A Natural Language Processing Technique to Identify Exaggerated News Titles

Tshephisho Joseph Sefara[0000−0002−5197−7802] and Mapitsi Roseline Rangata[0000−0002−7624−2415]

Council for Scientific and Industrial Research, Pretoria, South Africa
{tsefara,mrangata}@csir.co.za

Abstract. Exaggerated news titles are used to deceive news readers and spread misleading information. This paper presents a new natural language processing (NLP) technique that identifies exaggerated news titles. The technique uses Jaccard similarity as a pre-processing step to filter out unrelated articles. The technique then applies text summarisation on the content of the news article to create a new title. Lastly, the technique applies cosine similarity to compare similar articles between the article title and the newly generated titles. The output is the classification of the news articles using the output of cosine similarity. This technique performed well in major South African news articles.

Keywords: Text similarity, text summarisation, natural language processing, cosine similarity, Jaccard similarity.

1 Introduction

With a rapid increase in online migration from traditional physical newspapers. Exaggerated news titles are sometimes used to deceive readers and spread misleading information. Many online news sites have appeared, providing great convenience and information to many readers and attracting many readers. While traffic attraction remains a big business opportunity for organisations, readers are being deceived and shown more advertisements. The identification of exaggerated titles remains an open investigation, while the detection of fake news is actively investigated [14]. Using NLP methods and machine learning algorithms, such titles can be classified on a given labelled dataset. Currently, there are no open-source datasets that contain exaggerated news articles. This paper proposed a technique that can label such a dataset using the combination of text summarisation [9] and text similarity [13].

Text similarity measurement is a natural language processing (NLP) task that is used primarily in information extraction, automatic question answering, machine translation, text classification, document matching, text summarisation, and others[13]. Text similarity is used to identify similar documents using the same content to measure similarity scores. Text similarity is divided into two aspects: Lexical similarity and semantic similarity. Lexical similarity is measured

mainly by string-based algorithms to measure the similarity between words. Semantic similarity is measured using a corpus-based algorithm [3].

One of the simplest ways to measure whether the title of the article is related to the content is to apply text similarity between the two. While the length of the content is longer than the title, performance will be negatively affected. Applying text summarisation to longer articles can help summarise shorter titles [9].

Text summarisation is an NLP task that deals with shortening longer articles into shorter sentences while keeping the meaning of the original article. Automatic text summarisation methods are needed to overcome the increasing amount of textual data that are accessible on the internet. This can help discover and consume the appropriate data. Text summarisation can be approached using two methods, namely extraction and abstraction. Extractive methods choose a subset of the sentence of the original article, while abstractive methods create a semantic representation and use language generation methods to create a summary [4].

This paper proposes a new unique method that uses NLP tasks such as text similarity and text summarisation to detect exaggerated news titles.

The main contributions of this paper are as follows.

- This paper proposes a simple, lightweight, and unique method for identifying exaggerated news titles.
- We release the source code to GitHub[1] to maintain further development of the proposed method and to help reproduce the results for comparisons.
- The method can be used to identify fake news and disinformation in news.
- This method can be used to label a new dataset of exaggerated articles.

This paper is organised as follows. The next section discusses the literature review. Section 3 discusses the approach taken to implement the proposed architecture. Sections 4 contain the results and analysis. Sections 5 conclude the paper with future work.

2 Literature Review

Sepúlveda-Torres et al. [11] propose a tool known as HeadlineStanceChecker that uses Natural Language Processing methods to detect disinformation in the headline, a method that will classify the news article with respect to the body text and headline. This method was developed to help identify misleading news articles. The first created a summary of the body of the news article that reveals important information in the article using text summarisation. Second, they configure the relatedness of the headline with respect to its body text using text similarity methods, such as cosine similarity and overlap coefficient (to measure how similar the headline and body summary are). RoBERTa was used to classify the headline in relation to its text body in 4 classes:

1. Agree: The text of the body coincides with the title

[1] `https://github.com/JosephSefara/exaggerated-news-titles`

2. Disagree: The text of the body contradicts the title of the article
3. Discuss: The body text discusses the same subject as the title
4. Unrelated: The text of the body is not related to the title.

