Enhanced Twitter Sentiment Classification Using Contextual Information

Soroush Vosoughi
The Media Lab
MIT
Cambridge, MA 02139
soroush@mit.edu

Helen Zhou
The Media Lab
MIT
Cambridge, MA 02139
hlzhou@mit.edu

Deb Roy
The Media Lab
MIT
Cambridge, MA 02139
dkroy@media.mit.edu

Abstract

The rise in popularity and ubiquity of Twitter has made sentiment analysis of tweets an important and well-covered area of research. However, the 140 character limit imposed on tweets makes it hard to use standard linguistic methods for sentiment classification. On the other hand, what tweets lack in structure they make up with sheer volume and rich metadata. This metadata includes geolocation, temporal and author information. We hypothesize that sentiment is dependent on all these contextual factors. Different locations, times and authors have different emotional valences. In this paper, we explored this hypothesis by utilizing distant supervision to collect millions of labelled tweets from different locations, times and authors. We used this data to analyse the variation of tweet sentiments across different authors, times and locations. Once we explored and understood the relationship between these variables and sentiment, we used a Bayesian approach to combine these variables with more standard linguistic features such as n-grams to create a Twitter sentiment classifier. This combined classifier outperforms the purely linguistic classifier, showing that integrating the rich contextual information available on Twitter into sentiment classification is a promising direction of research.

1 Introduction

Twitter is a micro-blogging platform and a social network where users can publish and exchange short messages of up to 140 characters long (also known as tweets). Twitter has seen a great rise in popularity in recent years because of its availability and ease-of-use. This rise in popularity and the public nature of Twitter (less than 10% of Twitter accounts are private (Moore, 2009)) have made it an important tool for studying the behaviour and attitude of people.

One area of research that has attracted great attention in the last few years is that of tweet sentiment classification. Through sentiment classification and analysis, one can get a picture of people's attitudes about particular topics on Twitter. This can be used for measuring people's attitudes towards brands, political candidates, and social issues.

There have been several works that do sentiment classification on Twitter using standard sentiment classification techniques, with variations of n-gram and bag of words being the most common. There have been attempts at using more advanced syntactic features as is done in sentiment classification for other domains (Read, 2005; Nakagawa et al., 2010), however the 140 character limit imposed on tweets makes this hard to do as each article in the Twitter training set consists of sentences of no more than several words, many of them with irregular form (Saif et al., 2012).

On the other hand, what tweets lack in structure they make up with sheer volume and rich metadata. This metadata includes geolocation, temporal and author information. We hypothesize that sentiment is dependent on all these contextual factors. Different locations, times and authors have different emotional valences. For instance, people are generally happier on weekends and certain hours of the day, more depressed at the end of summer holidays, and happier in certain states in the United States. Moreover, people have different baseline emotional valences from one another. These claims are supported for example by the annual Gallup poll that ranks states from most happy to least happy (Gallup-Healthways, 2014), or the work by Csikszentmihalyi and Hunter (Csikszentmihalyi and Hunter, 2003) that showed reported

happiness varies significantly by day of week and time of day. We believe these factors manifest themselves in sentiments expressed in tweets and that by accounting for these factors, we can improve sentiment classification on Twitter.

In this work, we explored this hypothesis by utilizing *distant supervision* (Go et al., 2009) to collect millions of labelled tweets from different locations (within the USA), times of day, days of the week, months and authors. We used this data to analyse the variation of tweet sentiments across the aforementioned categories. We then used a Bayesian approach to incorporate the relationship between these factors and tweet sentiments into standard n-gram based Twitter sentiment classification.

This paper is structured as follows. In the next sections we will review related work on sentiment classification, followed by a detailed explanation of our approach and our data collection, annotation and processing efforts. After that, we describe our baseline n-gram sentiment classifier model, followed by the explanation of how the baseline model is extended to incorporate contextual information. Next, we describe our analysis of the variation of sentiment within each of the contextual categories. We then evaluate our models and finally summarize our findings and contributions and discuss possible paths for future work.

