Information Gain Based Feature Selection for Improved Textual Sentiment Analysis

Madhumathi Ramasamy1 · A. Meena Kowshalya2

Accepted: 7 February 2022 / Published online: 13 March 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
Sentiment analysis or opinion mining is the process of mining the emotion from a given text. It is a text mining technique that effectively measures the inclination of public opinions and aids in analysing the subjective information from the given context. Sentiment analysis evaluates the opinion of a sentiment as either positive or negative or neutral. Sentiments are very specific and with respect to the underlying content, it plays a very crucial role in depicting the real-world scenario. Sentiment analysis can be performed at three levels namely document level, sentence level and feature level. This paper proposes a novel Information Gain based Feature Selection algorithm that selects highly correlated features by removing inappropriate content. Using this algorithm, extensive sentimental analysis is performed at the document level, sentence level and feature level. Datasets from Cornell and Kaggle are exploited for experimental purposes. Compared to other baseline classifiers experimental results show that the proposed Information Gain based classifier resulted in an accuracy of 95, 96.3 and 97.4% for document, sentence and feature levels respectively. The proposed method is also tested with higher dimensional datasets namely Movielens 1M, 10M and 25M datasets. Experimental results proved that the proposed method works better even for high dimensional datasets.

Keywords Information gain · Textual sentiment analysis · Feature selection · Classification

1 Introduction
Sentiment analysis uses Natural Language Processing (NLP) and Machine Learning (ML) algorithms that directly quantifies the sentiments towards products, people or their ideas and assist in revealing the contextual polarity. Sharing and expressing people’s views and opinions is as simple as possible in the current era. Social media generates a large volume of data and this enormous volume of data is very difficult to intercept, describe and analyse. Sentiment analysis is one effective solution to identify,
extract or mine measure and study the context of these big data. It is a computational approach where the sentiments, opinions and subjectivity of the text is analysed [1, 2]. The applications of sentiment analysis extend widely from healthcare, business, e-commerce, financial services to even social networks and marketing [3, 4]. It is also used for classifying product reviews, movie reviews, social media news, online reviews, English texts in documents and many more [5, 6].

Document level sentiment analysis determines the overall opinion about the document. It assumes that each document express opinions on a single entity. The entire document is classified as positive and negative [7]. For example, if a customer reviews a product, the system determines the overall opinion and generates a prediction model. One drawback with document level sentiment analysis is that multiple entry evaluation and comparison is not possible [8]. Sentence level sentiment analysis determines whether each sentence has expressed an opinion. It discriminates the objective sentences expressing factual information with the subjective sentences expressing opinions. To measure the effectiveness of sentence level sentiment analysis, the following is done.

i. Identify whether the sentence has expressed an opinion or not.
ii. Then, assess the polarity of the opinion

One major drawback of sentence level sentiment analysis is that prediction becomes difficult if the objective sentence carries opinions. Sentiment analysis at the feature level or aspect level requires a finer level of analysis with the help of NLP techniques. The opinions are characterized by polarity and a target of opinions. To measure the effectiveness of feature level sentiment analysis, the following is done.

i. Identify the entity and aspect of the entity in question.
ii. Then, access the opinion on each aspect. Feature level in general has three stages namely identification and extraction of object features, determination of polarity of opinions and estimation of sentiments. Feature level sentiment analysis will provide more fine-grained analysis on opinion targets and has a wider range of applications in e-commerce [9, 10].

This paper proposes a novel Information Gain based Feature Selection algorithm that selects highly correlated features for classification. Using this algorithm, sentiment analysis is performed at document, sentence and feature levels. The key contribution of this paper is,

i. build an efficient Information Gain based Feature Selection algorithm.
ii. Gather features for extensive sentiment analysis.
iii. Perform sentiment analysis at document, sentence and feature levels using the proposed algorithm.
iv. Test the performance on high dimensional datasets.

