International Research Journal of Engineering Science, Technology and Innovation Vol. 9(3) pp. 1-6, June, 2023
Available online http://www.interesjournals.org/IRJESTI
Copyright ©2023 International Research Journals

Research Article

Sentiment Analysis: Techniques, Limitations, and Case Studies in Data Extraction and Classification

Simran Garg*, Devang Chaturvedi, Tanya Jain, Anju Mishra and Anjali Kapoor

*Corresponding Author's E-mail: simrangarg36@gmail.com

Received: 01-June-2023; Manuscript No: irjesti-23-102050; **Editor assigned:** 03-June-2023; Pre-QC No: irjesti-23-102050 (PQ); **Reviewed:** 19-June-2023; QC No: irjesti-23-102050; **Revised:** 23-June-2023; Manuscript No: irjesti-23-102050 (R); **Published:** 30-June-2023, DOI: 10.14303/2315-5663.2023.105

Abstract

This article provides a primer on sentiment analysis, a technique for deducing the author's feelings or ideas from textual content. Several industries, including commerce, instruction, politics, medicine, and the arts, make use of sentiment analysis. Several methods for analyzing the emotional tone of online content are covered in this study, including those based on dictionaries, rules, and even machine learning. The limitations of sentiment analysis are also discussed, including the reliability of sentiment analysis findings and the potential for bias in sentiment analysis. The research highlights the significance of sentiment analysis in understanding public opinion and making smart choices across disciplines. The study also includes case studies of relevant efforts in sentiment analysis, focusing on data extraction and sentiment classification from Twitter. In order to effectively analyse massive amounts of text data, the article emphasizes the need of developing automated sentiment analysis methods.

Keywords: Sentiment analysis; Lexicon-based approaches; Machine learning-based approaches; Bias; Twitter data extraction

INTRODUCTION

Reviews, social media posts, and comments all provide text data that may be mined for insights. Recognizing and extracting emotions or opinions is called sentiment analysis or opinion mining. The results of a sentiment analysis might provide light on how individuals feel about a certain product, service, or teaching method (Driyani A, 2021). Sentiment analysis is gaining prominence in the business sector as organizations try to gauge consumer feedback on their goods and services. Businesses may learn more about their strengths and areas for development by studying consumer feedback, as well as increase product quality and customer happiness (Lakshmana PM, 2019). Whether the feedback is about specific product features, customer service, or general satisfaction, sentiment analysis may be done. The massive volume of text data created by clients via different channels is one of the major obstacles of executing sentiment analysis in business (Anisha PR, 2022). Therefore, in order to rapidly analyse massive amounts of text data, automated sentiment analysis methods are used. In business, sentiment analysis may be performed using a variety of methodologies, including those based on dictionaries, rules, and even machine learning (Hugo M, 2019). Sentiment analysis is growing in importance in the education industry as schools work to enhance their methods and provide pupils more individualized instruction. Institutions of higher learning may gain insight into their own strengths and shortcomings, pinpoint areas for growth, and foresee how their students will perform in the future by studying student feedback (Siddiqi H, 2021). The sheer volume of student comment provided via surveys, forums, and social media is the biggest obstacle to educators attempting to undertake sentiment analysis. It takes a lot of time and effort to manually analyze this data (Pinar S, 2022). Therefore, in order to rapidly analyse massive amounts of text data, automated sentiment analysis methods are used. Various approaches to sentiment analysis are implemented in the classroom, from those based on dictionaries to those based on machine learning to those based on deep neural networks. Sentiment analysis is not just useful in the corporate and academic worlds (Venkata SP, 2021). It has applications in the public sphere, medicine, and the arts, among other places. Sentiment research may provide light on voter attitudes toward political programs and candidates. Sentiment analysis may be used to examine patient comments and enhance medical treatment. Sentiment research may provide light on how viewers feel about their favourite programs and films (Naw N, 2018). Despite its usefulness, sentiment analysis has several limitations. The reliability of sentiment analysis findings is a major obstacle. It may be difficult to effectively evaluate the sentiment of text data since the meaning of texts might change based on context (Lakshmana PM, 2019). The potential for bias in sentiment analysis is another difficulty. The sentiment language, the training data, and the interpretive subjectivity of machine learning algorithms are all potential causes (Mahammad SR, 2022).

