Twitter as a Corpus for Sentiment Analysis and Opinion Mining

Abstract

The paper discusses how to sentimentally analyze Twitter posts and classify them to one of three classes: positive sentiment, negative sentiment, and objectively neutral. The researchers show their three-step process, from corpus collection, through linguistic analysis of the corpus, and finally to corpus classification. Moreover, the used data was evaluated and the achieved results were prompted through a discussion. And eventually, the future work was stated.

Introduction

Microblogging platforms have become more prominent than ever nowadays. A gigantic number of users post millions of microblogging posts every day. Thus, data generated from microblogging platforms is a rich source for analysis and exploring different sentiment patterns. Data generated varies from expressing personal opinions regarding a movie, a book, or an electronic device, to official statements made by well-known corporations and presidents.

Pak and Paroubek -the researchers of the discussed paper- decided to use Twitter over other microblogging platforms, such as Facebook and Tumblr, for analysis after finding out about the following advantages: huge number of plain-text-written posts compared to Facebook and Tumblr, the diversity in terms of culture and ethnicity in Twitter was a key advantage, and also the wide range of different types of users (celebrities, companies, presidents, etc).

Definitions

Here are some technical terminologies that may be mentioned later on. First, the word corpus, which is simply a collection of plain text. Here it is referred to it as “Twitter as a corpus” which means that the collection can be thought of as a group of Twitter posts (tweets). Second, n-gram, the word gram is meaningful when a prefix is added to it. This prefix can be, for example, “uni” , “bi”, “tri”, which indicates that a “unigram” is one word, “bigram” constitutes of two words and “trigram” as three words. Third, POS (Part-of-Speech) tagging, which means that given a sentence, each word is identified and given a tag. This tag can be a verb. Noun. preposition, etc.

Motivation

Human beings can easily recognize whether a post of plain text is subjective, in terms of positive and negative emotions, or whether it is just an objective type of post. However, it is more challenging when such an observation of distinction in sentiment is done by a machine.
Since manufacturing companies are looking forward to gaining their customers’ trust and satisfaction, politicians are trying to analyze which age-group they should target, and celebrities seeking more fans and followers, an automation of such a process of sentiment analysis through a machine has never been as important as it is now. Moreover, it is more cost-efficient, quicker, and more efficient in terms of performance in terms of repetitive tasks, since a human is subject to tiresome and boredom, while a machine can work maintaining the same level as long as there is enough power supply. The aforementioned factors are key drivers for such an automation of the sentiment analysis and opinion mining process.
So, (Pak and Paroubek)’s aim was to develop a platform that is able to classify Twitter posts into one of three main classes: positive sentiment, negative sentiment, objective neutral.

State of the Art

Pak and Paroubek started by inspecting the work done by (Pang and Lee, 2008) which was an overview of the approaches and techniques used in

the sentiment analysis process and what fields had been analyzed through such techniques. Weblogs were inspected by researchers a few times and even less were the microblogging platforms inspected by researchers. From this point on, they decided to review the work done by (Yang et al., 2008) who were those among the few who decided to tackle the problem of sentiment analysis in weblogs. (Yang et al., 2008) collected data based on emoticons (emotion icons) since these emoticons are used as indicators of user's mood. Then, SVM (Support Vector Machine) and CRF (Conditional Random Field) were used in the classification of such weblogs posts into one of the sentiment classes (positive or negative). The main challenge that was present in such research was how to get the overall sentiment of a large weblog post. Many strategies were tested, and they found out that the last sentence in a weblog post is always an overall indicator of the mood of the user in such a post.

The last research reviewed by Pack and Paroubek was the one done by (Go et al., 2009). And this research Twitter in specific was being analyzed as a rare example of microblogging platforms. Twitter posts were collected using emoticons as well, like (Go et al., 2009). The classification was done in terms of two main classes: positive and negative. Naive Bayes classifier was used in the classification process (81% accuracy). However, the main problem that the researchers were not able to solve was that the introduction of the objective neutral class significantly degraded the accuracy.

Implementation

The implementation phase consists of three main sub-phases: corpus collection, corpus linguistic analysis, and finally corpus classification.

Corpus Collection

Twitter’s API was used to support the researchers with posts expressing positive and negative sentiments. Smiley faces such as: “:-)”, “:D”, etc. were indicators for positive sentiment posts. Sad faces such as: “:(“, “:-(“,

etc. were indicators of negative sentiment posts. Neutral objective posts were collected from well-known newspapers accounts such as “New York Times”, “Washington Post”, etc. Twitter’s limitation that a user is only allowed to write up to 140 characters was an advantage for the researcher since they considered a presence of an emoticon as an overall sentiment of the whole post.

Corpus Linguistic Analysis

The main aim of such a sub-phase as linguistic analysis for the collected corpus is to be able to generalize the results obtained from the technique deployed in this research paper.