The rest of the paper is organized as follows: Section 2 describes the related work. Section 3 explains the proposed Information Gain technique. In Sect. 4, the results of the proposed algorithm are discussed and Sect. 5 concludes the paper.
2 Related Work

Sentiment analysis in textual context classifies the subjective information based on the correlation between the selected features. Many machine learning algorithms are used for sentiment analysis which improves the accuracy of document, sentence and feature level textual data [11]. A three-step approach was used for analysis in [12] which uses different preprocessing, feature extraction and postprocessing techniques. The authors in [13] presented an approach which recognizes the polarity of a sentence as positive, neutral and negative according to the semantic information learned from news articles and blogs. Sentence level analysis obtains the relevant categories of a single sentence relating to grammar, content and punctuation. A method was proposed in [14] to extract sentence level features and domain sensitive features based on a disambiguation word technique. The performance of the supervised analysis increased when combining sentence level features and domain sensitive features. The authors in [15] proposed a feature extraction technique using Information Gain for aspect level sentiment analysis. The textual data is preprocessed and the information related to the lemma, grammatical relations and the Part-Of-Speech (POS) tags were obtained. Sentiment analysis are performed for movie reviews and twitter data based on different algorithms. In [16], the unigram and bigram frequency values of the movie reviews were explored using Naïve Bayes (NB), maximum entropy and Support Vector Machines (SVM) and in [17] twitter data was analyzed using Naïve Bayes classification algorithm with the features obtained by Information Gain based approach. The sentiments for twitter messages are also classified by exploring the utility of the linguistic features obtained at the feature selection level [18].

The authors in [19] proposed a text classification process using machine learning techniques. Various classification algorithms namely NB, SVM, Logistic Regression (LR), Decision Tree (DT) and K-Nearest Neighbor (KNN) are used for scientific textual sentiment analysis [20]. The authors in [21] proposed an opinion question answering system for differentiating news stories in document and sentence levels using Bayesian classifiers. Along with machine learning, deep learning methods are also used to solve sentiment polarity problems as discussed in [22]. Deep convolutional neural networks based on Stanford Sentiment Treebank (SSTb) and Stanford Twitter Sentiment corpus (STS) parse tree methods were used by the authors in [23] for sentence level analysis. The authors in [24] proposed an approach for sentiment analysis in which the utility of linguistic features is explored for analyzing Twitter messages. Various algorithms namely ID3, NB, DT and Random Forest are used in [25] for classifying sentiments in English texts in both balanced and unbalanced datasets. The authors in [26] proposed a classifier in which polarity of subjective phrases are predicted using Dictionary of Affect in Language (DAL) for lexical scoring with WordNet.

Sentiment analysis detects opinions, attitudes, and feelings expressed in the text using automated tools [27]. The authors in [28] proposed an experimental work for aspect-level analysis on movie reviews based on a new kind of domain-specific feature-based heuristic approach by selecting linguistic features comprising of adverbs, adjectives and n-gram features. Latent Factor Models (LFM) is based on Collaborative Filtering (CF) gives good prediction accuracy [29]. Explicit Factor Model (EFM) is used to produce recommendations with better prediction accuracy. It is essential to consider the importance of sentences when classifying documents. A sentence classification model using deep neural networks was proposed in [30] where a gate mechanism was used for automatically determining the degrees of each sentence. Hence for analyzing the sentiments with good accuracy an Information Gain based Feature Selection
algorithm is proposed which assess the sentiments at all three levels namely document, sentence and feature levels.

3 Proposed Work

The paper proposes a simple and novel feature selection algorithm based on the concept of Information Gain theory. The proposed Information Gain technique is able to extract/select highly correlated feature vectors at optimal time and this enables more accurate sentiment classification. The proposed algorithm fits and works well when tested with document, sentence and feature levels. The algorithm selects the feature that has highest relevance and occurrence to the target class and best application performance is rendered because most of the irrelevant and redundant features are removed.

- The algorithm estimates the mutual information between each feature and the class label.
- The features with the highest correlations are extracted.
- The relevance between the after bites are measured using Information Gain as shown in Eq. (7).

For any two random variables R and S where \( R = \{r_1, r_2, r_3, \ldots, r_N\} \) and \( S = \{s_1, s_2, s_3, \ldots, s_M\} \), the entropy of R is denoted in Eq. (1) as \( H(R) \). The conditional entropy is defined in Eq. (3) and joint entropy of R and S is defined in Eq. (4) as \( H(S, R) \).

\[
H(R) = - \sum_{i=1}^{N} p(r_i) \log(p(r_i)) \tag{1}
\]