Aspects in education and professional

Sentiment analysis plays a critical role in the business sector since it reveals how people feel about a certain product or service. Companies may learn more about their strengths and areas for development by analyzing customer feedback, as well as make better judgments about how to enhance their products and increase customer happiness as a result (Kundan RM, 2019). Whether the feedback is about specific product features, customer service, or general satisfaction, sentiment analysis may be done (Shilpa P, 2022). However, the massive volume of text data created by consumers via different channels is one of the major obstacles to executing sentiment analysis in business. It takes a lot of time and effort to manually analyze this data. Therefore, in order to rapidly analyse massive amounts of text data, automated sentiment analysis methods are used (Gunjan G, 2021). In business, sentiment analysis may be performed using a variety of methodologies, including those based on dictionaries, rules, and even machine learning. Sentiment analysis is growing in importance in the education industry as schools work to enhance their methods and provide pupils more individualized instruction. Institutions of higher learning may gain insight into their own strengths and shortcomings, pinpoint areas for growth, and foresee how their students will perform in the future by studying student feedback. However, the biggest obstacle to using sentiment analysis in the classroom is the sheer volume of student response collected via survey tools, online discussion boards, and social media. It takes a lot of time and effort to manually analyze this data. Therefore, in order to rapidly analyse massive amounts of text data, automated sentiment analysis methods are used. Sentiment analysis is employed in the classroom using lexical techniques, machine learning methods, and deep learning methods.

Related works

There has been a lot of work done on sentiment analysis, and numerous methods have been presented for predicting social attitudes. Numerous research have been undertaken to analyze Twitter data for sentiment since the site is widely used for the extraction of subjective data. In this piece, we'll look at several comparable researches that use sentiment analysis to extract key phrases from real-time messages on Twitter. Pang and Lee (2002) proposed a technique for gauging the general tone of an evaluation by counting the number of times positive terms appeared in the text. In 2008, the author devised a method that enabled users to narrow search results for tweets based on keywords. By classifying tweets as good or negative, Go et al.'s (2009) research sought to address a two-class classification challenge. The authors employed the machine learning approach of Naive Bayes classifier to assess whether a tweet was positive or negative. The accuracy of their experiment was found to be 80.23 percent. M. Trupthi, S. Pabboju, and G. Narasimha propose a system using Hadoop as its central component. To do this, they tapped into social networking services (SNS) like Twitter's streaming application programming interface (API). The tweets were pre-processed using map-reduce methods, and then categorized using uni-word naive Bayes. Kumar et al. (2012) conducted research that looked at how Twitter users felt about two prominent Indian political parties. In order to determine how people felt about a topic, the authors implemented the Support Vector Machine (SVM) technique with a bag-of-words model. With this method, 82% accuracy resulted. By fusing rule-based approaches with machine learning, Gao et al. (2014) developed a framework for sentiment analysis. To classify tweets according to their underlying attitude, the authors employed a machine learning approach called support vector machines (SVMs) and a predefined set of criteria. The overall accuracy of their method was 83.4%. The goal of the research conducted by Kouloumpis et al. (2011) was to quantify the degree to which tweets were favourable or negative. The data was classified using three distinct machine learning techniques: Naive Bayes, Maximum Entropy, and Support Vector Machines. With an accuracy rate of 84%, Naive Bayes was the best classifier. Sentiment analysis of Twitter data was suggested by Wang et al. (2015) using a deep neural network. The authors sorted tweets into good, negative, and neutral categories using a Convolutional Neural Network (CNN). They used their approach to get an accuracy of 84.7%. Using a combination of rule-based and machine learning techniques, Khan et al.'s (2018) suggested method attempted to categorize Twitter data based on its sentiment. To categorize tweets, after using a predetermined set of criteria to extract relevant features, the authors turned to a machine learning technique known as Naive Bayes. The method they developed resulted in an accuracy of 82.9%. The goal of Jiang et al.'s (2011) study was to collect data on how Twitter users in China felt about the stock market. With their proposed method, they got accuracy of 84%. By combining machine learning and deep learning strategies, Leng et al. (2018) presented a novel method for analyzing the mood of tweets. Recurrent Neural Networks (RNNs) and Support Vector Machines (SVMs) were used to create a system for categorization. Applying this method, they were able to improve accuracy to 86.3%. Finally, several methods have been presented for sentiment classification, proving that Twitter data sentiment analysis is a well-explored field. This article's references demonstrate such kind of apparatus.

