Hyperpartisan News and Articles Detection Using BERT and ELMo

Gerald Ki Wei Huang
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
geraldkwei@outlook.com

Jun Choi Lee
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
jclee@unimas.my

***Abstract*— Fake news and articles are misleading the readers. This leads to the increasing studies of fake news article detection over the decades. Hyperpartisan news is news riddled with twisted and untruth and extremely one-sided. This news can spread more successfully than others. Besides that, hyperpartisan news can mimic the form of regular news articles. This study aims to identify and classify the hyperpartisan news with BERT and ELMo. Two distinct models, BERT and ELMo, were created to classify hyperpartisan news from two datasets, namely by-article and by-publisher. Few other models with different settings and training designed to test and optimise the performance of both models. The results of the optimised BERT and ELMo models can achieve 68.4% and 60.8%, respectively.**

Keywords—Natural Language Processing, Classification, Hyperpartisan

I. INTRODUCTION

In this era, there has been concern about fake news or articles that are misleading the readers. The problem inspires various studies in identifying and classifying the fake and bias news over the decades. Hyperpartisan news is news riddled with untruth and twisted statements of information. This type of news spread more successfully than others. Hyperpartisan news not only can mislead readers but also can cause polarisation within a community or society.

According to [2], there is less research conducted for Clickbait and hyperpartisan news. Hence previous work of [2] was to help close this gap from both ends. This study aims to identify and classify hyperpartisan news using BERT (Bidirectional Encoder Representations from Transformers) [3] and ELMo (Embeddings from Language Models) [4]. BERT and ELMo recently emerge and gain impactful ground in Natural Language Processing (NLP) community on the given dataset from SEMEVAL 2019 [2]. In this paper, we will use the dataset by-article and by-publisher to train the BERT and ELMo model separately. The BERT model and ELMo model were created to identify and classify the hyperpartisan news. Both the model uses the same treated dataset that after the data cleaning process. The data cleaning process includes the removal of HTML tags and advertisements, text splitting using NLP toolkit spaCy and scikit-learn. The by-article dataset is split into training and validation sets. Both the BERT and ELMo models are pre-trained and fine-tuned to perform the classification task. Several BERT and ELMo Models were created during the study. These models are created through the initial performance testing, optimization process, and to further investigate the performance of the model.

II. RELATED WORK

There are many previous attempts to classify and identify fake or misleading news and articles; the approaches consist of from knowledge-based [3] towards style-based [1] [4]. [5] built a neural network to predict the political ideology of news articles to be either left, right, or center. The researchers combined the information from the headlines, the links found within an article, and the content. They use a CNN [6] for the headlines, a Node2Vec [7], to model the links and a hierarchical attention network [8] to extract features from the content. They compared the model with several difference baselines, which include Bag of Words with Logistic Regression model, fully connected feedforward network, or networks with only the individual components. The proposed model performs very well. However, their system was trained and evaluated on only data with publisher labels, which have randomly split into training and testing sets with overlapping publishers.

[9] [10] [11] exploited the idea to detect rumours on Sina Weibo with several new features. [10] proposed a framework for real-time news certification. The researchers gather related microblogs based on the keywords of an event using a distributed data acquisition system for real-time processing. An ensemble model has been build that combined user-based, propagation-based, and content-based model. The result has shown that the ensembled model is able to boost the performance, able to show a response at 35 seconds on an average per query, which is very crucial for a real-time system.

[1] used the BuzzFeed-Webis News corpus that encompasses the output of 9 publishers on seven workdays, 19 to 23,26, and 27, close to the US presidential elections 2016. Among the selected publishers, there were six prolific hyperpartisan, which were three left-wing, three right-wing, and the other three were mainstream. Every news and linked news article have been manually fact-checked by 4 BuzzFeed journalists. Each article had reviewed once, and articles were assigned round-robin. The researchers then classify the hyperpartisanship vs mainstream articles. The results found out that news articles that convey a hyperpartisan world view can be distinguished from more balanced news by writing style alone. The researchers concluded that they found quantifiable evidence that the writing styles of news of the two opposing orientations were in fact very similar. It has appeared to be a typical writing style of left and right extremism.

978-1-7281-6593-6/19/$31.00 ©2019 IEEE