International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 18th Feb 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 02

DOI: 10.48047/IJIEMR/V12/ISSUE 02/60

Title **Lexicon and Machine Learning Based Comparative Analysis to Classify the Students Opinions on Covid-19 Pandemic**

Volume 12, ISSUE 02, Pages: 380-387

Paper Authors
Dr.A. Angelpreethi

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Lexicon and Machine Learning Based Comparative Analysis to Classify the Students Opinions on Covid-19 Pandemic

Dr.A. Angelpreethi
Assistant Professor, Department of Computer Science, St. Joseph's College, Tiruchirappalli, TN, India

Abstract
Everything is online in this digital world. People express their feelings in the form of reviews on social media sites such as Twitter, Facebook, LinkedIn, and YouTube. With the use of the Internet, the voices of users are increasing day by day. Opinion mining or sentiment analysis plays an important role in classifying opinions according to user perceptions. During the pandemic (COVID-19), everything has gone digital, especially with most students suffering from it. The purpose of this study was to examine student feedback for online education systems using lexicon and machine learning based approach. The proposed Senti_Lexi and Senti_Mac approaches classify the student's opinion into positive, negative and neutral based on the polarity value. This research work is used for the educators to understand and classify the student attitudes towards online education.

Key Words: **Sentiment Analysis, Opinion Mining, Education, Twitter.**

Introduction:
At the end of 2019, the novel coronavirus (COVID-19) began to spread around the world. Most people are affected by the virus. Most teachers and students struggle with mental health issues. This leads to distance learning. Several countries have taken initiatives using other methods to improve the distance learning experience, such as social media, emails, phone calls, and even posts. Coronavirus is impacting face-to-face education systems in developing countries. Therefore, the countries need to improve their infrastructure for broadcast education, online education, and virtual classes.

Students used social networks, online games, chats and WhatsApp communications[15]. Especially during his COVID-19 pandemic, the high degree of accessibility via smartphones has greatly increased the impact of abuse on the Internet. They began expressing their feelings and opinions in the form of text messages and memes on social media platforms such as Twitter. This research focuses on using sentiment analysis to analyse student's sentiments towards online education.

Sentiment Analysis
An opinion is an individual's point of view about an entity or event. Opinion Mining (OM) is also known as sentiment analysis (SA). With the rapid emergence of microblogging sites such as Twitter in recent years, Twitter-based sentiment analysis applications have attracted a great deal of interest from online consumers seeking information about products and businesses that need to respond quickly to user opinions. This data has many features, making it difficult to analyse opinion using existing approaches. Therefore, it is an important issue to develop a method to automatically classify tweets into positive, negative, and neutral to extract and analyse Twitter data.

Opinion mining has two approaches: lexicon-based and machine learning-based approach. There are three levels 1. document-level 2. sentence-level and 3. aspect-level sentiment analysis. This paper uses lexicon based and machine learning based approaches to make

accurate predictions of student's opinions at sentence level sentiment classification.

Lexicon based Approach

This approach involves calculating orientation for a document from the semantic orientation of words or phrases in the document. The classification techniques of text involve creating classifiers from the labelled instances of texts and sentences essentially for supervised classification task. There are many lexicons existing for sentiment analysis such as WordNet, SentiWordNet (SWN), SentiTFIDF, SentiFul, SenticNet, etc. SentiTFIDF is based on term frequency and inverse document frequency across positively tagged document and negatively tagged document to classify the term as positive or negative.

Machine Learning Approach

Machine learning is one of the mounting areas of computer science that provides ability to computers to learn without programming explicitly. It is also used in sentiment analysis in which we can identify that the text generated by the user is positive, negative or neutral. Machine learning generally distinguishes two types of learning- supervised, and unsupervised learning. Supervised learning can be divided into regression and classification. some common supervised learning algorithms are Logistic Regression, Decision Tree, KNN algorithm, Simple Linear regression etc. un-supervised learning algorithms use association and clustering.

