Hashtag Sentiment Analysis using Tweets for the Ternary Classification

Jael Jayakumar , Indira Tamilselvam , Blessy Selvam

Department of Computer Science, University College of Engineering
Trichy, India

Abstract

To analyze the utility of the linguistic features for detecting the sentiment of the given Twitter messages. Evaluating the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in micro blogging. In addition to,apply for a supervised approach to the problem, but to make use of the 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 micro blogging platforms such as Twitter. By its growth, it has spurred the companies and media organizations to larger extent and are increasingly seeking ways to mine Twitter for more information about the people thoughts and how they feel about their products and services. Twitter is one among the many services such as tweet feel , and Social Mention are just a few who advertise Twitter sentiment analysis.

There has been a fair amount of research on how sentiments are expressed in genres such as the online reviews and news articles. How sentiments are expressed are given in the informal language and message-length constraints of blogging sites has been much less studied. Some of the Features such as automatic part-of-speech tags and resources such as sentiment lexicons have proved useful for sentiment analysis in other domains.

Another challenge of micro blogging is that, it is not an exaggeration to include that people tweet about anything and everything. Therefore, to be able to build systems to mine sentiment about any given topic, in need of a method for quickly identifying data that can be held for building the training data set. To invent the method of twitter sentiment using Twitter hash- tags (e.g., #njoymentfun, #epicmatch, #info) so as to identify positive, negative, and neutral tweets . used to train the obtained data tweets. There are three way sentiment classifiers.

Related Work

The Sentiment analysis is a growing area of NLP(Natural Language Processing) learning the polarity of words and phrases (e.g Esuli and Sebastiani 2006). Classifying the sentiment of Twitter messages and limiting the number of words for the tweets(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 micro blogging domain.

In reference to the Agarwal et al, it is defined the task of mining sentiment from the twitter sentiment , in a 3-way task of classifying sentiment that can be given as ternary classes. Three types models have been used for this experimental analysis: unigram model, a feature based model and a tree kernel based model. For the tree kernel based model they designed a new tree representation for tweets. The feature based model that uses 100 features and the unigram model uses over 10,000 features. The combination of polarity of words and parts-of-speech tags forms the most important part for the classification task. The tree kernel based model outperformed the other two.

Researchers have also begun to investigate various ways for 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 and non-sentiment classification, rather than 3-way polarity classification, as we do.

Data

Making use of three different frame of Twitter messages in the experiments. For the development and training process, the hash tagged data set (HASH), that can be compiled from the Edinburgh Twitter corpus , and the emoticon data set (EMOT) from Twitter sentiment analysis ,twitter data set.

	Positive	Negative	Neutral	Total
HASH	31,861	64,850	125,859	222,57
EMO	230,811	150,570	–	381,38
ISIEV	1,520	200 (5%)	2,295	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

Positive	#iloveitwhen, #thingsilike, #bestfeel- ing, #bestfeelingever, #omgthatssotrue
Negative	#fail, #epicfail, #nevertrust, #worst, #worse, #worstlies, #imtiredof, #itsno- tokay, #worstfeeling, #notcute, #somethin gaintright
Neutral	#job, #tweetajob, #omgfacts, #news, #lis- teningto, #lastfm, #hiring, #cnn

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

For the evaluation, it should make use of a manually annotated data set produced by the iSieve Corporation (ISIEVE). The number of Tweets messages and all of the distributed tweets across classes are given.

Dataset:

HASHTAGGED

The hashtagged data set is a subset of the Edinburgh Twitter corpus. The reference and related works contains more than 97 million tweets collected over a period . To create the hash- tagged data set, first filter out duplicate tweets, non- English tweets, and tweets that do not contain hashtags. From the remaining set (about 4 million), investigate the distribution of hashtags and identify what the 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, #musicmon- day), 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 it 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.

EMOTICON

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 the experiments, to 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

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) toward

PREPROCESSING

Data preprocessing takes place in the following 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 s the tweet's topic. This data set is exclusively used for evaluation.

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. The hope is that this normalization improves the performance of the POS tagger, which is the last preprocessing step.

Features

To make use a variety of features for the classification experi- ments. To use unigrams and bigrams. Also include features typically used in sentiment analysis, namely features representing information from a sentiment lexicon and POS features. To include the features to capture some of the more domain-specific tweets.

n-gram features

The foremost objective is to remove all the stop words from the tweets to process it into the analysis. The proceed the details by neglecting the "not" preceding or succeeding the words attached to it. then finally make use of the unigrams or bigrams for ranking the obtained tweets lexicons.. To use the top 1,000 n-grams in a bag- of-words fashion.

Lexicon features

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

Part-of-speech features

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

Micro-blogging features

To 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 rep- etitions). For the emoticons and abbreviations, and use the Internet Lingo Dictionary (Wasden 2006) and various internet slang dictionaries available online.

Experiments and Results

The main reachout for this experiment is two-fold. First, to evaluate whether the training data with labels derived from hashtags and emoticons is useful for training sentiment classifiers for Twitter. Second, 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 does it get from the domain-specific features?

For first set of experiments , use the HASH and EMOT data sets. 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 tweets from the training data and use this data to train AdaBoost.MH (Schapire and Singer 2000) models with 500 rounds of boosting. We repeat this process ten times and average the performance of the models.

