Twitter Sentiment Analysis: The Good the Bad and the OMG!

Efthymios Kouloumpis*
i-sieve Technologies
Athens, Greece
epistimos@i-sieve.com

Theresa Wilson*
HLT Center of Excellence
Johns Hopkins University
Baltimore, MD, USA
taw@jhu.edu

Johanna Moore
School of Informatics
University of Edinburgh
Edinburgh, UK
j.moore@ed.ac.uk

Abstract

In this paper, we investigate the utility of linguistic features for detecting the sentiment of Twitter messages. We evaluate the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in microblogging. We take a supervised approach to the problem, but leverage existing hashtags in the Twitter data for building training data.

Introduction

In the past few years, there has been a huge growth in the use of microblogging platforms such as Twitter. Spurred by that growth, companies and media organizations are increasingly seeking ways to mine Twitter for information about what people think and feel about their products and services. Companies such as Twitratr (twitratr.com), tweetfeel (www.tweetfeel.com), and Social Mention (www.socialmention.com) are just a few who advertise Twitter sentiment analysis as one of their services.

While there has been a fair amount of research on how sentiments are expressed in genres such as online reviews and news articles, how sentiments are expressed given the informal language and message-length constraints of microblogging has been much less studied. Features such as automatic part-of-speech tags and resources such as sentiment lexicons have proved useful for sentiment analysis in other domains, but will they also prove useful for sentiment analysis in Twitter? In this paper, we begin to investigate this question.

Another challenge of microblogging is the incredible breadth of topic that is covered. It is not an exaggeration to say that people tweet about anything and everything. Therefore, to be able to build systems to mine Twitter sentiment about any given topic, we need a method for quickly identifying data that can be used for training. In this paper, we explore one method for building such data: using Twitter hashtags (e.g., #bestfeeling, #epicfail, #news) to identify positive, negative, and neutral tweets to use for training three-way sentiment classifiers.

*Work performed while at the University of Edinburgh

Related Work

Sentiment analysis is a growing area of Natural Language Processing with research ranging from document level classification (Pang and Lee 2008) to learning the polarity of words and phrases (e.g., (Hatzivassiloglou and McKeown 1997; Esuli and Sebastiani 2006)). Given the character limitations on tweets, classifying the sentiment of Twitter messages is most similar to sentence-level sentiment analysis (e.g., (Yu and Hatzivassiloglou 2003; Kim and Hovy 2004)); however, the informal and specialized language used in tweets, as well as the very nature of the microblogging domain make Twitter sentiment analysis a very different task. It's an open question how well the features and techniques used on more well-formed data will transfer to the microblogging domain.

Just in the past year there have been a number of papers looking at Twitter sentiment and buzz (Jansen et al. 2009; Pak and Paroubek 2010; O'Connor et al. 2010; Tumasjan et al. 2010; Bifet and Frank 2010; Barbosa and Feng 2010; Davidov, Tsur, and Rappoport 2010). Other researchers have begun to explore the use of part-of-speech features but results remain mixed. Features common to microblogging (e.g., emoticons) are also common, but there has been little investigation into the usefulness of existing sentiment resources developed on non-microblogging data.

Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers rely on emoticons for defining their training data (Pak and Paroubek 2010; Bifet and Frank 2010). (Barbosa and Feng 2010) exploit existing Twitter sentiment sites for collecting training data. (Davidov, Tsur, and Rappoport 2010) also use hashtags for creating training data, but they limit their experiments to sentiment/non-sentiment classification, rather than 3-way polarity classification, as we do.

Data

We use three different corpora of Twitter messages in our experiments. For development and training, we use the the hashtagged data set (HASH), which we compile from the Edinburgh Twitter corpus[1], and the emoticon data set (EMOT) from http://twittersentiment.