2 Related Work

Sentiment analysis and classification of text is a problem that has been well studied across many different domains, such as blogs, movie reviews, and product reviews (e.g., (Pang et al., 2002; Cui et al., 2006; Chesley et al., 2006)). There is also extensive work on sentiment analysis for Twitter. Most of the work on Twitter sentiment classification either focuses on different machine learning techniques (e.g., (Wang et al., 2011; Jiang et al., 2011)), novel features (e.g., (Davidov et al., 2010; Kouloumpis et al., 2011; Saif et al., 2012)), new data collection and labelling techniques (e.g., (Go et al., 2009)) or the application of sentiment classification to analyse the attitude of people about certain topics on Twitter (e.g., (Diakopoulos and Shamma, 2010; Bollen et al., 2011)). These are just some examples of the extensive research already done on Twitter sentiment classification and analysis.

There has also been previous work on measuring the happiness of people in different contexts (location, time, etc). This has been done mostly through traditional land-line polling (Csikszentmihalyi and Hunter, 2003; Gallup-Healthways, 2014), with Gallup's annual happiness index being a prime example (Gallup-Healthways, 2014). More recently, some have utilized Twitter to measure people's mood and happiness and have found Twitter to be a generally good measure of the public's overall happiness, well-being and mood. For example, Bollen et al. (Bollen et al., 2011) used Twitter to measure the daily mood of the public and compare that to the record of social, political, cultural and economic events in the real world. They found that these events have a significant effect on the public mood as measured through Twitter. Another example would be the work of Mitchell et al. (Mitchell et al., 2013), in which they estimated the happiness levels of different states and cities in the USA using Twitter and found statistically significant correlations between happiness level and the demographic characteristics (such as obesity rates and education levels) of those regions.

In this work, we combined the sentiment analysis of different authors, locations, times and dates as measured through labelled Twitter data with standard word-based sentiment classification methods to create a context-dependent sentiment classifier. As far as we can tell, there has not been significant previous work on Twitter sentiment classification that has achieved this.

3 Approach

The main hypothesis behind this work is that the average sentiment of messages on Twitter is different in different contexts. Specifically, tweets in different spatial, temporal and authorial contexts have on average different sentiments. Basically, these factors (many of which are environmental) have an affect on the emotional states of people which in turn have an effect on the sentiments people express on Twitter and elsewhere. In this paper, we used this contextual information to better predict the sentiment of tweets.

Luckily, tweets are tagged with very rich metadata, including location, timestamp, and author information. By analysing labelled data collected from these different contexts, we calculated *prior* probabilities of negative and positive sentiments for each of the contextual categories shown below:

- The states in the USA (50 total).

- Hour of the day (HoD) (24 total).

- Day of week (DoW) (7 total).

- Month (12 total).

- Authors (57710 total).

This means that for every item in each of these categories, we calculated a probability of sentiment being positive or negative based on historical tweets. For example, if seven out of ten historical tweets made on Friday were positive then the prior probability of a sentiment being positive for tweets sent out on Friday is 0.7 and the prior probability of a sentiment being negative is 0.3. We then trained a Bayesian sentiment classifier using a combination of these prior probabilities and standard n-gram models. The model is described in great detail in the "Baseline Model" and "Contextual Model" sections of this paper.

In order to do a comprehensive analysis of sentiment of tweets across aforementioned contextual categories, a large amount of labelled data was required. We needed thousands of tweets for every item in each of the categories (e.g. thousands of tweets per hour of day, or state in the US). Therefore, creating a corpus using human-annotated data would have been impractical. Instead, we turned to distant supervision techniques to obtain our corpus. Distant supervision allows us to have noisy but large amounts of annotated tweets.

