Classification of user Opinions from tweets using Machine Learning Techniques

Poongodi S*
Research Scholar
PSGR Krishnammal College for Women
Peelamedu, Coimbatore
India

Radha N
Assistant Professor
GRG School of Applied Computer Technology
PSGR Krishnammal College for Women
Peelamedu, Coimbatore, India

Abstract—Online Social Network is a standard platform for collaboration, communication where people are connected to each other for sharing their opinion. In general, opinions can be articulated about anything like products, surveys, topics, individuals, organizations and events. There are two main types of textual information in web like facts and opinions. Facts can be expressed in defined terms by the user implicitly. To mine opinion, from the user defined facts is intellectually very demanding. User opinion is valuable data, which can be used for marketing research in business during decision making process. So opinion mining and classification plays a vital role in predicting what people think about products. In this work, basic Natural Language Processing (NLP) techniques and hash tag segments, emoticons are used for classification. The performance comparison of Support Vector Machine (SVM), Naïve bayes (NB) and Multilayer Perceptron (MLP) are done using weka. It is observed that the MLP gives better accuracy to classify the opinion from tweets.

Keywords — Opinion Mining, Machine Learning Methods, NLP techniques, Online Social Network, WEKA.

1. INTRODUCTION

Opinion mining allows concerns to obtain detailed information on what people are saying about products, and take action immediately in relation to negative feedback in their business. Online social network are becoming very trendy. Nowadays People are sharing their opinion, attitudes and connect to the people through social media’s like Facebook, Google+, MySpace and Twitter. Users are create status messages are called tweets in twitter microblog service. Around the world 340+ million of tweets (short text messages) are created and sent per day [2]. These tweets sometimes express opinions about different topics, which are small in length and not to exceed 140 characters.

Research on opinion mining started with the identification of opinion related words, such as grand, superb, wonderful, sad and bad, and looking at their semantic context (i.e. positive or negative). Opinion mining also involves the identification of objectivity or subjectivity in statements. The identification and handling of expression of negative emotions and normalize the short text messages are the major problems in tweets. The aim of this paper is to find out whether the opinions expressed about a product is either positive or negative in an aggregated collection of tweets.

Machine learning algorithms such as Naïve Bayes (NB), Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) are implemented and a comparative analysis is made. Related work is discussed in Section 2. Details of dataset and the pre-processing steps involved are discussed in section 3. Feature Extraction method is discussed in section 4. Experimental Results are presented in Section 5. And in Section 6 conclusion is discussed.

2. LITERATURE REVIEW

Pang et al. (2002) proposed machine learning as the method to classify movie review data. Its observation leads to a two class classification problem which can be performed after determining subjective text. SVM classifier achieved 81.6% accuracy in unigram presence feature set. Go et al. (2009) proposed for automatically classifying the sentiment of Twitter messages, which are classified as either positive or negative with used to a query term. The result of machine learning algorithms is to classifying Twitter message sentiments. The training data consists of Twitter messages with emoticons, which are used as noisy labels. The result showed above 81.9% accuracy for unigram + POS features.

Agarwal et al. (2011) performed three class (positive, negative and neutral) classifications of tweets. Datasets are collected from Twitter stream API and asked human judges to annotate the data into three classes. POS prior polarity features are introduced, which is included with twitter specific features. Unigram + senti features are achieved 75.39% accuracy. A. Bakkiwal et al. (2011) proposed to classify the sentiment using machine learning methods on movie and product review data set. Sentiment classified using basic NLP Techniques such as NGram, POS-Tagged Ngram. Reviews are classified as positive or negative for use several machine learning algorithms Naïve Bayes (NB), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM). Here negation is also studied and handled to improve accuracy in classification. MLP Algorithm achieved the accuracy of 78.32% on Movie data set and achieved 70.06% accuracy on Multi category dataset.
N Bora (2012) created dataset based on noisy labels. A list of 40 words (positive and negative) were created and used to identify the polarity of tweet. A combination of a minimum word frequency threshold and Categorical Proportional difference used as a feature selection method and achieved the highest accuracy of 83.33% on a hand labelled test dataset. A Bakliwal (2012) introduced Senti-feature Identification and Popularity Score to calculate positive or negative opinion, which performed well on twitter datasets. In feature vector approach, the NLP techniques and Twitter specific features are explained. Machine learning techniques implemented and achieved highest accuracy 87.64% on SVM. Research is motivated from each of this work. This paper is extended from A Bakliwal (2012) feature vector approach and implemented machine learning algorithms to classify the opinion.

