JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Distinct Analysis Techniques for Sentiment Analysis in Social Media Based on Multilingual Process.

**Deepak Kumar, Research Scholar, MSIT, MATS University, Raipur, India**

**Dr. Bhavana Narain, Professor, MSIT, MATS University, Raipur, India**

## Abstract

The internet has been used by countless businesses to sell their goods and services. Online shoppers consistently read other consumers' opinions about a product or service before deciding to purchase it or watch a movie. Based on their comments, the business must analyze the mood and feelings of its customers. Businesses can quickly determine whether user expressions are more favorable or negative based on the sentiment analysis results. There are several different sentiment analysis methods on the market right now. However, this study will only employ three (3) methods: Meaning Cloud, Miopia, and Python NLTK Text Classification. These methods are employed to analyze the sentiment of English-language reviews and comments on social media.

Two thousand four hundred datasets from IMD, Amazon, and Kaggle were utilized to examine how accurate various methods were. According to these analyses, the average accuracy for sentiment analysis utilizing the Python NLTK Text Classification method is 74.5%, while the Miopia technique only achieves 73% accuracy. When compared to other methods, the Meaning Cloud methodology yields the best accuracy of 82.1%. This demonstrates that the hybrid approach provides the highest level of accuracy for social review sentiment analysis.

**Keywords:** Multilingual Sentimental Analysis, English Language, Social media, Gmail, WhatsApp.

## I. Introduction

In the modern era, the general public uses the internet extensively. People use the internet to find books, movies, and recreations. Additionally, they convey their feelings about such viewpoints in the meantime with the intention that It is not difficult for others to identify what exactly occurs and whether anything is worthwhile to buy or try based on the opinion of the questionnaires. The business might use sentiment analysis to examine its clients' attitude and sentiment for their goods and services. Different sentiments Analytical methods already available on the market to assist businesses in conducting the sentiment analysis for the reviews and comments left by their customers. The company's goods and services are improved as a result. Enlisting a man is incredibly difficult and tiresome to analyze the data in order to determine whether the majority of clients provide a

favorable or bad audit based on numerous comments. In this way, machine learning offered that is reliant on sentiment analysis. It speeds up the data mining process. Nevertheless, earlier research revealed that the sentiment analysis methods develop There is only one language in their model [1–3]. Multilingual individuals are able to understand many different languages. They can post reviews of their customers on social media. utilizing many languages. If the sentiment analysis technique's model is limited to The sentiment analysis technique finds it difficult to extract data in a single language.

` Twitter, for instance, generates a vast volume of opinion tweets that include several languages. Sentiment analysis with a single language model is very challenging. method for text extraction and sentiment analysis. The business is unable to evaluate their client testimonials in several languages [2, 4, 5]. Next, the use of words in colloquial terms and emoticons when leaving comments on social media. As an illustration, some the "can't, wouldn't" in colloquial English is not detectable by the sentiment analysis technique. However, some sentiment analysis techniques are able to identify it. This will result in the in-sentiment analysis, positive and negative classification will be distinct. In addition, on social media, some users frequently express themselves with emoticons (Table 1).

**Table 1: The Summarization of Multilingual processing in sentiment analysis**

| Multilingual Processing in Sentimental Analysis | Explanation |
|---|---|
| Corpus-Based with interpretation and machine learning | ▪ Sentimental analysis with python NLTK text classification<br>▪ Solve the issue of knowledge procurement issue of rule-based machine interpretation |
| Lexicon with rule-based machine translation (RBMT) approach | ▪ Denotes machine translation systems based on linguistic information about target and source language<br>▪ Miopia |
| Hybrid Machine Interpretation | ▪ Mixing the two other frameworks<br>▪ Using meaning cloud |

Emoji commenting is a popular practice among media social users. This makes it challenging for certain sentiment analysis techniques to identify the emojis. This will affect how the sentiment analysis method is evaluated. This will have an impact on the sentiment analysis's results. As a result, sentiment analysis's positive and negative groupings will change [2, 5]. Finally, the sentiment analysis's problem is the yield's correctness. No sentiment analysis method can achieve one hundred percent accuracy. Various sentiment analysis techniques will be applied in various contexts to improve sentiment analysis accuracy. The general public will wonder which technique to use and under what circumstances, which will result in the time-consuming process of selecting the appropriate strategy [1, 9] (Table 2).