There are different methods of obtaining labelled data using distant supervision (Read, 2005; Go et al., 2009; Barbosa and Feng, 2010; Davidov et al., 2010). We used emoticons to label tweets as positive or negative, an approach that was introduced by Read (Read, 2005) and used in multiple works (Go et al., 2009; Davidov et al., 2010). We collected millions of English-language tweets from different times, dates, authors and US states. We used a total of six emoticons, three mapping to positive and three mapping to negative sentiment (table 1). We identified more than 120 positive and negative ASCII emoticons and unicode emojis[1], but we decided to only use the six most common emoticons in order to avoid possible selection biases. For example, people who use obscure emoticons and emojis might have a different base sentiment from those who do not. Using the six most commonly used emoticons limits this bias. Since there are no "neutral" emoticons, our dataset is limited to tweets with positive or negative sentiments. Accordingly, in this work we are only concerned with analysing and classifying the polarity of tweets (negative vs. positive) and not their subjectivity (neutral vs. non-neutral). Below we will explain our data collection and corpus in greater detail.

Positive Emoticons	Negative Emoticons
:)	:(
:-)	:-(
:)	: (

Table 1: List of emoticons.

4 Data Collection and Datasets

We collected two datasets, one massive and labelled through distant supervision, the other small and labelled by humans. The massive dataset was used to calculate the prior probabilities for each of our contextual categories. Both datasets were used to train and test our sentiment classifier. The human-labelled dataset was used as a sanity check to make sure the dataset labelled using the emoticons classifier was not too noisy and that the human and emoticon labels matched for a majority of tweets.

4.1 Emoticon-based Labelled Dataset

We collected a total of 18 million, geo-tagged, English-language tweets over three years, from January 1st, 2012 to January 1st, 2015, evenly divided across all 36 months, using Historical PowerTrack for Twitter[2] provided by GNIP[3]. We created geolocation bounding boxes[4] for each of the 50 states which were used to collect our dataset. All 18 million tweets originated from one of the 50 states and are tagged as such. Moreover, all tweets contained one of the six emoticons in Table 1 and were labelled as either positive or negative based on the emoticon. Out of the 18 million tweets, 11.2 million (62%) were labelled as positive and 6.8 million (38%) were labelled as negative. The 18 million tweets came from $7,657,158$ distinct users.

---

[1]Japanese pictographs similar to ASCII emoticons

[2]Historical PowerTrack for Twitter provides complete access to the full archive of Twitter public data.
[3]https://gnip.com/
[4]The bounding boxes were created using http://boundingbox.klokantech.com/

18

4.2 Human Labelled Dataset

We randomly selected 3000 tweets from our large dataset and had all their emoticons stripped. We then had these tweets labelled as positive or negative by three human annotators. We measured the inter-annotator agreement using *Fleiss' kappa*, which calculates the degree of agreement in classification over that which would be expected by chance (Fleiss, 1971). The *kappa* score for the three annotators was 0.82, which means that there were disagreements in sentiment for a small portion of the tweets. However, the number of tweets that were labelled the same by at least two of the three human annotator was 2908 out of of the 3000 tweets (96%). Of these 2908 tweets, 60% were labelled as positive and 40% as negative.

We then measured the agreement between the human labels and emoticon-based labels, using only tweets that were labelled the same by at least two of the three human annotators (the majority label was used as the label for the tweet). Table 2 shows the confusion matrix between human and emoticon-based annotations. As you can see, 85% of all labels matched ($\frac{1597+822}{1597+882+281+148} = .85$).

	Human-Pos	Human-Neg
Emot-Pos	1597	281
Emot-Neg	148	882

Table 2: Confusion matrix between human-labelled and emoticon-labelled tweets.

These results are very promising and show that using emoticon-based distant supervision to label the sentiment of tweets is an acceptable method. Though there is some noise introduced to the dataset (as evidenced by the 15% of tweets whose human labels did not match their emoticon labels), the sheer volume of labelled data that this method makes accessible, far outweighs the relatively small amount of noise introduced.

