Classification of Exaggerated News Headlines

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

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

Abstract. The amount of data online is increasing as companies generate news articles daily. These news articles contain headlines that have a level of exaggeration aimed to win the readers. In addition, these companies are competing against one another; hence creating appealing and exaggerated news headlines is one of the options to win the readers. Some of the exaggerated headlines contain some level of misleading information. Hence, this paper aims to apply machine learning methods and natural language processing to detect and identify exaggerated news headlines in South African context. Machine learning models such as logistic regression, decision trees, support vector machines, and XGBoost are trained on data that contain labelled news headlines as binary classification. The models produced good results, with XGboost and SVM obtaining 70% in terms of accuracy. Furthermore, the F measure was used to evaluate the models and decision trees obtained 56% followed by SVM with 53%. The classification of exaggerated news headlines is a difficult task. Therefore, we oversampled the data to obtain balanced labels. The performance of the models was increased. SVM obtained 84% followed by logistic regression, XGBoost, and decision trees with accuracy of 78%, 72% and 71%, respectively.

Keywords: Classification · News headlines · Machine learning · Natural language processing · Exaggerated News

1 Introduction

Text classification is a method or technique that automatically assigns labels or categories to given text or documents by utilising machine learning and statistical methods. Text classification is one of the important parts of natural language processing (NLP). It contains subtasks such as text-based language identification, topic classification, authorship classification, document classification, news classification, and more. The aim of text classification is to categorise given texts or documents by labels. This plays an important role in systems such as information retrieval and organisation [8].

The classification of news is becoming an urgent research area. Since 2016 US elections, the classification of news as real or fake gained popular awareness. Fake news detection can be done using machine learning models and NLP techniques [4]. Zhang et al. [23] discussed a comprehensive overview of fake news online,

the characterisation of the impact of fake news, and the methods used to detect fake news. These methods involve practical-based approaches such as online fact checking and research-based approaches, which involve the use of machine learning.

The media is one of the most significant sources for providing information about what is happening in the world. However, with the large number of users of social networks or the Internet, the media tend to oversell or mislead readers with exaggerated news titles. This poses the credibility of the news and makes it difficult for readers to identify non-exaggerated news. Researchers have explored the classification of misleading or fake news with supervised and unsupervised machine learning algorithms [22]. Machine learning algorithms are the most commonly and widely used by researchers to predict or classify a given set of data. The aim of this paper is to classify news headlines from the South African news data as exaggerated or non-exaggerated using machine learning models. The data set is acquired from the study by Sefara et al. [18], where the authors proposed a method to label exaggerated news headlines.

The contributions of the paper are as follows:

- We transform the data using Term Frequency-Inverse Document Frequency (TFIDF) to generate features.
- We train machine learning models on the data.
- We propose the baseline classification results of the models on the proposed data.
- We publish the code and results on Github[1] to allow future benchmarking.

This paper is organised as follows. The background is discussed in the next section. Section 3 explains the proposed architecture. Section 4 explains data collection and data engineering. The methods used to build machine learning models are discussed in Section 5. The evaluation of the machine learning models is discussed in Section 6. The findings and analysis of the results are discussed in Section 7, while Section 8 concludes the paper with future work.

2 Background

There is little or limited research that asserts the study of exaggerated news; however, most studies have explored research on fake news detection and classification. Researchers such as Snell et al. [20] took the liberty to manually classify news articles as fake or real. They began to identify news articles on different topics. Then manually evaluate each article to verify whether the article is real or fake by looking at several aspects such as capitalisation in headlines, comparison of an article from a different source to a trusted source, and others. The final step in their method was for each article to be validated by a different team member. Jehad et al. [5] classified news articles as fake or real using machine learning models; Random Forest and Decision Tree, and incorporated TF-IDF

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

as a feature extractor to improve model performance. Decision Tree appeared to perform better with feature extractor than without feature extractor.