The method was developed to identify the best news articles from various sources on United States news sites. They first created a headline and link extractor analyser on selective websites to examine ten news sites based in the United States for three months. To measure the similarity between different news sites, they recovered a three-month archived website. Cosine similarity measurement was applied to the collection of articles. The results of their similarity calculations show that the number of articles used in the calculation directly affects the overall similarity score. They can effectively identify events such as presidential elections and national holidays using similarity points. They also observed that political events had the greatest influence on the similarity score. After an eventful day, the similarity score decreases, which shows that the news site is following other stories[1].

Wang and Dong [13] conducted a study of text similarity to provide an overview of the development of text similarity measures. Their research focusses on two elements of similarity measure, text distance, and text representation. Text distance is classified into three types: length distance, distribution distance, and semantic distance. Text representation is classified as string-based, corpus-based, and others. Their findings indicate a link between text distance and text representation and that text representation provides a solid basis for the calculation of text similarity. Online news scraping was conducted to assess the semantic similarities between two distinct languages, Hindu and English, of identical news articles. The data sets used are collected from Google News and Google Translator, which is used to translate Hindi into English. The purpose of their study is to introduce three similarity measures to calculate the similarity of two articles in identical news contexts in two distinct languages. In their study, the term frequency-inverse document frequency (TF-IDF) and bag-of-words are used to generate features from articles. The generated features are used to measure the cosine similarity, Jaccard similarity, and Euclidean distance between the two articles. All three methods produced good similarity results, but Combined Cosine similarity using TF-IDF as a feature generator is the leading performer with higher accuracy, recall, and F measurement [12].

Jang et al. [8] proposed a method to model fake news analysis by identifying various features in Twitter data, such as Uniform Resource Locators (URLs), responses, followers, and tweets. They first used Kaggle data to perform statistical analyses to search for the main features of fake news. These are the main identified features of their analysis: URLs, replies, followers, tweets, and quote retweets. Analysis also indicates that false news spreads longer than real news. They conducted a Neural Network fake news classification to compare three models using different features. Model 1 used the main feature proposed by [2], Model 2 used all the features of the tweets, and Model 3 used only the main feature identified in their analysis. From each classification model, 405 fake news cases and 2085 real news cases were evaluated. The results of Model 3 indicate

that the main features reflect the attributes of false news well and that the effect of improving the classification accuracy of false news is produced by slightly reducing the classification accuracy of real news. Consequently, the use of the main features will contribute to improving the classification accuracy, compared to the use of all features of Twitter data if the focus is more on detecting false news.

Islam et al. [5] have developed a Bengali headline analysis using several machine learning algorithms. They determine the sentiment of the headline of the news from 0 to 1, where 0 is negative and 1 is positive news. For analysis, data were collected from Bengali newspapers, containing 1,019 data, 1,009 positive headlines, and 510 negative headlines with minimum and maximum words of 1 and 14. To classify Bengali news headlines by sentiment, they trained and tested several supervised machine learning algorithms on the collected data, including the Support Vector Machine (SVM), Naive Bayes, Decision Tree, Random Forest, and K-Nearest Neighbour (KNN), to find the best algorithm for headline classification. In their findings, Naive Bayes and SVM were more accurate, with 75% of SVM, 73% of Nave Bayes, 69% of Random Forests, 68% of KNN, and 60% of Decision Trees. Also, for a specific headline, SVM and naive Bayes and Decision Tree provide real predictions, whereas KNN and Random Forest provide incorrect predictions [15]. They also used the naive Bayes classification and TF-IDF and the count vectorizer to detect fake news and used Kaggle data sets to train naive Bayes multinomials to detect fake news. Their results showed that the model evaluation of the dataset is higher in TF-IDF, with 94% accuracy in detecting fake news and 93% accuracy in the count vectorizer. This shows that naive Bayes and TF-IDF detect fake news well together. Labelled data enable the use of machine learning to make predictions. The data in this work do not contain labels that indicate which article is exaggerated, hence, machine learning models were not explored. This paper proposes a technique to identify or classify exaggerated news titles using NLP methods.

