iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65276

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

A Framework for Sentiment Analysis of Online News Articles

Sagar Vakhare[1], Aditya Yadav[2], Amarjeet Kannaujiya[3], Ashutosh Pandey[4]
[1, 2, 3, 4]*Assistant Professor, Rai School of Engineering, Rai University, Ahmedabad*

Abstract: ***In many respects, the traditional way of life has altered as a result of the current innovation era. The vast amount of data that is being constantly and continuously disseminated by a large number of users via online journals, reviews, comments, news blogs, microblogging websites, social media, and other platforms has overflowed into information technology. Owing to the sheer volume of evaluation of valuable web assets, sentiment analysis is a major focus of the research. The process of identifying and removing abstract data from content using text analysis and natural language processing techniques is known as opinion analysis. News analysis can be used to map a company's behavior over time and provide important, useful information about companies. In order to characterize the general inclination or state of mind of customers as reflected in online life toward a specific brand or organization and determine whether they see positively or negatively, sentiment analysis is also useful in web-based life checking. For some users these days, reading the news online is a daily exercise. People's opinions will generally change depending on the news they experience. News reports on events involving emotions, whether positive, negative, or neutral. Sentiment analysis is a useful tool for studying the human feelings present in textual data. There are numerous difficulties in locating the sentiments news articles.***

Keywords: ***Natural Language Processing, News Analysis, Opining Mining, Sentiment Analysis, Text Classification, Text Mining.***

I. INTRODUCTION

The public sphere, everyday life, and choices of common people have undergone remarkable transformation since the advent of the internet. Everyone reads news articles online and watches advertisements for movies, products, and books these days before actually spending money on them. It affects a person's public behavior in the same way that it has altered their way of life. The way the new web-based world is presented—for example, through websites, news channels, message sheets, and news articles—is influencing people's public behavior and how they glance at different things in their environment. People will generally experience a shift in opinion depending on the news they read. The web's signature space has now been absorbed by the life on the web. There is constantly an enormous amount of information available on the new client-driven web. Customers are not just latent buyers anymore; they are also co-creators of content on the internet. The client is currently expanding their web presence by adding blog entries, news, audits, photographs, videos, and articles. With this update, a large portion of the content on the internet is now unstructured. Access to popular communication platforms has made it possible for people to form opinions, judgments, assumptions, assessments, and frames of mind about a variety of subjects, including objects, events, problems, relationships, individuals, services, and their attributes. These days, such data has a huge capacity. We divide this data as a part of our daily lives through the use of Online Social Networks (OSN) to facilitate better learning and coordination within our global community. Owing to the daily influence of these networks, we begin to replace our meaningful decisions and actions with particular suggestions already in place from other people's audits. Because of this, it's interesting and challenging when patterns in these kinds of OSNs shift periodically. Extensive analysis can lead to reasonable expectations. It examines the value of sentiment analysis, a tool designed to distinguish opinions, and offers an advantage to text mining. The rapid development of the web, portable technology, and the internet has altered how people read the news. Physical newspapers and traditional magazines have been supplanted by online news sources and weblogs. The two factors that are driving readers' interest in shifting to online news are promptness and intuition, according to [1]. These days, people must consume as much news as they can from the various sources regarding topics that interest them or are important to them. Intelligence refers to the innate tendency that most people exhibit, which leads them to spend money on their advantage. Quickness is a factor that makes people more concerned with getting the news quickly these days [2]. Because of the innovation, we adapt, allowing people to benefit by providing them with all the information they require on a daily basis. Online news sources have developed strong strategies to attract readers' attention [3]. Online news writers express their emotions about news items, which can include people, places, or even objects, while also writing about recent events [4].