The application of deep learning to fake news has sparked considerable interest; therefore, authors such as Mehta et al. [9] have proposed a bidirectional encoder representation (BERT) method for the classification of fake news in two separate datasets, which has demonstrated a significant improvement in performance compared to supervised machine learning classifiers such as Support Vector Machine (SVM) and others. Furthermore, Aggarwal et al. [1] proposed the BERT model to classify political news articles as fake or real. They also compared BERT to XGBoost and LSTM for performance evaluation in terms of accuracy, and BERT was the top performer with higher accuracy. Furthermore, a method for combining two word embedding models and their related entities with deep learning-based MLP method was proposed by [21] with their results suggesting that the count vectorizer and Glove together with MLP perform better than the Random Forest classifier when used with TF-IDF as a feature extractor. Jehad et al. [6] proposed a method based on TF-IDF and Multi-Layer Perceptron (MLP) to classify news as real or fake. Their results showed higher precision compared to related work covered in their study. Samadi et al. [16] have explored the use of different features such as topic extraction or categories, sentiment in text, and the use of various entities. These features were fed into a neural network classifier with semantic-based features to classify news articles as fake or real. In their analysis, the use of these features has been shown to improve the performance of the neural network classifier in terms of accuracy. Lai et al. [7] made a comparison between several machine learning models and neural network models for classification in twitter news data as fake or real. In their findings, the overall neural networks models seemed to perform better than the machine learning models. In a nutshell, there is more literature on fake news classification, with most applications based on deep learning approaches, while the application on classification of exaggerated news titles is limited. Hence, in this paper, we conduct research for the classification of exaggerated news titles in the South African context.

3 Proposed Architecture

The high-level architecture of the proposed method is illustrated in Fig. 1. For training the models, the news headline data is preprocessed to eliminate features that are not important for model training. This includes the removal of numbers, the removal of special characters, the removal of links, and the removal of stopwords. Stopwords are functional words that frequently occur within the data. Removing these words helps the model distinguish and learn the importance of features for each predicted label. Features are extracted in a form of n-gram of sizes between 1 and 3. N-grams are sequences of N words or tokens in a document that may overlap. At this stage, a vocabulary is being created. A vocabulary contains a list of features. This list can be later used to identify the importance of features. The data are divided into 80% for training the machine

learning models and 20% for testing the models. The test data are not used during model training. The data are normalised to eliminated outliers which are caused by high values of frequent words. For testing the models, the test data is used to extract the features. The features are normalised, and model prediction begins. The predicted labels and the original labels are used to evaluate the models' performance using the accuracy and the F measure.

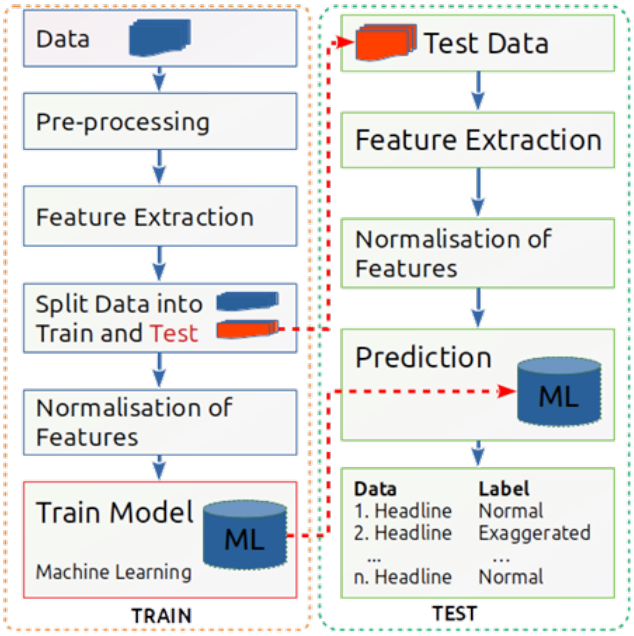


Fig. 1: The data flow diagram of the proposed method.

4 Data

This section discusses data, data processing, feature extraction, feature normalisation, and split of train and test data. The data consists of news headlines in the South African context. Data were acquired from Sefara et al. [19] in a comma separated value format (CSV). The data contains the following columns: publisher, article title, category, article content, link, publish date, label, and more. This paper focusses on the title, the content of the article, and the label. The label is the column that is used for the prediction based on the title and content of the article.

