topics and word classes across the various slices in each demographic category. For this, I perform (1) feature analysis and (2) qualitative and quantitative analyses of the various topics and word classes appearing in the surrounding contexts of target words.

4.5.1 Feature Analysis

First, an ablation study is conducted to examine how the individual feature types perform by themselves in comparison to the baseline, as well as when using the entire feature spectrum. Then, a closer investigation is performed to study the impact of context size on classification performance, and underline what parameters work better for a given demographic. Since socio-linguistic signals encompass several sources (such as LIWC or OpinionFinder), the modeling ability of each source is examined, to identify the factors that allow usage differences to emerge across demographics.

4.5.1.1 Performance of Individual Feature Types

To assess the encoding ability of each feature type with respect to word usage differences, the word models are retrained using each of the four feature types introduced in Section

4.3.

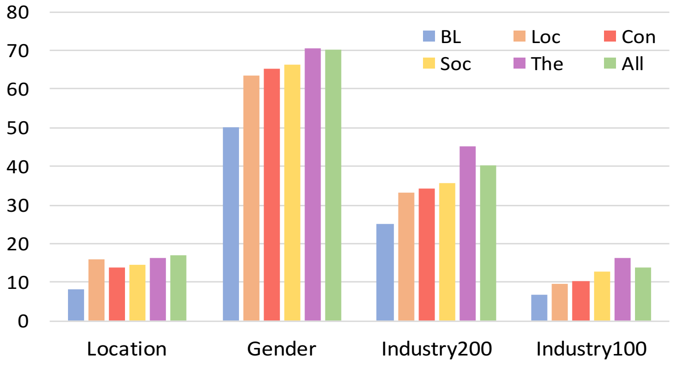


Figure 4.5: Ten-fold cross-validation feature ablation NB accuracies averaged over 450 target words with local (Loc), contextual (Con), socio-linguistic (Soc) and topical (The) features.

Figure 4.5 shows the accuracies of the NB classifier for each feature type, as well as for all features. The values are averaged over the individual accuracies pertaining to the 500 target words in each demographic category. BL represents the majority class baseline. It can be seen that for every demographic, there is an increasing trend. The baseline has the lowest accuracy, which every single one of our feature types surpasses. Coincidentally, the order in which the features are introduced also matches their discerning ability as it pertains to word usage, with local features being the worst performing (with the sole exception of *location*), contextual features being the next one up, socio-linguistic features achieving an even higher accuracy after that, and culminating with topic-based features, which display a striking and resilient performance. This implies that authors use words that are associated with different topics across demographics.

Interestingly enough, the information that each of these feature types encodes does not seem to be orthogonal, as combining all feature types result in lower accuracies for all demographics with the exception of *location*; the latter seems to be boosted by the interaction between the local and topic based features, in order to achieve an accuracy that is higher than that achievable by topic features alone, by a small margin of 0.7% in accuracy. It can be said that local features are able to perform well for the location demographic as they are able to capture regional usage for words, while for gender and industry, the discriminating power (at least using a context of 5 words to the left and right of the target word) is not sufficient to surpass that using features extracted from contexts of 100 words to the left and right of the target word.

Comparison with baseline. With respect to the baseline, improvements of 7.97% for location, 20.72% for gender, 20.07% for industry with 200 users, and 9.44% for industry with 100 users, are obtained using topic-based features alone.

Comparison with second best. The topic-derived features are also compared against any of the second best performing feature types. In the case of location, it is 0.07% better than the second best single feature type (local). For the remaining demographics, socio-linguistic features form the second best single feature type. Topic-based features for gender are better by 4.3%, and for industry they are 3.6% better with 200 users, and 9.58% better with 100 users.

To examine the effect of context size on the individual feature type performances, NB classifiers are trained with varying context lengths ranging from 5 to 100 words on the left and right of each target word. Figure 4.10 shows the results for each demographic category. Overall, increasing the context length in general increases the performance in the case of local, contextual and socio-linguistic features for all the demographic categories. However, it is interesting to note that in the case of topic features, the performance is highest when only five context words are used for *location* and *industry*. This generates instances in the dataset that are very sparse (as the topics are extracted cumulatively from a narrow window of context), and are easy for the classifier to assign to a given class, even in the case of 15 categories in the Industry_{100} dataset.

For *gender*, however, sparsely populated topic-based instances do not offer sufficient discriminative power, even though binary classification is being considered; gender seems to be best represented by the interplay of topic features extracted from lengthy contexts. In the gender experiments, by increasing the window for topic extraction from 5 to 100, the accuracy improves from 53% to 71%. In conclusion, for *location* and *industry*, the topics extracted from the immediate context surrounding a target word are the best predictors of word usage differences. The lengthier the context, the more degradation in signal is observed, with the steepest drop in performance occurring when increasing the surrounding context from 5 to 20 words left and right of the target word. For *gender*, the trend is exactly opposite, with the steepest increase in performance observed between 5 and 20 context words, and ultimately reaching the peak at 100 words. This implies that location and industry are best predicted by word *meaning*, namely the sense that a word takes given the topics that co-occur with it.

A word's neighboring context (when 5 words to the left and right are considered) usually captures the function of the word, and not its meaning. For this reason, for *location*, the prediction accuracy increases from 16% for local features to 27% for topic features for the same context of 5. For *industry*, the trend is even steeper: for Industry_{200}, local features have a prediction accuracy of 33% compared to 57% for topic features, and for Industry_{100}, local features have a prediction accuracy of 9% compared to 32% for topic features. For

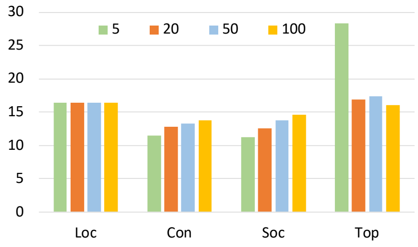


Figure 4.6: Location

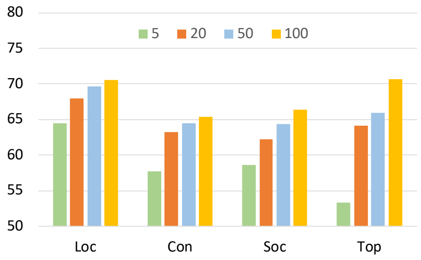


Figure 4.7: Gender

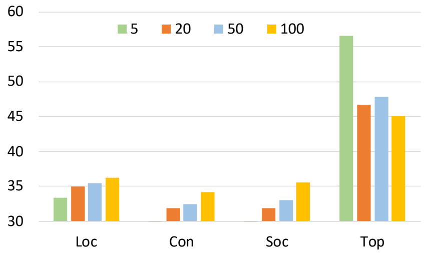


Figure 4.8: Industry$_{200}$

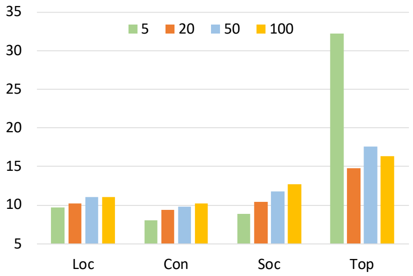


Figure 4.9: Industry$_{100}$

Figure 4.10: NB accuracies for each feature type with varying context lengths.

gender, the word itself is not used with a markedly different sense by males and females, rather what generally males and females talk about allows us to identify whether the word appears in the male or female writings. For this reason, local features at a context of 100 display the same performance as topic features extracted from the same context length (with an accuracy of 71%); it is to be noted that the number of features is extremely large in the first case, generating sparse instances, as the entire vocabulary of 100 words to the left and right of the target word is considered during classification, while for topic features, a dense 10 feature space is achieved for every instance.

4.5.1.2 Socio-linguistic Features Analysis

Figure 4.11 shows the classification accuracies obtained when using each socio-linguistic lexicon that was included under the umbrella of socio-linguistic features, namely LIWC, OpinionFinder (OF), Morality (ML), and WordNet Affect (WNA), as well as the four lexicons combined (Soc).

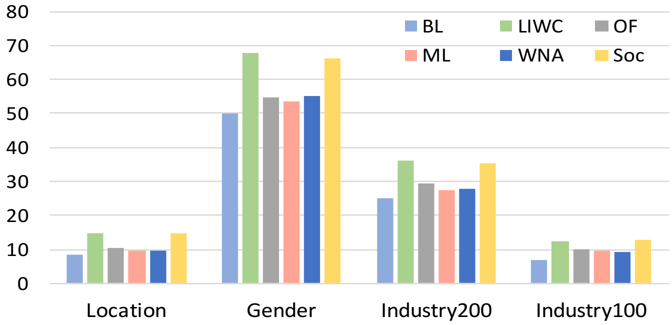


Figure 4.11: Ten-fold cross-validation NB accuracies averaged over 450 target words using LIWC, OpinionFinder, Morality Lexcion, and WordNet Affect.

The accuracies in the figure suggest that LIWC features contribute the most to the overall socio-linguistic classification performances, driving the accuracies when classifying over the entire feature set. OpinionFinder, Morality, and WordNet Affect, all have significantly lower accuracies. While including the other socio-linguistic features does not negatively impact the overall accuracy, there are slight drops in accuracy for location (0.13%), gender (1.21%) and industry$_{200}$ (0.59%). This implies that using LIWC features alone is sufficient to obtain the best results for the socio-linguistic feature type across all the demographics.

Accordingly, another set of experiments are conducted, where the overall classification performance is examined over all the feature types combined (local, context, socio-linguistic and topic), but this time replacing socio-linguistic features with LIWC features alone. Figure 4.12 has the results. It shows that there is no impact when simplifying the methodology and including LIWC features alone to model socio-linguistic aspects, as the overall accuracies remain the same.

By pairing the findings pertaining to context length (Figure 4.10) with the accuracies observed per feature type and overall (Figure 4.5), it can be concluded that the strongest

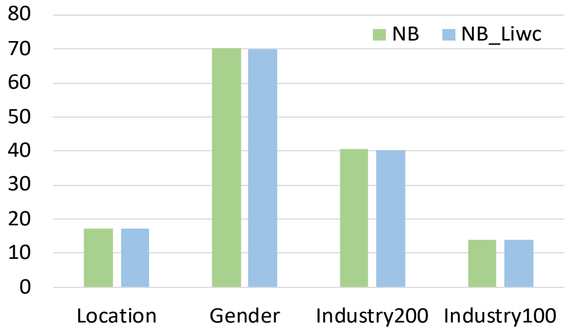


Figure 4.12: Ten-fold cross-validation NB accuracies averaged over 450 target words using local, contextual, LIWC and topical features together.

signal in differentiating word usages across demographics comes from topic modeling, and that a short window of 5 surrounding words to extract the latent topics is optimal for differentiating across *location* (accuracy of 28% compared to BL of 7%) and *industry* (for 4 industries: accuracy of 57% compared to BL of 25%, and for 15 industries: accuracy of 33% compared to BL of 8%). A wider window is necessary to extract topics that encode differences across gender usage, with 100 words to the left and right of the target word being optimal. For *gender*, the accuracy using topics is 70% compared to the baseline of 50%. To conclude, it can be emphasized that for every demographic category, at least a 20% improvement in accuracy can be achieved compared to the baseline, which is a notable performance improvement in itself, and even more commanding when bearing in mind that this improvement is achieved over demographic categories such as location with 12 different classes or industry with 15 different classes (Industry$_{100}$).

Given that the performances of the models are driven by topic-based features, and that these are quite interpretable, in the next section, the focus will be on a quantitative and qualitative analysis exploring the topic model's ability to encode usage differences across demographics.

4.5.2 Analyses

As topic features contribute the most to identifying usage differences, further investigations are conducted to examine these differences for a given target word, by analyzing the various latent topics that are covered in the contexts associated with a word's occurrences by bloggers from various demographic slices. Since examining the cross-demographic topics for all the 500 target word datasets qualitatively is tedious, the top 30 words that have the most significant improvements over the BL are considered for each demographic. These can be considered as words with *strong demographic bias* in their use.

Since the focus here is to aggregate the latent topics employed by the bloggers to explore differences over the various demographic slices, for each instance in the word dataset, the topic that has the highest preponderance in the topic distribution predicted by LDA is determined. This is called *the dominating topic* in this work. The instances and their dominating topic are then tallied across each slice, to obtain a word-centered topic distribution per slice. Using this experimental setup, the quantitative and qualitative analyses are performed below. Since these analyses are carried out in view of the dominating topic, the original 100 word context is used that was employed in all the analyses other than those pertaining to context length.

4.5.2.1 Quantitative Analysis

To gauge how similar or different the various demographic slices are with respect to word usage, the Pearson correlation is computed between the topic distributions for each pair of demographic slices (i.e. all combinations of 2 slices taken from the total number of slices in a given demographic) for the 30 top target words. Hence, when two slices have a high correlation, they are more similar, while when they have a low correlation, they are more dissimilar in terms of the primary topic with which a word is associated.

Location Table 4.7 shows the pair-wise Pearson correlation for the top 30 words with a strong demographic bias for every pair of demographic locations. Australia and Nigeria are the least correlated, while the highest correlation is observed between Australia and New Zealand. It should be emphasized that this trend holds even among the 30 target words which have the least NB improvements over the baseline. While a rushed conclusion may be that the disparities are caused by differences in the topics covered by a demographic slice, and not differences in topics as they are associated with word usage, Section 4.5.3 shows that the randomly sampled words are associated with a plethora of dominant topics for the same slice. Also, from the earlier analysis regarding context length (see Section 4.5.1.1), it is known that the shorter the window from which the topics are extracted for *location* and *industry*, the stronger is the discriminative power of the model.

India, Nigeria, and Pakistan all have negative correlations with 9 other countries, and positive correlations with each other, indicating that word usage differences are quite minor among them. This could be because India and Pakistan were a single country until 1947, so their English usage is quite similar; nonetheless, for India, the correlation is higher with Nigeria (at 0.66), compared to that with Pakistan (at 0.63), while for Pakistan, the correlation is higher with India (at 0.63), compared to that with Nigeria (at 0.61). While there are many people of Indian and Pakistani ethnicity living in Nigeria, this can only partially explain the strong correlation between word usage in an African country compared to the South Asian countries. What is even more interesting is that Nigeria experiences a negative correlation in word usage with South Africa (at -0.23), both being ex-African colonies that were part of the British Empire.

Looking at the 30 words with the least location bias, Philippines, South Africa, and United States have only positive correlations with the other countries, and hence do not differ from them in their distribution patterns, while Pakistan differs from the highest number of countries. The demographic locations can be grouped into the following clusters: (1) India, Pakistan and Nigeria, (2) Australia, Canada, New Zealand, United Kingdom, and Ireland.

	AU	**CA**	**IN**	**IE**	**NZ**	**NG**	**PK**	**PH**	**SG**	**ZA**	**GB**	**US**
AU	1	0.77	-0.50	0.53	**0.89**	*-0.55*	-0.52	0.54	0.40	0.63	0.81	0.55
CA	0.77	1	-0.45	0.69	**0.80**	-0.39	*-0.47*	0.51	0.26	0.52	0.73	0.77
IN	*-0.50*	-0.45	1	-0.49	-0.47	**0.66**	0.63	-0.17	-0.28	-0.06	*-0.50*	-0.20
IE	0.53	0.69	-0.49	1	0.60	-0.27	-0.35	0.39	0.02	0.37	**0.74**	0.57
NZ	**0.89**	0.80	-0.47	0.60	1	-0.53	*-0.57*	0.51	0.35	0.69	0.75	0.65
NG	*-0.55*	-0.39	**0.66**	-0.27	-0.53	1	0.61	-0.29	-0.32	-0.23	-0.43	-0.19
PK	-0.52	-0.47	**0.63**	-0.35	*-0.57*	0.61	1	-0.31	-0.40	-0.38	-0.43	-0.38
PH	0.54	0.51	-0.17	0.39	0.51	-0.29	*-0.31*	1	**0.60**	0.42	0.36	0.48
SG	0.40	0.26	-0.28	0.02	0.35	-0.32	*-0.40*	**0.60**	1	0.18	0.13	0.33
ZA	**0.63**	0.52	-0.06	0.37	0.69	-0.23	*-0.38*	0.42	0.18	1	0.49	0.44
GB	**0.81**	0.73	*-0.50*	0.74	0.75	-0.43	-0.43	0.36	0.13	0.49	1	0.50
US	0.55	**0.77**	-0.20	0.57	0.65	-0.19	*-0.38*	0.48	0.33	0.44	0.50	1

Table 4.7: Overall pair-wise Pearson correlations between the dominating topic distributions for the top 30 words with the highest demographic bias for location. The country codes can be found here: http://www.nationsonline.org/oneworld/country_code_list.htm. Correlations less than 0.6 are light gray in color, while the highest correlations in each row are shown in boldface and the least are shown in italics.

Gender The pair-wise Pearson correlation for the top 30 words with demographic bias between the topic distributions for males and females is -0.75, which suggests that the topics of interest for the two genders are largely different. This is supported by the analyses based on context length from Section 4.5.1.1, where it is also seen that the combinations of topics extracted from lengthy contexts of 100 words are the best predictors of gender.

Industry Table 4.9 shows the pair-wise Pearson correlations for the top 30 words with a pronounced demographic bias between the topic distributions for every pair of industries.

In the case of industry with at least 100 users (15 industrial groups), among the 30 target words with the highest industrial usage bias, Internet and Technology industries have the highest pair-wise correlation among all pairs, while Arts and Consulting are the least correlated. Religion is negatively correlated with 9 other industries making it the most "different" industry from the majority of other industries. On the other hand, Publishing is positively correlated with 13 other industries, making it the most "similar" industry in comparison to the others.

Looking at the set of 30 words with the lowest industrial usage bias, Publishing and Communications or Media have the highest pair-wise correlation, while Arts and Consulting are the least correlated with each other. Similar to the case with words with high industrial usage bias, Religion is negatively correlated with the majority of the other industries, while Engineering and Publishing are closer in word usages to the other industries. Some of these industrial groups can be clustered as: (1) Arts, Fashion and Science, (2) Engineering, Internet, Technology, Business Services and Consulting, (3) Religion and Non-Profit, (4) Publishing and Communications or Media.

Industry with at least 200 users offers a more coarse-grained view, taking only 4 industrial groups into account. Table 4.8 shows the Pearson correlation between all possible pairs of these 4 industries. Fashion and Art show no correlation of latent topics with respect to each other, at -0.01, while Art is negatively correlated with Technology (-0.49), and Fashion with Communication and Technology (-0.47, -0.47, respectively). Communication and Technology display the strongest correlation of topics across all groups that are explored (0.21).

	Art	**Com**	**Fas**	**Tec**
Art	1	-0.06	-0.01	-0.49
Com	-0.06	1	-0.47	0.21
Fas	-0.01	-0.47	1	-0.47
Tec	-0.49	0.21	-0.47	1

Table 4.8: Overall pair-wise Pearson correlation for the top 30 words with the highest demographic bias for industry with 200 users.

4.5.3 Qualitative Analysis

To get a qualitative overview of the usage differences in terms of the overall topics across various demographic slices, the dominating topics are listed in each demographic slice for a set of words randomly chosen from the top 30 words with demographic bias. For each

	Adv	**Art**	**Bus**	**Com**	**Con**	**Eng**	**Fas**	**Int**	**Mar**	**Non**	**Pub**	**Rel**	**Sci**	**Tec**	**Tou**
Adv	1	0.37	0.14	**0.51**	0.07	-0.07	0.45	0.09	0.32	-0.15	0.46	-0.18	0.08	-0.05	-0.12
Art	0.37	1	-0.34	0	-0.7	-0.34	**0.72**	-0.16	0.09	-0.17	0.18	-0.18	0.48	-0.26	0.05
Bus	0.14	-0.34	1	-0.01	**0.67**	0.22	-0.04	0.13	0.06	-0.03	-0.13	-0.24	0.1	0.23	-0.09
Com	0.51	0	-0.01	1	0.28	0.24	-0.26	-0.01	0.28	0.24	**0.85**	-0.06	-0.2	-0.1	0
Con	0.07	-0.7	**0.67**	0.28	1	0.43	-0.5	0.13	0.07	0.19	0.11	0.11	-0.34	0.27	-0.18
Eng	-0.07	-0.34	0.22	0.24	0.43	1	-0.48	0.34	0.09	0.39	0.17	0.23	-0.24	**0.56**	0.21
Fas	0.45	**0.72**	-0.04	-0.26	-0.5	-0.48	1	-0.11	-0.04	-0.32	-0.19	-0.21	0.53	-0.22	-0.08
Int	0.09	-0.16	0.13	-0.01	0.13	0.34	-0.11	1	0.6	-0.52	-0.02	-0.4	-0.26	**0.87**	-0.04
Mar	0.32	0.09	0.06	0.28	0.07	0.09	-0.04	**0.6**	1	-0.4	0.32	-0.44	-0.18	0.37	0.26
Non	-0.15	-0.17	-0.03	0.24	0.19	0.39	-0.32	-0.52	-0.4	1	0.23	**0.67**	0.05	-0.35	0.16
Pub	0.46	0.18	-0.13	**0.85**	0.11	0.17	-0.19	-0.02	0.32	0.23	1	-0.01	-0.04	-0.1	-0.03
Rel	-0.18	-0.18	-0.24	-0.06	0.11	0.23	-0.21	-0.4	-0.44	**0.67**	-0.01	1	-0.2	-0.22	-0.24
Sci	0.08	0.48	0.1	-0.2	-0.34	-0.24	**0.53**	-0.26	-0.18	0.05	-0.04	-0.2	1	-0.26	0.11
Tec	-0.05	-0.26	0.23	-0.1	0.27	**0.56**	-0.22	0.87	0.37	-0.35	-0.1	-0.22	-0.26	1	-0.08
Tou	-0.12	0.05	-0.09	0	-0.18	0.21	-0.08	-0.04	**0.26**	0.16	-0.03	-0.24	0.11	-0.08	1

Table 4.9: Overall Pearson correlation for the top 10 target words with the highest demographic bias for industry with 100 users. Correlations less than 0.2 are light gray in color, while the highest correlations in each row are shown in boldface and the least are shown in italics

word, in each demographic slice, labels are associated to the hidden topics based on the corresponding top words associated with each topic. In most cases, these labels are picked from the words associated with the topics, while in other cases, they are given based on the overall themes in the topics. In addition to the assigned labels, three words representative of the labels are mentioned for each topic, excluding the target word itself, in case it happens to occur among them.

