E3 SemEval 2016: Elementary Twitter Sentiment Analysis

Emily Ahn, Emily Chen, Elizabeth Hau
Department of Computer Science
Wellesley College
{eahn, tchen4, hhau}@wellesley.edu

Abstract

We modified a SemEval 2016 challenge (Task 4-C) of classifying Twitter tweets such that given a tweet about a Public Figure, we can classify its sentiment on a 3-point polarity scale. We used elementary features such as scoring a tweet with the AFINN sentiment lexicon, evaluating emoticons, and counting the number of words in capital letters. We trained an SVM classifier and tested it on a total of 4348 tweets. Our results gave us an accuracy of 48.3%, given a baseline of 45.8%.

Introduction

Over the past few years, social media has become a powerful broadcaster of the public's sentiments around topics ranging from celebrity gossip to current events. Twitter, in particular, has been an excellent example of this by providing a platform for users to quickly translate their thoughts into an informal, 140-character tweet. It is important to evaluate what the makeup of the Twitter community is in order to be aware of the biases that can emerge during our analysis. In the United States, Twitter accounts for 23% of adult internet users, or 19% of the entire adult population (PEW Research Center, 2015). So although the content found on the site does not represent the majority of society, it still embodies the opinions of a significant portion of the population. As of 2014, 24% of male internet users and 21% of female internet users were on Twitter. A significant number of users are also young professionals (pages 18-29), have graduated college, and are making average if not higher income (PEW Research Center, 2015). From these demographics (see Figure 1), we can conclude that a meaningful amount of tweets are expressed by educated, young, working professionals with a fairly even gender split. Of course, there is still a lot to be said about the variety of user backgrounds in the Twitter community, but these figures help give a general idea of what kind of user base our data represents.

Twitter users

Among online adults, the % who use Twitter

	2013	2014
All internet users	*18%*	*23%**
Men	17	24*
Women	18	21
White, Non-Hispanic	16	21 *
Black, Non-Hispanic	29	27
Hispanic	16	25
18-29	31	37
30-49	19	25
50-64	9	12
65+	5	10*
High school grad or less	17	16
Some college	18	24
College+ (n= 685)	18	30*
Less than $30,000/yr	17	20
$30,000-$49,999	18	21
$50,000-$74,999	15	27*
$75,000+	19	27*
Urban	18	25*
Suburban	19	23
Rural	11	17

Source: Pew Research Center's Internet Project September Combined Omnibus Survey, September 11-14 & September 18-21, 2014. N=1,597 internet users ages 18+. The margin of error for all internet users is +/- 2.9 percentage points. 2013 data from Pew Internet August Tracking Survey, August 07 – September 16, 2013, n= 1,445 internet users ages 18+.

Note: Percentages marked with an asterisk (*) represent a significant change from 2013. Results are significant at the 95% confidence level using an independent z-test.

PEW RESEARCH CENTER

Figure 1: Twitter User Demographics in the USA in 2014

Untangling the surface sentiment behind these tweets is valuable to determine how people convey their impressions of different topics on social media. Since the variety of topics discussed on Twitter is so broad, we have decided to focus our analysis only on tweets about public figures. The public figures include politicians, Hollywood celebrities, and athletes. Our goal is to be able to classify these public figures tweets on a three-point classification scale ranging

from negative to positive, and thus build a resource for learning about the language of sentiment. More information about the details of the challenge presented by the SemEval project can be found here: http://alt.qcri.org/semeval2016/task4/

Although tweets are short, sweet, and to the point, processing them is a significant challenge. Often, these sentence fragments contain slang, creative spelling or punctuation, misspellings, URLs, abbreviations, or word formats that are specific to Twitter (e.g. hashtags, "RT"). We need to filter past these attributes to effectively categorize tweets into three categories: positive, neutral, and negative. These labels are represented in the data by the numeric values: 1, 0, and -1 respectively. Our challenge can be accomplished by examining certain features of the tweets such as looking at certain words used in the tweet, the context in which they are placed, and other factors that could lead to predicting a tweet's sentiment.

Every day, we are presented with vast amounts of information ranging from ads on the subway to news articles popping up on our phone. We are able to access more information, more conveniently, than ever before. Our minds process this information uniquely, given different biases in the things we read or hear, as well as from our own personal experiences. This then turns into an opinion that we form about a certain subject and often times, we choose to share that view publicly on social media or among our friends. Sometimes, that perspective evolves as we are presented with more information or opinions from other individuals about the same matter influence our ideas.

Twitter has risen to become one of the most popular forms for people to express their voice due to its quick and easy posting system. Users can post a short tweet about their thoughts on practically any topics. Then, other users can either retweet it, favorite it, or reply to it. This streamlined system has allowed discussion to generate freely and instantaneously between members of the Twitter community, which extends internationally and across a range of demographics.