Where, \( p(r_i) = \frac{\text{Number of instants with value } r_i}{\text{Total number of instants of } R} \) \tag{2}

\[
H(S|R) = - \sum_{r_i \in R} \sum_{s_j \in S} p(s_j, r_i) \log(p(s_j|r_i)) \tag{3}
\]

\[
H(S, R) = - \sum_{r_i \in R} \sum_{s_j \in S} p(s_j, r_i) \log(p(s_j, r_i)) \tag{4}
\]

The mutual information between variable S and R is denoted as \( I(S;R) \) as shown in Eqs. (5) and (6).

\[
I(S;R) = \begin{cases} 
H(S) - H(S|R) \\
H(R) - H(R|S) \\
H(S) + H(R) - H(S, R)
\end{cases} \tag{5}
\]

\[
I(S;R) = \sum_{r_i \in R} \sum_{s_j \in S} p(s, r) \log \frac{p(s, r)}{p(s)p(r)} \tag{6}
\]

With respect to a Class C and attribute A, the Information Gain I is calculated according to Eq. (7).
Information Gain Based Feature Selection for Improved Textual…

\[ I(C, A) = 1 - H(C|A) \]  

(7)

Where \( H(C|A) \) is the conditional entropy of the class \( C \) given any attribute \( A \). when \( H(C|A) = 1 \), then \( I(C, A) \) will be equal to 1.

**Algorithm**

Input: Set of features for training

Output: Optimal subset of feature 'S'

Let A represent the instant set of features, C represent the class table f{} be an empty set

1. Begin
2. for each set of features
3. Do
4. Case: Find Information Gain \( IG(f_i, c) \) as per Eq. (7)
5. Find the first feature \( f_i \) that maximizes \( IG(f_i, c) \)
6. \( S = S \cup f_i \)
7. \( A=A-f_i \)
8. Repeat until optimal subset reached
9. For all pairs of features (\( f_i, f_s \))
10. Compute \( IG(f_i, c) \)
11. Select subset of features such that \( f_i \) maximizes \( IG(f_i, c) - \sum_{f_s \in S} IG(f_i, f_s) \times IG(f_i, f_s) \)
12. Return S
13. End

The algorithm finds initial set of features and computes IG for all of them. The first feature is taken and the algorithm appends the highly correlated features to it. This procedure is repeated until an optimal subset is reached. The IG is computed and the final set of features obtained are used to predict the sentiments at document, sentence and feature levels.

**3.1 Document Level Sentiment Analysis**

For performing document level sentiment analysis, the documents are aggregated and preprocessed. A document level sentiment analysis considers the entire document as a single entity and classifies the document as either positive or negative. This paper considers unigrams, bigrams and trigrams for evaluating the underlying sentiment. Term Frequency-Inverse Document Frequency (TF-IDF) is used to calculate the n-gram features vector as shown in Eq. (8).

\[ N_{id} = \log \left( 1 + TF_{id} \right) \times \frac{\log |D|}{df_i} \]  

(8)

Where \( TF_{id} \) is the occurrence of the \( i^{th} \) feature in the document \( D \), \( |D| \) is the total number of documents, \( df_i \) is the number of documents containing \( i^{th} \) feature.
Algorithm: Document level sentiment analysis
Input: Document D
Output: Positive, Negative
1. Begin
2. For the given set of Documents D
3. Do
4. Extract opinion nodes exploiting POS tagger
5. Generate seed list S according to Eq. (8)
6. Identify candidate words using N-gram
7. End for
8. For all the feature list in D
9. Do
10. Calculate TF-IDF
11. If frequency $\geq 0.2$ then
12. Feature set (FS) $\leftarrow$ FS + feature
13. End for
14. For every feature in FS
15. Do
16. Calculate IG(C/A)
17. End for
18. For every IG score in IG(C/A)
19. Do
20. If IG(C/A) $= 0.5$ then
21. Vocabulary V $\leftarrow$ V + A
22. If $P(A) = P(A/C) + A$
23. $FS_{positive} \leftarrow FS_{positive} + A$
24. Else
25. $FS_{negative} \leftarrow FS_{negative} + A$
26. End if
27. End if
28. End for
29. End

The algorithm calculates document level analysis by extracting the opinion words obtained from POS tagging technique for a set of documents. A seed list is prepared and candidate words are searched using n-grams according to Eq. (8). The TF-IDF is calculated and the IG is computed for all set of features in the feature set. The resultant attribute is added to the vocabulary from which the positive and negative feature sets are obtained.
3.2 Sentence Level Sentiment Analysis

