NLP-Fake News Detection

Apostol Alin, Ciuperceanu Vlad, Pîrvulescu Daria

May 2025

Introduction

This paper aims to highlight the importance of recognizing synthetic online content through the use of **Natural Language Processing.**

In today’s society, disinformation can range from images and videos altered using AI tools to entirely fabricated news articles and election posts. That being said, by leveraging **NLP techniques**, we can develop automated systems capable of identifying patterns of synthetic content.

Related Work

FakeBERT:

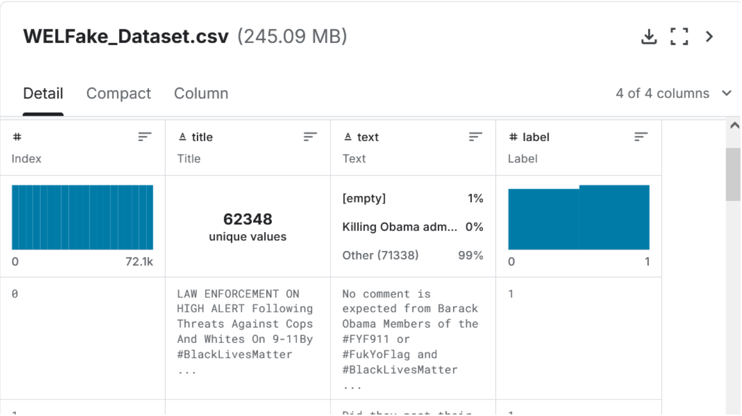
In the paper *’Fake news detection in social media with a bert-based deep learning approach. Multimedia tools and applications*’, Rohit Kumar Kaliyar, Anurag Goswami and Pratik Narang introduce an improved BERT, which was used to classify fake news from real news.

The results where impressive:

’Our proposed model (FakeBERT) produced more accurate results as compared to existing benchmarks with an accuracy of **98.90** *%.’*

Dataset

The dataset used in our experiments was obtained from Kaggle and originates from the study: Welfake: Word embedding over linguistic features for fake news detection.



WELFake_Dataset.csv (245.09 MB)

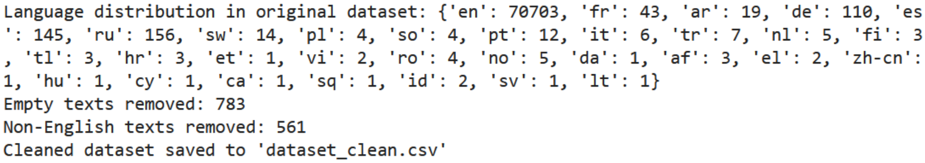
Detail Compact Column 4 of 4 columns

# Index	title Title	text Text	# label Label
0 72.1k	62348 unique values	[empty] 1% Killing Obama adm... 0% Other (71338) 99%	0 1
0	LAW ENFORCEMENT ON HIGH ALERT Following Threats Against Cops And Whites On 9-11By #BlackLivesMatter ...	No comment is expected from Barack Obama Members of the #FYF911 or #FukYoFlag and #BlackLivesMatter ...	1
1			1

Figure: Dataset View

Dataset-Analysis

Eliminating: empty strings, non-English texts, hyperlinks and emojis



```
Language distribution in original dataset: {'en': 70703, 'fr': 43, 'ar': 19, 'de': 110, 'es
': 145, 'ru': 156, 'sw': 14, 'pl': 4, 'so': 4, 'pt': 12, 'it': 6, 'tr': 7, 'nl': 5, 'fi': 3
, 'tl': 3, 'hr': 3, 'et': 1, 'vi': 2, 'ro': 4, 'no': 5, 'da': 1, 'af': 3, 'el': 2, 'zh-cn':
1, 'hu': 1, 'cy': 1, 'ca': 1, 'sq': 1, 'id': 2, 'sv': 1, 'lt': 1}
Empty texts removed: 783
Non-English texts removed: 561
Cleaned dataset saved to 'dataset_clean.csv'
```

Figure: Dataset Distribution

Dataset-Statistic

As seen in the graphics, country names and president names are the most frequent and we expected it to affect the training of the model.

