Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media

Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language

Yuheng Hu　**Kartik Talamadupula**　**Subbarao Kambhampati**
Department of Computer Science, Arizona State University, Tempe AZ 85287
{yuhenghu, krt, rao}@asu.edu

Abstract

Twitter has become the *de facto* information sharing and communication platform. Given the factors that influence language on Twitter – size limitation as well as communication and content-sharing mechanisms – there is a continuing debate about the position of Twitter's *language* in the spectrum of language on various established mediums. These include SMS and chat on the one hand (size limitations) and email (communication), blogs and newspapers (content sharing) on the other. To provide a way of determining this, we propose a computational framework that offers insights into the linguistic style of all these mediums. Our framework consists of two parts. The first part builds upon a set of linguistic features to quantify the language of a given medium. The second part introduces a flexible factorization framework, SocLin, which conducts a psycholinguistic analysis of a given medium with the help of an external cognitive and affective knowledge base. Applying this analytical framework to various corpora from several major mediums, we gather statistics in order to compare the linguistics of Twitter with these other mediums via a quantitative comparative study. We present several key insights: (1) Twitter's language is surprisingly more conservative, and less informal than SMS and online chat; (2) Twitter users appear to be developing linguistically unique styles; (3) Twitter's usage of temporal references is similar to SMS and chat; and (4) Twitter has less variation of affect than other more formal mediums. The language of Twitter can thus be seen as a projection of a more formal register into a size-restricted space.

1 Introduction

Given their ubiquity, immediacy and accessibility, social media channels such as Twitter have emerged as the *de facto* medium for information sharing, and communication about various topics from breaking news to personal stories. Twitter houses many features that make its language distinct. On the one hand, unlike on traditional media like blogs, magazines and newspapers, posts on Twitter (tweets) are inherently much shorter and constrained by a hard 140 character limit. As a result, the language of Twitter is widely believed to be highly compact and brief. On the other hand, while Twitter does share a brevity of expression with mediums like text messages (SMS) and online chat, it also encourages discussion on a much wider variety of topics and information (e.g. news events) than the "one-to-one" near-synchronous modalities afforded by SMS and chat. Additionally, Twitter has unique communication mechanisms where users can build a following and follow other users' tweets, which provides an opportunity to re-post and hence edit the content of others' tweets too. These external (character limit) and internal (content) differences contribute to the variations of Twitter's linguistics.

The subject of Twitter's language has received a lot of attention in the popular press; a continuing debate is on its position in the spectrum of well established "casual communication mediums" like SMS and chat on the one hand, and more formal mediums like emails, blogs, magazines and newspapers on the other. One argument is that the severe length restrictions on tweets induce a grammatically incorrect and aberrant language riddled with acronyms, hashtags etc. that has similarities to the language used in SMS and chat. An alternate view is that Twitter is really a length-restricted version of the language of more formal media. Evaluating the relative accuracy of these alternative theories of Twitter language is of importance to various applications in anthropology, communication studies, sociology and many subareas within computer science – text mining, computational linguistics, and machine translation.

Given the important role that Twitter usage is increasingly playing in daily life, a growing body of literature has emerged in the social network / media research community that aims to mine Twitter content, or to evaluate the linguistic aspects of that content in order to better understand the dynamics of content on Twitter (Golder and Macy 2011). However, thus far, researchers have predominantly looked at analyses that focus almost exclusively on either textual content, or network analysis, or simple and specific linguistic aspects of tweets such as length and hashtags. The characteristics of language on Twitter, and how different it is from other mediums, is an under-explored area. A primary research challenge, therefore, is to find a principled way to identify a set of aspects of Twitter's language that can be used to compare them with content from other mediums. In this study, we add to the budding work on the nature of Twitter language by setting out to discover what this language is really like. In specific, we are interested in comparing Twitter and other media in terms of linguistic and psycholinguistic features of the language used on them.

Our Methodology Since our larger goal is to place Twitter in the linguistic spectrum alongside other established mediums, we contribute a novel methodological approach to the study of computer-mediated communication (CMC) by utilizing comparative methods for analyzing linguistic and psycholinguistic features against multiple corpora including SMS, chat, email, blogs, magazines and newspapers. We propose a computational framework to gain insights into the linguistic styles of these different mediums (with central focus on Twitter). Rooted in the linguistics literature, our framework consists of two parts. The first part builds upon a set of orthographic and grammatical elements to quantify the linguistic style of the input medium. In the second part, we devise a flexible factorization method – SocLin – which infers the psycholinguistic aspects of the input medium by seeking low-rank representations of words and documents of the given medium on specified categories with the help of available external cognitive and affective knowledge from the 'Linguistic Inquiry Word Count' (LIWC) dictionary (Pennebaker, Francis, and Booth 2001). This framework enables a much deeper understanding about the underlying language of a medium, providing data to answer questions such as "what is the stylistic difference between Twitter and SMS", or "what is the most prevalent emotion on Twitter"?

Our Results Applying this analytical framework to the various corpora, we gather statistics to compare the linguistics of Twitter with these other mediums via a quantitative comparative study. Several key insights are revealed: (1) We find that Twitter in general is surprisingly more conservative, formal and less conversational than SMS and online chat although it shares a similar brevity and interactivity. Its primary usage is to convey information (either for sharing news or broadcasting self-status). (2) Twitter users appear to be developing linguistically unique styles when compared against other mediums – for example, both first-person and third-person pronouns are extensively used, whereas other mediums tend to stick to one type of pronoun. (3) We find that Twitter exhibits usage of temporal references that is similar to SMS and online chat. (4) Twitter has less variations of affect when compared to email, blog, slate and news, and it tends more toward positive moods than other mediums.

We conclude that the language of Twitter is a highly dynamic repository of linguistic mores, and that while it tends to mimic the linguistic practices of traditional media in an unremarkable manner by certain measures, it also exhibits a proclivity to adapt to communication needs by exhibiting more than a passing similarity to the language of newer and less orthodox mediums. In short, we find that the language of Twitter is a projection of the language of more formal media down into a space restricted by size. To the best of our knowledge, this work and the results herein are the first quantitative comprehensive study of Twitter's linguistics with respect to other mediums.

The rest of this paper is organized as follows: first, we look at related work from linguistics as well as work which considers Twitter in specific. We then present a way to characterize linguistic styles, and the methods and models that we use to measure linguistic and psycholinguistic features. This is followed by an exposition of our results, and a detailed discussion of these as well as the implications of these results on understanding the language of Twitter. We conclude by looking at future work.