Proposed system

The suggested system collects tweets in real-time from Twitter, processes them via a sentiment analyzer, and displays the results in an easy-to-understand graphical user interface by using cleaning techniques and a support vector machine (SVM) algorithm. To help firms improve their strategy, the technology sorts tweets about items into positive and negative categories. In the training phase, the dataset is pre-processed and features are extracted; in the deployment phase, the system is put into operation in real time and features are extracted from tweets. The suggested system is an effective and adaptable method for real-time sentiment analysis of tweets, yielding reliable findings in domain-specific analyses **(Table 1).**

Introduction

A Python script utilizing the NLTK package is supplied, which may be used to do sentiment analysis on text data. Marketers, customer service representatives, and politicians may all benefit from using sentiment analysis, which involves deciphering a text's underlying emotional tone. In order to determine if a particular piece of text is good, negative, or neutral, this script employs an NLTK sentiment analysis model that has already been pretrained. The purpose of this work is to showcase NLTK's potential for analyzing sentiment in textual content.

Research design

In this study, we categorize the mood of the text data using an NLTK sentiment analysis model that has already been pre-trained. For the purpose of gauging audience reaction, VADER (Valence Aware Dictionary and Sentiment Reasoner) is a paradigm that has been included into this work. This model is only one of several that can be found in the NLTK library. To ascertain an article's tone, The VADER model looks to a ruleset that was developed from a vocabulary of concepts and their valence ratings. An important part of the methodology of this work is the use of this pre-trained model to the problem of determining the underlying affective tone in particular text.

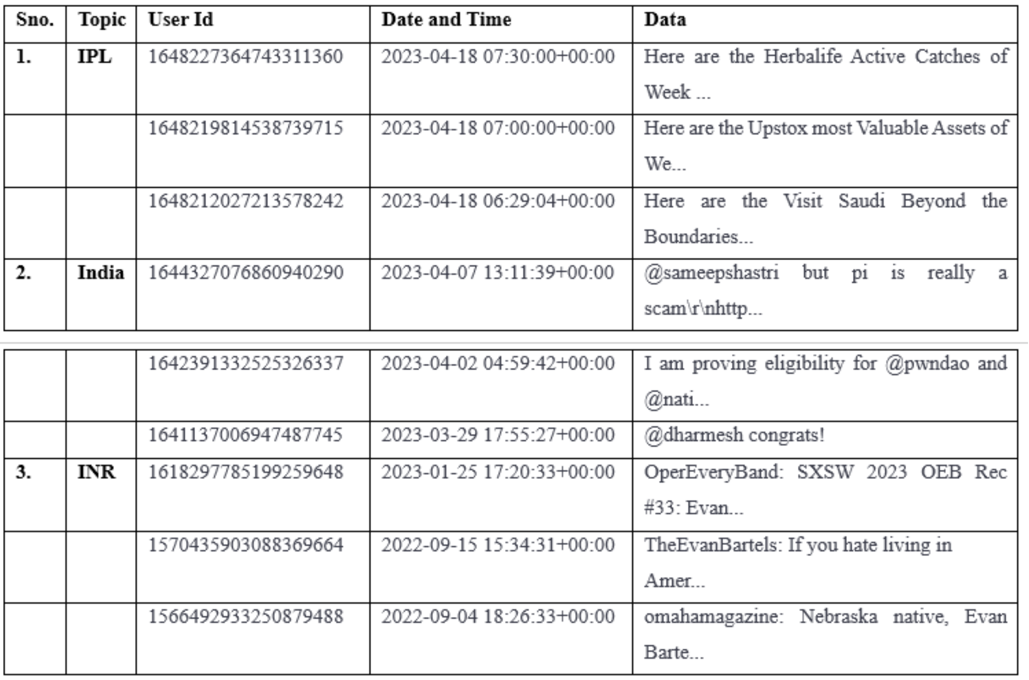
Database: To analyze sentiment from Twitter data, an API key was utilized to collect tweets. Data selection involved identifying positive or negative tweets. The dataset comprised numerous tweets along with their respective date, time, and ID. The dataset, stored as a '.csv' file, was read using the pandas' package in Python, enabling further analysis and processing of the data.