Table 4.10 shows the dominating topics for seven target words randomly chosen from the top words with demographic bias across the 12 locations under consideration. For instance, the word *national* is dominantly used to describe (1) nature parks by the bloggers from Australia, New Zealand and South Africa, (2) live shows by those from Canada, (3) health systems from India, (4) sports teams from Ireland and Nigeria, (5) art museums from Philippines, and (6) security from Pakistan. Another interesting example is the word *Wednesday*. Australians and Canadians talk about Wednesdays to post blogs and photos, while Indians think about market shares and companies in relation to Wednesdays. It is used to describe evening shows and events by those in Ireland, school work by those in New Zealand, politics and news reports by those in Nigeria and Pakistan respectively, friends by those in Philippines, Singapore, United Kingdom and United States, and God by South Africans.

For gender, Table 4.11 shows the dominating topics for the seven randomly chosen target words from those with the highest gender bias across genders. Consider the word *party* for instance. While females predominantly use it for celebrating birthdays and Christmas with their families, males tend to talk about it to describe the various political parties and elections. Another example is the word *oil*. Females use this word largely for cooking discussions, while males use it to talk about political issues such as wars.

The dominating topics for seven randomly picked target words with industrial bias are listed in Table 4.12. For instance, the word *result* is used in terms of (1) research by the writers from Business Services and Consulting, (2) elections by those from Communications or Media, Engineering and Publishing, (3) search by those from Internet and Technology, and (4) medical test by those from Marketing and Tourism. Another interesting example is the word *follow*. It is used to describe (1) game shows and films by the bloggers from Advertising and Marketing, (2) company plans and money by those from Business Services and Consulting, (3) political parties and members by those from Communications or Media and Publishing, (4) online posts and tweets by those from Engineering, Internet and Technology, and (5) God by those from Religion and Non-Profit.

COUNTRY	TARGET WORD						
	group	national	wednesday	south	town	market	store
IN	BUSINESS (company, business, service)	HEALTH (bank, health, system)	MARKET (company, market, share)	WAR (war, india, pakistan)	PROJECT (city, business, project)	STOCK (price, stock, bank)	APP (phone, app, free)
IE	THEATER (show, music, art)	TEAM (year, team, win)	SHOW (pm, show, event)	SHOW (show, film, art)	FILM (show, play, game, film)	BUY (product, buy, wear)	PRODUCT (buy, product, price)
NZ	WORK (group, work, school)	PARK (park, area, walk)	SCHOOL (school, week, work)	DIRECTION (north, road, east)	CITY (city, road, park)	(city, shop, visit)	(place, house, visit)
NG	POLITICS (state, party, government)	TEAM (year, team, win)	POLITICS (state, party, meeting)	WAR (war, india, pakistan)	ATTACK (police, kill, attack)	(state, public, party)	NEWS (news, school, fire)
PK	POLITICS (state, party, government)	SECURITY (people, security, war)	NEWS (news, police, report)	WAR (war, india, pakistan)	ATTACK (police, kill, attack)	PHONE (game, phone, design)	APP (phone, app, free)
SG	FRIENDS (people, friend, life)	(day, time, good)	FRIEND (good, friend, night)	KOREA (korea, asia, market)	(shop, food, buy)	FOOD (Food, buy, farmer)	(work, buy, start)
ZA	(back, walk, head)	PARK (park, area, walk)	GOD (god, church, walk)	FOOD (food, eat, indian)	(photo, picture, dress)	(city, shop, visit)	PRODUCT (buy, product, price)
GB	(back, walk, head)	(day, time, good)	FRIEND (good, friend, night)	DIRECTION (north, road, east)	(photo, picture, dress)	(city, shop, visit)	(place, house, visit)

Table 4.10: Seven sample words with significant usage difference among the various countries.

GENDER	TARGET WORD						
	product	oil	card	cut	party	box	purchase
Female	BEAUTY (skin, face, beauty, eye)	COOKING (olive, onion, garlic, cook)	DESIGN (stamp, cut, image)	PAPER (card, paper, piece)	CELEBRATION (birthday, family, christmas)	GIFT (card, paper, gift)	SHOP (shop, wear, buy)
Male	TRADE (market, increase, trade)	POLITICAL (government, national, war)	ELECTION (report, party, office)	TAX (pay, tax, price)	POLITICAL (state, political, election)	GAME (film, game, play)	STOCK (market, company, stock)

Table 4.11: Seven sample words with significant usage difference between male and females.

4.6 Conclusions

This chapter extends the previous work on identifying and understanding demographic differences in word usage between people belonging to various demographic groups. Specifically, it focuses on three demographics namely, location, gender and industry, and analyzes blog posts written by people from 12 countries, 2 genders and 15 industries. 500 target words are selected for each of the three demographics based on their occurrence frequency in the blog posts, and studies are carried out to understand the demographic differences in the usage of these words. For that, classifiers are developed for each of the selected target words based on linguistic features including local and global contexts, and psycholinguistic and topic-based word classes. The classification performances for all demographics are at least 20% higher than the baseline, and topic features are best in predicting how words are employed by the various demographic slices. It is also shown that the immediate context (of 5 words to the left and right of the target word) is best for extracting topics that differentiate between demographic slices for *location* and *industry*, as the meaning of the word can be directly encoded by very few topics, while for *gender*, the interplay of topics extracted from a wide context window of 100 words to the left and right is the best predictor.

Looking at the topical differences for the top target words with demographic biases, their dominating topics are examined in each demographic group resulting from topic modeling, both qualitatively and quantitatively. The correlations between the topic distributions for these groups suggest similarities and differences between the demographic groups, indicating that there are indeed differences in the ways various locations, genders and industries view the world.

The demographic word datasets used in the experiments reported in this paper are available at http://lit.eecs.umich.edu/downloads.html.

INDUSTRY	TARGET WORD						
	create	daily	remove	follow	service	provide	result
Adv	DESIGN (design, art, style)	REPORT (news, report, national)	GROUP (member, report, police, party)	SHOWS (show, game, film, win)	PUBLIC (state, public, government)	REPORT (country, report, security)	WORK (work, show, picture, design)
Bus	COMPANY (company, business, service)	MONEY (money, company, service)	GROUP (member, report, police, party)	PLAN (plan, company, money)	CUSTOMER (company, customer, product)	SUPPORT (support, project, customer)	RESEARCH (process, research, project)
Com	GROUP (state, group, law)	TEAM (year, team, win)	GROUP (member, report, police, party)	POLITICS (member, party, issue)	PUBLIC (state, public, government)	REPORT (country, report, security)	ELECTION (govt, million, election, report)
Con	COMPANY (company, business, service)	MONEY (money, company, service)	GROUP (member, report, police, party)	PLAN (plan, company, money)	CUSTOMER (company, customer, product)	SUPPORT (support, project, customer)	RESEARCH (process, research, project)
Int	ONLINE (blog, follow, website)	READ (read, post, facebook)	FILE (post, image, site)	ONLINE (post, link, twitter)	ACCESS (account, free, online, website)	INFORMATION (free, information, site)	SEARCH (post, search, google, site)
Mar	BOOK (book, story, write)	HEALTH (skin, food, body)	COOKING (add, water, oil, salt)	SHOWS (show, game, film, win)	CUSTOMER (company, customer, product)	STORY (event, story, write)	TEST (skin, test, hair, health)
Non	GOD (God, life, love)	GOD (god, life, give)	CLEARING (house, car, tree)	GOD (god, life, jesus)	SERVE (god, people, church)	LIFE (life, god, child)	RUN (home, run, hour)
Pub	BOOK (book, story, write)	REPORT (news, report, national)	GROUP (member, report, police, party)	POLITICS (member, party, issue)	PUBLIC (state, public, government)	STORY (event, story, write)	ELECTION (govt, million, election, report)
Rel	GOD (God, life, love)	GOD (god, life, give)	(people, god, life)	GOD (god, life, jesus, fat)	SERVE (god, people, church)	LIFE (life, god, child)	LIFE (life, god, child, family)
Tec	ONLINE (blog, follow, website)	MONEY (money, company, service)	FILE (post, image, site)	ONLINE (post, link, twitter)	ACCESS (account, free, online, website)	INFORMATION (free, information, site)	SEARCH (post, search, google, site)

Table 4.12: Seven sample words with significant usage difference among the various industries.

CHAPTER 5

Demographic-Aware Word Associations

5.1 Introduction

"Objects once experienced together tend to be associated in the imagination, so that when any one of them is thought of, the others are likely to be thought of also, in the same order of sequence or coexistence as before [123]."

Understanding the associations that are formed in the mind is paramount to understanding the way humans acquire language throughout a lifetime of learning [124, 125]. Furthermore, word associations are believed to mirror the mental model of the conceptual connections in a human mind, and constitute a direct path to assessing one's semantic knowledge [39, 40]. Psycholinguistic studies focused on word associations have been conducted since early 1900s, and the field is actively pursued in Psychology, with the latest study to my knowledge being a spiking neuron model of word association [126].

Word associations start forming early in life, as language is acquired, and one learns based on the environment where concepts lie in relation to each other. For example, one may learn to associate "mother" with "warmth," or "fire" with "burn." Yet, this mental model is not static but highly dynamic, and is shaped by new experiences over a lifetime. For instance, [41] showed that word associations change with age, and that for respondents in younger age groups their variability is lower, while for those in older age groups the variability is higher, as their life experiences modify the commonality between respondents from the same group.

Computational linguistics has traditionally taken the "one-size-fits-all" approach, with most models being agnostic to the background of the speakers behind the language. With the introduction and adoption of Web 2.0, there has been an exponential increase in the availability of digital user-centric data in the form of blogs, microblogs and other forms of online participation. Such data often times can be augmented with demographic or other user-focused attributes, whether these are user-provided (e.g., from a user's online profile), or labeled using an automatic system. This enables computational linguists to go beyond generic corpus-based metrics of word associations, and attempts to extract associations that pertain to given demographic groups that would not have been possible without administering time-consuming and resource-intensive word association surveys.

While current NLP methods generally deal with more advanced tasks (relation extraction, text similarity, etc.), at their very core many of these tasks assume some way of drawing connections (or associations) between words. Therefore, as a step toward demographic-aware NLP, I choose to work on the core task of "word association." The algorithms that are introduced can be immediately applied to demographic-aware word similarity, and with some minor changes to demographic-aware text similarity. Future stages could also include demographic-aware labeled associations, and more advanced applications such as information retrieval (which relies heavily on word associations/similarity), demographic-aware keyword extraction, dialogue personalization, and so forth. Note that a few other researchers have explored demographic-aware NLP models with promising results, primarily focusing on the use of demographics for various forms of text classification [31], or sentiment and subjectivity classification [32].

This chapter makes several main contributions. First, a novel dataset of demographic-aware word associations is created, consisting of approximately 300 stimulus words along with 800 responses per word collected from a demographically-diverse group of respondents, for a total of 228,800 responses. Removing spam responses results in 176,097 responses. Analyses that are performed on this dataset demonstrate that indeed word associations vary across user dimensions.[1] Second, it is shown that the associations obtained follow the same pattern as those elicited in traditional classroom surveys. Third, an evaluation metric is proposed that suits the free association norms task. Fourth, a demographic-aware model based on a skip-gram architecture is introduced, and through several comparative experiments, it is shown that their performance can surpass the performance attainable on demographic agnostic models.

I specifically focus on two demographic dimensions: location and gender. For location, India and United States (US) are considered, choice made primarily because these two countries have a large English-speaking population, represented both on social media and on crowdsourcing platforms.

[1] This work is not centered around comparing different word forms, as one would encounter for example in British English and American English, but rather around different word associations that people with a particular demographic characteristic are inclined to make, e.g., "health" in India is more strongly associated with "wealth", while in the United States it is more strongly associated with "sick."

5.2 Word Associations Dataset

Word association data collection typically consists of providing participants with a list of words, also known in the psycholinguistics literature as *stimulus words*, and asking them to provide the first word that comes to mind in response to each stimulus [127, 41]. For instance, given a stimulus word such as *cat*, one would expect answers such as *dog* or *mouse*. Earlier work on word associations administered the tests in classroom settings, with 100 words per survey, and the results were compiled into tables of norms of word associations [66, 39].

Since the goal is to explore the effect of demographics on word associations, a task is created on Amazon Mechanical Turk (AMT) that can reach a wide and demographically-diverse audience. The survey is structured into two sections: a word association survey, followed by a demographic survey. Given the online nature of the survey, and the aim to obtain a high quality dataset, each participant is presented with a set of 50 stimulus words at a time (instead of 100 as done in earlier studies). The demographic section consists of seven questions covering gender, age, location, occupation, ethnicity, education, and income.

Stimuli. The stimulus list consists of a set of approximately 300 words. Among these, 99 words are sourced from the word list proposed by Kent and Rosanoff [66] (*standard* list).[2] The remaining words are identified using the method for finding word-usage differences between two groups introduced in Chapters 3 and 4 [33], which relies on large collections of texts authored by the two groups to identify words that can be accurately classified by an automatic classifier as belonging to one group versus another. Using this method, 100 words are chosen as the top most different words between US and India (*culture* list), and another set of 100 words as the top most different words between male and female (*gender* list). The union of these three lists results in 286 stimulus words for which the word associations are collected. Examples are shown in Table 5.1.

Responses. The task was published separately for respondents from US and India, as AMT has an option of only presenting the survey to people from a preselected geographical location. Six different surveys, each including approximately 50 stimulus words, are administered for each region. The survey is conducted in English for both countries, noting that one of the official languages of India is English (alongside Hindi). Each survey also includes four spam-checking questions with previously known answers (e.g., *What is the*

[2] Note that this list originally included 100 words. The word "foot" was however misspelled in my survey, and instead gathered answers for "food."

color of the sky?, with five options *blue*, *red*, *pink*, *green*, *yellow*), which were used to filter out respondents who were filling out the survey without reading the questions.

For each set, 400 responses are collected per region, resulting in 800 responses for both US and India. After removing the respondents who did not pass the spam-checking questions, an average of 752 responses per word are left, which are then balanced by gender, to retain an equal number of Indian women, Indian men, US women, and US men. This results in 492 and 480 responses for the two sets of 50 *standard* stimulus words, 436 and 468 for the *culture* words, and 440 and 432 for the *gender* words. Similar to [68], all the responses are normalized (i.e. plural was mapped to singular, gerund to infinitive, etc.); in this case, the Stanford CoreNLP Lemmatizer [121] is used, ultimately aggregating the responses into a gold standard.

Table 5.1 shows the top associations for a few sample stimuli, as collected from India and US, and males and females. Finer-grained qualitative analyses also reveal interesting distinctions. For instance *bath* is overwhelmingly associated by men with *water*, while US women associate it with *bubble*, and Indian women with *soap*. US men make several food-related associations, such as "mutton" with "chop" (US women associate it with "lamb"), "comfort" with "food" (US women associate it with "zone"), "use" with "consume" (while US women associate it with "tool"), or "whole" with "food" (while US women associate it with "half"). Interestingly, US men seem to provide responses based on collocations, e.g., they answer *Kane* for *citizen* (citizen Kane), *weight* for *heavy* (heavyweight), or *lion* for *mountain* (mountain lion); on the contrary, women more often provide responses that consist of synonym or antonym words, e.g., *person* for *citizen*, *health* for *sick*, or *light* for *heavy*.

	GENDER		LOCATION	
WORD	MALE	FEMALE	INDIA	US
beautiful	girl, woman, pretty	pretty, girl, ugly	girl, nature, flower	pretty, girl, ugly
cheese	pizza, bread, milk	butter, mouse, pizza	pizza, butter, bread	cracker, swiss, cheddar
hard	soft, rock, work	soft, work, rock	work, stone, rock	soft, rock, time
health	good, wealth, care	good, wealth, sick	wealth, good, fitness	good, sick, care
range	distance, gun, shooting	gun, rover, mountain	price, rover, wide	gun, distance, rover
admit	hospital, guilt, card	hospital, confess, one	hospital, card, accept	guilt, one, confess
mix	tape, match, juice	cake, tape, stir	juice, tape, match	stir, tape, cake
organize	clean, arrange, party	clean, arrange, meeting	arrange, meeting, party	clean, sort, neat
stack	pile, book, box	book, pile, hay	book, queue, pile	pile, book, pancake

Table 5.1: Top three most frequent responses for sample stimulus words.

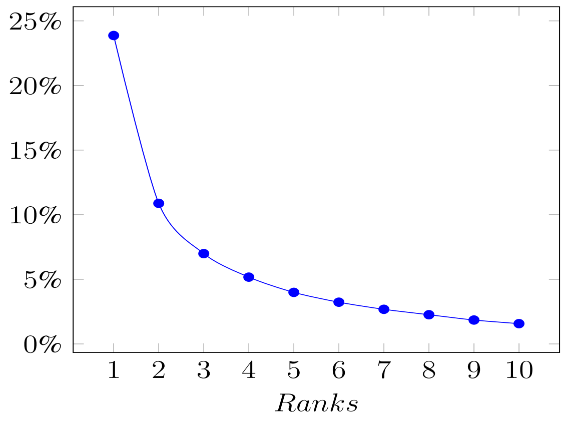


Figure 5.1: Primary response frequency (in percent) versus rank for the Standard word list.

For further insight, Table 5.2 shows the average number of unique responses obtained for a given stimulus word, with words with the least and highest variability.[3] The second column lists the correlations between the frequency of the primary response and the number of different responses, as also reported by [67]. This correlation is negative, as the more people agree on the primary response, the fewer overall unique answers for a stimulus word are provided. Additionally, Figure 5.1 shows the Zipfian distribution of average norm frequency; the most frequent response is given on an average by 24% of the respondents, while the third most frequent response is given by 7% of them.

DEMOGRAPHIC	AVERAGE	CORRELATION	LOWEST VARIABILITY	HIGHEST VARIABILITY
		STANDARD		
INDIA	60.88	-0.52	stove	city
US	51.19	-0.53	bath	trouble
MALE	61.63	-0.45	stove	city
FEMALE	56.75	-0.55	stove	city
		ALL		
INDIA	72.27	-0.59	stove	regardless
US	57.03	-0.56	east	basically
MALE	70.33	-0.52	stove	regardless
FEMALE	66.54	-0.59	east	respectively

Table 5.2: Average number of responses obtained for a given stimulus word, correlation between frequency of primary response and number of different responses, words exhibiting the least variability, and words with the highest variability.

[3] In several of the data analyses, in order to allow for a direct comparison with the word list from [66], in addition to showing statistics for the entire dataset (*All*), the statistics are shown separately compiled for the list from [66] (*Standard*).

Analyses of Demographic Variations. To model norm strength within a given demographic group or across groups, how often respondents from a group match the most frequent answer (*Primary*) or one of the most frequent ten answers for that group (*Top10*) is tabulated. That is, given the response for one stimulus word as provided by one held-out survey respondent at a time, it is determined whether that response matches the most frequent association of the *remaining* members of the same group (Table 5.3, *Primary* columns), or one of the top 10 associations pertaining to that same group (Table 5.3, *Top10* columns). Similarly, the match with the most frequent or the top 10 responses from the other group is measured, as shown in Table 5.4. As expected, the intra-group similarities are significantly higher than the inter-group similarities, which supports the hypothesis that different groups make different word associations, which tend to be more coherent within a group than across groups. While males and females have similar ranges for their agreement figures, it is noticed that on average US respondents have stronger intra-group agreements. Note also that inter-group similarities are asymmetrical, as multiple words may have the same association frequency for one group, yet for the complementary group that may not be the case.

As an additional analysis of demographic variations in the responses received, for each respondent, his / her demographic group is predicted using a majority vote conducted across all the user's responses using a simple rule-based system that assigns each response to the group having the highest frequency for that particular association. For instance, given the response *sun* obtained from a respondent for the stimulus *yellow*, the respondent is assigned to either India or US depending on the highest normalized frequency of the response *sun* for the same stimulus in each of those groups. A similar rule-based assignment is also used for gender. Thus, the response words and their normalized frequencies are computed based on the responses from 80% of the users chosen randomly, and accordingly the demographic groups for the remaining 20% of the users are predicted based on a decision, across the entire set of a users' responses. Table 5.5 shows the results of these predictions, which indicate high location variability (i.e., one can predict with high accuracy the location of a respondent), and medium gender variability.

5.3 Computational Models of Word Associations

First, a new model for measuring word associations is introduced that leverages a shallow neural net architecture to embed demographically-enriched words. Then, the performances of the predicted associations are compared to those resulting from other approaches, including traditional corpus-based measures such as mutual information or vector-space models,

Demographic	Standard		All	
	Primary	Top10	Primary	Top10
India-India	0.23	0.77	0.18	0.78
US-US	0.29	0.82	0.25	0.81
Male-Male	0.23	0.79	0.19	0.79
Female-Female	0.25	0.80	0.21	0.81

Table 5.3: Intra-group similarities (the higher the similarity, the more cohesive the group is).

Demographic	Standard		All	
	Primary	Top10	Primary	Top10
India-US	0.18	0.55	0.14	0.50
US-India	0.20	0.60	0.16	0.56
Male-Female	0.22	0.63	0.17	0.59
Female-Male	0.24	0.66	0.19	0.61

Table 5.4: Inter-group similarities (the higher the similarity, the less distinct the groups are).

as well as a recent distributional learning model with word embeddings. For each of these methods, generic associations (devoid of any demographic information) are predicted, evaluated, and compared to demographic-aware associations.

5.3.1 Composite Skip-gram Language Models

A new word association model is introduced, which relies on the skip-gram neural net architecture [128], and leverages its efficiency and ability to deal with less frequent words.

The skip-gram model tries to predict the context given a word, that is, for each word w_i in the input sequence $w_1, \ldots, w_T$, the model tries to predict w_{i-2}, w_{i-1}, w_{i+1} and w_{i+2}, assuming, for example, a sliding window of five words. Mathematically, the model maximizes the objective function

$$J = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=-c, j \neq 0}^{c} \log P(w_{i+j} | w_i) \tag{5.1}$$

where T is the number of tokens in the dataset, c is the number of context words on each side of the target word w_i, and $P(w_{i+j}|w_i)$ is the probability to observe word w_{i+j} in the context of word w_i.

