Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 61 (2013) 16765-16768

ISSN: 2229-712X

A case study of Sentiment Analysis of Twitter Data

Rakesh Chandra Balabantaray and Monalisa Swain

Department of Computer Science and Engineering, IIIT Bhubaneswar, Bhubaneswar-751003, India.

ARTICLE INFO

Article history:

Received: 3 June 2013;

Received in revised form:

24 July 2013;

Accepted: 1 August 2013;

Keywords

Twitter,

Sentiment analysis,

Sentiment classification,

SVM etc.

ABSTRACT

The purpose of this research is to classify sentiment of Twitter messages. We present a novel approach for automatically classifying the sentiment of Twitter messages. These messages are classified as either positive or negative. This is useful for consumers who want to do research on the user sentiment for products before purchase, or companies that want to monitor the public sentiment of their brands. Known supervised learning algorithms as support vector machines and Maximum Entropy are used to create a prediction model. Before the prediction model created, the data has pre-processed by using some properties of data such as username, hyperlink, and multiple occurrence of character in order to reduce feature space and achieve high accuracy.

Introduction

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [1]. It can be performed using NLP, statistics, or machine learning methods. Sentiment Analysis has any number of applications which are important for both the individuals and organizations. In the past whenever a customer was going to make purchase of any product he/she was asking for opinions from friends, families and neighbour about that product whether that is good or bad, whether it is worth to buy. Similarly whenever an organisation wanted to know opinions of the general public about its products and services, it was conducting survey. But now a days with the rich availability of social media content customer can easily get the review of different user about the product within a short time period from that social media content instead of taking opinions from the friends or family and also the organisations can get the product feedback, suggestion, area of problem, what can be done for better improvement according to user need from this social media content.

Now-a-days, various social networking sites like Twitter, Facebook, MySpace, YouTube have gained so much popularity. In this paper we have taken tweets of the twitter for sentiment analysis and classification. "tweets" is nothing but the status message created by user. In this paper, we apply common machine learning techniques such as SVM, Maximum Entropy Model to extract sentiment as positive or negative from a tweet.

Characteristics of Tweets

San Antonio-based market-research firm Pear Analytics analyzed 2,000 tweets (originating from the US and in English) over a two-week period in August 2009 from 11:00 am to 5:00 pm (CST) and separated them into six categories: [2]

- Pointless babble – 40%
- Conversational – 38%
- Pass-along value – 9%
- Self-promotion – 6%
- Spam – 4%
- News – 4%

Social networking researcher Danah Boyd responded to the Pear Analytics survey by arguing that what the Pear researchers labelled "pointless babble" is better characterized as "social grooming" and/or "peripheral awareness" (which she explains as persons "want[ing] to know what the people around them are thinking and doing and feeling, even when co-presence isn't viable").[3]

Tweets have many unique features

I. This text based message is limited within 140 characters.

II. User's are posting message from different media, including cell phones .Due to shorter length of the message people are using acronyms, emoticons and other characters that convey special meanings which causes the frequency of misspelling word more often in comparison with other domains and also makes it informal.

III. The amount of data available in twitter is more and millions of tweets can be collected with the use of Twitter API.

IV. Users are posting message about variety of topics instead of staying within limited specified topic like other sites.

Some of the basic formats used in tweets are

I. **Hash tags (#):** A hashtag (#) is a way to aggregate tweets that are appended with a hashtag. When user tweets and want that message to be part of a larger conversation beyond his/her followers, adds a relevant hashtag to the end of the message, and the message automatically reach anyone who is monitoring the same hashtag.

II. **@username:** a response to an existing tweet automatically begins with @username (the username of the person to whom you are replying).

III. **Retweet or "RT":** "RT" is an indication of repeat of someone else's earlier tweet.

IV. **Emoticons:** *emoticons* are glyphs designed to add emotion to plain text messages. Just as simple punctuation can convey surprise ! or pose a question ? , emoticons can convey happiness and joy :-), sadness :-(.

Tele:

E-mail addresses: rakeshbray@gmail.com

Data collection and Pre-processing

In order to train a classifier, supervised learning usually requires hand-labelled training data. With the large range of topics discussed on Twitter, it would be very difficult to manually collect enough data to train a sentiment classifier for tweets. So we have used the tweet corpus from Sentiment140. This training data has already post-processed with the following filters [4]:

Table: 1

Emoticons mapped to :)	Emoticons mapped to :(
:)	:(
:-)	:-(
:)	: (
:D	
=)	

I. Emoticons listed in Table 1 are stripped off. This is important for training purposes. If the emoticons are not stripped off, then the MaxEnt and SVM classifiers tend to put a large amount of weight on the emoticons, which hurts accuracy.

II. Any tweet containing both positive and negative emoticons is removed. This may happen if a tweet contains two subjects. Here is an example of a tweet with this property: Target orientation :(But it is my birthday today :). These tweets are removed because we do not want positive features marked as part of a negative tweet, or negative features marked as part of a positive tweet.

