Sentiment Analysis using Machine Learning Technique: A Literature Survey

Nirag T. Bhatt[1], Asst. Prof. Saket J. Swarndeep[2]

[1]*Department of computer engineering, L.J Institute of Engineering and Technology (Gujarat Technological University), Ahmedabad, Gujarat, India*
[2]*Prof. L.J Institute of Engineering and Technology (Gujarat Technological University), Ahmedabad, Gujarat, India*

---***---

Abstract: - *In this article there are different machine learning techniques which are used for sentiment analysis. Mostly sentiment analysis done by using machine learning classifier like SVM (support vector machine), Random forest, Naïve Bayes. In this we are seeing some paper which are help new researcher to found a proper path for their new research. In this there is a proposed method of new research program. Social media is biggest medium to share people's opinion on different topics. Sentiment analysis uses machine learning technique and without any human interruption machine will give and accurate sentiment of the people. Sentiment analysis turn text into positive, negative or neutral. So, any company or foundation or movie reviewer can take the opinion of the people and take further steps according that.*

Key Word: - *Sentiment analysis, SVM, Naïve Bayes, Social media, Twitter, Social media*

1. Introduction

Sentiment analysis is a machine learning tool which is used for analyze the texts for polarity from positive to negative. Machine automatic learn how to analyze the sentiment of the human without the human input or interruption. Nowadays social media is a part of the people's life; people uses social media for give their review over some political field, movie review or marketing area. There are many social media sites like Twitter, Facebook, Instagram etc. They use this social media sites as the medium to express their view on many topics. So, sentiment analysis analyzes the text which inputted by any person from the different country by using the training data set it will analyze the sentiment of that particular text by knowing the emotion of that people.

The application of the sentiment analysis very broad and powerful like Expedia Canada; Canadian take the advantage of sentiment analysis when they notice that people are giving negative comments on the music used by their television channel. Rather than chalking by negative comment, Expedia manages to take advantage of that negative comment and air all new soulful music in their channel.

1.1 Levels of Sentiment analysis

1. Document level: - Document level analysis used for whole document. In this level of classification, a document about single topic is included. Customer have mentality to compare two topics or two document it can't be done in document level analysis. The supervised and unsupervised both machine learning technique used for classification of Document level sentiment analysis.
2. Sentence level: - Subjectivity classification is closely related to sentence level sentiment analysis. The sentence level sentiment analysis is to find out the expression like positive, negative or neutral from the given sentence. All the classifier from document level sentiment analysis is used for sentence level sentiment analysis.
3. Aspect level: - The Aspect level sentiment analysis is used to find out sentiment on Aspect of those entities. "My car has good handling but it is little heavy" let's take this example. In this example there is an opinion on a car that handling of cat is positive but the weight of car is negative. The statement which are competitive is part of an Aspect level sentiment analysis.
4. Phrase level: - In the phrase where opinion words are found out their phrase level classification done. These has advantage and disadvantage both because advantage is their where the exact opinion about entity is there. But in disadvantage there is contextual polarity matter so result may not be accurate.
5. Feature Level: - Product feature is identifying as product attributes. In document Analyzing of these feature for identifying sentiments called as feature level sentiment analysis. Positive, negative or neutral opinion is identifying from extracted features.

1.2 Application of Sentiment Analysis

1. Monitoring market research: - To see what is the new thing that came into the market and what people wants in the market. After analyzing that you can change your business strategy according that.
2. To see the competition: - To see what your competitors are launching or which product they put into the market. To study the competitor's strategy according to people's opinion. That is one of the main application of sentiment analysis.
3. Product Analysis: -To find out what people say about the product after the launch or you can see people's reaction that you did never seen before. By searching the keyword for a products particular feature you can easily analyze the product review.
4. Social media monitoring: - People share their view on social media on any field like business, government, market or any other. By the sentiment analysis by searching some keyword you can easily monitor people's sentiment on individual point of view.
5. Customers feedback: - Customer feedback is most important in any market or business. By using sentiment analysis, a company easily see their customers review about that product and as per the review company can do changes in their product.

