Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media

Svitlana Volkova
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD
svitlana@jhu.edu

Theresa Wilson
Human Language Technology Center of Excellence
Johns Hopkins University
Baltimore, MD
taw@jhu.edu

David Yarowsky
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD
yarowsky@cs.jhu.edu

Abstract

Different demographics, e.g., gender or age, can demonstrate substantial variation in their language use, particularly in informal contexts such as social media. In this paper we focus on learning gender differences in the use of subjective language in English, Spanish, and Russian Twitter data, and explore cross-cultural differences in emoticon and hashtag use for male and female users. We show that gender differences in subjective language can effectively be used to improve sentiment analysis, and in particular, polarity classification for Spanish and Russian. Our results show statistically significant relative F-measure improvement over the gender-independent baseline 1.5% and 1% for Russian, 2% and 0.5% for Spanish, and 2.5% and 5% for English for polarity and subjectivity classification.

1 Introduction

Sociolinguistics and dialectology have been studying the relationships between language and speech at the phonological, lexical and morphosyntactic levels and social identity for decades (Picard, 1997; Gefen and Ridings, 2005; Holmes and Meyerhoff, 2004; Macaulay, 2006; Tagliamonte, 2006). Recent studies have focused on exploring demographic language variations in personal email communication, blog posts, and public discussions (Boneva et al., 2001; Mohammad and Yang, 2011; Eisenstein et al., 2010; O'Connor et al., 2010; Bamman et al., 2012). However, one area that remains largely unexplored is the effect of demographic language variation on subjective language use, and whether these differences may be exploited for automatic sentiment analysis. With the growing commercial importance of applications such as personalized recommender systems and targeted advertising (Fan and Chang, 2009), detecting helpful product review (Ott et al., 2011), tracking sentiment in real time (Resnik, 2013), and large-scale, low-cost, passive polling (O'Connor et al., 2010), we believe that sentiment analysis guided by user demographics is a very important direction for research.

In this paper, we focus on gender demographics and language in social media to investigate differences in the language used to express opinions in Twitter for three languages: English, Spanish, and Russian. We focus on Twitter data because of its volume, dynamic nature, and diverse population worldwide.[1] We find that some words are more or less likely to be positive or negative in context depending on the the gender of the author. For example, the word *weakness* is more likely to be used in a positive way by women (*Chocolate is my weakness!*) but in a negative way by men (*Clearly they know our weakness. Arggggg*). The Russian word достичь (achieve) is used in a positive way by male users and in a negative way by female users.

Our goals of this work are to (1) explore the gender bias in the use of subjective language in social media, and (2) incorporate this bias into models to improve sentiment analysis for English, Spanish, and Russian. Specifically, in this paper we:

- investigate multilingual lexical variations in the use of subjective language, and cross-cultural

[1] As of May 2013, Twitter has 500m users (140m of them in the US) from more than 100 countries.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1815–1827,
Seattle, Washington, USA, 18-21 October 2013.

emoticon and hashtag usage on a large scale in Twitter data;[2]

- show that gender bias in the use of subjective language can be used to improve sentiment analysis for multiple languages in Twitter.
- demonstrate that simple, binary features representing author gender are insufficient; rather, it is the combination of lexical features, together with set-count features representing gender-dependent sentiment terms that is needed for statistically significant improvements.

To the best of our knowledge, this work is the first to show that incorporating gender leads to significant improvements for sentiment analysis, particularly subjectivity and polarity classification, for multiple languages in social media.

2 Related Work

Numerous studies since the early 1970's have investigated gender-language differences in interaction, theme, and grammar among other topics (Schiffman, 2002; Sunderland et al., 2002). More recent research has studied gender differences in telephone speech (Cieri et al., 2004; Godfrey et al., 1992) and emails (Styler, 2011). Mohammad and Yang (2011) analyzed gender differences in the expression of sentiment in love letters, hate mail, and suicide notes, and emotional word usage across genders in email.

There has also been a considerable amount of work in subjectivity and sentiment analysis over the past decade, including, more recently, in microblogs (Barbosa and Feng, 2010; Bermingham and Smeaton, 2010; Pak and Paroubek, 2010; Bifet and Frank, 2010; Davidov et al., 2010; Li et al., 2010; Kouloumpis et al., 2011; Jiang et al., 2011; Agarwal et al., 2011; Wang et al., 2011; Calais Guerra et al., 2011; Tan et al., 2011; Chen et al., 2012; Li et al., 2012). In spite of the surge of research in both sentiment and social media, only a limited amount of work focusing on gender identification has looked at differences in subjective language across genders within social media. Thelwall (2010) found that men and women use emoticons to differing degrees on MySpace, e.g., female users express positive emoticons more than male users. Other researchers included subjective patterns as features for gender classification of Twitter users (Rao et al., 2010). They found that the majority of emotion-bearing features, e.g., emoticons, repeated letters, exasperation, are used more by female than male users, which is consistent with results reported in other recent work (Garera and Yarowsky, 2009; Burger et al., 2011; Goswami et al., 2009; Argamon et al., 2007). Other related work is that of Otterbacher (2010), who studied stylistic differences between male and female reviewers writing product reviews, and Mukherjee and Liu (2010), who applied positive, negative and emotional connotation features for gender classification in microblogs.

Although previous work has investigated gender differences in the use of subjective language, and features of sentiment have been used in gender identification, to the best of our knowledge no one has yet investigated whether gender differences in the use of subjective language can be exploited to improve sentiment classification in English or any other language. In this paper we seek to answer this question for the domain of social media.

3 Data

For the experiments in this paper, we use three sets of data for each language: a large pool of data (800K tweets) labeled for gender but *unlabeled* for sentiment, plus 2K development data and 2K test data labeled for both sentiment and gender. We use the unlabeled data to bootstrap Twitter-specific lexicons and investigate gender differences in the use of subjective language. We use the development data for parameter tuning while bootstrapping, and the test data for sentiment classification.

