Statistical Models for Linguistic Variation in Social Media

Vivek Kulkarni

May 2016

Department of Computer Science
Stony Brook University
Stony Brook, NY 11790

Thesis Committee:
Steven Skiena
Andrew H Schwartz
Niranjan Balasubramanian
David Bamman (U C Berkeley)

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

Abstract

Language on the Internet and social media varies due to time, geography and social factors. For example, consider an online chat forum where people from different regions across the world interact. In such scenarios, it is important to track and detect regional variation in language. A person from the UK, who is in conversation with someone from the USA could say *"he is stuck in the lift"* to mean *"he is stuck in an elevator"*, since the word `lift` means an `elevator` in the UK. Note that in the US, `lift` does not refer to an `elevator`. Modeling such variation can allow for applications to prompt or suggest the intended meaning to the other participants of the conversation.

In this thesis I conduct two related lines of inquiry focusing on (a) language itself and the variation it manifests and (b) the user and what we can infer about them based on their language use on social media.

First I develop computational methods to track and detect changes in word usage, including semantic and syntactic variation. I examine two modalities: time and geography. Specifically I outline methods to use distributional word representations (word embeddings) to detect semantic variation in word usage. Our methods are scalable to large datasets, making them particularly suited for social media. Second, I turn my attention towards users. In particular, I seek to model latent personality traits of users based on their language use on social media. I propose to develop generative latent factor models, that explicitly seek to build representations of each user based on their inferred latent personality traits. These models seek to capture latent personality traits that serve as useful co-variates for a wide variety of tasks like predicting what topics users like on social media and the number of friends in their social circle.

This work has broad applications in several fields like information retrieval, semantic web applications, socio-variational linguistics, computational social science including digital health care, psycho-linguistics and ad-targeting.

Contents

Publications

- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee, 2015
- Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. Freshman or fresher? quantifying the geographic variation of internet language. *arXiv preprint arXiv:1510.06786*, 2015
- Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. Polyglot-ner: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining*, 2015
- Eran Shor, Arnout van de Rijt, Alex Miltsov, Vivek Kulkarni, and Steven Skiena. A paper ceiling explaining the persistent underrepresentation of women in printed news. *American Sociological Review*, 80(5):960–984, 2015
- Bryan Perozzi, Rami Al-Rfou, Vivek Kulkarni, and Steven Skiena. Inducing language networks from continuous space word representations. In *Complex Networks V*. 2014
- Vivek Kulkarni, Jagat Sastry Pudipeddi, Leman Akoglu, Joshua T Vogelstein, R Jacob Vogelstein, Sephira Ryman, and Rex E Jung. Sex differences in the human connectome. In *Brain and Health Informatics*, pages 82–91. Springer, 2013
- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Exploring the power of gpu's for training polyglot language models. *CoRR*, abs/1404.1521, 2014
- Bryan Perozzi, Vivek Kulkarni, and Steven Skiena. Walklets: Multiscale graph embeddings for interpretable network classification. *Submitted*, 2016

Chapter 1

Introduction

Language exhibits a wide variety of variation. This variation is influenced by time, geography and other social variables like income, gender, and education to name a few [20, 33]. The evolution of online social media and the Internet at a global scale presents a potential to detect and analyze linguistic variation at a scale that is not possible using traditional methods like surveys and questionnaires. Furthermore with the evolution of the semantic web, tracking and detecting semantic variation in language is important for an enriched user experience. *How can we detect and track such linguistic variation on the Internet and online media? Furthermore what does language reveal about the users themselves?* In this thesis, I examine these questions and advocate for computational approaches and statistical methods to detect such variation.

First language itself varies across time and geography. To that end, I propose and investigate computational methods to detect changes in word usage (syntactic and semantic) across both these modalities. I examine methods that track variation in word usage by modeling word usage patterns, syntax and semantics (using distributional methods). I then investigate a second dimension: What does language inform us about the users or other social variables? In particular, I investigate what latent factors of people's personalities can be inferred from social media text. While addressing each of these questions, I advocate to capture significance of the observed variation by introducing null models to account for random variation due to chance, thereby reducing false positives.

Finally I evaluate my models and methods on several data-sets like Google Book Ngrams, Amazon Movie Reviews, Twitter and demonstrate that our proposed models can identify rich variation in word usage.

1.1 Completed Work

My research falls broadly in two areas (a) Statistical Models for Linguistic Variation and (b) Natural Language Processing.

Statistical Models for Linguistic Variation My primary work is on developing computational methods to detect linguistic variation in online media. I have developed computational methods to detect when words acquire new senses or change their usage over time [18]. I have also proposed methods to use word embeddings to detect regional variation of word usage [19].

Natural Language Processing I have contributed to several research projects on Natural Language Processing. We induce a network over distributed word representations using word embeddings and analyze several properties of the induced network structure[29]. I have also worked on developing a system of Named Entity Recognition in 40 languages without using explicit feature engineering or language specific features[1]. Our Named Entity Recognition system effectively uses word embeddings and linkages between entities in Wikipedia and Freebase to learn models for Named Entity Recognition in 40 languages.

I have also collaborated with sociologists to analyze the persistent under-representation of women in social media [32] by using LYDIA, a News and Blog Analysis system developed at our lab.

Miscellaneous Projects Some of my other projects include identifying sex differences in human connectomes modeling the connectome as a complex network [16] and understanding the stability of dropout in neural networks [13]. More recently, I have also worked on representation learning for online social networks[30].

1.2 Proposed Work

I intend to continue to work on projects around my primary research area of detecting and analyzing language variation, which I briefly summarize below:

Lingustic Variation I outline several extensions to my work on analyzing linguistic variation exhibited by words along both time and geography. I propose to work on extensions that explicitly model word senses, capture richer variation and improved change detection.

Inferring Latent Socio-linguistic Variables Inspired by my work on linguistic variation in time and geography, I seek to focus my attention to yet another modality:users and model linguistic variation that users exhibit and infer latent socio-linguistic variables. Specifically I propose to build Bayesian models to infer latent personality traits of users that are reflected in language on social media.

1.3 Impact

1.3.1 Academic Impact

My work on detecting semantic variation of word usage across time LANGCHANGETRACK [18] has already been cited 17 times since its publication in May 2015. I have accumulated 27 citations overall as of April 9,2016 with several forks of my code on `github`. Furthermore my work

[18] has been cited by leaders like Christopher Manning [21], in the field of Natural Language Processing. We have been invited to present our work at the following venues:

- Yahoo Labs, Sunnyvale, California
- Google, Mountain View, California
- NetSci 2015, Zaragoza, Spain
- Mid-Atlantic Student Colloquium on Speech, Language and Learning, University of Pennsylvania
- Institute of Advanced Computational Science, Stony Brook University

1.3.2 Press and Media Attention

My work has received considerable press and media coverage. In particular my work on detecting linguistic variation through time [18] has featured in The MIT Technology Review , ACM Tech News and even by the popular media VICE. My work with sociologists on analyzing the under-representation of women in news articles has been featured by The Guardian and in a press release by Stony Brook University.