2 Related Work

Considerable research has been performed in the fields of linguistics and communication to examine the styles and structure of language. Various analyses have been performed from different angles, such as the usage of grammars (Lakoff 1979), the cognitive process involved in picking words and the linguistic style (Flower and Hayes 1981), the variations across different registers (Biber 1991), and the correlation between style and gender (Carroll 2008). With the burgeoning use of computers and the Internet, research has turned more toward the language of CMC and that on the Internet (Crystal 2001). The results from this line of work mostly focus on characterizing linguistic styles on various platforms such as SMS (Thurlow and Brown 2003), IM (Tagliamonte and Denis 2008), emails (Baron 1998) and blogs (Herring et al. 2004).

Due to the popularity of Twitter and its freely available data, there is also a rich body of work that focuses on content and linguistic analysis of tweets. One particularly interesting type of linguistic activity in social media has to do with conversations, i.e., exchanges between one or more individuals. Java et al. (2007) found that 21% of users in their study used Twitter for conversational purposes and that 12.5% of all posts were part of conversations. Similar investigations were also conducted in (Naaman, Boase, and Lai 2010). Danescu-Niculescu-Mizil et al. studied linguistic style accommodation on Twitter (Danescu-Niculescu-Mizil, Gamon, and Dumais 2011). Eisenstein et al. investigated the role geographic variation plays on language in Twitter (Eisenstein et al. 2010) while the work by Hong et al. focuses on the cultural differences in Twitter's language (Hong, Convertino, and Chi 2011). In terms of content analysis, Hu et al. developed a topical model to correlate the dynamics of Twitter posts to news events (Hu et al. 2012). In addition to this, various learning-based approaches have been proposed to enhance textual understanding of tweets (Ritter et al. 2011; Owoputi et al. 2013; Hu et al. 2013).

Limitations of Prior Work In spite of the great progress in comparing linguistic styles, the characteristics of language on Twitter remain largely unexplored. Although a lot of this can be explained by the presence of properties that are special to Twitter – the strict length limitation, hashtags, and the underlying network structure that enables the sharing of content – it is still unclear how these differences manifest themselves in comparison with other mediums. Apart from these superficial features, there has also been very little analysis of whether there are deeper interactions among linguistic features within Twitter that distinguish the language from other mediums.

3 Linguistic Analysis

In this section, we present a framework geared towards characterizing the linguistic style and psycholinguistic aspects on a given medium. This framework is then applied in a comparative study to position Twitter's language in a wide spectrum of languages on various established mediums. Our framework consists of two parts. The first part builds upon a set of linguistic features to quantify the language of a given medium. The second part introduces a flexible factorization framework, SocLin, which conducts a psycholinguistic analysis of a given medium with the help of external cognitive and affective knowledge from the LIWC dictionary.

3.1 Quantifying Linguistic Style

We first turn our attention towards the style of the language used in the various corpora that we consider. The literature on linguistics and communication (Wardhaugh 2011) posits that the style of a language can be evaluated from two different perspectives: (1) Orthographic, which includes features such as singular pronouns, word frequency and lexical density; and (2) Grammatical, which includes the usage of personal pronouns, intensifiers and temporal references.

Word Frequency Word frequency is widely used to estimate the difficulty and readability of words, sentences and documents (Breland 1996). We measure average word frequency using a list of the top 100,000 most frequent words from the Corpus of Contemporary American English (COCA) (Davies 2010). In specific, for a given corpus A, we calculate the average word frequency as $WF_A = \sum_{w_i \in A} WordFreq(w_i)/Size_A$, where $WordFreq(w_i)$ is obtained by looking up each word in the documents of corpus A and summing up the log word frequencies; $Size_A$ represents the total number of words in A. Based on the common perception, we conjuncture that the language of Twitter– and SMS and online chat – contains higher average word frequencies and fewer "difficult" or uncommon words than blogs and news articles, which are written after considerable deliberation and using structured processes by professional writers. We evaluate this in our experiments, presented in the next section.

Lexical Density Although analyzing the word frequency of various corpora is initially revealing, such an analysis is incapable of characterizing important linguistic issues such as the register and genre of text or discourse (e.g. *formal* or *intimate*). To enable this, we considered another important orthographic metric – lexical density. Lexical density captures the stylistic difference between various documents by measuring the proportion of the lexical words over the total words. Lexical words are mostly made up of verbs, nouns, adjectives and adverbs (the so-called content or information carrying words). Consider the following example from Halliday (Halliday and Matthiessen 2004) – given the two documents:

D1: *Investment in a rail facility implies a long term commitment.*

D2: *If you invest in a rail facility this implies that you are going to be committed for a long time.*

the first one has a higher lexical density (LD = 0.7) than the second (LD = 0.35). The second sentence represents spoken communication much better, whereas the first one seems more like a written communication. Similar to our calculation of word frequency, we obtain the average lexical density for a given corpus medium by summing up the lexical density for each document in that corpus and then dividing by the total number of documents.

In general, it is widely believed that high lexical density indicates a large amount of information-carrying words and a low lexical density indicates relatively fewer of these words. It may therefore be expected that well-written or organized pieces of text such as blogs and newspapers will exhibit much higher lexicon density than tweets. Besides, we also conjecture that tweets will have higher lexical density than SMS and chat because they are less conversational. However, one must also recognize the underlying tension between this and Twitter's use as a medium to convey information by projecting content from longer media while modifying it to fit within the 140 character limit, which can make tweets less formal and decrease their lexical density.

Personal Pronouns In addition to the two orthographic features described above, we also consider several grammatical features of a language. The most widely-adopted grammar measure is the usage of personal pronouns. Past work states that first and second-person pronouns are more frequently used in conversation-based mediums (e.g. speech) than in writing. More interestingly, Yates found that the language used in computer-mediated communication resembles speech much more than it does writing (Yates 1996). Note that SMS and online chat are considered as computer-mediated communication (CMC).

In order to position Twitter along the media spectrum, we calculated the percentage of personal pronouns as $PP_A(P_i) = \sum_{w \in P_i} Freq(w)/Size_A$, where $P_i \in P = \{P_1, P_2, P_3\}$ is a class of pronouns for each person, e.g., P_1 contains first-person pronouns such as *I*, *me*, and *myself* etc. $Freq(w)$ represents the number of occurrences of a given personal pronoun inside a corpus A, and $Size_A$ represents the total size of corpus A. For this feature, we conjecture that the language of Twitter will turn out closer to mediums that exhibit speech-like modalities in the usage of personal pronouns, since conversation is one of the key parts of Twitter.