First, the distribution of the frequencies of the words was analyzed. It was found that it follows Zipf’s law which applies to English language in general. What Zipf’s law states is that the most frequent word in English language (the article “the”) happens as twice as many as the second most frequent word in the English language (the preposition “of”) and three times as many as the third most frequent word (the article “a”). The same statistics were also present in the corpus collected which consists of 300,000 Twitter posts.

Second, the “TreeTagger” software was used to show the distribution of POS tagging. For example, given a sentence such as “TreeTagger is easy to use”, the software’s output states that that word “TreeTagger” is a noun, the word “is” is a verb, etc. To statistically differentiate the POS tagging distribution between different sets, the following equation was used to compare between subjective (positive and negative) posts and objective ones.

$$P^{T}_{1,2} = (N^{T}_{1} - N^{T}_{2}) / (N^{T}_{1} + N^{T}_{2})$$

Where N^{T}_{1} and N^{T}_{2} are the counters of the tags frequencies in subjective and objective sets respectively.

It was observed that more proper nouns were used in case of objective posts. However, personal pronouns were used more often in case of

subjective posts. Users of objective texts use verbs in third person. On the other hand, users of subjective texts (whether positive or negative) use verbs in first or second person forms. Results considering the aforementioned two sets can be viewed in the following graph.

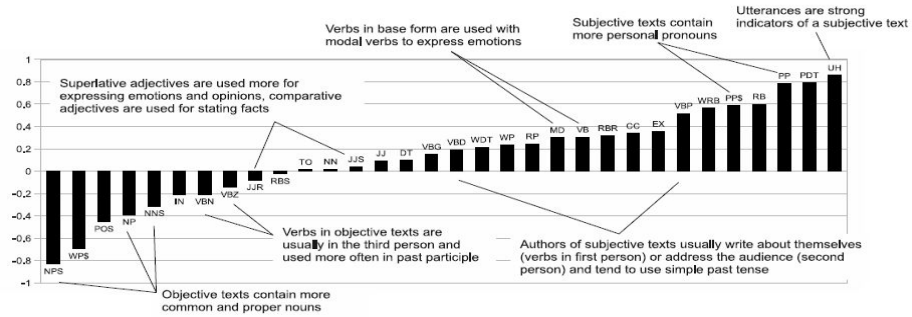


Figure 2: P^T values for objective vs. subjective

Moreover, Pak and Paroubek decided to expand their linguistic analysis process to include the difference in POS tagging between positive and negative subjective texts. It was found that superlative adverbs are more used in case users want to express positive sentiment such as “most and “best”, however, negative sentiment in posts was expressed more by users using verbs in past tense forms to express their loss and disappointment such as “missed” and “lost”. More results considering POS tagging between positive and negative sentiment classes can be viewed in the following graph.

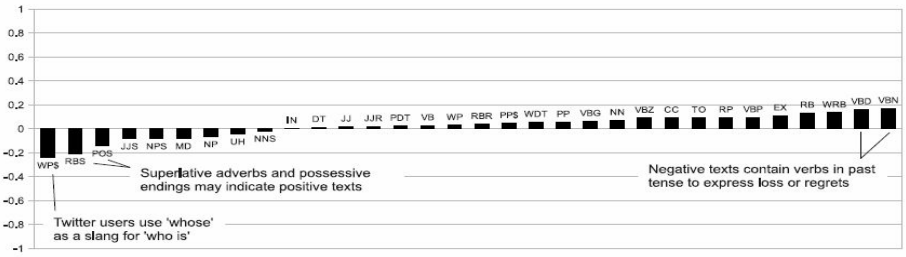


Figure 3: P^T values for positive vs. negative

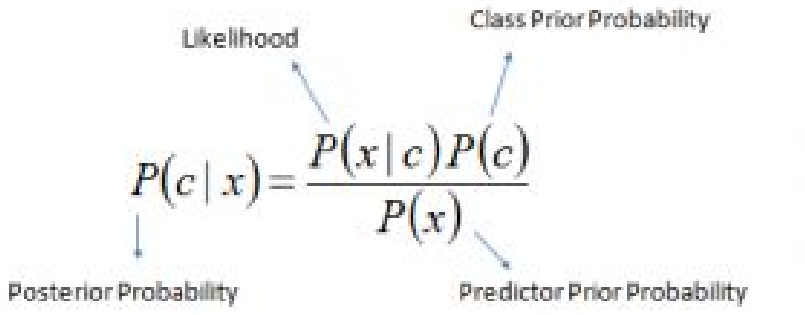
Corpus Classification

Supervised Machine Learning model was used. The basic process of such a process is to have a classifier that takes as input a huge set of data with labels corresponding to each post. In this case, one can assume that posts with positive sentiment are labeled with class “1”, negative sentiment with “2” and objective neutral with class ”3”. This huge set of data is used to train classifier and produce a predictive model. The predictive model is later on used on new examples (Twitter posts in this case) to classify them to one of the desired classes. Such a process is constituted of three main steps: Feature Extraction, Classification, and Increasing Accuracy.