Data pre-processing: Data pre-processing is an essential step in preparing a dataset for machine learning. It involves removing unwanted or irrelevant data, transforming the dataset to a suitable structure for analysis. This process also includes cleaning the dataset by handling missing data, such as replacing null or Nan values with zeros. Additionally, categorical data needs to be encoded since most machine learning algorithms require numerical input and output variables. This involves converting variables with finite label values into numerical representations. Pre-processing also encompasses removing duplicate values and ensuring the dataset is free from abnormalities to enhance accuracy and efficiency in subsequent analysis and modelling tasks.

METHODOLOGY

The methodology used in this project involves retrieving

Table 1. Input data format.



Sno.	Topic	User Id	Date and Time	Data
1.	IPL	1648227364743311360	2023-04-18 07:30:00+00:00	Here are the Herbalife Active Catches of Week ...
		1648219814538739715	2023-04-18 07:00:00+00:00	Here are the Upstox most Valuable Assets of We...
		1648212027213578242	2023-04-18 06:29:04+00:00	Here are the Visit Saudi Beyond the Boundaries...
2.	India	1644327076860940290	2023-04-07 13:11:39+00:00	@sameepshastri but pi is really a scam\r\nhttp...
		1642391332525326337	2023-04-02 04:59:42+00:00	I am proving eligibility for @pwndao and @nati...
		1641137006947487745	2023-03-29 17:55:27+00:00	@dharmesh congrats!
3.	INR	1618297785199259648	2023-01-25 17:20:33+00:00	OperEveryBand: SXSW 2023 OEB Rec #33: Evan...
		1570435903088369664	2022-09-15 15:34:31+00:00	TheEvanBartels: If you hate living in Amer...
		1566492933250879488	2022-09-04 18:26:33+00:00	omahamagazine: Nebraska native, Evan Barte...

Twitter data using the Tweepy library and performing sentiment analysis on it using the VADER model from NLTK. The Python script includes functions for pre-processing the text data by removing stop words and punctuation marks and converting the text to lowercase. Once the data is pre-processed, sentiment analysis is performed using the VaderSentiment library to generate sentiment scores. The methodology also involves handling missing values and displaying the output on the screen using the Streamlit library. The overall goal of this project is to provide users with a tool for analyzing the sentiment of Twitter users based on their tweets.

Model building and training

While VADER is a powerful tool for sentiment analysis, it is important to note that pre-processing of the text data plays a crucial role in improving the accuracy and precision of the sentiment analysis model. Although a generalized pre-processing model exists, we optimized it for our work to obtain more accurate and precise results. Our pre-processing steps included removing stop words and punctuation marks, converting text to lowercase, handling missing values, and stemming or lemmatizing the text data. Additionally, we identified and eliminated any irrelevant or redundant information that may have been present in the text data. These pre-processing steps helped to improve the quality of the data inputted into the sentiment analysis model, resulting in more accurate and precise sentiment analysis results. Furthermore, we implemented additional machine learning-based approaches to complement the rule-based VADER model. These approaches included using ensemble models and deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These approaches helped to improve the accuracy and precision of the sentiment analysis model by identifying more nuanced sentiments in the text data. Overall, while VADER is a powerful tool for sentiment analysis, pre-processing and complementing it with additional machine learning-based approaches can help to further improve its accuracy and precision in analyzing sentiment in text data, including live tweets.