For English, we download tweets from the corpus created by Burger et al. (2011). This dataset contains 2,958,103 tweets from 184K users, excluding retweets. Retweets are omitted because our focus is on the sentiment of the person tweeting; in retweets, the words originate from a different user. All users in this corpus have gender labels, which Burger et al. automatically extracted from self-reported gender on Facebook or MySpace profiles linked to by the Twitter users. English tweets are identified using a compression-based language identification (LID)

[2]Gender-dependent and independent lexical resources of subjective terms in Twitter for Russian, Spanish and English can be found here: `http://www.cs.jhu.edu/~svitlana/`

tool (Bergsma et al., 2012). According to LID, there are 1,881,620 (63.6%) English tweets from which we select a random, gender-balanced sample of 0.8M tweets. Burger's corpus does not include Russian and Spanish data on the same scale as English. Therefore, for Russian and Spanish we construct a new Twitter corpus by downloading tweets from followers of region-specific news and media Twitter feeds. We use LID to identify Russian and Spanish tweets, and remove retweets as before. In this data, gender is labeled automatically based on user first and last name morphology with a precision above 0.98 for all languages.

Sentiment labels for tweets in the development and test sets are obtained using Amazon Mechanical Turk. For each tweet we collect annotations from five workers and use majority vote to determine the final label for the tweet. Snow et al. (2008) show that for a similar task, labeling emotion and valence, on average four non-expert labelers are needed to achieve an expert level of annotation. Below are the example Russian tweets labeled for sentiment:

- **Pos**: Как же приятно просто лечь в постель после тяжелого дня... (It is a great pleasure to go to bed after a long day at work...)
- **Neg**: Уважаемый господин Прохоров купите эти выборы! (Dear Mr. Prokhorov just buy the elections!)
- **Both**: Затолкали меня на местном рынке! но зато закупилась подарками для всей семьи :) (It was crowded at the local market! But I got presents for my family:-))
- **Neutral**: Киев очень старый город (Kiev is a very old city).

Table 1 gives the distribution of tweets over sentiment and gender labels for the development and test sets for English (EDEV, ETEST), Spanish (SDEV, STEST), and Russian (RDEV, RTEST).

Data	Pos	Neg	Both	Neut	♀	♂
EDEV	617	357	202	824	1,176	824
ETEST	596	347	195	862	1,194	806
SDEV	358	354	86	1,202	768	1,232
STEST	317	387	93	1203	700	1,300
RDEV	452	463	156	929	1,016	984
RTEST	488	380	149	983	910	1,090

Table 1: Gender and sentiment label distribution in the development and test sets for all languages.

4 Subjective Language and Gender

To study the intersection of subjective language and gender in social media, ideally we would have a large corpus labeled for both. Although our large corpus is labeled for gender, it is not labeled for sentiment. Only the 4K tweets for each language that compose the development and test sets are labeled for both gender and sentiment. Obtaining sentiment labels for all tweets would be both impractical and expensive. Instead we use large multilingual sentiment lexicons developed specifically for Twitter as described below. Using these lexicons we can begin to explore the relationship between subjective language and gender in the large pool of data labeled for gender but unlabeled for sentiment. We also look at the relationship between gender and the use of different hashtags and emoticons. These can be strong indicators of sentiment in social media, and in fact are sometimes used to create noisy training data for sentiment analysis in Twitter (Pak and Paroubek, 2010; Kouloumpis et al., 2011).

4.1 Bootstrapping Subjectivity Lexicons

Recent work by Banea et.al (2012) classifies methods for bootstrapping subjectivity lexicons into two types: corpus-based and dictionary-based. Corpus-based methods extract subjectivity lexicons from unlabeled data using different similarity metrics to measure the relatedness between words, e.g., Pointwise Mutual Information (PMI). Corpus-based methods have been used to bootstrap lexicons for ENGLISH (Turney, 2002) and other languages, including ROMANIAN (Banea et al., 2008) and JAPANESE (Kaji and Kitsuregawa, 2007).

Dictionary-based methods rely on relations between words in existing lexical resources. For example, Rao and Ravichandran (2009) construct HINDI and FRENCH sentiment lexicons using relations in WordNet (Miller, 1995), Rosas et. al. (2012) bootstrap a SPANISH lexicon using SentiWordNet (Baccianella et al., 2010) and OpinionFinder,[3] Clematide and Klenner (2010), Chetviorkin et al. (2012) and Abdul-Mageed et. al. (2011) automatically expand and evaluate GERMAN, RUSSIAN and ARABIC subjective lexicons.

[3] www.cs.pitt.edu/mpqa/opinionfinder

We use the corpus-based, language-independent approach proposed by Volkova et al. (2013) to bootstrap Twitter-specific subjectivity lexicons. To start, the new lexicon is seeded with terms from the initial lexicon L_I. On each iteration, tweets in the unlabeled data are labeled using the current lexicon. If a tweet contains one or more terms from the lexicon it is marked subjective, otherwise neutral. Tweet polarity is determined in a similar way, but takes into account negation. For every term not in the lexicon with a frequency threshold, the probability of that word appearing in a subjective sentence is calculated. The top k terms with a subjective probability are then added to the lexicon. Bootstrapping continues until there are no more new terms meeting the criteria to add to the lexicon. The parameters are optimized using a grid search on the development data using F-measure for subjectivity classification. In Table 2 we report size and term polarity from the initial L_I and the bootstrapped L_B lexicons. Although more sophisticated bootstrapping methods exist, this approach has been shown to be effective for atomically learning subjectivity lexicons in multiple languages on a large scale without any external, rich, lexical resources, e.g., WordNet, or advanced NLP tools, e.g., syntactic parsers (Wiebe, 2000) or information extraction tools (Riloff and Wiebe, 2003).