Islam et al. [6] propose a Bangla text classification analysis to detect YouTube videos with exaggerated titles. Authors trained and tested the collected data on six classification models known as Decision Tree, Random Forest, Logistic Regression, Neural Networks, and Convolutional Neural Networks (CNN). Their results showed that CNN performed better among all trained classification models in classifying exaggerated YouTube video titles.

3 Methodology

This section discusses the architecture, acquired data, and the methods used to implement the proposed technique.

This paper proposes the architecture illustrated in Fig. 1. The architecture requires structured data in the form of news titles and their content. The initial step is to filter out unrelated articles by applying the Jaccard similarity method. Articles with Jaccard similarity of zero are removed from the dataset and will not be used in the experiments. Articles with Jaccard similarity of more than

zero are kept and will be used in the experiments. The second step is to apply the text summarisation method to news content to produce shorter titles that contain one sentence. The third step is to apply cosine similarity to the title of the article and to the new titles generated by the text summarisation. This step will produce a cosine similarity score for each article. Using a threshold of zero, when an article contains zero similarity is labelled as exaggerated and, otherwise, it is a normal article.

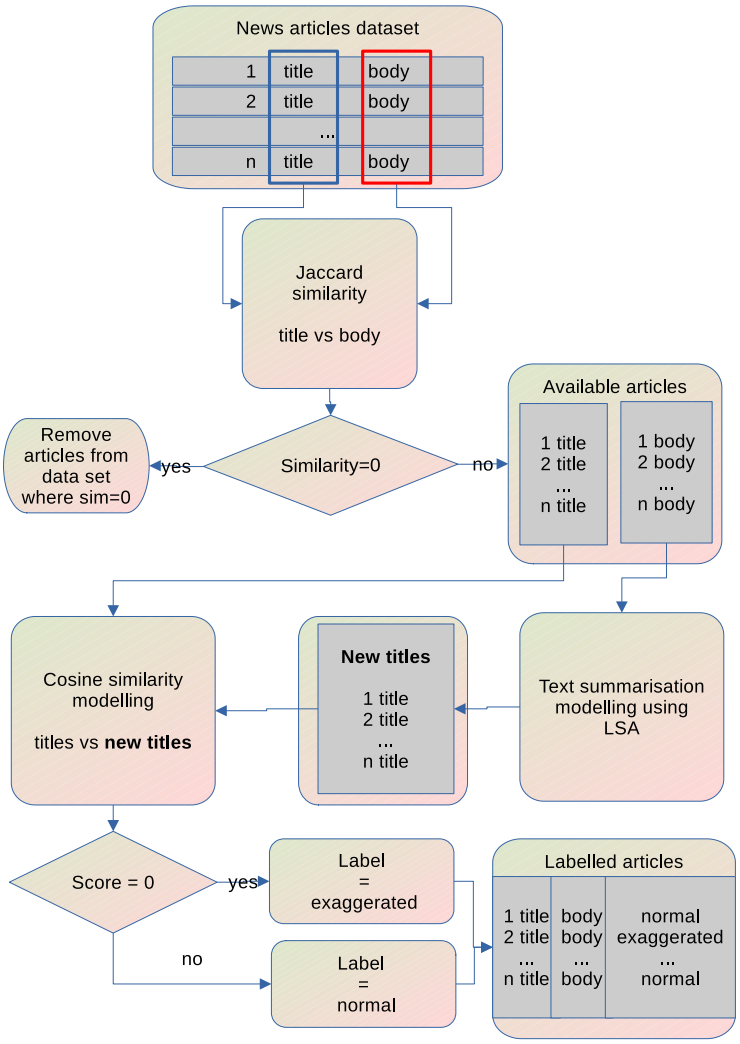

Fig. 1. Architecture of the proposed method.