First, Feature Extraction, the first challenge the researchers had to face is whether to use the presence of an n-gram as a feature or the frequency of such an n-gram. Based on the work done by (Pang and Lee, 2008), it is found that the presence of an n-gram is more efficient in sentiment detection. The second challenge was whether to use unigrams or higher order grams as features. Based also on reviewing the overview done by (Pang and Lee, 2008), it was found that unigrams were better indicators in case of movie reviews, however, higher order grams were more indicative in terms of rating products. So, Pak and Paroubek decided on testing both approaches and presented the results, which will be discussed in the results’ section. Before moving on to the classification sub-phase, a pre-processing step was performed, which is a four-step filtering process. First, username, external links and “RT” (which indicates that this post is a retweet, re-shared from another user) were removed from the analysis process. Second, spaces and punctuation marks were removed, however, a word as “I’ll” was still treated as a single unigram without the removal of the apostrophe. Third, stop-words such as: the, a, an, etc. were removed since they are not indicators of any of the three classes. Fourth, the formation of n-grams should handle the case of attaching the negation words such as “not” to attach it to the previous or next word to it and both

being treated as unigrams.

Second, the classification sub-phase, which was based on the Naive Bayes classifier. Naive Bayes classifier is based on the Bayes theorem, which assumes the independence between the input features that are used as inputs to the classifier. It then performs the classification based on the following equation that calculates a probability considering the belonging to one of the desired classes. Such an equation is shown below.



$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Pak and Paroubek used to classifiers, one that takes as input the presence of an n-gram, while the other takes as input distribution information about POS tagging between different sets. The probability generated by both classifiers is calculated using the following equation.

$$P(s|M) \sim P(G|s) \cdot P(T|S)$$

Where P(s | M) is the probability of a sentiment (s) give a twitter message (M), P(G | s) is the probability of the n-grams presence (G) given a sentiment (s) and P(T | S) is the probability of POS tag (T) given a sentiment (S).

In order to normalize the values generated from the previous equation, the following equation was used.

$$L(s|M) = \sum_{g \in G} log(P(g|s)) + \sum_{t \in G} log(P(t|s))$$

Where L(s | M) is the likelihood of a sentiment (s) given a Twitter message (M).

The normalization process is performed in order to disperse the contiguous values more in order to be normally distributed over the range between 0 and 1.

Increasing Accuracy

The results achieved were good but it can be improved by increasing the accuracy of classification. The researchers decided to inspect the usage of two strategies: Entropy and Salience.

First, Entropy, which represents whether an n-gram is evenly distributed over posts expressing the three classes of emotions or whether such an n-gram is indicative exclusively to one of the classes. Higher Entropy values indicated the former theory. On the other hand, lower Entropy values indicated the latter one. Entropy is calculated as follows.

$$entropy(g) = H(p(S|g)) = -\sum_{i=1}^{N} p(S_i|g) \log p(S_i|g)$$

A threshold is placed, so that all values above that threshold are discarded from the analysis process, and those below are taken into consideration.

Second, Salience, which expresses the importance of an n-gram with

respect to the context it is being represented in. High salience values indicated high importance, and low values indicated low importance with respect to the context. It is calculated as follows:

$$salience(g) = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} 1 - \frac{\min(P(g|s_i), P(g|s_j))}{\max(P(g|s_i), P(g|s_j))}$$

Same concept of deploying a threshold is being used where n-grams with values above are taken into consideration and n-grams with values below are being discarded.

Here is an example of high salience n-grams and low entropy ones.

N-gram	Salience
so sad	0.975
miss my	0.972
so sorry	0.962
love your	0.961
i'm sorry	0.96
sad i	0.959
i hate	0.959
lost my	0.959
have great	0.958
i miss	0.957
gonna miss	0.956
wishing i	0.955
miss him	0.954
can't sleep	0.954

N-gram	Entropy
clean me	0.082
page news	0.108
charged in	0.116
so sad	0.12
police say	0.127
man charged	0.138
vital signs	0.142
arrested in	0.144
boulder county	0.156
most viewed	0.158
officials say	0.168
man accused	0.178
pleads guilty	0.18
guilty to	0.181

Table 2: N-grams with high values of salience (left) and low values of entropy (right)

Results

216 Twitter posts were being tested on. The classification of such posts can be shown in the following table.

Sentiment	Number of samples
Positive	108
Negative	75
Neutral	33
Total	216

Table 3: The characteristics of the evaluation dataset

Two entities were used to analyze the results obtained. First, the accuracy which is represented as the ratio between the number of correct classifications and the number of all classifications. Second, the decision which is the ratio between the number of retrieved documents so far over the number of all documents being retrieved. The reason behind measuring the accuracy against decision is that when using Machine Learning algorithms the accuracy tend to degrade as the number of retrieved documents increase, however, how much the accuracy degrades in this case is an indicator of the overall accuracy.