To make this model demographic-aware, two variations are proposed, which are refered to as *composite skip-gram language models* ($C - SGLM$). In the first variation ($EMB1$), the target word w_i is tagged with a demographic label L (e.g., gender). For example, con-

DEMOGRAPHIC	STANDARD	ALL
GENDER	0.60	0.56
LOCATION	0.94	0.94

Table 5.5: Predictions based on similarity to group.

sider the female-authored text: "If your baby prefers warm formula or milk place a filled bottle in a bowl of warm water and let it stand for a few minutes." For the target word "formula$^{L=\text{female}}$" a high probability is to be predicted for "baby" and "milk" occurring in the neighboring context. The underlying reasoning is that tagged words that appear in similar contexts will be nudged toward each other, while those that do not, will further distance themselves. This allows discrepancies to emerge between how the words are embedded given a demographic dimension.

In the second variation ($EMB2$), the demographic label is also introduced in the context. That is, for each skip-gram $(c_{i,\text{left}}, w_i, c_{i,\text{right}})$ the following three skip-grams are generated:

$$\begin{aligned}
&(c_{i,\text{left}}^{\text{label}}, \quad w_i, \quad c_{i,\text{right}}) \\
&(c_{i,\text{left}}, \quad w_i^{\text{label}}, \quad c_{i,\text{right}}) \\
&(c_{i,\text{left}}, \quad w_i, \quad c_{i,\text{right}}^{\text{label}})
\end{aligned} \tag{5.2}$$

The two models seek to capture different scenarios. In the first model, where the demographic label is added to only the target word, the embedding of the labeled word is optimized with respect to the generic embedding of the context. In the second model, the optimization is rather symmetric, allowing tagged and generic embeddings to influence each other. Thus, the optimization function seeks to predict both tagged and untagged words in the vicinity given a target word, instead of only focusing on predicting untagged words like that in $EMB1$. The embeddings resulting from such a model are expected to allow for more accurate representations across the tagged and untagged vocabulary, where for example the word "mother" uttered by a female would be close to the word "mother" (regardless of author gender).

In both scenarios, the embedding space accommodates both tagged and untagged words at the same time, being very computationally robust, and allowing comparisons across the tagged version of words, as well as between generic words and their tagged surrogates. For both variations, the cosine similarity is computed between the embeddings of the stimulus word and each of the vocabulary words (whether generic or demographic-enhanced), and the closest unique candidates are retained (after dropping their demographic tag).

5.3.2 Other Word Association Models

Mutual Information (MI). First, the information theoretic measure, proposed by Church and Hill [38], is implemented. It is defined as follows:

$$I(x, y) = log_2 \frac{P(x, y)}{P(x)P(y)} \tag{5.3}$$

This measure compares the probability of observing words x and y together (the joint probability) with the probabilities of observing x and y independently. The joint probability, $P(x, y)$, is generally estimated by counting the number of times x is followed by y in a window of w words, and normalizing this count with the size of the corpus. Similar to [38], the window size w is set to five, as it is large enough to capture verb-argument constraints, and not so large to restrict to strict adjacency. For a given stimulus word, (1) the entire corpus is used to compute the generic MI word associations with the rest of the vocabulary, and get the top associations according to their MI scores; and (2) the section of the corpus obtained for a given demographic is used to determine the top demographic-aware MI word associations.

Vector-Space Model (VSM). The next implementation is of the traditional vector-space model, where each word is represented by a $tf.idf$ weighted vector inside the term-document matrix (representing term occurrences inside the documents in the corpus), with a length equal to the number of documents D in the corpus [129]. For a given stimulus word, cosine similarities are computed with all the remaining word vectors in the vocabulary, and those words having the highest similarity are considered as the top responses. Similar to MI, all the documents in the corpus are used to produce generic word associations, while only those documents pertaining to a specific demographic value are utilized to derive demographic-aware associations.

Skip-Gram Language Model. Finally, the distributional representation technique of word embeddings ($SGLM$) proposed by Bamman et al. [79] is used. Specifically, information about the speaker (geography, in their case) is used while learning the vector-space representations of word meanings from textual data that is supplemented with meta-data about the authors. In addition to the *global* embedding matrix W_{main} that contains low-dimensional representations for every word in the vocabulary [128], this approach has an additional $|C|$ matrices $\{W_c\}$ of the same size as W_{main}, where $|C|$ denotes the number of values the demographic variable has in the data (e.g., if gender is the demographic variable, $C = \{female, male\}$ and $|C| = 2$). Each of these $|C|$ matrices captures the effect that each demographic variable value has on each word in the vocabulary. To index the embedding of a stimulus word $w \in \mathbb{R}^{|V| \times k}$, the hidden layer h is computed as the sum of the matrix multiplications with each of the independent embeddings:

$$h = w^T W_{main} + \Sigma_{c \in C} w^T W_c \tag{5.4}$$

It then predicts the value of the context word y using another parameter matrix $X \in \mathbb{R}^{|V| \times k}$ based on a softmax function $o = softmax(Xh)$, where $o \in \mathbb{R}^{|V| \times k}$. Backpropagation using (input x, output y) word tuples learns the values of the various embedding matrices W and parameter matrix X, which maximize the likelihood of context words y conditioned on the stimulus word x.

This approach is used in its original implementation, provided by [79], to compute the word embedding vectors for all the words in the vocabulary.

Given a stimulus word, the closest vocabulary words with the highest cosine similarity are retained as the top association predictions for the given stimulus word.

5.4 Experiments

All the models require textual data with demographic information for training them. The data used and the adopted metrics for evaluation are introduced below.

Data. Given the requirement of having gender and location information associated with the data, blogs collected from Google Blogger[4] a large set of blog posts authored between 1999 and 2016, are used. Table 5.6 shows the breakdown of the raw blog counts per demographic category. From these, only those posts with non-empty content are retained, and preprocessed by removing HTML tags, converting all the tokens to their lemmatized forms,[5] and discarding those lemmas with a frequency less than 10, in order to avoid misspellings and other noise characteristic to social media content.

From the above pool of blog posts, two datasets are created with complementary demographic classes (1) location: India-US, and (2) gender: male-female.

Each of these datasets is processed so that they are profile-balanced with no peaks for any specific years, by applying several heuristics: **(1)** Compute the minimum number of users n over all the classes (e.g., Indian and US authors in the case of the location dataset). **(2)** From each class, select the top n users based on the number of years they were blogging, and the number of posts they wrote.[6] This ensures that the maximum amount of data will

[4] www.blogger.com

[5] The word forms are normalized using the Stanford CoreNLP lemmatizer [121].

[6] Prolific users will be chosen first. For a class with exactly n users, all users will be chosen.

DEMOGRAPHIC	RAW		BALANCED		
	PROFILES	POSTS	PROFILES	POSTS	TOKENS
INDIA	1,520	339,624	1,520	34,987	16,884K
US	3,273	825,093	1,520	32,782	11,706K
MALE	2,031	597,935	1,818	44,299	21,971K
FEMALE	1,818	321,779	1,818	45,980	17,070K

Table 5.6: Raw and balanced blog dataset statistics.

be available for the selected users. **(3)** For each of these n users, pick at most 50 posts in a round-robin fashion from the years in which they blogged. **(4)** Let M be the total number of posts collected in this manner from all the classes. In order to avoid having most of the posts coming from a small number of years, set a cutoff X as a fraction of M. For each year, a maximum of X posts will be chosen from the set of M posts ($X = 0.1M$). **(5)** To ensure that all the users get to contribute posts, and that the contribution of prolific writers is kept in check, maintain user participation scores:

$$p(user) = \frac{\text{posts collected from user}}{\text{total number of posts collected}} \tag{5.5}$$

These scores are updated after every year is processed, as explained further. **(6)** Sort the years in increasing order of number of posts and iterate through them; identify the lowest number of posts contributed by the least prolific writer, then collect the minimum number of posts from all users who published in that year in a round-robin manner. Then, select additional posts from users in increasing order of participation scores, until the number of posts for the year reaches the cutoff X. **(7)** After each year, update the user participation scores. Table 5.6 shows the number of users and posts retained after balancing. This particular composition is used in the *location* dataset (consisting of India and US posts) and *gender* dataset (consisting of females and males posts).

The *generic* datasets are obtained by combining half of each of the two location datasets, or the two gender datasets.

Metrics. Given that the word association task is relatively similar to the lexical substitution task, in terms of having an open vocabulary and lack of a “right” answer, the *best* and out-of-ten (*oo10*) evaluation metrics that are traditionally used for the latter [130], yet corrected for weight [131], are used for evaluation. Briefly, these measures take the best (or top ten) responses from a system, and compare them against the gold standard, while accounting for the frequencies of the responses in the gold standard. In addition, since Figure 5.1 shows that the top three ranking norms are provided as answers by approximately 42% of the respondents, with the remaining norms following a long Zipfian distribution in terms of frequency of appearance, the out-of-three (*oo3*) measure is also computed, as it represents a more focused approximation of the ability to predict human associations (note that out-of-ten covers 62% of the responses). Several recent works on word associations evaluated their models indirectly via Pearson or Spearman correlation performance on word similarity tasks [132, 133]; I choose instead to evaluate word associations directly, by using metrics that more closely align with the evaluations performed in the field of psychology, where the best output of a system is compared against the most frequent human response [134, 135].

For a given stimulus word w with human responses H_w, suppose a system returns a set of answers S_w. It is estimated how well this system can find a *best* substitute for w using Equation 5.6, where the function $freq_w(s)$ returns the count of a system response s in H_w, and $maxfreq_w$ returns the maximum count of any response in H_w.

$$best(w) = \frac{\Sigma_{s \in S_w} freq_w(s)}{maxfreq_w \times |S_w|} \tag{5.6}$$

$$oo\mathbf{n}(w) = \frac{\Sigma_{s \in S_w^n} freq_w(s)}{|H_w|} \tag{5.7}$$

Equation 5.7 measures the coverage of a system by allowing it to offer a set S_w^n of *n* responses for w, where each response s is weighted by its frequency $freq_w(s)$ in H_w.

5.5 Evaluations and Discussions

Evaluations are conducted using all the word association models described in Section 5.3. The results using the *best*, *out-of-three*, and *out-of-ten* evaluation metrics are listed in Table 5.7. For all the embeddings experiments, 300 latent dimensions are used. The Gen variation uses the demographic-blind dataset, whereas the DA variation uses the demographic-aware dataset.[7]

The MI and VSM models do not perform well in the word association prediction task, irrespective of considering the generic or the demographic-aware data. It should be emphasized, however, that the generic version of these models is able to consider co-occurrences across the entire generic datasets, while the demographic-aware co-occurrences can only

[7] To place the results in this table in perspective, it is important to note that results for this task are traditionally low. Given that the most frequent response is selected on an average by 24% of respondents (see Figure 5.1), it can be seen that even for humans, the highest score would be around 0.24.

Method	Type	Best		Oo3		Oo10		Best		Oo3		Oo10	
		IN	US	IN	US	IN	US	M	F	M	F	M	F
MI	Gen	0.00	0.00	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	DA	0.00	0.00	0.01	0.00	0.02	0.02	0.00	0.01	0.01	0.01	0.01	0.02
VSM	Gen	0.00	0.00	0.01	0.01	0.03	0.03	0.00	0.00	0.02	0.02	0.05	0.05
	DA	0.00	0.00	0.02	0.01	0.04	0.02	0.00	0.01	0.02	0.01	0.04	0.06
SGLM	Gen	0.02	0.02	0.03	0.03	0.06	0.05	**0.13**	0.13	**0.18**	*0.18*	0.20	0.21
	DA	0.05	0.01	0.07	0.02	0.11	0.03	0.10	0.13	0.16	*0.18*	0.18	0.20
C-SGLM	Gen	0.05	**0.04**	0.07	**0.07**	0.11	**0.10**	*0.11*	0.13	*0.17*	0.17	0.20	0.21
	EMB1	*0.08*	*0.03*	*0.12*	**0.07**	*0.18*	**0.10**	**0.13**	*0.14*	**0.20**	**0.20**	**0.25**	**0.26**
	EMB2	**0.09**	0.02	**0.14**	*0.04*	**0.19**	*0.06*	0.10	**0.16**	*0.17*	**0.20**	*0.23*	*0.25*
CSGLM-raw	EMB1	0.11	0.13	0.17	0.15	0.21	0.17	0.09	0.16	0.17	0.18	0.21	0.23
	EMB2	0.10	0.08	0.15	0.12	0.19	0.15	0.09	0.14	0.15	0.16	0.18	0.20

Table 5.7: Best, out-of-three (oo3), and out-of-ten (oo10) scores across the various methods. IN: India, US: United States, M: Male, F: Female. The numbers in bold mark the highest scores, those in italics, the second highest.

be computed from the section of the dataset that matches a particular demographic; as such, these latter models are placed at a disadvantage.

Perhaps not surprisingly, the neural network skip-gram-based architectures, whether SGLM [79] or C-SGLM, always achieve better results when compared to MI or VSM. The demographic-aware variation proposed in [79] uses an extended skip-gram architecture that encodes a generic embedding, and several demographic-based filters per class, which in the current scenario translates into three matrices of 300 dimensions each, the first for the generic words, and the subsequent ones for skews to be applied to the generic words in order to render the embedding through the lens of a given demographic. $SGLM - Gen$ in this case are the predictions based on the generic matrix, while $SGLM - DA$ are the predictions modified along the lines of a particular demographic.

The composite skip-gram models (C-SGLM) encode a single matrix that contains a mix of demographic-aware and generic words expressed as 300 latent dimensions. For both gender and location, the gender-aware models ($EMB1$ and $EMB2$) surpass the SGLM gender-aware model. Surprisingly, while SGLM was never meant to be generic, the predictions based on its generic embedding matrix prove to be a difficult baseline to surpass, similar to generic C-SGLM. Nonetheless, the composite skip-gram models ($EMB1$ and $EMB2$) do achieve best and second best rankings in the vast majority of cases (when compared to the best among all the other methods), with $EMB1$ being the more robust variation, performing well both for gender and for location.

Focusing on the performance of $EMB1$, the highest gains are observed for India-based predictions, for *best* (from 0.05 to 0.08) and *out-of-three* (from 0.07 to 0.12); for male-based predictions increasing from 0.11 to 0.13 for *best*, and from 0.17 to 0.20 for *out-of-three*; and for female-based predictions, increasing from 0.13 to 0.14 for *best*, and from 0.17 to 0.20 for *out-of-three*. US-based associations are the hardest to predict, probably because of the diverse makeup of society; additional evaluations are needed to pinpoint the exact cause.

To determine how susceptible the embedding model is to skewed, but larger training data, a separate experiment is run on the entire raw set of blogs that are collected (described on the left side of Table 5.6), where the $EMB1$ and $EMB2$ models are re-generated. While the entire dataset is significantly larger than the balanced set, it is also significantly skewed: the data in the India-US dataset was skewed in a proportion of 1:0.48 tokens, while for Female-Male the proportion was 1:0.41 tokens. As is the case for the balanced dataset, the $EMB1$ model is still the most robust (see the bottom section in Table 5.7), and it achieves significant gains when compared to its balanced counterpart, in particular for *best* (for the US demographic from 0.03 to 0.13, and for India from 0.08 to 0.11), and for *out-of-three* (for US from 0.07 to 0.15, and for India from 0.12 to 0.17), which suggests that as an avenue for future research, the use of significantly larger but unbalanced datasets can be explored to train the association prediction models.

5.6 Conclusion

In this chapter, the task of demographic-aware word associations is introduced. To understand the various ways in which people associate words, a new large demographic-enhanced dataset is collected. It consists of approximately 300 stimulus words and their associated norms compiled from 800 respondents, resulting in a total of 176,097 non-spam responses, and it shows that for people of different demographics, associations do differ with gender and location.

A new demographic-aware word association method is proposed based on composite skip-gram models, that can jointly embed generic and gender tagged words. This method improved over its generic counterpart, and also outperformed previously proposed models of word association, thus demonstrating that it is useful to account for the demographics of the people behind the language when performing the task of automatic word association prediction. I regard this as a first step toward demographic-aware NLP, and in the upcoming chapters, I plan to address other more advanced NLP tasks while accounting for demographics.

The findings of this chapter are published here: [44]. The word association dataset introduced in this chapter is publicly available from http://lit.eecs.umich.edu/downloads.html.

CHAPTER 6

Demographic-Bias in Part-of-Speech Tagging and Syntactic Parsing

6.1 Introduction

Language variation goes beyond the lexical level – syntax plays a crucial role in the systemic variation among groups. Sociolinguistic studies have shown that people use grammatical features to signal the speakers' membership in a demographic group, with a focus on gender [82, 11, 83]. Several works in the NLP community use syntactic features from text, such as part-of-speech tags and dependency relations, to improve data-driven dependency parsing [84], or sentiment classification [85, 86], among other tasks. Various works have also explored the influence of demographics on syntax [11, 12, 13]; for example, [11] does a systemic analysis of the usage of various types of clauses and their positions among men and women, stating that women have a higher usage of adverbial (*accordingly*, *consequently*[1]), causal (*since*, *because*), conditional (*if*, *when*) and purpose (*so*, *in order that*) clauses, while men tend to use more concessive (*but*, *although*, *whereas*) clauses. Similar results hold across various languages in [13].

This correlation between grammatical features and demographics has important ramifications for statistical models of syntax: if the training sample is unbalanced, these differences inadvertently introduce a strong demographic bias into the training data. Such demographic imbalances are amplified by the model [136], which in turn can be detrimental to members of the underrepresented demographic groups [137, 138, 139]. Since several works use syntactic analysis to improve tasks ranging from data-driven dependency parsing [84] to sentiment classification [85, 86], underlying model biases end up affecting the performance of a wide range of applications. While data bias can be overcome by accounting for

[1] The conjunctions or conjunctive adverbs that introduce and link in a subordinating relationship the given type of subordinate clause are exemplified in paratheses.

demographics, and can even improve classification performance [32, 31, 140, 141, 136, 87], there is still little understanding on the amount and sources of bias in most training sets.

In order to address gender bias in part-of-speech tagging and dependency parsing, first an adequate data set is required which is labeled for a) *syntax* along with b) *gender* information of the authors. However, existing datasets fail to meet both criteria: datasets with gender information are either too small to train on, lack syntactic information, or are restricted to social media; sufficiently large syntactic datasets are not labeled with gender information and rely (at least in part) on news genre corpora such as the Wall Street Journal (WSJ). To address this problem, I augment the WSJ subset of the Penn Treebank corpus with gender,[2] based on author first name. To my knowledge, this is the first work that explores syntactic tagging while accounting for gender in the Penn Treebank corpus.

Contributions. The main contributions of this chapter are as follows:

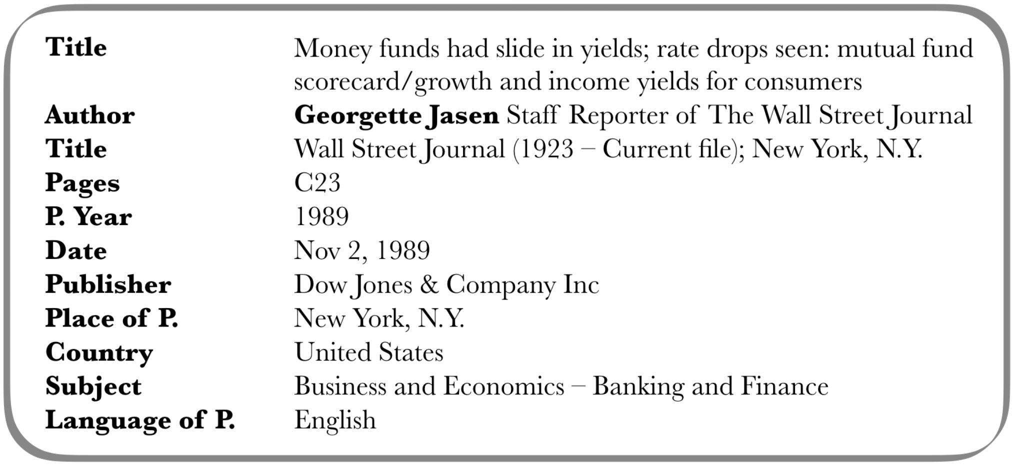
- A standard POS-tagging and dependency parsing dataset is augmented with gender information of the authors.
- Experiments are conducted to study the role played by gender information in POS-tagging and syntactic parsing.
- The relation of gender with various part-of-speech tag and syntactic differences is explored.

6.2 Annotating Penn Treebank for Gender

The Penn Treebank [142] is the de facto data set used to train many of the POS taggers [143, 144, 145, 109] and syntactic parsers [110, 146, 147]. It contains articles published in the Wall Street Journal in 1989, as well as a small sample of ATIS-3 material, totalling over one million tokens, and manually annotated with part-of-speech tags and syntactic parse trees. Since the ATIS-3 data contains dialogue acts between humans and automated systems [148], this study focuses on adding author gender information to only the Wall Street Journal subset of the Treebank.

The WSJ articles are supplemented with metadata from the ProQuest Historical Newspapers database, which indexes, among others, WSJ articles released between 1923 and 2000, and provides fields such as author names. Out of the original 2,499 WSJ articles,

[2] Since all the editors of WSJ belong to United States, the study is restricted to gender as the demographic dimension.



Title	Money funds had slide in yields; rate drops seen: mutual fund scorecard/growth and income yields for consumers
Author	**Georgette Jasen** Staff Reporter of The Wall Street Journal
Title	Wall Street Journal (1923 – Current file); New York, N.Y.
Pages	C23
P. Year	1989
Date	Nov 2, 1989
Publisher	Dow Jones & Company Inc
Place of P.	New York, N.Y.
Country	United States
Subject	Business and Economics – Banking and Finance
Language of P.	English

Figure 6.1: Example Wall Street Journal article metadata from ProQuest (P. refers to Publisher).

1,814 are found in ProQuest and their metadata is retrieved. Figure 6.1 shows the metadata for a sample WSJ article. The *Author* tab contains the name of the reporter who published the story. 556 articles with an empty *Author* field are removed, resulting in 1,258 WSJ articles with author information. Using a combination of regular expressions and manual verification, the author names are extracted for 1,006 articles (the remaining 252 articles do not have actual author names).

