Journal of Advances and Scholarly Researches in Allied Education
Vol. 20, Issue No. 2, April-2023, ISSN 2230-7540

A Study the review of Sentiment Analysis of Social Media Data Using Deep Learning

Priyesh Upadhyay[1]*, Dr. Ravindra Tiwari[2]

[1] Research Scholar, LNCT University, Bhopal

[2] Associate Professor, LNCT University, Bhopal

Abstract - In the current years, social media became one of the most important sources of data for different data analytic purposes. One of the most important issues is how to map different trends in social media and define the relation between different groups based on their sentiment or interests. SA, recommendation systems, event tracking, summarization, etc. all rely heavily on it as a component of NLP. The proliferation of social media platforms, the ease with which users can maintain a constant online presence, & rise of Web 2.0 have all contributed to a meteoric rise in the volume of digital social media data. Microblogging sites like Twitter, blogs, review collection websites, online forums have been contributing to the large scale of big data.

Keywords - Sentiment Analysis, Big Data, Deep Learning, Social Media, Sentiment Subject

----------X----------

INTRODUCTION

Sentiment analysis (SA) is the practice of gleaning meaning from the words people use to express their thoughts, feelings, and attitudes on social media. There are several other ways in which this text data could be found, including reviews, blogs, news, & comments. Many businesses all around the world have adopted the ability to derive insights from this kind of data. It has numerous & strong applications. Opinion mining, sentiment mining, & subjectivity analysis are a few names for sentiment analysis. This enormous amount of information is analysed using a variety of Natural Language Processing (NLP) activities. Notably, Sentiment Analysis (SA) is a work that is becoming more popular with the aim of categorising thoughts & attitudes expressed in text that is available on various social media platforms.

SOCIAL MEDIA

According to Castillo et al. (2011), social networking sites like Facebook, Twitter, Instagram, Google +, LinkedIn, & blogs all have the same corporate objectives: to draw in more users to foster social relationships and human engagement in virtual environments in all kinds of civilizations. They identified the characteristic that people have a propensity to communicate information about their opinions, states of mind, lifestyles, emotions, & sentimental feelings through a variety of channels. As a result, social media gives people a platform to share & see interactive content [Culotta, 2010]. They include appealing features that let users edit their own postings and make it easy to store links for later reading or sharing. With these features, shared information is accessible to everyone with access to that particular social media platform at any time and from anywhere [Fink, 2013].

Many academics use text mining tools & machine learning techniques in social networking platforms. An intriguing and enduring concept is sentiment analysis and opinion mining. In the last several years, a lot of research has been conducted in the fields of NLP, web mining, data mining, text mining, and others. All were carried out with the same objective in mind: to compile & evaluate various words of views & opinions about various subjects.

Information can be correctly categorized & human ideas may be examined by using sentiment analysis. This is vitally crucial to do in order to improve an NLP tool & effectively complete mining tasks. The work of NLP has been the subject of various improvements, although it is still constrained by significant difficulties [Grossman (2004)]. Investigators from a variety of fields are interested in exploring the hidden knowledge by using sophisticated data analysis techniques due to the development of social media data. Facts & views are the two basic categories of textual content on the internet. The assumption of truth in facts contrasts with the subjective nature of opinions on a particular person, thing, or issue.

GENERAL FLOW FOR SA

It takes specialist algorithms to recognise, categorise, extract, & summarise opinions from text,

which is a difficult & time-consuming procedure. The SA procedure is broken down into the subsequent steps. In Figure, they are depicted. This procedure is separate from the techniques utilized to carry out SA. These steps might be considered as characterising the process itself rather than how it is carried out. The actions are:

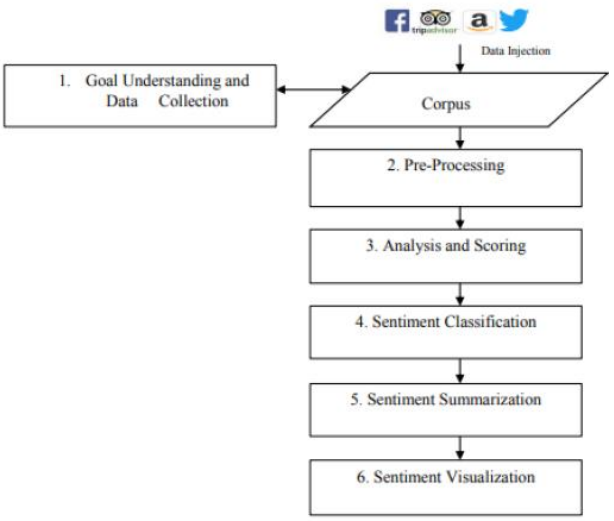

Figure 1: General Flow for Sentiment Analysis

SENTIMENT SUBJECT

The focus of the investigation into the sentiment is crucial. When an opinion construction conforming to the main sentiment evolves, the Subject holder is the most important factor. Those who transfer an opinion to another entity or the writer themselves may be the subject of sentiment analysis. The sentiment of a person can be discovered from two angles: from the viewpoint of the opinion holder who expresses an opinion on something, and from the viewpoint of the study of the response from opinion holders to such expression. The opinion reader may have a different or similar perspective on the opinion bearer. For instance, when home prices decline, it is not seen favourably from an economic standpoint because it may be terrible news for sellers while being excellent news for purchasers (Liu B., 2012). A select instances, such as financial news about publicly traded companies, where reader sentiment is reflected in stock market results, make it possible to acquire information about how viewpoints might reflect on readers and their sentiment. Similar to how the emotion in financial news would be used, this feasibility is used in this study.

LEVELS FOR SENTIMENT ANALYSIS

Three levels of SA are possible: document level, sentence level, & feature level.

Document-level

The document-level analysis offers sentiment analysis of a whole document & makes the assumption that the whole thing is about a single thing. For instance, on a product review website, the evaluation is about a specific entity, & entire article solely discusses the evaluation of that entity, which offers a general opinion about the product that is either favorable or bad.

Sentence-Level

The task is broken down to the individual sentences in the sentence-level analysis, which then identifies the attitude conveyed about each sentence as a favorable, negative, or neutral opinion. This level of analysis carefully analyses the sentence's direction by taking into account each word. It is simple to discern between subjective (contains an opinion) & objective statements using this level of analysis (factual sentence or sentence which nullifies the opinion). For instance, "I just bought an LED TV, & display is amazing." The short text primarily uses sentence-level analysis.

Feature-level

The aspect-level or entity-level analysis, also called feature-level analysis, is based on keywords (features can be thought of as keywords). The feature-level analysis can forecast a person's preferences with regard to a specific objective. If someone says, "Even though the service isn't that fantastic, I still love this Vistara Airline," they may not be pleased with the services, but they still appreciate the airline. In contrast to document- or sentence-level analysis, this form of analysis is referred to as "fine-grained analysis." This type of analysis can be utilized to perform a quality analysis with quantification. Additionally, each level analysis could be applied in one of three ways: supervised, unsupervised, or hybrid models.

TYPES OF MODELS FOR SA

Supervised Models

The supervised models involve the use of the training sets and testing set for capturing the sentiments as shown in figure1. Since SA is a text classification problem, therefore we can apply any existing supervised machine learning method, e.g., Multinomial Naïve Bayes, naïve Bayes classification, Logistic Regression, and support vector machines (SVM). This has evolved as a dominant approach in sentiment classification also. The only drawback of machine learning is that they required prior knowledge in the form of a labeled dataset for training, and a supervised algorithm trained on a labeled source domain does not generalize well on an unlabeled target domain which deteriorates the performance of cross-domain models (Sharma, 2018). In supervised models, this generally happens because the model fails to understand the basic patterns in the data. Even a minor difference in the available test data can make a large difference in predicting the sentiment pattern for a given target data.

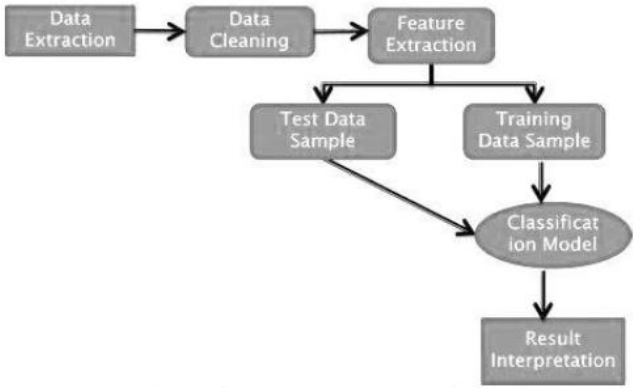

Figure 2: Supervised Learning Model

Unsupervised Models

The unsupervised models involve the use of a standardized dictionary instead of training sets, as shown in figure 2. This provides freedom from using the training sets which require domain-specific knowledge; the unsupervised models (Lexicon model) make use of lexicon dictionaries to calculate the polarity score for each feature. The numbers of dictionaries are already available for handling the score of word, emoticon, and slang, etc. and still researches are ongoing for improving the lexicon-based classification based on these factors using different paradigms like by normalizing the scoring calculation function, negation handling, intensifier, and diminishers handling, abbreviations handling, spelling errors handling properly. Nevertheless, with the supervised models all the above factors can be ignored because once the machine is trained with the features, it doesn't bother about all the above facts.

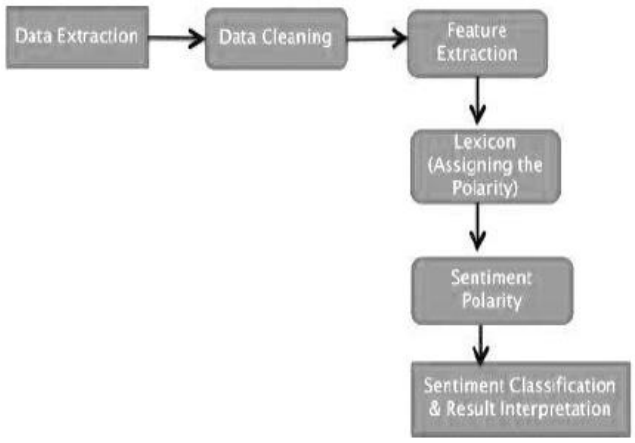

Figure 3: Unsupervised Learning Model

One more important factor that gauges the unsupervised model performance is the availability of the language-specific dictionaries for calculating the overall score of the given sentence. The lexicons for different native languages(other than English) are designed based on translations and translations are never be consistent and reliable(Basiri 2017). These dependencies make it hard to maintain domain independence (Hamdan, Bellot 2015) in unsupervised models.

Semi-supervised Models

It cites the concerns related accuracy of the model. The latest researches with semi-supervised and hybrid approaches have come up to overcome the above limitations (Asghar, 2018 c). Semi-supervised models are a combination of supervised and unsupervised learning models, where capabilities of both these models are taken together for the development of a new accurate model. The semi-supervised model can be used in different ways; the unsupervised clustering algorithms can be used with the supervised classification algorithm to reduce the training time. It can be used to create the labeled dataset and can be used in the design of a completely new lexicon required for sentiment analysis (Baccianella S 2010). Hence it is already being used by different researchers in different for giving accuracy in classification models and to mitigate the existing gaps of unsupervised and supervised models.

Hybrid Models

This is another approach used for improving accuracy in domain adaptable sentiment classifiers. Hybrid models are nothing but the combination of different models used for improving the overall accuracy of the model. Hybrid can be a combination of various supervised models only, it can be a combination of only unsupervised models also or can be the combination of unsupervised, supervised and other supportive model required for getting the accuracy in a model (Asghar 2015).

The key concern for the implementation of any of the above models is to minimise the total training time, which is usually accomplished by choosing the appropriate characteristics from which the model can learn the classification pattern.

SENTIMENT ANALYSIS

Understanding the tone of a document by sentiment analysis, a text mining technique, is becoming increasingly common. In some contexts, the term "opinion mining" may also be used. SA employs textual elements at various textual levels, such as the entire document, paragraph, phrase, or just a text window, to correctly determine & attribute the particular mood & feelings to the relevant entities, also termed objects [Liu B 2010]. Understanding text information, assigning the individual polarized information, & interpreting human thought is complex, making this analysis a difficult task to automate. The crux of the issue is that, unlike humans, computers lack interpretive skills and must rely on model-based learning in order to comprehend complex relationships. Due to this constraint, SA requires processing of unstructured data into a structured & machine-readable form.

Lexical resources that provide information on which particular words could be allocated to what emotions are used for the actual evaluation & analysis of texts in many applications. It was published in 2012 in the journal Cambria E. The overall tone of a document, set of papers, or body of text can be altered by associating its constituent parts with specific emotions. Many different approaches have been developed for the purpose of SA. Document-level SA is commonly used as a first step in textual analysis [Turney PD, 2002]. Token frequency analysis is one example of an alternative method that generates attribute-value representations for the purpose of easy comparison & categorization [Joachims T 1998]. More granular SA typically involves document segmentation for easier analysis at the paragraph level. The lexical assessment, often known as a bag of words or dictionary-based approach [Liu B 2012], is a popular method in SA. In this way, the document's pre-processing results in a more precise evaluation of the text's constituent parts. Multiple SA tools make advantage of lexical assessment [Cambria E 2013].

Furthermore, some analysis algorithms use even finer layers of the text to finally allocate the polarized text content direct to the entities. In this scenario, the entity is recognized at the sentence-level on the basis of sentence structure (Part-of-Speech) (Part-of-Speech). Adjectives, adverbs, verbs, and nouns make up the standard sentence components that convey emotion. The article in question can be found at [Benamara 2007]. Common word combinations (n-grams) can be used in some analytic techniques as well. The tokenization method did not remove the context here. In [Pang B 2002], for instance, n-grams, along with bi- and unigrams, are analyzed to see which yielded better results during the classification of movie reviews. Trigrams, on the other hand, are [Dave K 2003]'s preferred method of categorizing customer feedback on products. It is common practice to employ N-grams when analyzing literature for phrases that serve as fixed expressions & appear numerous times throughout. Artificial intelligence models can also be used to spot these kind of recurring themes in written work. All methods and studies of emotion distinguish between subjective and objective forms of feeling [Wiebe J 2004]. It is a universal truth that all methods of SA require extensive & intricate document preprocessing. Textual information needs to be formatted in a certain way before it can be processed, and this is something that is lacking in unstructured data. Therefore, the messages are processed utilizing procedures from NLP. For the analysis, this essentially entails an identification of text levels, usually the endings of sentences, breaking text into n-grams or individual words [Dave K 2003]. A number of techniques, including those based on word associations [Uhr P 2014], make use of windows in which text can be entered. After the sentence structure is used to identify entities, the text is broken down into tokens. The textual data is structured & machine-readable form after other NLP techniques are performed, such as the removal of stop words (usually a list of non meaningful terms).

Putting texts into positive, negative, or neutral categories is a common method. The classification procedures in this scenario can perform lexical evaluations, summaries, and case-by-case distinctions on the tokenized bits of the original text. To find the centers of categories, other classification algorithms typically look for textual similarities & use distance measures to differentiate between them [Weiss SM, 2010]. Further, Bayes classifiers are widely utilised probabilistic classification [Das SR, 2010]. SA, sadly, has not yet overcome all of its shortcomings. The analysis has certain unresolved concerns with the usual approaches, in addition to its restricted learning capacity. There are still open questions about the resolution of ambiguities in word sense disambiguation & negation within sentiment analysis [Adam, 2010]. It is extremely challenging to make sense of a text that has been broken down to its most basic elements (tokens). Particularly challenging for the SA are items that are ironic or satirical, even though the human mind recognizes them as such; this is a problem that cannot be overcome without adequate artificial intelligence. Indicators of sarcasm in punctuation are common. Punctuation is typically stripped away during the initial processing of text, making it nearly impossible to recreate or decipher. Not only are that, but shorthand words and slang frequently employed when conversing on social media. The SA tools are lacking in their understanding of these terms and phrases and thus require regular updates. And it's through these emoji-like symbols that the emotion of social media is communicated. These strings are typically stripped away during preprocessing since they are made up almost entirely of special characters (ASCII). The utilization of hashtags is encouraged on certain social media sites. Since the pound sign (#) is not considered part of the text, it will be removed during the pre-processing phase if no changes are made. Since hashtags are intended to be used in a succinct manner to refer to a linked topic, person, or trend, they should be treated as a distinct entity in any system developed to analyze information from social networks. According to [Asur 2010], this sort of short writing (like a Twitter post with a maximum of 140 characters) makes for a fascinating foundation since it is straightforward to evaluate & forces authors to get to the point quickly. Because of the rise of social media, another issue has emerged: so-called opinion spam, which is triggered by a wide range of emotional remarks that either distort or hide the true opinion [Jindal N 2008]. Intentional fraudulent claims or advertising must be detected as such and deleted from the analysis if an accurate result is to be achieved.

DEEP LEARNING

In the field of artificial intelligence, deep learning is both a subfield and the dominant current trend (Zhang et al. 2018). The goal of deep learning is to provide computers the ability to observe, learn from, & respond to complex circumstances by employing multilayer neural networks to perform both linear &

nonlinear transformations on data in order to extract its features (Deng 2014; Day 2017). AlphaGo is the paradigmatic instance of this type. Recurrent neural networks (RNNs) for machine translation services based on natural language processing & statistical methodologies (Krizhevsky et al., 2012) & CNNs for computer vision and image recognition are two popular deep learning methods (Cho et al. 2014). Even though deep learning has been shown to produce good results in many applications, there are still a number of issues that need to be improved, including exploding & vanishing gradients, difficulties in model interpretation, connected parameter settings, increasingly complex model training as the number of network layers increases, and how to preserve a certain accuracy rate to enhance training speed. In order to progress in the field of deep learning, several challenges must be investigated and resolved. Hochreiter proposed the LSTM, or long short term memory, which is an expansion of the RNN architecture (1997). Like RNN, LSTM has a basic structure that consists of three parts: input gate, output gate, & forget gate (Zhang et al. 2018). These gates have their own weights and are activated or deactivated based on the incoming signal. These gates also act as filters, evaluating the intensity and content of incoming signals before deciding whether or not to pass them. When RNN fails to train as a result of vanishing or bursting gradients, the forget gate of LSTM can choose both remembered & forgotten input (Day 2017). Because of its long-term memory capabilities, LSTM has been shown to be very effective in learning a wide range of sequence modeling problems with uncertain lengths (Zhang et al. 2016). But LSTM only takes into account unidirectional context messages. Therefore, Graves used the Bidirectional Long Short-Term Memory (Bi-LSTM) architecture to enhance the performance of the standard LSTM model by extracting finer-grained features in 2005. In order to get the forward & reverse context messages, this design employs two LSTMs running in opposing directions: a forward layer and a backward layer. Each output is the cumulative result of adding the outputs of two LSTMs, one forward and one backward. Speech classification (Lehner et al. 2015), scene recognition (Chen et al. 2017), stock market price fluctuation analysis, time series forecasting, healthcare monitoring & human behavior and motion recognition are just some of the areas that LSTM-based research method has explored (Fok et al. 2018).

DATA MINING CONCEPTS

We live in a data age, where massive amount of data is being generated with every passing second. However, all this data is of no use, if it is not converted to a form that is beneficial and meaningful for the recipient to do something beneficial. Data mining is the process of finding the interesting patterns or information from the data. All business organizations these days tend to stay updated not only about what has happened in the past and why it has happened, but also about the things that are going on at present and those which are going to happen in future. All this involves the analysis of data. Data Analysis, which involves a series of stages, is used to extract meaningful information from the data [Han 2011]. Data analysis process is sometimes taken as a substitute for KDD- Knowledge Discovery from Data. It involves various steps and one among these steps is Data Mining [wang 2010]. These steps are briefly put down as below:

1. Data Cleaning: To remove and other impurities from the data.
2. Data Integration: In this step various data sources are combined.
3. Data Selection: This step involves selection or retrieval of the relevant data from the database.
4. Data Transformation: In this step, the conversion of data into forms suitable for mining is carried out.
5. Data Mining: This step involves use of intelligent methods to extract the interesting patterns out of data.
6. Pattern Evaluation: The extracted data patterns are evaluated in this step.
7. Knowledge Representation: Various visualizing tools are used in this step to present the data mining results.