Performing sentiment analysis at sentence and sub sentence level identifies fine grained expressions, polarities, strength of the expression and its relationship to the subject. Each sentence is POS tagged and has a dependency tree that is created using Stanford hex parser [31]. The proposed sentence level sentiment analysis algorithm outputs the overall sentiment of the sentence by completely analyzing phrases within sentences. The polarity of individual phrases is determined by Default Polarity Calculation (DPC) with WordNet and this method starts by identifying positive or negative words using the General Inquirer (GI) [32]. In case of a found word in the GI list, the WordNet dictionary searches for semantic corrections with respect to the GI list.

The sentence level sentiment algorithm performs POS tagging and creates a dependency tree for a given set of input sentences. The individual phrases are identified and their polarities are calculated using DPC method. The algorithm finds IG for each feature and the same procedure is repeated as discussed in Sect. 3.1 for finding the positive and negative sentiments.
Algorithm: Sentence level sentiment analysis
Input: Sentence S
Output: Positive, Negative polarities

1. Begin
2. For all Sentences S = \{S_1, S_2, \ldots, S_n\}
3. Do
4. Perform POS tagging
5. Generate Dependency tree
6. Identify individual phrases P = \{P_1, P_2, \ldots, P_n\}
7. Calculate Polarity of P using DPC
8. End for
9. For every feature in S
10. Do
11. Calculate TF-IDF
12. If frequency \geq 0.2 then
13. Feature Set (FS) \leftarrow FS + feature
14. End if
15. End for
16. For every feature in feature subset
17. Do
18. Calculate IG(C/A)
19. End for
20. For every IG score in IG(C/A)
21. Do
22. If IG(C/A) = 0.5 then
23. Vocabulary V \leftarrow V + A
24. If P(A) = P(A/C) + A then
25. FS_{positive} \leftarrow FS_{positive} + A
26. Else
27. FS_{negative} \leftarrow FS_{negative} + A
28. End if
29. End if
30. End for
31. End

3.3 Feature Level Sentiment Analysis

In feature level analysis, the TF-IDF is used for feature selection. The Document Frequency (DF) thresholding technique already proved in literature suffer from overfitting while selecting the maximum appearing features in the training data. Since DF thresholding suffers from overfitting, the conventional method is chosen. The proposed feature level sentiment analysis algorithm results in a reduction of redundant and unrelated features thus classifying the sentiments effectively.
Algorithm: Feature level sentiment analysis
Input: Attributes A and class C
Output: Positive, Negative feature set

1. Begin
2. For every feature
3. Do
4. Calculate TF-IDF
5. If frequency $\geq 0.2$ then
6. Feature Set (FS) $\leftarrow$ FS + feature
7. End if
8. End for
9. For every feature in feature subset
10. Do
11. Calculate IG(C/A)
12. End for
13. For every IG score in IG(C/A)
14. do
15. If IG(C/A) = 0.5 then
16. Vocabulary V $\leftarrow$ V + A
17. If P(A) = P(A/C) + A then
18. FS$_{positive}$ $\leftarrow$ FS$_{positive}$ + A
19. Else
20. FS$_{negative}$ $\leftarrow$ FS$_{negative}$ + A
21. End if
22. End if
23. End for
24. End

3.4 Classification

The Classification algorithm classifies the sentiments based on various levels as positive or negative.
Table 1  Dataset for sentiment analysis

| S. No | Document Level | Sentence Level | Feature Level |
|-------|---------------|----------------|---------------|
| 1.    | 2000 documents | 10000K sentences | 70000K features |

Fig. 1  Precision

![Precision graph](image)

Fig. 2  Recall

![Recall graph](image)

Fig. 3  F1-Score

![F1-Score graph](image)
Algorithm: Classification
Input: Sentiment Feature Vector (V x D)
Output: Sentiments (Positive, Negative)

1. Begin
2. For each vocab in V do
3. If vocab is positive feature then
4. Positive ← positive + 1
5. Else
6. Negative ← negative + 1
7. End if
8. If positive > negative then
9. Class label_{positive} ← Class label_{positive} + ‘positive’
10. Else
11. Class label_{negative} ← Class label_{negative} + ‘negative’
12. End if
13. End for
14. End

4 Experimental Results

The datasets for document level, sentence level and feature level sentiment analysis are taken from the Cornell repository consisting of 2000 documents, 10000K sentences and Kaggle repository consisting of 70000K features respectively. Table 1 shows the details of the datasets used.