The first names are isolated using regular expressions, and [149] is followed to automatically assign gender and compute a gender ambiguity score taking into consideration: (1) the list of first names obtained based on Facebook profiles by [150]; and (2) the Social Security Administration's (SSA) baby names data set.[3] The Facebook list has male and female assignment scores for each name, while the SSA maintains a data set of counts for baby names and gender for each year since the 1880s, for names with at least five occurrences. If both databases agree in their gender assignment, that assignment is used as the final label. 987 articles are assigned author genders with agreement between both the databases. For the remaining 19 articles, the author gender is manually identified by cross-referencing the names online. 5 of these only have a first name initial, and thus could not be resolved and were discarded. The gender mapping results in 1,001 gender-tagged Wall Street Journal articles. Discarding 115 articles with joint authorship and considering only articles with both part-of-speech tags and parse trees results in a final set of **886 articles** from the Penn Treebank.

The final set of articles includes 379 unique authors, with a heavy gender imbalance of

[3] http://www.ssa.gov/oact/babynames/limits.html

1 to 3 (96 female and 283 male authors). The total number of sentences in female articles is 7,282, with a mean of 21.17 tokens per sentence ($\sigma = 10.03$), while the male articles consist of 19,400 sentences, with a mean of 20.99 tokens per sentence ($\sigma = 10.52$). This is similar to the findings in [151], which also notes a lengthier utterance mean for women versus men (the study focuses on adolescents).

Tthe Universal Dependencies (UD) v1.4 [152] annotation guideline is used for part-of-speech tags and parse trees, and accordingly, the constituency trees from the Penn Treebank format are converted to the CoNNL format.[4] The part-of-speech tags are then mapped to the universal part-of-speech tag set.[5]

6.3 The Effect of Gender on Part-of-Speech Tagging and Dependency Parsing

To assess whether author gender affects parsing performance, the state-of-the-art transition-based neural network model SyntaxNet[6] [153] is trained on the data (with default parameters), and test whether stratified training can alleviate these effects. The performance is evaluated for individual part-of-speech tags and dependency relations, as well as over all the tags and relations.

Stratifying the Training Data. Since the Wall Street Journal data has a heavy gender imbalance (1:3 female to male articles), the data is stratified by discarding male examples so that the number of female and male sentences and tokens do not differ by more than 15%, according to the following heuristics: (1) The female and male Wall Street Journal sentences are sorted in the descending order of the number of tokens. (2) For each female sentence F_i with f_i number of tokens, a male sentence M_j is selected such that the number of tokens $m_j \in [0.85f_i, 1.15f_i]$. (3) If no more male sentences that qualify for this condition remain, the next male sentence is chosen in decreasing order of the number of tokens $m_j \in [5, 30]$. Table 6.1 shows the number of sentences and tokens in the Wall Street Journal data before and after balancing for gender.

SyntaxNet is trained in three scenarios: (1) on female-only data, (2) on male-only data, and (3) on generic data containing an equal number of male and female sentences. All the three datasets have an equal number of sentences.

[4] `https://nlp.stanford.edu/software/stanford-dependencies.shtml`.

[5] The datasets are annotated with the 16 universal part-of-speech tags; `conj` is used for both `sconj` and `conj` tags.

[6] `https://github.com/tensorflow/models/tree/master/syntaxnet`

	Raw		Balanced	
Gender	Sentences	Tokens	Sentences	Tokens
Female	7,282	175,107	7,282	175,107
Male	19,400	461,742	7,282	202,144

Table 6.1: Number of sentences and tokens in the raw and balanced Wall Street Journal data.

Evaluation. The standard metrics for part-of-speech tagging and dependency parsing are used for evaluation – *accuracy* (Acc): the percentage of tokens that have a correct assignment to their part-of-speech (for part-of-speech tagging); and *labeled attachment score* (LAS) – the percentage of tokens that have a correct assignment to their heads *and* the correct dependency relation [154] (for dependency parsing).

In each training scenario, five training-test splits are created with 90:10 ratio by random sampling on the Wall Street Journal data. In order to derive parameters for SyntaxNet, each train split is further randomly split into five training-validation folds, and the results are averaged across the folds. When creating the folds, it is ensured that sentences authored by the same author are not shared across splits to avoid over-fitting the models to the writing styles of individual authors, rather than learning the underlying gender-based differences as they pertain to syntax.

Train:	Generic	Female	Male
Test	POS Accuracy		
Generic	95.81*	95.49	95.74*
Female	**95.96***	95.90*	95.47
Male	95.47	95.03	**96.08***
	Dependency LAS		
Generic	83.03	82.01	83.11
Female	**83.46***	83.17	83.12
Male	82.53	81.15	**83.21***

Table 6.2: Results for part-of-speech tagging (Acc) and dependency parsing (LAS) on Wall Street Journal test data. The highest values in each row are in **bold**, and are statistically significant compared to all other values using McNemar's test ($p < 0.05$).

In each training scenario, the models are evaluated on: (1) female-only data, (2) male-only data, and (3) generic data containing an equal number of male and female sentences, such that all test settings share the same number of sentences (10% of 7,282 = 728; 364 sentences from each gender for the generic test dataset). Since there are give test folds, and each fold in turn has five validation folds (for parameter tuning), results are averaged over the 25 total runs and reported to ensure robustness.

6.4 Results and Discussion

Table 6.2 (top) shows the part-of-speech tagging accuracies for labeling the Wall Street Journal test data. Note that while accuracy differences may be relatively small, they are within the margins of recent state-of-the-art improvements [153] in a task that achieves extremely high accuracy and where further improvement can only be incremental. Considering performance across the three different training scenarios, the female test data sees a slight benefit from a mixed training set, achieving its highest accuracy of 95.96%, while male test data only achieves the highest performance (96.08%) with a model trained on male-only data, representing a relative error rate reduction[7] of 13.46% when compared to the generic model.

The setting closest to the current part-of-speech tagging setups is embodied by training on the generic model. In this case, the female test data achieves its highest accuracy (95.96%), but the male test data achieves only a second best performance (95.47%). This difference suggests an area of possible improvement in performance for off-the-shelf part-of-speech taggers.

A similar pattern is observed in dependency parsing (Table 6.2, bottom), where the female test set achieves the highest LAS on the mixed training set (83.46%). The male test set obtains its highest accuracy when the training is performed on male-only data, with a relative error reduction of 3.89% as compared to training on generic data.

It seems that female writings are more diverse, with a complexity that can best be approximated with mixed-gender training samples. This setting improves performance by relative error reductions of 1.46% and 1.72% respectively for ACC and LAS when compared to training on female-only data, and 10.82% and 2.01% respectively for ACC and LAS when compared to training on male-only data. The male test sentences appear to display less variability, and therefore cannot benefit the same amount of information from the spectrum displayed by female training data; in fact, whenever female-authored sentences are present in the training set (either in female-only setting or generic setting), performance drops significantly on male test data.

When comparing male and female-only training sets and their ability to generalize to the opposite gender, it is noticed that male training data is more malleable and lends itself better to be used when testing on female samples, but not the reverse.

[7] Relative error rate reduction between two figures X and Y is $\frac{Y-X}{100-X}$.

Note that the Wall Street Journal exemplifies a highly formal and scripted newswire genre where gender differences are likely less pronounced, yet they still surface. It can be expected to find even stronger gendered language differences in a large, informal dataset (from social media) comprising both gender and syntactic information, where people use syntactic structures that best articulate their thoughts without any restrictions. These differences can be leveraged to achieve better performances for core NLP tasks.

Train:	Generic	Female		Male	
	Acc	Acc	Err	Acc	Err
		Male Test			
`noun`	93.74	92.51	-19.63	**94.23**	7.92
`det`	99.09	99.09	-0.13	**99.13**	4.08
`num`	99.23	99.34	15.35	**99.35**	16.60
`propn`	93.97	90.10	-64.14	**95.26**	21.41
		Female Test			
`pron`	98.91	**99.12**	18.99	98.97	4.64
`aux`	98.60	**98.77**	12.12	98.39	-14.75
`adj`	92.12	**92.62**	6.37	92.36	3.06
`propn`	94.66	**94.97**	5.76	91.60	-57.33

Table 6.3: Tag-wise results for part-of-speech tagging on Wall Street Journal test data; Accuracies (Acc) and relative error reduction rates (Err) versus generic models are reported. Values in **bold** are statistically significant ($p < 0.05$) using McNemar's test.

Clear gender-based performance improvements are observed at the tag level (Table 6.3). For instance, models trained on male-only data perform better on the male test data for the tags `nouns`, `determiners`, `numerals` and `proper nouns`, compared to models trained on mixed data (with a relative error rate reduction between 2.75% and 21.41%). Similarly, female-trained models perform better on female test data for the tags `pronouns`, `auxiliaries`, `adjectives`, and `proper nouns`, compared to models trained on mixed data (with a relative error rate reduction between 5.76% and 18.99%). For 8 out of the 16 part-of-speech tags, mixed training achieves best results for either female or male test data.

In dependency parsing (Table 6.4), models trained on female data perform better on female test sets for the tags `amod`, `cop`, `appos`, and `cc:preconj` (with a relative error rate reduction between 3.11% and 22.96% compared to generic models). Similarly, male-trained models are able to outperform mixed models on male test data for the tags `csubj`, `iobj`, `acl`, `compound`, `xcomp`, `dobj`, `conj` and `nummod` with a relative error rate reduction between 2.11% and 14.61%. In dependency parsing, mixed training never achieves

TRAIN:	GEN.	FEMALE		MALE	
	LAS	LAS	ERR	LAS	ERR
		MALE TEST			
csubj	25.20	27.89	3.60	**36.13**	14.61
iobj	47.11	40.61	-12.29	**48.59**	2.80
acl	63.93	60.47	-9.60	**66.09**	5.99
compound	75.06	72.95	-8.45	**77.26**	8.83
xcomp	74.39	72.26	-8.30	**75.38**	3.85
dobj	84.48	82.13	-15.17	**85.20**	4.66
conj	82.45	80.74	-9.77	**82.82**	2.11
nummod	92.00	91.24	-9.42	**93.08**	13.53
		FEMALE TEST			
amod	91.18	**91.46**	3.11	91.08	-1.18
cop	92.78	**93.89**	15.47	92.80	0.34
appos	79.44	**80.31**	4.21	80.13	3.38
cc:preconj	54.68	**65.09**	22.96	50.78	-8.60

Table 6.4: Tag-wise results for dependency parsing on Wall Street Journal test data; LAS and relative error reduction rates (ERR) versus generic models are reported. Values in **bold** are statistically significant ($p < 0.05$) using to McNemar's test.

the best per tag results for either male or female test sets.

These tag-wise gender-specific results suggest that leveraging the idiosyncrasies for specific tags displayed by each gender could help create gender-aware models that leverage the syntactic strengths of each gender, and improve prediction accuracy for both genders. Note that there is a heavy topic overlap between the male and female Wall Street Journal articles, with a Pearson correlation of 0.85 between the male and female topic distributions,[8] indicating that the differences in performances between male and female models on various test sets are not due to topical shifts, but are due to syntactic variations.

6.5 Conclusion

The experiments in this study show that women's syntax displays resilience: part-of-speech taggers and dependency parsers trained on any data perform well when tested on female writings. Male syntax, on the other hand, is parsed or tagged best when sufficient male-authored data is available in the training set. This suggests that men "lucked out" with respect to the gender imbalance in the WSJ training data: a more balanced or more female-

[8] The topic distributions were extracted using Latent Dirichlet Allocation [118]. The LDA implementation included with the Python Gensim library [122] is used with 10 topics.

heavy data set could have caused significant drops in the performance of automatic syntax analysis for male writers. While the current study provides insights into the relationship between gender and syntax in a general newswire domain with stringent stylistic restrictions, a more in-depth analysis would require labeled data from social media domains, where people are at liberty to express their personal writing styles. Nonetheless, the gender-annotated Wall Street Journal data provides a starting point for leveraging gendered grammatical differences and the development of better and fairer models for syntactic annotations, as well as for the many downstream NLP tasks that use syntactic information in their models.

The Penn Treebank author gender information is publicly available from `http://lit.eecs.umich.edu/downloads.html`. This work is published here: [47].

CHAPTER 7

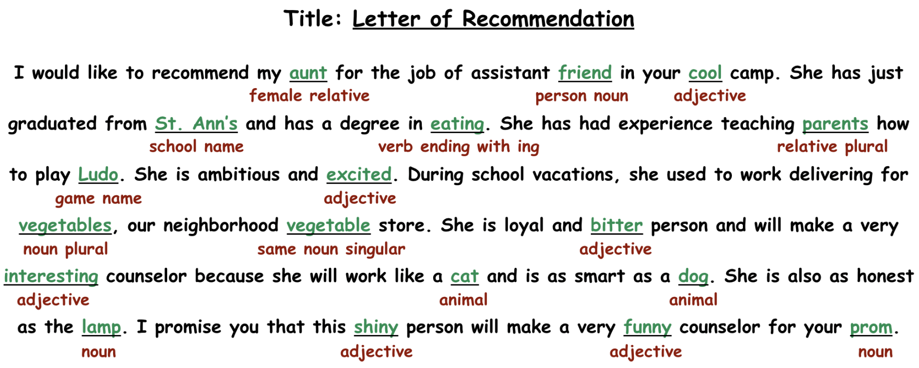
Automatic Demographic-Aware Humor Generation for Mad Libs

7.1 Introduction

Humor can be defined as the tendency to provoke laughter and provide amusement. It is a universal phenomenon that is employed by people all over the world, across different countries, genders, and age groups [155]. Despite this, there are differences in how humor is enacted and understood due to demographic differences [156, 157, 158, 107, 105, 159]. For instance, [105] indicates that women share humor to build solidarity, whereas men employ humor to impress and emphasize similarities. Similarly, humor styles vary across nationalities, and the differences can be significant [107]. Specifically, [107] indicates that humorous communications from diverse nationalities share certain universal cognitive structures, while also containing content that is nation specific.

Computer-generated humor is an essential aspect in developing personable human-computer interactions. However, generating humor is typically considered a difficult natural language problem, with some researchers considering this task AI-complete [97]. Despite years dedicated to theories and algorithms for humor, the best automated humor is still mediocre compared to the one crafted by humans. Humor is subjective and it can be interpreted in different ways by different people. Humor requires creativity, world knowledge, and cognitive mechanisms, which are extremely difficult to model theoretically.

In computational linguistics, a large body of humor research involves *humor recognition*, which refers to the task of determining whether a given context expresses a certain degree of humor. Research in humor recognition is mostly limited to specific types of jokes [160, 161, 162, 163, 164, 162]. *Humor generation*, on the other hand, involves generating content that expresses a certain degree of humor. Similar to humor recognition, this task is also largely limited to specific joke types, while also being limited to short texts, such as riddles [96], acronyms [97], or one-liners [98].



Title: Letter of Recommendation

I would like to recommend my aunt for the job of assistant friend in your cool camp. She has just graduated from St. Ann's and has a degree in eating. She has had experience teaching parents how to play Ludo. She is ambitious and excited. During school vacations, she used to work delivering for vegetables, our neighborhood vegetable store. She is loyal and bitter person and will make a very interesting counselor because she will work like a cat and is as smart as a dog. She is also as honest as the lamp. I promise you that this shiny person will make a very funny counselor for your prom.

Figure 7.1: Sample Mad Lib story with blanks filled-in by a human player.

In 1958, a phrasal template party game called Mad Libs was introduced, where one player prompts the others for a list of words to substitute for blanks in a story, before reading the story aloud [165]. The resulting story is often comical due to the nonsensical filling of words. Figure 7.1 shows a sample Mad Lib story with its blanks filled-in by a human player. In a recent study, Hossain et al. [100] suggest and evaluate a word candidate selection method with the aim of helping human players fill in the blanks to make stories of the Mad Lib format funny. Inspired by this line of research and seeking to find a solution to generate demographic-compatible humor, in this chapter, I develop and evaluate a demographic-aware humor generation framework that is fully-automated and that seeks to fill in the blanks in a Mad Libs story by mimicking word choices made by a player from a given demographic background.

In particular, this framework has three components:

1. A demographic-specific language model that generates words to fill in the sentences in each Mad Lib story, while also accounting for a desired demographic slant to the stories,

2. A demographic-specific sentence pair classifier to assess if a filled-in or "transformed" sentence is a funny version or "transformation" of the original sentence template for the demographic group under consideration, and

3. A story coherence component to join individual funny sentences so as to form complete Mad Lib stories that are humorous.

The goal of this study is to examine if a fully-automated approach without human intervention for filling-in the Mad Lib stories can generate humor better than simple baselines, and if people from a specific demographic background prefer humor created by a matching demographic versus a dissimilar demographic.

This chapter makes three main contributions.

First, a novel dataset for demographic-specific humor generation is created, which is used to study demographic-specific preferences for humor between the groups of interest. This dataset is based on Mad Libs-like humorous stories, which so far have been fully completed only by human players. It is a challenging task in that a story template cannot be filled-in using trivial strategies, such as substituting random words, while achieving an amused reaction from an audience. The dataset is annotated and judged on Amazon Mechanical Turk (AMT) using an appropriate setup to avoid spam responses and reduce deviation in the human judgements for humor.

Second, a demographic-specific humor generation framework is developed by slanting a neural-net based language model through refining its predictions on texts culled from social media platforms that are authored by members of the demographic group in question. Furthermore, the framework is able to learn what makes a story funny by fine-tuning a classifier on the humor prediction task leveraging the data annotated via AMT. In most cases, the framework generates funnier Mad Lib stories than those created by human players.

Third, qualitative and quantitative analyses are conducted to explain what makes the stories generated by the proposed framework humorous. I explore how the stories differ from those composed by humans and those obtained using a general baseline method. Furthermore, I investigate the preferences of people from a given demographic toward stories completed using a matching-demographic model versus a miss-matching demographic model, to gauge whether the framework is able to generate demographically slanted humorous Mad Libs.

I focus on culture as a demographic dimension, selecting India and the United States (US) for the study, as both countries have a large English-speaking population, who richly participates not only on social media platforms, but also on crowd-sourcing portals, such as AMT.

While the previous work ("Visual Madlibs") done by [102] is titularly similar to the current work in this chapter, that work augmented an image dataset with fill-in-the-blank questions such as "This object is a chair," whereas the current work aims to generate humorous new Mad Lib stories with a demographic slant. This is similar to the work by [101] for the Story Cloze Test, where the goal was to choose a satisfying ending to stories. This study is most closely related to [100], in which the task of generating funny stories via filling-in the blanks in stories following a Mad Lib format was introduced, and a semi-automated humor generation framework is developed to aid humans to create humorous stories. Unlike their work, the framework I propose is fully automated and seeks to select candidates for the blanks that speak to a person's demographic background, enabling the generation of demographically-aware humor.

To the best of my knowledge, this is the first study in NLP that aims to develop a completely automated approach to generate humor by filling-in Mad Lib stories, while also accounting for the demographic component in humor. The annotated data collected in this study will be released to the research community to encourage further exploration in this direction.

7.2 Mad Libs®

Invented by Leonard Stern and Roger Price in 1953, Mad Libs is a fill-in-the-blank game intented to create humorous stories [165]. A Mad Lib consists of a title and a short story of several sentences in length. In each story, some of the words are replaced with blanks, where each blank is accompanied by a hint type, prompting for a grammatical and coherent replacement in context. The task of the players is to fill in the blanks with words compatible with the hint types. Some of the hint types include part-of-speech tags (noun, verb, adjective, adverb), place, celebrity names, part of the body, and food type. Only the story's title and the hint types are shown to the players, without revealing the context around the blanks. Only after the blanks are filled in, the story is revealed, resulting in a generally funny atmosphere, with the humor aspect coming from the nonsensical filled-in words in an otherwise coherent and sensible story.

Figure 7.1 shows a sample Mad Lib story created from a template describing a letter of recommendation.

7.3 Adapting Mad Libs-like Stories to Humor Generation

Filling-in Mad Lib-like stories was introduced as a semi-automatic humor generation task in [100], where the task was to produce candidate word suggestions for the blanks, to aid human players in developing funnier stories.

In addition to providing the title and hint type for each blank, as done in the original Mad Libs game, the words surrounding the blanks in the story are also utilised in this task, so as to provide contextual information that an algorithm can learn from while making the word suggestions. While the former problem can be approached by choosing *a priori* funny words according to a funny lexicon, the latter problem is richer, as the players and models can learn from the surrounding context to choose the appropriate words to fill in the blanks.

7.4 Overview of Fun Libs

Instead of using the original Mad Lib stories, this study utilizes the “Fun Lib” stories created by [100] for the following reasons: (1) Mad Libs are copyrighted, and hence it is difficult to release datasets obtained by using stories from the Mad Lib books, and (2) experimentation with the stories used by [100] allows a comparison between the proposed framework in this chapter with the one proposed by [100]. After studying 50 original Mad Libs [166], The Fun Libs have been carefully selected and curated by [100] through several human pilot studies, ensuring that they are meaningful and diverse. Some of the hint types from the original Mad Lib setting are discarded in Fun Libs, as they restricted the variety of humor that can be generated when filling in their blanks, either because they do not allow for many substitutions (*color*, *exclamation*, *silly word*) or subtlety in humor (*number*), or because they require the players to have knowledge of cultural references and specifics (*celebrity name*, *place*). Further, I discard 4 of the 50 Fun Lib stories as they are about topics more widely known to US audiences and possibly less known to Indians (*Kim Kardashian*, *Baseball*, *Boston Tea Party* and *The Statue of Liberty*). Instead, I replace them with 4 new culture-neutral stories following the guidelines devised by [100].

Type	Mad Libs		Fun Libs	
	Mean	Std Dev	Mean	Std Dev
Blanks	16.00	2.25	14.74	1.16
Words	114.84	20.58	120.08	18.37
Sentences	9.04	2.38	7.68	1.65

Table 7.1: Mean and standard deviation of the number of words, sentences and blanks per story in Mad Libs and Fun Libs datasets.

Type	Mad Libs	Fun Libs
Noun	7.06	6.86
Adjective	4.06	3.22
Verb	1.22	3.16
Adverb	0.44	0.38
Animal	0.20	0.36
Type of Food	0.22	0.20
Part of the Body	0.98	0.42
Type of Liquid	0.18	0.14
All Blanks	*16.00	14.74

Table 7.2: Mean hint types per story in Mad Libs and Fun Libs datasets. *The mean number of blanks in the Mab Libs dataset is computed based on 14 hint types.