4.1 Data Exploration

Data exploration helps to understand the nature of the data. The data contain 49772 titles with their corresponding labels. The distribution of labels is shown in Table 1 with most titles labelled as not exaggerated. This shows that the data are unbalanced and that one needs to use the F score metric to evaluate the models' performance. Table 2 shows the word and character distributions of the data. The words *South, Cup, Africa, SA* and *World* appeared frequently in the news headlines. This validates the localisation of the data that is specific to South Africa.

Table 1: Label Frequency

Label	Frequency
Exaggerated	16401
Non-exaggerated	33371
Total	49772

Table 2: Top 5 characters and words frequency of the data

Words	Frequency	Characters	Frequency
South	3043	e	241856
Cup	1832	a	181982
Africa	1822	s	167685
SA	1765	i	151829
World	1649	r	147380

Figure 2 shows a graphical representation of the word distribution in the data. Larger words appeared more frequently in the data. The word cloud shows that the data are localised to South Africa and shows some of the companies and political organisations of South Africa.

Fig. 2: Word cloud of the data.

4.2 Data Pre-processing

Data pre-processing includes the cleaning of the news headlines to remove noise. Noise can be expressed in terms of characters that may confuse machine learning models. Such characters include numbers, links, emojis, and stopwords. Stopwords are function words that exist in the data and do not provide meaningful information. Removal of noise results in smaller data in terms of size, and then reduces the model training time due to a lower number of tokens in the training data. Table3 shows some of the 145 stopwords identified in the data, and it shows the frequency of special characters identified in the data. Data preprocessing is an important step in NLP to help determine which type of token should be removed from the data without losing the quality of the data. The tokens shown in Table 3 do not help the models learn significant features for each label; therefore, we removed these tokens.

4.3 Feature Extraction

Feature extraction in NLP involves the transformation of natural language into feature representation that the models can understand. There are different types of feature extraction technique including:

- Word embeddings (BoW)
- TF-IDF

Word embedding is a learnt token representation in which each token is represented as a vector in a lower-dimensional space. Word embeddings maintain the context and relationship of words in such a way that it is simple to find word similarity. Word embeddings can be implemented in various ways, including Word2vec [10], GloVe [13], FastText [2] and more. There are two approaches

Table 3: Frequency of stopwords and special characters.

Stopwords	Frequency	Special characters	Frequency
to	14194	:	12717
in	10870	’	11808
of	8784	,	9655
the	8662	...	4843
for	7704	‘	4039
and	4867	’	3239
on	4623	–	1925
a	4326	?	1673
as	3029	[	1668
with	2870	]	1668

to implement word embedding, namely, skip-gram, which predicts the context using target word, and continuous bag-of-word, which predicts the target word using the context.

TF-IDF is a statistical measure that evaluates word relevance for a document in a set of documents. TF-IDF is used mostly in NLP applications such as:

- text summarisation - a process of summarising large text into summaries
- information retrieval - a process of identifying and obtaining relevant information
- text categorisation - a process of classifying text into relevant categories, and
- sentiment analysis - a process of analysing sentiments in a text.

TF-IDF highlights significant terms that are not too frequent in the data but are of great importance. This paper explores the use of the TF-IDF method to extract features using n-gram of size 3. A Python library scikit-learn is used to implement feature extraction [12].

4.4 Train and Test data

After features are extracted for each title of the article, we divided the news headlines data into a train and test set to avoid overfitting. The train set contained 80% of the news headlines, while the test set has 20% of the news headlines. The news headlines in the test set are used only to evaluate the trained model to measure the performance and quality of the model. There are two main reasons to split the data.

- **To avoid overfitting**. Overfitting occurs when a machine learning model learns the training data too well and fails to generalise to new data. By splitting the data into a training set and a test set, we can train the model on the training set and evaluate its performance on the test set. This helps us to ensure that the model is generalising well and is not simply memorising the training data.

– **To get an unbiased estimate of the performance of the model**. If we train and evaluate the model on the same data, we will get an over-optimistic estimate of its performance. This is because the model will already have seen the data on which it is being evaluated. By dividing the data into a training set and a test set, we can ensure that the model is being evaluated on data that it has never seen before. This gives us a more unbiased estimate of the performance of the model on real-world data.