Review of Literature

Generally, the user reviews consist of slang words, acronyms, emoticons and sentiments. Major reason for using slang word is to reduce the classification time. There are many works concerning the user's sentiment posted on social media networks with the focus on classifying the opinion as positive, negative and neutral opinions. Few are discussed here.

Alattar.F et al [1] have developed a Filtered Latent Drichlet Allocation Framework for inferring variations in the user opinions from the collected tweets. The proposed framework applied several sets of multiple parameters for taking the reasons of the candidate that cause sentiment variations. The old topics of the tweets were removed. However, this framework failed to support the sentiment analysis of emoticons.

Hassan Raza H et al [2] proposed scientific text sentiment analysis using machine learning techniques. The proposed framework automatically classifies the given online news articles using existing approaches. Many classifiers were verified to get high accuracy. With the help of Bayesian classifier, the proposed work got higher accuracy.

Jawad Khan et al [9] used four sentiment classifiers for better optimization. The authors used naïve bayes, OneR, J48 and BFTree.The results were obtained with good number of correctly classified, F-measure and precision. Output reveals that OneR produced more promising results compared with other classifiers. Maite Taboada et al [16] planned a multi classification mechanism for classifying tweets into positive, negative and neutral. The authors used VADER tool to categorize the tweets on the 2016 US election. The authors suggested VADER tool is the good choice for twitter data for sentiment classification. Huge volume of data could be classified with the help of VADER.

Alani et al., [17] proposed a lexicon-supported technique for Twitter based data analysis that captures the sentiment class of words in numerous contexts and updates the sentiment scores accordingly. Their approach is based on co-occurrence word signals in different domains at both the tweet and entity levels. The proposed approach was evaluated using three Twitter datasets and achieved 4–5% higher accuracy than the comparison methods.

Preethi et al [8] discussed the different sentiment slang dictionaries like derivative slang, acronym slang and shortened slang dictionaries using lexicon-based dictionary approach. Also, the authors used emoticon and SWN dictionary for sentiment classification. The proposed approach is tested with twitter microblogging datasets.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Siddarth et al [3] discussed feature extraction, data pre-processing and various approaches of machine learning. The author used twitter data to predict the emotions using machine learning algorithms. In the future prevailing models can be enhanced by increasing semantic knowledge.

Ms. Bhumika Gupta et al. [4], in their paper, discussed the correctness of the model by putting the trained data in a machine learning algorithm which is used in the future to predict the result or to analyse the sentiment of the different dataset.

Proposed Senti_Lexi Approach

To find the effectiveness of the research the proposed work is first tested with lexical based approach (Senti_Lexi). The obtained results were compared with (Senti_Mac) Machine learning based approach.

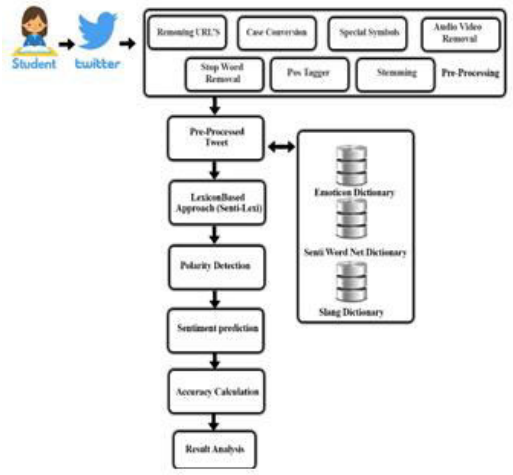


Figure 1. Methodological diagram of Senti_Lexi approach

The proposed work consists of two types of approaches namely

1. **Senti_Lexi** – A Lexicon Based Sentiment Classification Approach
2. **Senti_Mac**- A Machine Learning Based Sentiment Classification Approach

In the proposed work the tweets are collected from the students using Twitter API. The collected tweets are pre-processed and stored in the desired format. Methodological diagram of the Senti_Lexi approach is presented in Figure1. The pre-processing step includes tokenization, Stemming and lemmatization, Removal of irrelevant content, Transliteration, and POS Tagging. These processes are briefly explained in the following subsections.