The results of document, sentence and feature level sentiment analysis are validated and verified using the metrics such as Precision, Recall, F1-Score and Accuracy. The experiments proved that the proposed algorithm outperformed all the other algorithms by providing the highest level of accuracy at all the three levels. Figs 1, 2 and 3 shows the Precision, Recall and F1-Score of the proposed and other existing classifiers.
### Table 2 Comparison of the Proposed Method and Other Baseline Classifiers for Movielens 25M Dataset

| Algorithm     | Precision | Recall | F1-Score |
|---------------|-----------|--------|----------|
|               | Document level | Sentence level | Feature level | Document level | Sentence level | Feature level | Document level | Sentence level | Feature level |
| Proposed Algorithm | 0.96        | 0.95    | 0.97     | 0.96        | 0.96    | 0.98    | 0.96        | 0.95    | 0.97     |
| Random Forest | 0.84        | 0.56    | 0.92     | 0.84        | 0.55    | 0.95    | 0.84        | 0.55    | 0.93     |
| SVM           | 0.83        | 0.57    | 0.55     | 0.85        | 0.56    | 0.58    | 0.84        | 0.56    | 0.56     |
| Naive Bayes   | 0.81        | 0.52    | 0.51     | 0.82        | 0.56    | 0.57    | 0.81        | 0.54    | 0.54     |
| KNN           | 0.85        | 0.60    | 0.63     | 0.81        | 0.60    | 0.62    | 0.83        | 0.60    | 0.62     |
Fig. 4 compares the accuracy of the proposed algorithm with other baseline classifiers. This infers that the proposed Information Gain based classifier resulted in the highest accuracy at document, sentence, and feature levels.

We attempted to analyze the performance of the proposed Information Gain based Feature Selection algorithm on document level, sentence level, and feature levels for high dimensional datasets. The Movielens dataset with 1 million, 10 million, and 25 million ratings were chosen for experimental purposes. Table 2 presents the Precision, Recall, and F1-Score of the proposed and the existing classifiers using Movielens 25M dataset.

Figure 5 shows the accuracy of Movielens 1M, 10M, and 25M datasets of the proposed algorithm for all three levels. Table 3 depicts the comparison of proposed and existing algorithms for the accuracy of Movielens 1M, 10M, and 25M for document level, sentence level, and feature level respectively. For higher dimensional datasets, the proposed method works better and outperforms the other existing baseline classifiers in terms of Precision, Recall, F1-Score, and Accuracy.

5 Conclusion

Sentiment analysis is a frequently used text classification tool that evaluates the underlying sentiment and classifies it as positive or negative. This paper proposes an Information Gain based Feature Selection algorithm. Using this algorithm, an extensive sentiment analysis is performed at document, sentence, and feature levels. The proposed Information Gain based Feature Selection method selects the features that have greater Information Gain with

\[
\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (9)
\]

\[
\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (10)
\]

\[
\text{F1 - Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (11)
\]

\[
\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \quad (12)
\]
| Algorithm       | Document level |                | Sentence level |                | Feature level |                |
|-----------------|----------------|----------------|----------------|----------------|---------------|----------------|
|                 | Movielens 1M (%) | Movielens 10M (%) | Movielens 25M (%) | Movielens 1M (%) | Movielens 10M (%) | Movielens 25M (%) | Movielens 1M (%) | Movielens 10M (%) | Movielens 25M (%) |
| Proposed Algorithm | 94             | 93             | 94             | 97             | 96             | 95             | 94             | 97             | 97             |
| Random Forest    | 81             | 80             | 80             | 91             | 89             | 57             | 88             | 89             | 96             |
| SVM             | 83             | 83             | 85             | 52             | 51             | 54             | 56             | 52             | 57             |
| Naive Bayes     | 80             | 80             | 81             | 52             | 55             | 52             | 55             | 52             | 55             |
| KNN             | 86             | 86             | 87             | 55             | 54             | 58             | 62             | 67             | 60             |
the highest occurrence. Experimental results proved that the proposed Information Gain model produced high accuracy of 95%, 96.3% and 97.4% for document, sentence and feature levels than the existing baseline classifiers. Also, the proposed method is tested with Movielens dataset with 1 million, 10 million and 25 million ratings. Despite the size of the datasets, the proposed algorithm outperformed the state of art classifiers.