**Table 2: The Comparison of Sentiment Analysis Techniques**

| Sentimental Analysis Technique | Explanation |
|---|---|
| Naïve Bayes | ▪ Apply Bayes theorem to arrange the information with assumption that the traits are restrictively autonomous |
| SVM | ▪ Divide information into paired classes<br>▪ It uses the hyperplane<br>▪ It determines linear separators in the search space |
| Neural Network | ▪ Hardware or programming designed after the assignment of neurons in the human mind in data innovation<br>▪ Deep learning technology |
| Python NLTK text Network | ▪ Open source natural language processing (NLP) platform<br>▪ Capable of textual tokenization, parsing, classification, stemming, tagging and semantic reasoning |
| General Bayesian Network | ▪ It is the probabilistic graphical model |
| MIOPIA | ▪ Evaluate the dependencies among sentences<br>▪ Do the calculation on their sum weighting |
| SentiStrength | ▪ Estimate the sentiment analysis polarity in short messages |
| Semantic Orientation Calculator | ▪ Calculate entire sentiment analysis polarity and strength of message |
| Meaning Cloud | ▪ Facebook/Twitter (Hybrid machine interpretation)<br>▪ Classification Model<br>▪ Use hybrid algorithm |

Natural Language Processing (NLP) enables viable human-machine communication. Numerous applications of NLP have been made in recent decades. A significant percentage of these are profitable in day-to-day living. Take a machine that can understand rules by voice, for example. NLP is a product of human intellect. In order to establish a connection with Personal Computers (PCs) in both written and spoken contexts using human languages, it is able to deconstruct, comprehend, and generate the languages that people often use [15]. The word restrictions are typically mixed, and the phrases that are valued are distinct. At the accompanying level, the dialect's sentence structure leads us to infer more meaning from the way the words are combined. It can be challenging to set up a program that values regular dialect. There are vast numbers of common dialects. They have a wide variety of sentences. Furthermore, ordinary language has a lot of ambiguity. Many words have a couple of possibilities, and sentences have several meanings depending on the context. This makes it a demanding task to create projects that understand a distinctive vernacular. Word restrictions are typically blended, and the sentences understood are distinct. At the accompanying level, the language's phrase structure leads us to choose how the words came together to create deeper meanings. Developing a program that recognizes that regular language is a problematic problem. Numbers of common languages are huge. They have a wide variety of sentences. Additionally, there is a lot ambiguity in everyday languages. Many terms have a few recommendations and the meaning of statements varies depending on the context. This makes project creation easier. that understands a distinctive language, a test task. Three categories of multilingual processing techniques that combine interpretation and corpus-based Lexicon with Rule-based Machine Interpretation, Machine Learning, and Hybrid Interpretation of Machines [16–19]. This work aims to investigate the multilingual processing utilized in NLP. analysis, particularly sentiment analysis. The comparison of various patterns of social media sentiment analysis will be taken into account. NLP creates a human voice. the machine's feasibility. Numerous applications of NLP have been made in recent decades. In everyday life, a significant number of these are highly profitable.

For example, a voice-activated rule-making machine [20]. There are multiple goals to accomplish, which is to investigate and assess Social Review in English utilizing three several methods of sentiment analysis, such as the Python NLTK Text Classification, Miopia and Meaning Cloud tools for classifying texts according to their polarity (positive or negative categorization), as well as to compare the accuracy performance of these three methods.

## II. Methodology

The research's scope is divided into five (5) categories:

(a) Three sentiment analysis techniques (Miopia, MeaningCloud, and Python NLTK Text Classification) will be utilized to compare the accuracy of the results in the English language.

(b) Positive, neutral, and negative classifications form the basis of the classification.

