Study of different approaches for Sentiment Analysis

Anuj Kumar Gupta

Professor, CSE Dept., Chandigarh Group of Colleges, Mohali, PB, India

Abstract: **We live in an era of social media and the Internet. The development of social media and digital technologies has changed how individuals share their opinions on things. Huge amounts of pertinent content have been produced as a result, which may now be analysed by corporations and the affected parties for their own gains on the Internet and in social media. Nowadays, people express their thoughts about goods, services, etc. via blogs, Twitter, Facebook, and other online forums. These reviews can be helpful to customers in deciding whether to purchase a good or use a service, as well as to businesses in gaining open input and advancing the planning process. There must be some effective automated procedures to analyse the enormous volume of pertinent data that can be used for financial gain. Sentiment analysis is aprocedure of determining a person's genuine moods or emotions from a highly unstructured tweet on Twitter or a status post on Facebook. This research compares several estimation mining methods such as Naive Bayes,Machine Learning etc. with a specific focus on Twitter data. Additionallyhighlighted are thesentiment analysis process and several approaches.**

Keywords: **Sentiment Analysis, Machine Learning, Naive Bayes, Twitter.**

I. Introduction

Users, these days, like to express their views on micro blogging sites such as tweeter and other social media about anything and everything. It has become a universal medium of expression to the extent that now businesses across the globe want to draw the benefits out of it by carrying out the in depth analysis of data or tweets with a purpose of focussing their marketing strategies based on this analysis and thereby attracting a huge customer base. On twitter, there are handles of every organisation and its products and users post their views on it in the form of tweets having the size up to 148 characters. Social media is producing an enormous amount of sentiment rich data every day and this very fact calls for the efficient techniques that can automate the sentiment analysis. The techniques aimed at extracting the textual information from the data rely on the analysis of facts. Facts provide objectivity to the data. Apart from this objective information, the data also possesses some subjective characteristics. These parts of the data are very hard to decipher as these may represent someone’s attitude, opinion, etc. This demands the development of applications that can accurately depict the sentiment behind the data.

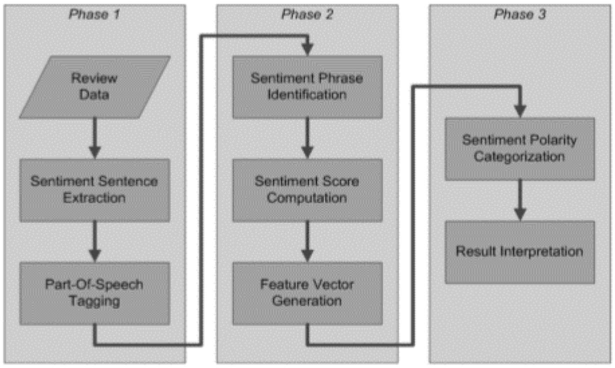
Phase 1
Review Data
Sentiment Sentence Extraction
Part-Of-Speech Tagging
Phase 2
Sentiment Phrase Identification
Sentiment Score Computation
Feature Vector Generation
Phase 3
Sentiment Polarity Categorization
Result Interpretation

Fig.1. Sentiment Polarity Categorization Process

Figure 1 above depicts how a sentiment analysis proceeds through different phases to achieve the required results.There are different terminologies used for sentiment analysis as given under:

<Sentence> = camera of a mobile phone is not good and has a worst picture quality.
<Opinion holder>=Author
<Object>=Mobile Phone
<Feature>= camera
<Opinion>= Not good & worst picture quality
<Polarity>= negative

II. Related work

Using Twitter opinion mining, Cho et al. devised a technique to display the sequential and altitudinal distribution of brand pictures [2]. They create a vocabulary of emotions for words in Korean [3]. This study [2] demonstrated how Twitter data may be utilised to analyse brand perception across space and time. Additionally, the brand association network's temporal fluctuations revealed which terms are the subject of peoples' awareness.

By summarising the reviews, Taysir et al. presented an opinion mining methodology to aid potential buyers in making a decision on a purchase [4]. By measuring the cosine similarity, they were able to categorise the review sentences of a merchandise according to its topographies [2]. Features and polarity were ranked in the study. A product class was categorised using its synonyms by the feature classification. According on the polarity of the sentence, sentences.

An idea for choosing the best classifier based on the characteristics and attributes of the database was provided by Yu Zhang and Pedro Desouza [2]. They assess the results of five classifiers using famous social media data sources of Twitter, Amazon etc. [5]. To increase the accuracy and predictive capacity of sentiment analysis, they also created a new sentiment analysis process.

Pablo et. al. [10] discussed a machine learning technique in which two variations of naive Bayes classifier are used. These variations are baseline and binary used for detecting the sentiment of English tweets. The feature vectors used for training set are: lemmas, polarity lexicons and multiwords.