1.4 Overview

The proposal is structured as follows: In Chapter 2 I briefly describe our work on detecting linguistic variation and propose several extensions. I follow up in Chapter 3 by describing ongoing work and proposed work and modeling latent personality traits from social media. I finally conclude in Chapter 4 by outlining a time-line of how I intend to complete my proposed work.

Chapter 2

Linguistic Variation in Time and Geography

Language evolves over space and time. Studies on linguistic variation have been traditionally conducted via questionnaires and surveys [20, 33]. However the evolution of online social media and the Internet enables one to study language at a scale that is not achievable using classical techniques. Furthermore, with the rise of the semantic web, there has been a shift in focus from keywords to semantics of user queries and an increasing focus on personalization [35]. Detecting and characterizing linguistic change can enable semantic web applications to return more relevant results [14].

In this chapter, we introduce computational methods to detect semantic variation in word usage in time and geography. We investigate techniques to analyze linguistic change ranging from elementary frequency analysis to effectively using distributional methods to capture word semantics. Furthermore while word embeddings have shown to be very useful as co-variates for several NLP tasks like Part of Speech (POS) Tagging, Named Entity Recognition (NER) and Semantic Role Labeling [9, 10, 29] we show how to use word embeddings to detect variation in word semantics.

2.1 Completed Work

2.1.1 Linguistic Variation in Time - WWW 2015 (LANGCHANGETRACK)

In this work, we introduce a technique to detect semantic change of words over time. At its heart, our technique has two key steps: (a) Model word usage over time by tracking some statistical measure to construct a time series and (b) Detect change points in the time series to estimate the time, the change became dominant.

Figure 2.1 demonstrates the variation captured by our method for the word `gay`. Note that in early 1900's `gay` is close to words like `cheerful, dapper, sublimely` indicating it was used to indicate a cheerful nature. However with time, we observe that `gay` starts to transition in this semantic space until the 1980's where the dominant sense is that of being a homosexual.

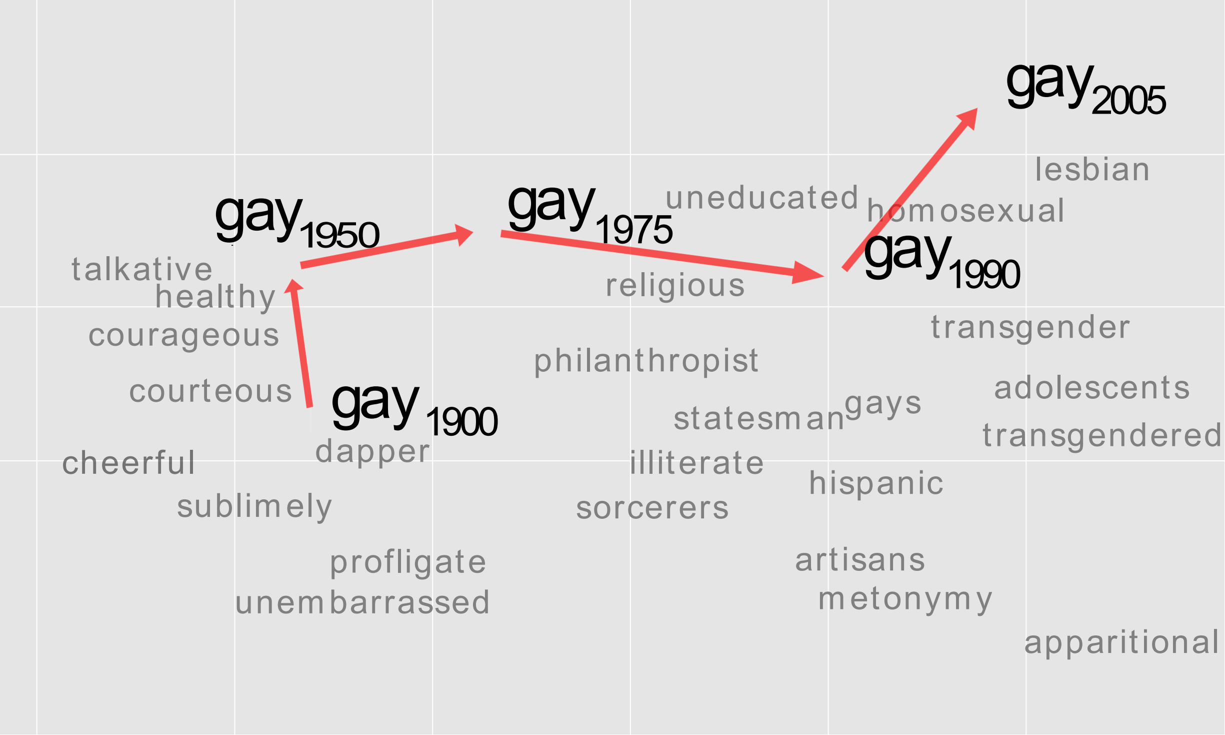


Figure 2.1: A 2-dimensional projection of the latent semantic space captured by our algorithm. Notice the semantic trajectory of the word gay transitioning meaning in the space.

Problem Definition We seek to quantify the linguistic shift in word meaning and usage across time. Given a temporal corpora $\mathcal{C}$ created over a time span $\mathcal{S}$, we divide the corpora into n snapshots $\mathcal{C}_t$ each of period length P. We construct a common vocabulary $\mathcal{V}$ by intersecting the word dictionaries that appear in all the snapshots (i.e, we track the same word set across time). This eliminates trivial examples of word usage shift from words which appear or vanish throughout the corpus. Specifically we pose the following questions:

1. How statistically significant is the shift in usage of a word w across time?
2. Given that a word has shifted, at what point in time did the change occur?

Key Idea First we model word evolution by constructing a time series. We construct a time series $\mathcal{T}(w)$ for each word $w \in \mathcal{V}$. Each point $\mathcal{T}_t(w)$ corresponds to statistical information extracted from corpus snapshot $\mathcal{C}_t$ that reflects the usage of w. Our method is generic enough to accommodate different statistical properties. For example, one might use the normalized frequency of word usage in the corpus at time t as a measure or a measure over the Part of Speech (POS) distribution of the word. Figure 2.2 shows an example time-series constructed by tracking the POS distribution of apple over time. While apple has primarily always been used as a common noun, observe the marked change in the distance measure for apple around the 1980's with an increased usage as a proper noun. This coincides with the rise of apple as a software company.