The Fun Lib stories are created from Wikipedia articles, as these also have a title and content, similar to Mad Libs. Sentences from Wikipedia articles are selected such that they have the potential to generate humor (with a few edits to reduce verbosity), and some of the words belonging to the available hint types in them are blanked out such that the sentence, word, blank and hint type distributions are similar between the original Mad Lib and new Fun Lib stories. Tables 7.1 and 7.2 show the mean and standard deviation values per Mad Lib (with 14 hint types) and Fun Lib story (with 8 hint types).

7.5 Culture-Specific Fun Libs Annotation

Starting with the 50 modified Fun Lib story templates, the Amazon Mechanical Turk (AMT) platform is used to create filled-in stories, such that the resulting stories are humorous. The general approach used for generating Fun Libs annotations by [100] is adopted in this study, with a few changes so as to include the culture[1]-specific annotations. The annotations are done in three stages. First, an initial set of *pre-filled stories* are used to select **qualified judges** on AMT. Second, a separate set of *story templates* are published on AMT to obtain completed stories, and **qualified players** are selected based on the quality of the word candidates they provided in generating a humorous story (as decided by the judges selected in the prior stage). Finally, the qualified players are given the 50 Fun Lib story templates to complete, resulting in a dataset of stories with filled-in blanks. The judges then label the ensuing stories with funniness grades for each filled-in word independently, as well as for the story overall. All the turkers involved in each of these stages are native English speakers from India or the United States, have a HIT approval rate of at least 97%, and have completed at least 10,000 HITs. Following are detailed descriptions of each of the above three stages.

7.5.1 Judge Selection

Grading the funniness of filled-in stories requires judges who are unbiased referees of humor. The selection of judges is challenging, as humor is subjective. To do so, a qualification test is published on AMT with clear instructions as to what qualities make a desirable judge. The task for the turkers is to grade for funniness a set of seven stories, three of when are direct excerpts from Wikipedia (without further modifications), except for underlining some of the words to make it seem as if though they were filled-in, and the blanks in the

[1] From here on, the term “culture” is used instead of “demographic” as it is the only demographic category considered in this study.

remaining four stories are filled-in by English speakers from the US and India, who were instructed to create funny stories. This task is published on AMT twice, once for Indian turkers and once for US turkers, to select Indian and US judges, respectively. The turkers are asked to select for each of the seven stories a grade on a Likert scale between 0 and 3, with the following interpretations:

0: Not Funny

1: Slightly Funny

2: Moderately Funny

3: Funny

In addition to the questions regarding the funniness grades, each story in the task is followed by a question that can only be answered after reading the entire story. The task ends with a demographic survey containing questions regarding the turker's background information, such as age group, nationality, gender, education, occupation, and income level.

The ground truth grades for the four US filled-in stories are obtained from five student volunteers from my research group, while the grades for the India filled-in stories are obtained from five students from India. The mean value assigned by each of these groups is considered as the ground truth grade for that culture.

60 turkers are enlisted to do the task from each country, and I select those turkers who: (i) assign 0 to the three Wikipedia stories, (ii) assign a grade from $\{1, 2, 3\}$ to at least three out of the remaining four stories, (iii) answer the story-based verification questions correctly, and (iv) spend at least 4 minutes of time to complete the task. The turkers are then sorted in the decreasing order of the Euclidean distance of their grades from the ground truth. From the 60 initial turkers, 42 of them are qualified based on the above strategy from US, while only 27 are qualified from India. To select a few more judges so as to account for diversity in judges, the tasks are republished with 15 more turkers from US and 30 more turkers from India, and 10 and 16 additional judges are qualified from US and India, respectively, according to the above strategy. Thus, a total of 43 Indian and 50 US judges are selected using this qualification task, such that they are all careful, consistent, and representative in objectively judging the funniness of a filled-in story

7.5.2 Player Selection

The goal of player selection is to qualify turkers from each culture who are creative, and are good at creating funny stories by filling-in the blanks. To do so, four additional stories are obtained from Wikipedia, some selected words in them are marked as blanks, and the hint types for each of the blanks are provided next to the blanks. The four stories are divided into two tasks on AMT, with each task requiring the filling-in of the stories so as to make them as funny as possible for a general audience, and each of the tasks is provided with the following further instructions for the turkers: the filled-in words must (1) be grammatically correct, not be colloquial, and be found in an English dictionary, (2) have no more than one word, and with alphabetic characters only (thus multi-word expressions such as “grow up” or “panic-stricken” are not allowed), (3) agree with the hint types provided, and (4) not be slang words, sexual references or be bathroom-related humor, as they are crude ways of obtained funniness, and do not require creativity.

The four stories are divided into two tasks (instead of having all of the four stories in one task) to avoid player fatigue and to allow for variety and creativity in the generated humor. Toward the end of each story, a question is asked for the turkers to self-grade their filled-in stories for their funniness, to ensure that they are conscious of creating a funny version rather than just filling-in the blanks, and to see if they find their own story funny. In addition, a demographic survey similar to the one presented during the judge selection stage is added at the end of the task to obtain the background information of the players. 30 turkers are required to complete each task from each culture, thus making a total of 60 turkers being examined from each culture in this qualification test via the two AMT tasks.

Each of the filled-in stories is graded for its funniness on the 0-3 scale by 5 qualified judges (selected in the judge selection stage) to reduce the effect of variations in humor preferences and to be representative of an audience rather than an individual. Those players are selected who have an average judgement ≥ 1 for at least one of the two filled-in stories.Very rarely do stories have drastically different judgements (such as $\{0, 3, 3, 3, 3\}$ or $\{0, 0, 0, 1, 3\}$), and the corresponding turkers are qualified or rejected as players by ignoring the judgements in minority. The turkers are ranked in the decreasing order of the **mean funniness grade** (the average over the five judgements for their stories), and top ranked turkers are selected as qualified players. To obtain a comparable number of Indian players, an additional task is published on AMT requiring 20 more turkers to fill in the blanks from India. A total of 30 Indian and 26 US players are selected via this process. Some of the selected players are also qualified as judges, hence in the data annotation to label Fun Lib stories, it is ensured that each turker can participate as either a player or a judge but not both

7.5.3 Labeling Fun Libs

The stories are filled-in by the qualified players, and the judgements are obtained from the selected judges to assess the overall funniness and other aspects of humor. For the training stories, in addition to obtaining the overall funniness, the humor contribution of each filled-in word is also asked for, and these scores are later used in the training process. The players are again required to self-grade their filled-in stories.

Each story is filled-in by 3 players from each culture following the instructions used for player selection, and similarly the judgements are obtained for each filled-in story from 5 judges from the corresponding cultures. The judges are required to answer the following questions for the training set stories:

- Humor contribution of each filled-in word (from {'funny', 'not funny'}).
- A question about the **overall funniness** of the filled-in story (from {0, 1, 2, 3}).
- A question about how **coherent** the filled-in story is (from {0, 1, 2, 3}). Coherence is the quality of being **logical** and **consistent**.
- A question about the extent of **deviation** the filled-in words caused to the original story (as suggested by the title) from {0, 1, 2, 3}.
- A question whether the **incongruity theory** of humor generation was applied by the player (from {'Yes', 'No'}). Incongruity theory states that a joke is funny when it has a surprise, often at the end, that violates the conventional expectation, often set up at the start [90, 93].
- A **verification question** which can be answered only after reading the entire story.

The test stories are also filled-in by 3 players from each culture, and the judges are asked to answer questions 2-6 above. In addition to obtaining the same culture judgements that are asked for the training stories, the filled-in test stories are also judged by judges from the opposite culture to allow for cross-cultural analyses. The overall funniness score is used to evaluate the filled-in stories in the evaluation phase, while the coherence is questioned so as to study the role of coherence in this humor generation task, as it is expected that incongruity plays a major role in generating humor in Mad Libs due to the nature of the task. Also, a question about the deviation is to understand how topic changes in the stories contribute to humor. The verification question is to ensure that the judges read the entire story before providing their judgements. The Krippendorf's alpha [167] values for inter-rater reliability for India and US judgements are 0.214 and 0.173 respectively, which indicate positive agreements among AMT judges, and are comparable to those obtained by [168], who crowd-sourced a dataset of humorous edited news headlines on the same funniness scale.

The players are remunerated with \$0.50 during the player selection and the data annotation phases, while the judges receive \$0.25, \$0.10 and \$0.50 for the judge selection, the player selection and the data annotation phases, respectively. The total Mechanical Turk cost is approximately US \$1,200.

7.6 Culture-Specific Humor Generation

The proposed framework for automatic culture-specific humor generation for Fun Libs has three main stages:

1. A **candidate selection** stage in which possible word replacements for the blanks in each sentence of the Fun Lib stories are generated using a language model, also taking into the account the desired cultural slant.

2. A **candidate ranking** stage in which the selected candidates are ranked according to their funniness contributions to the sentences they occur in, using culture-specific machine learning classification.

3. A **story completion** stage in which funny Fun Lib stories are created by selecting the top ranked humorous transformations for each sentence, and concatenating them to obtain complete stories.

Each stage is described in detail below.

7.6.1 Candidate Selection

The first stage of the framework involves generating candidate words for each blank in the stories. To provide reasonable candidates, the BERT (Bidirectional Encoder Representations from Transformers) masked language model [169] is used. BERT is based on the multi-layer bidirectional Transformer [170] architecture, and is trained on plain text for masked word and next sentence prediction tasks. All the BERT models are pre-trained on the English Wikipedia (2,500M words) and the Book Corpus (800M words) [171] datasets. Throughout this chapter, the $\text{BERT}_{\text{base}}$ model is used, which has the following model size: L=12, H=768, A=12, Total Parameters=110M, where L denotes the number of layers, H the hidden size, and A the number of self-attention heads. In order to predict candidate word fillings for the blanks for each culture, the pre-trained BERT model is trained further[2] using the pre-training objective (combination of masked language modeling and next sentence prediction loss) on culture-specific text datasets authored by people of Indian or US origin. The datasets used for fine-tuning consist of the blog posts obtained from Google Blogger from the years 2000 to 2016 authored by Indian and US bloggers, and are obtained in a similar way as in Section 5.4.

To generate candidate words for each blank, all the blanks in each sentence are masked,[3] and the resulting masked sentences are given as input to the fine-tuned culture-specific language model, and the probability scores are obtained for all the words in the vocabulary, estimating how well each of them fits into the input sentence at the corresponding position. The candidates are sorted according to their scores in decreasing order, and the top $k = 10,000$ words are selected for each blank. The top k words are further filtered so as to obtain a cleaner candidate list adhering to the hint types and other restrictions imposed in the data annotation stage, according to the following filters.

- The candidate must belong to WordNet [172] or any other English dictionary.
- The candidate must contain only alphabetic characters (no spaces, numbers or other special characters).
- The candidate must have a part-of-speech tag that agrees with the hint type (*animal*, *body part*, *food* and *liquid* are considered to be nouns).
- For candidates which are nouns, the plurality should match with that of the hint type, and should fit in the context accordingly (*The cats eats the food* is changed to *The cat eats the food*).
- For verbs, the candidates must fit in the contexts in terms of the verb tenses (*The cat is throwing the food* is valid).
- The candidates must not be slang, adult or bathroom-related words (filtered using existing word lists for profanity), as they are very crude, and result in shallow, easy and repetitive humor.

In sentences with more than one blank, the candidate selection happens in a left-to-right manner where the candidates for first blank are chosen first, and from the next blank

[2] All the fine-tuning experiments in this work are done using the Pytorch framework https://github.com/huggingface/pytorch-transformers

[3] The blanks in each sentence are replaced with [MASK] symbol to have the sentences resemble the pre-training data for BERT.

onwards, candidates are chosen after filling in all the previous blanks. It is to be noted that with different word fillings for the previous blanks, different sentences are obtained, and hence different candidates are predicted for the next blanks. For each blank, instead of considering all the k candidates, the candidates are ranked using a humor classifier trained for each culture separately on culture-specific humor datasets, and the top-ranked $n = 100$ candidates are chosen to fill in each blank, and the hence obtained sentences are used for candidate selection for next blank using the language model. This interplay between the candidate selection for each blank using language model and candidate ranking for that blank using a classifier is repeated, until all the blanks are filled-in in each sentence. This kind of candidate selection in a left-to-right manner enhances the overall coherence of the resulting funny sentence, as the previously selected candidates are considered in selecting next candidates; otherwise, each filled-in word in a sentence may contribute to the overall humor, but they may be incongruous, resulting in a random sentence.

For a given context, it is expected that the candidates with high scores from the language model are very good fits, and hence are less humorous (*The cat is eating the food*), while those candidates with low scores are more likely to be incongruous, and hence may generate humor in the form of surprise in the given context (*The cat is preparing the food*). The classifier ranking is described in detail in the next section.

7.6.2 Candidate Ranking

Examining the position of the candidates in their corresponding lists or their scores from the language model is not sufficient to determine if they are funny substitutes for the given blanks. The second stage of the framework involves ranking the candidates selected for each blank based on their humor contributions to the containing sentence. For this, a sentence-level machine learning classification model is developed by fine-tuning the BERT model, and the objective of this model is to predict whether a sentence filled-in with candidates from the language model is a humorous transformation of the original masked sentence. BERT is the first fine-tuning based representation model that achieves state-of-the-art performances on several sentence-level and token-level tasks, even outperforming many task-specific frameworks. Moreover, the limited sizes of the annotated humor datasets make BERT a suitable framework to build on.

Specifically, the BERT model is fine-tuned for the sentence pair classification task[4] by adding an additional layer for classification, where the input sequence pair (<sentence `A`, sentence `B`> from pre-training) corresponds to <`A`: masked sentence, `B`: filled-in

[4] The name for this task in BERT terminology is MRPC (Microsoft Research Paraphrase Corpus), as this task was originally proposed for paraphrase classification task on the Microsoft Research paraphrase corpus.

SENTENCE A	SENTENCE $\mathrm{B}_{\text{India}}$	SENTENCE B_{US}
Richard Nixon, who was the [MASK], watched the [MASK] from the White House.	Richard Nixon, who was the janitor, watched the rocket from the White House.	Richard Nixon, who was the weasel, watched the debacle from the White House.
Kangaroos are marsupials because they [MASK] their young in a/an [MASK] pouch on their bodies.	Kangaroos are marsupials because they hate their young in a rigid pouch on their bodies.	Kangaroos are marsupials because they beat their young in an ugly pouch on their bodies.
Amazon Mechanical Turk is a crowd-sourcing site enabling [MASK] and businesses to coordinate the use of human intelligence to perform [MASK] that computers are currently unable to do.	Amazon Mechanical Turk is a crowd-sourcing site enabling potatoes and businesses to coordinate the use of human intelligence to perform yoga that computers are currently unable to do .	Amazon Mechanical Turk is a crowd-sourcing site enabling introverts and businesses to coordinate the use of human intelligence to perform atrocities that computers are currently unable to do.

Table 7.3: Training examples with input sequence pairs that belong to the **funny** class.

sentence>. Internally, the fine-tuning model takes as input the sequence <[CLS] A [SEP] B>, where [CLS] is a special classification token used in BERT and [SEP] is a special token to separate the sentences in the input pairs and the output is the prediction $c \in \{$funny, not funny$\}$ whether the sentence pair corresponds to a funny transformation. BERT takes the final hidden vector $h \in \mathbb{R}^{\mathrm{H}}$ corresponding to the first input token [CLS] as the representation for the whole sequence. The only new parameters added during fine-tuning are classification layer weights $W \in \mathbb{R}^{\mathrm{C} \times \mathrm{H}}$, where C is the number of labels. The probability of each label c is

$$p(c|h) = \mathrm{softmax}(Wh) \tag{7.1}$$

All the parameters from BERT and W are fine-tuned jointly by maximizing the log-probability of the correct label.

7.6.2.1 Culture-Specific Humor Classification

Of the 50 Fun Lib stories, 40 stories are used for training and the evaluations are carried out on the remaining 10 stories. The training and test samples are the same as those used by [100], so as to allow for comparison between studies.[5] The full training dataset includes 40 Fun Lib stories, each filled-in by 3 players from each culture, and each of the filled-in stories graded by 5 judges from the corresponding culture. For each culture, the dataset is divided into 30 training stories and 10 validation stories, with the same division as in

[5] The 4 new stories that are created are added to the training set, as the replaced stories also occur in the training set.

[100], resulting in 90 filled-in stories for training and 30 filled-in stories for validation. The stories are further split into sentences,[6] and the training and validation datasets consist of sentence pairs and labels from the 90 and 30 stories respectively. Labels are assigned to each sentence pair using **majority vote** on the funniness contribution judgements for each of the filled-in words: a filled-in word is considered "funny," if it has three of more "funny" judgements (out of the 5 judgements), else it is considered "non-funny." A sentence is considered **funny** if at least 50% of the filled-in words in it are funny (if there is only one filled-in word in the sentence, then it has to be funny), else the sentence is considered **not funny**. Sentences that do not contain blanks are not considered for training, as they do not add any information about humor. Table 7.3 shows a few examples of input sentence pairs, where sentence B is filled-in by the qualified players on AMT from India or the US, and belongs to funny class.

TYPE	FUNNY	NOT FUNNY
	INDIA	
TRAINING	566	130
VALIDATION	173	49
TEST	137	94
	US	
TRAINING	574	122
VALIDATION	193	29
TEST	210	21

Table 7.4: Culture-wise, label-wise statistics on the humor classification datasets.

Table 7.4 shows the number of funny and non-funny examples in the training, validation and test sets for India and the US. Since the datasets are collected by instructing the players on AMT to fill in the blanks to make the stories humorous, they are skewed towards the funny class. To balance the datasets, they are augmented with additional input sentence pairs that belong to the not-funny class. For this, sentences are randomly sampled from the English Wikipedia dataset, such that their word counts are in the range $[\mathrm{M}-5, \mathrm{M}+5]$, where M is the median of the word counts of the existing sentences. The input sentence pairs are obtained by masking one word belonging to one of the four part-of-speech tags[7] in the Wikipedia sentences. Instead of augmenting sentence pairs with the exact Wikipedia sentences to the Indian and US datasets (and thus resulting in partial dataset overlap between them), **modified Wikipedia sentences** are augmented, which are essentially the

[6] Sentence breaks are pre-defined in the Fub Lib stories created in [100].

[7] The other hint types are not considered as it is not trivial to identify words belonging to them. Part-of-speech tags, on the other hand, can be identified using a standard tagger.

original sentences with the masked word replaced with the top-scored word predicted by the culture-specific fine-tuned language model (the word with the highest score). By doing so, the augmented non-funny sentence pairs are not only different for each culture, but also have culture-specific word replacements in them. This results in an equal number of funny and non-funny examples in the training and validation datasets for each culture (1,132 and 346 India; 1,148 and 386 for US).

For each culture, the BERT model is fine-tuned for the sentence-pair classification task on the training data from that culture with a batch size of 32, vocabulary size of 30,522 and the `gelu` [173] activation function. BERT can process sequences of length no more than 512, and all the input sentences have less than 512 tokens. As used in the pre-trained BERT, the Adam optimizer is used for fine-tuning, and a learning rate of 1e-5 (from 5e-5, 1e-5, 5e-6, 1e-6) is selected using the validation set. The number of epochs is set to 10, as the highest validation accuracies are achieved using 10 epochs for both the cultures. The resulting humor classifier is called **FunnyBERT** for the rest of this chapter.

Metric	Train	Val
FunnyBERT$_{\text{India}}$		
Precision (Funny)	99.82	81.48
Recall (Funny)	100.00	89.02*
F1 score (Funny)	99.91	85.08
Accuracy	99.91	**84.39**
Accuracy (Exact Wikipedia sentences)	99.91	80.06
FunnyBERT$_{\text{US}}$		
Precision (Funny)	100.00	91.16*
Recall (Funny)	100.00	85.49
F1 score (Funny)	100.00	88.24*
Accuracy	100.00	**88.60***
Accuracy (Exact Wikipedia sentences)	100.00	**89.12***
Majority (Baseline)		
Accuracy	50.00	50.00

Table 7.5: Culture-wise training and validation accuracies of BERT for the sentence-pair classification task ($p < 0.05$).

Table 7.5 shows the training and validation accuracies of FunnyBERT and majority vote classifier as baseline. Since both the training and validation sets are class-balanced, the corresponding baseline accuracies are 50%. The results that are significantly better than the corresponding ones from the other culture according to the two-sample t-test are marked with *. As can be seen from Table 7.5, the validation accuracies (in bold) for both cultures are significantly higher than those of the baseline. For India, the funny-class recall is higher and precision is lower than those for US. This may be due to a slightly lower quality of funny sentence completions from Indian players compared to those from US players, resulting in the India-specific FUNNYBERT predicting a non-funny sentence also as funny in most cases, thus resulting in a high recall and low precision. This is also reflected in the lower validation accuracy for India (84.39%) compared to that for US (88.60%). It is to be noted that the datasets that BERT is pre-trained on are more closely aligned to US English than to Indian English, and this may also have contributed to the lower validation accuracy and F1 score of FUNNYBERT for India.

Also, when the exact Wikipedia sentences are augmented to the non-funny class without culture-specific word replacements to class-balance the training datasets, the validation accuracy for India (80.06%) is significantly lower than when the replacements are included (84.39%), while for US the validation accuracy is slightly higher (89.12%), though this improvement is not statistically significant (compared to 88.60%). This supports the claim that the datasets used to pre-train BERT are biased towards US English, and hence India-specific replacements, however minor they are (as only one word in each sentence is replaced), result in an improved classifier performance for India.

The FUNNYBERT classifier is used to rank the candidate filled-in sentences for each masked sentence based on their funniness levels (in terms of the softmax probabilities from the output of the classification layer). The top sentences ranked by the classifier are either used for candidate selection of the next blank if it exists ($n = 100$), or are fixed as the top humorous transformations of the original sentence if the sentence has no more blanks, which are later used in the next stage to form complete stories that are both funny and coherent.

7.6.3 Story Completion

The final stage of the framework involves forming complete stories from the top funny transformations for each of constituting masked sentences. Similar to candidate selection, the story completion is done in a left-to-right manner, with the following two constraints for selecting top candidates for each subsequent sentence: (1) they must be funny, and (2) they should be logically the next possible sentences after the previously selected sentences, to ensure that the resulting stories are both funny and coherent. To ensure (1), the funny transformations corresponding to each subsequent sentence, as obtained from candidate ranking, are considered. To address (2), for each sentence, the funny transformations are ranked based on their semantic similarity to the previously selected sentences, and the top

$N = 100$ funny and similar transformations are selected, and this process of sentence selection is repeated until the story is complete.