Xia et. al. [11] proposed a framework in which he used different feature sets(part-of-speech tags and relation between the words) and three types of classifiers(naive Bayes, support vector machine, maximum entropy). The performance of these three classifiers is compared and different approaches like fixed combination, weighted combination are applied to achieve better results.

Po-Wei Liang et. al. [12] proposed a technique to carry out the sentiment analysis of twitter data. First of all, the tweets are aggregated using twitter API. The data used for training is divided into three categories: camera, movie and mobile. Then the unigram naive Bayes model is employed and the features that are not relevant are eliminated using chi square technique and mutual information feature extraction method.

III. Sentiment Analysis Process

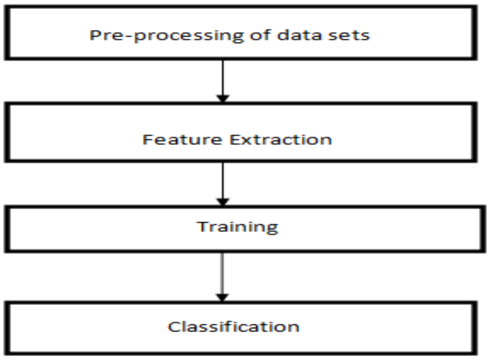

Fig. 2. Sentiment Analysis Process

Following are the phases to carry out the sentiment analysis for twitter data:

Pre-processing of datasets

Before analysing the sentiments of tweets, pre-processing is carried out. This is because the tweets are the raw data and contains inconsistencies and redundancy which has to be removed to make it suitable for analysis. Pre-processing of the tweets consists of the following:

1. Removal of hash tags and URLs.
2. Removal of stop words.
3. Removal of non-English words.
4. Removal of punctuations, symbols and numbers.
5. Replace all the emoticons with their sentiments.

A. Feature Extraction

Pre-processed tweets possess various features which can be extracted in order to determine the positive and negative polarity of a sentence. After extracting the features, the entire process becomes easy and can be done using the models like bigram or n-gram model[6]. Machine learning techniques require these features to be extracted so that they can act as a feature vectors that can be used as a training set in the learning.

The features that can be extracted are:

1. ***Parts of speech tags***: The parts like adverbs, nouns can be extracted to determine the subjectivity in the tweets. Syntactic dependency can be generated based on these using parse trees.
2. ***Words and their frequencies***: The frequency counts of the words in the text can be effectively used as a feature which can be used in the models like unigram, bigram and n-gram models.
3. ***Position of the terms***: The relative positioning of terms and words in a phrase can act as a useful feature which can determine of the opinion of the text.
4. ***Specific opinion words and phrases***: In the context of the situation, specific opinion words and phrases tend to repeat more frequently than the others. These words can be identified and used as a feature.

B. Training

After extracting the features, the training data set is created according to which the classifier is trained and then it is able to work for unknown data.

C. Classification

Classification is divided in two types as follow:

i. Naive Bayes:

It assumes the availability of a set of articles of which the opinions and reality labels have already been assigned [7]. Based on this data, it allocates a document d, a category c which maximizes its probability since naive Bayes is a probabilistic classifier:

$$P\left(\frac{c}{d}\right) = \frac{P(c)P\left(\frac{d}{c}\right)}{P(d)} \qquad (1)$$

*ii. Support Vector Machine**:***

SVM classifies a huge number of instances into different groups after receiving them as input. It uses training data, which is a lot of points in an infinitely dimensional space, as examples. These points are a part of one of the numerous classes known as hyper plane that are separated by the greatest distance conceivable. New examples must fall into at least one of the classes and be spaced apart by the margin when they are fed into the SVM [13]. Greater the margin, the lower the classifier's generalisation error would be [8].

IV. Approaches for Sentiment Analysis

There are primarily two approaches for carrying out the sentiment analysis of twitter data:

A. *Machine learning approaches:*

Such approaches classify the data into classes [14]. There are two machine learning techniques:

i.Unsupervised learning:

It is a learning technique without the training data and relies mainly on clustering. It is regarded to learn by observation not by example [9].

Table 1. Comparison of various machine learning techniques

Technique	Conception	Extensions	Accuracy	Benefits	Difficulties
SVM (Support Vector Machine)	Grounded on decision plane that defines the borders of decision	1. Soft Margin 2. Non Linear 3. Multi Class Sum	With unigram 82.91%	1. Low Dependency on data set 2. Decent for investigational result	1. Reads pre-processing for lost values 2. Understanding is problematic
MLP (Multi-Layer Perception)	1 or N layer exist for input/output. Named feed-forward neural network	2 phase: 1-forward phase input-to-output layer 2-alter the weight and bias value error	84.25% - 89.50%	1-Act as a common function 2-Can learn each association	1-Needs extra time for implementation 2-Considered as a composite black box
Naive Bayes Classifier	For 2 event conditional prob.– P(e1/e2)=P(e2/e1)P(e1)/e2	Use exactness precision recollection reference	0.79939	1-Simple to device 2-EffectiveCalculation	1-Assumption of qualities being autonomous which may not be necessarily effective

ii.Supervised learning:

It is regarded as to learn by example. A training data set is created and fed into the system to obtain the meaningful outputs. This helps in decision making.The efficiency of the supervised learning techniques is based on the fact that how accurately the features are extracted from the pre-processed data and fed as the feature vectors into the system in order to detect the sentiment.