(c) The sentiment polarity dataset version 2, which includes 32938 positive and 31783 negative reviews, is the source of the datasets used. Additionally, the datasets from the reviews by Bo Pang and Lillian Lee are used.

(d) The IMDb, Yelp, and Amazon datasets were also used.

(e) Three sentiment analysis methods were employed in this study: MeaningCloud, Miopia, and Python NLTK Text Classification.

The technique flow used to complete this study is depicted in Figure 1. In order to structure the model, the first step is to store the English data in the database. This is due to the construction of the parser, the part-of-speech tag (POS-tagging), and the information control. The Kaggle, Amazon, Yelp, and IMDb databases were used in this study. The Python NLTK Text Classification, Miopia, and MeaningCloud tools are the three sentiment analysis methods employed in this study.

The list of sentiment data extracted from each sentiment analysis is shown in Table 3. The methods in the preliminary analysis will be tested using these sentiment data. Entering the English language reviews for testing is the second stage. They do the sentiment analysis. Following the sentiment analysis procedure, the reviews are categorized as neutral, negative, or favorable. The confusion matrix formula is used to determine the accuracy of sentiment analysis. Three English-language sentiment analysis approaches are used to collect the accuracy of the results. The assessment method, which compares the precision of several sentiment analysis approaches in the English language, is the final phase.

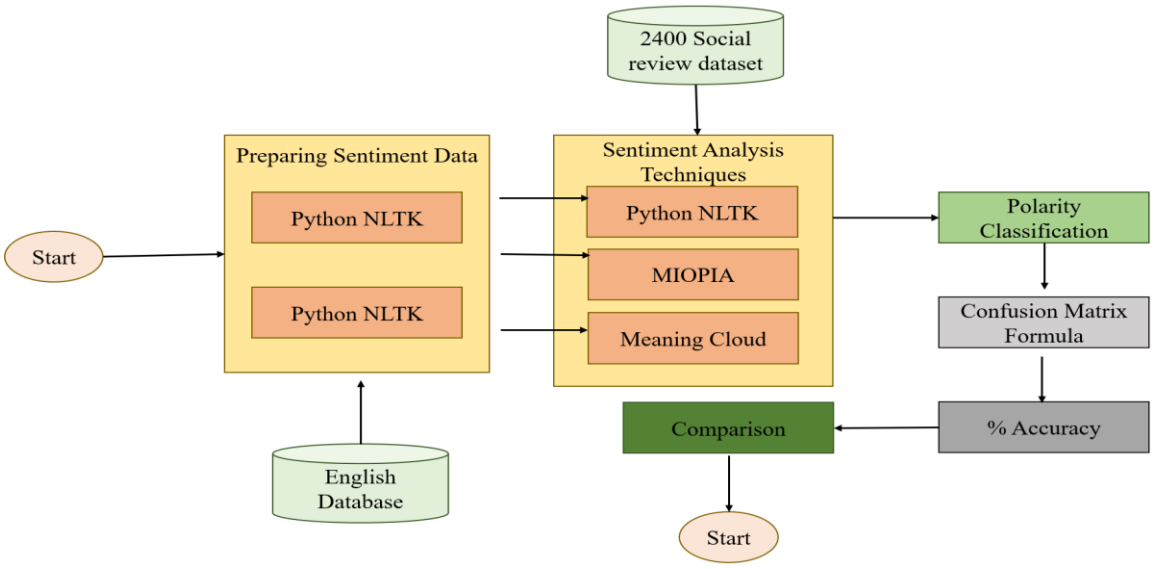



**Fig. 1: Flow Chart for the implemented Study**

# III. Result and Discussion

Table 4 demonstrates that the accuracy of the Python NLTK Text Classification (Corpus-based approach) is higher than that of the Miopia tool (Lexicon-based approach), with 74.50% and 73.00% separately. This is because the corpus-based approach uses machine learning to automatically acquire knowledge from text, requires a consistent sample corpus, and can be used for testing across a wide range of datasets. However, the Lexicon-based approach has a knowledge acquisition bottleneck.