Intensifiers We also consider intensifiers of a language. Grammatically speaking, intensifiers are adverbs that maximize or boost meaning, as the following examples with intensifiers marked in bold demonstrate:

D3: my clean room is **so** weird

D4: haha it was kinda creepy, but **very** cool!!

D5: that meal was **really** awesome ...

Intensifiers are an interesting area of grammar, partly because of a speaker or writer's desire to be "original" to demonstrate *verbal* skills, and to capture the attention of an audience (Ito and Tagliamonte 2003). This presents us with an ideal feature to investigate and test the common hypothesis that Twitter is similar to the speech-like mediums like SMS and online chat. Similar to the calculation of the statistics for personal pronoun usage, we compute frequency of intensifiers on our collected corpora as: $INT_A(Int_i) = \sum_{w \in Int_i} Freq(w)/Size_A$, where $Int = \{Int_1, Int_2, ..., Int_{25}\}$ consists of 25 the most commonly used intensifiers listed in (Quirk et al. 1985).

In past research on intensifier usage, Ito and Tagliamonte have discovered that the usage of "*very*" – the most frequent intensifier in contemporary English usage (Bäcklund 1973) – is prevalent only among older individuals. In contrast, the newer variant "*really*" is increasingly used by the younger generation (Ito and Tagliamonte 2003). Given that Twitter is mostly used by the youth (age from 18 - 29) (Smith and Brenner 2012), we conjecture that *really* should be used much more frequently on Twitter. In contrast to this, newspapers and blogs, which are written by and for the consumption of older people may feature higher usage of the intensifier *very*. We also evaluate the usage of the intensifier *so*.

Temporal References The last grammatical characteristic of language that we consider is the usage of temporal references – particularly, references to the future. Tenses can be understood to indicate the location of an event or state on a time axis relative to a reference time, which is usually taken as the writing time. When an event or state takes place or holds before the time of speech, the tense is past tense; in the reversed situation, the tense is future tense; and when the process or state overlaps with the speech time, the tense is present. The usage of temporal references can thus be used to gauge underlying activities. In order to investigate whether a medium is more about the present or the future, we collect words corresponding to five temporal reference categories: *going to, gonna, will, shall* and the simple present tense. Similar to previous analysis for the intensifier and personal pronouns, we calculate the frequency of each temporal category as as $TR_A(T_i) = \sum_{w \in T_i} Freq(w)/Size_A$, where T is our temporal categories.

Since Twitter is home to the unfolding of many a breaking news event – which tend to happen in the present – we expect that references to the present are more common on Twitter than in other media like blogs and newspapers, where future references may hold sway.

3.2 Psycholinguistic Analysis

In addition to the linguistic analysis using orthographic and grammatical features, it is also important to analyze the psycholinguistics of the language, namely, cognitive and affective aspects of the language in use – factors that influence the generation of content. A large body of research shows that people tend to organize their thoughts via cognitive processes before choosing words, linguistic styles and affects in their writings (Rosenwasser and Stephen 2011; Magnifico 2010; Flower and Hayes 1981). For example, when people write a tweet or blog-post pertaining to a presidential campaign, their understanding of that campaign (i.e., the cognitive process) plays an important role in picking the right affect (e.g. positive, negative, sad) and the proper style (formal, casual etc.).

We investigate whether there are underlying cognitive and affective aspects that differentiate Twitter from the other mediums. In order to find the right set of words to measure these affects, we follow a linguistic methodology used in a variety of applications known LIWC. LIWC was designed to facilitate the understanding of individuals' cognitive and emotional status through text analysis. As a result, most of the categories in LIWC concern mental activity, with over 3,700 words related to cognitive and affective aspects.

LIWC is a dictionary. In order to infer cognitive and affective aspects for a document (e.g., a tweet, a blog-post, a newspaper article), we need an effective way to transform the word-level aspects into that of document-level. One straightforward solution is to first calculate statistics for all cognitive and affective-oriented words in a document and later aggregate them for that document. In practice, however, word occurrences can be sparse especially for short-text like tweets, SMS and online chat (which mention only a handful of words). Therefore, a counting based approach may generate highly unreliable statistics.

Inspired by the work in sentiment analysis, which also suffers from similar sparsity issues (e.g., sentiment lexicons are not found in most documents), we propose a framework called SocLin for inferring the psycholinguistic properties of documents. In specific, SocLin seeks a low-rank representation of the collected corpus by factorizing a term-document matrix into two major factors corresponding to term-aspects and document-aspects. Cognitive and affective prior knowledge from LIWC is leveraged to provide supervision on the term-aspects factor. SocLin advances the counting-based approaches in that it takes the contextual information of the collected documents into account (this information is embedded in a term-document matrix) and propagates it across documents during factorization. As a result of this, although a document may not contain any cognitive or affective-oriented words, SocLin may still able to infer such aspects as long as that document shares some context with another document which has these kind of words. Considering aggregated contextual information has proved successful in sentiment analysis and has achieved significantly better results than counting-based methods (see (Li, Zhang, and Sindhwani 2009) for example).

Formally, let a corpus for a medium consist of n documents (e.g., tweets, emails), contributing to a vocabulary of N terms. SocLin takes these documents (in terms of the term-document matrix $\mathbf{X}$) as input and decomposes them into three parts (which include one smoothing factor) that specify soft membership of documents and terms in each latent psycholinguistic dimension. The supervision from LIWC is enforced as constraints on the learning process of our model. In other words, our basic model tries to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{T},\mathbf{G}} \quad & \mathcal{J} = \left\| \mathbf{X} - \mathbf{T}\mathbf{S}\mathbf{D}^\top \right\|_F^2 \\ & + \alpha Tr\left((\mathbf{T} - \mathbf{T}_0)^\top \mathbf{\Lambda} (\mathbf{T} - \mathbf{T}_0) \right) \\ s.t. \quad & \mathbf{T} \geqslant 0, \mathbf{S} \geqslant 0, \mathbf{D} \geqslant 0 \end{aligned} \tag{1}$$

