EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Effective preprocessing and feature analysis on Twitter data for Fake news detection using RWS algorithm

M. Santhoshkumar*,[1], V. Divya[2]

[1]School of Computing Sciences, VISTAS, Pallavaram, Tamilnadu, India & Department of Computer Science, Sri Venkateshwaraa Arts & Science College, Dharmapuri, Tamilnadu, India
[2]Department of Information Technology, School of Computing Sciences, VISTAS, Pallavaram, Tamilnadu, India

Abstract

The tremendous headway of web empowered gadgets develops the clients dependably strong in virtual redirection affiliations. Individuals from social affairs getting moment notices with respect to news, amusement, training, business, and different themes. The development of artificial intelligence-based classification models plays an optimum role in making deeper analysis of text data. The massive growth of text-based communication impacts the social decisions also. People rely on news and updates coming over in social media and networking groups. Micro blogs such as tweeter, facebooks manipulate the news as faster as possible.

The quality of classification of fake news and real news depends on the processing steps. The proposed articles focused on deriving a significant method for pre-processing the dataset and feature extraction of the unique data. Dataset is considered as the input data for analyzing the presence of fake news. The extraction of unique features from the data is implemented using Bags of relevant tags (BORT) extraction and Bags of relevant meta words (BORMW).

Keywords: Machine learning, Fake news detection, Data analytics, Data science, Feature analysis

Received on 15 November 2023, accepted on 09 February 2024, published on 20 February 2024

doi: 10.4108/eetsis.5157

1. Introduction

The emerging growth of the internet offered an easily accessible and flexible platform for communication. A massive amount of users keep on utilizing the internet and its services for various real-time tasks. In the current scenario, text data is communicated over the internet for conveying information, commenting on a particular event, exposing thoughts, etc. People often utilize microblogs, Facebook, and social networking groups for reading the daily news, understanding current affairs, etc. The versatility of the internet provides numerous users to get involved in accessing micro blogs social networking groups, etc [1]. The spread of rumors on particular events, and news keeps on increasing these days due to vulnerable users involved in the accountability forums. The amount of fake news spread is equally important to consider as similar to that of the real news. The sensitive information may create a major impact on the readers in situations like disasters, natural consequences, social struggles, etc. Text analysis through natural language processor (NLP) through tokenization, stemming, stop word removal and frequency of word analysis are considered as the primary steps in text analysis. The NLP tools [2] automatically clean up the text data before it is opted for classification. The growth of machine learning, and deep learning algorithms impact the steps involved in tuning the text data for classification. Supervised learning algorithms [3] such as linear regression, and the k-nearest neighbour algorithm are utilized for the primary analysis. Unsupervised learning models such as Gaussian mixture models, random forest algorithms, and gradient boost models are utilized to make reliable processing of unbalanced datasets [4]. The detection of multiple language-enabled text analysis is discussed in existing articles, utilizing a multi-modal [5] feature extraction technique. During the COVID [6] pandemic situations, people for anonymous accounts take

*Corresponding author. Email: santhokeyan@gmail.com over the control of tweets on disease and spread more fake news on the internet. The author from the existing article discussed the COVID-19 impact by utilizing the keyword called #Covid. The news is posted with more frequently viewed tags. These tags are accessed by millions of people from the internet platform. When considering the global text analysis methods, when coming to social networking groups, forums analysis, micro blogs analysis Hashtags plats an optimum role.

- ✓ The proposed system discusses the feature extraction method that includes the benchmark model of Bags of word (BOW) approach as elaborated technique named as BORT and BORMW technique etc.
- ✓ The method involved recursively searching the relevant tags from the global database. These tags include millions of viewers globally.
- ✓ The proposed technique focused on analysing the existing PHEME dataset and reducing the features into 22 features from 54 feature attributes or columns.
- ✓ These 22 attributes are classified with respect to datatype as well as the complete format tested. The source of the data posted is also considered as important feature.