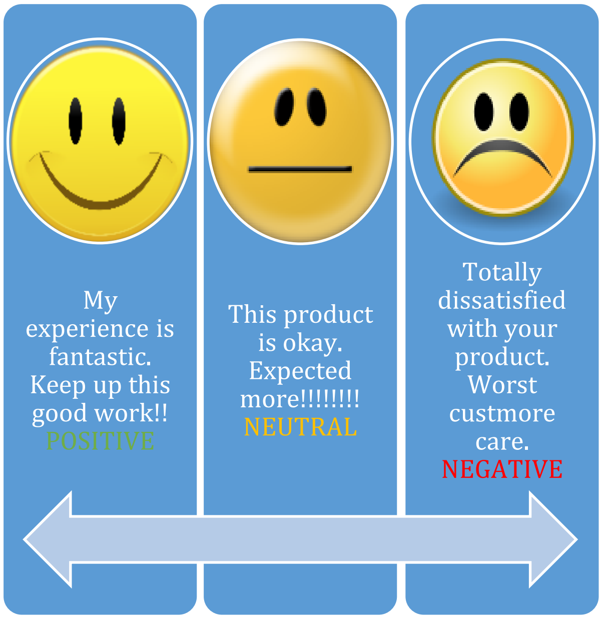
My experience is fantastic. Keep up this good work!!
POSITIVE
This product is okay. Expected more!!!!!!!!
NEUTRAL
Totally dissatisfied with your product. Worst custmore care.
NEGATIVE

Figure 1 Sentiment analysis

1.3 Advantages of sentiment analysis

1. Lower cost than customer insight support.
2. It is the faster way to collect a customer insight data.
3. It will be easy to act on the customer suggestion using sentiment analysis
4. It will become very easy to identify a strength or weaknesses of other organization or company.
5. The customer opinion will be more accurate.

2. Literature Survey

In this paper 1) Tweets are classify into the positive or negative comments using machine learning algorithm such as Naïve Bayes, Random forest (RF), Support vector machine (SVM), Unigram with Sentiwordnet and unigram with Sentiwordnet including negations are using as the input in this paper. Author derived three thousand one hundred eighty-four (3184) tweets using the tweeter API. Nine hundred fifty-four (954) positive, one thousand eighteen (1318) negative, 145 stop words have been identifying from 3184 tweets. Using. Author used feature of sentiment analysis like Bag of words (BOW), Term frequency vs Inverse document frequency (TF-IDF), Unigram with Sentiwordnet, Unigram with Sentiwordnet including negation words as an input. Author gets a conclusion that all the classifier with Unigram with Sentiwordnet and Unigram with Sentiwordnet including negation word shows higher accuracy the Bags of words (BOW) and term frequency vs Inverse document frequency (TF-IDF). Random forest algorithm with Unigram with Sentiwordnet including negation words get highest accuracy of 95.6%.

In this paper 2) authors try to use machine learning algorithm for Arabic customer's feedback. They study two different type of methods which are voting and meta-classifier combination. They collecting data using Tweepy Application Programing Interface (API)17. There are many sarcastic and neutral tweets with the positive and negative tweets. Total 438,931 tweets were collected from that 75,774 are positive and 75,774 negative. Removing the al noisy data from the tweets like pictures, hashtags, retweets, emotions; second tokenization removing non Arabic letters, normalizing Arabic analogue letters. 10 classifiers NB, ME, LR, RR, PA, MNB, SVM, SGD and Ada boost BNB were used to extract and discover the polarity of given tweets. The highest accuracy achieved by PA and RR is 99.96%. Lowest accuracy achieved by Ada boost, LR and BNB which is less than 60%.

This paper 3) uses Amazon customer review data to find out the positivity, negativity and neutrality on customer's review. In this they compare two machine learning algorithms Naïve Bayes algorithm and Support vector machine (SVM). The input is the customer review of the Amazon products. The review maybe negative, positive or neutral. Apriori algorithm is used to extract the frequently used aspects from the input dataset. Sentiwordnet is used to calculate positivity, negativity and neutrality score and after that the classifier will apply. The comparison of the algorithm based on the performance can be calculated by using the Accuracy, Precision, Recall and F-1 Measure of each classification. By the experimental result Naïve Bayes classification is batter accuracy then Support vector machine (SVM). Calculation were done by True positive sample (TP), False positive samples (FP), True negative samples (TN) and False negative samples (FN).

In this paper 4); There are many unsolicited email campaigns are one of the biggest threats affecting the users. Author combine both Sentimental analysis and personality recognition for analyze the email content. They use two different datasets to validate the proposed method. The first dataset is original dataset (CSDMC 2010 dataset) and second dataset validation dataset (TREC 2007). CSDMC 2010 spam corpus: - This composed 2949 emails messages to carry out original experiments. TREC 2007 public corpus: - In this there are 75419 emails in which 25220are legitimate 50199 spam emails. This method validated in two different datasets improving the best accuracy in the both the cases (from 99.15% to 99.24% and 98.98% to 99.18%). Further this method is also using for different validation like SMS and social media validation.