Examining how society expresses their reactions towards various topics across social media is critical to understanding what contributes to public opinion and how the language of sentiment is built. Sentiment analysis can be used to measure public opinion quickly and cheaply, thus making it meaningful due to its instantaneous value since the data gathered from it is relevant to users live. The insight gathered from this study identifies what indicators in language express one's sentiment, which can then be utilized by companies, public figures, and other entities that use Twitter as a platform to represent themselves, to design their marketing or product campaigns to better suit consumers. The results from Twitter classification can also reveal insights into the mechanisms of the human mind and the how the public processes and expresses information.

Related Work (existing approach to similar problems)

In order to get a sense of where to start, we read papers from several teams that participated in previous SemEval conferences. Our subtask this year is new, but it relates to 2015's Task 10, specifically a mixture of subtasks C and E (Rosenthal et al., 2015). 10-C was trying to classify a Tweet on a 3-point scale of the polarity of a topic. 10-E was trying to determine the degree of the polarity (given that it's positive) of a topic. We hoped start out with a system like 10-C first, then incorporate degrees of a positive or a negative sentiment from 10-E next. However by the end of our project, we made a decision to only use a 3-point scale as our final data set—essentially to build a system that was just the 10-C task.

A team from IIIT, Dalmia et al. (2015), worked on 10-B (which is a 3-way message polarity classification, regardless of topic) and provided some strategies that we wanted to try. They used a few sentiment lexicons and then created a couple of their own dictionaries (including one for emoticons and one for acronyms). They made sure to preprocess the data to make it cleaner, like taking out URLs, non-English text, etc, in addition to tokenizing it. They then built feature vectors for emotion tokens, and added on word counts and scores from the sentiment lexicons. They chose the SVM classifier to learn the data and make predictions, which was successful about 65% of the time. Dalmia et al. were detailed and to-the-point in their methodology.

Another system called TwitterHawk (Boag et al., 2015) placed 1st in the 10-C classification from SemEval 2015, and they described many other techniques such as: running a spell checker from a library, segmenting hashtags, and normalizing for negation. They also used a linear SVM classifier. Then we found that what made their features so robust was using a variety of lexicons that processed the polarity scoring as different types of features. They cite their most success as coming from analyzing hashtags, and also from their models that they built from other subtasks of Task 10. Since their system is so complex, we only focused on a few elements of their system.

There are a variety of other papers out there, and we used them for inspiration. One of these papers discussed several teams' methods of classifying tweets from SemEval 2013, Task 2. Hagen et al. (2015) compared 4 teams' features, and we saw that many of them included a Polarity Dictionary—which is what we implemented first.

Data

This section describes the type of data we used, how we collected them, statistics about the data, and the data preprocessing steps.

The organizers of SemEval provided the data. The files we were provided with were in three datasets: train, dev, and devtest, all of which contain Twitter status IDs, a topic, and the label, formatted as:

id<TAB>topic<TAB>sentiment label

where the sentiment labels were on a five-point scale (subtask C): 2 (very positive), 1 (positive), 0 (neutral), -1 (negative), -2 (very negative). To obtain the tweets that correspond to the Twitter status IDs, we ran a downloading script for the dev, devtest, and train text files that downloaded the datasets with the actual tweets as .tsv files from Twitter. Each of these files obtained from the download contain Twitter status IDs, a topic, the label (sentiment) for the tweet, along with the tweet itself, formatted as:

id<TAB>topic<TAB>sentiment label<TAB>tweet

There are exactly 100 tweets provided for each topic and a total of 100 topics, giving us a total of 10,000 tweets. The 100 topics provided are split between the three datasets so that 60 topics are in train, 20 in dev, and 20 topics are in devtest. However, some of the tweets were not available at the time of download due to deletion of the tweet, change on the privacy settings, or deactivated accounts, so we wrote a script to remove the tweets that are "Not Available," results in Table 1.

Table 1 shows that although we were provided with 10,000 tweets to begin with, after removing tweets that were not available at the time of the download, we end up with 9049 tweets and 9020 distinct tweet IDs. This indicates that the tweets and the tweet IDs also do not have a one-to-one mapping, which means we cannot simply store all the data in a dictionary with the tweet IDs as the keys.