where $\| \cdot \|_F$ is the Frobenius Matrix Norm and $tr(\cdot)$ is the matrix trace. $\mathbf{T} \in \mathbb{R}^{N \times k}$ indicates the assignment of each document to the relevant k aspects based on the strength of their associations. That is, the i-th row of $\mathbf{T}$ corresponds to the posterior probability of word i referring to k aspects defined in LIWC. Similarly, $\mathbf{D} \in \mathbb{R}^{n \times k}$ indicates the posterior probability of a document n belonging to the k aspects. $\mathbf{S} \in \mathbb{R}^{k \times k}$ provides a condensed (smoothed) view of X. We encode prior information from LIWC in a term-aspect matrix $\mathbf{T}_0$, where $\mathbf{T}_0(i,j) = 1$ if the a word i belongs to the j-th category in LIWC, and $\mathbf{F}_0(i,j) = 0$ if not. $\alpha > 0$ is the parameter which determines the extent to which we enforce $\mathbf{T} \approx \mathbf{T}_0$. $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$ is a diagonal matrix, indicating the entries of $\mathbf{T}_0$ that correspond to labeled entities. As a result of this factorization, we can readily determine whether a document has cognitive and affective aspects from the factorization result $\mathbf{D}$. Note that the non-negativity in SocLin makes the factorized factors easy to interpret.

Conceptually, our basic matrix factorization framework is similar to the probabilistic latent semantic indexing (PLSI) model (Hofmann 1999). In PLSI, $\mathbf{X}$ is viewed as the joint distribution between words and documents, which is factorized into three components: $\mathbf{W}$ is the word class-conditional probability, $\mathbf{D}$ is the document class-conditional probability and $\mathbf{S}$ is the condensed view of $\mathbf{X}$.

Model Inference The coupling between $\mathbf{T}$, $\mathbf{S}$, and $\mathbf{D}$ makes it difficult to find optimal solutions for all factors simultaneously. We adopt an alternative optimization scheme (Ding et al. 2006) for Eq. 1, under which we update $\mathbf{T}$, $\mathbf{S}$ and $\mathbf{D}$ alternatingly with the following multiplicative update rules. Due to the space limit, the detailed inference procedure is omitted.

First, for the tweets-segment matrix $\mathbf{T}$, we have:

$$T_{ij} \leftarrow T_{ij} \sqrt{\frac{(\mathbf{X}\mathbf{D}\mathbf{S}^\top + \alpha\mathbf{\Lambda}\mathbf{T}_0)_{ij}}{(\mathbf{T}\mathbf{T}^\top\mathbf{X}\mathbf{D}\mathbf{S}^\top + \alpha\mathbf{\Lambda}\mathbf{T})_{ij}}} \tag{2}$$

Next, for the tweets-segment matrix $\mathbf{S}$, we have:

$$S_{ij} \leftarrow S_{ij} \sqrt{\frac{(\mathbf{D}^\top\mathbf{X}\mathbf{T})_{ij}}{(\mathbf{D}^\top\mathbf{D}\mathbf{S}\mathbf{T}^\top\mathbf{T})_{ij}}} \tag{3}$$

Last, for the document-segment matrix $\mathbf{D}$, we have:

$$D_{ij} \leftarrow D_{ij} \sqrt{\frac{(\mathbf{X}\mathbf{D}\mathbf{S}^\top)_{ij}}{(\mathbf{D}\mathbf{D}^\top\mathbf{X}^\top\mathbf{T}\mathbf{S})_{ij}}} \tag{4}$$

Our learning algorithm (see below) consists of an iterative procedure using the above rules until convergence. The correctness and convergence of Algorithm 1 can be proved based on Karush-Kuhn-Tucker (KKT)(Nocedal and Wright 2000) conditions of Eq.(2), Eq.(3), Eq.(4). Since the term-document matrix $\mathbf{X}$ is typically very sparse with $z \ll n \times N$ non-zero entries. k is typically also much smaller than n and N. Thus, our approach can scale to large datasets.

SocLin in Practice In practice, we extract 3,712 words from 14 cognitive and affective categories of LIWC (such as *Positive emotion*, *Anxiety*, *Anger*, *Causation*, *Inhibition*) to form the psycholinguistic knowledge $\mathbf{T}_0$. With this prior, we apply SocLin to infer cognitive and affective aspects of each document in the given corpus. As a result, we get the document-aspect factor $\mathbf{D}$, where each row of $\mathbf{D}$ represents the probability distribution that a document d belongs to those 14 psycholinguistic categories. We assign d to the category with the highest probability, and compute the statistics for each category following this.

Algorithm 1: Aspects Factorization with LIWC.

```
input  : $\alpha$
output : $\mathbf{T}$, $\mathbf{S}$, $\mathbf{D}$
Initialize $\mathbf{T} \geqslant 0, \mathbf{D} \geqslant 0, \mathbf{S} = (\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{T}^\top\mathbf{X}\mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}$
while Algorithm Not Converges do
    Update $\mathbf{T}$ with Eq.(2) while fixing $\mathbf{S}$,$\mathbf{D}$
    Update $\mathbf{S}$ with Eq.(3) while fixing $\mathbf{T}$,$\mathbf{D}$
    Update $\mathbf{D}$ with Eq.(4) while fixing $\mathbf{T}$,$\mathbf{S}$
end
```

4 Results

In this section, we quantitatively investigate the linguistic style and psycholinguistic aspects of Twitter's language, with comparisons to other mediums like SMS and chat on one side and email, blogs and newspapers on the other.

Datasets We used seven large-scale datasets in our study:

Tweets: We collected over 45 million tweets in Oct 2012 with the Twitter Firehose access level, which (according to Twitter) returns all public statuses. Non-English tweets were removed since they are outside the scope of this paper. Our data was randomly sampled from Twitter and *was not* biased toward news organizations, celebrities or other "top" users.
SMS: We used the English SMS dataset provided by the NLP group at the National University of Singapore[1]. These SMS texts were collected from 2010 to 2012. Most users were native English speakers (average age = 21.7).
Online Chat: We used the standard NPS chat corpus[2], which was collected in 2006 from various online chat rooms. The corpus is organized into 15 files, where each file contains several hundred textual posts from age-specific chatrooms (teens, 20s, and 30s). We merge these files into one file where each line represents one post. In total, we have roughly over 10,000 messages.
Email: This dataset is a part of the Enron Dataset[3], which contains over 200,000 email communications between employees of Enron. For our experiments, we eliminated blank and duplicated emails.
Blogs: We obtained this data from the ICWSM 2011 Spinn3r dataset[4]. Only English blog-posts that were posted between Jan 13th and Feb 14th of 2011 were considered for our experiments.
Magazine (Slate): Slate is a United States based English language online magazine that deals with current affairs and

[1] http://wing.comp.nus.edu.sg:8080/SMSCorpus/

[2] http://faculty.nps.edu/cmartell/NPSChat.htm

[3] http://www.isi.edu/ adibi/Enron/Enron.htm

[4] http://www.icwsm.org/data/

culture issues. We obtained this dataset from the Open American National Corpus (OANC)[5].