**Funding** Not applicable.

**Data Availability** Enquiries about data availability should be directed to the authors.

**Declarations**

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical Approval** This article does not contain any studies with human participants performed by any of the authors.

**References**

1. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011
2. Yaakub, M. R., Latiffi, A., Iqbal, M., & Safra, L. (2019). A review on sentiment analysis techniques and applications. *IOP Conference Series: Materials Science and Engineering*. https://doi.org/10.1088/1757-899X/551/1/012070
3. Kowshalya, A. M., Madhumathi, R., & Gopika, N. (2019). Correlation based feature selection algorithms for varying datasets of different dimensionality. *Wireless Personal Communications, 108*(5), 1977–1993. https://doi.org/10.1007/s11277-019-06504-w
4. Kowshalya, A. M., & Valarmathi, M. L. (2018). Evaluating twitter data to discover user’s perception about social Internet of Things. *Wireless Personal Communications*, *101*(2), 649–659. https://doi.org/10.1007/s11277-018-5709-2
5. Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and Its application: Systematic literature review. *Procedia Computer Science, 161*, 707–714. https://doi.org/10.1016/j.procs.2019.11.174
6. Vamsi, B., Suneetha, N., Sudhakar, Ch., & Amaravati, K. (2017). Sentiment analysis on online reviews using supervised learning: A survey. *International Journal of Control Theory and Applications, 10*(30), 143–152.
7. Behdenna, S., Barigou, F., & Belalem, G. (2018). Document level sentiment analysis: A survey. *EAI Endorsed Transactions on Context-Aware Systems and Applications, 4*(13), 1–8. https://doi.org/10.4108/eari.14-3.2018.154339
8. Nicholls, C., & Song, F. (2010). Comparison of feature selection methods for sentiment analysis.” In: *Proceedings of Canadian Conference on Artificial Intelligence*, Springer, pp. 286–289, 2010. https://doi.org/10.1007/978-3-642-13059-5_30.
9. Quan, C., & Ren, F. (2016). Feature-level sentiment analysis by using comparative domain corpora. *Enterprise Information Systems, 10*(5), 505–522. https://doi.org/10.1080/17517575.2014.985613
10. Pratiwi, A. I., & Adiwijaya,. (2018). On the feature selection and classification based on information gain for document sentiment analysis. *Applied Computational Intelligence and Soft Computing, 2018*, 1–5. https://doi.org/10.1155/2018/1407817
11. Gupta, S. L., & Baghel, A. S. (2018). Efficient feature extraction in sentiment classification for contrastive sentences. *International Journal of Modern Education and Computing, 5*, 54–62. https://doi.org/10.5815/ijmecs.2018.05.07
12. Jagdale, R. S., Shirsath, V., & Deshmukh, S. (2019). Sentiment analysis on product reviews using machine learning techniques. *Advances in Intelligent Systems and Computing, 768*, 639–647. https://doi.org/10.1007/978-981-13-0617-4_61
13. Shirsat, V., Jagdale, R., Shende, K., Deshmukh, S. N., & Kawale, S. (2019). Sentence level sentiment analysis from news articles and blogs using machine learning techniques. *International Journal of Computer Sciences and Engineering, 7*(5), 1–6.

14. Rintyarna, B. S., Sarno, R., & Fatchah, C. (2019). Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks. *Journal of Big Data, 6*, 1–19.

15. Schouten, K., Frasincar, F., & R. Dekker, R. (2016). An information gain-driven feature study for aspect-based sentiment analysis. In: *Proceedings of International Conference on Applications of Natural Language to Information Systems*, pp. 48–59, 2016. https://doi.org/10.1007/978-3-319-41754-7_5.

16. Franky, & Manurung, R. (2008). “Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews. In: *Proceedings of International Conference on Advanced Computational Intelligence and Its Applications (ICACIA)*, Depok, Indonesia, Jan, 2008.

17. Singh, M., & Gupta, S. (2020). Sentiment analysis using Naive Bayes classifier and information gain feature selection over twitter. *International Journal of Computer Trends and Technology, 68*(5), 84–91.