4.3 Data Preparation

Since the data is labelled using emoticons, we stripped all emoticons from the training data. This ensures that emoticons are not used as a feature in our sentiment classifier. A large portion of tweets contain links to other websites. These links are mostly not meaningful semantically and thus can not help in sentiment classification. Therefore, all links in tweets were replaced with the token "URL". Similarly, all mentions of user-

names (which are denoted by the @ symbol) were replaced with the token "USERNAME", since they also can not help in sentiment classification. Tweets also contain very informal language and as such, characters in words are often repeated for emphasis (e.g., the word *good* is used with an arbitrary number of *o*'s in many tweets). Any character that was repeated more than two times was removed (e.g., *goooood* was replaced with *good*). Finally, all words in the tweets were stemmed using *Porter Stemming* (Porter, 1980).

5 Baseline Model

For our baseline sentiment classification model, we used our massive dataset to train a negative and positive n-gram language model from the negative and positive tweets.

As our baseline model, we built purely linguistic bigram models in Python, utilizing some components from NLTK (Bird et al., 2009). These models used a vocabulary that was filtered to remove words occurring 5 or fewer times. Probability distributions were calculated using Kneser-Ney smoothing (Chen and Goodman, 1999). In addition to Kneser-Ney smoothing, the bigram models also used "backoff" smoothing (Katz, 1987), in which an n-gram model falls back on an $(n-1)$-gram model for words that were unobserved in the n-gram context.

In order to classify the sentiment of a new tweet, its probability of fit is calculated using both the negative and positive bigram models. Equation 1 below shows our models through a Bayesian lens.

$$\Pr(\theta_s \mid W) = \frac{\Pr(W \mid \theta_s)\Pr(\theta_s)}{\Pr(W)} \quad (1)$$

Here θ_s can be θ_p or θ_n, corresponding to the hypothesis that the sentiment of the tweet is positive or negative respectively. W is the sequence of ℓ words, written as w_1^ℓ, that make up the tweet. $\Pr(W)$ is not dependent on the hypothesis, and can thus be ignored. Since we are using a bigram model, Equation 1 can be written as:

$$\Pr(\theta_s \mid W) \propto \prod_{i=2}^{\ell} \Pr(w_i \mid w_{i-1}, \theta_s)\Pr(\theta_s) \quad (2)$$

This is our purely linguistic baseline model.

6 Contextual Model

The Bayesian approach allows us to easily integrate the contextual information into our models. $\Pr(\theta_s)$ in Equation 2 is the prior probability of a tweet having the sentiment s. The prior probability ($\Pr(\theta_s)$) can be calculated using the contextual information of the tweets. Therefore, $\Pr(\theta_s)$ in equation 2 is replaced by $\Pr(\theta_s|C)$, which is the probability of the hypothesis given the contextual information. $\Pr(\theta_s|C)$ is the posterior probability of the following Bayesian equation:

$$\Pr(\theta_s \mid C) = \frac{\Pr(C \mid \theta_s)\Pr(\theta_s)}{\Pr(C)} \qquad (3)$$

Where C is the set of contextual variables: $\{State, HoD, Dow, Month, Author\}$. $\Pr(\theta_s|C)$ captures the probability that a tweet is positive or negative, given the state, hour of day, day of the week, month and author of the tweet. Here $\Pr(C)$ is not dependent on the hypothesis, and thus can be ignored. Equation 2 can therefore be rewritten to include the contextual information:

$$\Pr(\theta_s \mid W, C) \propto \prod_{i=2}^{\ell} \Pr(w_i \mid w_{i-1}, \theta_s) \\ \Pr(C \mid \theta_s)\Pr(\theta_s) \qquad (4)$$

Equation 4 is our extended Bayesian model for integrating contextual information with more standard, word-based sentiment classification.

7 Sentiment in Context

We considered five contextual categories: one spatial, three temporal and one authorial. Here is the list of the five categories:

- The states in the USA (50 total) (spatial).

- Hour of the day (HoD) (24 total) (temporal).

- Day of week (DoW) (7 total) (temporal).

- Month (12 total) (temporal).

- Authors (57,710 total) (authorial).

We used our massive emoticon labelled dataset to calculate the average sentiment for all of these five categories. A tweet was given a score of -1 if it was labelled as negative and a score 1 if it was labelled as positive, so an average sentiment of 0 for a contextual category would mean that tweets in that category were evenly labelled as positive and negative.