As a result, various news sites' channels offer intuitive feeling rating facilities, i.e. e- News can be neutral or positive, and it can be of any opinion [5]. Sentiment analysis, also known as opinion mining, is the process of identifying the quality or extreme of an opinion expressed in written content [6-7]. The manual marking of feeling words will be a laborious process. There are two well-established approaches for sentiment analysis. The first step makes use of a dictionary of weighted words, and the second step relies on machine learning strategies. In order to determine the polarity of words, the second methodology compares sentimental words with a given lexicon dictionary using a word lexicon method. This methodology does not require pre-processing of data or model training, in contrast to AI techniques [8]. Opinion mining is a type of natural language processing tool that tracks what people think about a particular topic or item. It is also known as sentiment analysis; it involves creating an application to collect and observe different sentiments about the text of different tweets, polls, comments, and blog posts. Mood analysis can be used in various fields, to for example, to launch a new element on the market. It is also Help in the decision -making process Advertising campaign. It allows to discover the types of items that are widespread on the market and even to distinguish what is particularly liked or disliked by socio-economic groups. Sentiment analysis also encounters certain difficulties. A specific sensation can be positive Each one and other situations are negative. People do not express the same opinion The atmosphere is in each case. In Opinion Mining, in any case, "The pictures were great" is not at all similar to "The pictures were not great." People can be inconsistent in what they say. After analysis, we can see that Most of the reviews contain both positive and negative remarks. While facts are verified in traditional text analysis, sentiment analysis focuses on human relationships. Sentiment analysis covers many research areas including opinion generalization, sentiment classification, feature-based sentiment classification. Features You can use sentiment classification to manage your entire report. Specific articles as indicated by their sentiment towards.

II. RELATED WORK

Opinion mining is a broad and developing field of study these days. Text may convey both objective and subjective feelings. Subjective text is defined as the presence of an individual's diverse linguistic expressions, including beliefs, evaluations, emotions, sentiments, speculations, and opinions [2-5]. Sentiment analysis is a modern field of information retrieval that focuses more on the viewpoint expressed than the subject matter of the document [9]. A broad spectrum of human emotions can be captured through sentiment analysis, and the majority of sentiment analysis research focuses on identifying the polarity of a given text. This implies that a specific message about a topic is automatically classified as belonging to positive or negative sentiments [10]. There are a number of uses for polarity analysis, particularly in relation to news articles. Pang and Lee (2012) distinguish between two categories of sentiment analysis techniques: machine learning-based and lexical-based. Machine learning techniques employ supervised classification algorithms, which classify sentiments as either positive or negative. This approach uses labelled data to train the classifier [11]. 69907 headlines from different news websites, including The New York Times, Reuters, the BBC, and the Daily Mail, were used in an experiment utilizing a variety of sentiment analysis techniques. By extracting the properties from the text of news headlines, they looked into the opinions regarding the polarity of these headlines. According to their research, the polarity of the headline affected the news article's level of fame. According to the study's findings, people are less interested in neutral news headlines than in positive and negative ones. Godbole and associates. created a lexicon-based algorithm in 2007 that uses the appearance of the entity in the same sentence to identify the feeling words and objects assigned in news texts and blogs [7]. From seven distinct domains—politics, business, sports, crimes, health, and general—they chose news articles and blogs. The experiment consists of two primary tasks: the Subjectivity task measures the amount of sentiment an entity holds, while the Polarity task measures the positive and negative sentiment related to each entity. Both the Subjectivity and Polarity calculation scores have performed. Instead of using dynamic corpora, they use static corpora that have been crawled online. Islam along with others. (2017) put forth a system for categorizing news articles found online. Sentence levelsentiment analysis has been performed, and sentiment polarity has been determined using a dynamic dictionary that includes a selection of words that are both positive and negative [8]. The following tasks were completed in this experiment. The process involves selecting and extracting sentences from news websites, looking for positive and negative words within the sentences, describing their polarities, and calculating the final polarity of the news article by calculating the polarity of all the sentences. With this experiment, 91 percent accuracy has been attained. Meyer et al (2017) use machine learning and a lexicon-based approach to analyze sentiment in financial news articles. Specific findings and results were obtained from eight experiments [12]. Sentiment polarities have found using a lexicon-based approach, which uses the General Inquirer Lexicon (H4N) alongside the Bag of Words (BOW) model. Machine learning approach was used for Parts of Speech (POS) syntactic model and using this approach better accurate result was obtained. Agarwal et al. (2016) employ Python libraries for conducting sentiment analysis to categorize words, and to pinpoint words that are positive or negative, utilizing SentiWordNet 3.0 [14].