2. Background Study

Wei et al, (2022) [7] have discussed fake news detection through modality and event adversarial networks considering the modality invariant features from image data and text data. A decoder is developed to extract the discriminant variables from the datasets. A dual discriminator is utilized here. Reduction of information loss is considered an important criterion for making preprocessing which is achieved here.
Rong et al, 2022) [8] proposed on dense text localization network to extract the intermediate unique context present in the data. The presented approach continuously scales the data effectively from long text and clusters the text patterns. The quantitative evaluations in the given articles ensure the retrieval process for better results.
Yang et al, (2022) [9] a system where multiple perspective-enabled feature extraction technique is focused. The discriminative feature is analyzed with multiple text datasets such as ICDAR2015, CTW1500, and MSRA-TD500 using a concentric mask network. The major challenge in the presented work is the limitation in features extracted from the dataset. Each data is unique. Commanding on feature extraction process that deep digs the feature ensures a better quality of classification.
Abonizio et al, (2022) [10] a sentiment analysis framework using NLP. The presented system considers back translation, pre-trained text augmented using a random forest algorithm. The major problem of an unbalanced dataset reduces the quality of the classification process. The presented approach produces a sparse solution to handle the limitations of labels. Further testing the relevant system with multiple datasets is helpful to comparatively validate the algorithm which seems a drawback here.
Li et al, (2023) [11] presented article deals with internet fraudulent data analysis through a graph learning approach. The input data is collected from various financial statements, and transaction hints, where the structuring of text is difficult. Here structure to vector *Struc2Vec model is considered. Compared with traditional approaches, the proposed model through the graph learning method improves the accuracy, precision, and F1-score of the prediction process.

Various existing articles are considered for analysing the implementations done so far on unique feature extraction from the text data. From the background study, the following are considered as the critical challenges in making text analysis.
The limitations in the text data collected from massive platforms, micro blogs which are not labelled, provides uncertain results.
The data cleaning process of an unbalanced dataset needs more complex structures. The feature extraction should meet the expected criteria of the classification model.
The removal of sparse data during multiple-feature extraction is important. The clustering of massive convergence data has certain drawbacks on missing out on data due to heterogeneity.
Considering these constraints, the presented system created a robust method for feature extraction through multiple channels of feasibility. The feasible methods enhance the nature of the feature extraction process and further improve the quality of the prediction process.

3. Dataset descriptions

PHEME Dataset

The PHEME dataset is a massive collection of tweeter data in terms of rumors on breaking news. The dataset contains 60000+ informative rows on events such as German wing crah and Charlie Hebdo. The dataset contains a totally of 9 events-related information containing rumours and non-rumors. The features of the dataset are extracted from the meta keywords in the source of the dataset. The peak value on repeated keywords, meta tags are helpful to analyze the text data better.

Table 1 PHEME sample dataset

Attributes	**Values**	**Type**
contributors	Empty	Integer
truncated		Integer
text	Charlie Hebdo became well known for publishing the Muhammed cartoons two years ago	String
in_reply_to_status_id		Integer

id	5.52785E+17	Integer
favorite_count	41	Integer
source	<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>	String
retweeted	FALSE	String
entities	1	Integer
symbols	#	String
user_mentions	21	Integer
hashtags	#breakingnews	String
urls	"http///	String
id_str	15464d564	Integer
retweet_count	202	Integer

Table 1 exhibits the PHEME sample dataset contents with 22 attributes are available with unstructured manner. This dataset is cleaned up and opted for a feature extraction process. The important attributes considered here are hashtags, retweet counts, source of the input, reply_id_counts, text data, etc. From the raw text data RWS algorithm is implemented. The preprocessing stage removes the unwanted attributes as well as cleans up the text data [12]. NLP steps are utilized to formulate the attribute. Hashtags are considered a crucial feature for relevant bags of word extraction. Table 2 enumerates the various types of attributes present inside the PHEME dataset.

Table 2 Attribute type

Data Type	Count
Hexadecimal	10
Integer	13
String	18
Float	1
Logical	12

4. Preprocessing and Feature extraction

The fake news detection system gathers the data from various articles that are labeled and unlabeled in nature. The data is fetched into the preprocessing stage. Preprocessing of text data includes stemming [13], tokenization [14], stop word removal and transforming the data into readable numeric values [15] to make classification better. Other feature extraction techniques include the BoW approach, metadata analysis, etc.

Sample data:
Charlie Hebdo became well-known for publishing the Muhammed cartoons two years ago.