7.1 Spatial

All of the 18 million tweets in our dataset originate from the USA and are geo-tagged. Naturally, the tweets are not evenly distributed across the 50 states given the large variation between the population of each state. Figure 1 shows the percentage of tweets per state, sorted from smallest to largest. Not surprisingly, California has the highest number of tweets ($2,590,179$), and Wyoming has the lowest number of tweets ($11,719$).
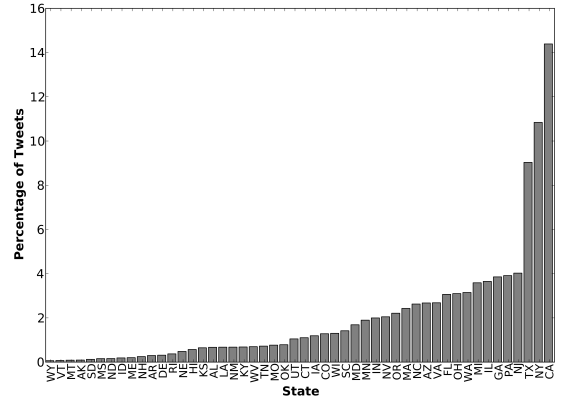

Figure 1: Percentage of tweets per state in the USA, sorted from lowest to highest.

Even the state with the lowest percentage of tweets has more than ten thousand tweets, which is enough to calculate a statistically significant average sentiment for that state. The sentiment for all states averaged across the tweets from the three years is shown in Figure 2. Note that an average sentiment of 1.0 means that all tweets were labelled as positive, -1.0 means that all tweets were labelled as negative and 0.0 means that there was an even distribution of positive and negative tweets. The average sentiment of all the states leans more towards the positive side. This is expected given that 62% of the tweets in our dataset were labelled as positive.

It is interesting to note that even with the noisy dataset, our ranking of US states based on their Twitter sentiment correlates with the ranking of US states based on the well-being index calculated by Oswald and Wu (Oswald and Wu, 2011) in their work on measuring well-being and life satisfaction across America. Their data is from the behavioral risk factor survey score (BRFSS), which is a survey of life satisfaction across the United States from 1.3 million citizens. Figure 3 shows this correlation ($r = 0.44$, $p < 0.005$).
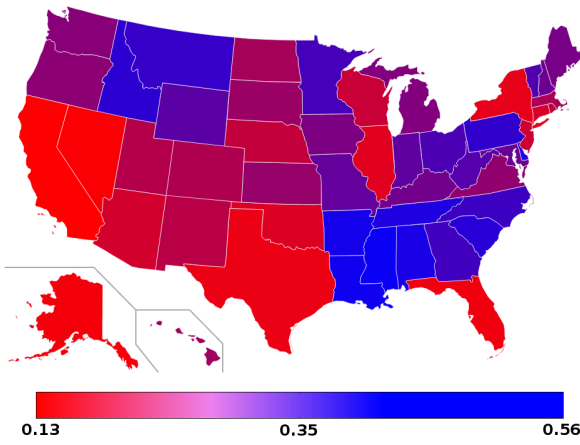
Figure 2: Average sentiment of states in the USA, averaged across three years, from 2012 to 2014.
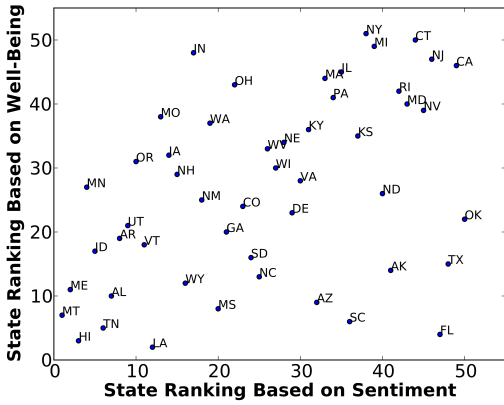

Figure 3: Ranking of US states based on Twitter sentiment vs. ranking of states based on their well-being index. $r = 0.44$, $p < 0.005$.