For English, seed terms for bootstrapping are the strongly subjective terms in the MPQA lexicon (Wilson et al., 2005). For Spanish and Russian, the seed terms are obtained by translating the English seed terms using a bi-lingual dictionary, collecting subjectivity judgments from MTurk on the translations, filtering out translations that are not strongly subjective, and expanding the resulting word lists with plurals and inflectional forms.

To verify that bootstrapping does provide a better resource than existing dictionary-expanded lexicons, we compare our Twitter-specific lexicons L_B to the corresponding initial lexicons L_I and the existing state-of-the-art subjective lexicons including:

- 8K strongly subjective English terms from SentiWordNet χ^E (Baccianella et al., 2010);
- 1.5K full strength terms from the Spanish sentiment lexicon χ^S (Perez-Rosas et al., 2012);
- 5K terms from the Russian sentiment lexicon χ^R (Chetviorkin and Loukachevitch, 2012).

For that we apply rule-based subjectivity classification on the test data.[4] This subjectivity classifier predicts that a tweet is subjective if it contains at least one, or at least two subjective terms from the lexicon. To make a fair comparison, we automatically expand χ^E with plurals and inflectional forms, χ^S with the inflectional forms for verbs, and χ^R with the inflectional forms for adverbs, adjectives and verbs. We report precision, recall and F-measure results in Table 3 and show that our bootstrapped lexicons outperform the corresponding initial lexicons and the external resources.

	English		**Spanish**		**Russian**	
	L_I^E	L_B^E	L_I^S	L_B^S	L_I^R	L_B^R
Pos	2.3	16.8	2.9	7.7	1.4	5.3
Neg	2.8	4.7	5.2	14.6	2.3	5.5
Total	**5.1**	**21.5**	**8.1**	**22.3**	**3.7**	**10.8**

Table 2: The initial L_I and the bootstrapped L_B (highlighted) lexicon term count ($L_I \subset L_B$) with polarity across languages (thousands).

	$Subj \geq 1$			$Subj \geq 2$		
	P	R	F	P	R	F
χ^E	0.67	0.49	0.57	0.76	0.16	0.27
L_I^E	**0.69**	0.73	0.71	**0.79**	0.34	0.48
L_B^E	0.64	**0.91**	**0.75**	0.7	**0.74**	**0.72**
χ^S	**0.52**	0.39	0.45	**0.62**	0.07	0.13
L_I^S	0.50	0.73	**0.59**	0.59	0.36	0.45
L_B^S	0.44	**0.91**	**0.59**	0.51	**0.71**	**0.59**
χ^R	0.61	0.49	0.55	0.74	0.17	0.29
L_I^R	**0.72**	0.34	0.46	**0.83**	0.07	0.13
L_B^R	0.64	**0.58**	**0.61**	0.74	**0.23**	**0.35**

Table 3: Precision, recall and F-measure results for subjectivity classification using the external χ, initial L_I and bootstrapped L_B lexicons for all languages.

4.2 Lexical Evaluation

With our Twitter-specific sentiment lexicons, we can now investigate how the *subjective use* of these terms differs depending on gender for our three languages. Figure 1 illustrates what we expect to find. $\{F\}$ and $\{M\}$ are the sets of subjective terms used by females and males, respectively. We expect that some terms will be used by males, but never by females, and vice-versa. The vast majority, however, will be used by both genders. Within this set of shared terms, many words will show little difference

[4] A similar rule-based approach using terms from the MPQA lexicon is suggested by (Riloff and Wiebe, 2003).

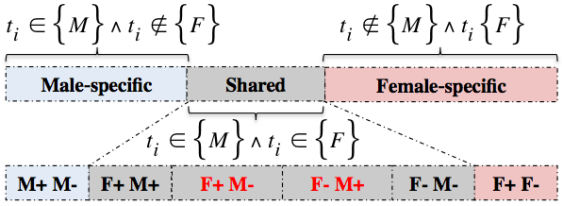


Figure 1: Gender-dependent vs. independent subjectivity terms (+ and - indicates term polarity).

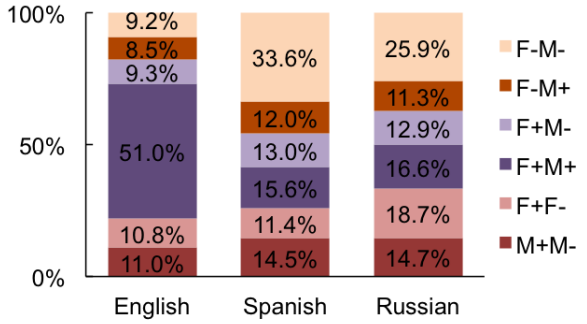


Figure 2: The distribution of gender-dependent $GDep$ and gender-independent $GInd$ sentiment terms.

in their subjective use when considering gender, but there will be some words for which gender will have an influence. Of particular interest for our work are words in which the polarity of a term as it is used in context is gender-influenced, the extreme case being terms that flip their polarity depending on the gender of the user. Polarity may be different because the concept represented by the term tends to be viewed in a different light depending on gender. There are also words like *weakness* in which a more positive or more negative word sense tends to be used by men or women. In Figure 2 we show the distribution of gender-specific and gender-independent terms from the L_B lexicons for all languages.

