Identifying Characteristic Features of Fake News Articles for Deep Learning-Based Identification

Arjun Sharma

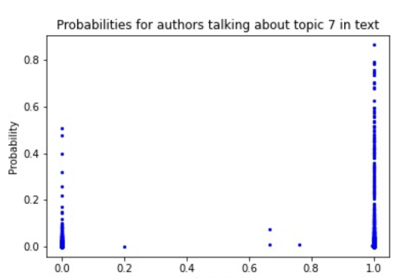
***Abstract* — Fake news articles rapidly spread online, spreading misinformation and distorting trust in institutions. This research identifies characteristic stylistic features in the text of fake news articles to build a deep learning-based binary classifier of news articles. These features are topics, sentiment, and length of titles and bodies of articles. An experimental approach is taken to rank the effectiveness of these features as input data for a neural network. The top-performing set of features results in a classifier that achieves 89.7% accuracy on a testing set.**

I. METHODS

A dataset consisting of 20 800 articles was obtained from Kaggle [1]. In it, 10387 articles were flagged as unreliable (labeled as 0), and the remaining 10413 were flagged as reliable (labeled as 1). The title, author name, and body text were also present for each article. After cleaning the dataset through lemmatization, lowercasing, and stop word removal, three categories of stylistic features are extracted for article titles and bodies.

Topics: Two separate Latent Dirichlet Allocation (LDA) models are used to identify topics in both the titles and bodies. Each model identified 20 topics in its respective corpus. All tokens in both corpora are given a probability of being associated with each of the 20 topics of that corpus. The probability of a document (a title or body) being associated with a topic is found by calculating the average association of its tokens with that topic. A probability vector is stored for each document, where each of its 20 values represents the probability of the document being associated with the corresponding topic [2]. To analyze the relationship between topics and article reliability, the topics most frequently covered by an author can be compared to the average reliability of that author. Their reliability is calculated as the fraction of articles written by the author that are labeled as reliable. The probability of writing about a topic is calculated as the average probability of their documents being associated with it. Table 1 compares author reliability to the average probability they discuss a particular topic in their article body.

TABLE 1. RELATIONSHIP BETWEEN AUTHOR TOPICS AND RELIABILITY

<table>
<tr><th>Scatterplot</th><th>Tokens</th><th>Trend</th></tr>
<tr><td rowspan="5">
</td><td>Game</td><td rowspan="5">Tend to write reliable articles</td></tr>
<tr><td>Team</td></tr>
<tr><td>Player</td></tr>
<tr><td>Season</td></tr>
<tr><td>Last</td></tr>
</table>

Sentiment: The VADER sentiment analyzer [3] is used to extract sentiment scores for the titles and bodies. It returns a compound score representing the overall sentiment of a document based on its preconfigured lexicon, ranging from -1 (negative) to 1 (positive); as well as ratios representing the proportions of words in the text which fall under positive, negative, and neutral thresholds in its lexicon. Table 2 compares sentiment polarity scores of reliable and unreliable articles for article titles and bodies.

TABLE 2. MEAN ARTICLE SENTIMENT SCORES COMPARISON

Sentiment score	**Title**		**Body**	
	Reliable	Unreliable	Reliable	Unreliable
Compound	-0.139	-0.165	0.175	0.00309
Positive (%)	9.23	11.5	8.65	8.90
Negative (%)	15.9	20.2	7.62	8.37
Neutral (%)	74.9	68.3	83.7	82.7

Two trends emerge here: a) Unreliable articles tend to use less neutral language in their titles, compared to reliable articles, by 6.6% on average. b) Reliable articles tend to be significantly more positive on average, with an average compound sentiment score greater than that of unreliable articles by approximately 0.144.

Length: Titles of reliable articles tend to be longer than those of unreliable articles, with their median length being 16 characters longer. Reliable article bodies are significantly longer than unreliable article bodies by ~95%. Table 3 compares statistical features of the length of reliable and unreliable articles for article titles and bodies.

TABLE 3. RELATIONSHIP BETWEEN AUTHOR TOPICS AND RELIABILITY

Length	Title		Body	
	Reliable	Unreliable	Reliable	Unreliable
Mean	79.6	65.2	5214.1	3875.9
Median	81	65	4591	2351
Standard deviation	15.6	31.0	4313.8	5753.6

II. EXPERIMENTAL DESIGN

At this stage, the following features are present for each article: 1) Topic probability vectors for the title and body each; 2) Four polarity scores for the title and body each; 3) Length of the title and body each. The number of articles for which topics, sentiment, and length are available is 9698. This set is split into training and testing sets, consisting of 90% and 10% of the articles, respectively. Models are trained with a 15% validation split. To identify the combination features that most effectively characterize fake news articles when used as input features for a binary classification model, multiple neural networks are trained, each of which uses one combination of features: 1) All features, 2) Title features (its topics, sentiment, and length) 3) Body features (its topics, sentiment, and length) 4) Six different combinations, each of which uses all but one of the six features. To find the optimal hyperparameters for each of the nine neural networks, the Keras tuner [4] with a hyperband search algorithm trains different variations of a base neural network architecture: A dense input layer consisting of 12 neurons; 2-20 hidden layers, each of which consists of 32-512 neurons which use a relu activation function (the number of hidden layers and neurons are optimized by the tuner for each model); and an output layer which uses a sigmoid activation function to predict a probability between 0 (reliable) and 1 (unreliable) [5] for the article. Each variation is trained for ten epochs. After the search is complete, a model with the same hyperparameters as the best-performing combination is trained for 50 epochs and saved.

