# Fuzzy Based Latent Dirichlet Allocation in Spatio-Temporal and Randomized Least Angle Regression

D. Nithya(✉) and S. Sivakumari

Department of Computer Science and Engineering, School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India
nithya.apcse@gmail.com, prof.sivakumari@gmail.com

**Abstract.** Due to the emerging growth of Social Networks and web blogs, several news providers share their news articles on different web sites and web blogs. It is also used to get public opinion about the articles. Twitter is one of the popular microblogs which act as intermediate for publics to distribute their thoughts. Our intention is to find Twitter data related to the news in the web news articles are also used to enhance the performance of Evolving Fuzzy System-Penguins Search Optimization Algorithm (EFS-PeSOA) based web news mining. In this paper, Latent Dirichlet Allocation (LDA) is used to model the topics within the text of tweets. Twitter-specific tokenizer, part-of-speech tagger, and snowball stemmer are used to generate terms from the twitter data. Term-Frequency and Inverse-Document-Frequency (tf-idf) of each term in tweets are calculated along with the terms in web news articles for the creation of evolving fuzzy rules. Based on the evolving fuzzy rules, web news articles are categorized. In order to enhance the efficiency of categorization of web news articles, a Spatio-Temporal Generalized Additive Model (STGM) is developed where the spatial and temporal information of the tweets are considered for categorization. However, it generates different terms in tweets. So a Randomized Least Angle Regression (RLAR) is used to choose the most significant terms in tweets and only the selected term's tf-idf values are used in EFS for categorization of web news articles.

**Keywords:** Web news articles · Evolving Fuzzy System · Latent Dirichlet Allocation · Spatio-Temporal Generalized Additive Model · Randomized Least Angle Regression

## 1 Introduction

Text classification [1, 2] is the process of assigning a text document to one or more decided categories. This will help the users to find desired information easily by searching only the relevant categories. The importance of text classification is even more apparent when the information space is large. Because of increasing the rate of web pages in the World Wide Web (WWW), the classification of text documents is more difficult. So, machine learning methods have been introduced for automating the


P. Karrupusamy et al. (Eds.): ICSCN 2019, LNDECT 39, pp. 497–508, 2020.
https://doi.org/10.1007/978-3-030-34515-0_52

classification process. Online news articles represent a type of web information that is frequently referenced. Nowadays, online news is provided by many dedicated newswires such as PR Newswires and Reuters. It will be useful to gather news from newswires sources and classify the news accordingly for ease reference.

A method based on Evolving fuzzy system [3] was proposed to categorize the web news articles according to the text content of the articles. The most relevant terms of each article were obtained by using tokenization, stopword elimination and stemming processes. Based on the relevance value of each term, a large number of terms were pruned by a user-specified pruning threshold. Then, evolving fuzzy rules were created based on the tf-idf value of terms. A Gaussian membership function was utilized in EFS to define the nearness of the article to the prototype. However, the problem of selection of the pruning threshold was solved by using a Penguins Search Optimization Algorithm (PeSOA) [4]. Instead of Gaussian membership function, a bell-shaped membership function was utilized to describe the nearness of the article to the prototype. Because sometimes the Gaussian membership function is hard to justify. Classification accuracy of web news mining is improved by utilizing PeSOA.

In this paper, the web news mining based on EFS is improved by analyzing the Twitter data which are related to the news articles. The topics of the Twitter data are modeled by LDA [5] model. Then the terms in tweets are generated by using twitter-specific tokenizer, part of speech tagger and stemming. The Term Frequency and Inverse Document Frequency (tf-idf) are calculated for each generated terms and prune the terms which have tf-idf value lesser than the optimized pruning threshold. Then, Evolving Fuzzy Rules are created based on the tf-idf of web news article terms and tf-idf of twitter terms which effectively categorize the web news articles. In addition to this, a Spatio-Temporal Generalized Additive Model (STGM) is developed to analyze the Twitter terms based on user's location and time when they tweet about the news. It generates different terms, so the most important terms are selected by using RLAR. Thus, the tf-idf value of the most important twitter terms is used in the creation of evolving fuzzy rules which improves the web news articles categorization accuracy.