Finally, given such a time series we quantify the significance of the shift that occurred to the

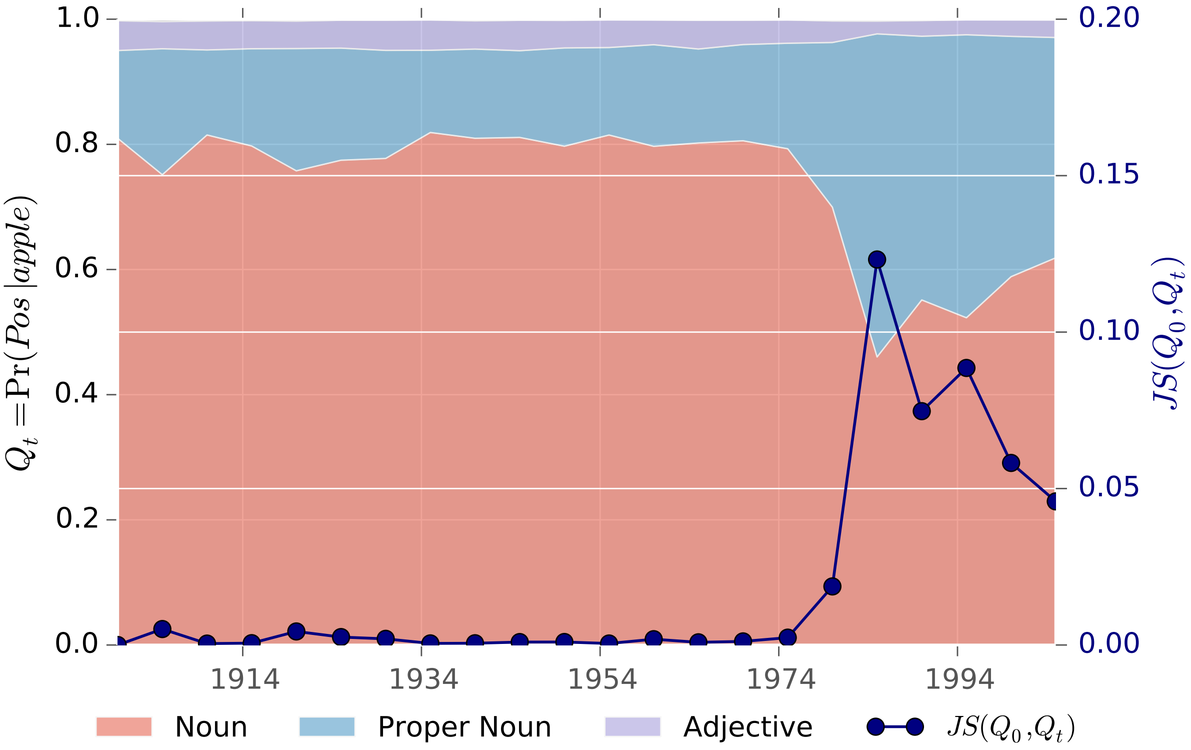


Figure 2.2: Part of speech tag probability distribution of the word apple (stacked area chart). Observe that the Proper Noun tag has dramatically increased in 1980s. The same trend is clear from the time series constructed using Jenssen-Shannon Divergence (dark blue line).

word in its meaning and usage. We draw on techniques from time series analysis, specifically change point detection to achieve this. Our method to detect changes in time series relies on a *mean shift model* which intuitively tracks changes in the mean of the time-series to detect changes. We provide a detailed description in our paper [18].

Distributional Method We briefly describe how to use word embeddings to construct a time-series to capture word semantics through time. Here is how it works:

1. **Word Embeddings Construction** We first train word embeddings for each time point independently using the Skipgram model outlined by [27].
2. **Constructing a Unified Vector Space** Since the word embeddings trained above for each time point are in different vector spaces, we cannot yet induce a distance measure over them. Therefore, we propose a method to unify word embeddings from different time points to a joint space primarily using a piece-wise linear model that seeks to preserve local neighborhood structure.
3. **Inducing a distance measure** In the joint space, we now can induce a distance measure between word embeddings for different time points. We use the COSINEDISTANCE to track the displacement of a word over time. In particular at each time point t we compute

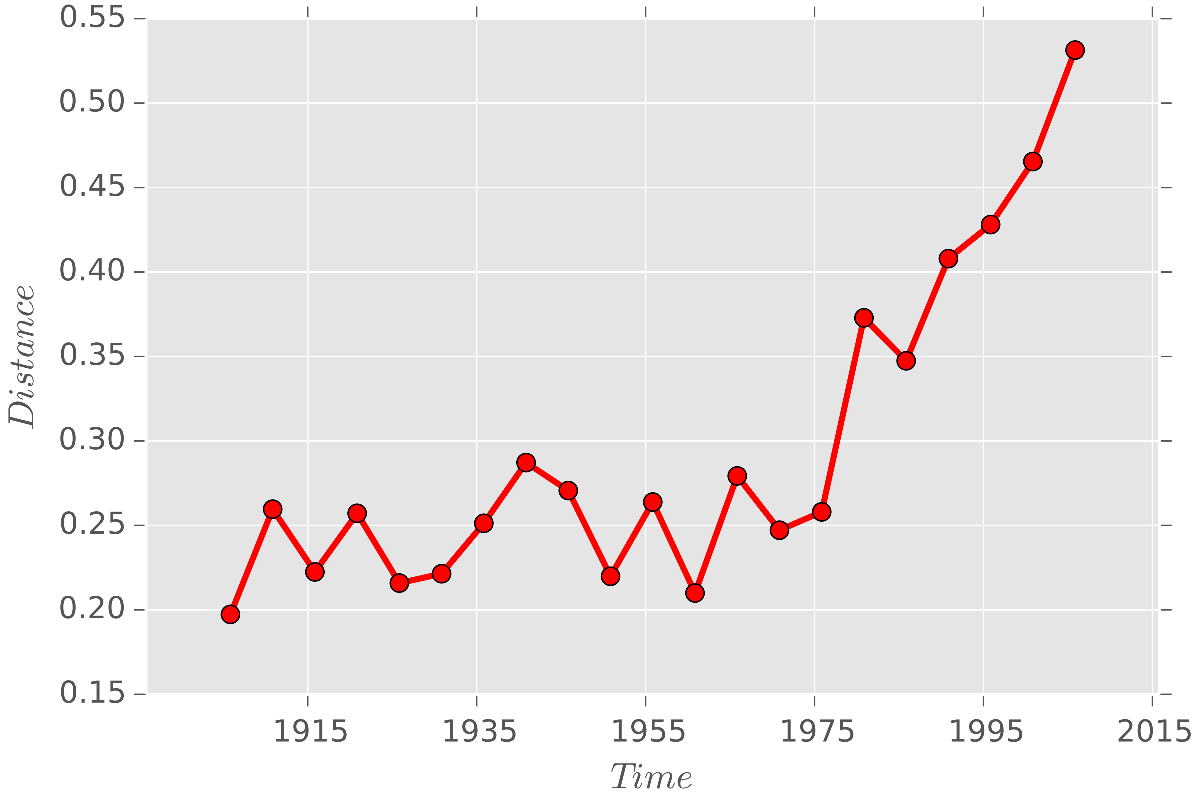


Figure 2.3: Tracking the displacement of `gay` over time using word embeddings with CosineDistance as the metric.

