A COMPREHENSIVE STUDY OF SENTIMENT ANALYSIS PREDICTED ON MACHINE LEARNING CLASSIFIERS FOR VARIOUS DATASETS

Abstract

Opinion mining refers to the analysis of sentiment as most important intelligence tools. It analyzes subjective information in an expression to determine the emotional tone. It can be either optimistic, pessimistic or unbiased. Today, a huge amount of data is available on the internet as emails, social media reviews, comments etc. Sentiment analysis helps to determine the author's attitude or the public opinion on a particular topic. There are a variety of applications for sentiment analysis that will be explored in this chapter.

Keywords: Sentiment analysis, NLP, ML

Authors

Priyanka S. Hase
Research Scholar
Department of Computer Engineering
MET's Bhujbal Knowledge City,Nashik
SPPU, Pune (India)
aher.priyanka1@gmail.com

Dr. Baisa L. Gunjal
Research Supervisor
Professor and Head,
Department of I.T
AVCOE,Sangamner
SPPU, Pune (India)
hello_baisa@yahoo.com

I. INTRODUCTION

Sentiment analysis is the Artificial Intelligence tool. It is expert system which encourages to find out the emotions of a person or public. Sentiment analysis is a vital method which analyzes a huge amount of documents that are widespread and continuously increasing.[1] Sentiment analysis is an effective tool of intelligent retrieval that analyses the text data into different groups. It helps to analyze the customer's polarity of information, finding the sentiments and customers reviews.

II. MOTIVE FOR SENTIMENT ANALYSIS

Nowadays people are using social media to convey their view about everything. There is a huge amount of data created every day on the internet in the form of news, announcements, posts, messages etc. The data which is collected from all these resources is raw data which is used for analyzing the content. [2]. some regards like: brand monitoring, recommendation system, social media monitoring, and product monitoring and market analysis

III. WORKFLOW OF SENTIMENT ANALYSIS

1. **Data Collection:** Each analysis task starts with the collection of data. Nowadays many social media platforms are easily and freely available which provides a large amount of data. Real time twitter data can be accessed using tweepy library in python. Similarly, beautifulSoup is a web scrap technique which extracts the review of amazon products. Flipkart-scraper is a chrome browser extension used on flipkart.com which generates a summary of product reviews with a single click.

2. **Pre-Processing:** fter the data collection, pre-processing is a major task in sentiment analysis. Noisy data is present in the collected data. It is necessary to clean the data for further processing. Various methods are used such as removing punctuations, numbers and symbols, converting all the letters to lowercase. Tokenization is applied as it replaces a piece of sensitive data with a non-sensitive substitute, known as a token. After this lemmatization technique is applied to convert it to the base form. [3]

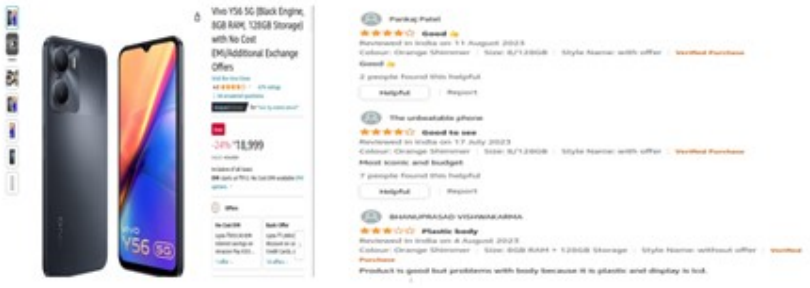

Figure 1: Example of customer review about product

3. **Vectorization:** The ML model does not understand the textual data. It is the process to convert text data into numerical vectors. Two popular techniques of vectorization are Bag of Words and Term frequency Inverse document frequency.