3. PRE-PROCESSING THE DATASET

Twitter dataset are collected from twitter search API and snapbird tool, which are related to mobile product. Table 1 shows some example tweets.

<table>
<thead>
<tr>
<th>TABLE 1 COLLECTED TWEETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sana: I dnt want Samsung galaxy …so bad its killing me #bad #samsunggalaxy #s4 :(</td>
</tr>
<tr>
<td>Coulo_: Welcome back RT&quot;@phumzilee: The Samsung Galaxy S4 has 2 b d most disappointing &quot;smartphone&quot; I've ever owned. #Samsung #Fail #GalaxyS4 :(</td>
</tr>
<tr>
<td>ofie82: OMG OMG OMG MY MUM SAY THAT MY DAD BOUGHT FOR ME SAMSUNG GALAXY S4!!!!!!!!! :)</td>
</tr>
<tr>
<td>&quot;AzemVerc: &quot;&quot;@FaizAkmalYT: #IWishICould have an iPhone 5&quot;</td>
</tr>
<tr>
<td>EastMjebha: @igalaxys4 @itsmovies. Wow micromax the best :)</td>
</tr>
<tr>
<td>Chaiwei1114: Interested in nokia? Spot me at first avenue</td>
</tr>
</tbody>
</table>

3.1 STOP WORDS REMOVAL

A List of stop words are created like he, she, at, on, a, the, etc. to be removed, which are not used for classify the opinion.

3.2 STEMMING PROCESS

Stemming process is to find the root of the words with suffix with ing, ed such as amazing, liked. Porter stemmer algorithm is used.

3.3 SPELL CHECKING PROCESS

Tweet is not following a standard structure. To avoid this, spell correction is made in the user generated content to normalize the tweets. Spell correction algorithm is used to replace the suggested words using wordnet dictionary. For example the words ‘gooood’ is replaced by ‘god’, ‘good’, ‘goods’, ‘goof’.

![Diagram](image.png)

Fig 1 Flow diagram for opinion classification
Other type of mistakes includes skipping some of characters in words, such as “good” in short words is written as “gud”. Such types of mistakes are handled by Short text dictionary, which is created to use for short text by replacing correct terms.

4. Feature Extraction

Text, smileys, emotions and punctuations are combination of tweets. The collected tweet class labels are annotated manually. These features are extracted using NLP techniques for handling text and twitter features include hashtag segments, emoticons, and URL.

4.1 NLP Techniques

4.1.1 Unigram

List of Positive and negative words are collected. A particular tweet has positive and negative label. Positive contains number of the positive opinion words and negative contains number of the negative opinion words. Table 2 gives some example of positive and negative frequency words in training data set.

<table>
<thead>
<tr>
<th>Positive words</th>
<th>Negative words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Want</td>
<td>bad</td>
</tr>
<tr>
<td>Amazing</td>
<td>boring</td>
</tr>
<tr>
<td>Awesome</td>
<td>slow</td>
</tr>
<tr>
<td>Like</td>
<td>worst</td>
</tr>
</tbody>
</table>

4.1.2 Bigram

A Negation problem is an important one to classify the opinion, which may change the polarity of a sentence. Additionally looking for negation words in the context of a word is also calculated. Negation words can be never, not, don’t, won’t...etc, these type of word presented in a sentence are labelled as a negative. Table 3 gives some example of bigram positive and negative frequency word in training data set.