The similarity between any two filled-in sentences is measured in terms of the cosine similarity between the corresponding sentence embeddings. For each sentence, the average of the embeddings of the words in it obtained from culture-specific fine-tuned BERT model is considered as the sentence embedding, throughout this chapter. Since BERT has $L = 12$ layers, each word token has 12 separate vectors, each of length $H = 768$. Different layers of BERT encode different kinds of information, so the appropriate aggregation strategy depends on the specific task of interest. Two variations of obtaining the final word vectors are considered: (i) summing the embeddings from the last 4 hidden layers, and (ii) considering the embeddings from only the second to last hidden layer. The second strategy is employed from here on, as the story completion done using sentence embeddings obtained from the **second last hidden layer** resulted in better quality stories for the validation set.[8]

To illustrate this, let us consider a story S made up of three sentences, $S =< S1, S2, S3 >$. For $S1$, the top $N = 100$ funny transformations are selected from candidate ranking. To obtain the possible variations with $S2$, for each selected $S1$, the $N = 100$ most similar candidates for $S2$ are selected, resulting in a total of $100,000$ possible $< S1, S2 >$ combinations. From these, again the top $N = 100$ combinations based on the similarity scores are advanced to the next stage of selection for $S3$, and this process is repeated until the variations for the last sentence are processed. This constraint of having only the top $N = 100$ combinations at the end of processing each sentence is to avoid the exponential time complexity in considering all possible combinations. For each original story, the culture-specific humor generation framework results in the top $N = 100$ funny stories, with each of them having the coherence constraint imposed through the left-to-right candidate selection and story completion strategies.

7.7 Evaluation

In this section, three approaches for generating humor in Fun Libs are evaluated,[9] and the results are compared.

1. **FREETEXT**: Players fill in the blanks with the constraint of making the stories humorous. These stories are collected in the data collection stage.

[8] The quality of stories is determined by human subjects, as this task is subjective.

[9] Each of these approaches is analogous to those in [100], except that these approaches (except for FREETEXT) are fully automated, while the ones in [100] require human assistance.

2. **MLM**: For each blank in a story, the word-replacement is selected based on its probability score using the pre-trained BERT masked language model (without any culture-specific fine-tuning).

3. **PROPOSED**: The stories are generated using the proposed framework, as described in Section 7.6.

For evaluation, the 10 test Fun Lib stories are used. The stories from FREETEXT (by players) and PROPOSED approaches are generated in culture-specific settings, while the MLM approach does not take into account the desired cultural slant. In the MLM approach, stories are ranked based on their filled-in word probabilities of fitting in the corresponding sentences. Hence, the top-ranked stories in this approach are expected to be more congruous, and hence less humorous than the low-ranked stories. The stories obtained using the PROPOSED approach are first sorted in the decreasing order of their **story funniness**, and further sorted in the decreasing order of their **average filled-in word coherence**. **Story funniness** is defined as the mean of the funniness scores (obtained from the classifier's softmax probabilities) pertaining to each constituting sentence. **Average filled-in word coherence** is the average of the pair-wise similarities between word embeddings of the filled-in words, obtained using the second last hidden layer from the culture-specific fine-tuned BERT.

For each test story, the grading is done by judges from both the cultures on (i) the 3 filled-in stories from FREETEXT approach, (ii) the top 10 generated stories from MLM, to verify if funniness increases as less congruous words are filled-in, and (iii) the top 10 generated stories from the culture-specific PROPOSED approach. More number of stories are graded from the PROPOSED approach for two reasons: (1) unlike the FREETEXT stories, which need human assistance and hence time and money, these stories are generated automatically, and hence require minimal effort (only requires to obtain the training data to train the FUNNYBERT component of PROPOSED), and (2) both the story funniness and the average word coherence scores for the top 10 stories differ only in their fourth decimal place, and hence together are treated as the best diverse humorous variations generated for the given stories (BEST10).

Table 7.6 shows the averages of the mean funniness grades of the stories obtained using each of the three approaches. The mean funniness grade for a story is the mean over the judgements given to that story by 5 judges, and it ranges from 0 (not funny) to 3 (funny). The averages are presented for two settings for the MLM and PROPOSED approaches.

1. 30 STORIES: For MLM, the top three generated stories (based on the filled-in word probabilities) are considered. For PROPOSED approach, those three filled-in stories

APPROACH	INDIAN JUDGES	US JUDGES
FREETEXT$_{\text{India}}$	$1.17^{\ddagger}$	$\underline{1.39}^{\ddagger\P}$
FREETEXT$_{\text{US}}$	$1.57^{*\ddagger}$	$1.41^{\ddagger}$
	30 STORIES	
MLM	0.70	0.68
PROPOSED$_{\text{India}}$	$\underline{\mathbf{1.94}}^{*\dagger\ddagger\P}$	$1.56^{*\ddagger}$
PROPOSED$_{\text{US}}$	$\underline{\mathbf{2.03}}^{*\dagger\ddagger\P}$	$\mathbf{1.77}^{*\dagger\ddagger\S}$
	100 STORIES	
MLM	0.91	0.84
PROPOSED$_{\text{India}}$	$\underline{1.60}^{*\ddagger\P}$	$1.32^{\ddagger}$
PROPOSED$_{\text{US}}$	$\underline{\mathbf{1.70}}^{*\dagger\ddagger\S\P}$	$1.48^{*\ddagger\S}$

Table 7.6: Averages of the mean funniness grades for the stories generated via the three humor generation approaches. In each column, the scores that are significantly higher than (i) FREETEXT$_{\text{India}}$ values are marked with *, (ii) FREETEXT$_{\text{US}}$ with †, (iii) MLM with ‡, and (iv) PROPOSED$_{\text{India}}$ with §. In each row, the values that are higher than those in the other column are marked with ¶ ($p < 0.05$). The column-wise significantly highest values are in **bold** font, and the row-wise highest values are underlined.

among the BEST10 for each story title are considered which have the highest mean funniness grades (as all the 10 of them have very similar funniness scores assigned by the framework).

2. 100 STORIES: The top 10 (based on filled-in word probabilities) and BEST10 generated stories for each of the 10 titles are considered in this setting for the MLM and PROPOSED approaches respectively.

As can been seen from Table 7.6, the humor generated by both the FREETEXT$_{\text{X}}$ and PROPOSED$_{\text{X}}$ ($\text{X} \in \{\text{India}, \text{US}\}$) approaches are preferred to that by MLM, by both Indian and US judges. This is expected, as MLM fills the stories with the most plausible words, and hence does not introduce any surprise or creativity that is essential for humor generation. The average mean funniness grade for MLM increases significantly in the 100 STORIES setting, confirming that the stories become more humorous as more incongruous words are selected for filling-in the blanks (with lower probability scores). In the FREETEXT approach, both Indian and US judges prefer US humor (1.57, 1.41) to Indian humor (1.17, 1.39), though for US judges, the difference is not significant. On the other hand, between the Indian and US judges, Indian humor is liked more by US judges (1.39), while US humor is preferred by Indian judges (1.57) (though the latter score is not significantly higher than that by US judges 1.41). Hence, in general US humor is preferred to Indian humor by an average judge on AMT, whereas within the culture-specific settings, Indian humor is liked more by US judges and US humor by Indian judges. That is, between India and US, an average US turker is better at generating humorous stories than an average Indian turker, and judges in general enjoy humor generated by turkers from different culture more than that by those from their own culture. This is in contrast to the general expectation that people prefer humor originated from their own culture due to a better understanding of the various culture-specific subtleties, suggesting that seeing things in a new light is possibly contributing to the surprise and creativity factors of humor generation.

In the machine-only PROPOSED approach as well, US humor is in general preferred to Indian humor by both Indian (2.03, 1.70) and US (1.77, 1.48) judges, in both 30 STORIES and 100 STORIES settings. This may be due to the better performance of FUNNYBERT$_{\mathrm{US}}$ in comparison to FUNNYBERT$_{\mathrm{India}}$ in humor classification of sequence pairs, resulting in a better ranking of funny sentences for story completion. Also, as seen for the FREETEXT approach, the stories filled-in by US turkers are on an average of better quality humor than those by Indian turkers, and this may have led to better quality training datsets for FUNNYBERT$_{\mathrm{US}}$, resulting in its better performance. Almost always, the PROPOSED approach outperforms FREETEXT approach in generating humor for both Indian and US judges (more so in the 30 STORIES setting for US judges, and both the settings for Indian judges), indicating that the fully-automated PROPOSED approach not only learns and generates quality humor, but also does so in a much better and creative manner than humans. Moreover, this approach also outperforms Libitum [100], which is a semi-automated computer-aided framework to help humans fill in the Fun Libs, in the 30 STORIES setting. Libitum has an average mean funniness grade of 1.51, which is significantly lower than that given by US judges to US humor (1.77).[10]

Within the culture-specific scenarios in the PROPOSED approach, both the Indian and US humor are liked more by Indian judges than by US judges, in both the settings (underlined in Table 7.6). This suggests that an average Indian on AMT has a more lenient outlook towards humor, whether it is from Indian or US origin, and this may have resulted in the lower quality of humor generated by Indian turkers, whereas an average US turker has a stricter and more extreme perspective towards it, resulting in the more enjoyable humor by US turkers.

Approach	Indian judges			US judges		
	Coh	Incon	Dev	Coh	Incon	Dev
FreeText$_{\text{India}}$	**0.396**	0.245	**0.587**	0.013	**0.675**	0.187
FreeText$_{\text{US}}$	0.225	0.239	**0.620**	**0.767**	**0.826**	**-0.592**
MLM	**-0.567**	**0.742**	**0.656**	**-0.640**	**0.359**	**0.718**
Proposed$_{\text{India}}$	**0.219**	**0.413**	**0.519**	0.152	0.117	**0.222**
Proposed$_{\text{US}}$	0.074	**0.274**	**0.405**	0.116	**0.267**	0.046

Table 7.7: Correlations of coherence (Coh), incongruity (Incon) and deviation (Dev) values with the corresponding mean funniness grades. The correlations that are statistically significant with $p < 0.05$ are in **bold**.

7.8 Discussion

Table 7.7 shows the correlations of coherence, incongruity and deviation scores with the corresponding mean funniness grades for the test stories generated via the three approaches. For MLM and Proposed approaches, the correlations are computed over the 100 generated stories for the given 10 test stories. In the FreeText approach, coherence plays a very important role in generating humor for audience from the same cultural background (correlation of 0.396 for India and 0.767 for US). In other words, humans have the striking ability to apply coherence while generating humor, and this is particularly appreciated by audience from a similar cultural background, possibly due to a consistent use of culture-specific information to generate humor. This can be seen in the judgements of US judges on Indian humor, where funniness has no correlation with coherence, and high correlation with incongruity, indicating that US judges find those Indian authored stories funny which contain seemingly unexpected words.

On the other hand, coherence is negatively correlated with the graded humor for the MLM approach, indicating the difficulty in generating humor via coherent words by a language model. Instead, most of the words generating humor in this approach are incongruous, that introduce a shift (deviation) in the original topic of the stories, as given by their titles, indicated by the statistically significant correlations of incongruity and deviance with the mean funniness grades. In the Proposed approach, coherence plays very little role in generating humor, and most of the funniness is achieved via the use of incongruous words and topic deviations, though Indian judges, to an extent, find automatically generated Indian humor to be coherent (0.219).

[10] Since Libitum was trained and evaluated on data collected from USA, Canada, Great Britain and Australia, it is closer to the US humor judged by US judges setting, and hence is compared against Proposed$_{\text{US}}$ only. Also, Libitum generated and evaluated 30 stories for the given 10 test stories, hence the comparison is done for the 30 Stories setting.

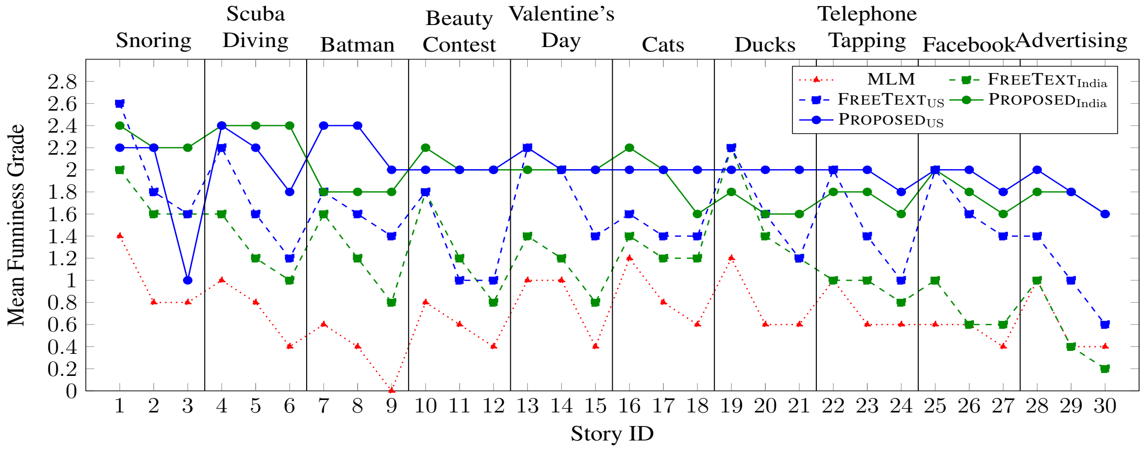


Figure 7.2: Mean funniness grade for the 30 test stories using each of three approaches graded by Indian judges.

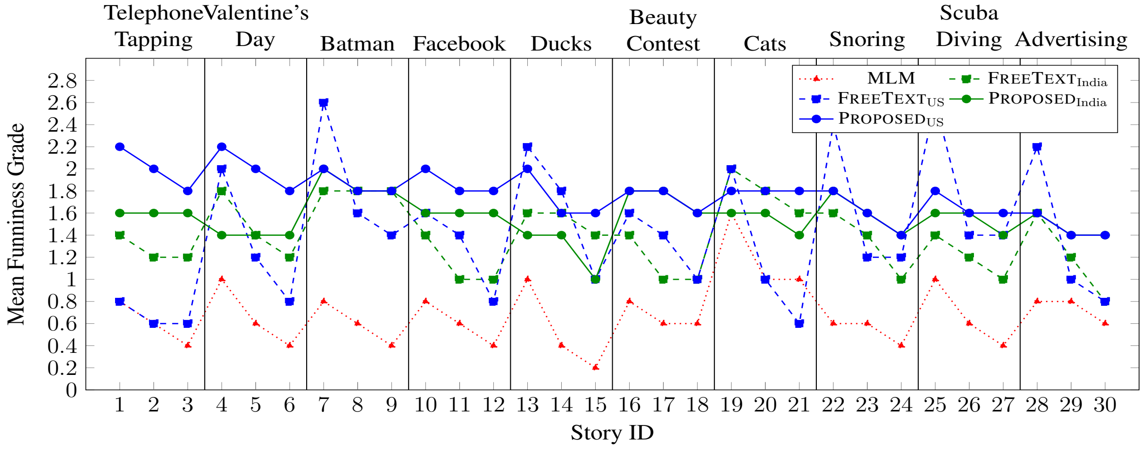


Figure 7.3: Mean funniness grade for the 30 test stories using each of three approaches graded by US judges.

In general, humor in most cases is positively correlated with incongruity and topic deviation for all the approaches, indicating that funniness is in general achieved by using incongruous words, and changing the overall topics of the stories from those suggested by their titles. The only case when topic deviation negatively affects the funniness grades is for US humor via the FREETEXT approach judged by US judges. Also, skilled humans are capable of generating meaningful and coherent humor, something that the PROPOSED approach finds very difficult to achieve.

GRADE	FREETEXT		MLM	PROPOSED	
	IN HUMOR	US HUMOR	-	IN HUMOR	US HUMOR
IN JUDGES					
0	39	18	59	9	12
1	64	55	39	31	34
2	30	58	46	68	75
3	17	2	6	42	29
US JUDGES					
0	20	27	59	10	13
1	46	52	49	82	62
2	63	54	39	50	62
3	21	17	3	8	13

Table 7.8: Funniness judgement counts from Indian (IN) and US judges for the 30 test stories from the three approaches.

7.8.1 Qualitative Analysis

Figures 7.2 and 7.3 show the mean funniness grades by Indian and US judges respectively for the stories generated via each of the three approaches in the 30 STORIES setting. For better visualization, the story titles are sorted in the decreasing order of their mean funniness grades from left to right. The FREETEXT approach consistently outperforms the MLM approach for both Indian and US judges, while the PROPOSED approach outperforms FREETEXT for 28 and 23 stories respectively for Indian and US judges. When judged by Indian judges, the average improvement in the mean funniness grades of the PROPOSED stories over FREETEXT stories is 0.79 for Indian humor and 0.46 for US humor. By US judges, the corresponding average improvements are lower (0.17 and 0.37 respectively). This suggests that an average Indian judge, though preferring coherence as well, finds stories with less coherence and more incongruous words and topic deviations also funny, and since these are further incorporated by the PROPOSED approach, the improvements in the Indian judgements are higher. On the other hand, US judges prefer coherence along with incongruity, and hence the improvements in the US judgements are lower from FREETEXT to PROPOSED stories. This can be further illustrated in the fact that the highest mean funniness grades by Indian judges are for the PROPOSED approach stories (*e.g.*, “Snoring”, “Scuba Diving”, “Batman”), while those by US judges are for the FREETEXT stories (*e.g.,* “Batman”, “Snoring”, “Scuba Diving”).

Table 7.8 shows the number of judgements given by Indian and US judges for each of the four grades for the evaluated stories. For the FREETEXT stories, Indian judgements ’0’ and ’1’ are more in number for Indian humor (written by Indian turkers on AMT), while

Scuba Diving is a sport where people can swim under *beer* / feather for a long time, using a tank filled with compressed *nitrogen* / cigarette. The tank is a *hollow* / dusty cylinder made of steel or *rubber* / star. A scuba diver *walks* / burns underwater by using fins attached to the *legs* / windows. They also use *coconut* / hospitality such as a dive mask to *hide* / burn underwater vision and equipment to control *air* / crouch. A person must take a *course* / cigarette class before going scuba diving. This proves that they have been trained on how to *bunk* / punch the equipment and dive *deep* / cold. Some tourist attractions have a *simple* / deadly course on certification and then the instructors *bunk* / sing the class in a *simple* / textile dive, all in one day.
Cats are the most *naughty* / foolish pets in the world. They were probably first kept because they ate *dogs* / lawyers. Later cats were *praised* / disgusted because they are *ferocious* / psychotic and they are good *thinkers* / radicals. Cats are active carnivores, meaning they hunt *small* / wicked prey. They mainly prey on small mammals, like *dogs* / dice. Their main method of *hunting* / baking is stalk and *kiss* / kid. While dogs have great *allergy* / gore and they will catch prey over long distances, cats are extremely *romantic* / creepy, but only over short distances. The cat creeps towards a chosen victim, keeping its *mouth* / buster flat and near to the *dog* / booth so that it cannot be *sensed* / cashed easily, until it is close enough for a rapid *kiss* / neutron or pounce.

Table 7.9: Example stories with highest improvements of PROPOSED over FREETEXT with filled-in words in the following order: *FreeText* / Proposed for Indian humor, and *FreeText* / Proposed for US humor.

the judgement '2' is higher for US humor. The US judgements '0' and '1' are more in number for US humor, and '2' and '3' for Indian humor. This further demonstrates the previous observation that Indian humor written by Indian turkers on AMT is liked more The judgements from Indian judges on the proposed stories are skewed towards the grades '2' and '3', while those on the stories from MLM are towards the grades '0' and '1'.

Table 7.9 shows two stories, one of Indian humor and one for US humor, for which the PROPOSED approach outperforms the FREETEXT approach by a large margin when graded by Indian judges. As expected, incongruity is significant in generating humor in the stories via the PROPOSED approach (*e.g., tank filled with* nitrogen vs *tank filled with* cigarette in "Scuba Diving," *they ate* dogs *vs they are* lawyers in "Cats"), while the FREETEXT stories, though being less humorous, are slightly more coherent (*e.g., scuba diver* walks *underwater using fins attached to* legs in "Scuba Diving," *main method of* hunting *is stalk and* kiss *... cats are extremely* romantic in "Cats").

Tables 7.10 and 7.11 show a few Indian and US stories that received the highest mean funniness grades via the FREETEXT approach. Note that these grades are by US judges, as Indian judges frequently give higher grades to stories obtained via the PROPOSED approach. The strategy used by the Indian turker in Table 7.10 is to consistently portray Valentine's Day as a training process that happens for dramas, and it is named after a mad bishop who performs dramas for people. Similarly, US turkers in Table 7.11 describe Snoring as a disturbing and murderous noise that is symbolic of stupidity and lack of valor, and

Valentine's Day is a *training* / pigeon that happens on February 14. It is the day of the year when lovers show their *drama* / bird to each other. This can be done by giving *sticks* / sports, flowers, Valentine's cards or just a/an *bad* / ultraviolet gift. Some people *kill* / drink one person and call them their "Valentine" as a gesture to show *love* / nero and appreciation. Valentine's Day is named for the *mad* / kind Christian saint named Valentine. He was a bishop who performed *dramas* / strings for couples who were not allowed to get married because their *friends* / goddesses did not agree with the connection or because the bridegroom was a soldier or a/an *joker* / eel, so the marriage was *solved* / chickened. Valentine gave the married couple flowers from his *kitchen* / name. That is why flowers play a very *tasty* / mythological role on Valentine's Day. This did not *attract* / bat the emperor, and Valentine was *praised* / melted because of his Christian *drama* / bath.

Table 7.10: The best "Valentine's Day" story obtained for Indian humor. Filled-in words in the following order: *FreeText* / Proposed.

Scuba Diving as a violent sport involving deadly dives and fatal courses that takes place under lava, and intended to destroy the equipment and avoid panicking while doing so. Batman is portrayed as a bland and uninteresting person who is weak, cowardly and moronic, living in a disconcerted city of Gotham. Humor in these stories is generated by portraying the title concepts as notions that are surprising and unexpected, and doing so consistently. It is also interesting that Batman is depicted with two main characteristic traits consistently, one with respect to the bland and moronic person that he is, and the other about his opposition to romance in Gotham due to a robber kissing his parents. This is striking, as it illustrates how good skilled humans of both Indian and US origin can be at generating humor via multiple coherent and meaningful concepts, a feature that can be very difficult for machines to achieve.