the CosineDistance between the word embedding at time t and the word embedding at the source time point. Figure 2.3 shows such a time series constructed for `gay` using our method. Observe the noticeable increase in distance around 1980 corresponding to the `gay` transitioning dominantly to its current meaning.

Results We apply our method on three datasets - years of micro-blogging from Twitter, a decade of movie reviews from Amazon, and a century of written books using the Google Books Ngram Corpus.

- The Google Books Ngram Corpus project enables the analysis of cultural, social and linguistic trends. It contains the frequency of short phrases of text (*ngrams*) that were extracted from books written in eight languages over five centuries [26]. These ngrams vary in size (1-5) grams. We use the 5-gram phrases which restrict our context window size m to 5. We focus on the time span from $1900 - 2005$, and set the time snapshot period to 5 years (21 points). We obtain the POS Distribution of each word in the above time range by using the Google Syntactic Ngrams dataset [23].
- Amazon Movie Reviews dataset consists of movie reviews from Amazon. This data spans August 1997 to October 2012(13 time points), including all 8 million reviews. However,

	Word	**ECP**	***p*-value**	**Past ngram**	**Present ngram**
Distributional better	recording	1990	0.0263	*to be ashamed of recording that*	*recording, photocopying*
	gay	1985	0.0001	*happy and gay*	*gay and lesbians*
	tape	1970	<0.0001	*red tape, tape from her mouth*	*a copy of the tape*
	checking	1970	0.0002	*then checking himself*	*checking him out*
	diet	1970	0.0104	*diet of bread and butter*	*go on a diet*
	sex	1965	0.0002	*and of the fair sex*	*have sex with*
	bitch	1955	0.0001	*nicest black bitch* (Female dog)	*bitch* (Slang)
	plastic	1950	0.0005	*of plastic possibilities*	*put in a plastic*
	transmitted	1950	0.0002	*had been transmitted to him, transmitted from age to age*	*transmitted in electronic form*
	peck	1935	0.0004	*brewed a peck*	*a peck on the cheek*
	honey	1930	0.01	*land of milk and honey*	*Oh honey!*

				Past POS	**Present POS**
Syntactic better	hug	2002	<0.001	Verb (*hug a child*)	Noun (*a free hug*)
	windows	1992	<0.001	Noun (*doors and windows of a house*)	Proper Noun (*Microsoft Windows*)
	bush	1989	<0.001	Noun (*bush and a shrub*)	Proper Noun (*George Bush*)
	apple	1984	<0.001	Noun (*apple, orange, grapes*)	Proper Noun (*Apple computer*)
	sink	1972	<0.001	Verb (*sink a ship*)	Noun (*a kitchen sink*)
	click	1952	<0.001	Noun (*click of a latch*)	Verb (*click a picture*)
	handle	1951	<0.001	Noun (*handle of a door*)	Verb (*he can handle it*)

Table 2.1: Estimated change point (ECP) as detected by our approach for a sample of words on Google Books Ngram Corpus. *Distributional* method is better on some words (which *Syntactic* did not detect as statistically significant eg. sex, transmitted, bitch, tape, peck) while *Syntactic* method is better on others (which *Distributional* failed to detect as statistically significant eg. apple, windows, bush).

we consider the time period starting from 2000 as the number of reviews from earlier years is considerably small. Each review includes product and user information, ratings, and a plain-text review.

- A Twitter dataset consisting of Tweets of that spans 24 months starting from September 2011 to October 2013. Each Tweet includes the Tweet ID, Tweet and the geo-location if available.