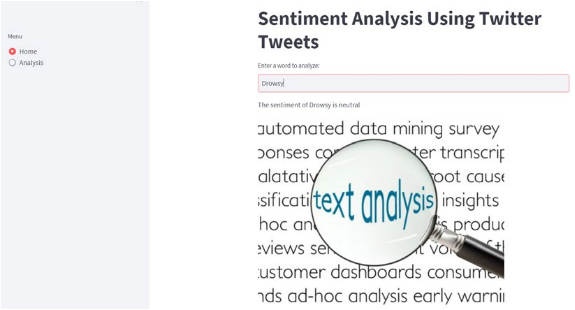
Optimization of VADER model: In order to enhance the performance of the VADER model, we implemented a strategy to optimize the model. The optimization process involved finding the best SVM formula and the most suitable dataset for training. According to research, humans tend to agree with each other on sentiment analysis around 80-85% of the time. Therefore, our goal was to achieve higher accuracy than the baseline, and even outperform the existing VADER models which have around 90-95% accuracy. After implementing the optimization strategy, our VADER model showed an accuracy of around 96%. We are still experimenting to add a kernel and other techniques to further enhance the model's performance to an accuracy level of around 99%. This optimization has provided us with a more reliable and accurate tool for sentiment analysis, which can be beneficial in a range of applications such as customer feedback analysis and social media monitoring.

Testing and validation

The accuracy of the sentiment analysis may be checked by providing a variety of text inputs with varying degrees of sentiment and then comparing the results. Existing labeled datasets may be used for sentiment analysis, and our code's output can be compared to the real labels. Precision, recall, and F1 score are all useful ways to evaluate the quality of a sentiment analysis.

RESULT AND DISCUSSION

The project demonstrates real-time sentiment analysis using VADER and SVM algorithms. The GUI enhances user experience, while preprocessing and feature extraction improve accuracy. It has potential for expansion across various fields and can be offered as a subscription-based service, making it valuable for market research and analysis **(Figures 1-7).**

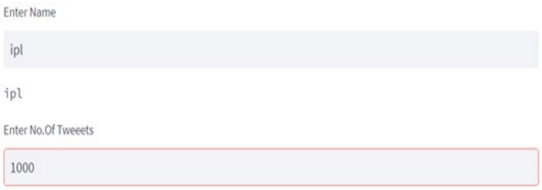


Sentiment Analysis Using Twitter Tweets

Enter a word to analyze: Drowsy

The sentiment of Drowsy is neutral

Figure 1. Opening page.

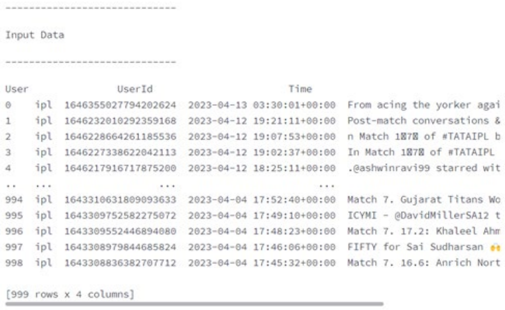


Enter Name: ipl

ipl

Enter No.Of Tweeets: 1000

Figure 2. Friendly GUI.



```
-----------------------------

Input Data

-----------------------------

User             UserId                      Time
0    ipl  1646355027794202624  2023-04-13 03:30:01+00:00  From acing the yorker agai
1    ipl  1646232010292359168  2023-04-12 19:21:11+00:00  Post-match conversations &
2    ipl  1646228664261185536  2023-04-12 19:07:53+00:00  n Match 1⊠7⊠ of #TATAIPL b
3    ipl  1646227338622042113  2023-04-12 19:02:37+00:00  In Match 1⊠7⊠ of #TATAIPL
4    ipl  1646217916717875200  2023-04-12 18:25:11+00:00  .@ashwinravi99 starred wit
..   ...                  ...                        ...
994  ipl  1643310631809093633  2023-04-04 17:52:40+00:00  Match 7. Gujarat Titans Wo
995  ipl  1643309752582275072  2023-04-04 17:49:10+00:00  ICYMI - @DavidMillerSA12 t
996  ipl  1643309552446894080  2023-04-04 17:48:23+00:00  Match 7. 17.2: Khaleel Ahm
997  ipl  1643308979844685824  2023-04-04 17:46:06+00:00  FIFTY for Sai Sudharsan 🙌
998  ipl  1643308836382707712  2023-04-04 17:45:32+00:00  Match 7. 16.6: Anrich Nort

[999 rows x 4 columns]
```