4.5 Feature Normalisation

The significance of normalising the features is to improve the performance of machine learning models. Feature normalisation is a method used to scale the independent variable or data features. This method is an important step in data preprocessing. Both the train and test data set features are normalised using a standard scaler which is defined by the following equation:

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

where x represents the feature, μ represents the mean, and σ represents the standard deviation of the data [17]. The feature normalisation step may have advantages in domains such as speech [17], but in this study the models were unable to improve performance. Therefore, we chose to keep the original features that exhibit the characteristics and context of the data.

5 Machine Learning Models

The following techniques are implemented and used to build the classification model on the acquired data set.

– **Logistic Regression** is a supervised machine learning classifier that is used for predicted binary or multiclass variables. The binary logistic regression model performs classification using two classes, which provides probabilities ranging from 0 to 1 using the cross-entropy loss function. In this paper, we used logistic regression to classify exaggerated news titles as the model performs better in a binary classification.
– **Decision Tree** is a non-parametric supervised machine learning method that is utilised for classification and regression applications. This model uses a tree-like model to make decisions. In this paper, we used the popular decision tree to classify exaggerated news titles, as it performs better on a small data set.
– **SVM** is a supervised machine learning method that can be used for classification and regression applications. In classification problems, SVM labels the data by finding the optimal decision boundary that divides the labels. SVM has a set of mathematical functions that are called kernels. Kernels are used to transform the input data into the required form. In this paper, we used different SVM kernels to choose the best kernel to classify exaggerated news titles.

– **XGBoost** is a machine learning technique that uses a group of decision trees with gradient boosting to make predictions [15]. It can be used in regression, binary classification, and multiclass classification. In this paper, it is used for binary classification.

6 Model Evaluation

Model evaluation helps to understand the performance, weaknesses, and strengths of the model. The models are tested on an unseen test data set that was not used during model creation. We used the following performance metrics to measure the quality of the models.

– **Accuracy**: is the proportion of the number of correctly predicted headlines out of all the headlines.
– **F1 score** is used to measure the accuracy of the model by calculating the harmonic mean of recall and precision. This metric helps to obtain accurate results under the conditions of imbalanced data.
– **Confusion matrix**: is an X x X matrix used to measure the performance of a classification task, where X is the number of labels.

7 Results and Discussion

The machine learning models were trained using 80% of the news headlines and evaluated on the unseen news headlines, which is 20%. The results are shown in Table 5. The models were fitted to the training data with default parameters except SVM. The SVM was fitted with the parameter *kernel* set to the following values: *polynomial*, *linear*, *radial basis function (RBF)*, and *sigmoid*. Table 4 shows the results of the test of the best SVM kernel for the classification of exaggerated news headlines. Both *polynomial* and *RBF* obtained an accuracy of 70% while *linear* and *sigmoid* obtained an accuracy of 69%. Since the data were unbalanced, we considered measuring the quality of the models using the F score. As shown in Table 4, *linear* SVM outperformed other SVM kernels, obtaining 64% followed by *RBF*, *sigmoid* and *polynomial* with 63%, 63% and 62%, respectively.

Table 4: SVM kernels.

SVM Kernel	**Accuracy (%)**	**F1 score(%)**
Linear	**69**	**64**
RBF	70	63
Sigmoid	69	63
Polynomial	70	62

Other models included XGBoost, LR, and decision trees as shown in Table 5. These machine learning models were built and tested using the proposed data.

XGBoost obtained an accuracy of 70% followed by logistic regression and decision trees with an accuracy of 69%, 64%, respectively. Furthermore, we used the F1 score as one of the evaluation metrics for the models, as our dataset was unbalanced, the F1 score was calculated with linear SVM obtaining 64% followed by decision trees, XGBoost and LR with a score of 63%, 62% and 61%, respectively.

Table 5: Model prediction results in percentage based on the proposed data.