**Table 3: List sentiment data extract form each sentiment analysis techniques**

| Techniques | Sentimental Data |
|---|---|
| Python NLTK text Classification & Meaning Cloud | Noun: Perfection, Genius, Mastery, Advantage, Decay, Love, Demand<br>Verb: Culminate, Boost, Ruin, Excel, Betray, Assist, Decrease, Waste<br>Adjective: Calm, Finest, Right, Beloved, Slim, Fierce, Scientific, Clean, Bloody, Worthless<br>English Sentence (Corpus): "Boost The Performance", "I Love My Father" Etc |
| MIOPIA | Noun: Masterpiece, Perfection Excellence Angel Zenith Attempt, Complexity, Conflict, Congestion<br>Verb: Culminate, Exalt, Conquer, Assist, Capture, Comfort, Battle, Decrease, Complain, Mangle, Curse<br>Booster Words: The Least, Nowhere Near, Not That, Noticeably, Totally, Remarkable, Entirely<br>Adverbs: Pleasingly, Enjoyable, Courteously, Professionally, Intimately, Rarely, Plainly Indifferently, Clumsily<br>Adjective: Beloved Accessible, Elegant, Worthy, Infectious, Intentional, Foolish, Freaky, Rusty, Unsympathetic, Malicious<br>Emoticons: E, :\, XO, :_(, :[, X3?, =], .:D, ^_^ |

Although the Lexicon-based method is very accurate at identifying sentiment, it has issues when tested on a variety of datasets. Because of the knowledge acquisition barrier, that method's dataset transformation are insufficient, unfinished, and time-consuming. Consequently, in the testing process, it demonstrates that the Python NLTK Text Classification tool's (Corpus-based) accuracy method) is superior to the Miopia tool (based on Lexicon) when using Kaggle datasets. the datasets from Amazon, and IMDI.

**Table 4: Accuracy Percentage for three (3) Sentiment Analysis Techniques**

| Sentimental Analysis Technique | Accuracy (Testing Work) |
|---|---|
| **Python NLTK Text Classification** | 75.56% |
| **Miopia** | 74.01% |
| **Meaning Cloud** | 83.14% |

Although the Lexicon-based method is highly accurate in analyzing sentiment, it has issues when tested on a variety of datasets. Because of the knowledge acquisition bottleneck, the approach's dataset transformations are insufficient, time-consuming, and incomplete.

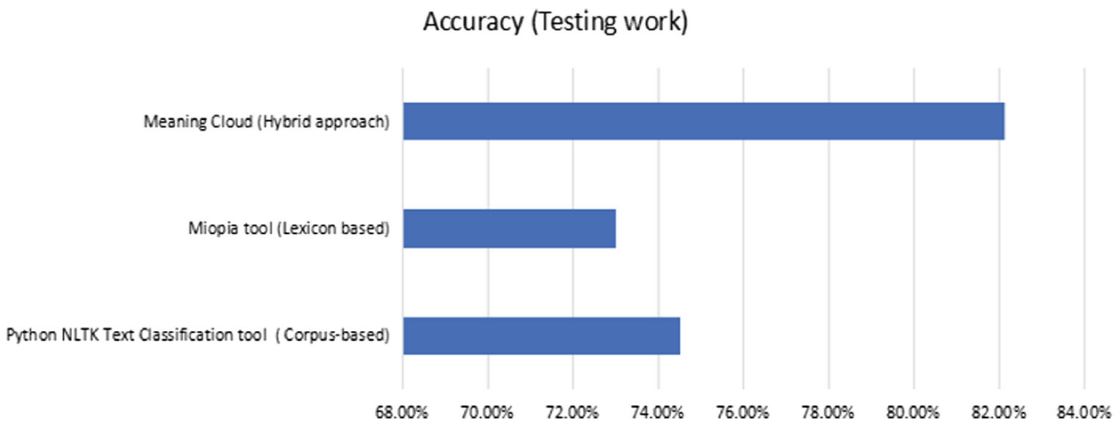



**Fig 2: Accuracy Percentage (in Bar Graph) Difference Between Three Different Techniques**

Thus, using datasets from Kaggle, Amazon, Yelp, and IMDI, the testing study demonstrates that the Python NLTK Text Classification tool (Corpus-based technique) has a greater accuracy than the Miopia tool (Lexicon-based approach).