18. Saif, H., Y. He, Y. & Alani, H. (2012). “Semantic Sentiment Analysis of Twitter”. In: *Proceedings of 11th International conference on The Semantic Web*, vol.1, pp 508–524, 2012. https://doi.org/10.1007/978-3-642-35176-1_32.

19. Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *SEAS Transactions on Computers, 4*(8), 966–974.

20. Raza, H., Faizan, M., Hamza, A., Mushraq, A., & Akhtar, N. (2019). Scientific text sentiment analysis using machine learning techniques. *International Journal of Advanced Computer Science and Applications, 10*(12), 157–165.

21. Yu, H., & Hatzivassiloglou, V. (2003). “Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In: *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Jul, 2003. https://doi.org/10.3115/1119355.1119372.

22. Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Multidisciplinary Digital Publishing Institute Journal, 9*, 1–29. https://doi.org/10.3390/electronics9030483.

23. dos Santos, C.N., & M. Gatti, M. (2014). “Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts”. In: *Proceedings of 25th International Conference on Computational Linguistics: Technical Papers*, COLING, pp. 69-78, Aug, 2014. http://creativecommons.org/licenses/by/4.0.

24. Kouloumpis, E., Wilson, T. & Moore, J. (2011). “Twitter Sentiment Analysis: The Good the Bad and the OMG!”. In: *Proceedings of International Conference on Weblogs and Social Media*, vol. 2, Jul, 2011.

25. Alshamsi, A., Bayari, R., & Salloum, S. (2020). Sentiment Analysis in English Texts. *Advances in Science, Technology and Engineering Systems Journal, 5*(6), 1683–1689.

26. Agarwal, A., Biadsy, F., & Mckeown, K.R. (2009). Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams. In: *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 24–32, Mar, 2009. https://doi.org/10.3115/1609067.1609069.

27. Lin, C., & He, Y. (2009). “Joint sentiment/topic model for sentiment analysis.” In: *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 375–384, Nov, 2009. https://doi.org/10.1145/1645953.1646003.

28. Singh, V.K., Piryani, R., Uddin, A., & Waila, P. (2013). “Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification”. In: *Proceedings of the International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, Mar, 2013. https://doi.org/10.1109/iMac4s.2013.6526500.

29. Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., Ma, S., (2014). “Explicit factor models for explainable recommendation based on phrase-level sentiment analysis”. In: *Proceedings of the 37th International ACM SIGIR conference on Research & development in information retrieval*, pp. 83–92, Jul, 2014. https://doi.org/10.1145/2600428.2609579.

30. Choi, G., Oh, S., & Kim, H. (2020). Improving document-level sentiment classification using importance of sentences. *Multidisciplinary Digital Publishing Institute Journal, 22*, 1–11. https://doi.org/10.3390/e22121336

31. The Stanford Natural Language Processing Group (http://nlp.stanford.edu/software/lexparser.shtml).
32. Stone, P. J., Dunphy, D., Smith, M. S., & Ogilvie, D. M. (1966). The general inquirer: A computer approach to content analysis. The MIT Press.

Publisher's Note  Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dr. R. Madhumathi working as an Assistant Professor in Computer Science and Engineering Department at Sri Ramakrishna Engineering College, Coimbatore. She received her PhD degree under Information and Communication Engineering at Anna University, Chennai in the year 2016, Master of Engineering in Software Engineering at Sri Ramakrishna Engineering College, Coimbatore in the year 2009 and Bachelor of Engineering in Computer Science & Engineering at Jayam College of Engineering & Technology, Dharmapuri in the year 2002. She has 11 years of academic experience and 8 years of research experience. Her area of specialization is Cloud Computing, Big Data, Data Analytics and Artificial Intelligence. She is a member of IEEE Women in Engineering, life member in ISTE, CSI and ISAE.

Dr. A. Meena Kowshalya graduated from Bharathiar University. She obtained her Masters and Ph.D degree in Computer Science and Engineering from Anna University. She is a gold medal winner for the academic year 2011. She is the winner of AICTE Viswakarma Awards for the year 2019. She currently serves as an Assistant professor in Government College of Technology, Coimbatore, TamilNadu, India. She has total of 14 years teaching experience. Her area of research interest includes Information Retrieval, Image Captioning, Artificial Intelligence and Machine Learning.