EVOLUTION OF SENTIMENTAL ANALYSIS

Renuka S. Deshmukh, Dr. Aashish A. Bardekar

[1]*Student, Dept. of Computer Science and Engineering, Sipna College of Engineering and Technology, Maharatra , India.*
[2]*Professor, Dept. of Computer Science and Engineering, Sipna College of Engineering and Technology, Maharatra, India.*

---***---

Abstract - *The sentiment analysis is the technique which can analyze the behavior of the user. The data which is analyzed is the twitter data. The four steps are followed for the sentiment analysis in the first step, the first step is applied in which data pre-processed. In the second step feature of the data will be extracted which is given as input to the third step in which data is classified for the sentiment analysis. In this paper, pattern based technique is applied for the feature extraction in which patterns are generated from the existing patterns which increase the accuracy of data classification. The proposed algorithm is been implemented in python using the nltk tool box and it is been analyzed that execution time is reduced and accuracy is increased at steady rate.*

Keywords: Feature extraction, pre-processing, pattern generation, sentiment analysis.

1. INTRODUCTION

Sentiment analysis intents to define the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual division or emotional response to a document, interaction, or event. It refers to the use of natural language processing, text analysis, computational semantics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is broadly applied to "voice of the customer" materials such as reviews and survey responses, as well as to online and social media. Sentiment analysis has claims in a variety of domains, ranging from marketing to customer service to clinical medicine. Sentiment analysis stands at the intersection of natural language processing and large-scale data mining.

Sentiment analysis has important applications in academia as well as commerce. The understanding of human language is a core problem in AI research. At the same time, with increasingly lowering barriers to the Internet, it is easier than ever for end-users to provide feedback on the products and services they use. This information is highly valuable to commercial organizations; however, the volume of such reviews is growing rapidly, necessitating an automated approach to extracting meaning from the huge volume of data. This automated approach is provided by sentiment analysis.

2. Litereature Review

Rincy Jose, et.al, most sentiment analysis systems use bag-of-words approach for mining sentiments from the online reviews and social media data. Rather considering the whole sentence/ paragraph for analysis, the bag-of-words approach considers only individual words and their count as the feature vectors. This may mislead the classification algorithm especially when used for problems like sentiment classification. Traditional machine learning algorithms like Naive Bayes, Maximum Entropy, SVM etc. are widely used to solve the classification problems [15]. Experiments conducted demonstrate that the semantics based feature vector with ensemble classifier outperforms the traditional bag-of-words approach with single machine learning classifier by 3-5%. It is observed that the ensemble method outperforms the traditional classification methods by about 3- 5%. Among the ensemble methods Extremely Randomized Trees classification performs better than others.

Nehal Mamgain, et.al, this paper additionally highlights a comparison between the results got by exploiting the following machine learning algorithms: Naïve Bayes and Support Vector Machine and an Artificial Neural Network model: Multilayer Perceptron [16]. Moreover, a contrast has been displayed between four distinct kernels of SVM: RBF, linear, polynomial and sigmoid. Multilayer Perceptron Neural Network surpasses the results yielded by the machine learning algorithms owing to its exceptionally accurate approximation of the cost function, ideal number of hidden layers and learning the relationship among input and output variables at every progression.

Aldo Hernández, et.al, this paper presents a sentiment analysis method on Twitter content to predict future attacks on the web [17]. The method is based on the daily gathering of tweets from two sets of users; the individuals who utilize the platform as a method for expression for views on relevant issues, and the individuals who utilize it to present contents identified with security attacks in the web. Daily information is converted into data that can be broke down statistically to predict whether there is a plausibility of an assault. The last is finished by investigating the aggregate sentiment of users and groups of hacking activists in response to a global event. The goal is to predict the response of specific groups involved in hacking activism when the sentiment is sufficiently negative among various