The proposed framework examines the component extraction strategy that incorporates the benchmark model of the BoW approach as elaborated procedure named as Sacks of BORT and Sacks of BORMW method and so forth. The strategy engaged with recursively searching the important tags from the worldwide data set. These labels incorporate a large number of watchers internationally. The proposed method zeroed in on breaking down the current PHEME dataset [16] and diminished the highlights into 22 features from 54 component ascribes or segments [17].
These 22 attributes are grouped regarding datatype as well as the total configuration is tested. The wellspring of the information posted is likewise viewed as a significant component.

Algorithm : RWS
Input : Raw PHEME dataset text
Output: features f(n)

```
Start
Input_data= PHEME_raw_data
pp=Stemming(Input_data);
pp2=Tokenization(pp);
Pp3=stop_word_removal(pp2);
x=pp3;
For ii=1:n(num_of_search_nodes)
For  jj=1:eop(end_of_paragraph)
      A(ii,jj)=weight(x(n))
If A(ii,jj)==BORW(ii,jj)
      Count=count+1
End
If A(ii,jj)== BORMW(ii,jj)
Meta_count=meta_count+1;
end
End
Fea_1=Count
Fea_2=Meta_count
End
```

Summary of algorithm
This algorithm appears to aim at analyzing the input PHEME dataset and generating features based on word weights and metadata counts. Below is a breakdown of your provided pseudo-code:

Input:
Raw PHEME dataset text: The input text data from the PHEME dataset.
Preprocessing:
pp = Stemming(Input_data): Perform stemming on the input data, reducing words to their base or root form.
pp2 = Tokenization(pp): Tokenize the stemmed data into individual words.
pp3 = stop_word_removal(pp2): Remove stop words (commonly used words that carry little meaning) from the tokenized data.
x = pp3: Set the variable x to the processed tokenized data.
Loop Over Nodes and Paragraphs:

For ii = 1:n (num_of_search_nodes): Loop over each search node.
For jj = 1:eop (end_of_paragraph): Loop over each paragraph within the node.
A(ii, jj) = weight(x(n)): Calculate the weight of the words in paragraph jj of node ii using a weight function on the processed text x.

Comparison and Counting:
If A(ii, jj) == BORW(ii, jj): Compare the calculated weight A with a predefined baseline weight BORW for the same paragraph and node. If they match:
Increment Count by 1.
Meta Count:
If A(ii, jj) == BORMW(ii, jj): Compare the calculated weight A with a predefined baseline meta weight BORMW for the same paragraph and node. If they match:
Increment Meta_count by 1.
Feature Extraction:
Fea_1 = Count: Set Fea_1 as the value of the Count variable, representing the count of matching weights.
Fea_2 = Meta_count: Set Fea_2 as the value of the Meta_count variable, representing the count of matching meta weights.
Output:
features f(n):
Fea_1: Feature representing the count of matching weights.
Fea_2: Feature representing the count of matching meta weights.
It seems that the algorithm aims to generate two features, Fea_1 and Fea_2, which are counts of matching word weights and matching meta weights, respectively, based on a predefined baseline. The algorithm involves comparing the calculated weights with predefined baseline values to determine the feature values.

5. Results and discussions

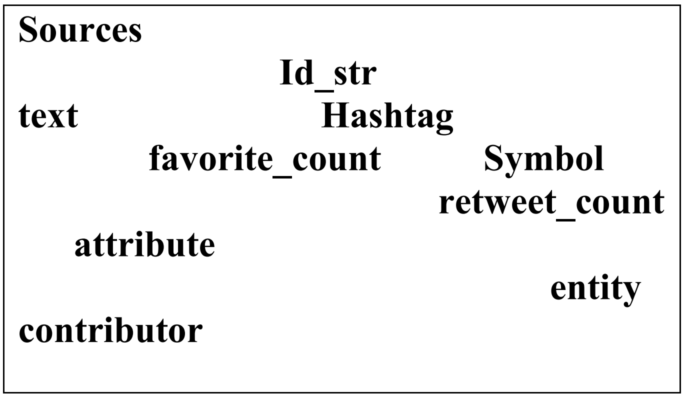

Figure 1 Word cloud of PHEME attribute

Figure 1 illustrates the word cloud of PHEME dataset after the preprocess stage. These attributes contain relevant data which is identified based on the data type mapping.