3.1 Data

The data contain newspaper articles published on the internet from 13 January 2021 to 24 March 2023 by the main publishers in South Africa shown in Table 1. The data were acquired by subscribing to the RSS feed of each publisher from 7 November 2022 up to 24 March 2023 and the data show that some articles have been updated, resulting with date before the subscription, and some articles are missing the date. The articles contained metadata that included the following:

- Title: the subject or topic of the news article.
- Publisher: the name of the publisher.
- Category: the category or class of the article.
- Link: the URL of the news article.
- Access date: the date the article was acquired.
- Published date: the date the article was published on the internet.
- Body: the content of the news article.

The dataset contained missing values which will affect the accuracy on the proposed technique. Hence, the dataset was pre-processed and filtered to reduce noise and evaluate the quality of the articles. The following steps were taken during data filtering:

- Step 1: Articles that have no content in the title or body should be removed from the dataset.
- Step 2: Articles that are dissimilar can be removed by applying the Jaccard similarity in the title and body of the article. This step also helps to remove fake news from the dataset.

Table 1. News articles after pre-processing

	Publisher	Total
1	The South African	11122
2	The Citizen	9033
3	News24	7203
4	Daily Maverick	5812
5	SABC	4632
6	EWN	4137
7	MyBroadband	2055
8	BusinessTech	1842
9	Mail & Gaurdian	1699
10	TechCentral	1018
11	Times Live	821
12	Sowetan Live	397
13	IOL	1
14	Total	49772

A total of 49772 articles were available after pre-processing and filtering. The news articles contain news categories that are shown in Table 2. The *general news* category has more articles than other categories, followed by the business category.

Table 2. Categories of news articles

Category	Total
News	35948
Business	4264
Sports	4711
Technology	3549
Politics	1300
Total	49772

3.2 Methods

This section discusses the approach taken to implement the proposed architecture. The following NLP tasks are used to build the proposed architecture.

- **Jaccard similarity** is a method of measuring similarity and diversity between two data points [7]. The Jaccard similarity index defined in Equation (1) is used to determine the similarity between the title of the article and the body of the article. This method helps to filter the data set to remove unnecessary articles, such as advertisements. The filtering rule is that at least one word in the title of the article must exist in the body of the article. The Jaccard similarity index states that:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

  where A represents the title of the article and B represents the body of the same article. $|A \cap B|$ is the set of the common words from A and B, and $|A \cup B|$ is the set of all the words from A and B.
- **Text summarisation** is an NLP task that reduces a long article to fewer sentences without losing the meaning of the original article. This paper adopts the text summarisation technique discussed in [9] using latent semantic analysis (LSA). This method uses extractive method to choose a subset of the sentence of the original article based on the importance. This technique is used to shorten the body of the article into a summarised title that will be used as input for cosine similarity.
- **Cosine similarity** is an NLP task that measures the similarity between documents. This paper uses cosine similarity to measure the similarity between the title of the article and the summarised title. Exaggerated news

titles are identified by the lower similarity score that is determined using equation (2) which is derived from the formula of the Euclidean dot product. Cosine similarity is the cosine of the angle between vectorised titles or the scalar product of the vectorised titles divided by the product of their magnitude. Cosine similarity of title A and B is defined as:

$$Sim(A, B) = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{2}$$

where A represents vectorised titles and B represents vectorised new titles generated from text summarisation. $A \cdot B$ represents the scalar product of A and B, and $\|A\|\|B\|$ represents the product of the magnitude of A and B. Both A and B were transformed using the TFIDF vectorizer[2] from the scikit-learn Python library [10]. The similarities between A and B for the whole dataset are discussed in the next section.