## 2 Literature Survey

A news article classification [6] method was proposed to classify the news articles based on the vector representation of articles. The pre-processing process was enhanced by introducing words which reduced the word space without compromising the information value. A vector representation was introduced to improve the classification accuracy and reduce the computational complexity of the classification process. After the preprocessing process, the news articles were classified using Naïve Bayes, K-Nearest Neighbor (K-NN) and decision trees.

A Support Vector Machine (SVM) [7] was introduced for twitter news classification. Initially, small messages were taken out from the Twitter microblog by choosing different active newsgroups. Then the short messages were manually classified into 12 groups. SVM got words of each small message as features and it was trained by using the manually classified news data. Based on the trained data, the SVM classified the

news into groups. The disadvantage of using the SVM algorithm is that it needs more memory for classification in many cases and involves high complexity.

Based on random forests and multimodal features, a news article classification framework [8] was introduced. In this framework, both visual features and textual features were used for category-based classification of news articles. The textual features and visual features were taken out from the textual part and image of the news articles respectively. Then, these two textual and visual features were combined using a late fusion strategy. Random forest classifier got fused features as input and classified the web news articles.

For web news classification, a deep learning technique [9] was proposed. Initially, a training database was created in which the class of each data was known. It was used to predict the class of online web news. Then, a neural network was employed for the classification of online web news. A goal was set for the neural network by the user. If the goal was not met, then change the weight of the neural network otherwise the training sides were reselected. Online web news was classified based on the weights. However, the user-specified goal greatly influences the online web news classification.

A method [10] was proposed for the classification of tweets about the newspaper reports. Initially, a dataset was formed by using the content of an article and the tweets having a link to the news. The unnecessary symbols and words in the collected data were removed in the pre-processing phase. Then, the stemming process was applied through morphological analyzers. A textual similarity approach was exploited to find the similarity between the tweets. Finally, a binary classifier called Naïve Bayes was applied to classify the tweets about the newspaper reports. The disadvantage of using Naive Bayes requires more data to get better results.

A Fuzzy Algorithm for Extraction, Monitoring, and Classification of Communicable Diseases (FAEMC-CD) [11] model was proposed based on evolving fuzzy model. This model extracts the information about communicable diseases from Twitter and news websites. Twitter and data in news websites were pre-processed using tokenizing, stemming, stop word filtering and term filtering. Then fuzzy rules were developed using fuzzy rule-based classification package.

## 3 Proposed Methodology

In this section, the proposed method for web news mining using web news articles and Twitter data is described in detail. The topics within the twitter data are modeled by the Latent Dirichlet Allocation (LDA). Then a Twitter-specific tokenizer and part of speech tagger are used to split the tweets into tokens and label the tokens. The terms of each tweet under different topics are pruned by a term reduction technique. If the terms have a relevance value more than the optimized pruning threshold, then those terms are removed from the twitter data. Based on the Term Frequency and Inverse Document Frequency (tf-idf) of terms in web news articles and twitter data, Evolving fuzzy rules are created where the web news articles are categorized. A Spatio-Temporal Generalized Additive Model (STGM) is developed where the location and time of the tweets are considered for efficient categorization of web news articles. For the efficient selection of dynamically changing terms of tweets, a Randomized Least Angle

Regression (LAR) is used. The tf-idf of selected terms of Twitter data is used in the Evolving fuzzy rules which enhance the accuracy of web news article categorization.