Pre-Processing:

This module can gain the data from the user's tweets using twitter API. These tweets contain many attributes, which include username, date and time, location, re tweet status, re tweet count. After collecting the data from the twitter, the data needs to be pre-processing. Data pre-processing is an essential task in opinion mining during this phase, the opinion mining system prepares collected data for further processing. This involves many steps.

Tokenization.

Tweets are divided or segmented into individual words.

Lowercasing.

Textual letters are changed in to lower case to matching the words in tweets.

Stemming and lemmatization.

Words in tweets are converted to their root word for example; "ordering," "ordered," and "orderrrr" is all converted to "order."

Removal of irrelevant content.

Punctuation and stop words, which are irrelevant for opinion mining, are removed to improve system response time and effectiveness.

Transliteration.

To address the issue of use of mixed Language in tweets the text is transliterated using Google Transliterate API.

POS Tagging:

POS Tagging is also called Term Category Disambiguation and Grammatical Tagging. The Procedure of classifying the text into their Parts of Speech and tokenize them consequently.

Experimental results of the Pre-Processing:

The reviews are collected from twitter microblogging website using its API. The reviews are pre-processed. After pre-processing, the Sent_lexi approach compared the words with SWN lexicon.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

The matched words are termed as subjective words remaining are termed as objective words. Summary of the dataset is presented in Table 1. In this dataset there are 2548 slang words, 246 emoticons and 4256 subjective words and 1478 objective words are found. Stop words, negations and irrelevant words are removed in the pre-processing stage.

Table 1. Summary of the Twitter Dataset

S. No.	Review Category Type	Count
1.	Total Tweets	10,058
2.	Slang Words	2548
3.	Subjective words	4256
4.	Objective words	1478
5.	Emoticons	246

The summary of the twitter dataset is diagrammatically represented in figure 2.

Approach	Precision	Recall	F-score	Accuracy
Senti_Lexi	80.71%	66.62%	72.75%	73.38%

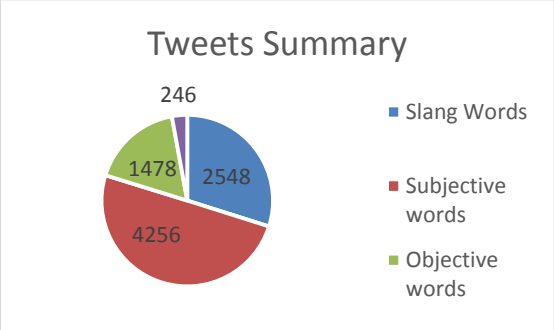


Figure 2 summary of the Twitter Dataset

The pre-processed tweets were given as an input to the Senti_Lexi Approach. Emoticon Dictionary, SentiWordNet Dictionary and Slang Dictionaries are interconnected with Senti_Lexi approach. Tweets are filtered based on SWN because SWN have only formal words. Based on the occurrence of the SWN, tweets are filtered by slang words, and emoticons. Each term of the tweet is searched whether it belongs to SWN. Based on the results of SWN polarity of the text is classified with positive, negative and neutral. SWN lexicon allocate sentiment score to each individual word. Based on the sentiment score polarity of the text is classified.

If Score>0 ➔ Positive sentiment

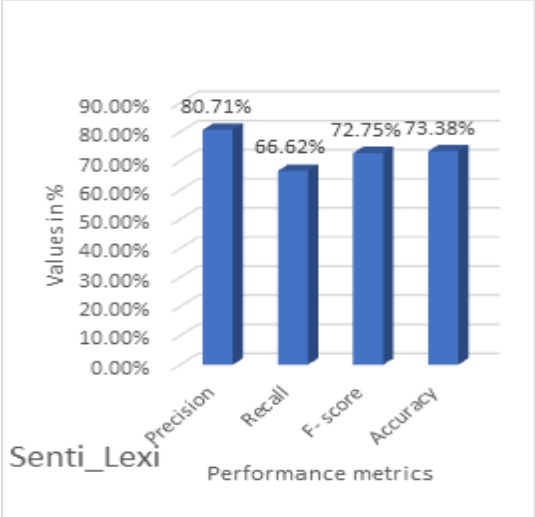
Score<0 → Negative Sentiment

Score =0 Neutral Sentiment

After classifying all the sentiments, the performance of the Senti_Lexi approach is measured based on the *following parameters such as precision, recall, F-Measure and accuracy* [5]. The overall performance and effectiveness of the Senti_Lexi approach is presented in Table 2 and diagrammatically represented in Figure 3. x axis represents the performance metrics and y axis represents the percentage of common measures.