3.3 Experiment environment

The proposed technique was applied to the given news articles. The Jaccard similarity method was quick to run on the data. The data were then passed to the text summarisation, which focusses on the content. This method took 10 minutes to complete. The text summarisation uses the LSA technique. The cosine similarity then took 5 minutes to complete. The experiment was conducted on a machine running Ubuntu operating system with an i7 processor and 16 Gig of memory.

4 Results and Discussions

This section discusses the results obtained from the proposed technique. The technique classified news articles into two categories, namely *exaggerated* and *normal*. Fig. 2 shows the classification of news articles, with 33% being exaggerated news articles and 67% being normal articles. Based on the analysis, we find that most published articles have been classified into the *normal* category. Fig. 3 shows that the *news* category has more than 10,000 news articles classified as exaggerated. The *sports* and *politics* category has fewer titles classified as exaggerated, this is because these categories have fewer readers.

We can also observe that *The South African*, *The Citizen*, and *Daily Maverick* are the top publishers with the highest news articles classified as exaggerated in Fig. 4. Fig. 5 focusses on the news articles per publisher in the *news* category and shows that the three publishers: *The South African*, *The Citizen*, and *Daily Maverick* are classified as the highest publishers with exaggerated news articles among the other publishers compared to the *normal* category with a slight difference from *Daily Maverick* publisher in the *exaggerated* and *normal* category.

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`

These results show that South African news articles have a level of exaggeration based on the proposed technique.

From this analysis, we can conclude that the proposed method was able to detect exaggerated and non-exaggerated news and that most of the news articles published by various publishers were classified into the *normal* category.