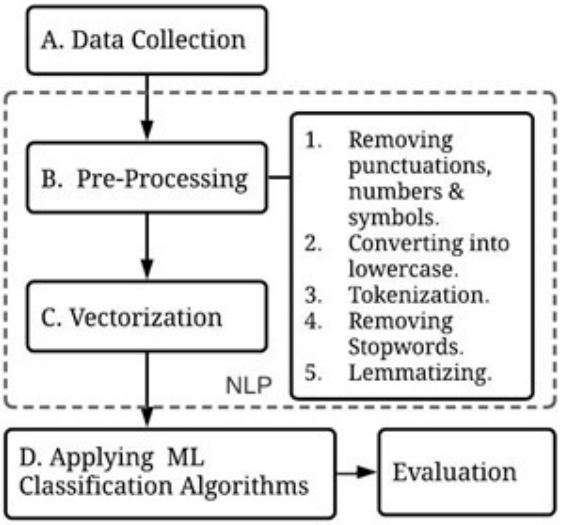

Figure 2: Sentiment Analysis Workflow

4. **Classification:** Different ML classification algorithms can be applied to determine the category of new observations based on the trained data.

IV. LITERATURE STUDY

In Ref. 8 three different techniques are used. Firstly the tf-idf algorithm is applied which extracts the keywords. Second, the PoS method is applied to extract emotions. And at last based on the emotional polarity reverse dictionary is built. Based on the results it constructs the sentiment dictionary effectively. Ref.9 proposes the analysis model which predicts sentiment lexicon and CNN. Sentiment features are enhanced in the reviews. CNN extracts the important features and classifies the features. Experiments are performed on the book reviews dataset. Results show that this model improves the performance of text mining. In Ref 10 different methods are used for working on the dataset. They used the flipkart dataset of laptop reviews. API is used to extract the data from flipkart. It used two different algorithms to classify the comments. Based on the reviews will get the idea which product is popular. The aim of Ref. 11 is to analyze the users view about using the ecommerce websites for shopping. For this two different companies are selected. The dataset used here is twitter dataset. It helps companies to identify their pros and cons and improves the marketing. In Ref.12 the algorithms are applied to extract product attributes and calculate sentiments based on the opinions from different platforms.

Table 1: Different Survey papers and accuracies of classifier

Applied Classifiers	Accuracy of Classifier	Dataset	Reference
Negation Phrase Identifier Algorithm	Polarity based sentiment analysis	Amazon.com	Ref. 13
Sentiment Scores	Positive 86%	Qualtrics.com	Ref. 14
MLRNN, MLLogistic Regression, MLSVM, MLTF-IDF	MLRNN – 54%	Manual Hindi Language Dataset	Ref. 15
SVM-TFIDF, SVM-TFIDF-Probabilistic Sentiment Score	SVM-TFIDF-Probabilistic Sentiment Score – 89.37%	AMT, Tripadvisor,Yelp	Ref. 16
SD-NB, N-Gram-NB, N-Gram_SVM, TextCNN	SD-NB- 88%	Danmaku	Ref. 17
CNN, Bi-LSTM, Bi-LSTM+A	Bi-LSTM+A – 91%	IMDB, SST,	Ref. 18
VADER	VADER – 81%	ABSA Movie Dataset	Ref. 19
CNN, GCAE, SAG CNN	SAG CNN – 81%	SemEval Restaurant Dataset	Ref. 20

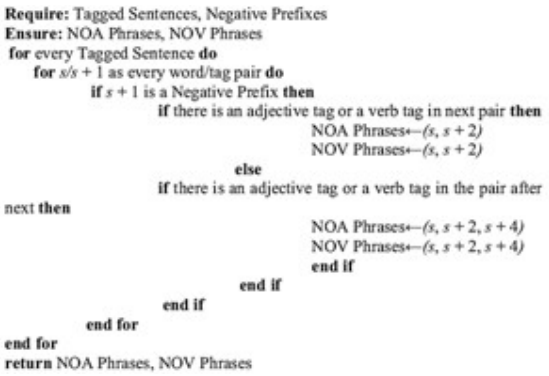

Figure 3: NPI Algorithm

J. Jabbar, N. Azeem worked on an e-commerce application.[4] The data required to work is collected from the Amazon dataset. This data is online product reviews. The main aim is to enhance the user experience in e-commerce applications based on the analysis of product reviews. The system builds the model which provides instantaneous analysis to the user by performing real time opinion mining on the reviews of products. Product reviews are the opinions given by the customers which represents thinking of people towards the