This paper 5) shows; During the pandemic of the COVID-19 whole word is suffering. Social media is the vast platform to share your thoughts any situations. Author uses the social media to analyze the people's reaction on this situation. Author portray the fact that how irrationally people are behaving in this situation. It would be easier for victim to gather some structured information from social media. Two sets of datasets have been used in this paper. #corona, #covid19, #coronavirus mostly used for this survey. In dataset-1 there were 2,26,668 tweets used as the preliminary for dataset-2 they use the tweets which were retweeted most. To fit in the model data have been categorized in train, validation and test sets. To show the accuracy unigram, bigram and trigram performed. The accuracy of dataset 1 is 81% and accuracy of dataset 2 is 75% using different classifiers. By the conclusion author came to know that social media is not useful enough to help people.

In this paper 22) author examine the Alzheimer disease stigma on twitter using machine learning technique. Machine learning technique modeled stigmatization expressed in 31150 Alzheimer disease-related tweets collected via tweeter API. In this 1% of the dataset used to train a classifier the tweet and rest 99% of the dataset. In this paper author discuss that how social media outlet affect attitude bearing in other development outcomes. Retweet were removed, other tweets which are not related to Alzheimer were removed, the keywords "alz", "Alzheimer", "dementia", "memory loss", "senility" which defined the sample of analysis. Lastly they removed the username which contain the topic name they removed. Two researcher manual coding and result are as follow: 43.41% informative, 23.79% joke, 21.22% metaphorical, 19.29% organization, 24.50% ridicule.

3. Problem Statement

Comparing different machine learning algorithm it will be easy to get an output that which algorithm perform batter in different features; So that we can get know about algorithm's accuracy.

4. Proposed System

As showing in the Figure 2

1. Data Collection: - Data collection is first step of sentiment analysis. There is different source of collecting a data like blog, movie review, social networking sites, product review. User have to collect a data using twitter API; for accessing twitter data users must have to create a twitter account which provide consumer key, consumer secret, access token.
2. Data preprocessing: -In preprocessing of data all the noise from the dataset has been removed like hashtags, URL and targeted name. Uppercase letter converted into lower case letters. Text tokenization has been done; tokenization is process which used to covert text into a token form.
3. Feature extraction: - Feature extraction is the most important task to classification. All the irrelevant term has been removed from the dataset like the word which do no express any sentiment. For feature extraction Unigram, term frequency vs Inverse document frequency (TF-IDF).
4. Classifier: - For determining the accuracy of a single classifier or for comparing the different classifiers the F-score is usually used. The formula of F-force is given in the formula. There is different classifier used for classification of data like SVM (support vector machine), Naïve Bayes, Random forest etc.

$$F=\frac{2pr}{p+r} \text{ p=precision, r=recall}$$

5. Result: - In result there will be comparison of different classifier to see how they get accuracy. Using the features like Unigram (n=1) with Sentiwordnet, Bags of word, term frequency vs Inverse document frequency (TF-IDF) etc. And there will be graph also to show the output and the working of the output.

5. Conclusion

Features like TF-IDF, Unigram and Bags of words by comparing them with the use of machine learning classifier like SVM (Support vector machine), Random forest and Naïve Bayes; it will be easy to show which feature will get the best accuracy out of them. It will give an output in graph as well as it will be shown in table format also.

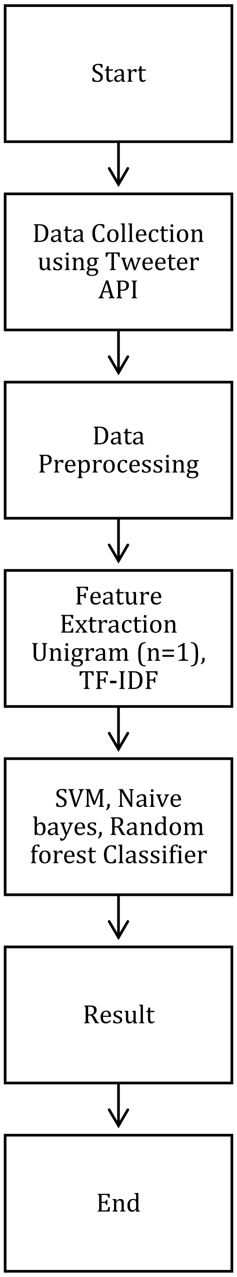
Start
Data Collection using Tweeter API
Data Preprocessing
Feature Extraction Unigram (n=1), TF-IDF
SVM, Naive bayes, Random forest Classifier
Result
End