To identify gender-influenced terms in our lexicons, we start by randomly sampling 400K male and 400K female tweets for each language from the data. Next, for both genders we calculate the probability of term t_i appearing in a tweet with another subjective term (Eq.1), and the probability of it appearing with a positive or negative term (Eq.2-3) from L_B.

$$p_{t_i}(subj|g) = \frac{c(t_i, P, g) + c(t_i, N, g)}{c(t_i, g)}, \quad (1)$$

where $g \in F, M$ and P and N are positive and negative sets of terms from the *initial* lexicon L_I.

$$p_{t_i}(+|g) = \frac{c(t_i, P, g)}{c(t_i, P, g) + c(t_i, N, g)} \quad (2)$$

$$p_{t_i}(-|g) = \frac{c(t_i, N, g)}{c(t_i, P, g) + c(t_i, N, g)} \quad (3)$$

We introduce a novel metric $\Delta p^+_{t_i}$ to measure polarity change across genders. For every subjective term t_i we want to maximize the difference[5]:

$$\Delta p^+_{t_i} = |p_{t_i}(+|F) - p_{t_i}(+|M)|$$
$$s.t.$$
$$\left|1 - \frac{tf^{subj}_{t_i}(F)}{tf^{subj}_{t_i}(M)}\right| \leq \lambda, \; tf^{subj}_{t_i}(M) \neq 0, \quad (4)$$

where $p(+|F)$ and $p(+|M)$ are probabilities that term t_i is positive for females and males respectively; $tf^{subj}_{t_i}(F)$ and $tf^{subj}_{t_i}(M)$ are corresponding term frequencies (if $tf^{subj}_{t_i}(F) > tf^{subj}_{t_i}(M)$ the fraction is flipped); λ is a threshold that controls the level of term frequency similarity[6]. The terms in which polarity is most strongly gender-influenced are those with $\lambda \to 0$ and $\Delta p^+_{t_i} \to 1$.

Table 4 shows a sample of the most strongly gender-influenced terms from the initial L_I and the bootstrapped L_B lexicons for all languages. A plus (+) means that the term tends to be used positively by women and minus (−) means that the term tends to be used positively by men. For instance, in English we found that *perfecting* is used with negative polarity by male users but with positive polarity by female users; the term *dogfighting* has negative polarity for women but positive polarity for men.

4.3 Hashtags

People may also express positive or negative sentiment in their tweets using hashtags. From our balanced samples of 800K tweets for each language, we extracted 611, 879, and 71 unique hashtags for English, Spanish, and Russian, respectively. As we did for terms in the previous section, we evaluated the subjective use of the hashtags. Some of these are clearly expressing sentiment (*#horror*), while others seem to be topics that people are frequently opinionated about (*#baseball, #latingrammy, #spartak*).

[5] One can also maximize $\Delta p^-_{t_i} = |p_{t_i}(-|F) - p_{t_i}(-|M)|$.

[6] $\lambda = 0$ means term frequencies are identical for both genders; $\lambda \to 1$ indicates increasing gender divergence.

English Initial Terms L_I^E	Δp^+	λ	**English Bootstrapped Terms** L_B^E	Δp^+	λ
perfecting	+ 0.7	0.2	pleaseeeeee	+ 0.7	0.0
weakened	+ 0.1	0.0	adorably	+ 0.6	0.4
saddened	– 0.1	0.0	creatively	– 0.6	0.5
misbehaving	– 0.4	0.0	dogfighting	– 0.7	0.5
glorifying	– 0.7	0.5	overdressed	– 1.0	0.3
Spanish Initial Terms L_I^S			**Spanish Bootstrapped Terms** L_B^S		
fiasco (fiasco)	+ 0.7	0.3	cafeína (caffeine)	+ 0.7	0.5
triunfar (succeed)	+ 0.7	0.0	claro (clear)	+ 0.7	0.3
inconsciente (unconscious)	– 0.6	0.2	cancio (dog)	– 0.3	0.3
horroriza (horrifies)	– 0.7	0.3	llevara (take)	– 0.8	0.3
groseramente (rudely)	– 0.7	0.3	recomendarlo (recommend)	– 1.0	0.0
Russian Initial Terms L_I^R			**Russian Bootstrapped Terms** L_B^R		
магическая (magical)	+ 0.7	0.3	мечтайте (dream!)	+ 0.7	0.3
сенсационный (sensational)	+ 0.7	0.3	танцуете (dancing)	+ 0.7	0.3
обожаемый (adorable)	– 0.7	0.0	сложны (complicated)	– 1.0	0.0
искушение (temptation)	– 0.7	0.3	молоденькие (young)	– 1.0	0.0
заслуживать (deserve)	– 1.0	0.0	достичь (achieve)	– 1.0	0.0

Table 4: Sample of subjective terms sorted by Δp^+ to show lexical differences and polarity change across genders (module is not applied as defined in Eq.1 to demonstrate the polarity change direction).

English	Δp^+	λ	**Spanish**	Δp^+	λ	**Russian**	Δp^+	λ
#parenting	+ 0.7	0.0	#rafaelnarro (politician)	+ 1.0	0.0	#совет (advise)	+ 1.0	0.0
#vegas	– 0.2	0.8	#amores (loves)	+ 0.2	1.0	#ukrlaw	+ 1.0	1.0
#horror	– 0.6	0.7	#britneyspears	+ 0.1	0.3	#spartak (soccer team)	– 0.7	0.9
#baseball	– 0.6	0.9	#latingrammy	– 0.5	0.1	#сны (dreams)	– 1.0	0.0
#wolframalpha	– 0.7	1.0	#metallica (music band)	– 0.5	0.8	#iphones	– 1.0	1.0

Table 5: Hashtag examples with opposite polarity across genders for English, Spanish, and Russian.

Table 5 gives the hashtags, correlated with subjective language, that are most strongly gender-influenced. Analogously to Δp^+ values in Table 4, a plus (+) means the hashtag is more likely to be used positively by women, and a minus (–) means the hashtag is more likely to be used positively by men. For example, in English we found that male users tend to express positive sentiment in tweets mentioning $\#baseball$, while women tend to be negative about this hashtag. The opposite is true for the hashtag $\#parenting$.