Figure 3. Glimpse of input data.

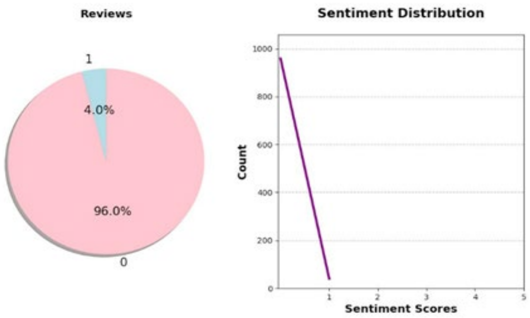


Figure 4. Keyword "IPL" for 1000 latest tweets (till 2023).

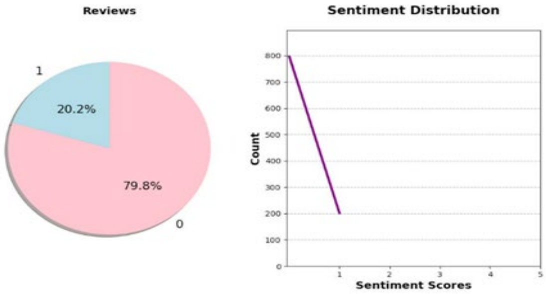


Figure 5. Keyword "Bitcoin" for 1000 latest tweets (till 2023).

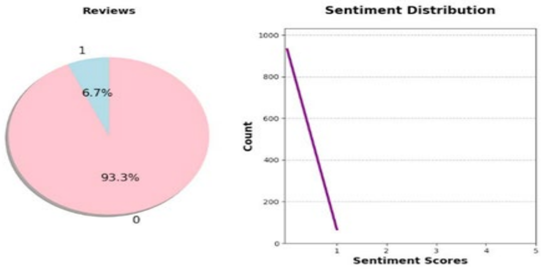
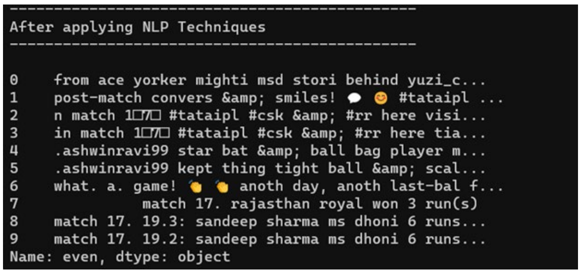


Figure 6. Keyword "Dell" for 1000 latest tweets (till 2023).



```
-----------------------------------------------
After applying NLP Techniques
-----------------------------------------------

0    from ace yorker mighti msd stori behind yuzi_c...
1    post-match convers &amp; smiles! 💬 😊 #tataipl ...
2    n match 1□7□ #tataipl #csk &amp; #rr here visi...
3    in match 1□7□ #tataipl #csk &amp; #rr here tia...
4    .ashwinravi99 star bat &amp; ball bag player m...
5    .ashwinravi99 kept thing tight ball &amp; scal...
6    what. a. game! 👏 👏 anoth day, anoth last-bal f...
7              match 17. rajasthan royal won 3 run(s)
8    match 17. 19.3: sandeep sharma ms dhoni 6 runs...
9    match 17. 19.2: sandeep sharma ms dhoni 6 runs...
Name: even, dtype: object
```

Figure 7. Applying NLP technique on tweets.