III. Retweets are removed. Retweeting is the process of copying another user's tweet and posting to another account. This usually happens if a user likes another user's tweet. Retweets are commonly abbreviated with "RT." For example, consider the following tweet: Awesome! RT @rupertgrintnet Harry Potter Marks Place in Film History http://bit.ly/Eusxi :). In this case, the user is rebroadcasting rupertgrintnet's tweet and adding the comment Awesome!. Any tweet with RT is removed from the training data to avoid giving a particular tweet extra weight in the training data.

IV. Tweets with ":P" are removed. At the time of this writing, the Twitter API has an issue in which tweets with ":P" are returned for the query ":(". These tweets are removed because ":P" usually does not imply a negative sentiment.

V. Repeated tweets are removed. Occasionally, the Twitter API returns duplicate tweets. The scraper compares a tweet to the last 100 tweets. If it matches any, then it discards the tweet. Similar to retweets, duplicates are removed to avoid putting extra weight on any particular tweet.

We have taken the training set from tweet corpus which contains 800,000 positive tweets, and 800,000 negative tweets for both training and testing classifiers. We have also pre-processed this data set for reductions of feature space.

Some of the properties that have taken for reduction are given below

I. **Usernames:** a response to an existing tweet automatically begins with @username (e.g. @monalisa). Here we have replaced all the token having @ symbol at the beginning with USRNAME.(e.g. @monalisa --> USRNAME)

II. **Hyperlink:** users are using hyperlink in their message very often, so we have replaced all these hyperlink with HYPERLINK (e.g. http://ping.fm/c2hPP --> HYPERLINK)

III. **Multiple Occurrence of Character:** In tweets most of the words are not in correct form people are writing according to their own interest. for example different from of "hello" written in tweet messages are "hello", "hellooooo", "hellllllo", "helllloooooooooo" and many more .so we have replaced the more than one conjugative occurrence of a single character with a single occurrence of that character.

So for example after replacing the above word becomes as below

hello ---> helo
hellooooo ---> helo
hellllllo ---> helo
Helllloooooooooo---> Helo

We have taken different combination of properties reduction and prepared training and testing data set for each combination separately and tested with the classifiers.

Classifiers

We have tested different classifiers such as maximum entropy, support vector machines and a statistical approach.

Support Vector Machines

SVMs are a machine learning classification technique which uses a function called a kernel to map a space of data points in which the data is not linearly separable onto a new space in which it is, with allowances for erroneous classification. Support Vector Machines is another popular classification technique [6]. We use the SVM light [7] software with a linear kernel.

Maximum Entropy

The idea behind Maximum Entropy models is that one should prefer the most uniform models that satisfy a given constraint [5]. MaxEnt models are feature-based models. In a two class scenario, it is the same as using logistic regression to find a distribution over the classes. MaxEnt makes no independence assumptions for its features. This means we can add features like bigrams and phrases to MaxEnt without worrying about features overlapping. The model is represented by the following:

$$P_{ME}(c|d,\lambda) = \frac{\exp[\Sigma_i \lambda_i f_i(c,d)]}{\Sigma_{c'} \exp[\Sigma_i \lambda_i f_i(c,d)]}$$

In this formula, c is the class, d is the tweet, and λ is a weight vector. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class. The weight vector is found by numerical optimization of the lambdas so as to maximize the conditional probability.

We use the Stanford Classifier to perform MaxEnt classification.

Our statistical Approach

Twittratr is a website that performs sentiment analysis on tweets. Their approach is to use a list of positive and negative keywords. As a baseline, we use Twittratr's list of keywords, which is publicly available. This list consists of 174 positive words and 185 negative words. For each tweet, we count the number of negative keywords and positive keywords that appear. If the count of positive keyword in a tweet is more than negative keyword we have consider that as positive tweet otherwise if the count of negative keyword in a tweet is more than positive keyword we have consider that as negative tweet and if both negative and positive count is same we have consider that as positive tweet . If both the count for negative and positive keyword in a tweet is zero we have considered that as neutral tweet.

Experimental Set up

In machine learning [8][9], there are basically two types of learning methods called as Supervised learning [8] and

unsupervised learning [8]. In supervised learning, the developer has to provide learning data to the system in order to train the system. In unsupervised learning, the system itself learns patterns from the data.we have taken supervised learning method for classification.

The features need to be extracted before the classification can start. The filtered dataset from the pre-processing is used to extract features, we have created feature vector based on word unigrams. Each feature is a single word found in a tweet. If the feature is present the value is 1 and if the feature is absent the value is 0. As stated above we have prepared training and testing data set by taking different combination of properties reduction as given below :

From the data set provided by Sentiment140, the training set which contains 800,000 positive tweets and 800,000 negative tweets is taken for both training and testing.