7.2 Temporal

We looked at three temporal variables: time of day, day of the week and month. All tweets are tagged with timestamp data, which we used to extract these three variables. Since all timestamps in the Twitter historical archives (and public API) are in the UTC time zone, we first converted the timestamp to the local time of the location where the tweet was sent from. We then calculated the sentiment for each day of week (figure 4), hour (figure 5) and month (figure 6), averaged across all 18 million tweets over three years. The 18 million tweets were divided evenly between each month, with 1.5 million tweets per month. The tweets were also more or less evenly divided between each day of week, with each day having somewhere between 14% and 15% of the tweets. Similarly, the tweets were almost evenly divided

between each hour, with each having somewhere between 3% and 5% of the tweets.

Some of these results make intuitive sense. For example, the closer the day of week is to Friday and Saturday, the more positive the sentiment, with a drop on Sunday. As with spatial, the average sentiment of all the hours, days and months lean more towards the positive side.
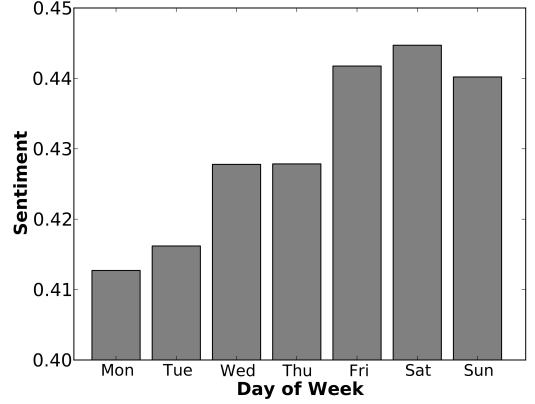

Figure 4: Average sentiment of different days of the week in the USA, averaged across three years, from 2012 to 2014.
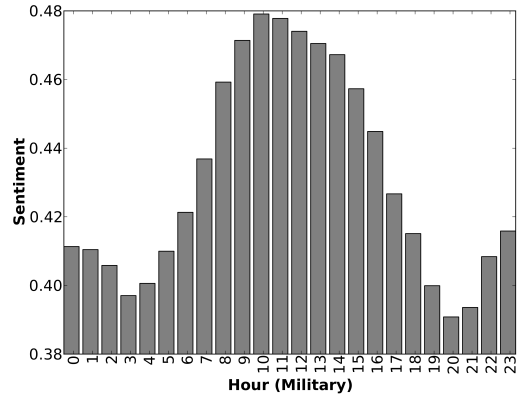

Figure 5: Average sentiment of different hours of the day in the USA, averaged across three years, from 2012 to 2014.

7.3 Authorial

The last contextual variable we looked at was authorial. People have different baseline attitudes, some are optimistic and positive, some are pessimistic and negative, and some are in between. This difference in personalities can manifest itself in the sentiment of tweets. We attempted to capture this difference by looking at the history of
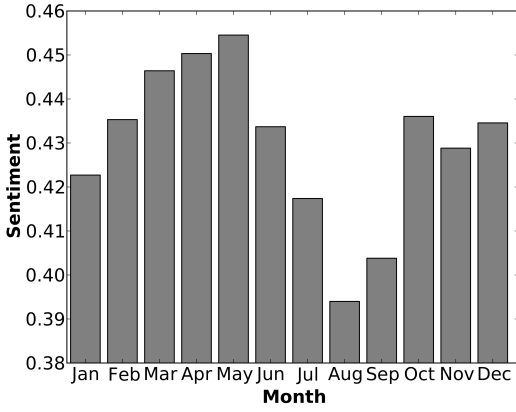
Figure 6: Average sentiment of different months in the USA, averaged across three years, from 2012 to 2014.
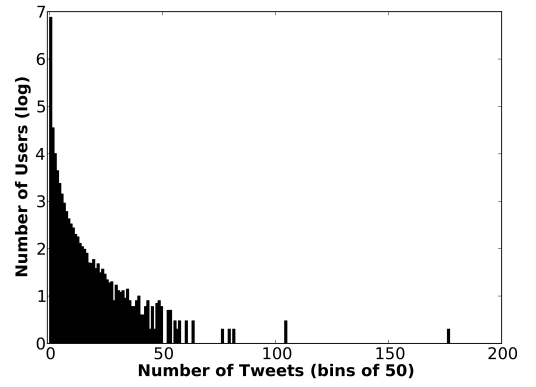