The total influence of news sentiments has been calculated with high accuracy, though this method is notably time-consuming. Lei et al. (2014) have developed a model for identifying human emotions triggered by news stories and tweets. The model incorporates various modules, including document selection, lexicon generation, and parts of speech (POS) tagging. Initially, the model creates a training set, after which it applies POS tagging and feature extraction methods. In their study, Fong et al. (2013) discuss various machine learning methods and compare different algorithms for efficient sentiment analysis [16]. The text is classified into positive, negative, and neutral classes. The researchers suggest that the Naïve Bayes classifier can yield better results compared to other classifiers such as c4.5, decision tree, maximum entropy, and winnow classifiers, as it offers higher accuracy. Zhou et al. (2013) have conducted a study where they developed the Tweets Sentiment Analysis Model (TSAM). This model allows the gathering of people's opinions and social interests regarding a specific social event [17]. The authors used the Australian federal elections in 2010 as an example dataset. They analyzed the emotions expressed and the sentiment of the text. The study suggests that it is more effective to only consider words that have some sentimental value, instead of using all words for sentiment analysis. To enhance the accuracy of the classifier, the researchers proposed a lexicon-based sentiment analysis system that incorporates various techniques, including Naïve Bayes. n their study, Li and Liu (2010) utilized a k-means clustering algorithm to create a method for sentiment analysis. To incorporate weighting on the raw data, they employed the TF-IDF technique [18]. Through the use of a voting mechanism, they were able to achieve a consistently improving clustering result. By employing multiple iterations of the clustering process, they were able to attain the desired outcome. Additionally, they utilized term scores to enhance the clustering result. The documents were categorized into positive and negative groups for further analysis. Popescu and Etzioni (2007) introduced a methodology termed OPINE, which constitutes an unsupervised information extraction system designed to isolate opinions and product attributes from consumer reviews [19]. In the initial phase, OPINE identifies noun phrases from the reviews and subsequently returns those phrases that exhibit a frequency surpassing a predetermined threshold; to accurately extract specific properties, these phrases are evaluated by OPINE's feature assessor. To determine the opinion lexicon, OPINE employs a set of manual extraction rules. Somprasertsri and Lalitrojwong (2010) formulated a methodology aimed at identifying the semantic relationships that exist between opinions and product features [20]. The core of their approach involves mining both opinions and item characteristics by leveraging semantic and syntactic information through the application of ontological knowledge and dependency relations, utilizing a probabilistic-based model. Theussl et al. (2009) employ the R package tm to introduce a framework for extensive sentiment analysis [21].

They derive sentiment scores by utilizing the polarity of annotated terms in conjunction with the static text corpus from the New York Times. The authors further elucidate the application of distributed text mining methodologies alongside the MapReduce paradigm. Additionally, they critically assess the contributions of related scholars and articulate the concept of deriving sentiment scores from articles published across various news websites, correlating these scores with economic conditions. Their examination encompasses the entirety of the text, neglecting the filtration of the corpus to specific relevant text and excerpts. A comparison of multilingual sentiment analysis using supervised learning and machine translation is conducted by Balahur et al. (2011) [22]. The piece explains techniques for classifying sentiment using supervised learning and collecting multilingual data via machine translation.

The work makes use of metaclassifiers, three distinct machine translation systems, algorithms, and other aspects. The multilingual data used in this study's sentiment analysis demonstrates that machine translation technologies are a superior option for deployment. An technique put forward by Goonatilake and Herath (2007) demonstrates the safe correlation between news items, economic indices, and oil prices [23]. They determined how news stories affected important indexes like the Dow Jones Industrial Index (DJIA), S&P 500, and NASDAQ. They divide the news stories into four groups and create a regression model to quantify the daily changes in stock prices. Breen (2011) suggested an airline satisfaction mechanism where the algorithm uses keywords to determine sentiment and polarity scores. Twitter stream tagging [24]. By employing this method, the writers effectively elucidated the concept of airline customer sentiment and satisfaction ratings.Yu et al. (2007) presented a text mining technique wherein the system learns the viewpoints of news articles and explains how they affect energy consumption [25]. By quantifying the news sentiment and representing them as time series, this method compares the market ups and downs of energy prices and requirements. An strategy that links online news websites' reporting with stock exchange trade numbers was presented by Agić et al. (2010) [26]. The authors clarified that attitudes expressed in news items that derive from a trend's period are linked to broader market movements. They discovered that the news stories' overall sentiment is controlled by the polarity terms they contain. Using a rule-based polarity word identification module, they also investigate the financial domain for categorization and automatic polarity detection in unread texts [26].