Twitter users. For two contextual analyses, it is demonstrated that having coefficients of determination greater than 44.34% and 99.2% can figure out whether a significant increase in the percentage of negative opinions is identified with attacks. Anurag P. Jain, et.al, this Paper presents approach for examining the sentiments of users utilizing data mining classifiers [18]. It additionally compares the performance of single classifiers for sentiments analysis over ensemble of classifier. Experimental results acquired demonstrate that k-nearest neighbor classifier gives high predictive accuracy. Results likewise demonstrate that single classifiers outperforms ensemble of classifier approach. It can be seen from the test results that data mining classifiers is a decent decision for sentiments prediction utilizing tweeter data. In experimentation, k-nearest neighbor (IBK) outperforms over every one of the three classifiers in particular RandomForest, baysNet, Naive Baysein. RandomForest additionally gives great prediction accuracy. There is a no compelling reason to utilization of ensemble of classifier for sentiments predictions of tweets as single classifier (i.e k-nearest neighbor) gives a better accuracy over all combinations of ensemble of classifier. Manju Venugopalan, et.al, the proposed work goes for building up a half and half model for sentiment classification that explores the tweet specific features and uses domain independent and domain specific lexicons to offer a domain oriented approach and thus investigate and extract the shopper sentiment towards popular smart phone brands in the course of recent years [19]. The analyses have demonstrated that the results enhance by around 2 points on an average over the unigram baseline. The SVM accuracy has improved in the range 1.5 to 3.5 and J48 could provide an accuracy improvement ranging from 1.5 to 4 points across domains. The improved lexicon which have adapted polarities learning the domain and the tweet specific features extracted have added to the improvement in classification accuracies.

3. Details of Topics

3.1 Fine-grained Sentiment Analysis

If polarity precision is important to your business, you might consider expanding your polarity categories to include:

- Very positive
- Positive
- Neutral
- Negative
- Very negative

This is usually referred to as fine-grained sentiment analysis, and could be used to interpret 5-star ratings in a review, for example:

- Very Positive = 5 stars
- Very Negative = 1 star

3.2 Emotion detection

This type of sentiment analysis aims at detecting emotions, like happiness, frustration, anger, sadness, and so on. Many emotion detection systems use lexicons (i.e. lists of words and the emotions they convey) or complex machine learning algorithms.

One of the downsides of using lexicons is that people express emotions in different ways. Some words that typically express anger, like bad or kill (e.g. your product is so bad or your customer support is killing me) might also express happiness (e.g. this is bad ass or you are killing it).

3.3 Aspect-based Sentiment Analysis

Usually, when analyzing sentiments of texts, let's say product reviews, you'll want to know which particular aspects or features people are mentioning in a positive, neutral, or negative way. That's where aspect-based sentiment analysis can help, for example in this text: "The battery life of this camera is too short", an aspect-based classifier would be able to determine that the sentence expresses a negative opinion about the feature battery life.

3.4 Multilingual sentiment analysis

Multilingual sentiment analysis can be difficult. It involves a lot of preprocessing and resources. Most of these resources are available online (e.g. sentiment lexicons), while others need to be created (e.g. translated corpora or noise detection algorithms), but you'll need to know how to code to use them.

Alternatively, you could detect language in texts automatically with MonkeyLearn's language classifier, then train a custom sentiment analysis model to classify texts in the language of your choice.

4. Purposed Work

As we are doing the sentimental analysis project we have to perform some of the algorithms to find the exact result or the approximate result to get the correct output we are using the following algorithms.

1. Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. ... Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

2. Support vector machine

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

3. Naive Bayes

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

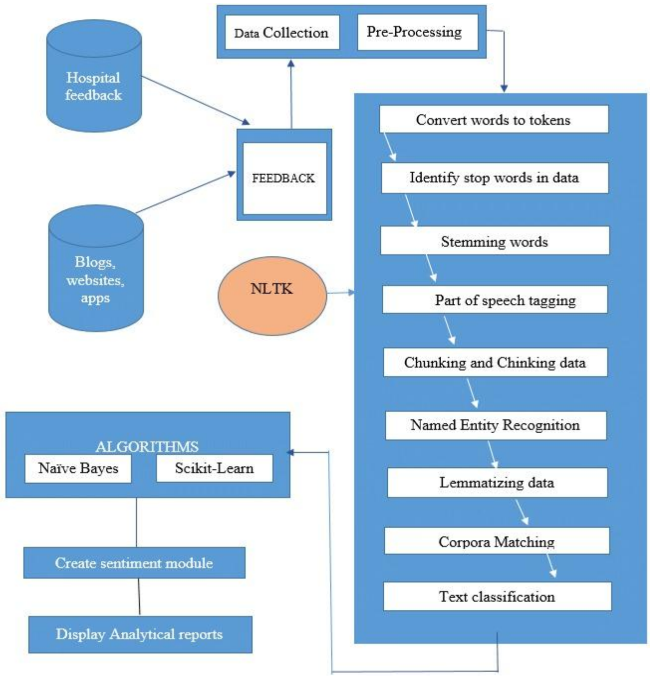
Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

4. Decision Tree algorithm.

Decision Tree algorithm belongs to the family of supervised learning algorithms. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

5. Hybrid Approaches

Hybrid systems combine the desirable elements of rule-based and automatic techniques into one system. One huge benefit of these systems is that results are often more accurate.