Figure 7: Number of users (logarithmic) in bins of 50 tweets. The first bin corresponds to number of users that have less than 50 tweets throughout the three years and so on.
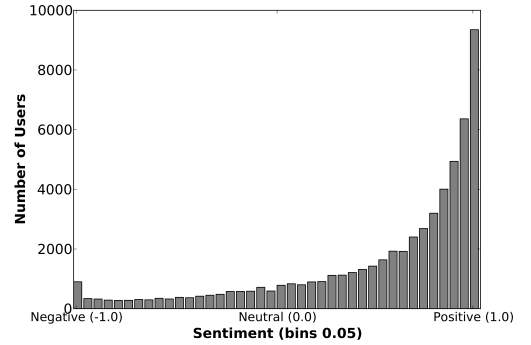

Figure 8: Number of users (with at least 50 tweets) per sentiment bins of 0.05, averaged across three years, from 2012 to 2014.

tweets made by users. The 18 million labelled tweets in our dataset come from $7,657,158$ authors.

In order to calculate a statistically significant average sentiment for each author, we need our sample size to not be too small. However, a large number of the users in our dataset only tweeted once or twice during the three years. Figure 7 shows the number of users in bins of 50 tweets. (So the first bin corresponds to the number of users that have less than 50 tweets throughout the three year.) The number of users in the first few bins were so large that the graph needed to be logarithmic in order to be legible. We decided to calculate the prior sentiment for users with at least 50 tweets. This corresponded to less than 1% of the users ($57,710$ out of $7,657,158$ total users). Note that these users are the most prolific authors in our dataset, as they account for 39% of all tweets in our dataset. The users with less than 50 posts had their prior set to 0.0, not favouring positive or negative sentiment (this way it does not have an impact on the Bayesian model, allowing other contextual variables to set the prior).

As it is not feasible to show the prior average sentiment of all $57,710$ users, we created 20 even sentiment bins, from -1.0 to 1.0. We then plotted the number of users whose average sentiment falls into these bins (Figure 8). Similar to other variables, the positive end of the graph is much heavier than the negative end.

8 Results

We used 5-fold cross validation to train and evaluate our baseline and contextual models, ensuring that the tweets in the training folds were not used in the calculation of any of the priors or in the training of the bigram models. Table 3 shows the accuracy of our models. The contextual model outperformed the baseline model using any of the contextual variables by themselves, with state being the best performing and day of week the worst. The model that utilized all the contextual variables saw a 10% relative and 8% absolute improvement over the baseline bigram model.

Because of the great increase in the volume of data, distant supervised sentiment classifiers for Twitter tend to generally outperform more standard classifiers using human-labelled datasets. Therefore, it makes sense to compare the performance of our classifier to other distant supervised

Model	Accuracy
Baseline-Majority	0.620
Baseline-Bigram	0.785
Contextual-DoW	0.798
Contextual-Month	0.801
Contextual-Hour	0.821
Contextual-Author	0.829
Contextual-State	0.849
Contextual-All	**0.862**

Table 3: Classifier accuracy, sorted from worst to best.

classifiers. Though not directly comparable, our contextual classifier outperforms the distant supervised Twitter sentiment classifier by Go et al (Go et al., 2009) by more than 3% (absolute).

Table 4 shows the precision, recall and F1 score of the positive and negative class for the full contextual classifier (Contextual-All).

Class	Precision	Recall	F1 Score
Positive	0.864	0.969	0.912
Negative	0.905	0.795	0.841

Table 4: Precision, recall and F1 score of the full contextual classifier (Contexual-All).