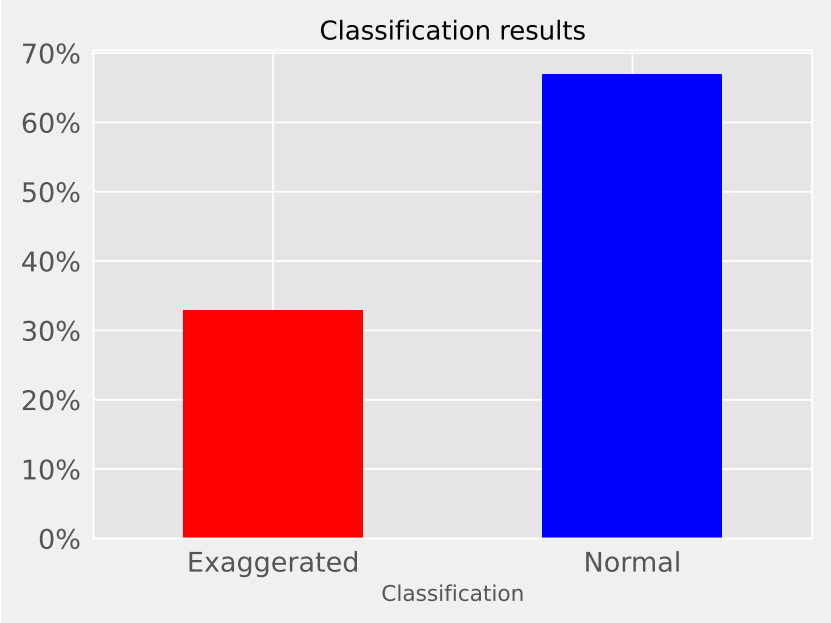

Fig. 2. Classification of news articles

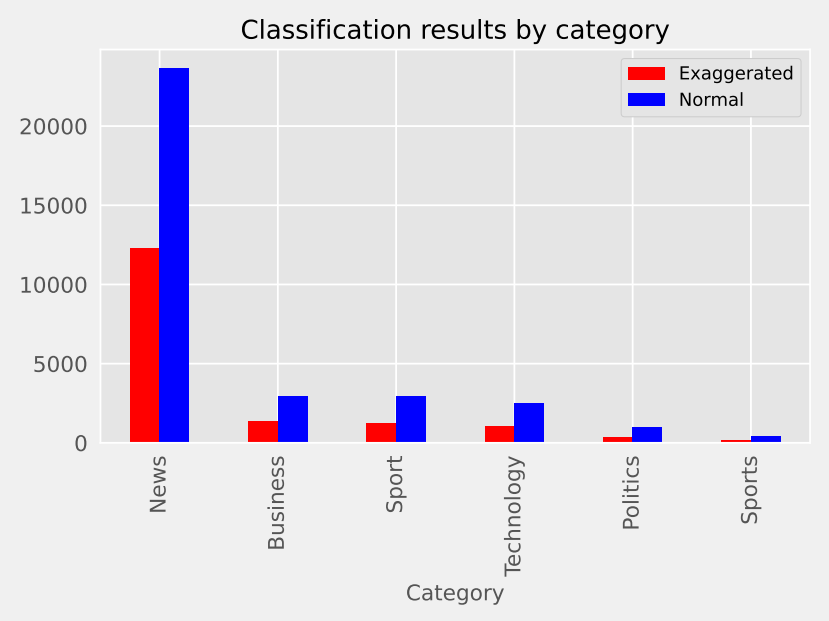

Fig. 3. Classification of news articles by category.

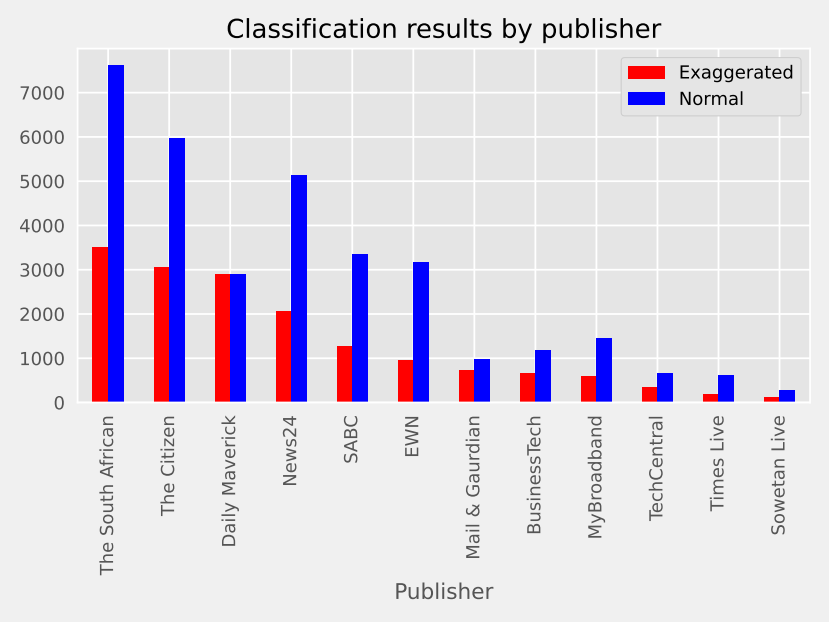

Fig. 4. Classification of news articles by publisher.