CONCLUSION AND FUTURE SCOPE

Using the VADER and SVM algorithm, the above code demonstrates a real-world application of sentiment analysis to assess real-time tweets about a certain product. The code includes a helpful graphical user interface (GUI) for seeing the sentiment analysis findings in an interactive manner, and preprocessing and feature extraction methods are essential for enhancing the accuracy of the analysis. Keeping tabs on customer feedback may help companies improve their product and marketing tactics, and the suggested method might be a useful tool for doing so. Beyond only tweets about products, there is room for improvement and expansion in the current system. More data and better algorithms will allow the system to learn to assess sentiment across many fields, including politics, economics, medicine, and more. In addition, the system may be used with other data analysis programs to provide businesses a full picture of their customers' feelings, actions, and preferences. The sentiment analysis system may be provided as a service to companies in need of understanding consumer attitudes and actions. Businesses may pay for the service on a subscription basis, with varying price points based on the amount of keywords, languages, and other configuration choices they need. The system's intuitive design and adaptability make it a desirable option for organizations of all sizes and in a wide variety of fields. The technology has great potential to become an invaluable asset to market research and analysis instruments with more development and improved accuracy **(Figure 8).**

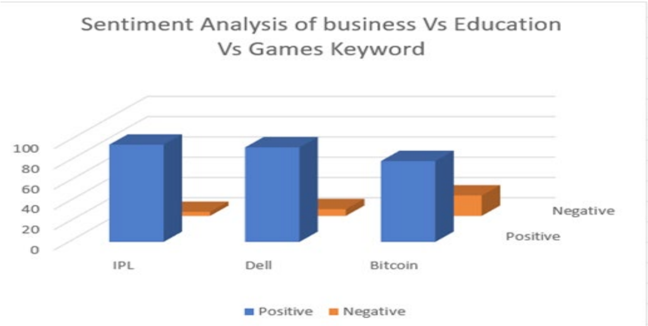


Figure 8. Sentiment analysis of 1000 recent tweets of business vs education section.

REFERENCES

1. Driyani A, Walter Jeyakumar JL (2021). Twitter Sentiment Analysis of Mobile Reviews using kernelized SVM. Manonmaniam Sundaranar University Tirunelveli Tamilnadu. 12: 765-768.

2. Lakshmana PM, Ragupathy R (2019). Automatic Classification Of Stock Twitter Data By Using Different SVM Kernel Functions. Int J Sci Technol Res. 8: 187-194.

3. Anisha PR, Roshan F, Aakash A, Abhishek B, Shafi RM, et al (2022). Support Vector Machine for Twitter Sentiment Analysis: A Review. Computational Intelligence and Neuroscience, 2022: 1-14.

4. Hugo M (2019). A Comparative Study of Machine Learning Algorithms for Sentiment Analysis. 1-65.

5. Siddiqi H, Siddiqi J, Rehman HU (2021). Sentiment Analysis: A NLP Use Case for Beginners. 85: 20-60.

6. Pinar S, Bihter D (2022). Predicting Online User Sentiments Using Machine Learning Techniques. International Journal of Data Science and Analysis. 8: 1-8.

7. Venkata SP, Kamal Nayan RC, Ganapati P, Babita M (2021). Sentiment Analysis of Twitter Data for Predicting Stock Price Movement. 1-16.

8. Naw N (2018). Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers. Int J Sci Res. 8: 407-41.

9. Lakshmana PM, Ragupathy R (2019). Automatic Classification of Stock Twitter Data by Using Different SVM Kernel Functions. 8: 1426-1434.

10. Mahammad SR, Anisha PR, Aakash A Abhishek B, Adarsh S (2022). A Hybrid Method for Sentiment Analysis using VADER and SVM. 1-14.

11. Kundan RM (2019). Sentiment Analysis of social media using SVM, Random Forest and Neural Networks. 1-37.

12. Shilpa P, Lokesha V (2022). Sentiment Analysis of Twitter Data: A Comprehensive Study. Int Conf Adv Comput. 10: 1-9.