News: Our last dataset was the popular Reuters news dataset, which has been used for evaluating the performance of text categorization algorithms. We merged the training and testing data together to form a larger corpus.[6]

	#Docs	ShortWords (per doc)	Length (by word)	Length (by chars.)
Twitter	46,480,800	7.60	12.21	53.74
SMS	51,654	8.08	10.88	40.65
Chat	10,567	2.56	3.81	18.72
Email	244,626	137.68	255.04	1306.34
Blog	24,004	147.96	269.75	1323.65
Mag.	186,020	382.43	682.09	3274.28
News	10,788	73.83	129.41	619.32

Table 1: *Statistics about our dataset; short words are defined as 3 characters or less.*

Basic Statistics We compiled some basic statistics about the dataset that we used, and the results are presented in Table 1. There is a clear demarcation in terms of the average document length (both by characters and by words) and the percentage of short words per document, between SMS and online chat on the one hand, and email, blogs, magazines and news articles on the other. Given the restriction on the size of tweets, it is not surprising that Twitter tends to cluster with the former on this measure. A more surprising observation is that even though the character limits for Twitter and SMS are 140 and 160 characters respectively, the corresponding average character length per document does not reach even half of this amount (53.74 for Twitter, 40.65 for SMS).

Experimental Setup and Procedure Within the proposed computational framework, two main tasks are undertaken to quantify the language on Twitter. First, we examine the linguistic style of Twitter and the other corpora, encompassing orthographic features and three grammatical measures. Then, using our factorization framework SocLin with supervision from LIWC, we explore the psycholinguistic aspects inherent in the language. We perform coarse parameter tuning for SocLin for every corpus: for a given corpus C, we vary the parameter α_c and choose the value that minimizes the reconstruction error in our training set for that corpus.

Research Questions We propose to perform experiments that will answer two key research questions that relate to the linguistic style and psycholinguistic aspects of language as it is used on Twitter, and compare this to its usage in other mediums.

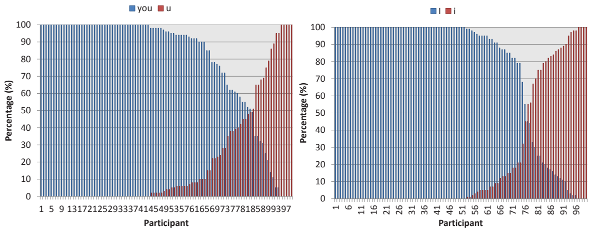
> **RQ1**: Is there a fundamental difference between linguistic styles on Twitter and other mediums? If yes, what is the reason for this contrast?

> **RQ2**: Are there distinct affective and cognitive processes underlying the generation of content on Twitter that are manifested via psycholinguistic factors?

4.1 Linguistic Styles

In order to study and classify the linguistic styles underlying Twitter and answer RQ1, we first consider three common orthographic features, followed by some components of grammar.

Usage of Singular Pronouns The use of singular pronouns is a quick and useful way of analyzing the language of Twitter due to the conversational nature of the medium. On a medium with emergent linguistic modalities like Twitter – and moreover one with a hard character limit – the intuitive assumption (and indeed one that is used often in the popular media) is that the regular, orthographic forms of singular pronouns like *you* and *I* are replaced by the more convenient (and short) *u* and *i* respectively. To analyze this assumption, we first choose 100 users at random for each pronoun such that each user had at least 100 tweets that contained one or the other form of that pronoun. We then compute the percentage of use of the regular form versus the informal (shortened or lowercase) form, and plot them as shown in Figure 1. The data comprehensively reject the intuitive assumption outlined above; if anything, the relatively low number of users who use *both* the regular and informal forms shows that the selection is more of an individual's stylistic choice, as Tagliamonte et al. (2008) found in the case of IM data. A stronger point that can be made is that this disproves the assumption that a shorter, space-restricted medium like Twitter leads to a degradation of language; if anything, the data only strengthen the argument that users retain their style on Twitter, and that the linguistics of Twitter are merely a projection of the language and style of longer media into a restricted character space.



(a) usage of *you* vs. *u* **(b)** usage of *I* vs. *i*

Figure 1: *Usage of Singular Pronouns*

Word Frequency This experiment studies yet another orthographic feature, word frequency, across different mediums. As mentioned in Section 3.1, word frequency is used to estimate the difficulty and readability of words and documents. The results are shown in the first row of Table 2. It is clear that in SMS and online chat people prefer to use highly frequent and easy words (i.e., their WF is high). This is expected since SMS and online chat are mostly casual and may not be very serious. In contrast, content for magazines and news articles is mostly generated by professional writers, who may use a much larger vocabulary consisting of harder words, thus leading to lower average word frequencies. The

[5] http://www.americannationalcorpus.org/

[6] Readers may notice that the datasets we use are varied in terms of the time periods that they cover. Although it would be ideal to have corpora from the exact same time, we decided to overlook this issue in the interests of a more wide-ranging analysis.

most interesting discovery comes from Twitter: although it shares the short, compact and interactive nature of SMS and online chat, the word choice in its language tends to exhibit more similarity toward email and blogs. We conclude that at least a considerable amount of tweets are written after some deliberation, making Twitter a more serious communicative platform than SMS and online chat.

