Domain-Specific Sentiment Classification for Games-Related Tweets

Trevor Sarratt, Soja-Marie Morgens, and Arnav Jhala
Baskin School of Engineering
University of California, Santa Cruz, Santa Cruz, CA 95064
{tsarratt, smorgens, jhala}@soe.ucsc.edu

Abstract

Sentiment classification provides information about the author's feeling toward a topic through the use of expressive words. However, words indicative of a particular sentiment class can be domain-specific. We train a text classifier for Twitter data related to games using labels inferred from emoticons. Our classifier is able to differentiate between positive and negative sentiment tweets labeled by emoticons with 75.1% accuracy. Additionally, we test the classifier on human-labeled examples with the additional case of neutral or ambiguous sentiment. Finally, we have made the data available to the community for further use and analysis.

Introduction

Sentiment classification has become a popular method of analyzing text in recent years. Twitter, one of the most popular social media websites used today, allows users to discuss topics in small messages, called "tweets", restricted to one hundred and forty characters. Given the widespread use of the service, analyzing the sentiment of these discussions can be a valuable method of determining how a particular product, company, celebrity, or event is viewed by users.

In recent years, researchers have utilized existing symbolic indicators to automatically identify the sentiment of a message. Read (2005) proposed using the presence of certain emoticons, a facial representation composed of symbols and characters, to generate a set of data for training text classifiers for Usenet newsgroups. Go et al. (2009) extended this approach to the analysis of Twitter data.

As text classification can be topic-dependent (Engström 2004) or domain-dependent (Read 2005), we are publicly releasing a set of Twitter data related to games for use by the community. To demonstrate one potential application of the data, we implement a text classifier in the same manner as (Read 2005; Go, Bhayani, and Huang 2009), with evaluation against an established sentiment analysis scoring algorithm as well as against human-labeled examples.

Data

We collected tweets using the streamR package (Barbera 2014) for access to the Twitter Streaming API. The data was collected hourly between the dates of May 2nd, 2014 and June 29th, 2014. We restricted tweets to those containing at least one of a list of specified keywords or hashtags from broad categories such as platform (e.g. "Xbox", "PlayStation"), title ("Call of Duy", "Skylander"), or journalism ("Polygon", "IGN").

In total, the 7,294,826 tweets in the data set contain 1,928,257 retweets, 75,760 tweets containing negative emoticons, and 101,855 tweets containing positive emoticons. Additionally, though it is not considered in this work, the tweets span a time leading up to and including the Electronic Entertainment Expo for 2014.

Positive	Negative
:) ;) :D ;D :P ;P :-) ;-)	:(:C D: :/ :'(:'C :@ :-(
:-D ;-D :-P ;-P =) =D =P	:-C D-: :-/ :-@ =(D= =C

Table 1: List of emoticons and their corresponding sentiment class.

Table 1 contains the sets of emoticons we associated with either positive or negative sentiment. These sets were constructed by adding several types of emoticons observed within the data set to those used in (Go, Bhayani, and Huang 2009) and (Read 2005).

Classification Model

We constructed a naive Bayes (NB) classifier using the frequencies of unigrams associated with emoticons, our label indicator, in our dataset. The probability of each potential classification is calculated as in Equation 1.

$$P(class|tweet) = \frac{P(class)P(tweet|class)}{P(tweet)} = P(class) \prod_{word \in tweet} \frac{P(word|class)}{P(word)} \quad (1)$$

$$P(word|class) = \frac{freq_{word} + 1}{\sum_{w \in \text{words assoc. with class}} (freq_w + 1)} \quad (2)$$

$$P(word) = \sum_{class \in \{+,-\}} P(word|class)P(class)$$

In order to prevent the collapse of probabilities to zero under the frequentist approximation, we use additive smoothing with a psuedocount value of one, as shown in Equation 2.We assume a uniform prior, despite the frequencies of tweets containing positive or negative emoticons in our dataset.

Evaluation

We evaluate our trained classifier against one established technique for analyzing Twitter sentiment using the AFINN affective word list (Nielsen 2011). For further validation of our classifier, we compare to human classified examples.