4.4 Emoticons

We investigate how emoticons are used differently by men and women in social media following the work by (Bamman et al., 2012). For that we rely on the lists of emoticons from Wikipedia[7] and present the cross-cultural and gender emoticon differences in Figure 3. The frequency of each emoticon is given on the right of each language chart, with probability of use by a male user in that language given on the x-axis. The top 8 emoticons are the same across languages and sorted by English frequency.

We found that emoticons in English data are used more overall by female users, which is consistent with previous findings in Schnoebelen's work.[8] In addition, we found that some emoticons like :-) *(smile face)* and :-o *(surprised)* are used equally by both genders, at least in Twitter. When comparing English emoticon usage to other languages, there are some similarities, but also some clear differences. In Spanish data, several emoticons are more likely to be used by male than by female users, e.g., :-o *(surprised)* and :-& *(tongue-tied)*, and the difference in probability of use by males and females is greater for the emoticons, as compared to the same emoticons for English. Interestingly, in Russian Twitter

[7]List of emoticons from Wikipedia http://en.wikipedia.org/wiki/List_of_emoticons

[8]Language and emotion (talks, essays and reading notes) www.stanford.edu/~tylers/emotions.shtml

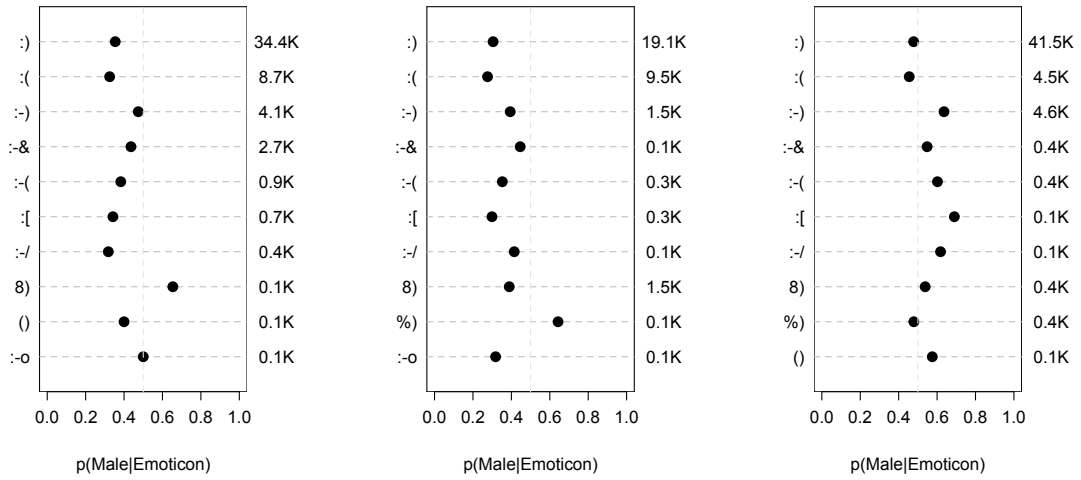


Figure 3: Probability of gender and emoticons for English, Spanish and Russian (from left to right).

data emoticons tend to be used more or equally by male users rather than female users.

5 Experiments

The previous section showed that there are gender differences in the use of subjective language, hashtags, and emoticons in Twitter. We aim leverage these differences to improve subjectivity and polarity classification for the informal, creative and dynamically changing multilingual Twitter data.[9] For that we conduct experiments using gender-independent $GInd$ and gender-dependent $GDep$ features and compare the results to evaluate the influence of gender on sentiment classification.

We experiment with two classification approaches: (I) rule-based classifier which uses only subjective terms from the lexicons designed to verify if the gender differences in subjective language create enough of a signal to influence sentiment classification; (II) state-of-the-art supervised models which rely on lexical features as well as lexicon set-count features.[10,11] Moreover, to show that the gender-sentiment signal can be learned by more than one classifier we apply a variety of classifiers implemented in Weka (Hall et al., 2009). For that we do 10-fold cross validation over English, Spanish, and Russian test data (ETEST, STEST and RTEST) labeled with subjectivity (pos, neg, both vs. neut) and polarity (pos vs. neg) as described in Section 3.

5.1 Models

For the rule-based $GInd^{RB}_{subj}$ classifier, tweets are labeled as subjective or neutral as follows:

$$GInd^{RB}_{subj} = \begin{cases} 1 & \text{if } \vec{w} \cdot \vec{f} \geq 0.5, \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where $\vec{w} \cdot \vec{f}$ stands for weighted set features, e.g., terms from L_I only, emoticons E, or different part-of-speech tags (POS) from L_B weighted using $w = p(subj) = p(subj|M) + p(subj|F)$ subjectivity score as shown in Eq.1. We experiment with the POS tags to show the contribution of each POS to sentiment classification.

Similarly, for the rule-based $GInd^{RB}_{pol}$ classifier, tweets are labeled as positive or negative:

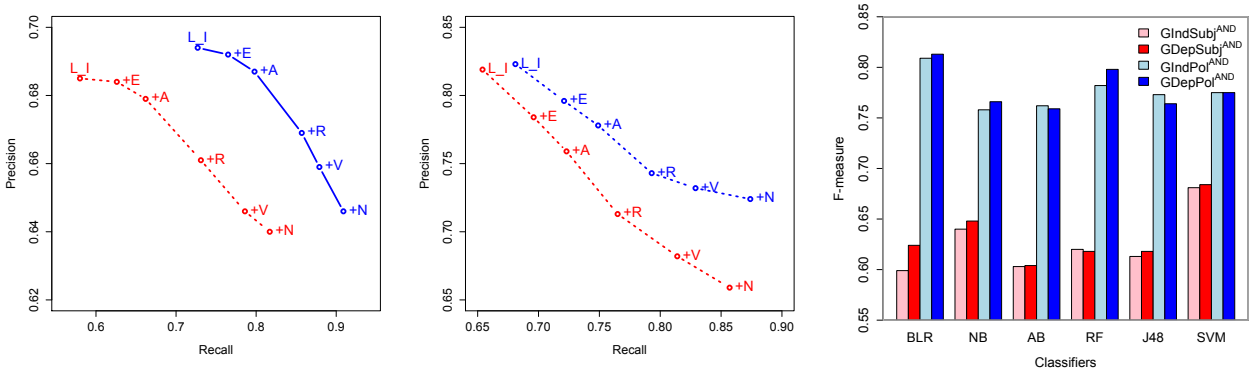
$$GInd^{RB}_{pol} = \begin{cases} 1 & \text{if } \vec{w^+} \cdot \vec{f^+} \geq \vec{w^-} \cdot \vec{f^-}, \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where $\vec{f}^+$, $\vec{f}^-$ are feature sets that include only positive and negative features from L_I or L_B; w^+ and w^-

[9] For polarity classification we distinguish between positive and negative instances, which is the approach typically reported in the literature for recognizing polarity (Velikovich et al., 2010; Yessenalina and Cardie, 2011; Taboada et al., 2011)

[10] A set-count feature is a count of the number of instances from a set of terms that appears in a tweet.

[11] We also experimented with repeated punctuation (*!!, ??*) and letters ($nooo, reealy$), which are often used in sentiment classification in social media. However, we found these features to be noisy and adding them decreased performance.



(a) Rule-based subjectivity (b) Rule-based polarity (c) SL subjectivity and polarity

Figure 4: Rule-based (RB) and Supervised Learning (SL) sentiment classification results for English. L_I - the initial lexicon, E - emoticons, A, R, V, N are adjectives, adverbs, verbs, nouns from L_B.

are positive and negative polarity scores estimated using Eq.2 - 3 such as: $w^+ = p(+|M) + p(+|F)$ and $w^- = p(-|M) + p(-|F)$.

The *gender-dependent* rule-based classifiers are defined in a similar way. Specifically, $\vec{f}$ is replaced by $\vec{f}^M$ and $\vec{f}^F$ in Eq.5 and $\vec{f}^-, \vec{f}^+$ are replaced by $\vec{f}^{M-}, \vec{f}^{F-}$ and $\vec{f}^{M+}, \vec{f}^{F+}$ respectively in Eq.6. We learn subjectivity $\vec{s}$ and polarity $\vec{p}$ score vectors using Eq.1-3. The difference between $GInd$ and $GDep$ models is that $GInd$ scores $\vec{w}$, $\vec{w^+}$ and $\vec{w^-}$ are not conditioned on gender.

For *gender-independent* classification using supervised models, we build feature vectors using lexical features V represented as term frequencies, together with set-count features from the lexicons:

$$\vec{f}^{GInd}_{subj} = [L_I, L_B, E, V];$$
$$\vec{f}^{GInd}_{pol} = [L^+_I, L^+_B, E^+, L^-_I, L^-_B, E^-, V].$$

Finally, for *gender-dependent* supervised models, we try different feature combinations. (A) We extract set-count features for gender-dependent subjective terms from L_I, L_B, and E jointly:

$$\vec{f}^{GDep-J}_{subj} = [L^M_I, L^M_B, E^M, L^F_I, L^F_B, E^F, V];$$
$$\vec{f}^{Dep-J}_{pol} = [L^{M+}_I, L^{M+}_B, E^{M+}, L^{F+}_I, L^{F+}_B, E^{F+}$$
$$L^{M-}_I, L^{M-}_B, E^{M-}, L^{F-}_I, L^{F-}_B, E^{F-}, V].$$

(B) We extract disjoint (prefixed) gender-specific features (in addition to lexical features V) by relying only on female set-count features when classifying female tweets; and only male set-count features for male tweets. We refer to the joint features as $GInd-J$ and $GDep-J$, and to the disjoint features $GInd-D$ and $GDep-D$.

5.2 Results

Figures 4a and 4b show performance improvements for subjectivity and polarity classification under the rule-based approach when taking into account gender. The left figure shows precision-recall curves for subjective *vs.* neutral classification, and the middle figure shows precision-recall curves for positive *vs.* negative classification. We measure performance starting with features from L_I, and then incrementally add emoticon features E and features from L_B one part of speech at a time to show the contribution of each part of speech for sentiment classification.[12] This experiment shows that there is a clear improvement for the models parameterized with gender, at least for the simple, rule-based model.

For the supervised models we experiment with a variety of learners for English to show that gender differences in subjective language improve sentiment classification for many learning algorithms. We present the results in Figure 4c. For subjectivity classification, Support Vector Machines (SVM), Naive Bayes (NB) and Bayesian Logistic Regression (BLR) achieve the best results, with improvements in F-measure ranging from 0.5 - 5%. The polarity classifiers overall achieve much higher scores, with improvements for $GDep$ features ranging from 1-2%. BLR with Gaussian prior is the top scorer

[12] POS from the Twitter POSTagger (Gimpel et al., 2011).