There are two types of data sets needed in the supervised learning algorithms:
Training set
Test set

There exists a variety of supervised learning techniques like Naive Bayes, SVM(support vector machine) etc.

B. *Lexicon-based approaches:*

In these approaches, the words are matched with their polarity present in the sentiment dictionary. The sentiment dictionary already contains the words along with their polarity which can be obtained from the other work or created by the author specific to its work.These approaches are based on a sentiment lexicon- a group of precompiled lexicon terms that are specifically created for a specific purpose.The lexicon-based approaches are further subdivided into two sub-approaches:

a) Dictionary based approach

The dictionary like WordNet which is created and annotated manually consists of various opinion words with pre assigned polarity. These can be grown by adding more synonyms and antonyms of the pre-existing words in the dictionary. But the major drawback of these techniques is that it can’t deal with the domain-specific words.

b) Corpus-based approach

Corpus of the domain-specific words is created starting from a set of seed words,this corpus is specific to a particular kind of work. Such type of dictionaries grow by adding more related terms or opinion words related to the field with the help of various semantic or statistical techniques.

Table 2. Comparison of sentiment analysis approaches

Technique	Method	Dataset	Accuracy
Machine Learning	SVM	More reviews	86.40%
	Co-Training SVM	Twitter	82.52%
	Deep Learning	Stanford Sentiment Tree Bank	80.70%
Lexical based	Dictionary	Amazon’s Mechanical Turk	----
	Corpus	Product reviews	74.00%

V. Conclusion

In addition to an in-depth look at machine learning approaches, this research aims to provide a comparative analysis of the current opinion mining techniques. The effectiveness of machine learning techniques like SVM and MLP, as well as their benefits and drawbacks, are also thoroughly examined. The study found that machine learning techniques are effective and yield superior outcomes. In order to validate the veracity of their findings, extensive research is also done in the field of lexicon-based techniques. Furthermore, efforts are being made to increase the accuracy of machine learning techniques up to bigram and trigram. The majority of persons in the web world rely on social networking websites to push their valuable information;

however studying the reviews from these blogs can produce a deeper understanding and support in decision-making.

References

[1] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Springer J. Big Data*, vol. 2, no. 1, p. 5, 2015.

[2] M. D. Devika, C. Sunitha, and A. Ganesh, "Sentiment Analysis: A Comparative Study on Different Approaches," in *Procedia Computer Science*, 2016, vol. 87, pp. 44–49.

[3] V. A. Kharde and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," *Int. J. Comput. Appl.*, vol. 139, no. 11, pp. 975–8887, 2016.

[4] Z. Hu, J. Hu, W. Ding, and X. Zheng, "Review Sentiment Analysis Based on Deep Learning," in *2015 IEEE 12th International Conference on e-Business Engineering*, 2015, pp. 87–94.

[5] M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," *J. King Saud Univ. - Eng. Sci.*, vol. 34, no. 4, 2016.

[6] Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment Treebank."Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2013.

[7] Q. Rajput, S. Haider, and S. Ghani, "Lexicon-Based Sentiment Analysis of Teachers ' Evaluation," *Hindawi Appl. Comput. Intell. Soft Comput.*, vol. 2016, no. 6, 2016.

[8] Tripathy, A. Agrawal, and S. K. Rath, "Classification of Sentimental Reviews Using Machine Learning Techniques," *Procedia Comput. Sci.*, vol. 57, pp. 821–829, 2015.

[9] V. M. Pradhan, J. Vala, and P. Balani, "A Survey on Sentiment Analysis Algorithms for Opinion Mining," *Int. J. Comput. Appl.*, vol. 133, no. 9, pp. 7–11, 2016.

[10] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

[11] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[12] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management, Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN:978-1-494673-6068-5.

[13] Kushawaha D., De D., Mohindru V., Gupta A.K. (2020) Sentiment Analysis and Mood Detection on an Android Platform Using Machine Learning Integrated with Internet of Things. In: Singh P., Kar A., Singh Y., Kolekar M., Tanwar S. (eds) Proceedings of ICRIC 2019. Lecture Notes in Electrical Engineering, vol 597. Springer

[14] AK Gupta, M Sharma, A Sharma, V Menon. "A Study on SARS-CoV-2 (COVID-19) and Machine Learning Based Approach to Detect COVID-19 Through X-Ray Images", International Journal of Image and Graphics, 2021, 21(1).