### 3.1 LDA Based Topic Modeling of Twitter Data

The news articles of different topic areas are extracted from the web and tweets containing a link to the news are collected from Twitter. The topics within the text of tweets are modeled by LDA model. A twitter dataset of unlabelled tweets is given as input to the LDA model. It found the hidden topics as distributions over the words in the vocabulary. The words are represented as observed random variables and the topics are represented as latent random variables. Initially, a topic weight vector is drawn that is modeled by a Dirichlet random variable. It finds out which topics are almost certainly to appear in the Twitter data. From the topic weight vector, a single topic was chosen for each word that is to appear in the Twitter data. By a different, randomly chosen topic each word in the Twitter data is generated. The available choices of topics for each tweet are, in order, drawn from a smooth distribution over the space of all possible topics. The LDA process is described as follows:

For each tweet indexed by $m \in \{1, 2, \ldots M\}$ in a twitter data:

1. Select a K-dimensional topic weight vector $\theta_m$ from the distribution $p(\theta|\alpha) = Dirichlet(\alpha)$. The Dirichlet distribution is given as follows:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta^{\alpha_i - 1} \tag{1}$$

   Equation (1), $\Gamma$ is the gamma function and $\alpha$ is a K-dimensional parameter.

2. For each word indexed by $n \in \{1, 2, \ldots N\}$ in a tweet:
   a. Select a topic $z_n \in \{1, 2, \ldots K\}$ from the multimodal distribution $p(z_n = k|\theta_m) = \theta_m^k$.
   b. Given the chosen topic $z_n$, draw a word $w_n$ from the probability $p(w_n = i|z_n = j, \beta) = \beta_{ij}$, where $\beta$ is a K-dimensional parameter.

### 3.2 Term Generation

The news articles undergo the term generation process where the tokenization, stop word elimination and stemming processes are carried out. A twitter-specific tokenizer is used to split the tweets into symbols, phrases and words are called tokens. In the Twitter-specific tokenizer, emoticons are treated as separate tokens. Since the emoticons in twitter data hold some semantic content that describes the user's emotional state. A part of speech tagger is used to label the tokens of twitter data as a common noun, proper noun, pronoun, proper noun + possessive, nominal + possessive, interjection, verb, adverb, adjective, hashtag, emoticon, proper noun + verbal, nominal + verbal and existential 'there' + verbal. The list of labelled tokens is typical to stop word elimination. Finally, the stemming process is carried out by Snowball Stemmer where the words are reduced into stems. It reduces the number of distinct terms in twitter.

### 3.3 Term Filtering and Web News Classification

After modelling the topics of twitter data, a term reduction technique is applied to reduce the number of terms associated with each tweet. The terms are pruned based on their frequencies of occurrence throughout the collection. A relevance value of each term is calculated by the Term Frequency and Inverse Document Frequency (tf-idf) metric. Remove the words from the dataset of news articles and Twitter when tf-idf value of those words is less than an optimized pruning threshold [4]. This process is called term filtering. The efficiency of EFS based web news mining approach is improved by including twitter feature with news article feature for classification of web news articles. Based on the tf-idf values of each term in news articles and twitter data, categorize the news articles based on the following fuzzy rules:

$$Rule_i = IF(A_1 \,\&\, T_1 \sim Prot_1)\, AND\, (A_2 \,\&\, T_2 \sim Prot_2) \ldots AND\, (A_n \,\&\, T_n \sim Prot_n) THEN\, Category = Category_j \tag{2}$$

Equation (2), $i$ denotes the number of rules, $n$ denotes the number of terms in the collection of news articles and tweets, the $A_i$ stores the tf-idf of the term $i$ of the article $A$, the $T_i$ stores the tf-idf of the term $i$ of the tweet $T$, the $Prot_i$ stores the tf-idf value of the term $i$ of one of the prototypes of the corresponding class and $Category_m \in \{set\ of\ different\ categories\}$. The category of new news articles $A_z$ is identified by comparing the article $A_z$ to all the prototypes using the cosine distance and it is given as follows:

$$Class(A_z) = Class(Prot^*) \tag{3}$$

$$Prot^* = Min_{i=1}^{NumP}(cosDist(Prot_i, A_z)) \tag{4}$$

where, $A_z$ denotes the non-categorized web news article to classify, $NumP$ denotes the number of existing prototypes, $Prot_i$ denotes the $i^{th}$ prototype and $cosDist$ denotes the cosine distance between two news articles.