<table>
<thead>
<tr>
<th>Positive words</th>
<th>Negative words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not bad</td>
<td>Not good</td>
</tr>
<tr>
<td>Never slow</td>
<td>Don’t want</td>
</tr>
<tr>
<td>don’t oppose</td>
<td>Doesn’t used</td>
</tr>
</tbody>
</table>

4.2 Hashtag Segments

Hash Tags polarities are decided based on the essential words of the hash tags. Hash tag symbol is removed and checks whether it contains positive or negative words, with a list opinion words, which are labelled as positive or negative.

4.3 Emoticons

Tweets contain positive and negative emoticons that are classified either as positive or negative which are collected from Wikipedia. URL is also one of the features in tweets, if an URL is presented in a sentence it is labelled as positive otherwise as negative.

5. Machine Learning Algorithm

Extracted features from the data set are used to classify the opinion as positive or negative. To the classified features machine learning algorithm such as NB, SVM and MLP are applied.

5.1 Naïve Bayes (NB)

Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes’ Theorem with strong independence assumptions. A more eloquent term for the underlying probability model would be independent feature model. This independence hypothesis of features make the features order is irrelevant and as a result that the presence of one feature does not affect other features in classification tasks. These make the computation of Bayesian classification approach more efficient. Naïve Bayes classifiers can be trained very powerfully by requiring a small amount of training data to estimate the parameters necessary for classification.

5.1 Support Vector Machine (SVM)

SVM takes a set of input data and predicts, for each given input, which of the two possible classes comprises the input, making the SVM a non-probabilistic binary linear classifier. Training set, each marked as belonging to one of two categories, and an SVM training algorithm builds a model that assigns new instances into one category or the other. An
SVM model is a representation of the instances as points in space, mapped so that the instances of the separate categories are divided with conditions. Testing instances are mapped into same conditions and predicted to belong to a category.

5.2 Multi Layer Perceptron

A Multi Layer Perceptron (MLP) is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron or processing element with a nonlinear activation function. MLP utilizes back propagation for training the network. This class of networks consists of multiple layers of computational units, and interconnected in a feed-forward way. Many applications the units of these networks apply a sigmoid function as an activation function.

6. Experimental Results

This section describes the experimental results of the classification algorithm using collected tweets in weka tool. Weka tool is an open source for machine learning, data mining, text mining, predictive analytics, and business analytics. The opinions are classified using machine learning techniques namely Naïve bayes, Multilayer Perceptron and Support Vector Machine. A comparative analysis is used for the above algorithms based on both NLP techniques and NLP with other features. From the result it is analysed that the Multilayer Perceptron produces better accuracy, when compared with Multi Layer Perceptron as shown in Table 4.

<table>
<thead>
<tr>
<th>Algorithm Used</th>
<th>NLP</th>
<th>NLP with emoticons</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>87.98%</td>
<td>90.41%</td>
</tr>
<tr>
<td>SVM</td>
<td>92.84%</td>
<td>97.28%</td>
</tr>
<tr>
<td>MLP</td>
<td>93.70%</td>
<td>99.57%</td>
</tr>
</tbody>
</table>

Fig 2 Comparisons charts for NB, SVM and MLP machine learning Algorithm

7. Conclusions

Opinion classification is an important and challenging task using twitter data set. A twitter micro blog suffers from various linguistic and grammatical errors. This paper provides how to pre-process the tweets for maximum information extraction from short text message and classify the opinion using machine learning algorithms. The experimental results are compared for three machine learning algorithms, where mobile product related tweets achieves 90.41% on Naïve Bayes, 97.28% on Support Vector Machine algorithm and 99.57% on Multi Layer Perceptron algorithm, in which Multi Layer Perceptron achieves more accuracy, when comparing with Naïve Bayes and Support Vector Machine algorithm.
REFERENCES