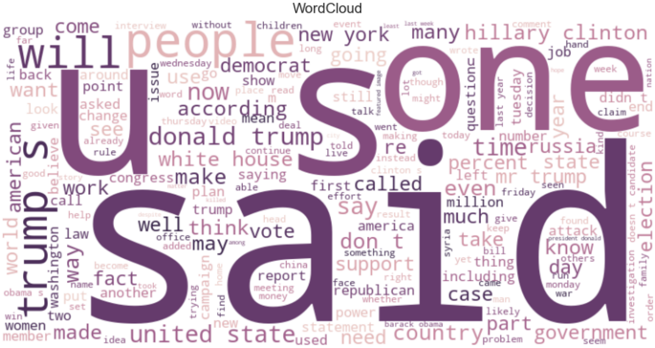


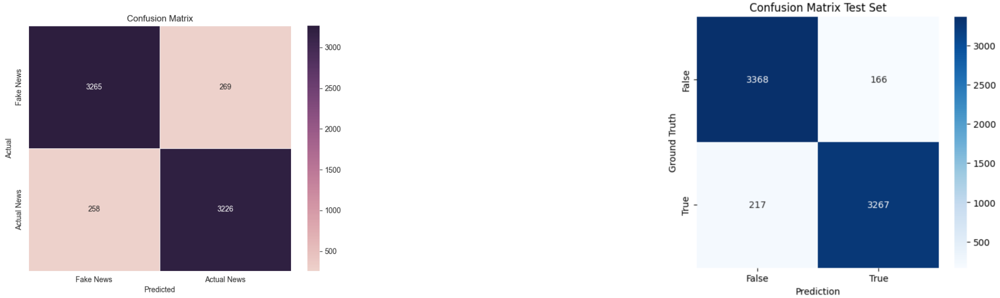
Figure: Word Cloud

Word-Embeddings

We used : Word2Vec-CBOW, Word2Vec-SkipGram and GloVe (*both in their pretrained form and trained from scratch*).

Classic-ML

As for comparing our two models the best combinations were: Random-Forest with Word2Vec-SkipGram trained from scratch with $92,49\%$ accuracy with Linear SVM Word2Vec-SkipGram trained from scratch with $94,54\%$ accuracy. It seems that **Skip-Gram** worked best for our large dataset.



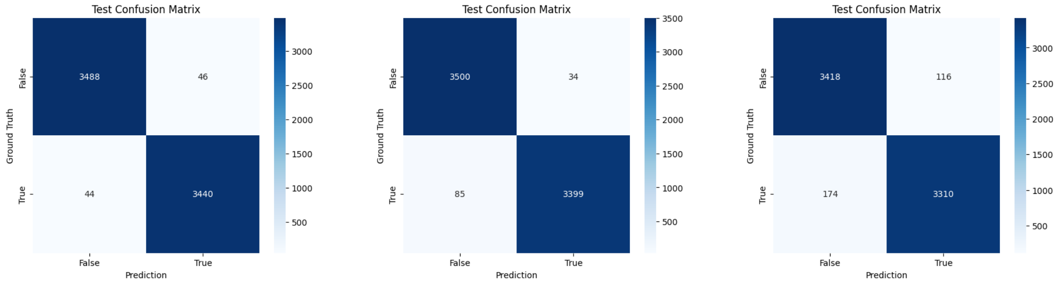
(a) Random Forest

(b) SVM

Figure: Confusion matrices for the best models

Transformers

BERT with 98.72%
DistilBERT with (98.30%)
TinyBERT with (95.87%).