III.PROPOSED WORK

The sentiment of an online news article has been analyzed using a technique based on a lexicon. Both methods that require supervised and those that do not can be applied to sentiment analysis. In the supervised method, a collection of labels is used as the training data to create a classifier model, which is then tested on data that does not have labels. Conversely, in the unsupervised method, no training data is used in the creation of the model. Instead, the polarity of words is utilized to determine the sentiment of the words.

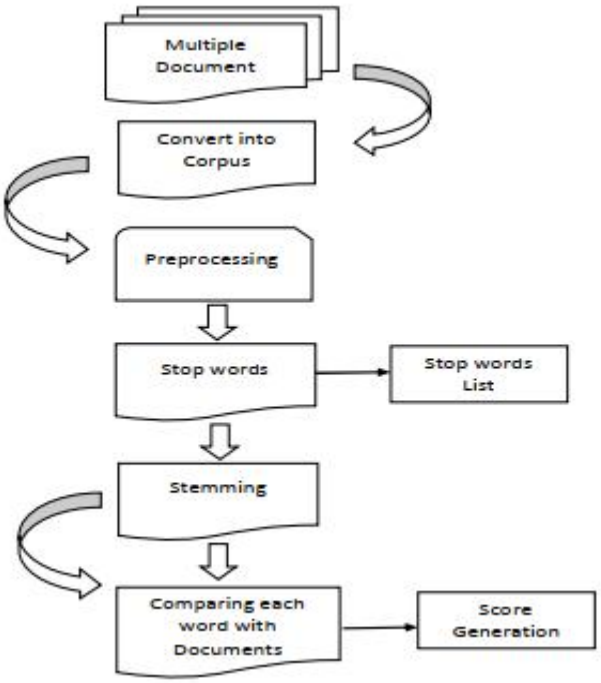


Figure 1: Proposed Approach

At the level of individual words, sentences, or entire documents, sentiment analysis can be applied. The current work concentrates on analyzing documents at the document level, identifying different sentiments such as positive, negative, or neutral within the articles from news websites. This method of sentiment analysis employs the WordNet lexical dictionary. The process is divided into five steps. Initially, data is gathered from various news websites. Following that, data preprocessing is carried out to reduce any inconsistencies. The WordNet lexical dictionary is utilized to determine the sentiment polarity of words. The details of these steps are outlined below.

A. Data Collection

The present research work used Indian Express new dataset which contains 10000 documents from the website corresponding to stories in five popular areas, i.e. Business, entertainment, education, Sport, Tech etc. each classes carry 2000 to 2500 article and the file format of these file is .txt.

B. Data Pre-processing

The text of the news article has undergone preprocessing to reduce inconsistencies, making the dataset more suitable for sentiment analysis. The initial step, "Tokenization," involves breaking a collection of sentences into individual symbols, phrases, or words, referred to as tokens. During this process, punctuation marks are eliminated. Following that, the "Filter Stop Words (English)" operator is applied to eliminate stop words. Finally, the text undergoes stemming, this operation reduces inflected or derived words to their base form.

C. Polarity Computation of Words

After the preprocessing is completed, TF-IDF is utilized to assess the significance of specific words within a document. The Term Frequency-Inverse Document Frequency method is employed to pinpoint the words that frequently occur in a manuscript, which are

then considered important, and corresponding weights are assigned to them. Following the identification of these important words, sentiment scores are attributed to them using a lexical resource. In this study, the WordNet dictionary, a lexical database for the English language, is utilized. The WordNet dictionary contains over 90,000 distinct word senses and 118,000 different word forms [17]. The choice of the WordNet dictionary is based on its capability to identify opinion words and allocate sentiment scores effectively.

D. Calculate Total Sentiment Score

Following the polarity analysis of individual words, phrases, or sentences, each document is associated with a specific polarity. In this context, every news article is regarded as a document. The polarity attributed to a document is obtained by summing the polarities of all words, phrases, and sentences within the news articles. Once the polarity is determined, the sentiment of the news articles is assessed. A sentiment score of +1 indicates a positive sentiment, while a score of -1 signifies a negative sentiment. A score of 0 denotes neutral sentiments. The sentiment scores are calculated utilizing the SentiWordNet 3.0.0 dictionary. This dictionary serves as an enhanced version of the WordNet dictionary. Synset IDs are employed to bridge the connection between WordNet and SentiWordNet. The overall sentiment of the news articles is computed using a scoring function based on the WordNet and SentiWordNet dictionaries.