Much of the humor generated via the PROPOSED approach is through incongruity – the filled-in words are funny as they do not match the expectation of the reader (*e.g., Snoring is the robot that people make when they are brooding, a scuba diver steals underwater using fins attached to the bombs, Batman is an abusive superhero*).

Occasionally, the PROPOSED approach is able to generate small snippets of coherent and funny phrases. For example, the tourist attractions of Scuba Diving are described as having risky courses where the instructors attack the class; Batman is described as an abusive superhero who was fiery as a child, and grew up learning different ways to glare; Valentine's Day is named after a kind bishop, who performed strings for couples whose marriages were not allowed due to the disagreement of their goddesses, and the flowers given by him play a mythological role.

Also, some of the culture-specific concepts surface in the stories generated via the PROPOSED approach. Some examples for Indian stories include Valentine's Day with reference

Snoring is the *disturbance* / robot that people often make when they are *relaxing* / brooding. It is often caused by a blocked *intestine* / laugh or throat. The noise is often *murderous* / societal, as it is made by *air* / madman passing through the *nasal* / presidential passages or the *mouth* / runway. Research suggests that snoring is one of the factors of *job* / fame deprivation. It also causes daytime *stupidity* / bum, irritability, and lack of *valor* / buck. Snoring can cause significant *moral* / naval and social damage to *sociopaths* / wolves. So far, there is no certain *milkshake* / available that can *completely* / bunny stop snoring.
Scuba Diving is a sport where people can swim under *lava* / genie for a long time, using a tank filled with compressed *dough* / brute. The tank is a *rusty* / stray cylinder made of steel or *coconut* / sailor. A scuba diver *sprints* / steals underwater by using fins attached to the *ears* / bombs. They also use *camouflage* / mister such as a dive mask to *inhibit* / pardon underwater vision and equipment to control *panicking* / bombing. A person must take a *skydiving* / crack class before going scuba diving. This proves that they have been trained on how to *destroy* / sniff the equipment and dive *violently* / willingly. Some tourist attractions have a *fatal* / risky course on certification and then the instructors *leave* / attack the class in a *deadly* / curly dive, all in one day.
Batman is a fictional character and one of the most *bland* / abusive superheroes. He was the second *wimp* / pest to be created, after Superman. Batman began in comic books and he was later *exposed* / fired in several movies, TV programs, and books. Batman lives in the *discombobulated* / tool city of Gotham. When he is not in *costume* / wink, he is Bruce Wayne, a very *dense* / pacific businessman. Batman's origin story is that as a *moronic* / fiery child, Bruce Wayne saw a robber *kiss* / spin his parents after the family left a *bakery* / teddy. Bruce decided that he did not want that kind of *romance* / wayne to happen to anyone else. He dedicated his life to *terrorize* / tolerate Gotham City. Wayne learned many different ways to *grow* / glare as he grew up. As an adult, he wore a *diaper* / bird to protect his *skin* / drill while fighting *nuns* / pneumonia in Gotham.

Table 7.11: The best stories for "Snoring," "Scuba Diving" and "Batman" obtained for US humor. Filled-in words in the following order: *FreeText* / Proposed.

to *gods*, *goddesses* and *mythology*, Batman fighting *bombay* in Gotham, etc., while for US stories, a Beauty Contest is described in terms of *yankee* contest, Batman fights *Roosevelt* in Gotham and so on. The relationship between culture-specific word occurrences in the generated stories and the culture of the target audience is to be explored in detail in the future.

7.9 Conclusion

In this chapter, I studied the problem of culture-specific automatic humor generation for the Mad Libs-style fill-in-the-blank game to make stories humorous in a culture-specific setting. For this, first a novel culture-specific humor dataset for stories in the Mad Libs format is created on Amazon Mechanical Turk, by selecting a pool of players and judges to obtain ground truth data. Next, an automated culture-specific humor generation framework is developed to first generate possible candidate words for filling-in the blanks, and then rank them based on their humor contributions to form funny sentences. Stories are completed by selecting the best transformations of the constituting sentences to create complete and funny stories. As evaluated by judges from different cultures, the proposed approach outperforms a simple language model, and is usually better than humans in generating humorous stories. Further qualitative and quantitative analyses are conducted to understand what makes the stories generated by the proposed framework humorous, and how they differ from those by a language model and humans. Further culture-specific observations are made about the preferences of Indian and US audience in the humor generated by both humans and the proposed approach.

The analyses show that US humor is in general preferred to Indian humor by both Indian and US audiences on Mechanical Turk, in both human and machine-only settings, indicating a certain edge in the US turkers and the machine learning model trained on the data provided by US turkers to generate funny stories. Indian judges on AMT have a more lenient outlook towards humor, while US judges have higher expectations in terms of the stories being meaningful and coherent, which is also reflected in the better quality of humor generated by both humans and machines in the US-specific setting. When turkers outperform the proposed framework for humor generation, they do so by producing stories that are not only unexpected and surprising but also coherent. This is also demonstrated in the high correlation of coherence and mean funniness grades in stories generated by turkers in culture-specific settings, indicating the tremendous potential of coherence in making stories humorous for Mad Libs. Though preliminary measures are taken to incorporate coherence in the proposed humor generation framework, much of the humor from this approach is achieved by filling-in incongruous and unexpected words that result in a shift in the topics from the original story titles. In the future, I aim to understand what makes textual humor coherent in general, and go beyond the word and sentence similarity measures used in the current framework to generate coherent and funny stories. Also, since the dataset created for this study is demographically rich, I aim to study the humor preferences and variations across other demographic dimensions as well, such as gender, occupation, education and income level.

CHAPTER 8

Conclusions

The goal of this thesis was to explore demographic differences in word usage using computational linguistics techniques, and to examine if these differences can be leveraged to develop better language representations for NLP tasks. In particular, the thesis presented several large scale corpus-based experiments to investigate the following hypotheses: (1) demographic differences exist in word use; and (2) demographic information can be used to develop better language representations. The findings of the thesis are summarized below:

8.1 Research Questions Revisited

At the beginning of this thesis, several research questions were formulated, which were addressed by various experiments conducted throughout the thesis.

1. **Are there significant differences in how words are used by different demographic groups in naturally occurring social media data? If yes, can they and the linguistic features causing them be characterized? How do the word usage differences vary for a wider variety of demographic categories?**
   Assuming the availability of a large collection of social media data, numerous experiments are conducted to determine if there exist words with significant usage differences between different demographic groups.
   Specifically, I studied the differences between Australia and United States based on the words they use frequently in their online writings on Google Blogger. Using a large number of examples for 1,500 words, covering different parts-of-speech, I showed that classifiers trained on linguistic features, such as local and global contexts, socio-linguistic word classes and syntactic relations, can differentiate between word usages of two nationalities (India and US) with an accuracy of 58.36%, compared to chance at 50%. Owing to the frequent usage of these words in personal writings, I consider the word usage differences as a reflection of the demographic

bias in peoples worldviews.

To better understand the nature of usage differences and to explore additional encoding ability by different feature types, I conducted topic modeling experiments on the datasets for the words with the most significant usage differences, to find the main topics of discussion for each of the words in the two countries. Through Feature ablation experiments, I identified contextual and socio-linguistic features as the most significant identifiers of these differences. Finally, a one-versus-all classification to compare Australia or United States against ten different countries suggested that United States is a more "typical" country when it comes to word usage.

Similar experiments are conducted on a larger scale to characterize demographic differences in word usage between a greater variety of countries, industries and genders. Specifically, I focused on three demographic categories, namely location, gender and industry. Using 500 selected target words, I analyzed the blog posts containing these words and written by bloggers from 12 countries, 2 genders and 15 industries. Using linguistic, socio-linguistic and topical features, the classifiers developed resulted in classification performances of at least 20% higher than the respective baselines for all the demographic categories. Topic-based features are the best indicators of the word usage differences, and the correlations between the topic distributions for the words between the different demographic groups indicated similarities and differences between the groups. These results indicate that there are indeed differences in the way the various countries, genders and industries view the world.

2. **Can the demographic information of the authors be utilized to develop a more refined demographic-aware prediction model for word associations?**

   A novel demographic-enhanced data set consisting of demographic-aware word associations is collected via Amazon Mechanical Turk (AMT). The data set consisted of 800 responses for 300 stimulus words, resulting in a total of 176,097 non-spam responses, obtained from a demographically-diverse group of respondents. A selected set of qualitative and quantitative analyses on the responses for the stimulus words showed that the associations varied across genders and geographic locations. Further, the thesis proposed a new demographic-aware association prediction framework based on composite skip-gram architecture, that take into account the demographics of the people behind the language. Through comparative experiments, I showed that this framework surpasses its generic counterpart by significant amounts (23% for gender ($0.13 -> 0.16$), 80% for nationality ($0.05 -> 0.09$)). I regard this as a first step towards demographic-aware NLP.

3. **Do demographics play a role in complex linguistic tasks such as part-of-speech (POS) tagging and syntactic parsing?**
   Existing datasets with demographic information are either too small to train on or lack syntactic information, while sufficiently large syntactic datasets are not labeled with demographic details of the authors. To study the role of demographics on POS tagging and syntactic parsing, an annotation framework is is presented to augment the Wall Street Journal (WSJ) subset of the Penn Treebank (PTB) with the gender information of its contributors.

   Gender-specific experiments show that women's syntax in WSJ displays resilience. That is, POS taggers and dependency parsers trained on WSJ articles authored by any gender perform well when evaluated on female writings. Male syntax, on the other hand, is tagged and parsed best when sufficient male-authored data is available in the training sets. Tag-wise tagging and parsing experiments reveal numerous POS tags and syntactic relations whose prediction performances benefit from the predominance of a specific gender in the training data. These results suggest that leveraging the idiosyncrasies displayed by each gender for specific tags could help develop gender-aware models that utilize the syntactic strengths of each gender, resulting in better models for syntactic tasks.

4. **Does demographic information play a role in humor generation? Do humor preferences vary across demographic groups?**
   With culture as the demographic dimension under consideration, the task of culture-specific humor generation for the Mad Lib-style fill-in-the-blank game is studied, with the aim to generate funny stories in a culture-specific setting. For this, a novel dataset for culture-specific humor generation is created on Amazon Mechanical Turk (AMT). Players and judges are selected from India and United States (US) on AMT to create the ground truth data for 50 stories that approximately match the characteristics of Mad Libs. A culture-specific humor generation framework for filling-in the stories is proposed, which takes into account the cultural background of the people filling-in the stories. Through human experiments, I showed that the proposed framework outperforms a general language model, and in most cases is better than humans in generating funny stories.

   Further analyses indicated that US humor is generally preferred to Indian humor by both Indian and US audience on AMT, in stories generated by either humans or the proposed approach. Indian judges on AMT have a more lenient outlook towards humor, while US judges on AMT are slightly harder to impress. Skilled players on AMT have a striking ability to fill-in the blanks with coherence, a feature that the proposed approach, though accounted for via word and sentence similarity measures, still found difficult to achieve. This sets the path for my future research in this area, where I aim to understand coherence in textual humor in-depth, and go beyond the word and sentence similarity measures to generate funny stories that are coherent as well.

8.2 Discussion

Potential problems of personalisation This thesis studies the differences in the language use between different groups of people, and examines the impact of accounting for their personal information (in the form of their demographic backgrounds) in solving various natural language processing (NLP) tasks. Personalization, in general, refers to fitting services to one's preferences and needs. It has been explored in various areas of NLP, such as machine translation [174], dialog systems [175], query completion [176], search [177], lexical simplification [178], and word associations [44], with some of them resulting in better performances than generic approaches [176, 177, 44]. When applied properly, such studies can be useful in understanding the preferences and behaviors of various groups of people, and can help creating content that caters to different groups of people, thus impacting the business value of various products and services.

However, such modeling can also lead to forming stereotypes for the different groups under consideration. Even simple classification models, developed to predict the author attributes from texts authored by people from different groups, pose a risk of invoking stereotyping and essentialism [139, 179]. Unlike studies conducted on anonymous corpora (newswire) or temporally distant works (novels), the stereotyping resulting from analyzing social media data can affect the people authoring content on these platforms, and can be used to unfairly discriminate against them [139, 180, 181]. However, not conducting further research into the similarities and differences in the language use across various groups is not the answer, because when carried out carefully, such studies can provide valuable insights into people's preferences, which could further positively impact various solutions for them. The said studies have to be conducted soundly, by also addressing the ethical and fairness concerns with respect to them [180].

Generalization of the proposed frameworks The frameworks proposed in this thesis are fairly generic, and can be used for other NLP tasks with limited modifications. The sentence-pair humor classification framework (FUNNYBERT) proposed for humor generation can be used for other monolingual translation tasks, such as paraphrase identification, text simplification, and others which involve transforming one sentence to another.[1] The composite skip-gram framework to predict word associations can be used to predict similar words, and for other downstream tasks such as word sense disambiguation and lexical substitution. In the future, I aim to extend the current frameworks to other language processing tasks, and study in detail the problems arising with respect to the efficiency of the proposed methods while considering composite demographic groups.

Future work and challenges In the future, I aim to extend the explorations in this thesis to account for a wider set of demographic categories, such as age group, education and income level, political affiliation, occupation, and personality type, and a larger collection of demographic groups within each of these categories (e.g., more countries in addition to India and United States). However, it is to be noted that while certain demographic categories (e.g., age group, education level, occupation) are more easily available (mostly from self-representation on social media platforms), obtaining certain others (e.g., political affiliation, personality type) is more challenging, and needs more carefully curated measures to do so. With sufficient data available from the various groups, the frameworks proposed for word associations, part-of-speech tagging and syntactic parsing, and humor generation, can be used for other such categories in a straight-forward manner, by training them on the demographic groups of interest. Future work in this direction also includes accounting for multiple demographic categories, which could be challenging, particularly with the composite skip-gram framework for predicting word associations, as when used for multiple demographic categories, the framework would explode in terms of the space requirements (e.g., *Indian female student* considers country, gender and occupation simultaneously, and hence the the size of the demographically-labeled vocabulary would be multiplied with the number of countries, genders and occupations).

In general, the collection of demographically-labeled data from AMT is limited by the correctness of the turkers about their backgrounds, and their consistency in performing more complex linguistic tasks. Such collection should be directed with carefully-written instructions to avoid any biases that can be imposed from the requesters. Some of the tasks studied in this thesis are subjective (e.g., humor generation), and hence the evaluation involved pre-selected human judges (on AMT) to manually rate the final outputs, whose judgements are again limited by their world knowledge. Since the tasks are subjective, each new task requires data annotation from human subjects belonging to the different de-

[1] In the humor generation work, the transforming involved is filling-in/replacing the blanks in sentences with appropriate words so that the resulting sentence is a funny transformation.

mographic groups of interest. This is challenging as it is expensive and time-consuming. I also aim to study the effect of demographic information on tasks not explored in this thesis (e.g., lexical substitution) for larger set of demographics. Further, it is to be noted that while tasks such as word associations and filling-in the blanks are fairly simple, annotating sentences with their syntactic parse trees requires linguistic expertise. Obtaining demographically-labeled datasets for such complex tasks (e.g., part-of-speech tagging, dependency parsing) requires going beyond AMT; I aim to explore in the future ideas for semi-automatic demographic data annotation for the part-of-speech tagging and dependency parsing works in social media settings.

The goal of this thesis was to explore whether various natural language processing (NLP) tasks are impacted by the demographic characteristics of the people behind the text, and investigate the word usage differences that emerge across these tasks. My observation is that demographic-based differences do surface in word associations, syntax and humor generation, motivating and paving the way for further studies and the migration of the current agnostic language representations to a demographic-aware/demographic-specific state. In general, this thesis supports the idea that language is not just the interaction between words, but rather, it is the interaction *between people*. For a better modeling of language, the backgrounds of the people behind it must be accounted for.

BIBLIOGRAPHY

[1] Shweder, R. A., Goodnow, J. J., Hatano, G., LeVine, R. A., Markus, H. R., and Miller, P. J., "The cultural psychology of development: One mind, many mentalities," *Handbook of child psychology*, 1998.

[2] De Secondat, C.-L. et al., *The Spirit of Laws*, Hayes Barton Press, 1748.

[3] Doyle, J. A. and Paludi, M. A., *Sex and gender: The human experience*, William C. Brown, 1991.

[4] Lakoff, R., "Language and woman's place," *Language in society*, Vol. 2, No. 1, 1973, pp. 45–79.

[5] Pennebaker, J. W., Rimé, B., and Blankenship, V. E., "Stereotypes of emotional expressiveness of northerners and southerners: a cross-cultural test of Montesquieu's hypotheses." *Journal of personality and social psychology*, Vol. 70, No. 2, 1996, pp. 372.

[6] Shweder, R. A., *Thinking through cultures: Expeditions in cultural psychology*, Harvard University Press, 1991.

[7] Boroditsky, L., "Does language shape thought?: Mandarin and English speakers' conceptions of time," *Cognitive psychology*, Vol. 43, No. 1, 2001, pp. 1–22.

[8] Kern, S., *The culture of time and space, 1880-1918: with a new preface*, Harvard University Press, 2003.

[9] Furnham, A. and Alibhai, N., "Cross-cultural differences in the perception of female body shapes," *Psychological medicine*, Vol. 13, No. 04, 1983, pp. 829–837.

[10] Boroditsky, L., Schmidt, L. A., and Phillips, W., "Sex, syntax, and semantics," *Language in mind: Advances in the study of language and thought*, 2003, pp. 61–79.

[11] Mondorf, B., "Gender differences in English syntax," *Journal of English Linguistics*, Vol. 30, No. 2, 2002, pp. 158–180.

[12] Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W., "Effects of age and gender on blogging." *AAAI spring symposium: Computational approaches to analyzing weblogs*, Vol. 6, 2006, pp. 199–205.

[13] Johannsen, A., Hovy, D., and Søgaard, A., "Cross-lingual syntactic variation over age and gender," *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 2015, pp. 103–112.

[14] Durmus, E. and Cardie, C., "Understanding the Effect of Gender and Stance in Opinion Expression in Debates on" Abortion"," *Proceedings of the Second Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media*, 2018, pp. 69–75.

[15] Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M., "Classifying latent user attributes in twitter," *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ACM, 2010, pp. 37–44.

[16] Burger, J. D., Henderson, J., Kim, G., and Zarrella, G., "Discriminating gender on Twitter," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2011, pp. 1301–1309.

[17] Flekova, L. and Gurevych, I., "Can we hide in the web? Large scale simultaneous age and gender author profiling in social media," *CLEF 2012 Labs and Workshop, Notebook Papers*, 2013.

[18] Koppel, M., Argamon, S., and Shimoni, A. R., "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, Vol. 17, No. 4, 2002, pp. 401–412.

[19] Mukherjee, A. and Liu, B., "Improving gender classification of blog authors," *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, Association for Computational Linguistics, 2010, pp. 207–217.

[20] Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., and Aletras, N., "Studying user income through language, behaviour and affect in social media," *PloS one*, Vol. 10, No. 9, 2015, pp. e0138717.

[21] Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P., "A latent variable model for geographic lexical variation," *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2010, pp. 1277–1287.

[22] Volkova, S., Coppersmith, G., and Van Durme, B., "Inferring User Political Preferences from Streaming Communications." *ACL (1)*, 2014, pp. 186–196.

[23] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H., "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one*, Vol. 8, No. 9, 2013, pp. e73791.

[24] Lampos, V., Aletras, N., Preoţiuc-Pietro, D., and Cohn, T., "Predicting and characterising user impact on Twitter," *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, 2014, pp. 405–413.

[25] Lampos, V., Aletras, N., Geyti, J. K., Zou, B., and Cox, I. J., "Inferring the socioeconomic status of social media users based on behaviour and language," *European Conference on Information Retrieval*, Springer, 2016, pp. 689–695.

[26] Coppersmith, G., Dredze, M., and Harman, C., "Quantifying mental health signals in twitter," *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 51–60.

[27] De Choudhury, M., Counts, S., and Horvitz, E., "Social media as a measurement tool of depression in populations," *Proceedings of the 5th Annual ACM Web Science Conference*, ACM, 2013, pp. 47–56.

[28] Preotiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H. A., and Ungar, L., "The role of personality, age and gender in tweeting about mental illnesses," *NAACL HLT*, Vol. 2015, 2015, p. 21.

[29] Paul, M. and Girju, R., "Cross-cultural analysis of blogs and forums with mixed-collection topic models," *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, Association for Computational Linguistics, 2009, pp. 1408–1417.

[30] Yin, Z., Cao, L., Han, J., Zhai, C., and Huang, T., "Geographical topic discovery and comparison," *Proceedings of the 20th international conference on World wide web*, ACM, 2011, pp. 247–256.

[31] Hovy, D., "Demographic factors improve classification performance," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL 2015)*, Beijing, China, 2015, pp. 752–762.

[32] Volkova, S., Wilson, T., and Yarowsky, D., "Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media," *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, No. October, Seattle, WA, USA, 2013, pp. 1815–1827.

[33] Garimella, A., Mihalcea, R., and Pennebaker, J., "Identifying cross-cultural differences in word usage," *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING 2016)*, Osaka, Japan, 2016, pp. 674–683.

[34] Culotta, A. and Sorensen, J., "Dependency tree kernels for relation extraction," *Proceedings of the 42nd annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2004, p. 423.

[35] Mintz, M., Bills, S., Snow, R., and Jurafsky, D., "Distant supervision for relation extraction without labeled data," *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, Association for Computational Linguistics, 2009, pp. 1003–1011.

[36] Mihalcea, R., Corley, C., Strapparava, C., et al., "Corpus-based and knowledge-based measures of text semantic similarity," *AAAI*, Vol. 6, 2006, pp. 775–780.

[37] Islam, A. and Inkpen, D., "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 2, No. 2, 2008, pp. 10.

[38] Church, K. W. and Hanks, P., "Word association norms, mutual information, and lexicography," *Computational linguistics*, Vol. 16, No. 1, 1990, pp. 22–29.

[39] Nelson, D. L., L., M. C., and A., S. T., "The University of South Florida free association, rhyme, and word fragment norms," *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, Vol. 36, No. 3, 2004, pp. 402–407.

[40] Mollin, S., "Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations," *Corpus Linguistics and Linguistic Theory*, Vol. 5, No. 2, Oct 2009, pp. 175–200.