### 3.4 Spatio-Temporal Generalized Additive Model for Web News Classification

The terms in tweets are changing over the time they tweets and their locations and a large number of terms are generated from the Twitter dataset. So a Spatio-Temporal Generalized Additive Model (STGM) is developed which considers both the location of user and time they tweets. The STGM model is built on spatial information of users with the temporal information encoded by dummy variables. This model is given as follows:

$$logit\left(p\left(Tweet_{s_i,t_j}\right)\right) = \sum_{n=1}^{N} f_n\left(Term_{n,s_i,t_j}\right) + \kappa_{s_i,t_j - t^0} \tag{5}$$

Equation (5), $p\left(Tweet_{s_i,t_j}\right)$ is the probability of tweet in the spatial $s_i$ at time $t_j$, $logit(p) = \log\left(\frac{p}{1-p}\right)$ is a logit link function, $N$ denotes the number of terms in a $Tweet$,

$Term_{n,s_i,t_j}$ is the $n^{th}$ term in tweet with location $s_i$ and time $t_j$, $f_n$ is the smooth function of the $n^{th}$ feature and $\kappa_{s_i,t_j-t^0}$ denotes the dummy variable representing the length of the continuous zeros that precede the current observation location $s_i$ and time $t_j$. However, the collection of tweets based on time and location, generate different terms for tweets.

In order to choose the most important terms in the tweets, a Randomized Least Angle Regression (RLAR) is used. RLAR is the extension of Least Angle Regression (LAR) [12] method which is a feature selection model. The RLAR method introduced randomness into the dataset to enhance LAR when the relationship between twitter terms might be nonlinear. RLAR calculates the order of terms going into the regression model and selects the best $m$ number of terms based on criteria like mean squared error. In the RLAR method, the $T$ Twitter Dataset $T^s_{m\times q}$ from $T_{n\times p}$, where $m<n$ ($n$ is the number of terms) and $q<p$ ($p$ is the number of predictors). LAR is applied to each $T^s$ to rank the terms. The term priority for $T^s$ is $rank_s = \langle term_{s1}, term_{s2}, \ldots term_{sq} \rangle$. Finally, term priorities are voted on by $\{rank_s\}$. The overall process of RLAR is given as follows:

**RLAR Algorithm:**

**Input:** Twitter $T = \{term, y\}$

**Output:** Ranked terms $vote_i$

1. for$s = 1$ to $S$
2. sample $D^s$ from $D$
3. Get ranked features $rank_s = \langle term_{s1}, term_{s2}, \ldots term_{sp} \rangle$ by applying LAR on $D^s$, $term_{sp}$ is added in the $i^{th}$ step of LAR algorithm
4. End for
5. for$i = 1$ to $p$
6. for$s = 1$ to $S$
7. if$term_i \in rank_s$
8. $vote_i = vote_i + r_{si}$, where $rank_s[r_{si}] == term_i$
9. $sample.time_i = sample.time_i + 1$
10. End if
11. End for
12. $vote_i = \frac{vote_i}{sample.time_i}$
13. End for

The smaller $vote_i$ is the more important term $term_i$ of tweets. The top most terms are considered for calculation of tf-idf. Based on the calculated tf-idf of twitter data and web news articles, web news articles are categorized by evolving fuzzy rules.

## 4 Result and Discussion

The efficiency of the News Article Categorization using Optimized Evolving Fuzzy System (NAC-OEFS), Twitter enriched News Article Categorization using Optimized Evolving Fuzzy System (TNAC-OEFS) and Spatio Temporal Twitter enriched News Article Categorization using Optimized Evolving Fuzzy System (STTNAC-OEFS) techniques is analyzed in terms of accuracy, precision, and recall. For the experimental purpose, 4000 tweets from the Twitter dataset are collected. The web news article is created from five different sets of data which is briefly explained in [4].

### 4.1 Accuracy

Accuracy is the percentage of web news articles that are predicted with the correct category.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (6)$$

Table 1 tabulates the accuracy of NAC-OEFS, TNAC-OEFS, and STTNAC-OEFS for five different sets of data.