AFINN

AFINN provides a list of over two thousand words associated with an affective valence or score. The scores range from -5, indicating a very negative sentiment, to 5, indicating a positive valence. In contrast to our classifier, scoring with AFINN allows analysis of magnitude of sentiment in addition to polarity. However, we compare only across classification in this paper.

Our test set contains one thousand tweets containing at least one of the emoticons in Table 1. The test set is comprised of five hundred positive examples and an equal number of negative examples. All tweets were stripped of their emoticons and punctuation, as we are interested in classification by text content. No stop words were removed.

Qualitatively, we can define ambiguity under each classification method. With our classifier, ambiguous cases result in a probability around 50%, either due to a lack of frequently associated words with a particular class or due to a balance of words associated with both classes. Similarly, under AFINN scoring, amiguity arises when few non-zero scored words are present or when the total score sums to a value close to zero. With this description in mind, we can compare the performance of the two approaches. As seen in Figure 1, AFINN scoring results in many more ambiguous cases, while our classifier appears to be more discerning, as indicated by its bimodal distribution of probabilities.

Additionally, the Bayesian classifier proved more accurate than AFINN scoring for classification. If we assign all tweets with an AFINN score greater than zero as positive and similarly all with negative score as negative, the scoring approach has a resulting accuracy of 29.2%. Comparing to similar thresholds (around 50%) in the probabalistic classification approach, we observe an accuracy of 75.1%. If an application of this approach required a greater confidence before appending a label, the probability thresholds could be restricted more severely to 5%/95% and retain an accuracy of 44.2%. The confidence/accuracy tradeoff for our approach is illustrated in Figure 2, where the probability threshold required for a label is varied.

Human Comparison

We asked seven participants to classify tweets into one of three categories: positive (3), neutral (2), or negative (1). Instructions were given to classify ambiguous cases as neutral.

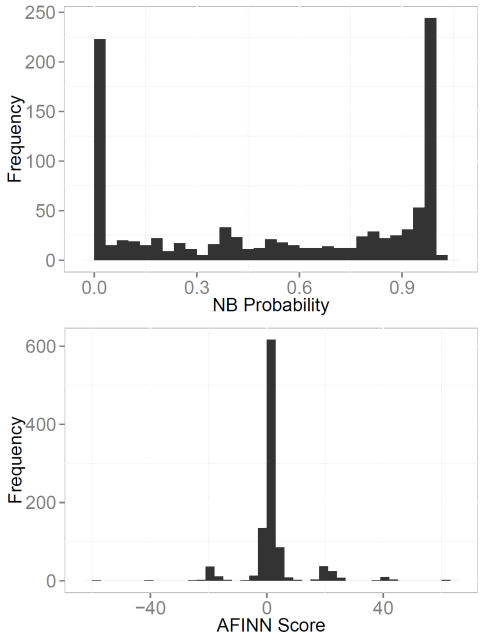

Figure 1: A comparison of the distributions of classification between AFINN and naive Bayes. AFINN scores over 60% of tweets as neutral, despite the test set being comprised entirely of tweets with positive or negative emoticons.

Emoticon Category	NB Accuracy	AFINN Accuracy
Positive	79.6%	32.2%
Negative	70.6%	26.2%
Overall	75.1%	29.2%

Table 2: Accuracy of naive Bayes and AFINN when classifying positive and negative sentiment tweets.

Due to our classifier's reliance on automatically labeled examples, via emoticons, it is only able to distinguish between the two opposing cases. However, we realize neutral cases also exist and, for this test, we classify tweets with resulting probability within the range 0.2 to 0.8 as ambiguous.

The participants were each given identical sets of one hundred tweets partitioned into three groups in proportions 30/30/40, representing those with positive, negative, and no emoticons respectively. The tweets were randomly ordered and stripped of the indicative emoticons to reduce bias in the results.