[1] http://demeter.inf.ed.ac.uk

	Positive	Negative	Neutral	Total
HASH	31,861 (14%)	64,850 (29%)	125,859 (57%)	222,570
EMOT	230,811 (61%)	150,570 (39%)	–	381,381
ISIEVE	1,520 (38%)	200 (5%)	2,295 (57%)	4,015

Table 1: Corpus statistics

Hashtag	Frequency	Synonyms
#followfriday	226,530	#ff
#nowplaying	209,970	
#job	136,734	#tweetajob
#fb	106,814	#facebook
#musicmonday	78,585	#mm
#tinychat	56,376	
#tcot	42,110	
#quote	33,554	
#letsbehonest	32,732	#tobehonest
#omgfacts	30,042	
#fail	23,007	#epicfail
#factsaboutme	19,167	
#news	17,190	
#random	17,180	
#shoutout	16,446	

Table 2: Most frequent hashtags in the Edinburgh corpus

appspot.com. For evaluation we use a manually annotated data set produced by the iSieve Corporation[2] (ISIEVE). The number of Twitter messages and the distribution across classes is given in Table 1.

Hashtagged data set

The hashtagged data set is a subset of the Edinburgh Twitter corpus. The Edinburgh corpus contains 97 million tweets collected over a period of two months. To create the hashtagged data set, we first filter out duplicate tweets, non-English tweets, and tweets that do not contain hashtags. From the remaining set (about 4 million), we investigate the distribution of hashtags and identify what we hope will be sets of frequent hashtags that are indicative of positive, negative, and neutral messages. These hashtags are used to select the tweets that will be used for development and training.

Table 2 lists the 15 most-used hashtags in the Edinburgh corpus. In addition to the very common hashtags that are part of the Twitter folksonomy (e.g., #followfriday, #musicmonday), we find hashtags that would seem to indicate message polarity: #fail, #omgthatsotrue, #iloveitwhen, etc.

To select the final set of messages to be included in the HASH dataset, we identify all hashtags that appear at least 1,000 times in the Edinburgh corpus. From these, we selected the top hashtags that we felt would be most useful for identifying positive, negative and neutral tweets. These hashtags are given in Table 3. Messages with these hashtags were included in the final dataset, and the polarity of each message is determined by its hashtag.

[2]www.i-sieve.com

Positive	#iloveitwhen, #thingsilike, #bestfeeling, #bestfeelingever, #omgthatssotrue, #imthankfulfor, #thingsilove, #success
Negative	#fail, #epicfail, #nevertrust, #worst, #worse, #worstlies, #imtiredof, #itsnotokay, #worstfeeling, #notcute, #somethingaintright, #somethingsnotright, #ihate
Neutral	#job, #tweetajob, #omgfacts, #news, #listeningto, #lastfm, #hiring, #cnn

Table 3: Top positive, negative and neutral hashtags used to create the HASH data set

Emoticon data set

The Emoticon data set was created by Go, Bhayani, and Huang for a project at Stanford University by collecting tweets with positive ':)' and negative ':(' emoticons. Messages containing both positive and negative emoticons were omitted. They also hand-tagged a number of tweets to use for evaluation, but for our experiments, we only use their training data. This set contains 381,381 tweets, 230,811 positive and 150,570 negative. Interestingly, the majority of these messages do not contain any hashtags.

iSieve data set

The iSieve data contains approximately 4,000 tweets. It was collected and hand-annotated by the iSieve Corporation. The data in this collection was selected to be on certain topics, and the label of each tweet reflects its sentiment (positive, negative, or neutral) towards the tweet's topic. We use this data set exclusively for evaluation.

Preprocessing

Data preprocessing consists of three steps: 1) tokenization, 2) normalization, and 3) part-of-speech (POS) tagging. Emoticons and abbreviations (e.g., *OMG, WTF, BRB*) are identified as part of the tokenization process and treated as individual tokens. For the normalization process, the presence of abbreviations within a tweet is noted and then abbreviations are replaced by their actual meaning (e.g., *BRB* $-$ $>$ *be right back*). We also identify informal intensifiers such as all-caps (e.g., *I LOVE this show!!!* and character repetitions (e.g., *I've got a mortgage!! happyyyyyy"*), note their presence in the tweet. All-caps words are made into lower case, and instances of repeated charaters are replaced by a single character. Finally, the presence of any special Twitter tokens is noted (e.g., #hashtags, usertags, and URLs) and placeholders indicating the token type are substituted. Our hope is that this normalization improves the performance of the POS tagger, which is the last preprocessing step.