I. **Set-1:** From the 800,000 positive tweets, and 800,000 negative tweets 799,000 positive tweets, and 799,000 negative tweets are taken for training without reducing any properties and 1,000 positive tweets, and 1,000 negative tweets are taken for testing without reducing any properties.

II. **Set-2:** From the 800,000 positive tweets, and 800,000 negative tweets 799,000 positive tweets, and 799,000 negative tweets are taken for training and 1,000 positive tweets, and 1,000 negative tweets are taken for testing after filtered by Usernames properties reduction.

III. **Set-3:** From the 800,000 positive tweets, and 800,000 negative tweets 799,000 positive tweets, and 799,000 negative tweets are taken for training and 1,000 positive tweets, and 1,000 negative tweets are taken for testing after filtered by both Usernames and Hyperlink properties reduction.

IV. **Set-4:** From the 800,000 positive tweets, and 800,000 negative tweets 799,000 positive tweets, and 799,000 negative tweets are taken for training and 1,000 positive tweets, and 1,000 negative tweets are taken for testing after filtered by both Usernames and Multiple Occurrence of Character properties reduction.

V. **Set-5:** From the 800,000 positive tweets, and 800,000 negative tweets 799,000 positive tweets, and 799,000 negative tweets are taken for training and 1,000 positive tweets, and 1,000 negative tweets are taken for testing after filtered by filtered by all the three properties, Usernames and Multiple Occurrence of Character and Hyperlink properties reduction.

These five set of training and testing data set is taken for training and testing maximum entropy, support vector machines. For Our statistical Approach we have taken the test data set from the data set provided by Sentiment140. we haven't taken the data set form training data set as in case of SVM and Maximum entropy , this is because in the training data set Emoticons are stripped off but the list of positive and negative keywords provided by Twittratr has already contains Emoticons. So in order to get the correct measure of our system we have used test data set from which Emoticons are not stripped off.

Results

In this study machine learning approach has performed better than our statistical approach. In our statistical approach we got the accuracy of 57%, but in case of MaxEnt and SVM for the different set of training and testing data as stated above we got good accuracy as mentioned in Table 2.

Table: 2

Data Set	Our statistical approach	MaxEnt	SVM
Test data set by Sentiment140	57%	N/A	N/A
Set-1	N/A	80.90%	81.60%
Set-2	N/A	82.12%	82.35%
Set-3	N/A	82.10%	82.11%
Set-4	N/A	81.82%	**82.51%**
Set-5	N/A	81.57%	82.41%

From the above it is concluded that the accuracy of SVM in case of Set-4 i.e. while system is trained and tested with dataset which is filter by both Usernames and Multiple Occurrence of Character properties reduction is more than all other results.

Conclusion

For sentiment classification, Twitter offers an entirely different challenge, which is largely created by the nonstandard and informal language posted by Twitter users. We implemented SVM and Maximum Entropy model to classify tweets and also implemented our statistical approach. We found that our SVM classifiers worked better than the Maximum Entropy model and much better than our statistical approach. We are trying to improve the result further by exploring richer linguistic analysis like Pos tag, Parsing, Semantic Analysis and Topic modeling.

Reference

[1] Bing Liu. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers ,May 2012**.**

[2] Ryan Kelly. Twitter Study – August 2009. Pear Analytics.

[3] Danah Boyd (August 16, 2009). "Twitter: "pointless babble" or peripheral awareness + social grooming?" Retrieved September 19, 2009.

[4] A. Go, R. Bhayani, L.Huang. Twitter Sentiment Classification Using Distant Supervision. Stanford University, Technical Paper, 2009.

[5] K. Nigam, J. La®erty, A. Mccallum. Using maximum entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61-67, 1999.

[6] N. Cristianini, J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, March 2000.

[7] T. Joachims. Making large-scale support vector machine learning practical. In B. SchÄolkopf, C. J. C. Burges, and A. J. Smola, editors, Advances in kernel methods: support vector learning, pages 169-184. MIT Press, Cambridge, MA, USA, 1999.

[8] N. J. Nilsson. Introduction TO MACHINE LEARNING. (Online book) http://ai.stanford.edu/people/nilsson/mlbook.html

[9] J.,. K. M. Han, Data Mining: Concepts and Techniques, 2nd ed. 2006.

[10] Akshi Kumar and Teeja Mary Sebastian. "Sentiment Analysis on Twitter" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012.

[11] Alec Go and Richa Bhayani. Exploiting the Unique Characteristics of Tweets for Sentiment Analysis. Technical report, Stanford University, 2010.

[12] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau. Sentiment Analysis of Twitter Data. In Proceedings of the ACL 2011 Workshop on Languages in Social Media, 2011, pp. 30–38.

[13] R. Parikh, M. Movassate. Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques. CS224N Final Report, 2009.

[14] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86, 2002.

[15] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau. Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011 Workshop on Languages in Social Media, 2011, pp. 30–38.