E. Sentiment Results

News articles have been categorized as having good, negative, or neutral attitudes based on their overall sentiment score. Next, by calculating the average of all the words sentiments, sentiment analysis of news stories was done.

IV. RESULT DISCUSSION

As mentioned in the previous section, news articles from the indian express news paper are used in this experiment. One month's worth of news stories were crawled by the experiment. Following the introduction of sentiment scores for news stories, a score of +1 is seen as positive, a score of -1 as negative, and a score of 0 as neutral. The experimental results are displayed in the table below. It has observed that maximum of the articles are belongs to both high quality or terrible sentiments, and most effective some articles belong to neutral sentiments. Most of the news articles withinside the crime and politics classes are belong to negative sentiments; however, the bulk of the news articles in entertainment, business, and sport belong to positive news. The politics, tech, and international information share an same a part of positive and negative sentiments. Graphical outcomes have proven in Table 1.

News Category	Positive Words	Negative Words	Score	Word Count
Business	11673	8803	2870	296873
Education	7467	3634	3833	187165
Sports	16133	11309	4824	277746
Entertainment	12423	7650	4773	235605
Technology	10589	6155	4434	202295

Table 1: Results of sentiment analysis of news articles.

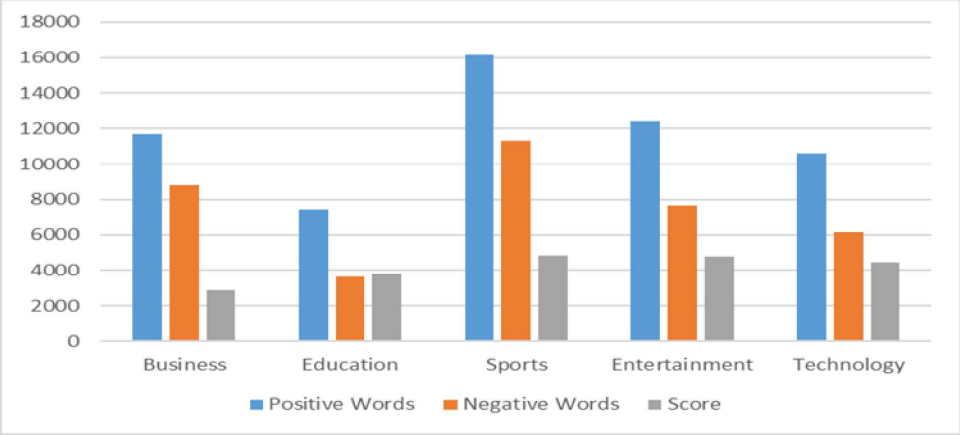


Figure 2: Comparison of categorized News articles with different moods.

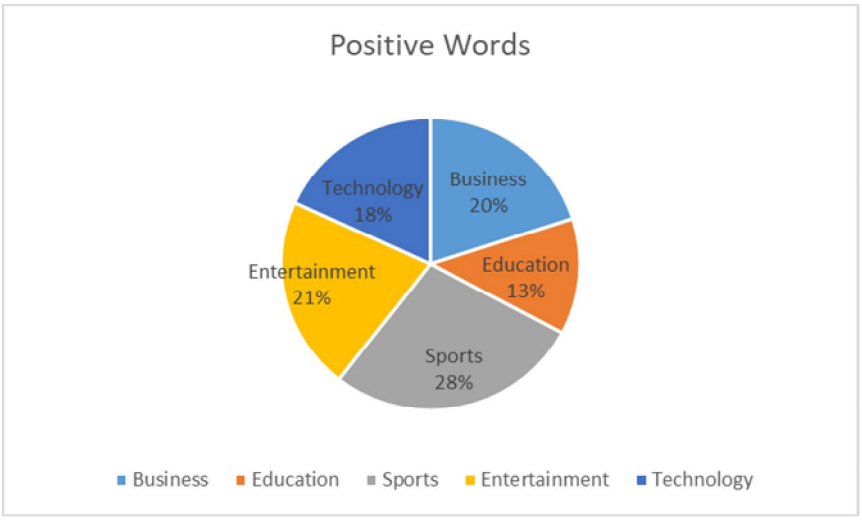


Figure 3: Percentage of every category under positive sentiments news articles.