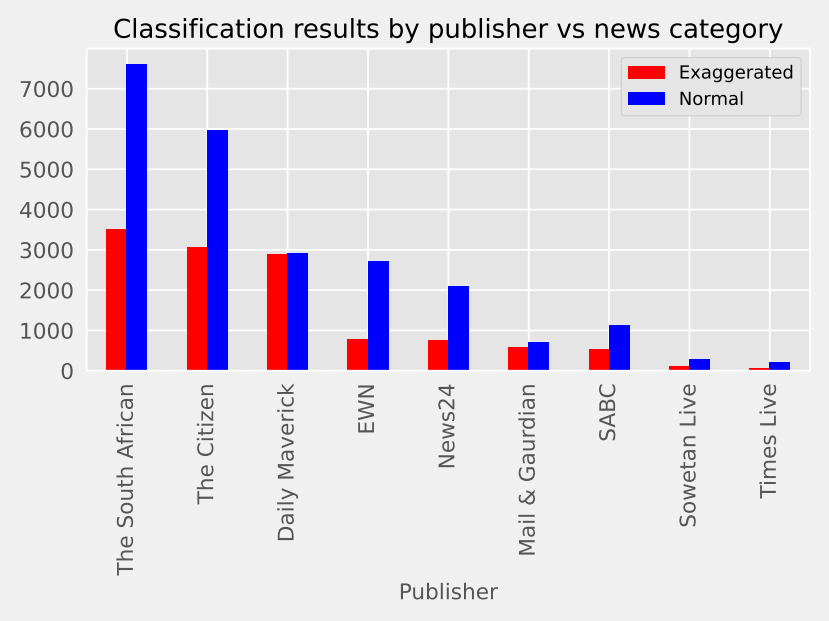

Fig. 5. Classification of news articles by publisher and *news* category.

5 Conclusions and Future work

The paper proposed the new NLP method to identify exaggerated news titles. The method can also identify fake news at the basic level. The literature was reviewed. The data was discussed and further filtered out unrelated news articles. The proposed method uses famous NLP tasks such as Jaccard similarity, text summarisation, and cosine similarity. The technique was evaluated on the acquired data set.

The future work will improve the method by including neural networks to compute the similarity and using advanced transformer models to implement the text summarisation. Furthermore, future work will find the labelled data set of exaggerated news articles to evaluate the performance of the model.

References

1. Atkins, G.C., Nwala, A., Weigle, M.C., Nelson, M.L.: Measuring news similarity across ten US news sites. arXiv preprint arXiv:1806.09082 (2018)
2. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: Proceedings of the 20th international conference on World wide web. pp. 675–684 (2011)
3. Gomaa, W.H., Fahmy, A.A., et al.: A survey of text similarity approaches. international journal of Computer Applications **68**(13), 13–18 (2013)

4. Gudivada, V.N.: Chapter 12 - natural language core tasks and applications. In: Gudivada, V.N., Rao, C. (eds.) Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications, Handbook of Statistics, vol. 38, pp. 403–428. Elsevier (2018). https://doi.org/https://doi.org/10.1016/bs.host.2018.07.010
5. Islam, M.M., Masum, A.K.M., Rabbani, M.G., Zannat, R., Rahman, M.: Performance measurement of multiple supervised learning algorithms for Bengali news headline sentiment classification. In: 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART). pp. 235–239. IEEE (2019)
6. Islam, M., Ria, N.J., Masum, A.K.M., Ani, J.F.: Performance comparison of multiple supervised learning algorithms for Youtube exaggerated Bangla titles classification. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). pp. 1–6. IEEE (2021)
7. Jaccard, P.: Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. Bull Soc Vaudoise Sci Nat **37**, 241–272 (1901)
8. Jang, Y., Park, C.H., Seo, Y.S.: Fake news analysis modeling using quote retweet. Electronics **8**(12), 1377 (2019)
9. Ozsoy, M.G., Alpaslan, F.N., Cicekli, I.: Text summarization using latent semantic analysis. Journal of Information Science **37**(4), 405–417 (2011). https://doi.org/10.1177/0165551511408848
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. the Journal of machine Learning research **12**, 2825–2830 (2011)
11. Sepúlveda-Torres, R., Vicente, M., Saquete, E., Lloret, E., Palomar, M.: HeadlineStanceChecker: Exploiting summarization to detect headline disinformation. Journal of Web Semantics **71**, 100660 (2021)
12. Singh, R., Singh, S.: Text similarity measures in news articles by vector space model using NLP. Journal of The Institution of Engineers (India): Series B **102**, 329–338 (2021)
13. Wang, J., Dong, Y.: Measurement of text similarity: a survey. Information **11**(9), 421 (2020)
14. de Wet, H., Marivate, V.: Is it fake? news disinformation detection on South African news websites. In: 2021 IEEE AFRICON. pp. 1–6 (2021). https://doi.org/10.1109/AFRICON51333.2021.9570905
15. Yuslee, N.S., Abdullah, N.A.S.: Fake news detection using naive Bayes. In: 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET). pp. 112–117. IEEE (2021)