5.1 Sentiment Analysis Challenges

Computer scientists have been trying to develop more accurate sentiment classifiers, and overcome limitations in recent years. Let's take a closer look at some of the challenges they face:

5.1.1 Subjectivity and Tone

The detection of subjective and objective texts is just as important as analyzing their tone. In fact, so called objective texts do not contain explicit sentiments. Say, for example, you intend to analyze the sentiment of the following two texts:

- The package is nice.
- The package is red.

Most people would say that sentiment is positive for the first one and neutral for the second one, right? All predicates (adjectives, verbs, and some nouns) should not be treated the same with respect to how they create sentiment. In the examples above, nice is more subjective than red.

5.1.2 Context and Polarity

All utterances are uttered at some point in time, in some place, by and to some people, you get the point. All utterances are uttered in context. Analyzing sentiment without context gets pretty difficult. However, machines cannot learn about contexts if they are not mentioned explicitly. One of the problems that arise from context is changes in polarity. Look at the following responses to a survey:

- Everything of it.
- Absolutely nothing!

Imagine the responses above come from answers to the question what did you like about the event? The first response would be positive and the second one would be negative, right? Now, imagine the responses come from answers to the question what did you DISlike about the event? The negative in the question will make sentiment analysis change altogether.

A good deal of preprocessing or postprocessing will be needed if we are to take into account at least part of the context in which texts were produced. However, how to preprocess or postprocess data in order to capture the bits of context that will help analyze sentiment is not straightforward.

5.1.3 Irony and Sarcasm

When it comes to irony and sarcasm, people express their negative sentiments using positive words, which can be difficult for machines to detect without having a thorough understanding of the context of the situation in which a feeling was expressed.

For example, look at some possible answers to the question, Did you enjoy your shopping experience with us?

- Yeah, sure. So smooth!
- Not one, but many!

What sentiment would you assign to the responses above? The first response with an exclamation mark could be negative, right? The problem is there is no textual cue that will help a machine learn, or at least question that sentiment since yeah and sure often belong to positive or neutral texts. How about the second response? In this context, sentiment is positive, but we're sure you can come up with many different contexts in which the same response can express negative sentiment.

5.1.4 Comparisons

How to treat comparisons in sentiment analysis is another challenge worth tackling. Look at the texts below:

- This product is second to none.
- This is better than older tools.
- This is better than nothing.

The first comparison doesn't need any contextual clues to be classified correctly. It's clear that it's positive.

The second and third texts are a little more difficult to classify, though. Would you classify them as neutral, positive, or even negative? Once again, context can make a difference. For example, if the 'older tools' in the second text were considered useless, then the second text is pretty similar to the third text.

5.1.5 Emoji's

There are two types of emojis according to Guibon et al.. Western emojis (e.g. :D) are encoded in only one or two characters, whereas Eastern emojis (e.g. ¯_(ツ)_/¯) are a longer combination of characters of a vertical nature. Emojis play an important role in the sentiment of texts, particularly in tweets.

You'll need to pay special attention to character-level, as well as word-level, when performing sentiment analysis on tweets. A lot of preprocessing might also be needed. For example, you might want to preprocess social media content and transform both Western and Eastern emojis into tokens and whitelist them (i.e. always take them as a feature for classification purposes) in order to help improve sentiment analysis performance.

Here's a quite comprehensive list of emojis and their unicode characters that may come in handy when preprocessing.

5.1.6 Defining Neutral

Defining what we mean by neutral is another challenge to tackle in order to perform accurate sentiment analysis. As in all classification problems, defining your categories -and, in this case, the neutral tag- is one of the most important parts of the problem. What you mean by neutral, positive, or negative does matter when you train sentiment analysis models. Since tagging data requires that tagging criteria be consistent, a good definition of the problem is a must.

Here are some ideas to help you identify and define neutral texts:

1. Objective texts. So called objective texts do not contain explicit sentiments, so you should include those texts into the neutral category.