**Table 1.** Accuracy model

| Methods | Dataset | | | | |
|---|---|---|---|---|---|
| | H-Sc | Sc-Te | H-Sc-Sp | B-He-Sc-Sp | A-B-H-Sc-Sp-Tr |
| NAC-OEFS | 80 | 97.27 | 70.45 | 60.7 | 42.45 |
| TNAC-OEFS | 84 | 91.24 | 76.12 | 65.98 | 51.78 |
| STTNAC-OEFS | 90 | 94.56 | 80.87 | 71.32 | 60.85 |

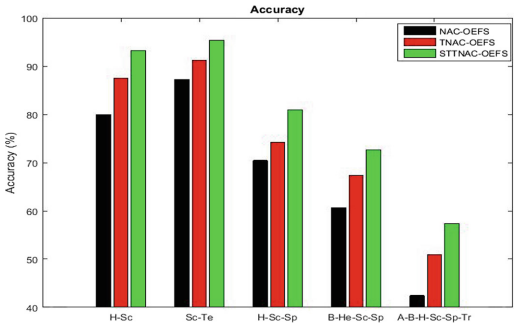


**Fig. 1.** Accuracy comparison model

Figure 1 shows the comparison of accuracy between existing News Article Categorization using Optimized Evolving Fuzzy System (NAC-OEFS), proposed Twitter enriched News Article Categorization using Optimized Evolving Fuzzy System (TNAC-OEFS) and Spatio Temporal Twitter enriched News Article Categorization using Optimized Evolving Fuzzy System (STTNAC-OEFS) methods for five different sets of data. The different sets of data are taken in X-axis and the accuracy value in % is taken in Y-axis. For Health vs. Science set of data, the accuracy of STTNAC-OEFS is 12.5% greater than NAC-OEFS and 7.14% greater than TNAC-OEFS method. By using the spatial information of Twitter users and temporal information of tweets for web news categorization, the accuracy of STTNAC-OEFS method is greater than the other web news categorization methods. From this analysis, it is known that the proposed STTNAC-OEFS has high accuracy than the other methods.

### 4.2 Precision

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \tag{7}$$

Table 2 tabulates the precision of NAC-OEFS, TNAC-OEFS, and STTNAC-OEFS for five different sets of data.

**Table 2.** Precision model

| Methods | Dataset | | | | |
|---|---|---|---|---|---|
| | H-Sc | Sc-Te | H-Sc-Sp | B-He-Sc-Sp | A-B-H-Sc-Sp-Tr |
| NAC-OEFS | 81 | 86.97 | 69.96 | 61.13 | 43.29 |
| TNAC-OEFS | 87.5 | 91.24 | 74.15 | 67.29 | 50.87 |
| STTNAC-OEFS | 93.28 | 95.38 | 80.96 | 72.69 | 57.36 |

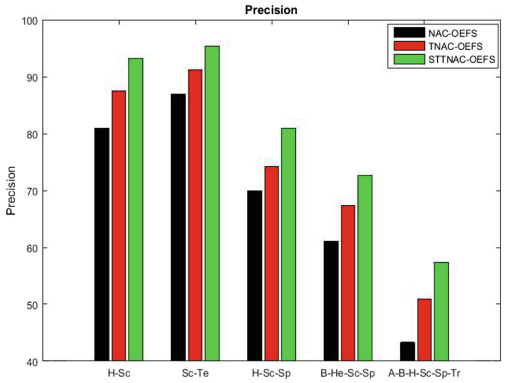


**Fig. 2.** Precision comparison model

Figure 2 shows the comparison of precision between existing NAC-OEFS, proposed TNAC-OEFS and STTNAC-OEFS methods for five different sets of data. The different sets of data are taken in X-axis and the precision value is taken in Y-axis. For Health vs. Science set of data, the precision of STTNAC-OEFS is 15.2% greater than NAC-OEFS and 6.6% greater than TNAC-OEFS method. By using twitter data, spatial and temporal information along with the web news articles, the true positive prediction of STTNAC-OEFS based web news categorization is high which also increase the precision value of STTNAC-OEFS method. From this analysis, it is proved that the proposed STTNAC-OEFS has high precision than the other methods.