[41] Tresselt, M. E. and Mayzner, M. S., "The Kent-Rosanoff word association: Word association norms as a function of age," *Psychonomic Science*, Vol. 1, No. 1-12, 1964, pp. 65–66.

[42] Church, K., Gale, W., Hanks, P., and Hindle, D., "Parsing, word associations and typical predicate-argument relations," *Proceedings of the workshop on Speech and Natural Language*, Association for Computational Linguistics, 1989, pp. 75–81.

[43] Wettler, M. and Rapp, R., "Computation of word associations based on the co-occurrences of words in large corpora," *Proceedings of the 1st Workshop on Very Large Corpora*, 1993, pp. 84–93.

[44] Garimella, A., Banea, C., and Mihalcea, R., "Demographic-aware word associations," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2275–2285.

[45] Voutilainen, A., "Part-of-speech tagging," *The Oxford handbook of computational linguistics*, 2003, pp. 219–232.

[46] Kübler, S., McDonald, R., and Nivre, J., "Dependency parsing," *Synthesis Lectures on Human Language Technologies*, Vol. 1, No. 1, 2009, pp. 1–127.

[47] Garimella, A., Banea, C., Hovy, D., and Mihalcea, R., "Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, July 2019, pp. 3493–3498.

[48] Cohen, D., Nisbett, R. E., Bowdle, B. F., and Schwarz, N., "Insult, aggression, and the southern culture of honor: An" experimental ethnography."," *Journal of personality and social psychology*, Vol. 70, No. 5, 1996, pp. 945.

[49] Street, B., "Culture is a verb: Anthropological aspects of language and cultural process," *Language and culture*, 1993, pp. 23–43.

[50] Doyle, J., *Sex and Gender: The Human Experience*, Wm. C. Brown Publishers, 1985.

[51] Chen, C., Lee, S.-y., and Stevenson, H. W., "Response style and cross-cultural comparisons of rating scales among East Asian and North American students," *Psychological Science*, 1995, pp. 170–175.

[52] Cox, T. H., Lobel, S. A., and McLeod, P. L., "Effects of ethnic group cultural differences on cooperative and competitive behavior on a group task," *Academy of Management journal*, Vol. 34, No. 4, 1991, pp. 827–847.

[53] Li, N. and Kirkup, G., "Gender and cultural differences in Internet use: A study of China and the UK," *Computers & Education*, Vol. 48, No. 2, 2007, pp. 301–317.

[54] Mulac, A., Seibold, D. R., and Farris, J. L., "Female and male managers and professionals criticism giving: Differences in language use and effects," *Journal of Language and Social Psychology*, Vol. 19, No. 4, 2000, pp. 389–415.

[55] Tannen, D., *You Just Don't Understand: Women and Men in Conversation*, London, Virago, 1991.

[56] Eckert, P. and McConnell-Ginet, S., *Language and gender*, Cambridge University Press, 2003.

[57] Kramsch, C., *Language and culture*, Oxford University Press, 1998.

[58] Ramírez-Esparza, N. and Pennebaker, J. W., "Do good stories produce good health?: Exploring words, language, and culture," *Narrative Inquiry*, Vol. 16, No. 1, 2006, pp. 211–219.

[59] Pennebaker, J. W. and Francis, M. E., "Cognitive, emotional, and language processes in disclosure," *Cognition & Emotion*, Vol. 10, No. 6, 1996, pp. 601–626.

[60] Pennebaker, J. W., Mayne, T. J., and Francis, M. E., "Linguistic predictors of adaptive bereavement." *Journal of personality and social psychology*, Vol. 72, No. 4, 1997, pp. 863.

[61] Rude, S., Gortner, E.-M., and Pennebaker, J., "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, Vol. 18, No. 8, 2004, pp. 1121–1133.

[62] Pennebaker, J. W., "The secret life of pronouns," *New Scientist*, Vol. 211, No. 2828, 2011, pp. 42–45.

[63] Pennebaker, J. W., "Using computer analyses to identify language style and aggressive intent: The secret life of function words," *Dynamics of Asymmetric Conflict*, Vol. 4, No. 2, 2011, pp. 92–102.

[64] Johnstone, B., "Language and place," 2010.

[65] Kurath, H., "A word geography of the Eastern United States." 1967.

[66] Kent, G. H. and Rosanoff, A. J., "A study of association in insanity," *American Journal of Psychiatry*, Vol. 67, No. 1, 1910, pp. 37–96.

[67] Jenkins, J. J. and Palermo, D. S., "Further data on changes in word-association norms," *Journal of Personality and Social Psychology*, Vol. 1, No. 4, 1965, pp. 303–309.

[68] Rosenzweig, M. R., "Comparisons among Word-Association Responses in English, French, German, and Italian," *The American Journal of Psychology*, Vol. 74, No. 3, Sep. 1961, pp. 347–360.

[69] Wettler, M. and Rapp, R., "A connectionist system to simulate lexical decisions in information retrieval," *Connectionism in Perspective. Amsterdam: Elsevier*, 1989, pp. 463–469.

[70] Dice, L. R., "Measures of the Amount of Ecologic Association Between Species," *Ecology*, Vol. 26, No. 3, 1945, pp. 297–302.

[71] Landauer, T. K. and Dumais, S. T., "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological Review*, Vol. 104, No. 2, 1997, pp. 211–240.

[72] Gabrilovich, E. and Markovitch, S., "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," *Proceedings of the 20th AAAI International Conference on Artificial Intelligence (AAAI 2007)*, Hyderabad, India, 2007, pp. 1606–1611.

[73] Hassan, S. and Mihalcea, R., "Measuring semantic relatedness using salient encyclopedic concepts," *Artificial Intelligence, Special Issue*, Vol. xx, No. xx, 2011.

[74] Leacock, C. and Chodorow, M., "Combining local context and WordNet similarity for word sense identification," *WordNet: An Electronic Lexical Database*, 1998, pp. 305–332.

[75] Lesk, M., "Automatic sense disambiguation using machine readable dictionaries," *Proceedings of the 5th annual international conference on Systems documentation (SIGDOC 1986)*, Toronto, Ontario, 1986, pp. 24–26.

[76] Jarmasz, M. and Szpakowics, S., "Rogets thesaurus and semantic similarity," *Proceedings of Recent Advances in Natural Language Processing (RANLP 2003)*, Borovetz, Bulgaria, 2003, pp. 111–120.

[77] Hughes, T. and Ramag, D., “Lexical semantic relatedness with random graph walks,” *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 2007)*, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 581–589.

[78] De Deyne, S., Navarro, D. J., and Storms, G., “Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations,” *Behavior Research Methods*, Vol. 45, No. 2, 2013, pp. 480–498.

[79] Bamman, D., Dyer, C., and Smith, N. A., “Distributed representations of geographically situated language,” *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL 2014)*, 2014, pp. 828–834.

[80] Boulis, C. and Ostendorf, M., “A quantitative analysis of lexical differences between genders in telephone conversations,” *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2005, pp. 435–442.

[81] Garimella, A. and Mihalcea, R., “Zooming in on gender differences in social media,” *PEOPLES 2016*, 2016, pp. 1.

[82] Vigliocco, G. and Franck, J., “When sex and syntax go hand in hand: Gender agreement in language production,” *Journal of Memory and Language*, Vol. 40, No. 4, 1999, pp. 455–478.

[83] Eckert, P. and McConnell-Ginet, S., *Language and gender*, Cambridge University Press, 2013.

[84] Gadde, P., Jindal, K., Husain, S., Sharma, D. M., and Sangal, R., “Improving data driven dependency parsing using clausal information,” *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 657–660.

[85] Moilanen, K. and Pulman, S., “Sentiment composition,” *Proceedings of the Conference on Recent Advances in Natural Language Processing*, Vol. 7, 2007, pp. 378–382.

[86] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C., “Recursive deep models for semantic compositionality over a sentiment treebank,” *Proceedings of the conference on empirical methods in natural language processing*, 2013.

[87] Lynn, V., Son, Y., Kulkarni, V., Balasubramanian, N., and Schwartz, H. A., “Human Centered NLP with User-Factor Adaptation,” *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1157–1166.

[88] Attardo, S. and Raskin, V., "Script theory revis (it) ed: Joke similarity and joke representation model," *Humor-International Journal of Humor Research*, Vol. 4, No. 3-4, 1991, pp. 293–348.

[89] Wilkins, J. and Eisenbraun, A. J., "Humor theories and the physiological benefits of laughter," *Holistic nursing practice*, Vol. 23, No. 6, 2009, pp. 349–354.

[90] Attardo, S., *Linguistic theories of humor*, Vol. 1, Walter de Gruyter, 2010.

[91] Morreall, J., "Philosophy of humor," 2012.

[92] O'Shannon, D., *What are You Laughing At?: A Comprehensive Guide to the Comedic Event*, A&C Black, 2012.

[93] Weems, S., *Ha!: The science of when we laugh and why*, Basic Books (AZ), 2014.

[94] Labutov, I. and Lipson, H., "Humor as circuits in semantic networks," *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, Association for Computational Linguistics, 2012, pp. 150–155.

[95] Raskin, V., *The primer of humor research*, Vol. 8, Walter de Gruyter, 2008.

[96] Binsted, K., Pain, H., and Ritchie, G. D., "Children's evaluation of computer-generated punning riddles," *Pragmatics & Cognition*, Vol. 5, No. 2, 1997, pp. 305–354.

[97] Stock, O. and Strapparava, C., "HAHAcronym: Humorous agents for humorous acronyms," *Stock, Oliviero, Carlo Strapparava, and Anton Nijholt. Eds*, 2002, pp. 125–135.

[98] Petrović, S. and Matthews, D., "Unsupervised joke generation from big data," *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2, 2013, pp. 228–232.

[99] Valitutti, A., Doucet, A., Toivanen, J. M., and Toivonen, H., "Computational generation and dissection of lexical replacement humor," *Natural Language Engineering*, Vol. 22, No. 5, 2016, pp. 727–749.

[100] Hossain, N., Krumm, J., Vanderwende, L., Horvitz, E., and Kautz, H., "Filling the blanks (hint: plural noun) for mad libs humor," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 638–647.

[101] Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., and Allen, J., "Lsdsem 2017 shared task: The story cloze test," *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, 2017, pp. 46–51.

[102] Yu, L., Park, E., Berg, A. C., and Berg, T. L., "Visual madlibs: Fill in the blank description generation and question answering," *Proceedings of the ieee international conference on computer vision*, 2015, pp. 2461–2469.

[103] Mundorf, N., Bhatia, A., Zillmann, D., Lester, P., and Robertson, S., "Gender differences in humor appreciation," *Humor-International Journal of Humor Research*, Vol. 1, No. 3, 1988, pp. 231–244.

[104] Hay, J., "Gender and humour: Beyond a joke," *Unpublished Masters thesis, Victoria University of Wellington, Wellington, New Zealand*, 1995.

[105] Hay, J., "Functions of humor in the conversations of men and women," *Journal of pragmatics*, Vol. 32, No. 6, 2000, pp. 709–742.

[106] Freud, S., *Der witz und seine beziehung zum unbewussten*, F. Deuticke, 1905.

[107] Alden, D. L., Hoyer, W. D., and Lee, C., "Identifying global and culture-specific dimensions of humor in advertising: A multinational analysis," *The Journal of Marketing*, 1993, pp. 64–75.

[108] Maples, M. F., Dupey, P., Torres-Rivera, E., Phan, L. T., Vereen, L., and Garrett, M. T., "Ethnic diversity and the use of humor in counseling: Appropriate or inappropriate?" *Journal of Counseling & Development*, Vol. 79, No. 1, 2001, pp. 53–60.

[109] Toutanova, K., Klein, D., Manning, C. D., and Singer, Y., "Feature-rich part-of-speech tagging with a cyclic dependency network," *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics, 2003, pp. 173–180.

[110] Klein, D. and Manning, C. D., "Accurate unlexicalized parsing," *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, Association for Computational Linguistics, 2003, pp. 423–430.

[111] Pennebaker, J. W., Francis, M. E., and Booth, R. J., "Linguistic inquiry and word count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, Vol. 71, 2001, pp. 2001.

[112] Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S., "OpinionFinder: A system for subjectivity analysis," *Proceedings of hlt/emnlp on interactive demonstrations*, Association for Computational Linguistics, 2005, pp. 34–35.

[113] Ignatow, G. and Mihalcea, R., "Injustice Frames in Social Media," Denver, CO, 2012.

[114] Strapparava, C., Valitutti, A., et al., "WordNet Affect: an Affective Extension of WordNet." *LREC*, Vol. 4, 2004, pp. 1083–1086.

[115] De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al., "Generating typed dependency parses from phrase structure parses," *Proceedings of LREC*, Vol. 6, 2006, pp. 449–454.

[116] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, Vol. 11, No. 1, 2009, pp. 10–18.

[117] Wei, X. and Croft, W. B., "LDA-based document models for ad-hoc retrieval," *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2006, pp. 178–185.

[118] Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent dirichlet allocation," *the Journal of machine Learning research*, Vol. 3, 2003, pp. 993–1022.

[119] Miller, G. A., "WordNet: a lexical database for English," *Communications of the ACM*, Vol. 38, No. 11, 1995, pp. 39–41.

[120] Heinrich, G., "Parameter estimation for text analysis," Tech. rep., Technical report, 2005.

[121] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D., "The Stanford CoreNLP Natural Language Processing Toolkit," *Association for Computational Linguistics System Demonstrations (ACL 2014)*, 2014, pp. 55–60.

[122] Řehůřek, R. and Sojka, P., "Software Framework for Topic Modelling with Large Corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[123] James, W., "The Principles of Psychology," *Psychology*, Vol. 2, 1890.

[124] Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K., *Rethinking innateness: a connectionist perspective on development*, The MIT Press, 1997.

[125] Rogers, T. R. and McClelland, J. L., *Semantic cognition: a parallel distributed processing approach*, The MIT Press, 2004.

[126] Kajić, I., Gosmann, J., Stewart, T. C., Wennekers, T., and Eliasmith, C., "A spiking neuron model of word associations for the remote associates test," *Frontiers in Psychology*, Vol. 8, No. February, 2017.

[127] Jenkins, J. J., "The 1952 Minnesota word association norms," *Norms of Word Association*, edited by L. Postman and G. Keppel, chap. Chapter 1, Academic Press, 1970, pp. 1–38.

[128] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., "Distributed representations of words and phrases and their compositionality," *Proceedings of the Neural Information Processing Systems Conference (NIPS 2013)*, 2013, pp. 3111–3119.

[129] Salton, G. and McGill, M. J., *Introduction to modern information retrieval*, McGraw-Hill, Inc., New York, NY, USA, 1986.

[130] McCarthy, D. and Navigli, R., “The English lexical substitution task,” *Language Resources and Evaluation*, Vol. 43, No. 2, 6 2009, pp. 139–159.

[131] Jabbari, S., Hepple, M., and Guthrie, L., “Evaluation metrics for the lexical substitution task,” *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*, 2010, pp. 289–292.

[132] Chaudhari, D. L., Damani, O. P., and Laxman, S., “Lexical co-occurrence, statistical significance, and word association,” *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, Edinburgh, Scotland, UK, 2011, pp. 1058–1068.

[133] Deyne, S. D., Perfors, A., and Navarro., D. J., “Predicting human similarity judgments with distributional models: The value of word associations,” *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING 2016)*, Osaka, Japan, 2016, pp. 1861–1870.

[134] Bel-Enguix, G., “Retrieving word associations with a simple neighborhood algorithm in a graph-based resource,” *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon*, Dublin, Ireland, 2014, pp. 60–63.

[135] Mohammad, S., “Colourful language: Measuring word-colour associations,” *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 2011, pp. 97–106.

[136] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W., “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints,” *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[137] Jørgensen, A., Hovy, D., and Søgaard, A., “Challenges of Studying and Processing Dialects in Social Media,” *Workshop on Noisy User-generated Text (W-NUT)*, 2015.

[138] Hovy, D. and Søgaard, A., “Tagging Performance Correlates with Author Age,” *ACL*, 2015.

[139] Hovy, D. and Spruit, S. L., “The Social Impact of Natural Language Processing,” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, Association for Computational Linguistics, 2016.

[140] Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., and Kalai, A. T., “Quantifying and Reducing Stereotypes in Word Embeddings,” *CoRR*, Vol. abs/1606.06121, 2016.

[141] Benton, A., Mitchell, M., and Hovy, D., “Multi-Task Learning for Mental Health using Social Media Text,” *Proceedings of EACL*, 2017.

[142] Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B., "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics*, Vol. 19, No. 2, 1993, pp. 313–330.

[143] Brill, E., "Some Advances in Transformation-based Part of Speech Tagging," *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, AAAI 1994, Menlo Park, CA, USA, 1994, pp. 722–727.

[144] Ratnaparkhi, A., "A maximum entropy model for part-of-speech tagging," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Vol. 1 of *EMNLP 1996*, Philadelphia, PA, 1996, pp. 133–142.

[145] Toutanova, K. and Manning, C. D., "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," *Proceedings of the 2000 Joint SIGDAT conference on Empirical Methods in Natural Language Processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, SIGDAT-EMNLP 2000, Hong Kong, China, 2000, pp. 63–70.

[146] Nivre, J. and Scholz, M., "Deterministic Dependency Parsing of English Text," *Proceedings of the 20th International Conference on Computational Linguistics*, COLING 2004, Geneva, Switzerland, August 2004.

[147] Chen, D. and Manning, C., "A Fast and Accurate Dependency Parser using Neural Networks," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, EMNLP 2014, Doha, Qatar, 2014, pp. 740–750.

[148] Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Pao, C., Rudnicky, A., and Shriberg, E., "Expanding the scope of the ATIS task: The ATIS-3 corpus," *Proceedings of the workshop on Human Language Technology*, Association for Computational Linguistics, 1994, pp. 43–48.

[149] Prabhakaran, V. and Rambow, O., "Dialog Structure Through the Lens of Gender, Gender Environment, and Power," *arXiv preprint arXiv:1706.03441*, 2017.

[150] Tang, C., Ross, K., Saxena, N., and Chen, R., "Whats in a name: a study of names, gender inference, and gender behavior in facebook," *16th International Conference on Database Systems for Advanced Applications*, 2011, pp. 344–356.

[151] Cornett, H. E., "Gender Differences in Syntactic Development Among English Speaking Adolescents," *Inquiries Journal/Student Pulse*, Vol. 6, No. 3, 2014, pp. 1–6.

[152] Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al., "Universal dependencies v1: A multilingual treebank collection," *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 1659–1666.

[153] Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M., "Globally Normalized Transition-Based Neural Networks," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2016, Berlin, Germany, 2016, pp. 2442–2452.

[154] Nivre, J., Hall, J., and Nilsson, J., "Memory-based dependency parsing," 2008.

[155] Apte, M. L., *Humor and laughter: An anthropological approach*, Cornell Univ Pr, 1985.

[156] Kramarae, C., "Women and men speaking: Frameworks for analysis." 1981.

[157] Duncan, W. J., Smeltzer, L. R., and Leap, T. L., "Humor and work: Applications of joking behavior to management," *Journal of Management*, Vol. 16, No. 2, 1990, pp. 255–278.

[158] Goodman, L., *Gender and humour*, na, 1992.

[159] Robinson, D. T. and Smith-Lovin, L., "Getting a laugh: Gender, status, and humor in task discussions," *Social Forces*, Vol. 80, No. 1, 2001, pp. 123–158.

[160] Mihalcea, R. and Strapparava, C., "Learning to laugh (automatically): Computational models for humor recognition," *Computational Intelligence*, Vol. 22, No. 2, 2006, pp. 126–142.

[161] Kiddon, C. and Brun, Y., "That's what she said: double entendre identification," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, 2011, pp. 89–94.

[162] Bertero, D. and Fung, P., "A long short-term memory framework for predicting humor in dialogues," *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 130–135.

[163] Raz, Y., "Automatic humor classification on Twitter," *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, Association for Computational Linguistics, 2012, pp. 66–70.

[164] Zhang, R. and Liu, N., "Recognizing humor on twitter," *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, ACM, 2014, pp. 889–898.

[165] Price, R. and Stern, L., *Mad Libs: World's Greatest Party Game: a Do-it-yourself Laugh Kit*, No. 1, Mad Libs, 1974.

[166] Price, R. and Stern, L., *Best of Mad Libs*, Penguin Group, 2008.

[167] Krippendorff, K., “Estimating the reliability, systematic error and random error of interval data,” *Educational and Psychological Measurement*, Vol. 30, No. 1, 1970, pp. 61–70.

[168] Hossain, N., Krumm, J., and Gamon, M., ““President Vows to Cut <Taxes> Hair”: Dataset and Analysis of Creative Text Editing for Humorous Headlines,” *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019, pp. 133–142.

[169] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019, pp. 4171–4186.

[170] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., “Attention is all you need,” *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[171] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S., “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[172] Fellbaum, C., “WordNet: An electronic lexical database and some of its applications,” 1998.

[173] Hendrycks, D. and Gimpel, K., “Bridging nonlinearities and stochastic regularizers with gaussian error linear units,” 2016.

[174] Rabinovich, E., Patel, R. N., Mirkin, S., Specia, L., and Wintner, S., “Personalized Machine Translation: Preserving Original Author Traits,” *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, Valencia, Spain, April 2017, pp. 1074–1084.

[175] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J., “Personalizing Dialogue Agents: I have a dog, do you have pets too?” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, July 2018, pp. 2204–2213.

[176] Jaech, A. and Ostendorf, M., “Personalized language model for query auto-completion,” *arXiv preprint arXiv:1804.09661*, 2018.

[177] Rohini, U., Ambati, V., and Varma, V., “Statistical machine translation models for personalized search,” *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008.

[178] Lee, J. and Yeung, C. Y., “Personalizing lexical simplification,” *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 224–232.

[179] Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W., “Gender differences in language use: An analysis of 14,000 text samples,” *Discourse Processes*, Vol. 45, No. 3, 2008, pp. 211–236.

[180] Koolen, C. and van Cranenburgh, A., “These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution,” *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Association for Computational Linguistics, Valencia, Spain, April 2017, pp. 12–22.

[181] Rudman, L. A. and Glick, P., *The social psychology of gender: How power and intimacy shape gender relations*, Guilford Press, 2012.