Model	**Accuracy (%)**	**F1 score (%)**
XGBoost [14]	70	62
Linear SVM [11]	69	64
LR [11]	69	61
Decision trees [5]	64	63

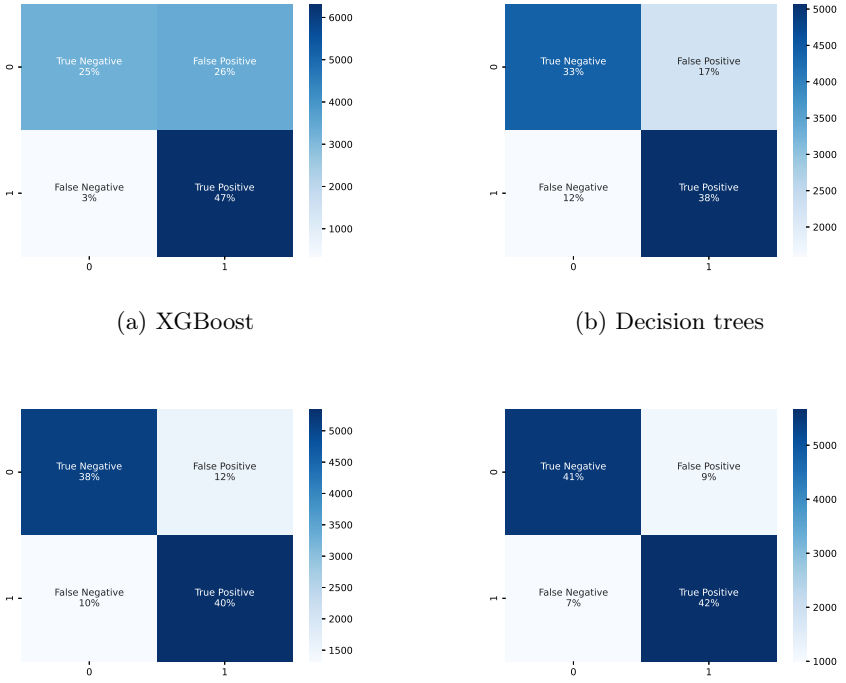
The performance of the models can be further improved by implementing data augmentation to overcome label imbalance. Data augmentation is a method to artificially increase the data size by introducing noise. Using the SMOTE oversampling method [3], we applied data oversampling on the minority label to have balanced labels. This increased the dataset by 33% to 66742 headlines. The results in Fig. 6 show the improved performance in all the models. SVM outperformed other models with accuracy and F score of 84% followed by LR, XGBoost and decision trees with accuracy of 78%, 72% and 71%, respectively. Data oversampling improved the performance of the models to reach state-of-the-art results. The models were fitted with the default parameters implemented on scikit-learn [12].

Table 6: Model prediction results after over sampling.

Model	**Accuracy (%)**	**F1 score (%)**
XGBoost	72	70
Linear SVM	84	84
LR	78	78
Decision Trees	71	71

To further evaluate the performance of the oversampled models, we computed confusion matrices in Fig. 3 for XGBoost, decision tree, logistic regression and linear SVM. The confusion matrices show that linear SVM outperformed other models in correctly predicting 84% of news headlines and missing only 16%. The second-best model is LR which correctly predicted 78% of the new titles and missing 22%. The third-best model is XGBoost in Fig. 3a which correctly predicted 72% of the news headlines and missing 28%. The decision tree model correctly predicted 71% news headlines and missing 29%. All the models were

able to predict the non-exaggerated news headlines. But linear SVM was able to predict exaggerated news headlines with an accuracy of 41% followed by LR with an accuracy of 38%. XGBoost confused 26% exaggerated news headlines with normal news, resulting in type I error, meaning that the model failed to predict the news headlines as exaggerated. In general, the four models performed better in predicting news headlines, as normal and linear models performed better in predicting exaggerated news headlines.



(a) XGBoost (b) Decision trees

(c) LR (d) Linear SVM

Fig. 3: The confusion matrices after oversampling

8 Conclusion and Future Work

This paper proposed the classification of exaggerated news headlines using machine learning algorithms applied to an exaggerated news dataset in the context of South Africa. The data contained news titles and labels. The data was preprocessed and cleaned. TFIDF was used to transform the data to generate feature sets. Several machine learning models were trained on the data and evaluated

using the accuracy, F score, and confusion matrix. The models were trained on 80% of the data and tested on the rest. The models produced baseline results for the proposed dataset. Linear SVM obtained better results with an accuracy and an F score of 84% after oversampling the data. The classification of news headlines remains a challenging task when applying machine learning. Therefore, we recommend the use of linguistic techniques to properly label the data. Future work will focus on improving the results using deep learning methods.