### 4.3 Recall

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \tag{8}$$

Table 3 tabulates the recall of NAC-OEFS, TNAC-OEFS, and STTNAC-OEFS for five different sets of data.

**Table 3.** Recall differentiation

| Methods | Dataset | | | | |
|---|---|---|---|---|---|
| | H-Sc | Sc-Te | H-Sc-Sp | B-He-Sc-Sp | A-B-H-Sc-Sp-Tr |
| NAC-OEFS | 83.45 | 85.87 | 73.14 | 64.21 | 45.78 |
| TNAC-OEFS | 87 | 89.21 | 77.62 | 70.13 | 51.27 |
| STTNAC-OEFS | 91 | 94.36 | 82.41 | 76.34 | 57.36 |

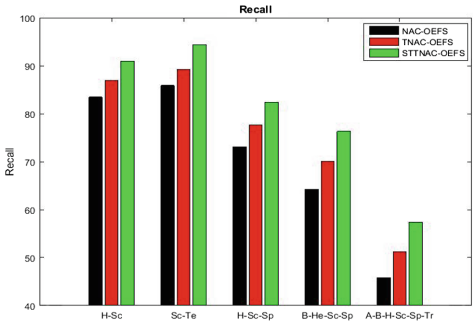


**Fig. 3.** Recall results for different methods

Figure 3 shows the comparison of recall between existing NAC-OEFS, proposed TNAC-OEFS and STTNAC-OEFS methods for five different sets of data. The different sets of data are taken in X-axis and the recall value is taken in Y-axis. For Health vs.

Science set of data, the recall of STTNAC-OEFS is 9.04% greater than NAC-OEFS and 4.6% greater than TNAC-OEFS method. From this analysis, it is known that the proposed STTNAC-OEFS has high recall than the other methods. The recall of STTNAC-OEFS is high because it categorizes the web news articles with the consideration of twitter data, spatial information of Twitter users and temporal information of tweets.

### 4.4 F-Measure

F-measure is the external measure for measuring goodness or accuracy of web news categorization methods. It depends on two factors are precision and recall. It is calculated as,

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

Table 4 tabulates the f-measure of NAC-OEFS, TNAC-OEFS, and STTNAC-OEFS for five different sets of data.

**Table 4.** F-measure analysis

| Methods | Dataset | | | | |
|---|---|---|---|---|---|
| | H-Sc | Sc-Te | H-Sc-Sp | B-He-Sc-Sp | A-B-H-Sc-Sp-Tr |
| NAC-OEFS | 82.21 | 86.42 | 71.51 | 62.63 | 44.5 |
| TNAC-OEFS | 87.25 | 90.21 | 75.85 | 68.68 | 51.07 |
| STTNAC-OEFS | 92.13 | 94.87 | 81.68 | 74.47 | 57.36 |

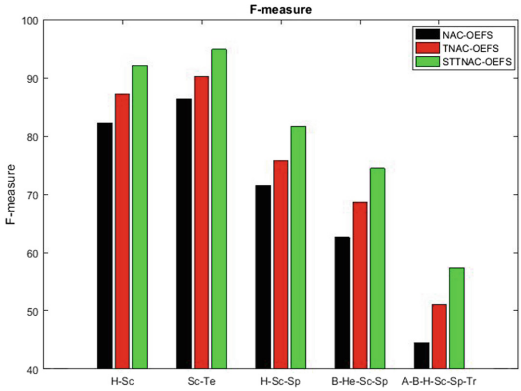


**Fig. 4.** F-measure analysis

Figure 4 shows the comparison of f-measure between existing NAC-OEFS, proposed TNAC-OEFS and STTNAC-OEFS methods for five different sets of data. The different sets of data are taken in X-axis and the f-measure value is taken in Y-axis. For Health vs. Science set of data, the f-measure of STTNAC-OEFS is 12.07% greater than NAC-OEFS and 5.59% greater than TNAC-OEFS method. From this analysis, it is known that the proposed STTNAC-OEFS has high f-measure than the other methods. By using efficient term filtering and efficient web news classification in STTNAC-OEFS, the f-measure of STTNAC-OEFS method is high.