	Initial number of tweets downloaded	Number of tweets after removing unavailable tweets	Number of Distinct Tweet IDs
Train	6000	5444	5420
Dev	2000	1814	1811
Devtest	2000	1791	1789
Total	**10,000**	**9049**	**9020**

Table 1: Distribution of tweets and tweet IDs

With hopes of getting better results given the time constraint, along with the fact that the labeled data provided was rarely on the extremes (-2 or 2), we decided to work on a 3-point scale instead of the given 5-point scale. We decided to reduce the scale because upon examining the distribution of the 5 labels, we discovered that amongst the 9049 tweets, only 1.5% of the tweets had the label -2 and 6.1% of the tweets had the label 2. However, instead of redownloading the data from the SemEval website for the subtask that classifies tweets on a 3-point scale, we wrote a script to change all negative labels in our current data that is on a 5-point scale to -1 and all positive labels to 1. Additionally, while going through the types of topics we are given, we noticed that the topics generally fall into one of the following topics: public figures, product, or company. Among the 100 topics provided, 49 topics fell into the public figures category, so we chose to only train and test on public figures because we predicted tweets within the same category would have similar content, and we could therefore extract features that are more relevant.

Before we started extracting features, we performed some data preprocessing to simplify the feature implementation process. Besides removing tweets that were not available at the time of the download, we used CMU's ARK natural language processing part-of-speech (POS) tagger (Owoputi et al., 2013; Gimpel et al., 2011) to tag the tweets. It is worth noting that the ARK NLP POS tagger is a Twitter specific tagger that tags emoticons, user mentions, and hashtags along with the usual part-of-speech tags (Dalmia et al., 2015). Therefore, using the tag tweets, we can remove URLs, user mentions, numbers, and hashtags from the tweet. We removed URLs because the URLs in the tweets are shortened and generally do not carry any sentiment about the tweet. Similarly, numbers and user mentions (indicated with an '@' followed by a username or an organization's name) are also thought to be parts of the tweets that do not contribute to determining the sentiment of a given tweet. Hashtags are also removed in this case because since hashtags are unspaced phrases, trying to decipher the words embedded within a hashtag would require us to segment the hashtag and use additional hashtag sentiment lexicons such as the NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013) used by Dalmia et al. and TwitterHawk, which we did not have time to get to this time.

After preprocessing the data, we divided the 49 public figure topics into training and testing data: the first 75% of the topics were used for training and the rest 25% of the topics were used for testing. This gave us 3259 tweets for train and 1089 tweets for test, with the frequency of labels [-1,0,1] being [533, 997, 1729] and [218, 372, 499], respectively. This means the distribution of tweets labeled as [-1,0,1] are [0.164, 0.306, 0.53] in the training data and [0.20, 0.342, 0.458] in the testing data.

Methodology

We used a variety of tools to aid us in analyzing Twitter sentiment. To first process the tweets, as mentioned in the Data section we used CMU's ARK natural language processing part-of-speech (POS) tagger with a Python port (Owoputi et al., 2013; Gimpel et al., 2011). This allowed us to access words by their tags and analyze them in a separate group. Additionally, we downloaded the AFINN lexicon, a lexicon that used a scaled polarity dictionary to score independently, based on its positive or negative sentiment (Nielsen, 2011). A single score for a word in its dictionary would be integer values ranging from -5 to 5. AFINN also had a separate emoticon dictionary that we used to create additional features.

As for the features themselves, we created 3 types: an AFINN total score, an AFINN Emoticon score, and the number of words in all-CAPS. For the first two types, we took the score based on the total sum of each word's polarized sentiment in the tweet. For the AFINN score itself, we would use binary features (i.e. whether the score was -5 or not, whether it was -4 or not, etc). We tried having a set of 21 binary AFINN score features, then realized this was too many, and collapsed it to 11 features, then 7 features (see the Results section below for how each feature size performed). In a similar manner, the Emoticon score took all emoticons (accessed via the tagger) from a single tweet and summed over their score (retrieved from the AFINN emoticon lexicon). We created 5 binary features for whether that total Emoticon score was -2 (or below), -1, 0, 1, or 2 (or above). Lastly, we counted how many of the words in the tweet were in all-CAPS; we used binary features to distinguish up to 6 CAPS words; anything more would have the value of 6. We predicted that this could help us discern between a user showing polarity versus being neutral.

We used an SVM classifier from the Scikit-learn Python library to train and test our data. More specifically, we used an rbf kernel and a C value of 0.75 as recommended by Dalmia et al. (2015). Lastly, we analyzed our results with the accuracy_score and confusion_matrix functions from the Scikit-learn metrics module.

Results

This section describes the combination of features that were run during testing and their respective results.

As mentioned in the data section above, the distribution of the labels in train and test were [0.16, 0.31, 0.53] and [0.20, 0.342, 0.458], respectively. Using that information, our baseline was 0.458, which is the proportion of test labels that were labeled 1. The results of running the SVM classifier on the test data using a combination of various features are shown in Table 2.