**Table 5: Comparison Between Reference and The Testing Work Using Miopia Tools**

| Miopia Tools | Accuracy |
|---|---|
| Testing | 73% |
| Reference | 74% |

The hybrid method for sentiment analysis is called MeaningCloud. The MeaningCloud tool has an accuracy rate of 82.13%. Out of the three methods for sentiment analysis, it is the most effective. This is due to the fact that it combines elements of both corpus-based and lexicon-based approaches. It can investigate the benefits of alternative strategies. When analyzing sentiment across large coverage (datasets), it can produce more accurate results. It is capable of achieving the highest level of categorization accuracy. As a result, the MeaningCloud tool has the highest accuracy percentage (82.13%) when compared to the Miopia tools (based on Lexicon) (73.00%) and the Python NLTK Text Classification (based on Corpus) (74.50%). Fig. 2 displays the outcome of this procedure.

Table 5 and Figure 3 demonstrate that the reference work's accuracy is higher than the testing work's, at 74.00% and 73.00%, respectively. This can be the case because there are more datasets utilized in reference work than there are datasets examined in testing work. According to Table 3, the reference work utilized a total of 2 252 adjectives, 1 142 nouns, 903 verbs, 745 adverbs, and 177 intensifiers.

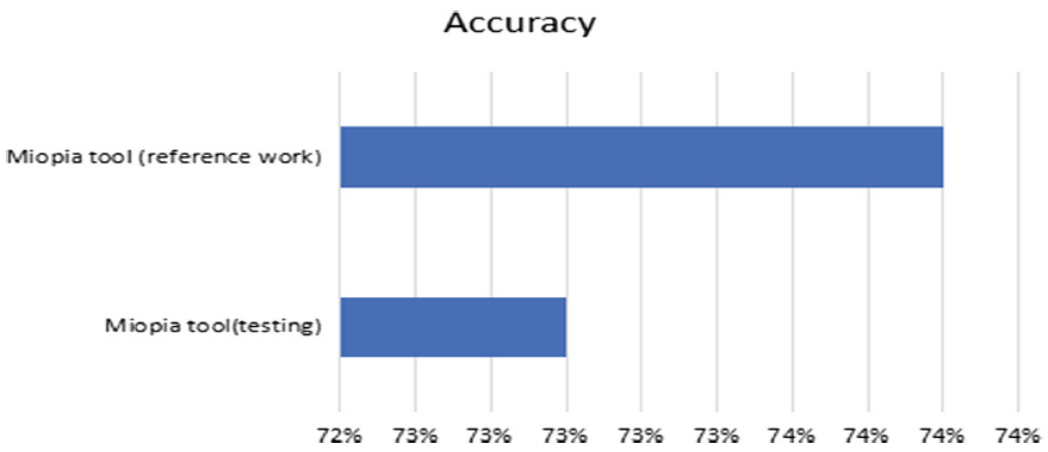



**Fig. 3: Comparison between Reference Work and Testing Work for Miopia Tools**

The more datasets that are used to test sentiment analysis systems, the more accurate the results will be. Overall, the Miopia sentiment analysis technique's accuracy was demonstrated in the testing work with an accuracy percentage of roughly 73%, which is probably comparable to the accuracy of the reference work with an accuracy percentage of roughly 74%.