III. MEASUREMENTS

The nine models can be compared by evaluating their performance when used to predict articles in the testing set in terms of 1) Their accuracy, i.e., the percentage of correctly predicted examples in the testing set; 2) The area under the curve (AUC) of the receiver operating characteristic (ROC) curve of the model, which compares the true positive rate (y-axis) to the false positive rate (x-axis) of the classifier at all classification thresholds [6]. Table 4 lists model accuracy and AUC for each of the nine models trialed.

TABLE 4. COMPARISON OF PERFORMANCE OF MODELS

Model No.	Features	Test Accuracy (%)	AUC
1	All features	88.7	0.957
2	Title features	80.0	0.865
3	Body features	85.2	0.934
4	Excluding title topics	84.7	0.938
5	Excluding title sentiment	88.9	0.956
6	Excluding title length	89.9	0.961
7	Excluding body topics	80.0	0.869
8	Excluding body sentiment	87.6	0.959
9	Excluding body length	88.7	0.958

IV. DATA ANALYSIS & DISCUSSION

First, we compare the importance of title and body features. Fig 1. compares ROC curves of the first three models.

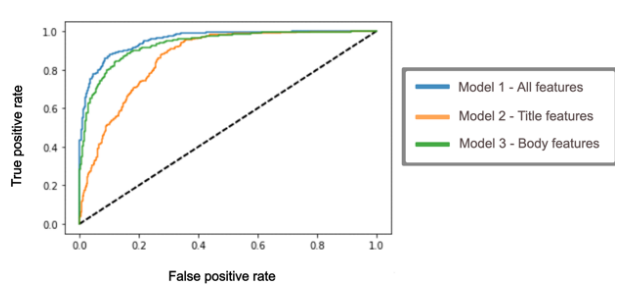


Figure 1. Comparison of ROC curves for models 1, 2, and 3

Fig. 1 indicates that body features create a more accurate classifier than title features. However, the best performing of the first three models is model 1, which achieves 88.7% accuracy on the testing set.

We then compare the importance of each of the six features by comparing the performance of model 1 with models which include all but one selected feature. Fig. 2 compares the ROC curve of the first model with ROC corves of models 4-9.

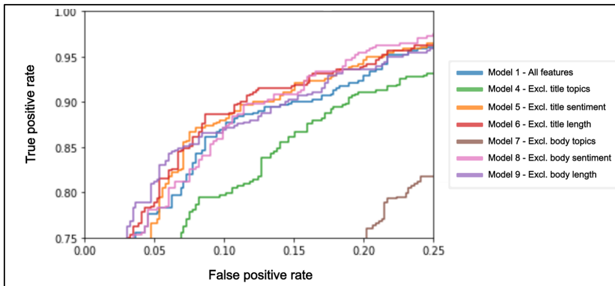


Figure 2. Comparison of ROC curves of model 1 and models 4-9, zoomed in to the top left of the graph

Fig 2. indicates that performance drastically falls when title and body topics are excluded. It also suggests a drop in performance when body sentiment is excluded. However, performance appears to improve when title sentiment is excluded (accuracy rises by 1.1% relative to model 1) and when title length is excluded (accuracy increases by 0.2%). Body length appears to have no statistically significant impact on the performance of the model.

V. CONCLUSION

The above findings indicate that the body of an article is a better predictor of the reliability of an article than its title. Furthermore, some stylistic features of the titles and bodies of articles are characteristic of fake news articles: title topics, body topics, and body sentiment. However, title sentiment and the length of both the title and body are not valid predictors of the reliability of an article. The best-performing neural network achieves 89.9% accuracy in predicting the reliability of articles in the test set. Future research may use other stylistic features, such as part-of-speech analysis; may use the propagation of fake news articles online or may analyze the behavior of the publishers of articles.

ACKNOWLEDGMENTS

I would like to thank to academic team at Polygence for their guidance and support in this research.

REFERENCES

1. UTK Machine Learning Club. "Fake News". Retrieved from kaggle.com/c/fake-news/data.
2. Řehůřek, R., & Sojka, P. (2010, Μάιος 22). Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45–50. Valletta, Malta: ELRA.
3. Hutto, C.J. & Gilbert, E.E. (2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
4. Chollet, F., & Others. (2015). Keras. https://keras.io
5. Google. (n.d.). Machine learning glossary. Google Developers. developers.google.com/machine-learning/glossary/
6. Google (n.d.) Classification: ROC Curve and AUC. Google Developers. developers.google.com/machine-learning/crash-course/classification/roc-and-AUC