TWEET VERACITY: A MACHINE LEARNING APPROACH FOR CLAIM CHECK WORTHINESS

AUTHORS

Amey Subash Pawar

Shreeyash Khalate

Swaib Ilias Mazumder

Rahul IIITD

CSE 556: NLP
GROUP NO: 32
PROF. MD. SHAD AKHTAR

This poster explores a machine learning approach to identify tweets containing claims that warrant fact-checking efforts. We present a novel system for prioritizing tweets based on their check-worthiness, aiding in the fight against misinformation on social media platforms.

INTRODUCTION

In today's digital landscape, the proliferation of misinformation and fake news on social media platforms poses a significant challenge. With Twitter emerging as a primary source of news and information for millions, the need for automated tools to assess the credibility of tweets is paramount. This study introduces a machine learning approach tailored to evaluate the claim check worthiness of tweets, aiming to combat the spread of false information on the platform.

OBJECTIVE

The objective of this study is to develop a machine learning model that can effectively determine the credibility of tweets through a systematic process of preprocessing, feature extraction, and model training. By leveraging techniques such as tokenization, stopword removal, and support vector classification, the aim is to provide a reliable tool for identifying misinformation and fake news within Twitter content.

RELATED LITERATURE

The initial works made use of classifiers such as SVMs, Decision Trees, Naive Bayes like by [1]. For feature extracion, the first attempts were made by using count based methods like TF-IDF, POS tags, sentiment scores etc. Additionally, works such as [2] added features such as average embedding vectors of a sentence, linguistic features, position awareness amongst sentences. Classifiers such as Deep Feed Forward Neural Networks were used. [3] used RNN based network where tokens were represented from a mix of embeddings, POS tags and one hot encoding of syntactic dependencies. [4] used n-gram features and used a KNN classifier.

METHODOLOGY

1. **Data Collection:** Use a publicly available dataset of labeled tweets.
2. **Preprocessing:** Tokenize, remove stopwords, punctuation, URLs, and perform spelling correction.
3. **Feature Extraction:** Extract POS tags, dependency relations, and create sentence vectors using a pre-trained Sentence Transformer model.
4. **Contextual Word Embeddings:** we use all-MiniLM-L6-v2 sentence transformer embeddings that are trained using the context of the word in a sentence
5. **Concatenation, Dimensionality Reduction and Classification:** To get the overall representation of the tweet, we concatenate all the syntactic features together with transformer-generated contextual features and then apply PCA for dimensionality reduction. SVM classifier is trained on the feature vectors of tweets to output a binary decision (check worthy or not).

CONCLUSION

In conclusion, the machine learning approach presented in this study offers a promising solution to the challenge of misinformation and fake news on Twitter. By leveraging advanced NLP techniques and robust classification algorithms, the model achieves high accuracy in identifying dubious content within tweet streams. Future research endeavors could explore the integration of additional features, such as named entity recognition, to further enhance the model's performance. Moreover, extending the scope of analysis to encompass other social media platforms would enable a comprehensive approach to combating misinformation in the digital sphere.

REFERENCES

[1]N. Hassan, C. Li, M. Tremayne, Detecting check-worthy factual claims in presidential debates, in: Proceedings of the 24th acm international on conference on information and knowledge management, 2015, pp. 1835–1838.

[2]P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, A context-aware approach for detecting worth-checking claims in political debates, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, 2017, pp. 267–276.

[3]C. Hansen, C. Hansen, J. G. Simonsen, C. Lioma, The Copenhagen Team Participation in the Check-Worthiness Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 CheckThat! Lab (2018) 8.

[4]B. Ghanem, M. Montes-y Gomez, F. Rangel, P. Rosso, UPV-INAOE-Autoritas - Check That: Preliminary Approach for Checking Worthiness of Claims (2018) 6.