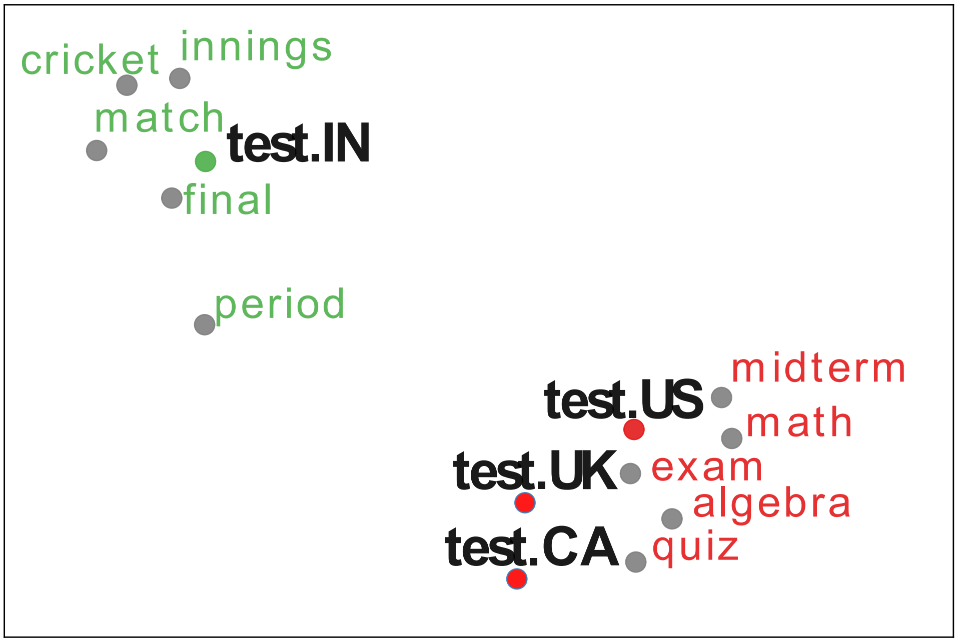


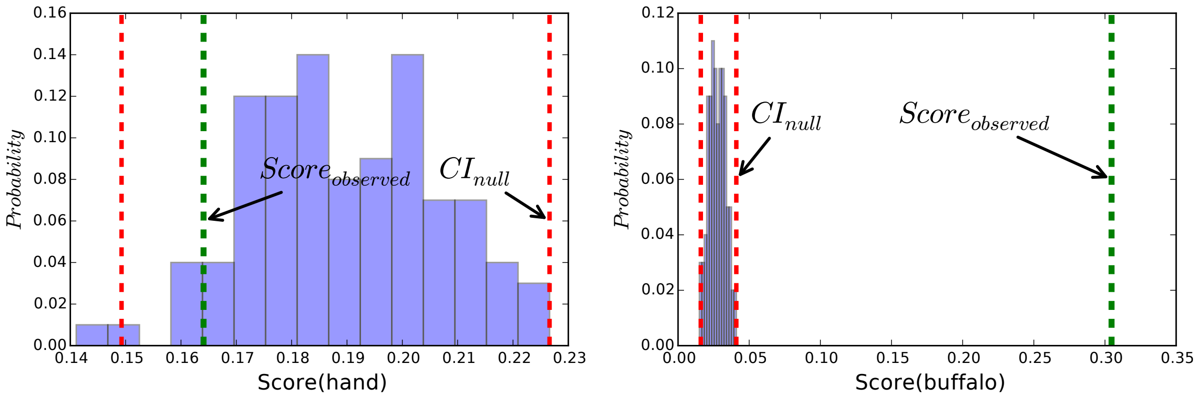
Figure 2.4: The latent semantic space captured by our method (GEODIST) reveals geographic variation between language speakers. In the majority of the English speaking world (e.g. US, UK, and Canada) a `test` is primarily used to refer to an `exam`, while in India a `test` indicates a lengthy cricket match which is played over five consecutive days.

We show a sample of the results in Table 2.1. These results indicate that our method is able to capture a variety of words that acquired new meaning over time. We also conduct a comprehensive quantitative evaluation of our method and the full set of results on the other datasets is available in [18]

2.1.2 Linguistic Variation in Geography - ICWSM 2016 (GEODIST)

We build on our previous work to detect and characterize linguistic variation arising due to geography. We propose a method GEODIST to use word embeddings to detect regional variation in word usage.

Problem Definition We seek to quantify shift in word meaning (usage) across different geographic regions. Specifically, we are given a corpus $\mathcal{C}$ that spans R regions where $\mathcal{C}_r$ corresponds to the corpus specific to region r. We denote the vocabulary of the corpus by $\mathcal{V}$. We want to detect words in $\mathcal{V}$ that have region specific semantics (not including trivial instances of words exclusively used in one region). For each region r, we capture statistical properties of a word w's



(a) Observed score for `hand`

(b) Observed score for `buffalo`

Figure 2.5: The observed scores computed by GeoDist (in 🟩) for `buffalo` and `hand` when analyzing regional differences between New York and USA overall. The histogram shows the distribution of scores under the null model. The 98% confidence intervals of the score under null model are shown in 🟥. The observed score for `hand` lies well within the confidence interval and hence is not a statistically significant change at this level. In contrast, the score for `buffalo` is far outside the confidence interval for the null distribution indicating a statistically significant change at this level.

usage in that region. Given a pair of regions (r_i, r_j), we then reduce the problem of detecting words that are used differently across these regions to an outlier detection problem using the statistical properties captured.

In summary, we answer the following questions:

1. In which regions does the word usage drastically differ from other regions?
2. How statistically significant is the difference observed across regions?

Key Idea We build on work proposed by Bamman et al. [4] to learn geographically situated word embeddings. First GeoDist learns region specific word embeddings for each word $w \in \mathcal{V}$. These region specific word embeddings are already in one unified space since they are learned jointly. GeoDist measures the difference in usage of a word in a given pair of regions (r_i, r_j) using the CosineDistance between the respective regional embeddings. Furthermore, in order to account for random variation due to chance, we propose an appropriate null model to assess the significance of the observed changes and effectively weed out spurious changes due to random chance. We briefly outline this step in Algorithm 1. In Figure 2.5 we demonstrate how the null model can effectively weed out observed variation due to random chance.

Results We apply GeoDist as well as other baseline methods outlined in our previous work [18] on two online data sets

- Google Books Ngram Corpus for English UK and English US
- Twitter: Twitter data partitioned accroding to (a) 50 states in the USA and (b) by countries - USA, UK, India.

Algorithm 1 SCORESIGNIFICANCE (C, B, α)

```
Require: $C$: Corpus of text with $R$ regions, $B$: Number of bootstrap samples, $\alpha$: Confidence Interval threshold
Ensure: $E$: Computed effect sizes for each word $w$, CI: Computed confidence intervals for each word $w$
// Estimate the NULL distribution.
$\text{BS} \leftarrow \emptyset$ {Corpora from the NULL Distribution}. $\text{NULLSCORES}(w)$ {Store the scores for $w$ under null model.}
repeat
    Permute the labels assigned to text of $C$ uniformly at random to obtain corpus $C'$
    $\text{BS} \leftarrow \text{BS} \cup C'$
    Learn a model $N$ using $C'$ as the text.
    for $w \in \mathcal{V}$ do
        Compute $\text{SCORE}(w)$ using $N$.
        Append $\text{SCORE}(w)$ to $\text{NULLSCORES}(w)$
    end for
until $|\text{BS}| = B$
// Estimate the actual observed effect and compute confidence intervals.
Learn a model $M$ using $C$ as the text.
for $w \in \mathcal{V}$ do
    Compute $\text{SCORE}(w)$ using $M$.
    $E(w) \leftarrow \text{SCORE}(w)$
    Sort the scores in $\text{NULLSCORES}(w)$.
    $\text{HCI}(w) \leftarrow 100\alpha$ percentile in $\text{NULLSCORES}(w)$
    $\text{LCI}(w) \leftarrow 100(1-\alpha)$ percentile in $\text{NULLSCORES}(w)$
    $\text{CI}(w) \leftarrow (\text{LCI}(w), \text{HCI}(w))$
end for
return $E, \text{CI}$
```

We show a sample set of results in Table 2.2. Observe that GEODIST detects words like `test` or `stand` which have different senses in India. In the US, `test` refers to an exam while in India `test` is dominantly used to refer to a form of cricket match. Similarly, in India `stand` can be used to mean a station where buses halt (a bus station) where as this sense of `stand` is not prevalent in the US. A full set of results is available in our preprint [19].