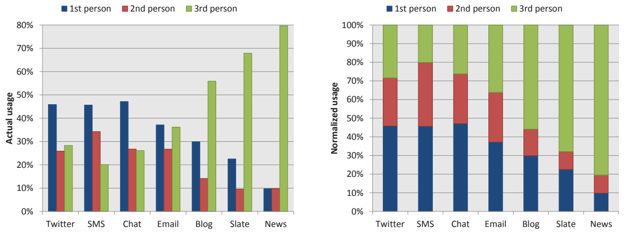
Lexical Density Lexical density (LD) is another effective orthographic feature that reveals the register and genre of a document (e.g., formal or informal), and can be used to quantify the stylistic differences between various mediums. We calculate lexical density using the procedure mentioned in Section 3.1. The averaged results for Twitter and other mediums are shown in row 2 of Table 2. Note that, in general, a higher LD indicates a larger usage of information-carrying words (i.e., lexical words) within the text (Wardhaugh 2011). This theory is borne out in our experiments: both SMS and online chat are relegated to the lower end of the spectrum where LD values are concerned. In contrast, blogs and news articles show a remarkably higher density, indicating that there is more of an emphasis on communicating information in an efficient, concise manner. Twitter falls in between these two extremes, yet is closer to blogs and news; we believe this to be a strong manifestation of the fact that although Twitter is used primarily as a medium to convey information, documents (tweets) must also necessarily stay within the 140 character limit imposed on them. The mere fact that the lexical density on Twitter is higher than SMS and online chat, and very nearly that of an online magazine like Slate, indicates that tweets are well-organized linguistically.

	Tw.	**SMS**	**Chat**	**Email**	**Blog**	**Slate**	**News**
WF	4.64	5.52	5.98	4.54	4.33	2.87	2.63
LD	0.47	0.42	0.40	0.44	0.54	0.48	0.58

Table 2: *Statistics about Word Frequency (WF) and Lexical Density (LD) across all the mediums. A one-way analysis of variance was conducted, and it showed a clear statistical difference ($p < 0.001$) between these seven corpora.*

Usage of Personal Pronouns In addition to the three orthographic features, we also explore the grammatical aspects of linguistics on Twitter and other mediums. According to past research, first and especially second-person pronouns are more frequently used in conversational circumstances (e.g., speech and debate) while third-person pronouns are more frequently used in reporting. Given the facts that conversation is highly prevalent on Twitter, and that the medium's open nature and follower/followee relationship model can set in motion a conversation thread, we wondered where Twitter would be positioned vis-à-vis other mediums. If Twitter is indeed a conversation-like medium, parallelism with SMS and online chat should be expected.

To answer this question, we extract all pronouns of personal reference in the corpora and measure their usage frequency. The results are shown in Figure 2. It is clear that on Twitter, personal reference is dominated by first-person pronouns at 45.8% (see Figure 2(a)), contributing to its placement on roughly the same end of the spectrum as mediums like SMS and online chat (whose first-person pronouns appear in 45.6% and 47.1% of instances, respectively). However, comparison of the usage of second and third-person pronouns reveals significant differences between Twitter and its short-text counterparts. In particular, people tend to use more third-person pronouns than second-person pronouns on Twitter (28.3% vs. 25.8%); this trend is reversed in SMS and online chat. The higher usage of third-person pronouns can be also found at the other end of the spectrum, in blogs and particularly in news corpora, where third-person pronouns are used to present news and other topics in an insightful and impartial manner (see Figure 2(b) for a normalized view).

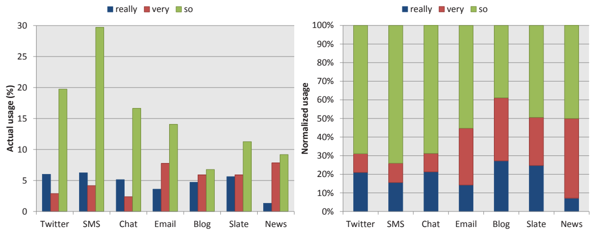


(a) Personal Pronoun Usage **(b)** Normalized Usage

Figure 2: *Usage of Personal Pronouns*

In fact, the heavy usage of first and third-person pronouns on Twitter conforms to previous research (e.g., (Naaman, Boase, and Lai 2010)) that Twitter is not just about updating self status, but also information sharing (e.g., breaking news). It is worth noting that unlike previous work which reveals this point from a topic-based perspective, we demonstrate that we can achieve the same finding from a linguistic perspective as well.

Usage of Intensifiers To study intensifier usage on Twitter and compare the results with other corpora, we extract the 25 most commonly used words in the corpus which are capable of being intensified and calculate their frequency using the procedure mentioned in Section 3.1.



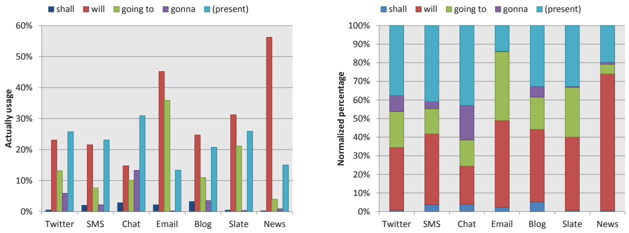
(a) Intensifier Usage **(b)** Normalized Usage

Figure 3: *Usage of Intensifiers*

The results are shown in Figure 3. There are two interesting observations: (1) At an aggregate level, mediums like SMS and online chat show a much higher percentage of intensifier use than other mediums such as email, blogs, Slate and News. A manual inspection of the content of SMS and online chat reveals that intensifiers are extensively used in order to emphasize a particular word in a running conversation, and their repeated usage is associated with a sparseness of vocabulary. This certainly explains the extreme divergence between traditional media, on the one hand, and conversation-oriented mediums on the other. (2) From Figure 3(b) it is clear that "*very*" is remarkably popular in email, blogs, slate and News.

In contrast, “*really*”, the newer variant of *very*, shows its dominance in mediums like Twitter, SMS and online chat whose average user age is lower. This conforms to previous findings that *very* is only frequently used by older individuals, and *really* is increasingly popular among the younger generation ((Ito and Tagliamonte 2003)).

Usage of Temporal References Temporal references can be used to discern the relation of text to the underlying event that precipitated it, with respect to the time axis. Figure 4 displays data on the usage of four temporal references–*going to*, *gonna*, *will*, *shall*, and one additional periphrastic present to denote references to the present. Note that the two future variants of *go*, *going to* and *gonna*, make up the bulk of the remainder, while the usage of the simple and periphrastic present is high; *shall* is virtually nonexistent. As with intensifiers, these relative frequencies parallel the findings of earlier research.



(a) Temporal Reference Usage **(b)** Normalized Usage

Figure 4: *Usage of Temporal References*

As far as differences between Twitter and other mediums are concerned, it can be clearly seen that while blogs, Slate, and particularly news articles indulge in heavy usage of future references (in specific, *will*), Twitter tends to have slightly more references to the present (39.1% vs. 33.2% for *will* and *present* in Figure 4(b)). This is entirely reasonable when one recollects that most of the content on Twitter is either related to breaking news and events as they are happening, or updating self status and having conversation with others in real-time. Similar behavior is manifested in SMS, and magnified in the case of online chat – over 30% usage as shown in Figure 4(a) and 42% in the normalized view in Figure 4(b), which are among the highest usage across all the corpora. The references to the present outnumber those to the future. Besides, the more vernacular variant *gonna* is significantly more frequent in online chat (20.3%) than Twitter (9.7%). This indicates that when people use Twitter, they tend more toward the standard variants than when they are talking to others. Twitter is, once again, more conservative than online chat.