Features	(1)	(2)	(3)
AFINN Score [-10,10]	✓		
AFINN Score [-5,5]		✓	
AFINN Score [-3,3]			✓
# words in tweet in all CAPS (max 6 words)	✓	✓	✓
Emoticon feature using AFINN emoticon scores [-2,2]		✓	✓
Accuracy	**0.4747**	**0.4812**	**0.4830**

Table 2: Feature results in tweet classification

Table 2 shows the combination of features we have tested and the accuracies we obtained. After obtaining the accuracies of our classifier, we found the confusion matrix for the best-performing feature combination (column (3) from Table 2) to be the following matrix:

$$\begin{bmatrix} 0 & 67 & 151 \\ 0 & 60 & 312 \\ 0 & 33 & 466 \end{bmatrix}$$

Figure 2: Confusion matrix without normalization

with a normalized form:

$$\begin{bmatrix} 0 & 0.31 & 0.69 \\ 0 & 0.16 & 0.84 \\ 0 & 0.07 & 0.93 \end{bmatrix}$$

Figure 3: Confusion matrix with normalization

Each row represents the distribution of how a correct label (starting with -1, then 0, then 1) is predicted by our classifier. The diagonal of the matrix in Figure 2 (0, 60, 466) indicates the number of tweets in the test that were correctly labeled as -1, 0, and 1, respectively. Figures 2 and 3 show that among all the -1 labels (row 1), 67 (31%) of the labels were incorrectly labeled as 0, and 151 (69%) of the labels were incorrectly labeled as 1. Among all the 0

labels, the classifier correctly predicts 60 (16%) of the labels and incorrectly predicts 312 (0.84) of the 0 labels as 1. Finally, the classifier only incorrectly predicts the label 1 as 0 33 (7%) of the time and correctly predicts the label 1 466 (93%) of the time. The high numbers in the last column indicate that our classifier tends to predict a 1 more often than any other label no matter what the actual label is. This could be because 45.8% of the tweets in the training data were labeled 1, skewing the classifier's predictions towards 1 whenever it is attempting a prediction. On the other hand, notice the first column of zeroes. This shows that the SVM classifier never predicts a -1 label for any of the tweets during the test. One possible reason for this is the low percentage (16.4%) of tweets having the label -1 in the training data. Interestingly, for the correct labels of -1 and 0, a higher percentage of 0's get classified as 1's, compared to the percentage of -1's classified as 1's. This is a good sign, showing that the SVM classifier is picking up on some gradient of sentiment. However, the line it draws is off-balanced.

Another possible explanation for our accuracies being just above baseline is that in our AFINN score features, we only take the sum of the sentiment scores across all words in a tweet. This could be a problem if there exists conflicting sentiment within one tweet. A better way to determine a label could be to count the number of positive tags and the number of negative tags in a tweet.

Both the all CAPS feature and the emoticon feature perform at baseline level with an accuracy of 45.8% when ran alone or with each other (not including the AFINN word sentiment score feature). This could be due to the small amount of training data we have to sufficiently train the classifier and make the features useful. For example, there are 70 words in the training data tagged as an Emoticon, and only 14 tagged Emoticons in the test data. Since the all CAPS feature was to determine whether a tweet carried sentiment or not, a different methodological approach could have helped us as well: having a two-step binary classifier by first distinguishing between {1, -1} and 0, then next distinguishing between a positive and negative label.

Moreover, as shown in Table 2, the accuracy increases slightly when we decrease the range of the binary features for the AFINN Score. Perhaps this is because a smaller set of possible features allows the SVM classifier to divide a data set that is more robust, without extraneous zero features. To produce better results, we need to be careful only to include the features that are robust and useful.

References

Boag, W., Potash, P., & Rumshisky, A. 2015. *TwitterHawk: A feature bucket based approach to sentiment analysis*. In Proceedings of the 9th International Workshop on Semantic Evaluation, pages 640–646. ACL.

Dalmia, A., Gupta, M., Varma, V. 2015. *SemEval 2015: Twitter Sentiment Analysis The good, the bad and the neutral*! SemEval 2015.

Duggan, M., Ellison, N., Lampe, C., Lenhart, A., & Madden, M. (2015, January 9). *Demographics of Key Social Networking Platforms*. Retrieved December 17, 2015, from http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein J., Heilman M., Yogatama D., Flanigan, J., & Smith, N. 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. Association for Computational Linguistics.

Nielsen, F. 2011. *A new anew: Evaluation of a word list for sentiment analysis in microblogs*. arXiv preprint arXiv:1103.2903

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. 2013. *Improved part-of-speech tagging for online conversational text with word clusters*. In HLT-NAACL, pages 380–390.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., Stoyanov, V. 2015. *Semeval-2015 task 10: Sentiment analysis in twitter*. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval 2015, Denver, Colorado.