## 5 Conclusion

In this paper, EFS based web news mining is improved by using Twitter data related to news in the web news articles. Initially, Twitter data containing a link to the news are collected along with news articles. The topics within the texts of tweets are modeled by using the LDA model. Then the terms in tweets under different topics are generated by twitter-specific tokenizer, part of speech tagger and Snowball Stemmer. A STGMmodel is developed which consider the location of user and time they tweets and the most important terms are selected by RLAR method. The tf-idf values of selected terms are calculated and it is used along with the tf-idf values of terms in the web news article to create the evolving fuzzy rules. Based on these rules, the web news articles are categorized. Finally, the efficiency of the proposed method has proved in terms of accuracy, precision and recall.

## References

1. Mirończuk, M.M., Protasiewicz, J.: A recent overview of the state-of-the-art elements of text classification. Expert Syst. Appl. **106**, 36–54 (2018). https://doi.org/10.1016/j.eswa.2018.03.058
2. Thangaraj, M., Sivakami, M.: Text classification techniques: a literature review. Interdiscip. J. Inf., Knowl. Manag. **13**, 118–135 (2018) https://doi.org/10.28945/4066
3. Iglesias, J.A., Tiemblo, A., Ledezma, A., Sanchis, A.: Web news mining in an evolving framework. Inf. Fusion **28**, 90–98 (2016). https://doi.org/10.1016/j.inffus.2015.07.004
4. Nithya, D., Sivakumari, S.: Categorizing online news articles using Penguin search optimization algorithm. Int. J. Eng. Technol. **7**, 2265–2268 (2018). https://doi.org/10.14419/ijet.v7i4.15607
5. Bastani, K., Namavari, H., Shaffer, J.: Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. arXiv preprint arXiv:1807.07468 (2018)
6. Kompan, M., Bieliková, M.: News article classification based on a vector representation including words' collocations. In: Third International Conference on Software, Services and Semantic Technologies, S3T 2011, pp. 1–8. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23163-6_1
7. Dilrukshi, I., De Zoysa, K., Caldera, A.: Twitter news classification using SVM. In: IEEE 2013 8th International Conference on Computer Science & Education, pp. 287–291 (2013). https://doi.org/10.1109/iccse.2013.6553926

8. Liparas, D., HaCohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., Kompatsiaris, I.: News articles classification using random forests and weighted multimodal features. In: Information Retrieval Facility Conference. Springer, Cham, pp. 63–75 (2014). https://doi.org/10.1007/978-3-319-12979-2_6
9. Kaur, S., Khiva, N.K.: Online news classification using Deep Learning Technique. Int. Res. J. Eng. Technol. (IRJET) **3**, 558–563 (2016)
10. Demirsoz, O., Ozcan, R.: Classification of news-related tweets. J. Inf. Sci. **43**, 509–524 (2017). https://doi.org/10.1177/0165551516653082
11. Jahanbin, K., Rahmanian, F., Rahmanian, V., Jahromi, A.S., Hojjat-Farsangi, M.: Application of Twitter and web news mining in monitoring and documentation of communicable diseases. J. Int. Transl. Med. **6**, 167–175 (2018). https://doi.org/10.11910/2227-6394.2018.06.04.03
12. Khan, J.A., Van Aelst, S., Zamar, R.H.: Robust linear model selection based on least angle regression. J. Am. Stat. Assoc. **102**, 1289–1299 (2007). https://doi.org/10.1198/016214507000000095
13. Malhotra, S., Dixit, A.: An effective approach for news article summarization. Int. J. Comput. Appl. **76**(16), 5–10 (2013)