4.2 Psycholinguistic Aspects

Our last experiment examines the cognitive and affective aspects of the language on various mediums in order to answer RQ2. To conduct this experiment, we apply our model SoCLin , as described in Section 3.2. The results are shown in Table 3. There are some interesting observations: first, we find that there is much more positive affect than negative on Twitter, indicating perhaps that people generally tweet about happier things. This conforms to previous research on sentiment analysis of tweets (Pak and Paroubek 2010). We also see that Twitter displays a much smaller variation in affect when compared to the other corpora (except SMS). This may seem counter-intuitive initially, but it must be kept in mind that out of all the mediums considered, Twitter affords users the most choice of what kind of affect they want to share. Other mediums like email and particularly news and magazines must necessarily carry content that relates to all kinds of aspects indiscriminately; email because it must carry responses to different subjects, and news media because they must report all kinds of stories. It is true that SMS and chat also afford users the choice on what to share; however, these mediums are much more private and involve one-to-one communication, whereas Twitter is a public broadcast medium. We believe that this leads to an overwhelming majority of Twitter users controlling (inadvertently or otherwise) the kind of content that they post, and exhibiting a disposition to lean towards positive affect, as shown in the first row of Table 3.

When it comes to cognitive aspects, the situation seems to be reversed – longer media, such as email, blogs and news display a much higher percentage of words from established cognitive classes like *causation*, *certainty*, *discrepancy* and *tentativeness*. At the other end of the spectrum, chat has a very low percentage of words from such categories. Twitter and SMS, whose numbers are not as low, still show much lesser usage of words from cognitive categories. Even within Twitter and SMS, the former tends to show a higher usage of words from the *certainty* category (e.g.: always, never), while the latter shows more *tentativeness* based words (e.g.: maybe, perhaps, guess). This seems to indicate two things: (1) that the language of Twitter is less about generating rationales from scratch; and (2) that Twitter, being less conversational than SMS, contains stronger opinions (more words from *certainty*) and fewer *tentativeness* words, which denote a back-and-forth discourse.

Affective Aspects							
C	**Tw.**	**SMS**	**Chat**	**Email**	**Blog**	**Sl.**	**News**
pose	0.48	0.57	0.22	0.68	0.34	0.39	0.26
nege	0.16	0.11	0.05	0.65	0.16	0.19	0.17
anx	0.01	0.02	0.01	0.12	0.09	0.13	0.05
anger	0.07	0.02	0.06	0.16	0.13	0.31	0.15
sad	0.03	0.03	0.01	0.16	0.14	0.17	0.21
Cognitive Aspects							
C	**Tw.**	**SMS**	**Chat**	**Email**	**Blog**	**Sl.**	**News**
insight	0.13	0.12	0.04	0.62	0.51	0.48	0.47
cause	0.12	0.11	0.04	0.37	0.60	0.57	0.85
discrep	0.21	0.18	0.07	0.89	0.69	0.74	0.71
tentat	0.04	0.21	0.07	0.81	0.77	0.76	0.92
certain	0.13	0.08	0.03	0.41	0.62	0.78	0.48
inhib	0.05	0.04	0.01	0.47	0.47	0.14	0.27
incl	0.32	0.22	0.07	0.68	0.61	0.98	0.62

Table 3: *Affective and cognitive aspects of various corpora. Statistics about the studied data: Tw. stands for Twitter, Sl. is Slate; pose: Positive emotion, nege: negative emotion, anx: Anxiety, anger: Anger, sad: Sadness, insight: Insight, cause: Causation, discrep: Discrepancy, tentat: Tentative, certain: Centainty, inhib: Inhibition, incl: Inclusive.*

5 Discussion

We now summarize the central findings of this work and re-visit the two core research questions posed in Section 4. Table 4 provides a quick summary of all the results related to the linguistic style and psycholinguistic aspects of Twitter's language. These findings confirm the conjectures that we make in Section 3.

Analysis	**Results**
WF	Similar to email and blog language; more sophisticated than SMS and chat.
LD	Close to blogs and news; tweets are used primarily to convey information, but are restricted by length.
PP	Mostly 1st and 3rd person, very distinct from other mediums; Tweets are about *self* as well as information sharing (e.g., breaking news). Much less conversational than SMS and chat.
INT	More usage of *really*, indicating a younger population of users than traditional mediums like email and news where *very* is mostly used. Higher net intensifier usage than chat.
TR	Highest number of references to the present; most content related to current events (real-time platform).
AA	Contains significantly more positive emotion than negative. Displays a much lesser variation of affect when compared to email, blogs, magazines and news.
CA	Contains less cognitive words than email and news. Contains strong opinions (more words on *certainty*) and lesser *tentativeness* than SMS and chat, meaning more information sharing than discourse.

Table 4: *Summary of Results for Twitter; WF: Word Frequency; LD: Lexical Density; PP: Personal Pronouns; INT: Intensifiers; TR: Temporal References; AA: Affective Aspects; CA: Cognitive Aspects*

Recall that the first question is about whether there is a difference between linguistic styles on Twitter and other mediums and what the reason for that contrast is. Our answer is that there is – we find that Twitter is markedly more standard and formal than SMS and online chat, closer to email and blogs, and less so than newspapers. In fact, we would argue that Twitter, as a new type of computer-mediated communication (CMC), is closer to traditional written language than it is to speech-like mediums such as SMS and online chat, although it shares their brevity and interactivity. This is in contrast to suggestions by some commentators that CMC is more biased in the direction of speech (Yates 1996). More generally, Twitter users appear to be developing linguistically unique styles when compared against other mediums. For example, both first-person and third-person pronouns are extensively used, whereas other mediums tend to stick to one type of pronoun. However, the data also give the impression that language use on Twitter is not too extreme in its uniqueness, given the prevalent use of standard grammatical constructions and lexical items; this goes against some claims in the popular media that Twitter is breaking down English.[7]

The second question pertains to the differences in language in terms of affective and cognitive aspects. Our results clearly show that Twitter's language makes more use of psycholinguistic aspects, particularly positive ones; whereas it does not seem to display the use of too many cognitive words. This seems to suggest that tweets are less about composing new ideas or content, and more about moulding opinions on such content.