2.2 Proposed Work

2.2.1 Detecting Word Sense Changes

While our current work detects changes in word semantics, it does not explicitly model the different senses words possess. This yields a natural extension namely: *How can we not only capture the semantic change in a word but also explicitly capture how the different senses were*

Word	Effect Size	CI(Null)	US Usage	IN Usage
high	0.820	(0.02,0.03)	*I am in high school*	*by pass the high way (as a road)*
hum	0.740	(0.03, 0.04)	*more than hum and talk*	*hum busy hain (Hinglish)*
main	0.691	(0.048, 0.074)	*your main attraction*	*main cool hoon (I am cool)*
ring	0.718	(0.054, 0.093)	*My belly piercing ring*	*on the ring road (a circular road)*
test	0.572	(0.03, 0.061)	*I failed the test*	*We won the test*
stand	0.589	(0.046, 0.07)	*I can't stand stupid people*	*Wait at the bus stand*

Table 2.2: Twitter: Differences between English usage in the United States and India (CI - the 98% Confidence Intervals under the null model)

used? Specifically, one could seek to model a word w as being associated with S senses (a word w can be viewed as a distribution over senses). As an example, we seek to model the fact that apple has two senses: (a) representing a fruit and (b) representing the software company with each of them being represented in the corpora with different proportions.

Recently there has been a surge of work in learning multi-prototype representations [6, 12, 34]. In particular, I propose a a deeper investigation of using multi-prototype word embeddings to explicitly capture multiple senses of words across time and geography. This involves inferring the number of senses a word has (perhaps using a Bayesian model) and then learning representations for each sense and tracking the evolution of senses across time or geography. I envision this work to have rich applications in automatically mapping words to their synsets (generation of dictionaries), word sense and word epoch disambiguation. For example, given a text like *"He was as gay as a lark"*, we should be able infer that the most likely sense of gay in this usage is that of happy, cheerful and the most likely period to place this text would be in the 1900's.

2.2.2 Enriched Linguistic Variation

While most of the methods to learn word embeddings model linear contexts [1, 27, 28] recent work like [22] propose to use arbitrary contexts (for example, contexts extracted by syntactic dependencies). Levy and Goldberg [22] demonstrate that using such targeted contexts produce word embeddings re-markedly different from the classical word embeddings. They observe that turing is associated with non-deterministic, finite-state when using the standard Skipgram CBOW model. However embeddings learned using syntactic contexts capture a different linguistic structure where turing is close to pauling, hamming. I propose to investigate variation that can be captured by using such arbitrary contexts since they capture linguistic cues not modeled by methods that only look at immediate local context.

2.2.3 Word Embeddings as Probability Distributions

There is a body of work on learning word embeddings which map a word w to a vector in $\mathcal{R}^d$ where d is the dimension of the vector space[1, 27, 28]. One limitation of representing each word as a vector is that it does not capture variance that might be associated with polysemy or how varied the usage of a word is. Recently [36] propose a method to learn word embeddings that departs from the prior work by proposing to embed each word as a normal probability distribution with an associated mean and variance. Modeling each word as a distribution thus models uncertainty and other asymmetric relationships like entailment etc. Inspired by this work, where the variance associated with the embedding for a word represents uncertainty in its contextual usage, we propose to investigate the following questions: *"How can we use word embeddings modeled as probability distributions to detect semantic variation in word usage?"* The idea is to investigate whether an increase in variance over time is indicative of a semantic change in word usage. Furthermore, by observing the variances over time, *can we predict whether a word will change and when the change point will occur?* I propose to investigate these questions not only to gain insight into linguistic variation but also investigate the properties of word embeddings that capture such uncertainty.

2.2.4 Improved change point detection

In our work, we propose null models to assess the significance of an observed change and demonstrate the effectiveness of our methods by evaluating on multiple datasets. Currently our null model assumes a non-parametric approach and therefore requires several bootstrapping iterations. This suggests an improvement: Can we design improved null models which are more computationally efficient and more stringently control the false discovery rate? One possible line of investigation towards a more effective null model is to ask "Are there assumptions about the modeled distribution that allow for fast parametric tests of significance?"

Secondly in our work we assume that a word exhibits only a single change-point. In time-series analysis relatively straight-forward extensions are outlined to generalize change point detection to detect potentially multiple change points. Furthermore, sophisticated techniques to more reliably detect change points (even multiple) are discussed by Killick [15].

In summary I propose to explore and incorporate improved change point detection techniques for detecting linguistic variation.

Chapter 3

Learning Latent Personality Traits from Social Media Text

In our previous work, we proposed methods to track and detect linguistic variation in time and geography. In this chapter, we shift our focus from language to people in social media.

People vary in their personalities and that influences their decisions, on-line behavior, and quality of life. Here, we ask *what are the basic traits that distinguish people?* Traditionally, social scientists have attempted to answer this question using surveys. The popular "Big 5" personality traits were derived by latent factor analysis over responses to *predefined questions*[25]. However recently, evolution of social media like Twitter and Facebook enables discovery of latent human differences based on *everyday behavior* (i.e. language use).

Problem Definition Several have sought to develop models for predicting the Big 5 personality factors in a supervised manner [3, 24, 31]. However we seek to *infer* a new construct (i.e. a set of factors) derived from user-level social media text. In particular we investigate several latent variable models with the goal of producing a new construct that is:

- *Generalizable*: Latent factors must be useful for a variety of predictive tasks – none of which the model is fit to directly.
- *Enduring*: These factors must be enduring and stable over individuals and time.
- *Interpretable*: We seek factors that can easily be understood and hence interpretable.

Key Insight The key idea is to observe that language can be a mirror into the personality of the author. For example, Schwartz et al. [31] observe that people who are neurotic tend to use phrases like `sick of` and `depressed` more frequently than others on social media. We adopt a Bayesian perspective where we posit that the social media text generated by users is a random variable which depends on latent hidden variables (factors) that reflect different personality traits. We therefore seek to develop latent variable models that can in infer a construct (a set of latent factors) based on social media text as evidence.

Data We use data that consists of 20 million Facebook status updates of $100K$ users. We also derive the "Big5" personality scores for these set of users by using the standard 20-100 item

	Questionnaire	Language-based			
	Big5	LDA5	LDA10	LDA25	SVD10
User Likes	0.51	**0.55***	**0.59***	**0.61***	**0.56***
80-questions	0.18	0.16	**0.18**	**0.21***	0.16
#(friends)	0.18	0.13	**0.20**	**0.28***	**0.19**

Table 3.1: Performance of various latent trait features on predicting other psychological variables of interest. (LDA**X**:LDA with X topics, SVD**X**:SVD with X factors) For User Likes we use AUC Score. Other variables we show correlation. * indicates significance at 0.05 level using Binomial Test. For #(friends) we used paired t-test. LDA25 is an upper bound.

questionnaire described by [11, 25]

Evaluation To evaluate our factors, we propose an evaluation on several fronts in order to ensure that these factors are generalizable and useful for several tasks. Specifically we seek to evaluate these factors on the following predictive tasks:

- **User Likes**: First we cluster likes of users into 100 clusters (categories). We now seek to predict based on our inferred latent factors, the cluster memberships of users.
- **Number of Friends**: Predict the number of friends a user has on the social network.
- **Psychological Questions**: We consider a set of responses to 80 psychological questions and seek to predict these responses (scores) using our latent factors. Below is a sample set of such questions that users report scores on a scale of 1-5:
    1. Am the life of a party.
    2. Am not interested in abstract ideas.
    3. Am filled with doubts about things.

