

A Review paper on Sentiment Analysis using Machine Learning Technique

Reshu Gupta¹, Aniket Kumar¹, Dr. Anil Rajput²

¹Shobhit Institute of Engineering and Technology

(Deemed to be University) Meerut

²Co-Supervisor: OSD, HED, Govt. of MP, Directorate Bhopal, MP

Abstract

This article discusses various machine learning methods for sentiment analysis. Generally, machine learning classifiers such as SVM (Support Vector Machine), Random Forest and Naive Bayes are used for hypothesis testing. This article contains a lot of information that a novice researcher can use to find the right path in their studies. Here is a suggestion for a new study. The most popular platform for people to share their opinions on various topics is social media.

Machine learning is used for emotional analysis, without human intervention the machine provides good human vision. With sentiment analysis, the text can be classified as positive, negative or neutral. So any business, foundation or film critic can take public opinion into account and assist in decision making.

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

I INTRODUCTION

A machine learning tool called sentiment analysis is used to analyze the polarity of text from positive to negative. Without human input or intervention, machines analyze human emotions. Today, social media has become a part of daily life. People use social media to express their opinions on many topics such as politics, movies and business. Many social media platforms include Twitter, Facebook, Instagram and more.

They express their views on various issues through social media platforms. Therefore, sentiment analysis evaluates text given by a person from a country that uses training data to understand sentiment and determine the sentiment of the text.

1.1 **Sentiment analysis** is the study of human behavior in which we extract users' thoughts and feelings from white text.

- Emotions, thoughts, feelings, likes/dislikes are good/bad.
- Sentiment analysis is also known as sentiment mining.
- Major business needs: brand analysis / understanding new products
- There are many data such as Internet Twitter dataset, movie review dataset etc.

1.2 Application of Sentiment Analysis

1. Monitoring market research: - To determine what new products have entered the market and what consumers demand. You can adjust your business plan in light of that analysis.
2. To see the competition: - To find out what products or launches your rivals are introducing to the market. to research the rival's tactics in light of public opinion. One of the key uses for sentiment analysis is that.
3. Product Evaluation: to learn what people think of the product after it has been released or to see reactions you have never seen before. You may quickly analyse a product review by searching the term for a certain characteristic of the product.
4. Social media monitoring: - People express their opinions on social media on a variety of topics, including business, politics, the market, and other topics. You can simply track people's sentiment on different points of view by using sentiment analysis and a keyword search.
5. Customer feedback: In any industry or firm, customer feedback is crucial. A corporation may quickly see customer reviews of a product by applying sentiment analysis, and based on those reviews, adjustments can be made to the product.



Satisfied	OK	Dissatisfied
Happy	Neutral	Sad

Figure 1 : Sentiment analysis

1.3 Advantages of sentiment analysis

1. Less expensive than assistance for customer insight.
2. It is the quickest method of gathering customer insight data.
3. Making use of sentiment analysis will make it simple to implement client suggestions.
4. Finding a company's or organization's strengths or shortcomings will get much simpler.
5. Consumer feedback will be more reliable.

II LITERATURE SURVEY

2.1 Need for Literature Review

A literature review is a survey of previous researches, published articles or research papers, by accredited scholars and researchers. It includes current knowledge, substantive findings, as well as theoretical and methodological contribution to a particular topic/subject.

Soumya, S. and K. J. I. E. Pramod (2020) Using machine learning algorithms like Naive Bayes, Random forest (RF), Support vector machine (SVM), Unigram with Sentiword net, and Unigram with Sentiword net incorporating negations, tweets are classified into positive or negative comments in this article. Three thousand one hundred eighty-four (3184) tweets were derived by the author using the tweeter API. Three thousand eight hundred forty-four (3184) tweets were analysed for nine hundred fifty-four (954) positive, one thousand eighteen (1318) negative, and 145 stop terms. The author comes to the conclusion that the classifier with the highest performance, 95.6%, is the Random Forest method with Unigram and Sentiword Net, which includes negation words.

Rani and Kumar (2017) propose using natural language processing and ML as tools to help university administrators process student feedback. They used NCR EmoLex to classify emotions and infer students' satisfaction and dissatisfaction with Coursera courses (coursera.org). They observed that class performance (course grade) highly correlates with student course survey results.

Vanaja, S. and M. Belwal (2018) determines the degree of positivity, negativity, and neutrality in customer reviews using data from Amazon customer reviews. In this, the Naive Bayes algorithm and the Support vector machine (SVM) are two machine learning algorithms that are compared. Customer reviews of Amazon products serve as the input. The review may be favourable, unfavourable, or neutral. The Apriori approach is used to extract the input dataset's commonly utilised features. Sentiwordnet is used to determine the score for positivity, negativity, and neutrality before the classifier is employed. 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).

Chakraborty, K., S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag and A. E. J. A. S. C. Hassanien (2020) demonstrates that the world is in pain due to the COVID-19 pandemic. Social media is a huge platform where you may express yourself in any situation. The author uses social media to examine how individuals are responding to this scenario. The author illustrates how unreasonably people are acting in this circumstance. For the victim, gathering organised information from social media would be simpler. In this paper, two sets of datasets were employed. This survey primarily used the hashtags #corona, #covid19, and #coronavirus. The 2,26,668 tweets from dataset-1 were utilised as a starting point for dataset 2, and those that received the most retweets were used. Data have been divided into train, validation, and test sets in order to fit the model. To demonstrate the precision of the trigram, bigram, and unigram. Using various classifiers, dataset 1's accuracy is 81% while dataset 2's accuracy

is 75%. By the end, the author realised that social media was not sufficient to assist individuals.

III PROBLEM STATEMENT

Comparing various machine learning algorithms will make it simple to determine which algorithm performs better over a range of features, allowing us to assess the correctness of the algorithm.

IV PROPOSED SYSTEM

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: -All noise from the dataset, such as hashtags, URLs, and targeted names, has been eliminated. translated from uppercase to lowercase letters. Tokenization, a method used to convert text into a token form, has been applied to the content.
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. There is different classifier used for classification of data like SVM (support vector machine), Naïve Bayes, Random Forest etc.

The formula of F-Score is $2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$.

5. Result: - In result there will be comparison of different classifier to see how they get accuracy.

V 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.

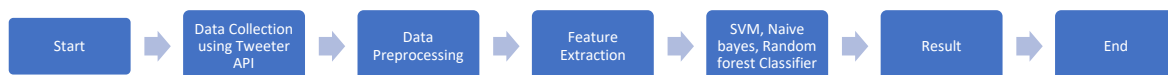


Figure 2 Flow Chart

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. Shamsirband 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) Rani, S., & Kumar, P. (2017). A Sentiment Analysis System to Improve Teaching and Learning. Computer, 50(5), 36–43. <https://doi.org/10.1109/MC.2017.133>