ANALYSIS

As we can see from the distributions of class labels that the dataset is imbalanced. So we have used weighted SVM on dataset which will give more weightage to the minority class during training
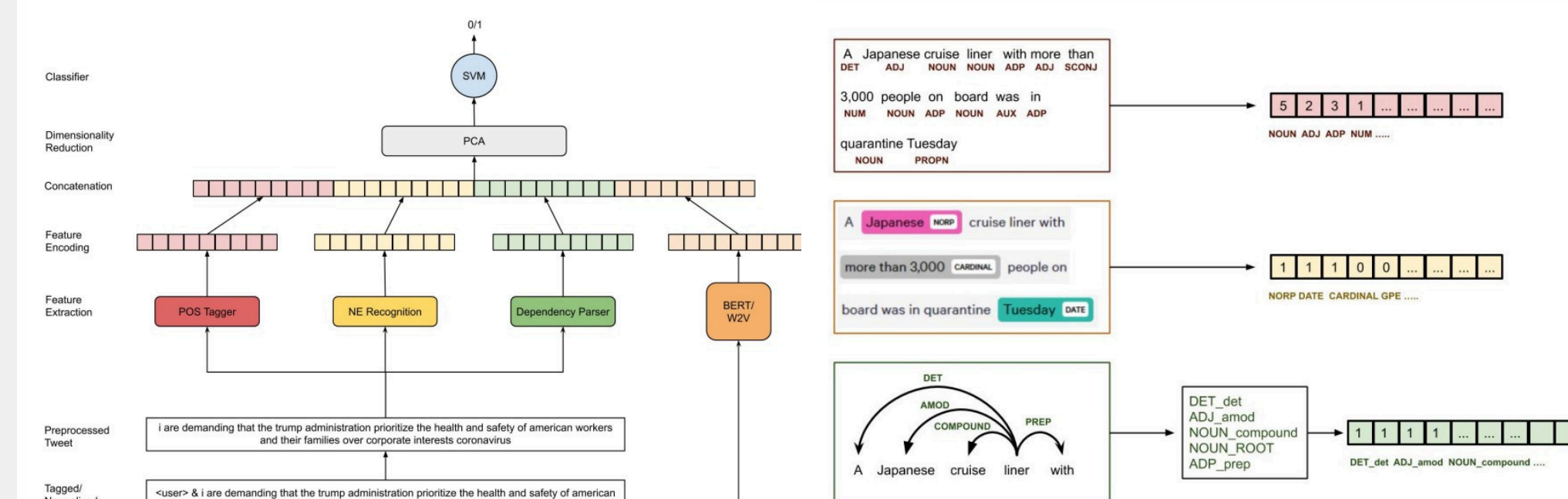

Fig. 2: Proposed Approach for Check-Worthiness Prediction

Fig.1: Syntactic feature extraction and encoding process. Feature vectors are based on the number of times it is seen in the given sentence.

Table 1: Sample tweets for Task-1 Check-Worthiness Prediction

Tweet	Claim	Check-Worthy
Dear @VP Pence: What are you hiding from the American people? Why won't you let the people see and hear what experts are saying about the #CoronaOutbreak?	0	0
Greeting my good friends from the #US the #Taiwan way. Remember: to better prevent the spread of #COVID19, say no to a handshake & yes to this friendly gesture. Check it out:	0	0
Corona got these flights cheap as hell I gotta job interview in Greece Monday	1	0
My mum has a PhD on Corona Virus from WhatsApp University	1	0
This is why the beaches haven't closed in Florida, and why they've had minimal COVID-19 prevention. Absolute dysfunction. <link>	1	1
COVID-19 cases in the Philippines jumped from 24 to 35 in less than 12 hours. This is seriously ALARMING. Stay safe everyone! <link>	1	1

```
Classification Report for Validation Set:

              precision    recall  f1-score   support

           0       0.92      0.78      0.85       503
           1       0.48      0.76      0.59       134

    accuracy                           0.78       637
   macro avg       0.70      0.77      0.72       637
weighted avg       0.83      0.78      0.79       637
```


Heatmap for Support Vector Classifier
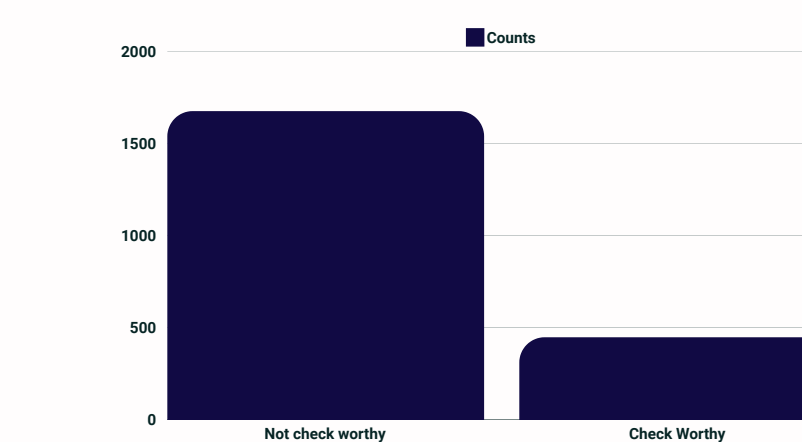
TRAINING ACCURACY   89%

VALIDATION ACCURACY   77%