**Table 6: Comparison between Reference and Testing Work Using Python NLTK Text Classification and Indian Tools**

| | Sentimental Analysis Technique | Accuracy |
|---|---|---|
| **Testing** | Python NLTK Text Classification Tool | 74.5% |
| **Reference** | India Author Tool | 74% |

The testing work's accuracy is higher than the reference work's, at 74.50% and 74.00%, respectively, as shown in Table 6 and Fig. 4. This is due to the fact that the reference work's datasets exclusively concentrate on product reviews. However, a variety of datasets are the focus of the testing work (Kaggle, Amazon, Yelp, and IMDb datasets). Another reason is because the author from Turkey employed the PMI calculation method. However, the Python NLTK Text Classification tool was employed for the testing. It is the comprehensive material that can be reliably deciphered, which will improve the sentiment analysis technique's accuracy. The disparity in accuracy between corpus-based approach sentiment analysis approaches can be attributed to these two variables.

Overall, the Python NLTK Text Classification tool's accuracy was demonstrated in testing, with an accuracy percentage of about 74.5%. This accuracy is probably comparable to the accuracy of the Turkish author reference work, which has an accuracy percentage of about 74%. The tests that relied on the suggested data (datasets) were carried out satisfactorily. During the testing, the sentiment grouping and accuracy calculation of the sentiment analysis approaches were carried out satisfactorily. The results of this testing are reliable because no significant problems arose during the experiment's execution. Three sentiment analysis techniques—Python Nltk Text Classification, MIOPIA, and MeaningCloud tool—are used to calculate and record the sentiment analysis accuracy.

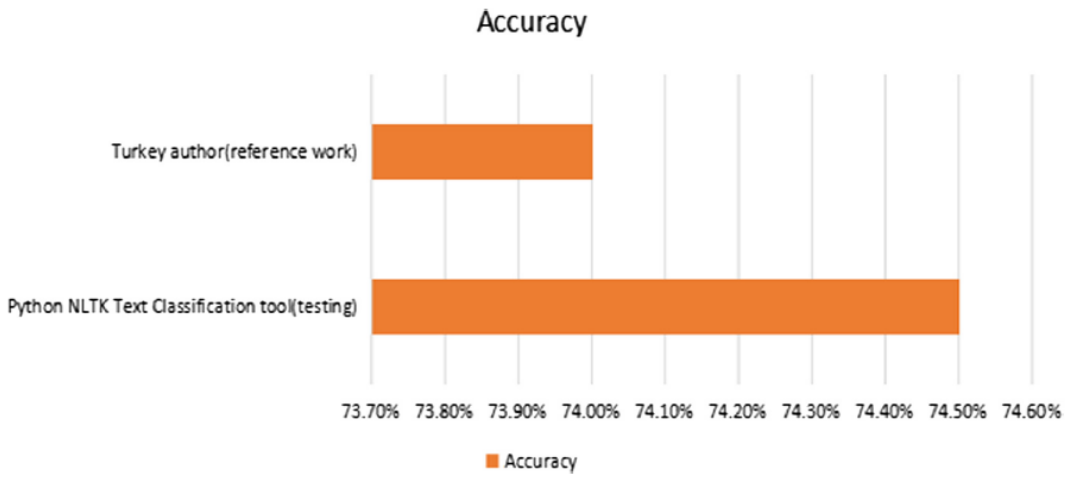



**Fig. 4: Comparison Between Reference Work and Testing Work for Python NLTK Text Classification**

The MeaningCloud tool, MIOPIA, and Python Nltk Text Classification have respective accuracy rates of 74.5%, 73%, and 82.125%. Of the three sentiment analysis methods, the MeaningCloud tool exhibits the best accuracy. This is because, in order to attain the best classification accuracy, the hybrid technique combines the features of corpus-based and lexicon-based approaches.