product.The opinions of the customers are in JSON format. For processing it needs to be converted in labeled form as amazon data is in unlabeled form. The Amazon dataset consists of 6 different attributes. The preprocessing phase consists of four different steps. The first step is to break the sentence into chunks as noun, symbols and expressions. It removes the character as semicolon, exclamation marks. In the second step it removes those words in a sentence which is not of any use. It helps to improve the efficiency of mining. The third step is to tag the words in sentences. It contains numeral, conjunction, pronoun and the subcategories of them. In the fourth step it removes the affixes from the words and creates the root form of the word. The next phase is opinion sentence extraction. In NLP it is a famous method to find the role of a word in a sentence. It also helps in differentiating the meaning of words that are used in different parts of sentences. Product opinion given by the customers contains at the minimum approach of the customer either positive or negative. Identification of the negative phrase is the next phase. Evaluation of the product's features is too difficult as customers use different phrases to express their opinions. The negative attitude can recognize the real situation based on the words used by customers. It is challenging so the NPI algorithm is used to recognize such phases.

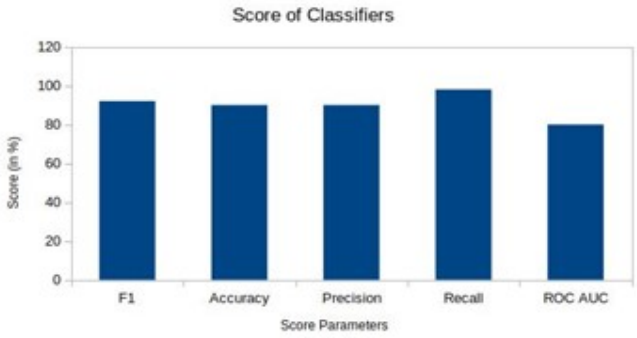

Figure 4: Classifier Scores

1. **To Evaluate the Responses of Product Reviews SVM Classifier is Applied:** HE, Zhou have explained the combination of textual analysis methods and ML algorithms. It consists of three different steps. The 1st step is to extract sentiment features. 2nd is to apply the ML algorithm to recognize sentiment conflicts of study. In the 3rd stage LDA model is applied to extract the sentiment topics from reviews.

2. **Preprocessing of Text:** Online reviews are collected from different internet platforms such as flipkart, amazon, snapdeal. Preprocessing of text mainly consists of preparation and representation of text using different methods.

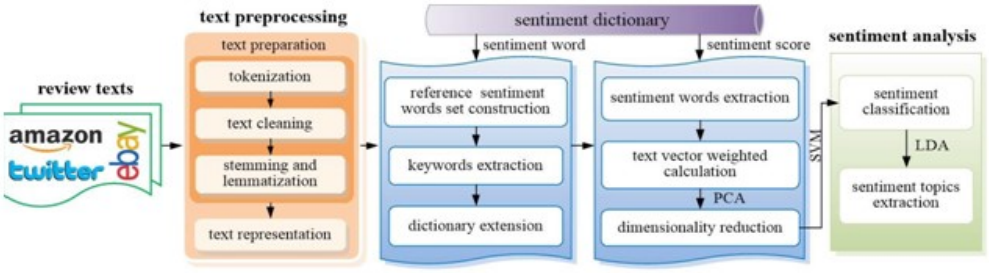

Figure 5: Research framework flow

3. **Vocabulary Extension:** As the word is internet based there is an increase in the number of the new words which are not included in previously established dictionaries. So working with the established dictionary will fail to give the correct sentiment features based on the review. So there is a need to extend the dictionary to adopt new words.

4. **Vectorization of Text:** After performing all above techniques on the text, vector p x q will be created for review. P is the number of words and q is the dimension. The formulas will be applied for vectorization.

$$\bar{v} = \sum_{i=1}^{m} \frac{w_i \times v(d_i)}{m}$$

$$w_i = \frac{2}{1 + e^{-5|s(i)/s_{max}|}} - 1$$

V. CONCLUSION

This literature study mainly focus on online shopping dataset where different classifiers and algorithms are used. We have studied papers on amazon, IMDB, Trip advisor, Flipkart etc. Each paper illustrated results depending upon different dataset word count and classifiers used. Different extraction methods and polarity is also discussed in this chapter. SVM, CNN, NB, VADER, CNN, Bi-LSTM classifier's accuracy has been discussed in this chapter.