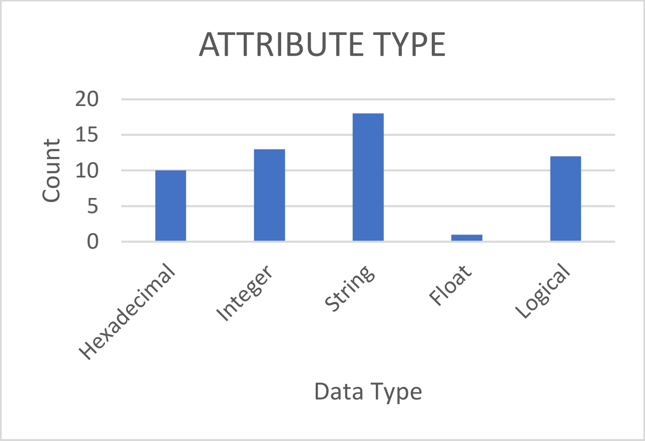

Figure 2 Attribute Types

Figure 2 depicts the attribute data type included in the PHEME dataset. The dataset is a kind of JSON file which is directly accessed by the software platform or opened using the Excel sheet.

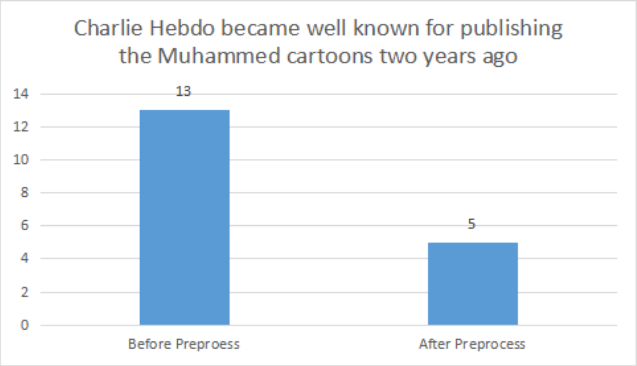

Figure 3 Pre-processed Data

Figure 3 exposes the sample text data before preprocess and after preprocess. The primary goal of preprocessing is to remove unwanted words, organize the required words, removal of stop words, etc.

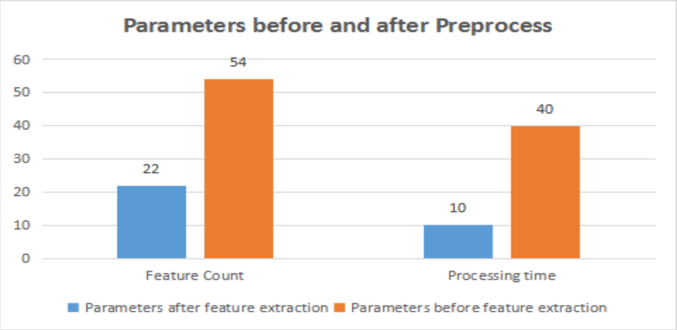

Figure 4 Feature Parameters before and after Preprocess

.

Figure 4 shows the feature parameters after and before the preprocess. The total features are 54 in number, the unstructured data opts for feature extraction process consumes 40 seconds approximately as processing time (t), the feature count or the attribute after the cleaning process removes the unfilled columns, empty spaces, unbalanced data hence the final consideration is 22 features. From the 22 features the unique word cloud is shown in Figure 1 the text data alone opted for the feature analysis process using the RWS algorithm.

When comparing the whole dataset, the feature extraction process using BORT, and BORMW process in the RWS model compares the extracted unique words and compares them with the global tags and words. Further, the cleaned data is utilized for the classification process using the novel machine learning model. The performance of the system is evaluated through accuracy and precision.