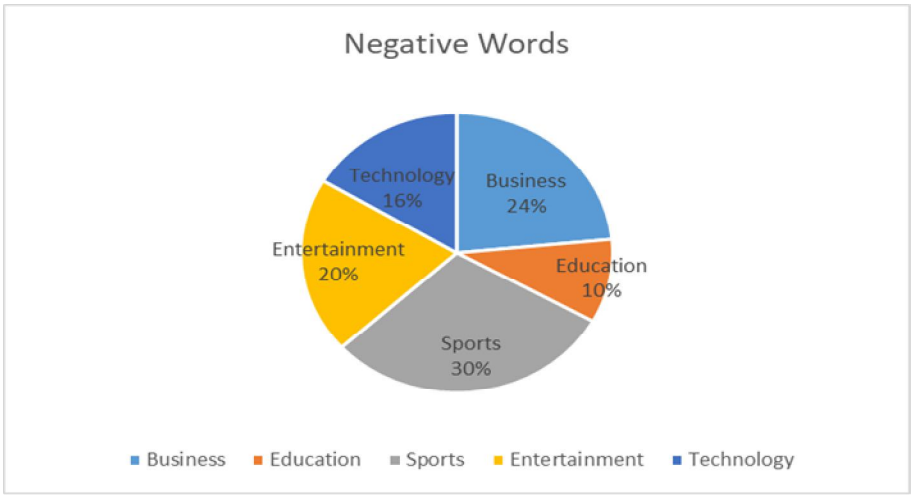


Figure 4: Percentage of every category under negative sentiments news articles.