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

This is equivalent to 10% of the training data. It experi- mented with different sample sizes for training the classifier, and this gave the best results based on the validation data.

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

To also be experimented with SVMs, which gave similar trends, but lower results overall.

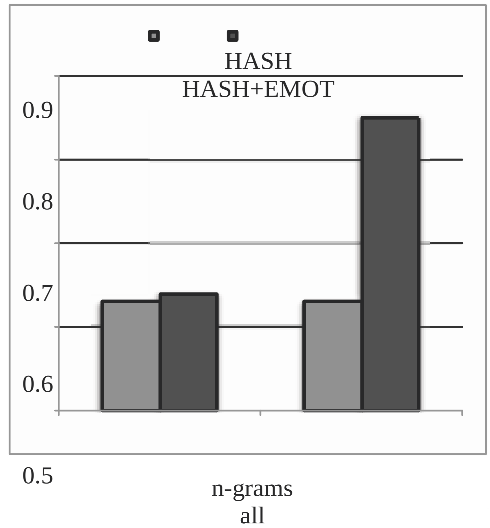

Figure 1: Average F-measure on the validation set over models trained on the HASH and HASH+EMOT data

Because the EMOT data set has no neutral data and experiments involve 3-way classification, it is not included in the initial experiments. Instead, to 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 , can expect for the HASH-trained models and whether including the EMOT data may yield improvements, to 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, 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 base- line 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 microblog- ging features (n-grams+lex+twit), and finally all the features combined. Figure 3 shows the accuracy for these same experiments.

The evaluation of the data tweets interestingly comes from the features of it. It also insist the part- of-speech features that actually gives a drop in performance. If this is due to the accuracy of the POS tagger on the tweets or whether Part of speech tags are less useful on online blogging site 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 re- sults on the evaluation data comes from the n-grams, lexical and Twitter features trained on the hashtagged data alone.

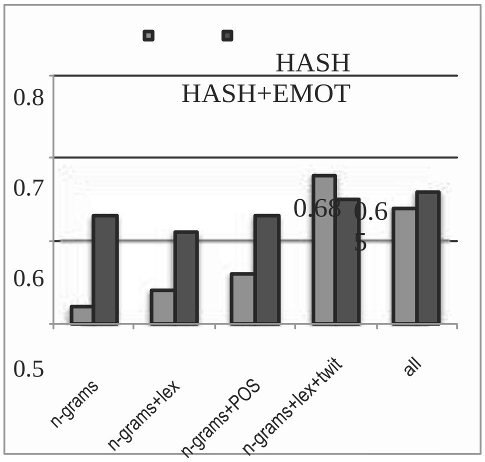

Figure 2: Average F-measure on the test set over models trained on the HASH and HASH+EMOT data

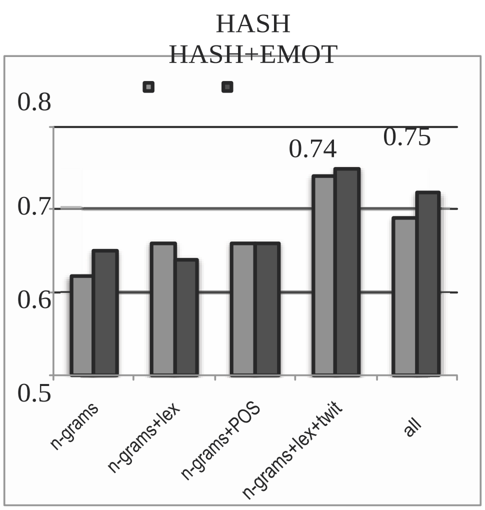

Figure 3: Average accuracy on the test set over models trained on the HASH and HASH+EMOT data

Conclusion

All 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,the collected training data set did prove the usefulness, as if did using the data collected based on positive and negative emoticons included, some of the emoticon data set benefits have been lowered.

References

1. Graduate School of Science and Technology, Keio University: A Pattern based approach for multiclass sentiment analysis in twitter [Accessed 7 Aug 2017]

2. Vidhya Content Team. 2015. Quick Guide: Steps To Perform Text Data Cleaning in Python /2015/06/quick-guide text-data-cleanin Goodfellow, [Accessed 20 May 2017].

3. M.Bouazizi and T.Ohtsuki,"Sentiment analysis:from binary to multiclass classification", in Proc.IEEE ICC,May 2016,pp.16.

4. L.Oneto, F. Bisio, e.Cambria, and D.Anguita, "Statistical learning theory and ELM for bid social data analysis ", IEEE Comp.Int.Mab., pp.45-55, 2016

5. A.Agarwal, B.Xie,I.Vovsha, O.Rambow ,R.Passonneau ,"Sentiment Analysis of twitter Data", In Proceedings of the ACL 2011 . The experimental analysis based on the use of three type classification.

6. 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.

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

8. 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 Tech- nology 60(11):2169–2188.

9. 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.

10. D.Davidov, O.Tsur and A. Rapport, "Supervised recognition of sarcastic sentences in twitter",in Proc. 14th Conf Comput.Natural Lang.Leran ., Jul .2010,.pp107-116.

11. Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval2(1-2):1–135.

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

13. 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.

14. 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.

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