	P	R	F	A	A^{rand}	P	R	F	A	A^{rand}
	English subj vs. neutral p(subj)=0.57					**English pos vs. neg p(pos)=0.63**				
$GInd_{LR}$	0.62	0.58	0.60	0.66	–	0.78	0.83	0.80	0.71	–
$GDep-J$	0.64	0.62	0.63	0.68	*0.66*	0.80	0.83	0.82	0.73	*0.70*
$\Delta R,\%$	+3.23	+6.90	+5.00	+3.03	*3.03*↓	+2.56	0.00	+2.50	+2.82	*4.29*↓
$GInd_{SVM}$	0.66	0.70	0.68	0.72	–	0.79	0.86	0.82	0.77	–
$GDep-D$	0.66	0.71	0.68	0.72	*0.70*	0.80	0.87	0.83	0.78	*0.76*
$\Delta R,\%$	–0.45	+0.71	0.00	–0.14	*2.85*↓	+0.38	+0.23	+0.24	+0.41	*2.63*↓
	Spanish subj vs. neutral p(subj)=0.40					**Spanish pos vs. neg p(pos)=0.45**				
$GInd_{LL}$	0.67	0.71	0.68	0.61	–	0.71	0.63	0.67	0.71	–
$GDep-J$	0.67	0.72	0.69	0.62	*0.61*	0.72	0.65	0.68	0.71	*0.68*
$\Delta R,\%$	0.00	+1.40	+0.58	+0.73	*1.64*↓	+2.53	+3.17	+1.49	0.00	*4.41*↓
$GInd_{SVM}$	0.68	0.79	0.73	0.65	–	0.66	0.65	0.65	0.69	–
$GDep-D$	0.68	0.79	0.73	0.66	*0.65*	0.68	0.67	0.67	0.71	*0.68*
$\Delta R,\%$	+0.35	+0.21	+0.26	+0.54	*1.54*↓	+2.43	+2.44	+2.51	+2.08	*4.41*↓
	Russian subj vs. neutral p(subj)=0.51					**Russian pos vs. neg p(pos)=0.58**				
$GInd_{LR}$	0.66	0.68	0.67	0.67	–	0.66	0.72	0.69	0.62	–
$GDep-J$	0.66	0.69	0.68	0.67	*0.66*	0.68	0.73	0.70	0.64	*0.63*
$\Delta R,\%$	0.00	+1.47	+0.75	0.00	*1.51*↓	+3.03	+1.39	+1.45	+3.23	*1.58*↓
$GInd_{SVM}$	0.67	0.75	0.71	0.70	–	0.64	0.73	0.68	0.62	–
$GDep-D$	0.67	0.76	0.71	0.70	*0.69*	0.65	0.74	0.69	0.63	*0.62*
$\Delta R,\%$	–0.30	+1.46	+0.56	+0.14	*1.44*↓	+0.93	+1.92	+1.46	+1.49	*1.61*↓

Table 6: Sentiment classification results obtained using gender-dependent and gender-independent joint and disjoint features for Logistic Regression (LR) and SVM models.

for polarity classification with an F-measure of 82%. We test our results for statistical significance using McNemar's Chi-squared test (p-value < 0.01) as suggested by Dietterich (1998). Only three classifiers, J48, AdaBoostM1 (AB) and Random Forest (RF) do not always show significant improvements for $GDep$ features over $GInd$ features. However, for the majority of classifiers, $GDep$ models outperform $GInd$ models for both tasks, demonstrating the robustness of $GDep$ features for sentiment analysis.

In Table 6 we report results for subjectivity and polarity classification using the best performing classifiers (as shown in Figure 4c) :

- Logistic Regression (LR) (Genkin et al., 2007) for $GInd-J$ and $GDep-J$ models.
- SVM model with radial-based kernel for $GInd-D$ and $GDep-D$ models. We use LibSVM implementation (EL-Manzalawy and Honavar, 2005).

Each $\Delta R(\%)$ row shows the relative percent improvements in terms of precision P, recall R, F-measure F and accuracy A for $GDep$ compared to $GInd$ models. Our results show that differences in subjective language across genders can be exploited to improve sentiment analysis, not only for English but for multiple languages. For Spanish and Russian results are lower for subjectivity classification, we suspect, because lexical features V are already inflected for gender and set-count features are down-weighted by the classifier. For polarity classification, on the other hand, gender-dependent features provide consistent, significant improvements (1.5-2.5%) across all languages.

As a reality check, Table 6 also reports accuracies (in A^{rand} columns) for experiments that use random permutations of male and female subjective terms, which are then encoded as gender-dependent set-count features as before. We found that all gender-dependent models, $GDep-J$ and $GDep-D$, outperformed their random equivalents for both subjectivity and polarity classification (as reflected by relative accuracy decrease ↓ for A^{rand} compared to A). These results further confirm the existence of gender bias in subjective language for any of our three languages and its importance for sentiment analysis.

Finally, we check whether encoding gender as a binary feature would be sufficient to improve sentiment classification. For that we encode fea-

	English		Spanish		Russian	
	P	R	P	R	P	R
(a)	0.73	0.93	0.68	0.63	0.66	0.74
(b)	0.72	0.94	0.69	0.64	0.66	0.74
(c)	0.78	0.83	0.71	0.63	0.66	0.72
(d)	0.69	0.93	0.71	0.62	0.65	0.76
(e)	**0.80**	**0.83**	**0.72**	**0.65**	**0.68**	**0.73**

Table 7: Precision and recall results for polarity classification: encoding gender as a binary feature vs. gender-dependent features $GDep - J$.

tures such as: (a) unigram term frequencies V, (b) term frequencies and gender binary $V + GBin$, (c) gender-independent $GInd$, (d) gender-independent and gender binary $GBin + GInd$, and (e) gender-dependent $GDep - J$. We train logistic-regression model for polarity classification and report precision and recall results in Table 7. We observe that including gender as a binary feature does not yield significant improvements compared to $GDep - J$ for all three languages.

6 Conclusions

We presented a qualitative and empirical study that analyses substantial and interesting differences in subjective language between male and female users in Twitter, including hashtag and emoticon usage across cultures. We showed that incorporating author gender as a model component can significantly improve subjectivity and polarity classification for English (2.5% and 5%), Spanish (1.5% and 1%) and Russian (1.5% and 1%). In future work we plan to develop new models for joint modeling of personalized sentiment, user demographics e.g., age and user preferences e.g., political favorites in social media.