It is also worth considering the possible methodological limitations of the study, and the implications that they may have on the results and any conclusions drawn. As mentioned in section 4, the datasets we used are varied in terms of the time periods that they cover. Although this is unavoidable given the available datasets, we would like to mention that more recent dataset (especially for online chat and email) may exhibit different linguistic characteristics. As various research studies show, there is an undergoing change in general English (both word usage and grammar), particularly among adolescents (Tagliamonte and Denis 2008). More and more words are being invented and some, which originally only existed in one medium, are now being used in another (e.g., Twitter's hashtags are now found in emails and other media[8]).

6 Conclusion and Future Work

In this paper, we proposed a two-part computational framework to offer insights into linguistic styles on Twitter, and other popular mediums. This framework was applied to various corpora from several major mediums to gather statistics, and compare and classify the linguistic styles of Twitter's language versus those other media. We concluded that the language of Twitter is highly dynamic, and that depending on the measure that is used, it shows similarities to different media. We believe that this proves – more than anything else – the fact that Twitter is a rich, evolving medium whose language is a projection of the language of more formal media like news and blogs into a space restricted by size, leading to adaptations that endow Twitter with characteristics that are similar to short media like SMS and chat as well.

Many useful extensions can be made starting out from the framework that we have proposed in this paper. Linguistic aspects that we have not considered in this work may be looked at in order to better understand and classify the language of Twitter. The number of datasets in use, as well as their variety, is another key factor that can be enhanced in order to obtain further results. In particular, we are in the process of obtaining data from the early days of Twitter in order to look at the evolution in language usage (if any) over time. Finally, other promising directions that we have started working towards are comparing linguistic aspects *within* Twitter data, and determining whether events, which drive information on Twitter; and social networks among Twitter users (Tang, Gao, and Liu 2012), which influence information flow; have an impact on the usage of language.

Acknowledgements This research is supported in part by ONR grants N000140910032 and N00014-13-1-0176, NSF grant IIS201330813 and a Google Research Award.

[7] http://www.telegraph.co.uk/culture/film/8853427/Ralph-Fiennes-blames-Twitter-for-eroding-language.html

[8] http://www.nytimes.com/2011/06/12/fashion/hashtags-a-new-way-for-tweets-cultural-studies.html

References

Bäcklund, U. 1973. The collocation of adverbs of degree in english.

Baron, N. S. 1998. Letters by phone or speech by other means: The linguistics of email. *Language and communication* 18:133–170.

Biber, D. 1991. *Variation across speech and writing*. Cambridge University Press.

Breland, H. M. 1996. Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science* 96–99.

Carroll, D. W. 2008. *Psychology of language*. Wadsworth Publishing Company.

Crystal, D. 2001. *Language and the Internet*. Cambridge University Press.

Danescu-Niculescu-Mizil, C.; Gamon, M.; and Dumais, S. 2011. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, 745–754. ACM.

Davies, M. 2010. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing* 25(4):447–464.

Ding, C. H. Q.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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

Flower, L., and Hayes, J. R. 1981. A cognitive process theory of writing. *College composition and communication* 32(4):365–387.

Golder, S. A., and Macy, M. W. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–1881.

Halliday, M. A., and Matthiessen, C. M. 2004. An introduction to functional grammar.

Herring, S. C.; Scheidt, L. A.; Bonus, S.; and Wright, E. 2004. Bridging the gap: A genre analysis of weblogs. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, 11–pp. IEEE.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50–57. ACM.

Hong, L.; Convertino, G.; and Chi, E. H. 2011. Language matters in twitter: A large scale study. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, volume 91, 518–521.

Hu, Y.; John, A.; Seligmann, D. D.; and Wang, F. 2012. What were the tweets about? topical associations between public events and twitter feeds. *Proc. ICWSM*.

Hu, X.; Tang, J.; Gao, H.; and Liu, H. 2013. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, WWW'13. ACM.

Ito, R., and Tagliamonte, S. 2003. Well weird, right dodgy, very strange, really cool: Layering and recycling in english intensifiers. *Language in Society* 32(2):257–279.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 56–65. ACM.

Lakoff, R. T. 1979. Stylistic strategies within a grammar of style. *Annals of the New York Academy of Sciences* 327(1):53–78.

Li, T.; Zhang, Y.; and Sindhwani, V. 2009. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of ACL*, 244–252. Association for Computational Linguistics.

Magnifico, A. M. 2010. Writing for whom? cognition, motivation, and a writer's audience. *Educational Psychologist* 45(3):167–184.

Naaman, M.; Boase, J.; and Lai, C.-H. 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM CSCW*, 189–192. ACM.

Nocedal, J., and Wright, S. J. 2000. *Numerical Optimization*. Springer.

Owoputi, O.; O'onnor, B.; Dyer, C.; Gimpel, K.; Schneider, N.; and Smith, N. A. 2013. Improved part-of-speech tagging for online conversational text with word clusters.

Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*, volume 2010.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*.

Quirk, R.; Greenbaum, S.; Leech, G.; Svartvik, J.; and Crystal, D. 1985. *A comprehensive grammar of the English language*, volume 397. Cambridge Univ Press.

Ritter, A.; Clark, S.; Etzioni, O.; et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1524–1534. Association for Computational Linguistics.

Rosenwasser, D., and Stephen, J. 2011. *Writing analytically*. Wadsworth Publishing Company.

Smith, A., and Brenner, J. 2012. Twitter use 2012. *Pew Internet & American Life Project*.

Tagliamonte, S. A., and Denis, D. 2008. Linguistic ruin? lol! instant messaging and teen language. *American Speech* 83(1):3–34.

Tang, J.; Gao, H.; and Liu, H. 2012. mTrust: Discerning multifaceted trust in a connected world. In *the 5th ACM International Conference on Web Search and Data Mining*.

Thurlow, C., and Brown, A. 2003. Generation txt? the sociolinguistics of young peoples text-messaging. *Discourse analysis online* 1(1):30.

Wardhaugh, R. 2011. *An introduction to sociolinguistics*, volume 28. Wiley-Blackwell.

Yates, S. J. 1996. Oral and written linguistic aspects of computer conferencing. *PRAGMATICS AND BEYOND NEW SERIES* 29–46.