Table 2 Performance of Senti_Lexi

Figure 3 Results of the Senti_Lexi Approach



Senti_Lexi approach is compared with the results of the existing literature. The comparative study of evaluation measures is depicted in Table 3.

Table 3 Comparison of Existing Works

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

S. No.	Approaches	Accuracy(%)
1.	*Alexander et al*	59.50
2.	Hussam Hamdan et al	64.27
3.	*Ayushi Dalmia et al*	67.04
4.	*Saprativa et al*	68.46
5.	**Senti_Lexi**	73.38

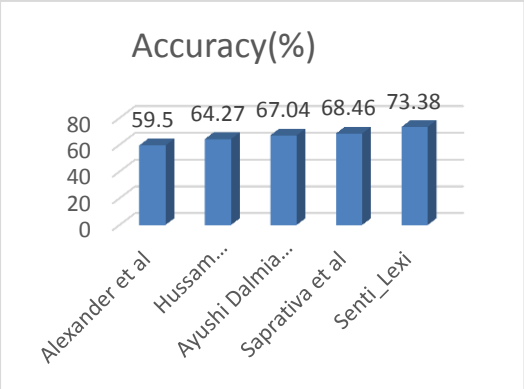


Figure 4 Comparison of Existing works

Figure 4 represents the comparison of existing works. The Senti_Lexi approach performs well and it got more accuracy. To measure the effectiveness of the research the proposed work is tested with machine learning based approach.

Machine Learning Based Classification (Senti_Mac):

The methodological diagram of the Senti_Mac is depicted in figure 5.

The pre-processed data were given as an input to the Senti_Mac Approach. The pre-processed data are to be converted into a matrix of tokens for vectorization. Because the machine cannot understand words and strings. To deal with textual data it must be represented using numbers to be understood by the machine. The count vectorizer is a technique to convert textual characters into numerical values.

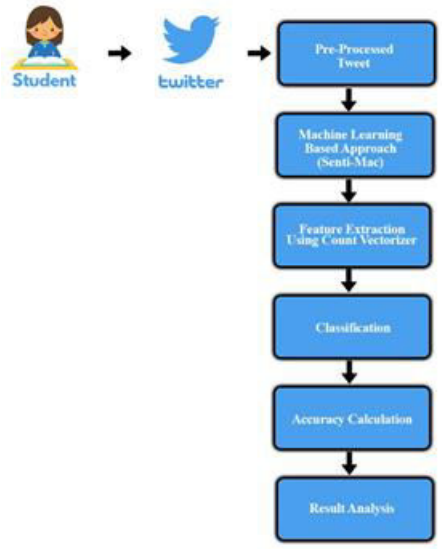


Figure 5. Methodological diagram of the Senti_Mac Approach

For example
Document 1= “The cat is White”
Document 2= “The cat is pink”
Document 3= “The cat drinks milk”
The above documents are given as
[““The cat is White”, “The cat is pink”, “The cat drinks milk”]

The above documents form a count vectorizer matrix of size 3*7 because there are three different documents and seven different features. If the feature is present 1 will be marked otherwise 0 will be marked in the matrix.

The table 4 represents the presents of features and documents. In the document 1 only four features. So, the first four features are indexed with 1 remaining are zeros. In the second document first three features are present there is no fourth feature and fifth is present. Remaining features in the second documents are zero. In Document 3 one, two, six and seventh features alone 1 remaining feature are zero. The pre-processed reviews are converted into vector values and then given as an input to the various machine learning algorithms.