(a) BERT (b) DistilBERT (c) TinyBERT

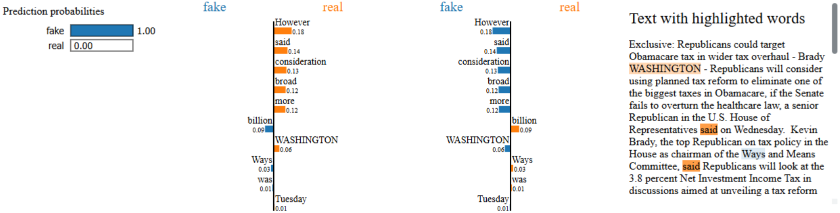
Figure: Confusion matrices for the best models

Fake part extraction

We aimed to extract the fake part of a fake news, or what out best performing models (DistilBERT and BERT) take most in consideration when making a prediction for real vs fake news.
For this we used two approaches: Lime and Captum Integrated Gradients.

Lime

Lime trains a simple, interpretable model on the outputs of the original model for different small variations of the original text.



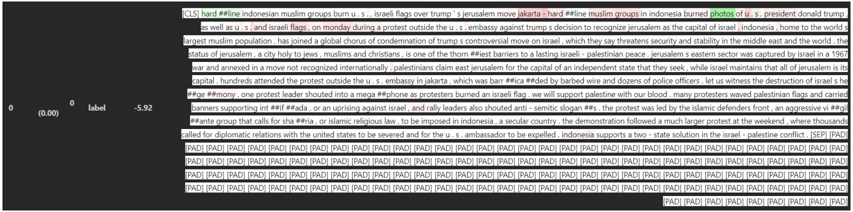
Text with highlighted words

Exclusive: Republicans could target Obamacare tax in wider tax overhaul - Brady WASHINGTON - Republicans will consider using planned tax reform to eliminate one of the biggest taxes in Obamacare, if the Senate fails to overturn the healthcare law, a senior Republican in the U.S. House of Representatives said on Wednesday. Kevin Brady, the top Republican on tax policy in the House as chairman of the Ways and Means Committee, said Republicans will look at the 3.8 percent Net Investment Income Tax in discussions aimed at unveiling a tax reform

Figure: Word importance analysis using Lime

Captum - Integrated Gradients

The Captum library provides the Integrated Gradients algorithm that calculates the integral of gradients with respect to inputs along the path from a given baseline to input.



0	0 (0.00)	label	-5.92	[CLS] hard ##line indonesian muslim groups burn u . s . , israeli flags over trump ' s jerusalem move jakarta - hard ##line muslim groups in indonesia burned photos of u . s . president donald trump , as well as u . s . and israeli flags , on monday during a protest outside the u . s . embassy against trump s decision to recognize jerusalem as the capital of israel . indonesia , home to the world s largest muslim population , has joined a global chorus of condemnation of trump s controversial move on israel , which they say threatens security and stability in the middle east and the world . the status of jerusalem , a city holy to jews , muslims and christians , is one of the thorn ##iest barriers to a lasting israeli - palestinian peace . jerusalem s eastern sector was captured by israel in a 1967 war and annexed in a move not recognized internationally . palestinians claim east jerusalem for the capital of an independent state that they seek , while israel maintains that all of jerusalem is its capital . hundreds attended the protest outside the u . s . embassy in jakarta , which was barr ##ica ##ded by barbed wire and dozens of police officers . let us witness the destruction of israel s he ##ge ##mony , one protest leader shouted into a mega ##phone as protesters burned an israeli flag . we will support palestine with our blood . many protesters waved palestinian flags and carried banners supporting int ##if ##ada , or an uprising against israel , and rally leaders also shouted anti - semitic slogan ##s . the protest was led by the islamic defenders front , an aggressive vi ##gil ##ante group that calls for sha ##ria , or islamic religious law , to be imposed in indonesia , a secular country . the demonstration followed a much larger protest at the weekend , where thousands called for diplomatic relations with the united states to be severed and for the u . s . ambassador to be expelled . indonesia supports a two - state solution in the israel - palestine conflict . [SEP] [PAD]

Figure: Word importance analysis using Integrated Gradients

Conclusions

While our models achieved **promising results**, we acknowledge the **limitations** inherent to this task, including dataset bias, linguistic ambiguity, and the evolving nature of deceptive content.

Our analysis for the fake part extraction shows that extracting the fake part from a fake news still has to be deeply studied, as we've only shown certain buzzwords that influence the model into taking a decision.