First observation is that less degradation occurs on using bigrams compared to unigrams and trigrams as depicted in the following graph.

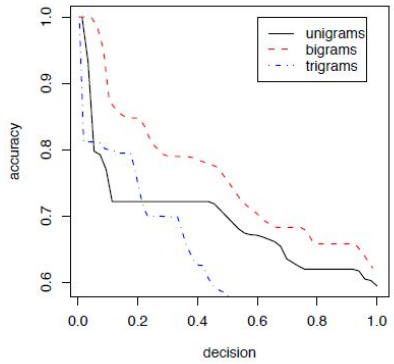


Figure 4: The comparison of the classification accuracy when using unigrams, bigrams, and trigrams

Second observation is that attaching the negation words such as “not” proved to be of more effectiveness than not attaching it. This result can be shown in the following graph.

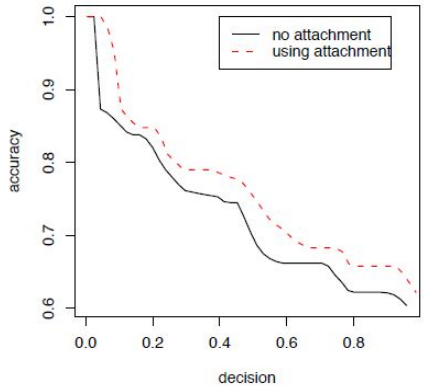


Figure 5: The impact of using the attachment of negation words

Third observation is the number of Twitter posts that should be used in the training phase of the Naive Bayes classifier.

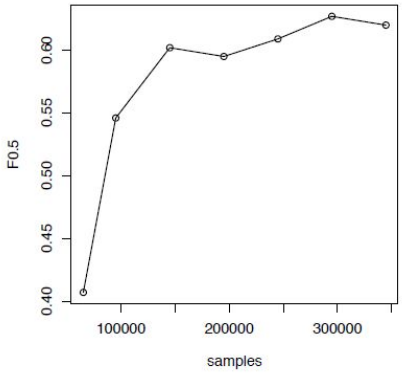


Figure 6: The impact of increasing the dataset size on the $F_{0.5}$-measure

As the graph shows, after retrieving the 300,000 Twitter posts the accuracy of classification and the generated predictive model almost flattens, so no

need to retrieve more posts.

Fourth observation considers whether to use salience or entropy to increase the accuracy of the obtained results. Salience was found to yield less degradation in accuracy against decision and hence better overall accuracy than Entropy. This is depicted in the following graph.

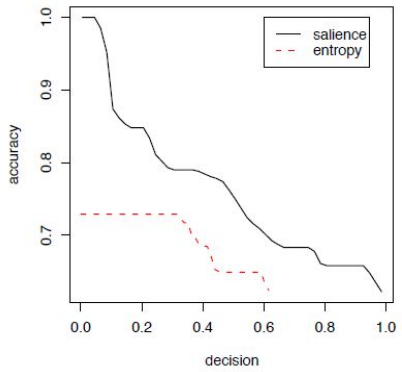


Figure 7: Salience vs. entropy for discriminating common n-grams

Conclusion

To conclude, microblogging platforms are a rich source when it comes to analyze sentiment of users posting such posts. The need arises to analyze sentiments present in posts to gain competitive advantage between companies and alliance in political races. Data was collected automatically using Twitter API and classified for training based on emoticons and newspapers accounts. Naive Bayes classifier yielded the best results when classifying to one of the three classes (positive, negative, objective neutral). Using bigrams was more efficient than unigrams and trigrams in this case of analyzing Twitter posts. Attaching the negation words was a better option than just discarding them from the analysis. Using Salience instead of Entropy increased the accuracy more. The results can be generalized since the corpus collected abided by the laws applied to English language such as Zipf's law. Moreover, the presence of n-grams and distribution of

POS tags across sets were good choice for features as input to the classifier.

As part of the future work, Pak and Paroubek suggest applying the used technique on other languages, which they see is doable since their approach is independent of the language being analyzed.

References

[1] A. Pak, P. Paroubek. 2009. Twitter as a Corpus for Sentiment Analysis and Opinion Mining.

[2] B.Pang, L. Lee. 2008. Opinion Mining and Sentiment Analysis.

[3] A. Go et al. 2009. Twitter Sentiment Classification using Distant Supervision.

[4] https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/