In Table 4, we provide the breakdown of our classifier's standard error against the average label given per tweet as well as the accuracy of matching the mode label given. Similarly, we compare AFINN labels against the participants' mode label. The naive Bayes classifier trained on our dataset performs comparably well to AFINN in matching human labels of tweets with positive emoticons. However, in the unlabeled and negative tweets, it performs poorly. Table 3 illustrates that both human participants and the AFINN scores prefer neutral categorization in the negative tweet and unla-

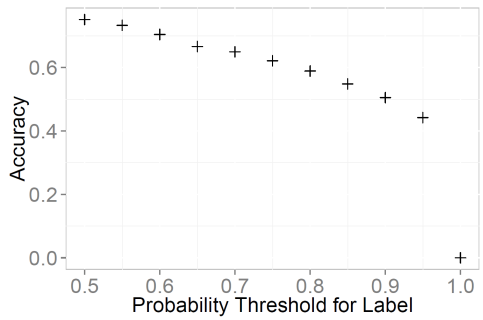

Figure 2: Accuracy tradeoff when restricting the required probability for labeling a tweet as having positive or negative sentiment.

Emoticon Category	Positive/Neutral/Negative Label Frequencies		
	Human Mode	Naive Bayes	AFINN
Positive	12/16/2	19/8/3	16/10/4
Negative	4/23/3	4/4/22	5/18/7
Unlabeled	6/31/3	9/14/17	8/26/6

Table 3: Breakdown of label frequencies given the original emoticon associations.

beled tweet categories. The NB classifier, however, rarely classifies a tweet as neutral or ambiguous due to the nature of the training data only having positive or negative examples.

Discussion

We observe that in many cases, participants formed a consensus that differed from the original category as indicated by the presence of the appropriate emoticons. Specifically, although thirty of the tweets originally had emoticons associated with negative sentiment, the participants only labeled three as such. Given that the participants received a modified version of the tweet missing such emoticons, this calls into question how much of the sentiment is contributed by emoticons. Though we do not examine the topic here, this may in part be due to sarcasm, where negative words can be offset by a positive smiley face and vice versa.

If the presence of emoticons are not a substantial part of the contributed sentiment, the human labels should be considered a gold standard evaluation, which suggests that emoticons in our dataset do not accurately distinguish sentiment classes. Further analysis on perceived sentiment and emoticon presence may be insightful.

Future Work

Comparing against AFINN illustrates how our domain-specific classifier can outperform approaches targeted at general Twitter data, but is restricted to cases where sufficient training examples can be obtained. Clearly, as our data set has no convenient indicator of neutral or ambiguous sentiment tweets, the naive Bayes classifier only performs well at discriminating between positive and negative cases. The reliance on uncertain probabilities does not reflect the true distribution of neutral sentiment examples. Training with a relevant set of neutral or ambiguous tweets may improve performance in the broader application.

Emoticon Category	NB Std. Error	NB Accuracy	AFINN Accuracy
Positive	0.801	46.7%	40.0%
Negative	1.057	16.7%	60.0%
Unlabeled	0.907	35.0%	67.5%
Overall	0.925	33.0%	57.0%

Table 4: Performance of the naive Bayes classifier and AFINN scoring against human labeling.

Additionally, a stronger evaluation between domain-specific and general classifiers would follow from collecting a similar amount of tweets on general topics, then compare performance between classifiers trained with two types of data. The comparison to AFINN is informative, yet it is not as strong a comparison as holding the classifier model constant and varying the training data.

Furthermore, more advanced classification or regression models may improve accuracy over that shown here. Particularly, larger n-gram features may perform more accurately than unigrams.

Conclusion

In summary, we present a dataset capable of training a classifier for sentiment analysis within the games domain. We demonstrate such capacity by employing naive Bayes and evaluating against both an established approach as well as human labeled examples. For further analysis or replication of these results, the data can be acquired at https://games.soe.ucsc.edu/ccs.

Acknowledgement

This work was partially supported by a grant from the Microsoft Studios University Program.

References

Barbera, P. 2014. streamr: Access to twitter streaming api via r.

Engström, C. 2004. Topic dependence in sentiment classification. *Unpublished MPhil Dissertation. University of Cambridge*.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1–12.

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

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