Features

We use a variety of features for our classification experiments. For the baseline, we use unigrams and bigrams. We also include features typically used in sentiment analysis, namely features representing information from a sentiment lexicon and POS features. Finally, we include features to capture some of the more domain-specific language of microblogging.

n-gram features

To identify a set of useful n-grams, we first remove stopwords. We then perform rudimentary negation detection by attaching the the word *not* to a word that preceeds or follows a negation term. This has proved useful in previous work (Pak and Paroubek 2010). Finally, all unigrams and bigrams are identified in the training data and ranked according to their information gain, measured using Chi-squared. For our experiments, we use the top 1,000 n-grams in a bag-of-words fashion.[3]

Lexicon features

Words listed the MPQA subjectivity lexicon (Wilson, Wiebe, and Hoffmann 2009) are tagged with their prior polarity: positive, negative, or neutral. We create three features based on the presence of any words from the lexicon.

Part-of-speech features

For each tweet, we have features for counts of the number of verbs, adverbs, adjectives, nouns, and any other parts of speech.

Micro-blogging features

We create binary features that capture the presence of positive, negative, and neutral emoticons and abbreviations and the presence of intensifiers (e.g., all-caps and character repetitions). For the emoticons and abbreviations, we use the Internet Lingo Dictionary (Wasden 2006) and various internet slang dictionaries available online.

Experiments and Results

Our goal for these experiments is two-fold. First, we want to evaluate whether our training data with labels derived from hashtags and emoticons is useful for training sentiment classifiers for Twitter. Second, we want to evaluate the effectiveness of the features from section for sentiment analysis in Twitter data. How useful is the sentiment lexicon developed for formal text on the short and informal tweets? How much gain do we get from the domain-specific features?

For our first set of experiments we use the HASH and EMOT data sets. We start by randomly sampling 10% of the HASH data to use as a validation set. This validation set is used for n-gram feature selection and for parameter tuning. The remainder of the HASH data is used for training. To train a classifier, we sample 22,247[4] tweets from the training data and use this data to train AdaBoost.MH (Schapire and Singer 2000) models with 500 rounds of boosting.[5][6] We repeat this process ten times and average the performance of the models.

[3]The number n-grams to include as features was determined empirically using the training data.

[4]This is equivalent to 10% of the training data. We experimented with different sample sizes for training the classifier, and this gave the best results based on the validation data.

[5]The rounds of boosting was determined empirically using the validation set.

[6]We also experimented with SVMs, which gave similar trends, but lower results overall.

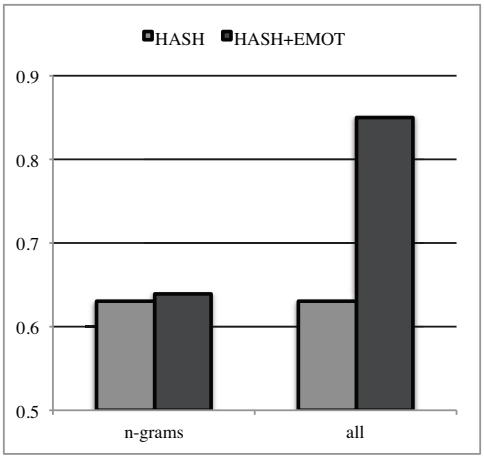

Figure 1: Average F-measure on the validation set over models trained on the HASH and HASH+EMOT data

Because the EMOT data set has no neutral data and our experiments involve 3-way classification, it is not included in the initial experiments. Instead, we explore whether it is useful to use the EMOT data to expand the HASH data and improve sentiment classification. 19,000 messages from the EMOT data set, divided equally between positive and negative, are randomly selected and added to the HASH data and the experiments are repeated.

To get a sense for an upper-bound on the performance we can expect for the HASH-trained models and whether including the EMOT data may yield improvements, we first check the results of the models on the validation set. Figure 1 shows the average F-measure for the n-gram baseline and all the features on the HASH and the HASH+EMOT data. On this data, adding the EMOT data to the training does lead to improvements, particularly when all the features are used.

Turning to the test data, we evaluate the models trained on the HASH and the HASH+EMOT data on the ISIEVE data set. Figure 2 shows the average F-measure for the baseline and four combinations of features: n-grams and lexicon features (n-gram+lex), n-grams and part-of-speech features (n-gram+POS), n-grams, lexicon features and microblogging features (n-grams+lex+twit), and finally all the features combined. Figure 3 shows the accuracy for these same experiments.