Acknowledgments

The authors thank the anonymous reviewers for helpful comments and suggestions.

References

Muhammad Abdul-Mageed, Mona T. Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, pages 587–591.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media (LSM'11)*, pages 30–38.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9). http://www.firstmonday.org/ojs/index.php/fm/article/view/2003/1878.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2012. Gender in Twitter: styles, stances, and social networks. *Computing Research Repository*.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2764–2767.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 36–44.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific Twitter collections. In *Proceedings of the Second Workshop on Language in Social Media (LSM'12)*, pages 65–74.

Adam Bermingham and Alan F. Smeaton. 2010. Classifying sentiment in microblogs: Is brevity an advantage? In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*, pages 1833–1836.

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in Twitter streaming data. In *Proceedings of the 13th International Conference on Discovery Science (DS'10)*, pages 1–15.

Bonka Boneva, Robert Kraut, and David Frohlich. 2001. Using email for personal relationships: The difference gender makes. *American Behavioral Scientist*, 45(3):530–549.

John D. Burger, John C. Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309.

Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira Jr, and Virgílio Almeida. 2011. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the17th International Conference on Knowledge Discovery and Data Mining (KDD'11)*, pages 150–158.

Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit P. Sheth. 2012. Extracting diverse sentiment expressions with target-dependent polarity from Twitter. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM'12)*, pages 50–57.

Ilia Chetviorkin and Natalia V. Loukachevitch. 2012. Extraction of Russian sentiment lexicon for product meta-domain. In *Proceedings of the 25rd International Conference on Computational Linguistics (COLING'12)*, pages 593–610.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 69–71.

Simon Clematide and Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for German. In *Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA'10)*, pages 7–13.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 241–249.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*, pages 1277–1287.

Yasser EL-Manzalawy and Vasant Honavar, 2005. *WLSVM: Integrating LibSVM into Weka Environment*. http://www.cs.iastate.edu/ yasser/wlsvm.

Teng-Kai Fan and Chia-Hui Chang. 2009. Sentiment-oriented contextual advertising. *Advances in Information Retrieval*, 5478:202–215.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 710–718.

David Gefen and Catherine M. Ridings. 2005. If you spoke as she does, sir, instead of the way you do: a sociolinguistics perspective of gender differences in virtual communities. *SIGMIS Database*, 36(2):78–92.

Alexander Genkin, David D. Lewis, and David Madigan. 2007. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49:291–304.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, pages 42–47.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)*, pages 517–520.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers age and gender. In *Proceedings of AAAI Conference on Weblogs and Social Media*, pages 214–217.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Exploratory Newsletter*, 11(1):10–18.

Janet Holmes and Miriam Meyerhoff. 2004. *The Handbook of Language and Gender*. Blackwell Publishing.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160.

Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'07)*, pages 1075–1083.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, pages 538–541.

Guangxia Li, Steven Hoi, Kuiyu Chang, and Ramesh Jain. 2010. Micro-blogging sentiment detection by collaborative online learning. In *Proceedings of IEEE 10th International Conference on Data Mining (ICDM'10)*, pages 893–898.

Hao Li, Yu Chen, Heng Ji, Smaranda Muresan, and Dequan Zheng. 2012. Combining social cognitive theories with linguistic features for multi-genre sentiment analysis. In *Proceedings of the 26th Pacific Asia*

Conference on Language,Information and Computation (PACLIC'12), pages 27–136.

Ronald Macaulay. 2006. Pure grammaticalization: The development of a teenage intensifier. *Language Variation and Change*, 18(03):267–283.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2012. Multilingual subjectivity and sentiment analysis. In *Proceedings of the Association for Computational Linguistics (ACL'12)*.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Saif Mohammad and Tony Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA'11)*, pages 70–79.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*, pages 207–217.

Brendan O'Connor, Jacob Eisenstein, Eric P. Xing, and Noah A. Smith. 2010. A mixture model of demographic lexical variation. In *Proceedings of NIPS Workshop on Machine Learning in Computational Social Science*, pages 1–7.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319.

Jahna Otterbacher. 2010. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*, pages 369–378.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1320–1326.

Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in Spanish. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 3077–3081.

Rosalind W. Picard. 1997. *Affective computing*. MIT Press.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pages 675–682.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the Workshop on Search and Mining User-generated Contents (SMUC'10)*, pages 37–44.

Philip Resnik. 2013. Getting real(-time) with live polling. http://vimeo.com/68210812.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, pages 105–112.

Harold Schiffman. 2002. Bibliography of gender and language. http://ccat.sas.upenn.edu/ haroldfs/popcult/bibliogs/gender/genbib.htm.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, pages 254–263.

Will Styler. 2011. The EnronSent Corpus. Technical report, University of Colorado at Boulder Institute of Cognitive Science. http://verbs.colorado.edu/enronsent/.

Jane Sunderland, Ren-Feng Duann, and Paul Bake. 2002. Gender and genre bibliography. www.ling.lancs.ac.uk/pubs/clsl/clsl122.pdf.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Sali A. Tagliamonte. 2006. *Analysing Sociolinguistic Variation*. Cambridge University Press, 1st. Edition.

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining (KDD'11)*, pages 1397–1405.

Mike Thelwall, David Wilkinson, and Sukhvinder Uppal. 2010. Data mining emotion in social network communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology*, 61(1):190–199.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, pages 417–424.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Proceedings of*

the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'10), pages 777–785.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual Twitter streams. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 505–510.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*, pages 1031–1040.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence (AAAI'00*, pages 735–740.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'05)*, pages 347–354.

Ainur Yessenalina and Claire Cardie. 2011. Compositional matrix-space models for sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 172–182.