References

1. Aggarwal, A., Chauhan, A., Kumar, D., Verma, S., Mittal, M.: Classification of fake news by fine-tuning deep bidirectional transformers based language model. EAI Endorsed Transactions on Scalable Information Systems **7**(27), e10–e10 (2020)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
4. Ibrishimova, M.D., Li, K.F.: A machine learning approach to fake news detection using knowledge verification and natural language processing. In: Barolli, L., Nishino, H., Miwa, H. (eds.) Advances in Intelligent Networking and Collaborative Systems. pp. 223–234. Springer International Publishing, Cham (2020)
5. Jehad, R., Yousif, S.A.: Fake news classification using random forest and decision tree (j48). Al-Nahrain Journal of Science **23**(4), 49–55 (2020)
6. Jehad, R., Yousif, S.A.: Classification of fake news using multi-layer perceptron. In: AIP Conference Proceedings. AIP Publishing (2021)
7. Lai, C.M., Chen, M.H., Kristiani, E., Verma, V.K., Yang, C.T.: Fake news classification based on content level features. Applied Sciences **12**(3), 1116 (2022)
8. Mao, K., Xiao, X., Zhu, J., Lu, B., Tang, R., He, X.: Item tagging for information retrieval: A tripartite graph neural network based approach. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2327–2336. SIGIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401438, https://doi.org/10.1145/3397271.3401438
9. Mehta, D., Dwivedi, A., Patra, A., Anand Kumar, M.: A transformer-based architecture for fake news classification. Social network analysis and mining **11**, 1–12 (2021)
10. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space (2013), http://arxiv.org/abs/1301.3781
11. Patel, A., Meehan, K.: Fake news detection on reddit utilising countvectorizer and term frequency-inverse document frequency with logistic regression, multinominalnb and support vector machine. In: 2021 32nd Irish Signals and Systems Conference (ISSC). pp. 1–6 (2021). https://doi.org/10.1109/ISSC52156.2021.9467842
12. 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)

13. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
14. Rao, V.C.S., Radhika, P., Polala, N., Kiran, S.: Logistic regression versus xgboost: Machine learning for counterfeit news detection. In: 2021 Second International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE). pp. 1–6 (2021). https://doi.org/10.1109/ICSTCEE54422.2021.9708587
15. Sahin, E.K.: Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using xgboost, gradient boosting machine, and random forest. SN Applied Sciences **2**(7), 1308 (2020)
16. Samadi, M., Momtazi, S.: Fake news detection: deep semantic representation with enhanced feature engineering. International Journal of Data Science and Analytics pp. 1–12 (2023)
17. Sefara, T.J.: The effects of normalisation methods on speech emotion recognition. In: 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC). pp. 1–8 (2019). https://doi.org/10.1109/IMITEC45504.2019.9015895
18. Sefara, T.J., Rangata, M.R.: A natural language processing technique to identify exaggerated news titles. In: International Conference on Information, Communication and Computing Technology. pp. 951–962. Springer (2023)
19. Sefara, T.J., Rangata, M.R.: A natural language processing technique to identify exaggerated news titles. In: Inventive Communication and Computational Technologies. Springer Nature Singapore (in press)
20. Shewalkar, A., Nyavanandi, D., Ludwig, S.A.: Performance evaluation of deep neural networks applied to speech recognition: Rnn, lstm and gru. Journal of Artificial Intelligence and Soft Computing Research **9**(4), 235–245 (2019)
21. Thilagam, P.S., et al.: Multi-layer perceptron based fake news classification using knowledge base triples. Applied Intelligence **53**(6), 6276–6287 (2023)
22. 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
23. Zhang, X., Ghorbani, A.A.: An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management **57**(2), 102025 (2020). https://doi.org/https://doi.org/10.1016/j.ipm.2019.03.004, https://www.sciencedirect.com/science/article/pii/S0306457318306794