3.1 Ongoing Work - Elementary Factor Models

While latent variable and factorization models, such as Latent Dirichlet Allocation (*LDA*) [8] or singular value decomposition (*SVD*), have a long tradition of use in NLP over *documents*, here we explore their use over *people* and how well they correlate with other psychological variables. As one of the first studies to examine, we start with standard models: *LDA* and *SVD* which will serve as strong baselines.

Figure 3.1 shows a set of sample topics derived using LDA. Observe that these topics include words like `friends, love, happy, god` that are indicative of emotional/psychological states. We evaluate the derived factors comprehensively based on how well they correlate with other variables as described in the evaluation section. Note that none of the derived factors were fit *a priori* to these tasks – an important prerequisite for generalizability. In Table 3.1 we present the correlations of derived latent features with different psychological variables. Observe that latent features derived from language (like the LDA model) correlate much more with several variables than the widely used “Big 5” factors.

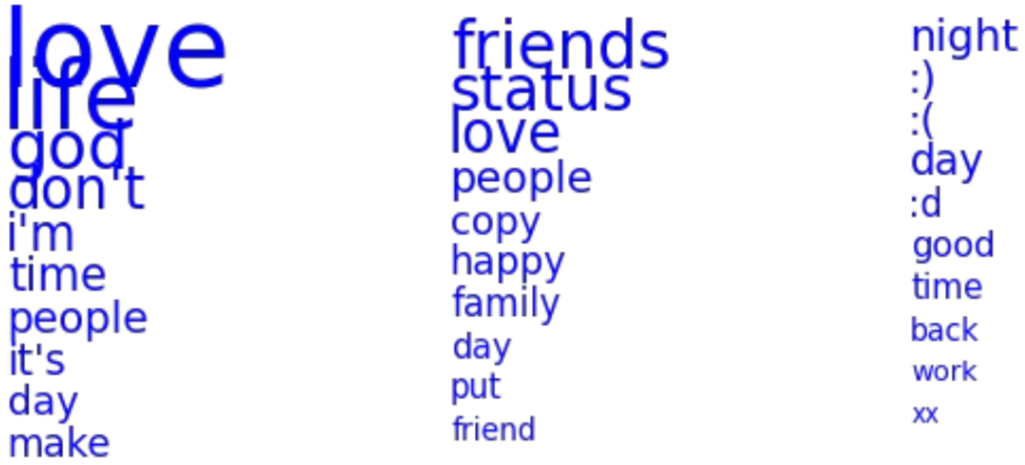


Figure 3.1: Sample set of LDA topics inferred from user status updates on Facebook using LDA with 10 topics (LDA10).

3.2 Proposed Work - Custom Factor Models

Finally, based on our preliminary yet promising results using elementary models we envision modeling personality traits using hierarchical probabilistic models to more effectively utilize deeper linguistic markers (like emotion words, semantic roles etc) to infer latent traits that are generalizable, enduring and interpretable.

One drawback of vanilla LDA is that inferred topics may not be indicative of personality traits at all (for example, there could be a topic about football or business). One straight-forward extension is to attempt to learn LDA topics on a corpus with irrelevant words filtered out based on a pre-constructed lexicon and use words that are relevant to personality. In order to do so, we could filter each users message to only the set of words contained in the PERMAv2 lexicon. The PERMAv2 contains words that known to be associated with positive/negative emotions, engagement, relationships, meaning (spirituality) and accomplishment. To illustrate, the PERMAv2 lexicon contains words like `acclaimed`, `adeptness` and `admirable` which are associated with positive emotion.

While we continue to work on developing hierarchical models, I describe one such potential model. Bamman et al. [5] propose a model to learn latent personas of film characters. They model each movie character as a set of distributions over inferred topics based on semantic roles associated with the character. In similar vein, we propose a model inspired by their work and outline this in Figure 3.2. In brief, the generative process for generating a word is as follows: Let there be U users, P personas and K topics. The personas and the topics are latent variables. We observe the words for each user and their respective annotations (emotion/sentiment tags). Each user is modeled as a probability distribution over P personas. This distribution for each user is drawn from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ which is then transformed by a logit transformation to yield a probability distribution over P personas. This allows for two things (a) users to have real valued factor scores and (b) model correlations between inferred

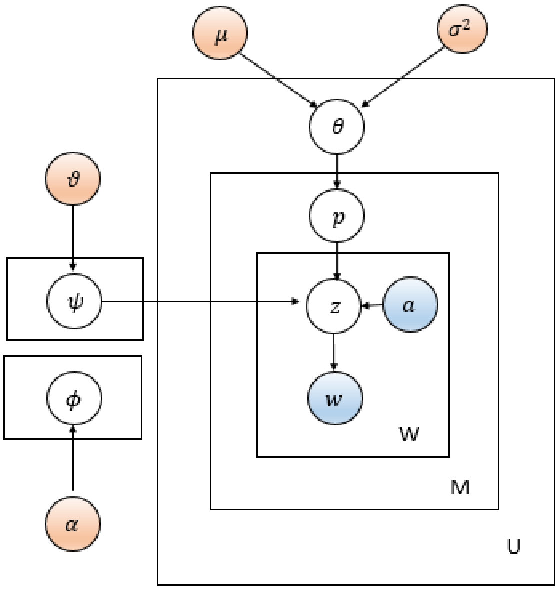


Figure 3.2: One of our potential models for capturing latent personality traits of users. Blue nodes indicate observed variables while red nodes indicate priors. We only observe words and their annotations.

factors. This insight is inspired by correlated topic models Blei and Lafferty [7]. Each message (or word) is then associated with a persona indicator p drawn from the P personas. A persona itself is a distribution over topics and is represented by ψ. Given a persona indicator p and the observed annotation for a word w, both of these select a persona from ψ and draw a topic indicator z (very much similar to Bamman et al. [5]) specific to that persona. Finally given a topic indicator z, we draw a word from the topic indexed by z from the topic set ϕ (similar to LDA Blei et al. [8]). Since conditional conjugacy does not hold in such a model, we are required to use variational inference to infer the latent parameters of the model. We are currently working on implementing this model in Stan, a probabilistic programming framework, the results of which are too preliminary to report here.