Figure 2 Flow chart

References

1) Soumya, S. and K. J. I. E. Pramod (2020). "Sentiment analysis of malayalam tweets using machine learning techniques."
2) Gamal, D., M. Alfonse, E.-S. M. El-Horbaty and A.-B. M. J. P. C. S. Salem (2019). "Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features." **154**: 332-340.
3) Vanaja, S. and M. Belwal (2018). Aspect-level sentiment analysis on e-commerce data. 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE.
4) Ezpeleta, E., I. Velez de Mendizabal, J. M. G. Hidalgo and U. J. L. J. o. t. I. Zurutuza (2020). "Novel email spam detection method using sentiment analysis and personality recognition." **28**(1): 83-94.
5) Chakraborty, K., S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag and A. E. J. A. S. C. Hassanien (2020). "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media." **97**: 106754.
6) Ahmad, M., S. Aftab, S. S. Muhammad and S. J. I. J. M. S. E. Ahmad (2017). "Machine learning techniques for sentiment analysis: A review." **8**(3): 27.
7) Arulmurugan, R., K. Sabarmathi and H. J. C. C. Anandakumar (2019). "Classification of sentence level sentiment analysis using cloud machine learning techniques." **22**(1): 1199-1209.
8) Chaturvedi, S., V. Mishra and N. Mishra (2017). Sentiment analysis using machine learning for business intelligence. 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), IEEE.
9) Hasan, A., S. Moin, A. Karim, S. J. M. Shamshirband and C. Applications (2018). "Machine learning-based sentiment analysis for twitter accounts." **23**(1): 11.
10) Hassan, A. U., J. Hussain, M. Hussain, M. Sadiq and S. Lee (2017). Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. 2017 International Conference on Information and Communication Technology Convergence (ICTC), IEEE.
11) Kamal, A. and M. Abulaish (2013). Statistical features identification for sentiment analysis using machine learning techniques. 2013 International Symposium on Computational and Business Intelligence, IEEE.
12) Mukhtar, N., M. A. J. I. J. o. P. R. Khan and A. Intelligence (2018). "Urdu sentiment analysis using supervised machine learning approach." **32**(02): 1851001.
13) Nasim, Z., Q. Rajput and S. Haider (2017). Sentiment analysis of student feedback using machine learning and lexicon based approaches. 2017 international conference on research and innovation in information systems (ICRIIS), IEEE.
14) Pang, B. and L. J. a. p. c. Lee (2004). "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Hammad, M. and M. Al-awadi (2016). Sentiment analysis for arabic reviews in social networks using machine learning. Information technology: new generations, Springer: 131-139.
15) Ahmad, M., S. Aftab and I. J. I. J. C. A. Ali (2017). "Sentiment analysis of tweets using svm." 177(5): 25-29.
16) Ahmed, E., M. A. U. Sazzad, M. T. Islam, M. Azad, S. Islam and M. H. Ali (2017). Challenges, comparative analysis and a proposed methodology to predict sentiment from movie reviews using machine learning. 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), IEEE.
17) Kolchyna, O., T. T. Souza, P. Treleaven and T. J. a. p. a. Aste (2015). "Twitter sentiment analysis: Lexicon method, machine learning method and their combination."
18) Li, Y. and H. Fleyeh (2018). Twitter sentiment analysis of new ikea stores using machine learning. 2018 International Conference on Computer and Applications (ICCA), IEEE.
19) Madhoushi, Z., A. R. Hamdan and S. Zainudin (2015). Sentiment analysis techniques in recent works. 2015 Science and Information Conference (SAI), IEEE.
20) Narendra, B., K. U. Sai, G. Rajesh, K. Hemanth, M. C. Teja, K. D. J. I. J. o. I. S. Kumar and Applications (2016). "Sentiment analysis on movie reviews: a comparative study of machine learning algorithms and open source technologies." 8(8): 66.
21) Ortigosa, A., J. M. Martín and R. M. J. C. i. h. b. Carro (2014). "Sentiment analysis in Facebook and its application to e-learning." 31: 527-541.
22) Oscar, N., P. A. Fox, R. Croucher, R. Wernick, J. Keune, K. J. J. o. G. S. B. P. S. Hooker and S. Sciences (2017). "Machine learning, sentiment analysis, and tweets: an examination of Alzheimer's disease stigma on Twitter." 72(5): 742-751.
23) Patil, G., V. Galande, V. Kekan, K. J. I. J. o. I. R. i. C. Dange and C. Engineering (2014). "Sentiment analysis using support vector machine." 2(1): 2607-2612.
24) Shi, H.-X. and X.-J. Li (2011). A sentiment analysis model for hotel reviews based on supervised learning. 2011 International Conference on Machine Learning and Cybernetics, IEEE