6. Conclusion

The colossal progress of web-engaged contraptions fosters the client's reliably solid virtual redirection affiliations. People from get-togethers get a second notification concerning news, entertainment, preparation, business, and various topics. The gigantic development of text-based correspondence influences social choices too. Individuals depend on news and updates coming over in web-based entertainment and systems administration gatherings. Miniature sites like Twitter, and Facebook control the news as quickly as could be expected. The nature of the characterization of phony news and genuine news relies upon the handling steps. The proposed articles zeroed in on determining a huge technique for pre-handling the dataset and highlighting the extraction of extraordinary information. PHEME tweeter dataset is considered as the information for investigating the presence of phony news. The extraction of remarkable highlights from the information is carried out utilizing BORT extraction and BORMW. Important weight search RWS calculation is executed here that incorporates the three-cycle preprocess, BORT, BORMW, and so forth.

7 References

[1] De Oliveira, N, R, Medeiros D, S, V, and Mattos, D, M, F.: A Sensitive Stylistic Approach to Identify Fake News on Social Networking, IEEE Signal Processing Letters. 2020; 27:1250-1254.

[2] Park, M, Chai, S.: Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques, IEEE Access. 2023; 11: 71517-71527.

[3] Radhika S, Prasanth A, An Effective Speech Emotion Recognition Model for Multi-Regional Languages Using Threshold-based Feature Selection Algorithm. Circuits, Systems, and Signal Processing. 2023; 1-22.

[4] Boualouache, A and Engel, T.: A Survey on Machine Learning-Based Misbehavior Detection Systems for 5G and Beyond Vehicular Networks, IEEE Communications Surveys & Tutorials. 2023; 25:1128-1172.

[5] Albalawi, R, M, Jamal, A, T, Khadidos, A, O and Alhothali, A, M.: Multimodal Arabic Rumors Detection, IEEE Access. 2023; 11:9716-9730.

[6] Elhadad, M, K, Li, K, F and Gebali, F.: "Detecting Misleading Information on COVID-19," in IEEE Access. 2020; 8:165201-165215.

[7] Wei, P, Wu, F, Sun, Y, Zhou, H and Jing, X, Y.: Modality and Event Adversarial Networks for Multi-Modal Fake News Detection, IEEE Signal Processing Letters. 2022; 29: 1382-1386.

[8] Rong, X, Yi, C and Tian, Y.: Unambiguous Text Localization, Retrieval, and Recognition for Cluttered Scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022; 44(3):1638-1652.

[9] Yang, C, Chen, M, Xiong, Z, Yuan, Y and Wang, Q.: CM-Net: Concentric Mask Based Arbitrary-Shaped Text Detection, IEEE Transactions on Image Processing. 2022; 31:2864-2877.

[10] Abonizio, H, Q, Paraiso, E, C and S. Barbon, S.: Toward Text Data Augmentation for Sentiment Analysis, in IEEE Transactions on Artificial Intelligence. Oct 2022; 3(5): 657-668.

[11] Li, R, Liu, Z, Ma, Y, Yang, D and Sun, S.: Internet Financial Fraud Detection Based on Graph Learning, in IEEE Transactions on Computational Social Systems. June 2023; 3:1394-1401.

[12] Alsuwaiket, M, A.: Feature Extraction of EEG Signals for Seizure Detection Using Machine Learning Algorthims, Engineering, Technology & Applied Science Research. Oct 2022; 12(5):9247–9251.

[13] Jayachitra, S, Prasanth, A, Rafi, A Hierarchical-Based Binary Moth Flame Optimization for Feature Extraction in Biomedical Application. Proceedings in 4th International Conference on Machine Learning, Image Processing, Network Security and Data Sciences. 2023; 27-38.

[14] Anwer, M, Khan, S, M, Farooq, M, U and Waseemullah.: Attack Detection in IoT using Machine Learning, Engineering, Technology & Applied Science Research, Jun. 2021; 11(3):7273–7278.

[15] Aramaki, Y, Matsui, Y, Yamasaki, T and Aizawa, K.: Text detection in manga by combining connected-component-based and region-based classifications, IEEE International Conference on Image Processing. 2016; 2901-2905.

[16] Mol, J, Mohammed, A and Mahesh, B, S.: Text recognition using poisson filtering and edge enhanced maximally stable extremal regions, International Conference on Intelligent Computing, Instrumentation and Control Technologies. 2017; 302-306.

[17] Satwashil, K, S and Pawar, V, R.: English text localization and recognition from natural scene image, 2017 International Conference on Intelligent Computing and Control Systems. 2017; 555-559.