Table 4. Count Vectorizer

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Features → Documents	1 (The)	2 (cat)	3 (is)	4 (white)	5 (pink)	6 (drinks)	7 (milk)
1	1	1	1	1	0	0	0
2	1	1	1	0	1	0	0
3	1	1	0	0	0	1	1

This senti_Mac used support vector machine algorithm. This algorithm categorizes the text. SVM achieves good performance in high dimensional feature space. SVM produces better results than the naïve bayes.

Thus, the output obtained from the SVM algorithm may predict that the given input documents is either positive or negative. Very few cases the results may predict that the review is neutral which means that there is no opinion or sentiment is present in the document.

Experimental results of the Senti_Mac Approach:

In our proposed Senti_Mac Approach used Support Vector Machine algorithm to classify the tweets. After classifying all the sentiments, the performance of the Senti_Mac approach is measured based on the *following parameters such as precision, recall, F-Measure and accuracy*[5]. The overall performance and effectiveness of the Senti_Mac approach is presented in Table 5.

Table 5 Performance of Senti_Mac

Approach	Precision	Recall	F- score	Accuracy
Senti_Mac	95.2%	82.3%	88.6%	96.4%

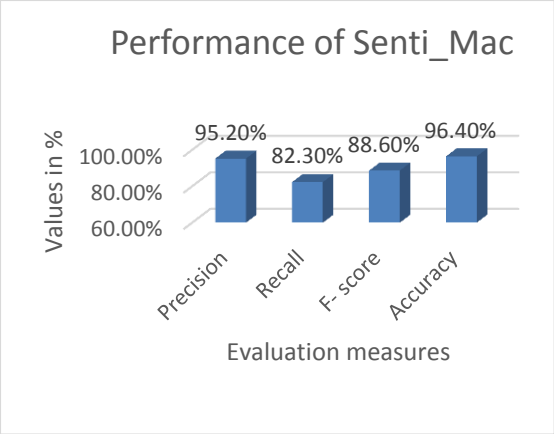


Figure 4 Results of the Senti_Mac approach

The Figure 4 represents the results of the Senti_Mac approach, where x axis represents the Evaluation measures and y axis represents the percentage of common measures.

Senti_Lexi approach is compared with the results of the Senti_Mac approach. The comparative study of evaluation measures is depicted in Table 6 Figure 7 respectively.

Table 6 Comparison of approaches

S. No.	Approaches	Accuracy(%)
1.	*Alexander et al*	59.50
2.	Hussam Hamdan et al	64.27
3.	*Ayushi Dalmia et al*	67.04
4.	*Saprativa et al*	68.46
5.	**Senti_Lexi**	73.38
6.	**Senti_Mac**	96.40

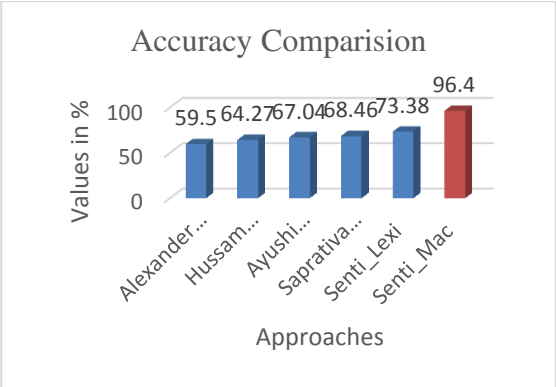


Figure 7 Comparative Results of the Approaches

From the above figure it is proved that Senti_Mac a machine learning based approach performs well compared with Senti_Lexi.

Conclusion:

The COVID-19 pandemic has impacted face-to-face education systems in most countries. This paper focuses on categorizing student learning experiences

in online modalities. Therefore, this study examined the polarity (positive, neutral, negative) of students' online learning experiences. Also, the results obtained from the proposed work based on the parameters such as precision, recall, F-Measure and accuracy. The comparative results obtained from Senti_Lexi and Senti_Mac show that Senti_Mac approach outperforms the best from the previous methods and it achieves 98.2% accuracy. So, machine learning approach (Senti_Mac) improves the accuracy of the user sentiments. Based on the results of the students the educators can able to identify the student's mentality. Hence, this study is used to classify the student mentality towards better learning and know about their status and promotes student achievement. Also the educators motivates the students to practice critical thinking at higher levels.

