Classifying Fake News on the Fake and Real news Dataset

Pedro Antônio Silva

I. Dataset

The dataset used was initially discussed in the context of bridging the gap between AI and human linguistic judgment, as highlighted in the paper ”AI vs linguistic-based human judgement: Bridging the gap in pursuit of truth for fake news detection” [1]. The business application, is therefore, to leverage AI models to assist human fact-checkers in quickly filtering out potentially harmful or misleading content.

II. Classification Pipeline

The classification pipeline began with text preprocessing using Lemmatization [2] to reduce words to their root form, then a TF-IDF (Term Frequency-Inverse Document Frequency) [3] vectorizer was applied to extract features from the text to highlight the importance of word frequency classification.

A Logistic Regression classifier [4] was then applied to the TF-IDF vectors. this model was used due to its interpretability and simplicity, where the learned coefficients indicate the weight given to each word in the classification process.

III. Evaluation

A balanced accuracy score was used to evaluate the classifier, ensuring the results weren’t biased by the imbalance between true and false articles. The classifier achieved a 94% balanced accuracy score, which aligns with findings from the paper ”Fake News Detection Platform—Conceptual Architecture and Prototype” [5], cited in the previously referenced work [1]. Significant words from each class were chosen as key examples of what the model considers during classification, based on the learned coefficients:

TABLE I
Most Significant words for classification

Score	***Word***	***Lemma***	***Specific Term***
-3.69	watch	O	X
-3.13	https	X	O
7.32	democratic	X	X
7.6	edt	X	O
-4.1	read	O	X

[a]This is a sample of the results

Negative scores, such as for the verbs ”watch” and ”read” indicate a higher likelihood of classifying an article as fake when these words are present, possibly due to their common use in online media. On the other hand, words like ”democratic” have higher positive scores, as their use is more commonly found in legitimate articles in the dataset. Additionally, terms related to media types and goals, such as ”edt” (a time zone) and ”https” (a link component), had scores reflecting where they commonly occur, with the first contributing positively and the second negatively.

IV. Data Size

As the dataset used has a considerable size and the model achieved high accuracy, a down sampling strategy was applied to seek understanding of the impacts and relevance the amount of data collected had in the model.

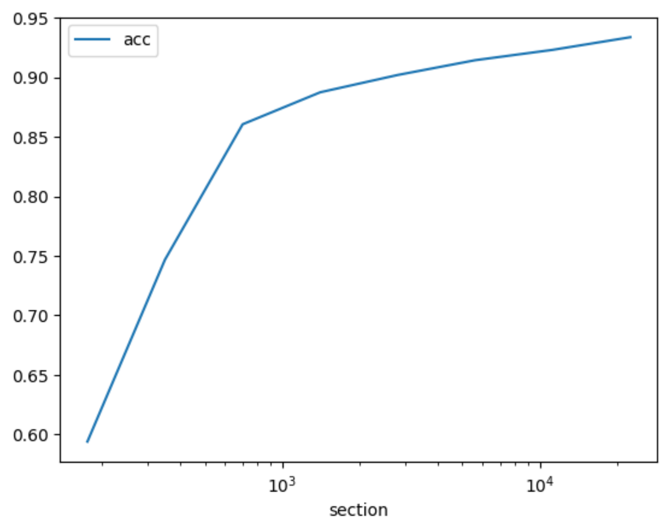

Fig. 1. Model Accuracy (%) in proportion to Dataset growth

Observing the results, it can be assessed that this model displays a considerable increase in accuracy as the sample size approaches 10^3 and stabilizes while approaching 10^4, with a noticeable diminished gain from sample size increases.

V. Topic Analysis

The analysis began by incorporating NMF (Non-negative Matrix Factorization) [6] between the vectorizer and classifier in the pipeline. After fine-tuning the model to determine the optimal number of topics, 10 topics were selected, this two-step configuration yielded a balanced accuracy score of approximately 88%.

One of the possible causes for this loss in accuracy could be the lack in overlap between the words in the topics and the words defined as relevant by the model, making this extra layer of abstaction an obstacle for classification..

REFERENCES

[1] Anonymous, "AI vs linguistic-based human judgement: Bridging the gap in pursuit of truth for fake news detection," unpublished.

[2] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, Inc., 2009.

[3] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Proceedings of the First International Conference on Machine Learning*, 2003, pp. 133–142.

[4] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[5] R. Kozik, M. Pawlicki, S. Kula, and M. Choraś, "Fake news detection platform—conceptual architecture and prototype," unpublished.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999, doi: 10.1038/4456.