REFERENCES

[1] B. Chen, "Sentiment Analysis From Machine Learning to Deep Learning," *2021- EIECS,* China, pp. 724-728, doi: 10.1109/EIECS53707.2021.9587971.
[2] M. Vikas, K. Balaji, "A Survey on Sentiment Analysis," *2021, 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2021, pp. 70-75.
[3] Tusar, Md. Khan & Islam, Md. Touhidul. (2021). A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data. 1-4.

[4] W. JunSheng and N. Azeem, "Real-time Sentiment Analysis On E-Commerce Application," *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, Banff, AB, Canada, 2019, pp. 391-396
[5] A. A, B. M, M. R, V. K and K. K. S R, "Sentimental Analysis for E-Commerce Website," 2022 10th (ICETET-SIP-22), Nagpur, India, 2022, pp. 1-4,
[6] G. Zhou, "Exploring E-Commerce Product Experience Based on Fusion Sentiment Analysis Method," in *IEEE Access*, vol. 10, pp. 110248-110260, 2022.
[7] M. O. Aftab, "Sentiment Analysis of Customer for Ecommerce by Applying AI," *(ICIC)*, pp. 1-7.
[8] Yanrong Zhang, Jiayuan Sun, "Sentiment Analysis of E-commerce Text Reviews Based on Sentiment Dictionary" , 2020 IEEE International Conference on Artificial Intelligence and Computer Applications.
[9] L. Yang et al.: "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning" , 2020 IEEE International Conference
[10] Santhosh Kumar K, "Sentiment Analysis of Customer Reviews on Laptop Products for Flipkart" , International Research Journal of Engineering and Technology , Volume: 05 Issue: 03 | Mar-2018
[11] Jenny Yow Bee Yin, "Exploring Sentiment Analysis on E-Commerce Business: Lazada and Shopee", Volume 11, Issue 4, pages 1508-1519
[12] Z. Yang, Q. Li, V. Charles, B. Xu and S. Gupta, "Online Product Decision Support Using Sentiment Analysis and Fuzzy Cloud-Based Multi-Criteria Model Through Multiple E-Commerce Platforms," in *IEEE Transactions on Fuzzy Systems*, doi: 10.1109/TFUZZ.2023.3269741.
[13] Xing, Justin,"Sentiment analysis using product review data",Journal of Big Data a SpringerOpen Journal, 2015, DOI 10.1186/s40537-015-0015-2
[14] Ding, Barnett,"Validating Sentiment Analysis on Opinion MiningUsing Self-reported Attitude Scores", 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS) | 978-0-7381-1180-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/SNAMS52053.2020.9336540
[15] Goel, Batra,"A Deep Learning Classification Approach for ShortMessages Sentiment Analysis", IEEE ICSCAN 2020, ISBN – 978-1-7281-6202-7
[16] Hasan, Islam,"Impact of Sentiment Analysis in Fake OnlineReview Detection", 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) | 978-1-6654-1460-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT4SD50815.2021.9396899
[17] Li,Jin,"Sentiment Analysis of Danmaku Videos Basedon Naïve Bayes and Sentiment Dictionary", IEEE Access, DOI 10.1109/ACCESS.2020.2986582
[18] Li,Liang, "Refining Word Embeddings Based on ImprovedGenetic Algorithm for Sentiment Analysis", 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) | 978-1-7281-5244-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ITAIC49862.2020.9339058
[19] Wang, She, Hu,"Importance Evaluation of Movie Aspects: AspectBased SentimentAnalysis", 020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 978-1-6654-2314-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMCCE51767.2020.00527
[20] Yang,"Aspect Based Sentiment Analysis with Self-Attention and Gated Convolutional Networks", 978-1-7281-6579-0/20/$31.00©2020 IEEE
[21] Hase Sudeep Kisan, Hase Anand Kisan, Aher Priyanka Suresh,"Collective intelligence & sentiment analysis of twitter data by using StandfordNLP libraries with software as a service (SaaS)", 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), DOI: 10.1109/ICCIC.2016.7919697,IEEE
[22] Sudeep Kisan Hase, Rashmi Soni,"Review of Sentiment Analysis on COVID-19 and Lockdown Twitter Data: Novel Techniques", Proceedings of 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication: MARC 2022, Lecture Notes in Electrical Engineering, vol 915. Springer, Singapore. https://doi.org/10.1007/978-981-19-2828-4_19