9 Discussions

Even though our contextual classifier was able to outperform the previous state-of-the-art, distant supervised sentiment classifier, it should be noted that our contextual classifier's performance is boosted significantly by spatial information extracted through geo-tags. However, only about one to two percent of tweets in the wild are geo-tagged. Therefore, we trained and evaluated our contextual model using all the variables except for state. The accuracy of this model was 0.843, which is still significantly better than the performance of the purely linguistic classifier. Fortunately, all tweets are tagged with timestamps and author information, so all the other four contextual variables used in our model can be used for classifying the sentiment of any tweet.

Note that the prior probabilities that we calculated need to be recalculated and updated every once in a while to account for changes in the world. For example, a state might become more affluent, causing its citizens to become on average happier. This change could potentially have an effect on the average sentiment expressed by the citizens of that state on Twitter, which would make our priors obsolete.

10 Conclusions and Future Work

Sentiment classification of tweets is an important area of research. Through classification and analysis of sentiments on Twitter, one can get an understanding of people's attitudes about particular topics.

In this work, we utilized the power of distant supervision to collect millions of noisy labelled tweets from all over the USA, across three years. We used this dataset to create prior probabilities for the average sentiment of tweets in different spatial, temporal and authorial contexts. We then used a Bayesian approach to combine these priors with standard bigram language models. The resulting combined model was able to achieve an accuracy of 0.862, outperforming the previous state-of-the-art distant supervised Twitter sentiment classifier by more than 3%.

In the future, we would like to explore additional contextual features that could be predictive of sentiment on Twitter. Specifically, we would like to incorporate the topic type of tweets into our model. The topic type characterizes the nature of the topics discussed in tweets (e.g., breaking news, sports, etc). There has already been extensive work done on topic categorization schemes for Twitter (Dann, 2010; Sriram et al., 2010; Zhao and Jiang, 2011) which we can utilize for this task.

11 Acknowledgements

References

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proc. COLING 2010*, pages 36–44. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proc. ICWSM 2011*.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Paula Chesley, Bruce Vincent, Li Xu, and Rohini K Srihari. 2006. Using verbs and adjectives to automatically classify blog sentiment. *Training*, 580(263):233.

Mihaly Csikszentmihalyi and Jeremy Hunter. 2003. Happiness in everyday life: The uses of experience sampling. *Journal of Happiness Studies*, 4(2):185–199.

Hang Cui, Vibhu Mittal, and Mayur Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *AAAI*, volume 6, pages 1265–1270.

Stephen Dann. 2010. Twitter content classification. *First Monday*, 15(12).

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proc. COLING 2010*, pages 241–249. Association for Computational Linguistics.

Nicholas A Diakopoulos and David A Shamma. 2010. Characterizing debate performance via aggregated twitter sentiment. In *Proc. SIGCHI 2010*, pages 1195–1198. ACM.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gallup-Healthways. 2014. State of american well-being. *Well-Being Index*.

Alec Go, Lei Huang, and Richa Bhayani. 2009. Twitter sentiment analysis. *Entropy*, 17.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proc. ACL 2011: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(3):400–401.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! *Proc. ICWSM 2011*.

Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5):e64417.

Robert J Moore. 2009. Twitter data analysis: An investor's perspective. http://techcrunch.com/2009/10/05/twitter-data-analysis-an-investors-perspective-2/. Accessed: 2015-01-30.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using crfs with hidden variables. In *Proc. NAACL-HLT 2010*, pages 786–794. Association for Computational Linguistics.

Andrew J Oswald and Stephen Wu. 2011. Well-being across america. *Review of Economics and Statistics*, 93(4):1118–1134.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proc. EMNLP 2002-Volume 10*, pages 79–86. Association for Computational Linguistics.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.

Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics.

Hassan Saif, Yulan He, and Harith Alani. 2012. Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org).

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proc. ACM SIGIR 2010*, pages 841–842. ACM.

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 *Proc. CIKM 2011*, pages 1031–1040. ACM.

Xin Zhao and Jing Jiang. 2011. An empirical comparison of topics in twitter and traditional media. *Singapore Management University School of Information Systems Technical paper series. Retrieved November*, 10:2011.