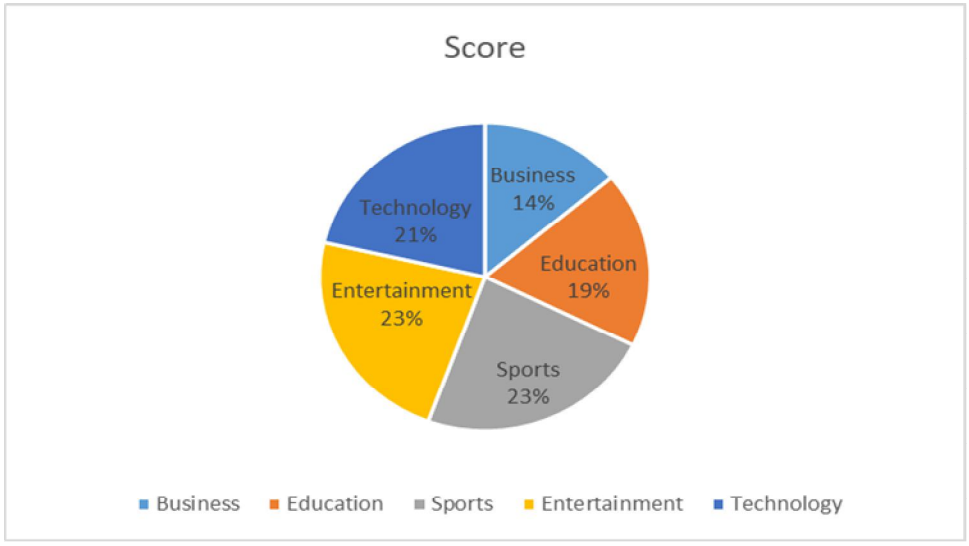


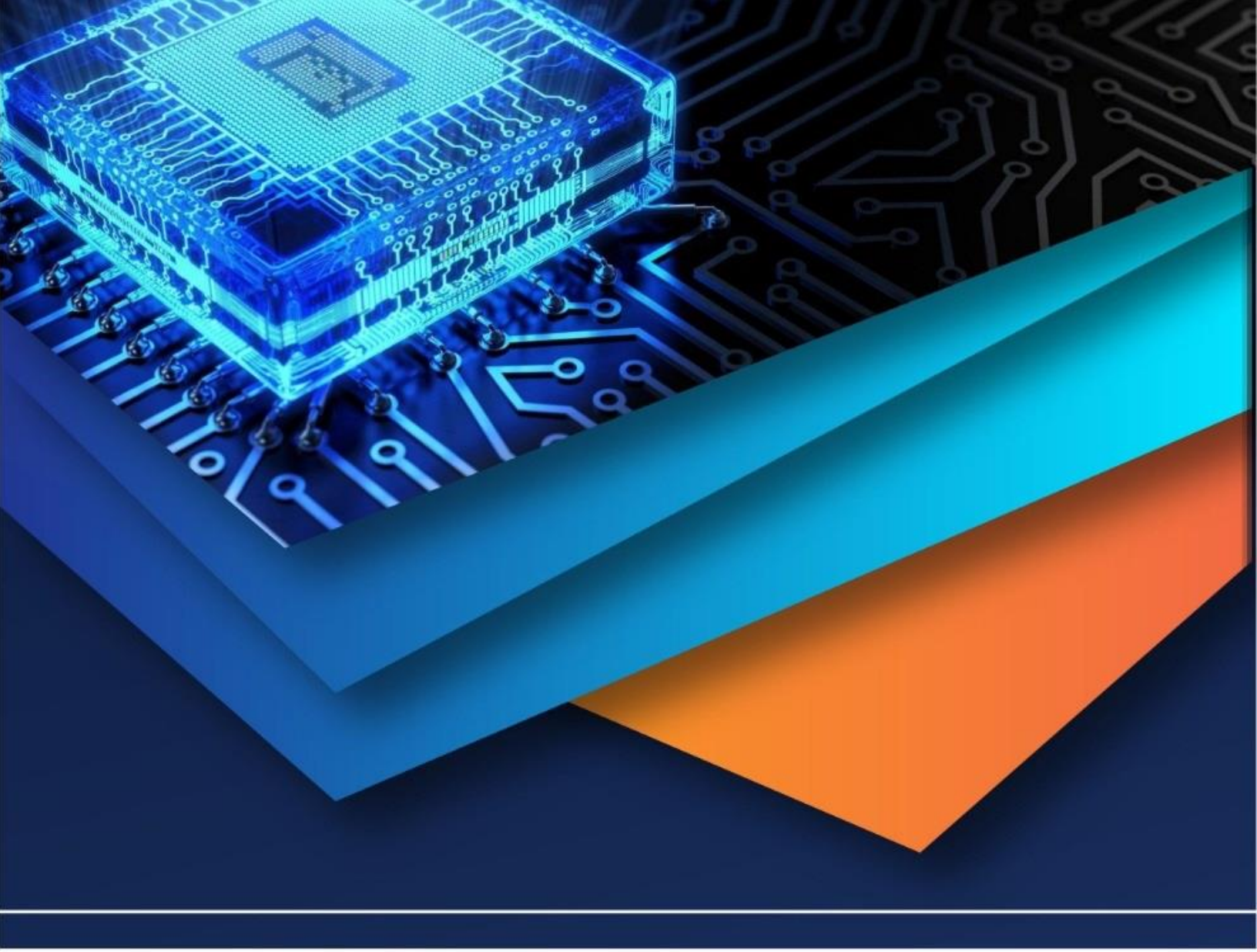
Figure 5: Percentage of every category under score sentiments news articles.

V. CONCLUSION

There are many applications of the information systems in which Sentiment Analysis have used, like summarizing review, opinion mining, classifying reviews, and various real-time applications. This paper applies sentiment analysis on news articles crawled from indian express newspaper from kaggle websites. One of the limitations of this approach is that it has noticed that sentiment analysis focused on English words and work on other languages like Hindi, Gujrati, and other languages need to perform. Many challenges may arise in dealing with other languages like implicit product features handling, sentence/ document complexity, negation expressions handling, and a summary of opinions of product features/attributes. Future work of this research could be implementing this sentiment analysis technique for other languages as well by overcoming the above mentioned challenges.