In a nutshell, I would like to continue to work on developing models that captures latent personality traits of users given their language use on the Internet.

Chapter 4

Conclusion

This thesis is concerned with detecting and analyzing linguistic variation on the Internet and online social media. In summary the main contributions are

- LANGCHANGETRACK: LANGCHANGETRACK develops methods to track and detect changes in word semantics/meaning over time.
- GEODIST: GEODIST proposes methods to use word embeddings to detect regional variational in word meaning and assessing their statistical significance by proposing a null model.
- LATENT PERSONALITY TRAITS We propose to develop latent variable models to infer a new construct (a latent set of factors) of personality traits that are generalizable, enduring and are interpretable.

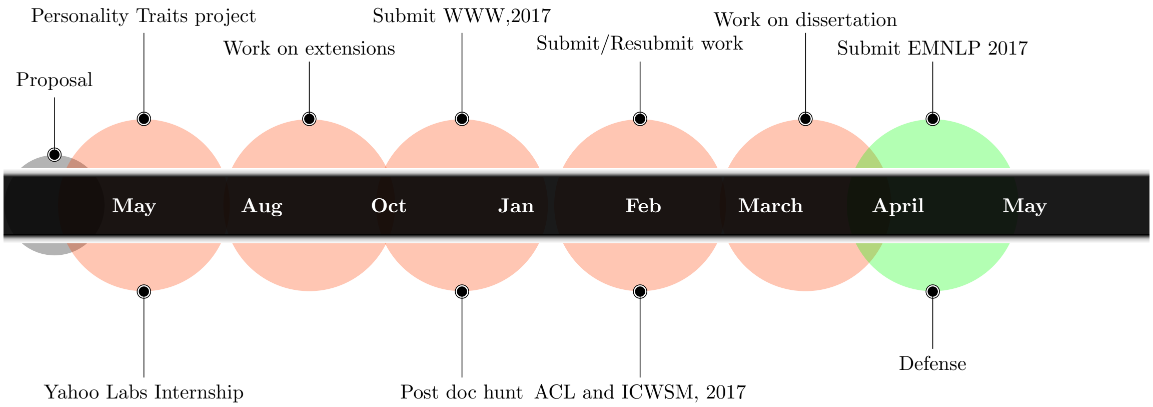


Figure 4.1: Estimated timeline of how I plan to work on my proposed work

We propose several extensions to LANGCHANGETRACK and GEODIST including focusing on signals derived from arbitrary contexts (like dependency parses) and explicitly modeling word senses. We also propose developing hierarchical probabilistic models to more effectively model socio-linguistic variables like personality traits etc.

Finally, I plan to schedule my work around deadlines for top conferences. I provide a brief summary of my estimated time-line in Figure 4.1. I account for the selectiveness of conferences and thus propose to target multiple conferences. I plan on pursuing a career in academia, and therefore as a next step, I would like to pursue a post-doc before I get on the academic job market. Consequently I have earmarked some time to look out for post-docs and travel.

Bibliography

[1] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*, 2013. 1.1, 2.2.2, 2.2.3

[2] Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. Polyglot-ner: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining*, 2015.

[3] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2): 119–123, 2009. 3

[4] David Bamman, Chris Dyer, and Noah A. Smith. Distributed representations of geographically situated language. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, June 2014. 2.1.2

[5] David Bamman, Brendan O'Connor, and Noah A Smith. Learning latent personas of film characters. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, page 352, 2014. 3.2

[6] Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. Breaking sticks and ambiguities with adaptive skip-gram. *arXiv preprint arXiv:1502.07257*, 2015. 2.2.1

[7] David Blei and John Lafferty. Correlated topic models. 3.2

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. 3.1, 3.2

[9] Yanqing Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226*, 2013. 2

[10] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *JMLR*, 2011. 2

[11] P.T. Costa, R.R. McCrae, and Inc Psychological Assessment Resources. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Psychological Assessment Resources, 1992. 3

[12] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012. 2.2.1

[13] Prateek Jain, Vivek Kulkarni, Abhradeep Thakurta, and Oliver Williams. To drop or not to

drop: Robustness, consistency and differential privacy properties of dropout. *arXiv preprint arXiv:1503.02031*, 2015. 1.1

[14] Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1191–1200. ACM, 2015. 2

[15] Rebecca Killick. Package changepoint. 2.2.4

[16] Vivek Kulkarni, Jagat Sastry Pudipeddi, Leman Akoglu, Joshua T Vogelstein, R Jacob Vogelstein, Sephira Ryman, and Rex E Jung. Sex differences in the human connectome. In *Brain and Health Informatics*, pages 82–91. Springer, 2013. 1.1

[17] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Exploring the power of gpu's for training polyglot language models. *CoRR*, abs/1404.1521, 2014.

[18] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee, 2015. 1.1, 1.3.1, 1.3.2, 2.1.1, 2.1.1, 2.1.2

[19] Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. Freshman or fresher? quantifying the geographic variation of internet language. *arXiv preprint arXiv:1510.06786*, 2015. 1.1, 2.1.2

[20] William. Labov. *Locating language in time and space / edited by William Labov*. Academic Press New York, 1980. 1, 2

[21] Deep Learning. Last words. *Computational Linguistics*, 41(4), 2015. 1.3.1

[22] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. 2.2.2

[23] Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, et al. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, 2012. 2.1.1

[24] François Mairesse and Marilyn Walker. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 85–88. Association for Computational Linguistics, 2006. 3

[25] Robert R McCrae and Paul T Costa. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81, 1987. 3, 3

[26] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011. 2.1.1

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 1, 2.2.2, 2.2.3

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors

for word representation. 2.2.2, 2.2.3

[29] Bryan Perozzi, Rami Al-Rfou, Vivek Kulkarni, and Steven Skiena. Inducing language networks from continuous space word representations. In *Complex Networks V*. 2014. 1.1, 2

[30] Bryan Perozzi, Vivek Kulkarni, and Steven Skiena. Walklets: Multiscale graph embeddings for interpretable network classification. *Submitted*, 2016. 1.1

[31] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013. 3, 3

[32] Eran Shor, Arnout van de Rijt, Alex Miltsov, Vivek Kulkarni, and Steven Skiena. A paper ceiling explaining the persistent underrepresentation of women in printed news. *American Sociological Review*, 80(5):960–984, 2015. 1.1

[33] Sali A. Tagliamonte. *Analysing Sociolinguistic Variation*. Cambridge University Press, 2006. 1, 2

[34] Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. A probabilistic model for learning multi-prototype word embeddings. 2.2.1

[35] Vassileios Tsetsos et al. Personalization based on semantic web technologies. *Semantic Web Engineering in the Knowledge Society*. 2

[36] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623*, 2014. 2.2.3