Interestingly, the best performance on the evaluation data comes from using the n-grams together with the lexicon features and the microblogging features. Including the part-of-speech features actually gives a drop in performance. Whether this is due to the accuracy of the POS tagger on the tweets or whether POS tags are less useful on microblogging data will require further investigation.

Also, while including the EMOT data for training gives a nice improvement in performance in the absense of microblogging features, once the microblogging features are included, the improvements drop or disappear. The best results on the evaluation data comes from the n-grams, lexical and Twitter features trained on the hashtagged data alone.

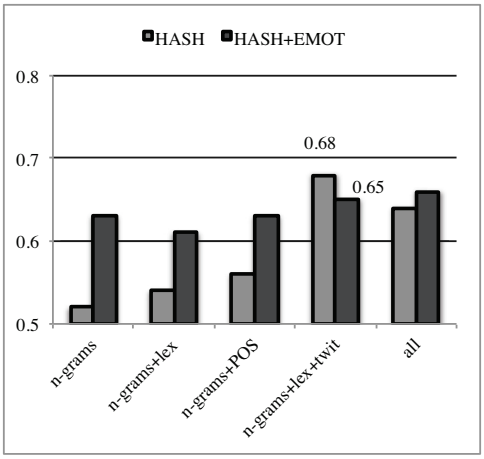

Figure 2: Average F-measure on the test set over models trained on the HASH and HASH+EMOT data

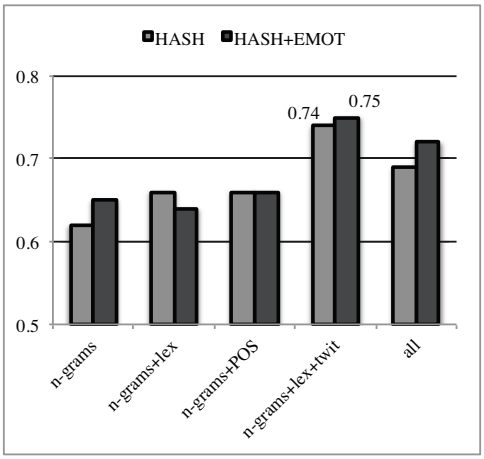

Figure 3: Average accuracy on the test set over models trained on the HASH and HASH+EMOT data

Conclusions

Our experiments on twitter sentiment analysis show that part-of-speech features may not be useful for sentiment analysis in the microblogging domain. More research is needed to determine whether the POS features are just of poor quality due to the results of the tagger or whether POS features are just less useful for sentiment analysis in this domain. Features from an existing sentiment lexicon were somewhat useful in conjunction with microblogging features, but the microblogging features (i.e., the presence of intensifiers and positive/negative/neutral emoticons and abbreviations) were clearly the most useful.

Using hashtags to collect training data did prove useful, as did using data collected based on positive and negative emoticons. However, which method produces the better training data and whether the two sources of training data are complementary may depend on the type of features used. Our experiments show that when microblogging features are included, the benefit of emoticon training data is lessened.

References

Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proc. of Coling*.

Bifet, A., and Frank, E. 2010. Sentiment knowledge discovery in twitter streaming data. In *Proc. of 13th International Conference on Discovery Science*.

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of Coling*.

Esuli, A., and Sebastiani, F. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*.

Hatzivassiloglou, V., and McKeown, K. 1997. Predicting the semantic orientation of adjectives. In *Proc. of ACL*.

Jansen, B. J.; Zhang, M.; Sobel, K.; and Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology* 60(11):2169–2188.

Kim, S.-M., and Hovy, E. 2004. Determining the sentiment of opinions. In *Proceedings of Coling*.

O'Connor, B.; Balasubramanyan, R.; Routledge, B.; and Smith, N. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of ICWSM*.

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

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

Schapire, R. E., and Singer, Y. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning* 39(2/3):135–168.

Tumasjan, A.; Sprenger, T. O.; Sandner, P.; and Welpe, I. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of ICWSM*.

Wasden, L. 2006. Internet Lingo Dictionary: A Parent's Guide to Codes Used in Chat Rooms, Instant Messaging, Text Messaging, and Blogs. Technical report, Idaho Office of the Attorney General.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics* 35(3):399–433.

Yu, H., and Hatzivassiloglou, V. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proc. of EMNLP*.