REFERENCES

[1] [online] 5G Radio Access, Research and Vision: http://www.ericsson.com/res/docs/whitepapers/wp- 5g.pdf [Last Accessed: 08-February-2014].
[2] Kanter, T., Forsström, S., Kardeby, V., Walters, J., Jennehag, U., & Österberg, P. (2012). MediaSense – an Internet of Things Platform for Scalable and Decentralized Context Sharing and Control. The Seventh International Conference on Digital Telecommunications, 27-32.
[3] [online] MediaSense | The Internet of Things Platform,http://www.mediasense.se/ [Last Accessed: 08-February-2014]
[4] Cugola, G., & Jacobsen, H. (2002). Using Publish/Subscribe Middleware for Mobile Systems. ACM SIGMOBILE Mobile Computing and Communications Review, 6, 25–33.
[5] Zarko, I. P., Antonic, A., & Pripužic, K. (2013). Publish/subscribe middleware for energy-efficient mobile crowdsensing. ACM conference on Pervasive and ubiquitous computing adjunct publication (UbiComp '13 Adjunct). Zurich. , 1099-1110.
[6] Dos Rieis, J. C. S., de Souza, F. B., de Melo, P. O. S. V., Prates, R. O., Kwak, H., & An, J. (2015). Breaking the news: First impressions matter on online news. In Ninth International AAAI conference on web and social media, 357-366.
[7] Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. Icwsm, 7(21), 219-222.
[8] Islam, M. U., Ashraf, F. B., Abir, A. I., & Mottalib, M. A. (2017). Polarity detection of online news articles based on sentence structure and dynamic dictionary. In 2017 20th International Conference of Computer and Information Technology (ICCIT), 1-5.
[9] Pereira, J., Fabret, F., Llirbat, F., &Shasha, D. (2000). Efficient matching for web-based publish/subscribe systems. In International Conference on Cooperative Information Systems, Eilat, Israel, September 6-8, 2000, 162-173.
[10] Fabret, F., Jacobsen, H. A., Llirbat, F., Pereira, J., Ross, K. A., & Shasha, D. (2001). Filtering algorithms and implementation for very fast publish/subscribe systems. In Proceedings of the 2001 ACM SIGMOD international conference on Management of data, 115- 126.
[11] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up Sentiment classi_cation using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing(EMNLP). (2002) 79-86.
[12] Meyer, B., Bikdash, M., & Dai, X. (2017). Finegrained financial news sentiment analysis. In Southeast Conference, 1-8.
[13] Shirsat, V. S., Jagdale, R. S., & Deshmukh, S. N. (2017). Document level sentiment analysis from news articles. In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), 1-4.
[14] Agarwal, A., Sharma, V., Sikka, G., & Dhir, R. (2016). Opinion mining of news headlines using SentiWordNet. In 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 1-5.
[15] Lei, J., Rao, Y., Li, Q., Quan, X., & Wenyin, L. (2014). Towards building a social emotion detection system for online news. Future Generation Computer Systems, 37, 438-448.
[16] Fong, S., Zhuang, Y., Li, J., & Khoury, R. (2013). Sentiment analysis of online news using mallet. In 2013 International Symposium on Computational and Business Intelligence, 301-304.
[17] Zhou, X., Tao, X., Yong, J., & Yang, Z. (2013). Sentiment analysis on tweets for social events. In Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 557-562.
[18] Li, G., & Liu, F. (2010). A clustering-based approach on sentiment analysis. 2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering, Hangzhou, 331-337.
[19] Popescu, A. M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In Natural language processing and text mining, 9-28. Springer, London.
[20] Somprasertsri, G., & Lalitrojwong, P. (2010). Mining Feature-Opinion in Online Customer Reviews forOpinion Summarization. J. UCS, 16(6), 938-955.
[21] Theussl, S., Feinerer, I., & Hornik, K. (2009). Distributed text mining with tm. In The R User Conference.
[22] Hofmarcher, P., Theußl, S., & Hornik, K. (2011). Do Media Sentiments Reflect Economic Indices?. Chinese Business Review, 10(7), 487-492.
[23] Goonatilake, R., & Herath, S. (2007). The volatility of the stock market and news, International Research Journal of Finance and Economics, 11, 53-65.
[24] Breen, J. (2011). R by example: Mining Twitter for consumer attitudes towards airlines. Boston Predictive Analytics Meetup Presentation.
[25] Yu, W. B., Lea, B. R., & Guruswamy, B. (2007). A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting. IJEBM, 5(3), 211-224.
[26] Agić, Ž., Ljubešić, N., & Tadić, M. (2010). Towards sentiment analysis of financial texts in croatian. In Proceedings of the Seventh International Conference on Language Resources and Evaluation.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY