

# Online news popularity prediction before publication: effect of readability, emotion, psycholinguistics features

Suharshala Rajagopal, Anoop Kadan, Manjary Gangadharan Prappanadan,  
Lajish Vimala Lakshmanan

Department of Computer Science, School of Mathematics and Computational Sciences, University of Calicut, Kerala, India

---

## Article Info

### Article history:

Received Oct 16, 2021

Revised 31 Dec, 2021

Accepted 20 Jan, 2022

---

### Keywords:

Emotion features

News popularity prediction

Online news media

Psycholinguistics features

Readability features

---

## ABSTRACT

The development of world wide web with easy access to massive information sources anywhere and anytime paves way for more people to rely on online news media rather than print media. The scenario expedites rapid growth of online news industries and leads to substantial competitive pressure. In this work, we propose a set of hybrid features for online news popularity prediction before publication. Two categories of features extracted from news articles, the first being conventional features comprising metadata, temporal, contextual, and embedding vector features, and the second being enhanced features comprising readability, emotion, and psycholinguistics features are extracted from the articles. Apart from analyzing the effectiveness of conventional and enhanced features, we combine these features to come up with a set of hybrid features. We curate an Indian news dataset consisting of news articles from the most rated Indian news websites for the study and also contribute the dataset for future research. Evaluations are performed over the Indian news dataset (IND) and compared with the performance over the benchmark mashable dataset using various supervised machine learning models. Our results indicate that the proposed hybrid of enhanced features with conventional features are highly effective for online news popularity prediction before publication.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



---

## Corresponding Author:

Anoop Kadan

Department of Computer Science, School of Mathematics and Computational Sciences, University of Calicut

Kerala-673635, India

Email: [anoopk\\_dcs@uoc.ac.in](mailto:anoopk_dcs@uoc.ac.in)

---

## 1. INTRODUCTION

The new era of technology and internet ready in hand, we are always connected to the world around us. Many social media platforms emerged as a result of this upheaval. People get engrossed themselves in these online social media platforms at their homes, work, while traveling and simply everywhere. One such most promising social assembling is at online news media. Busy schedules, less leisure time, and availability of mobile devices in plenty have paved ways for people to access the world with information about anything and everything at their fingertips. Thus, most of the prominent news media launched their online portals from which people can effortlessly access the news, best part of which is the access to up-to-date news. News has thus come out as one of the most consumed products. As in every business, rapid evolution in means and modes of communication with the introduction of new technologies has made news media businesses adapt to satisfy their consumers. While readers tend to visit the news portals which provide timely updated news, competition among the news media channels raised to great extends, resulting in these media making updates to popularize their news articles subsequently to please their readers. The role of online news popularity

prediction before publication has thus aroused. This area of research thus can help Journalists, content writers, and news editors by making them aware of writing content in such a way that news will gain more popularity among the readers.

Predicting news popularity is a broad area of research. Many studies have eventuated previously in this area, as many online news portals are ascending to get the idea of how the news gain popularity. News popularity prediction has been studied in various target domains of social media including, Facebook, and Twitter, [1]–[4], and in different languages [5]–[9], where, prediction can be performed either after or before publication. Most of the earlier works focus on news popularity prediction after publication. Assessing number of views, shares, or comments, and other metadata like title, time of publishing, and author, can help provide a better understanding of acceptance of the online news content among its users. But, the information like number of shares, views, or comments is only available if the news is already published. So, news popularity prediction after publication is facile with such available parameters. A large number of features are available for news popularity prediction after publication; while only a few are available for before publication. Hence, predicting news popularity before publication is a challenging task and as a result, the prediction results of the before publication systems are less than the after publication systems. Table 1 shows some state-of-the-art works in news popularity prediction before publication.

Table 1. Related works in news popularity prediction before publication

Work	Task setting	Feature domain	Language	Target domain
[5]	Supervised	Content, Metadata	Chinese	Freebuf, Chinese website
[7]	Unsupervised	News Comments, Temporal	Dutch	Dutch online news websites
[9]	Supervised, Unsupervised	Content, Temporal	English	Mashable news
[10]	Supervised	Content, Temporal, Emotion	English	BBC, Newyork times, Dailymail, Reuters, News websites
[11]	Supervised	News Comments	Dutch	Dutch online news websites
[12]	Supervised	Temporal	English	English online news websites
[13]	Supervised	Content, Temporal, Emotion	English	Mashable news
[14]	Supervised	Content, Metadata, Temporal, Emotion	English	Yahoo News, Twitter

In this work, we intend to state that, the set of hybrid features combining enhanced features with the conventional features like metadata, temporal, contextual, and embedding vector features, could improve the performance of online news popularity prediction before publication. And for that, readability, emotion and, psycholinguistics features are taken as the enhanced set of features in this work, where, psycholinguistic features, up to our best knowledge, have not until been used for online news popularity prediction before publication. Among the previous works in literature to predict online news popularity before publication, a vast majority utilize news title, content, author, and category, for analyzing popularity of news articles. Very few studies consider features like readability and emotion. A notable work in this regard would be that of Kate *et al.* [15] and Berger and Milkman [16], where, the importance of readability and emotion in content virality is studied. Unlike their work, this work explores NRC emotion lexicon [17], which is not yet used to study the effect of emotions on news popularity prediction before publication by extracting emotions in the news content. To understand the effectiveness of the features in news popularity prediction and to check the trends of prediction accuracies, the same feature-model combinations are realized over two datasets, our newly curated Indian news dataset and the benchmark mashable dataset. Our contribution of the Indian news dataset is publicly available at <https://dcs.uoc.ac.in/cida/resources/ind.html> to aid future research in this area.

## 2. RESEARCH METHOD

In this section initially we outline the two datasets used in our study. Later, we elaborate the features extracted from the datasets including the conventional, enhanced and hybrid features. Finally we detail the experimental settings followed in our empirical evaluations.

### 2.1. Dataset

We utilize two datasets, the newly curated Indian news dataset (IND) and the existing benchmark dataset, Mashable [18]. The details of our newly curated IND is described elaborately below, in section 2.1.1. The Mashable dataset contains URLs to 39,644 articles, collected during two years, along with 61 attributes describing different features of the corresponding news articles.

### 2.1.1. Indian news dataset (IND)

For the task of online news popularity prediction we newly curated a dataset named the IND. IND contains news data or articles from ten most rated Indian news websites (i.e, India Times, Firstpost, NDTV, The Indian Express, Times Now, One India, Hindustan Times, India TV, News18 and Zee News), with the main motive that they have news articles with a large number of views or shares which is a good indicator of news popularity among the readers. Considering the news genre common to all the websites, news articles are selected from the categories technology, election, sports, entertainment, and lifestyle. In total, 1,000 news articles were gathered from the websites, i.e., 100 news from each website with 20 articles belonging to each of the five different categories. The dataset was then labeled based on the number of shares. The news having least shares is labeled as ‘Unpopular’, whereas those with large number of shares are labeled as ‘Popular’. The news articles in the dataset are also appended with some additional associated information like the date of publishing a news, news category and name of the news portal. When compared to the existing datasets like mashable, IND provides the title and content of the news rather than some associated statistics or URLs to the news and hence, IND dataset can be considered as a much ready to use dataset. Table 2 shows the statistics of IND and Mashable dataset. The datasets are pre-processed using multiple methods including stop word removal, lemmatization, stemming, tokenization and part-of-speech tagging before admitting for feature extraction.

Table 2. Dataset statistics

Parameters	IND	Mashable
Number of news articles in the dataset	1000	39644
Average number of words in headlines	6	7
Average number of words in contents	1095	1048
Average number of sentences in contents	26	28

## 2.2. Feature extraction

After pre-processing, features are extracted from both the datasets. These features extracted are then fed to the supervised machine learning models. In this work we consider three major categories of features, conventional, enhanced, and hybrid set of features.

### 2.2.1. Conventional features

The conventional category of features includes metadata, temporal, contextual, and embedding vector features. Many previous works related to news popularity prediction before publication uses the conventional set of features. These features are extracted from the very immediate information available from the news.

- i) Metadata feature (MF): The metadata features author and category are good indicators of the popularity of online news articles [5]. If an author’s articles incessantly gain popularity, his/her articles are inclined to be popular in the future too. Similarly, in some season, specific categorical news will attain more popularity when compared to others [13], [14]. An example is people mostly accessing the Sports page during the seasons of football or cricket. In this work, for author score, each author is associated with the count of his/her popular and unpopular articles, and for category score, for each of the five categories considered (technology, election, entertainment, sports and lifestyle), the numbers 1 to 5 are assigned. The Mashable dataset does not include any details on author and category. hence, the MF are not extracted from this dataset.
- ii) Temporal feature (TF): Time is a significant factor in regulating news popularity. The interest of reading news fluctuates during various times. In this work, we explore how the weekends’ impact news popularity. For this, the news publishing dates are extracted and converted to the corresponding days. Later, the days are marked as Weekdays if they are the days Monday to Thursday, or as weekends if they are Friday to Sunday. The effect of this feature is studied based on a hypothesis that people will be having less time on weekdays and barely access online news, but on weekends, as they have enough leisure time, they might be able to spend more time accessing online news [5].
- iii) Contextual or content feature (CF): An attractive headline and polished content can inspire the enthusiasm of people prompting them to open and read a news article [5], [19], [20]. The news authors, therefore, ensure the content is sufficiently well so that people will read the content. Grammatical construction, length of content, and presence of some phrases, can improve the readability of an article. These comprise the linguistics and grammatical construct of the title and content of the news. The CF is a combination of N-grams, content length, POS tags, named entities, grammatical construct score and title punctuations [21], [22].

- iv) Embedding vector feature (VF): Word2Vec [23] and Doc2Vec [24] are the two embeddings performed on the news articles. Word2Vec is a two-layered neural network that processes text to group similar phrases together in a space producing vectors that are the numerical representations of word characteristics, features such as the meaning of single words. On the other hand, the Doc2Vec represents documents as a vector and is a generalization of the Word2Vec process. Doc2Vec aims to create numerical representations of text irrespective of its length. We combine these two vectors to form the embedding VF.

### 2.2.2. Enhanced features

The conventional set of features mentioned above, primarily only address the visible data. Other than these conventional features many other features can influence the popularity of online news articles like, readability features, sentiment or emotions in the content and title, and psychological pointers over the content linguistics. We explore the effect of such enhanced features in improving the popularity prediction of online news articles before publication.

- i) Readability feature (RF): Readability indicates the convenience with which a reader can comprehend a written text, depending on its content. The feature focuses on selected words and their transformation into sentences and paragraphs for the readers to understand. This feature combines the attributes associated with the content such as the number of characters, complex words, long words, syllables, word types, and paragraphs, and also several calculated readability metrics, including dale chall readability score, flesch reading ease, gunning fog, and smog formula [15].
- ii) Emotion feature (EF): Emotions are very fine indicators of admiration of an article. News articles encompassing highly positive or negative emotions acquire more viewers. Three attributes are gathered from the content, which are combined to form the Emotion feature. The first is the sentiment score given by the count of total positive and negative words in the content computed using a senti-lexicon, VADER [25], [26]. By using the NRC emolex [17], emotions in the content are classified into four categories joy, anger, fear, and sadness. The average intensity and the count of these emotions contained in the news are next two attributes.
- iii) Psycholinguistic feature (PF): The features affect, social, cognitive processes, perceptual processes, biological processes, drives, time orientation, relativity, personal concerns, and informal language, forms the PF. The presence of these features accelerates the assent of an article among the readers. The linguistic inquiry and word count (LIWC) lexicon [27] helps to extract the proportion of words falling within the psycholinguistic groups. LIWC is based on extensive word category lexicons that reflect psycholinguistic processes (e.g., positive emotions, perceptual processes), summary categories (e.g., words per sentence), as well as part-of-speech categories (e.g., articles, verbs).

### 2.2.3. Hybrid features

In this work we analyze whether a better outcome can be achieved by combining the enhanced features with the conventional ones. We analyze two sets of hybrid features. The first hybrid feature set is the combination of readability feature with the conventional features, examined individually. Since both the emotion and psycholinguistic features handle the affect involved in the content, their combination is considered as a single entity, named as the psycho-emotion feature (PEF). The combination of PEF with each of the conventional features and the readability feature is the second hybrid feature set.

## 2.3. Experimental settings

For the empirical evaluations, each of the two datasets are split into 75 percentage for training and the rest for testing. Four different supervised machine learning techniques, i.e., naive Bayes (NB), support vector machine (SVM), multi-layer perceptron (MLP), and decision tree (DT), are used to build the binary classifiers models. All models are implemented using the Scikit-learn machine learning library. The SVM implementation uses radial basis function kernel. The maximum depth of DT is set to 5, where the split is performed based on the criterion entropy. MLP is implemented using a 100 neuron hidden layer with Adam optimizer for 2,000 epochs. All other parameters of every models are set to the default values in our implementation. The percentage of documents correctly classified by the model in both the classes, ‘Popular news’ and ‘Unpopular news’ gives the model accuracy.

## 3. RESULTS AND DISCUSSION

Table 3 provides a record of the prediction accuracies in percentages for various supervised machine learning models. For IND and mashable dataset, temporal and embedding vector features of the conventional

category returns the highest two prediction results. MLP and SVM act as the best two classifier models for IND, whereas NB and DT works well for mashable, for the conventional category of features. For the enhanced feature category, psycholinguistic and readability features produce the best two prediction results. But unlike the conventional category, here the best two classifiers for both IND and Mashable comes as MLP and DT. The psycholinguistic features offer an exceptional accuracy of 79.2% for IND, an increase of 7.2 percentage points over the highest accuracy produced by the conventional features.

Table 3. Prediction results in percentages (the best two accuracies in boldface)

Features	IND				Mashable			
	NB	SVM	MLP	DT	NB	SVM	MLP	DT
Conventional features								
Metadata feature (MF)	50.0	60.4	63.0	56.4	-	-	-	-
Temporal feature (TF)	50.4	72.0	72.0	72.0	52.9	52.5	52.5	52.5
Content feature (CF)	50.0	54.0	55.6	54.0	51.1	47.1	49.0	49.4
Embedding vector feature (VF)	50.2	68.0	70.8	65.2	55.9	53.0	53.0	55.5
Enhanced features								
Readability feature (RF)	46.8	50.4	50.4	58.0	46.8	49.6	52.4	53.2
Emotion feature (EF)	46.0	48.4	49.2	51.2	46.0	46.0	50.3	50.4
Psycholinguistic feature (PF)	67.6	76.0	77.6	79.2	46.1	46.9	51.7	52.3
Hybrid features								
MF+RF	45.6	51.2	51.2	57.6	-	-	-	-
TF+RF	50.4	50.4	57.2	62.0	53.0	50.2	52.9	50.3
CF+RF	48.0	48.4	55.2	55.2	54.6	52.9	53.1	50.5
VF+RF	76.0	78.8	82.0	85.2	55.8	53.8	53.9	51.9
MF+PEF	51.2	69.2	69.6	72.4	-	-	-	-
TF+PEF	72.4	76.0	77.2	80.8	54.2	52.5	52.7	50.1
CF+PEF	64.4	68.5	68.8	76.0	49.4	48.9	49.2	49.0
VF+PEF	56.0	69.2	71.2	71.2	57.0	53.9	54.0	52.0
RF+PEF	60.8	60.8	65.6	76.0	53.4	53.1	53.1	50.2

In case of hybrid features, embedding vector with readability feature (VF+RF) produces an outstanding prediction accuracy of 85.2%, a noteworthy gain of 13.2 percentage points over the highest accuracy among the conventional features, and a gain of 6 percentage points over the highest accuracy among the enhanced features. VF+RF also produces good results for the mashable dataset. Another promising feature combination for IND is the hybrid of temporal with psycho-emotion feature (TF+PEF), which yields an accuracy of 80.8%. The best accuracy for mashable 57%, is also given by a hybrid feature, embedding vector with psycho-emotion (VF+PEF), even though the increase in prediction accuracy is not much appreciable in mashable. Mashable only consists of related attributes of a particular article, which might be the cause of substantially low accuracies compared to IND, while, IND is very structured and collected solely for news popularity prediction, unveiling better results than Mashable. The assumption that news popularity prediction before publication can be improved using enhanced features including readability, emotion and, psycholinguistics features elucidate from the experimental results where performances of the hybrid of enhanced with conventional features outperform the conventional features for IND particularly, and for mashable.

#### 4. CONCLUSION

This work proposes an improved online news popularity prediction system before publication using a hybrid of the enhanced features viz. Readability, emotion, and psycholinguistic features, with the conventional features, modeled using supervised machine learning. Our simple methodology demonstrates that the hybrid features obtained by combining the conventional features and the enhanced features like psycho-emotion features are exceptionally very much useful to achieve improved results. With limited inputs and a simple set of features based on the very primary details accessible within the content such as title, content, author, category, and time, this work arrives at good prediction results. We plan to expand the work by including a richer set of temporal features like detailed aspects of time instead of only considering the day of publishing, and also the content-based features related to trend analysis and readers' comments. Another direction of work would be to evaluate the effectiveness of the rank algorithms, extensively used in information retrieval, for the news popularity prediction task, with an assumption that news can be ranked based on the popularity it is going to achieve, and compare against the proposed supervised prediction system.

## ACKNOWLEDGEMENTS

The first author was supported by the University of Calicut M.Phil. Research fellowship. The second author was supported by Rajiv Gandhi National Fellowship (RGNF), University Grants Commission (UGC), India (RGNF-2014-15-SC-KER-79884) and the third author was supported by the Women Scientist Scheme-A (WOS-A) Fellowship, Department of Science and Technology, India (SR/WOS-A/PM-62/2018).

## REFERENCES




- [1] A. Pugachev, A. Voronov, and I. Makarov, "Prediction of News Popularity via Keywords Extraction and Trends Tracking," in *Recent Trends in Analysis of Images, Social Networks and Texts*, 2021, pp. 37–51, doi: 10.1007/978-3-030-71214-3\_4.
- [2] M. Nashaat and J. Miller, "Improving News Popularity Estimation via Weak Supervision and Meta-active Learning," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021, pp. 2679–2688, doi: 10.24251/HICSS.2021.327.
- [3] I. Heimbach, B. Schiller, T. Strufe, and O. Hinz, "Content Virality on Online Social Networks: Empirical Evidence from Twitter, Facebook, and Google+ on German News Websites," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, 2015, pp. 39–47, doi: 10.1145/2700171.2791032.
- [4] A. P. Sitorus, H. Murfi, S. Nurrohmah, and A. Akbar, "Sensing Trending Topics in Twitter for Greater Jakarta Area," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 1, pp. 330–336, Feb. 2017, doi: 10.11591/ijece.v7i1.pp330-336.
- [5] C. Liu, W. Wang, Y. Zhang, Y. Dong, F. He, and C. Wu, "Predicting the Popularity of Online News Based on Multivariate Analysis," in *IEEE CIT 2017 - 17th IEEE International Conference on Computer and Information Technology*, Aug. 2017, pp. 9–15, doi: 10.1109/CIT.2017.36.
- [6] A. Tatar, P. Antoniadis, M. D. de Amorim, and S. Fdida, "Ranking News Articles Based on Popularity Prediction," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Aug. 2012, pp. 106–110, doi: 10.1109/ASONAM.2012.28.
- [7] M. Tsagkias, W. Weerkamp, and M. de Rijke, "News Comments: Exploring, Modeling, and Online Prediction," in *Advances in Information Retrieval. ECIR*, 2010, pp. 191–203, doi: 10.1007/978-3-642-12275-0\_19.
- [8] H. M. Al-Mutairi and M. B. Khan, "Predicting the Popularity of Trending Arabic Wikipedia Articles Based on External Stimulants Using Data/Text Mining Techniques," in *2015 International Conference on Cloud Computing (ICCC)*, Apr. 2015, pp. 1–6, doi: 10.1109/CLOUDCOMP.2015.7149651.
- [9] Piotrkowicz A, Dimitrova V, Otterbacher J, and Markert K, "Headlines matter: Using headlines to predict the popularity of news articles on twitter and facebook," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2017, pp. 656–659, Accessed: Dec. 16, 2021. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/14951>.
- [10] Rieis J, de Souza F, de Melo PV, Prates R, Kwak H, and An J, "Breaking the News: First Impressions Matter on Online News," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2015, pp. 357–366, Accessed: Dec. 16, 2021. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/14619>.
- [11] M. Tsagkias, W. Weerkamp, and M. de Rijke, "Predicting the volume of comments on online news stories," in *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*, 2009, pp. 1765–1768, doi: 10.1145/1645953.1646225.
- [12] E. Hensinger, I. Flaouanas, and N. Cristianini, "Modelling and predicting news popularity," *Pattern Anal. Appl.*, vol. 16, no. 4, pp. 623–635, Nov. 2013, doi: 10.1007/s10044-012-0314-6.
- [13] M. T. Uddin, M. J. A. Patwary, T. Ahsan, and M. S. Alam, "Predicting the popularity of online news from content metadata," in *International Conference on Innovations in Science, Engineering and Technology, ICISSET*, Oct. 2016, pp. 1–5, doi: 10.1109/ICISSET.2016.7856498.
- [14] I. Arapakis, B. B. Cambazoglu, and M. Lalmas, "On the Feasibility of Predicting News Popularity at Cold Start," in *International Conference on Social Informatics*, 2014, pp. 290–299, doi: 10.1007/978-3-319-13734-6\_21.
- [15] Kate R et al., "Learning to predict readability using diverse linguistic features," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 546–554, Accessed: Dec. 16, 2021. [Online]. Available: <https://aclanthology.org/C10-1062>.
- [16] J. Berger and K. L. Milkman, "Emotion and Virality: What Makes Online Content Go Viral?," *GfK Mark. Intell. Rev.*, vol. 5, no. 1, pp. 18–23, May 2013, doi: 10.2478/gfkmir-2014-0022.
- [17] Mohammad SM, "Word affect intensities," in *Proceedings of the 11th International Conference on Language Resources and Evaluation*, 2018, pp. 174–183, Accessed: Dec. 16, 2021. [Online]. Available: <https://aclanthology.org/L18-1027>.
- [18] K. Fernandes, P. Vinagre, and P. Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News," 2015, doi: 10.1007/978-3-319-23485-4\_53.
- [19] Blood DJ and Phillips PC, "Economic headline news on the agenda: New approaches to understanding causes and effects," in *Communication and democracy: Exploring the intellectual frontiers in agendasetting theory*, 1st ed., Cowles Foundation for Research in Economics, 1997, pp. 97–113.
- [20] J. Holsanova, H. Rahm, and K. Holmqvist, "Entry points and reading paths on newspaper spreads: comparing a semiotic analysis with eye-tracking measurements," *Vis. Commun.*, vol. 5, no. 1, pp. 65–93, Feb. 2006, doi: 10.1177/1470357206061005.
- [21] R. Suharshala, K. Anoop, and V. L. Lajish, "Cross-Domain Sentiment Analysis on Social Media Interactions using Senti-Lexicon based Hybrid Features," in *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, Nov. 2018, pp. 772–777, doi: 10.1109/ICICT43934.2018.9034272.
- [22] V. L. Lajish, K. Anoop, and Gangan Manjary P, "The Impact of Online Product Reviews and Social Media Interactions on Consumer Purchase Behaviour Prediction: an Opinion Mining Perspective," in *Infinite: Frontiers of Research in Mathematical and Computing Science*, vol. 2, UGC-HRDC, University of Calicut, 2018, pp. 1–14.
- [23] D. Rahmawati and M. L. Khodra, "Word2vec semantic representation in multilabel classification for Indonesian news article," in *4th IGNITE Conference and 2016 International Conference on Advanced Informatics: Concepts, Theory and Application, ICAICTA 2016*, Aug. 2016, pp. 1–6, doi: 10.1109/ICAICTA.2016.7803115.
- [24] A. M. Dai, C. Olah, and Q. V. Le, "Document Embedding with Paragraph Vectors," Jul. 2015, Accessed: Dec. 16, 2021. [Online]. Available: <http://arxiv.org/abs/1507.07998>.
- [25] Hutto C and Gilbert E, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2014, pp. 216–225, Accessed: Dec. 16, 2021. [Online]. Available:

<https://ojs.aaai.org/index.php/ICWSM/article/view/14550>.




- [26] C. V. D, "Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 5, pp. 4452–4459, Oct. 2019, doi: 10.11591/ijece.v9i5.pp4452-4459.
- [27] Pennebaker JW, Francis ME, and Booth RJ, "Linguistic Inquiry and Word Count (LIWC): LIWC2001," in *Mahway: Lawrence Erlbaum Associates*, 2001, pp. 206–229.

## BIOGRAPHIES OF AUTHORS






**Suharshala Rajagopal**    received a Master of Philosophy (M.Phil.) in Computer Science from the University of Calicut, India in 2019. She holds her Master of Science (M.Sc.) and Bachelor of Science (B.Sc.) degrees in Computer Science from the same University in 2017 and 2015, respectively. Her research area is Big Data Mining and has publications in this field. She can be contacted at email: suharshala@gmail.com.






**Anoop Kadan**    is currently a Ph.D. research scholar in the Computer Science Department of the University of Calicut, India. Anoop holds a Master of Philosophy (M.Phil.) and Master of Science (M.Sc.) in Computer Science from the University of Calicut. His research interests are in Affective Computing, Fake News Detection, Natural Language Processing, and Machine Learning. He can be contacted at email: anoopk\_dcs@uc.ac.in.



**Manjary Praappanadan Gangadharan**    holds a Master of Philosophy (M.Phil.) in Computer Science from the University of Calicut and a Master of Technology (M.Tech.) from Amrita Vishwa Vidyapeetham. She is currently a Ph.D. research scholar in the Computer Science Department of the University of Calicut. Her research interests are in Digital Image Processing and Machine Learning. She can be contacted at email: manjaryp\_dcs@uc.ac.in.



**Dr. Lajish Vimala Lakshmanan**    has been associated with the Department of Computer Science, University of Calicut, India, since 17th January 2011. He has worked as Scientist R&D in the TCS Innovation Labs Mumbai, Tata Consultancy Services Ltd, during May 2007 – Jan 2011. His prime areas of research interests include Computational Intelligence, Data Analytics, Indian Language Speech and Script Technology Solutions. After his Masters in Computer Science from Vellore Institute of Technology, Lajish obtained his Ph.D in Computer Science from University of Calicut, Kerala in March 2007. Dr. Lajish has One US Patent, Two Indian patents, 2 books edited, 2 book chapters and more than 90 research publications in peer-reviewed International Journals and National/International Conferences to his credit. He has successfully guided 4 doctoral students and 14 MPhil scholars in Computer Science. He is associated various professional bodies including as the member of Association of Computing Machinery (ACM). He can be contacted at email: lajish@uc.ac.in.