2. Irrelevant information. If you haven't preprocessed your data to filter out irrelevant information, you can tag it neutral. However, be careful! Only do this if you know how this could affect overall performance. Sometimes, you will be adding noise to your classifier and performance could get worse.
3. Texts containing wishes. Some wishes like, I wish the product had more integrations are generally neutral. However, those including comparisons like, I wish the product were better are pretty difficult to categorize

6. CONCLUSIONS

It's the process of analyzing online pieces of writing to determine the emotional tone they carry. In simple words, sentiment analysis is used to find the author's attitude towards something. Sentiment analysis tools categorize pieces of writing as positive, neutral, or negative.

Thus, Opinion Mining and Sentiment analysis has wide area of applications and it also facing many research challenges. Since the fast growth of internet and internet related applications, the Opinion Mining and Sentiment Analysis become a most interesting research area among natural language processing community. A more innovative and effective techniques required to be invented which should overcome the current challenges faced by Opinion Mining and Sentiment Analysis.

Acknowledgement

A moment of pause, to express a deep gratitude to several individuals, without whom this project could not have been completed. We feel immense pleasure to express deep sense of gratitude and indebtedness to our guide **Dr. Aashish A. Bardekar**, for constant encouragement and noble guidance. We express our sincere thanks to **Dr.V. K. Shandilya**, Head of Department, Computer Science & Engineering, and the other staff members of the department for their kind co-operation.

We express our sincere thanks to **Dr. S.M.Kherde** Principal Sipna College of Engineering & Technology for his valuable guidance. We also express our sincere thanks to the library staff members of the college. Last but not the least we are thankful to our friends and our parents whose best wishes are always with us.

REFERENCES

1. Zhou Jin, Yujiu Yang, Xianyu Bao, Biqing Huang," Combining User-based and Global Lexicon Features for Sentiment Analysis in Twitter", 2016, IEEE, 978-1-5090-0620-5
2. Fadhli Mubarak bin Naina Hanif, G. A. Putri Saptawati," CORRELATION ANALYSIS OF USER INFLUENCE AND SENTIMENT ON TWITTER DATA", 2014, IEEE, 978-1-4799-7996-7
3. Vishal A. Kharde, S.S. Sonawane," Sentiment Analysis of Twitter Data: A Survey of Techniques", 2016, International Journal of Computer Applications, Volume 139 – No.11
4. Deepali Arora, Kin Fun Li and Stephen W. Neville," Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study", 2015, IEEE, 1550-445X
5. Richa Sharma1, Shweta Nigam2 and RekhaJain,"Supervised Opinion Mining Techniques: A Survey". International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 8 (2013), pp. 737-742.
6. Sindhu C1, Dr. S. ChandraKala, "A Survey on opinion mining and sentiment polarity classification". International Journal of Emerging Technology and Advanced Engineering. , Volume 3, Special Issue 1, January 2013).
7. Fadhli Mubarak bin Naina Hanif, G. A. Putri Saptawati," CORRELATION ANALYSIS OF USER INFLUENCE AND SENTIMENT ON TWITTER DATA", 2014, IEEE, 978-1-4799-7996-7
8. Zhou Jin, Yujiu Yang, Xianyu Bao, Biqing Huang," Combining User-based and Global Lexicon Features for Sentiment Analysis in Twitter", 2016, IEEE, 978-1-5090-0620-5
9. B. Liu, "Handbook Chapter: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing," Handbook of Natural Language Processing. Marcel Dekker, Inc. New York, NY, USA, 2009.
10. K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519–528.
11. Richa Sharma1, Shweta Nigam2 and RekhaJain,"Supervised Opinion Mining Techniques: A Survey". International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 8 (2013), pp. 737-742.
12. Sindhu C1, Dr. S. ChandraKala, "A Survey on opinion mining and sentiment polarity classification". International Journal of Emerging Technology and Advanced Engineering. , Volume 3, Special Issue 1, January 2013).

BIOGRAPHIES

Pursing M.E. Degree in Computer Science &Engineering From Sipna College of Engineering and Technology Amravati, Sant Gadge Baba Amravati University.

Dr.Aashish A.Bardekar
Assistant Professor,
Dept of Computer Science & Engineering, Sipna College of Engineering and Technology, Amravati. CSI, ACM, IETE.