To determine the correctness of each technique, calculations are done by hand. Following receipt of the total amounts of analyzed true positives and true negatives, consumers need to Using the confusion matrix formula, manually determine each technique's accuracy. The only polarity value that the Python Nltk Text Classification can determine is positive. The only polarity value that the Python Nltk Text Classification can determine is positive. subjectivity value and negative polarity value. In addition, the Miopia tool can only to determine the sentiment score total. Secondly, the MeaningCloud tool can only obtain the subjective/objective, polarity, confidence level, and overall mood. Consequently, the users must figure out how many true positives and true negatives there are overall. the accuracy of sentiment analysis methods by hand and documenting it with Microsoft Excel. There are certain ways that this research could be improved for the future. labor. First, a calculator that has been designed to display the total number of true positive, true the sentiment datasets' negative, false positive, and false negative results will be quite beneficial. for determining the accuracy of sentiment analysis methods. Secondly, datasets that are not This study can make use of datasets from Kaggle, Amazon, Yelp, and IMDb. This may be beneficial. the consumers to see how reliable the sentiment analysis methods are for examining the datasets with sentiment. Additionally, this can assist sentiment analysis methods in obtaining increased precision when analyzing sentiment datasets.

# IV. Conclusion

The sentiment categorization on the Kaggle, Amazon, Yelp, and IMDb datasets was the focus of this study. Using three (3) sentiment analysis techniques—the MeaningCloud tools, Miopia, and Python NLTK Text Classification—the initial goal is to assess the polarity classification for Social Media Review in English. Comparing the accuracy performance of these approaches is the second goal. The Python Nltk Text Classification is a corpus-based system that incorporates machine learning and interpretation. The MeaningCloud tool is a hybrid machine interpretation tool, while the Miopia tool is a Lexicon with Rule-Based Machine Translation (RBMT) Approach.
The Kaggle, Amazon, Yelp, and IMDi databases were used in this study. 800 datasets were selected at random

to be analyzed with these methods. Using Microsoft Excel, the text categorization analyses for these datasets were documented.

The results produced by the MeaningCloud, Miopia, and Python Nltk Text Classification tools were examined and contrasted. Of these methods, the MeaningCloud tool exhibits the best accuracy. This is because, in order to attain the best classification accuracy, the hybrid technique combines the features of the corpus-based and lexicon-based approaches. Some enhancements that can be made to this research for subsequent work include a programmed calculator, the use of various datasets in sentiment analysis techniques, and the advanced version of offline sentiment analysis techniques.

This research is subject to certain limitations. This research is subject to certain limitations. One important factor is the quantity of sentiment datasets employed; the evaluation of the outcome may be impacted by the overall number of datasets used. To get a more accurate assessment, the total number of sentiment datasets needed for sentiment analysis approaches should be as large as feasible of the outcome. However, only 2400 records were examined for this study.

## References

1. Bhuta, S., Doshi, A., Doshi, U., Narvekar, M.: A review of techniques for sentiment analysis of twitter data. In: 2014 International Conference on Issues and Challenge in Intelligent Computing Techniques (ICICT) (2014)

2. Sarlan, A., Nadam, C., Basri, S.: Twitter sentiment analysis. In: 2014 International Conference on Information Technology and Multimedia (ICIMU), Putrajaya, Malaysia (2014)

3. Cardona-Grau, D., Sorokin, I., Leinwand, G., Welliver, C.: Introducing the Twitter impact factor: an objective measure of urology's academic impact on Twitter. Eur. Urol. Focus. 2 (4), 412–417 (2016)

4. Dashtipour, K., et al.: Erratum to: multilingual sentiment analysis: state of the art and independent comparison of techniques. Cogn. Comput. 8(4), 772–775 (2016)

5. Neethu, M.S., Rajasree, R.: Sentiment analysis in Twitter using machine learning techniques. In: 4th ICCCNT, Tiruchengode, India (2013)

6. Bhavitha, B.K., Anisha, P.R., Chilunkar, N.N.: Comparative study of machine learning techniques in sentimental analysis. In: International Conference on Inventive Communication and Computational Technologies (2017)

7. Hailong, Z., Wengan, G., Jiang, B.: Machine learning and lexicon based methods for sentiment classification: a survey. In: 11th Web Information System and Application Conference (2014) 384 N. S. A. Zulkifli and A. W. K. Lee

8. Ehsan Basiri, M., Kabiri, A.: Translation is not enough: comparing lexicon-based methods for sentiment analysis in Persian. In: International Symposium on Computer Science and Software Engineering Conference (CSSE) (2017)