The graphical presentation of the overall KDD process showing the various steps in the extraction of interesting patterns out of data is put below in the following figure:

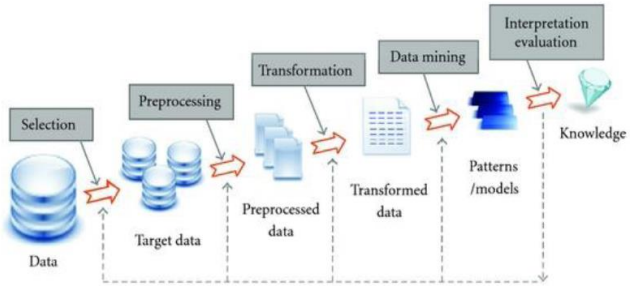

Figure 4: Knowledge Discovery Process

Data preprocesses involves the steps from 1 to 4 mentioned above. Since the data comes from multiple sources which may contains inconsistencies that need to be removed. As such, these steps prepare the data for mining. The next step involves mining the data which may be of the smaller size to make the entire process more efficient. Various data mining techniques are put into practice for mining of the data to extract interesting patterns. The patterns

extracted are then evaluated through certain rules. The final stage involves the presentation of the results through various visualizing tools like graphs, trees, charts, tables and others.

CONCLUSION

This work introduced the area of Sentiment Analysis (SA) on Social Media and projected our research analysis. . This research work is an attempt to address the issue of analyzing the sentiments of streaming text generated by social media platforms by proposing an adaptive big data model. For laying the foundation, a literature survey of both big data systems and sentiment analysis approaches has been conducted. Considering the state-of-the-art results obtained by deep learning networks over extensive data in an unsupervised setting, to work on streaming data & extract the subjects, we have decided to couple deep learning models with a topic modeling strategy.

REFERENCES

1. A. Almars, X. Li, and X. Zhao, “Modelling user attitudes using hierarchical sentiment-topic model,” Data Knowl. Eng., vol. 119, pp. 139–149, 2019.

2. A. Kolovou et al., “Tweester at SemEval-2017 Task 4: Fusion of Semantic-Affective and pairwise classification models for sentiment analysis in Twitter,” in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 675–682.

3. A. R. Pathak, “Social media data #bitcoin #ethereum #facebook”, Mendeley Data, v1, 2019,

4. A. R. Pathak, M. Pandey, and S. Rautaray, “Adaptive Model for Dynamic and Temporal Topic Modeling from Big Data using Deep Learning Architecture,” Int. J. Intelligent Systems and Applications. vol. 11, no. 6, pp. 13-27. 2019.

5. A. Rekik and S. Jamoussi, “Deep Learning for Hot Topic Extraction from Social Streams,” in International Conference on Hybrid Intelligent Systems, 2016, pp. 186–197

6. A. Saha and V. Sindhwani, “Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization,” in Proceedings of the fifth ACM international conference on Web search and data mining, 2012, pp. 693–702.

7. Abbasi-Moud, Z., Vahdat-Nejad, H., & Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. Expert Systems with Applications, 167, 114324.

8. Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research (2010) 52(1), PP. 11–19

9. Agbehadji, I. E., & Ijabadeniyi, A. (2020). Approach to sentiment analysis and business communication on social media. In Bio-inspired Algorithms for Data Streaming and Visualization, Big Data Management, and Fog Computing (pp. 169-193). Springer, Singapore.

10. Agüero-Torales, M. M., Salas, J. I. A., & López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. Applied Soft Computing, 107, 107373.

11. Ahmed, H. M., Javed Awan, M., Khan, N. S., Yasin, A., & Faisal Shehzad, H. M. (2021). Sentiment analysis of online food reviews using big data analytics. 20(2), 827-836.

12. Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. Journal of Travel Research, 58(2), 175-191.

13. Alarifi, A., Tolba, A., Al-Makhadmeh, Z., & Said, W. (2020). A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks. The Journal of Supercomputing, 76(6), 4414-4429.

14. Aletras and M. Stevenson, “Evaluating topic coherence using distributional semantics,” in Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)--Long Papers, 2013, pp. 13–22

15. Alsayat, A. (2022). Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model. Arabian Journal for Science and Engineering, 47(2), 2499-2511.

16. Amrutphale, Y., Vijayvargiya, N., & Malviya, V. (2020). A Novel Adaptive Approach for Sentiment Analysis on Social Media Data. In Social Networking and Computational Intelligence (pp. 717-726). Springer, Singapore.

17. Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2021). A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. Expert Systems with Applications, 164, 114006.

18. Asghar, M.Z., Kundi, F.M., Ahmad, S., Khan, A. and Khan, F., 2018. T-SAF: Twitter

sentiment analysis framework using a hybrid classification scheme. Expert Systems, 35(1), p.e12233.

Corresponding Author

Priyesh Upadhyay*

Research Scholar, LNCT University, Bhopal