References

1.Alattar, and k.shaalan, "Using Artificial Intelligence to Understand what causes sentiment changes on social media", in IEEE Access, vol 09, pp. 61756-61767,

2. H. Hassan Raza, M. Faizan, A. Hamza, A.Mushtaq .,N. Akhtar.," Scientific Text Sentiment Analysis using Machine Learning Techniques" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.10, No.12, 2019.

3. Siddharth, S., R. Darsini, and M. Sujithra. "Sentiment analysis on twitter data using machine learning algorithms in python." Int. J. Eng. Res. Comput. Sci. Eng 5, no. 2 (2018): 285-290.

4. Gupta, Bhumika, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, and B. Tech. "Study of Twitter sentiment analysis using machine learning algorithms on Python." International Journal of Computer Applications 165, no. 9 (2017): 29-34.

5. A. Angelpreethi and S. B. R. Kumar, "An Enhanced Architecture for Feature Based Opinion Mining from Product Reviews," in 2017 World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, Tamil Nadu, India, Feb. 2017, pp. 89–92.

6. Ameeta Agrawal, Aijun An "Kea: Sentiment Analysis of Phrases Within Short Texts", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 380–384, Dublin, Ireland, August 23- 24, 2014.

7. Dibakar ray, " Lexicon based sentiment analysis on twitter Data", International Journal for Research in Applied Science & Engineering Technology,Volume 5, Issue X, October 2017.

8. A.Angelpreethi, Dr.S.Britto Ramesh Kumar," Dictionary based approach to improve the accuracy of opinion mining on big data", International Journal of scientific research in computer science and management studies, volume 7, issue 5 (sep 2018).

9. Jawad Khan and Young-Koo Lee, "LeSSA: A Unified Framework based on Lexicons and Semi-Supervised Learning Approaches for Textual Sentiment Classification", Received: 21 November 2019; Accepted: 13 December 2019; Published: 17 December 2019.

10. Saprativa Bhattacharjee, Anirban Das, Ujjwal Bhattacharjee, Swanpan K. Parui and Sudipta Roy, "Sentiment Analysis using Cosine Similarity Measure", 2nd International Conference on Recent Trends in Information Systems, IEEE, 2015, pp. 27-32.

11. Alexander Hogenboom, Daniella Bal and Flavius Frasincar, "Exploiting Emoticons in Sentiment Analysis", Proceedings of the 28th Annual ACM Symposium on Applied Computing, ACM, 2013, pp. 703-710.

12. Ayushi Dalmia, Manish Gupta, and Vasudeva Varma, "IIIT-H at Sem Eval 2015: Twitter Sentiment Analysis The good, the bad and the neutral" Proceedings of the 9th International Workshop on Semantic Evaluation, 2015, pp. 520-526.

13. Irina Bogdana Pugna, Adriana Dut,escu and Oana Georgiana Stănilă, "Corporate Attitudes towards Big Data and Its Impact on Performance Management: A Qualitative Study", Journal of sustainability, Volume 11, Issue 686, 2019, pp. 1-26.

14. Jasmine Zakir, Tom Seymour and Kristi Berg, "Big Data Analytics", Issues in Information Systems, Volume 16, Issue 2, 2015, pp. 81-90.

15.Charles.S "An impact of intelligent quotient and learning behaviour of students in a learning environment", CIIT

International Journals, Volume 7, Issue1, 2015.
16.Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede; "Lexicon-Based Methods for Sentiment Analysis", ComputationalLinguistics 2011; Volume 37 Issue 2,Pages 267–307.
17. Saif, H., He, Y., Fernandez, M., and Alani, H, "Contextual semantics for sentiment analysis of twitter", Information Processing and Management, Volume 52, Issue 1,2016, pp.5–19.