9. Okpor, M.: Machine translation approaches: issues and challenges. Int. J. Comput. Sci. Issues 11(5), 159–165 (2014)

10. Al-Aidaroos, K., Abu Bakar, A., Othman, Z.: Naive Bayes variants in classification learning. In: International Conference on Information Retrieval & Knowledge Management (CAMP) (2010)

11. Jadav, B.M., Vaghela, V.B.: Sentiment analysis using support vector machine based on feature selection and semantic analysis. Int. J. Comput. Appl. 146(13) (2016)

12. Al-Saffar, M., Sabri, B., Tao, H., Awang, S., Abdul Majid, M., Al-Saiagh, W.: Sentiment analysis in arabic social media using association rule mining. J. Eng. Appl. Sci. 11(2), 3239–3247 (2016)

13. Ang, S.L., Ong, H.C., Low, H.C.: Classification using the general bayesian network. Pertanika J. Sci. & Technol. 24(1), 205–211 (2016)

14. Minn, S., Shunkai, F., Desmarais, M.C.: Efficient learning of general Bayesian Network Classifier by local and adaptive search. In: International Conference on Data Science and Advanced Analyrics (DSAA) (2014)

15. Noorhuzaimi, M.N., Ab Aziz, M.J., Mohd Noah, S.A., Hamzah, M.P.: Anaphora resolution of malay text: issues and proposed solution model. In: International Conference on Asian Language Processing (2010)

16. Ballabh, A., Jaiswal, U.C.: A study of machine translation methods and their challenges. Int. J. Adv. Res. Sci. Eng. 4(1), 423–429 (2015)

17. Al-Saffar, A., Awang, S., Tao, H., Omar, N., Al-Saiagh, W., Al-Bared, M.: Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. PLoS ONE 13(4), e0194852 (2018)

18. Yusof, N.N., Mohamed, A., Abdul-Rahman, S.: Reviewing classification approaches in sentiment analysis. In: Berry, M., Mohamed, A., Yap, B. (eds.) SCDS 2015. CCIS, vol. 545, pp. 43–53. Springer, Singapore (2015). https://doi.org/10.1007/978-981-287-936-3_5

19. Lo, S.L., Cambria, E., Chiong, R., Cornforth, D.: Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artif. Intell. Rev. 48(4), 499–527 (2017)

20. Foucart, A., Frenck-Mestre, C.: Natural language processing. In: The Cambridge Handbook of Second Language Acquisition, pp. 394–416. https://ieeexplore.ieee.org/document/667 8407

**Author Detail**

**Mr. Deepak Kumar**

**Asst.Professor , Computer Science and Application**
**Shri Shankaracharya Professional University, Bhilai, Chhattisgarh, India**



**Prof. (Dr.) Bhavana Narain**
**Professor, MATS School of Information Technology**
**MATS University, Raipur, Chhattisgarh, India**

**Prof. (Dr.) Bhavana Narain,** Professor, MSIT, MATS University Raipur. She has **23** years of work experience in the academic area and **04** years includes industrial experience. She has completed her Ph. D. in Computer Science and Application. Her research domain is Digital Image Processing, Wireless Networking and Cyber Security. She has undergone two minor projects as Co-PI and one in house project as PI. She has worked as a State student coordinator of Region 4 Chhattisgarh state in the computer society of India, from 2014 to 2020. Currently, she has been elected as secretary of the CSI Raipur chapter from 2021 to 2024.

She has published **fourteen books** in international and national publication. She has published more than **82 papers** in national and international journals. She has presented papers in national and international conferences, **seven patents** and life membership of **six societies**. She has been awarded by eight national and international bodies. She has delivered lectures as a keynote speaker and has been working as session chair in more than **four** organizations. Editor, reviewer and expert speaker in national and international platform. She is member of the University Board of Studies.

**Nine** PhD scholars and **Four** M.Phil. CS dissertations have been awarded under her supervision. Presently, **Three** PhD thesis has been submitted under her supervision and guiding **six** PhD research scholars.