DL-TXST FakeNews: Enhancing Tweet Content Classification with Adapted Language Models

Muhieddine Shebaro, Jason Oliver, Tomiwa Olarewaju, Jelena Tešić
Computer Science, Texas State University, San Marcos TX, USA
{m.shebaro, jasonoliver, tro24, jtesic}@txstate.edu

ABSTRACT

DL-TXST team participation runs submitted to the MediaEval Fake News task this year focused on improving the baseline benchmark pre-processing and modeling. We have introduced features learned from large, adapted language models. The predictive power of our pipeline was the strongest when we included the BERT model tuned to Tweet content. Subtask 1 on the test set had MCC 0.1.

1 INTRODUCTION

With today's modern technology, breaking news, from the latest celebrity gossip to updates on unprecedented events like the COVID-19 pandemic, are now available with just a few taps on your smartphone. As the availability and volume of readily available information has grown, so has the rise of misinformation. Fake news is specifically designed to plant a seed of mistrust and exacerbate existing social and cultural dynamics by misusing political, regional, and religious undercurrents [1]. "In 2019, 8 percent of engagement with the 100 top-performing news sources on social media was dubious. In 2020, that number more than doubled to 17 percent" [3]. Twitter's purpose has been advertised to the public as a platform that "uniquely provides its users the opportunity to discover what's happening in the world" [4]. Unique includes fake, so the Twitter platform has become an easy target for the rapid dissemination of skewed facts to the world, as seen with the attribution of the current COVID-19 pandemic to novel 5G technology. Topical automated classification systems with potent predictive power for innumerable conspiracies are urgently needed to curb the spread of inaccurate news. In this paper, we focus on content-based fake news detection strategies.

2 RELATED WORKS

This problem of misinformation in social media is universally faced by any user of a social media site. These users, as well as the private companies who run these social media sites, have a vested interest in ensuring that the information on the platform is beneficial to the consumer (the users). For most users, this means that information is accurate and can be trusted as valid. For example, rumors have surfaced in the past about McDonald's use of worm filler in its food. This has caused tremendous boycott threats [2].

3 DATA MANAGEMENT

The most recent data is collected from MediaEval's FakeNews: Coronavirus and 5G Conspiracy benchmark project [6] and is integrated with the data of the previous analysis and work that was retrieved using TwitterAPI. We used several pre-processing methods on the data. First, we used the baseline pre-processing [5], which included converting to lowercase; removing punctuation; preserving URLs; removing stop words; and normalizing terms ("u.k" to UK). Our pre-processing enhancements to the pipeline this year include removing usernames (Twitter handle); removing all special characters; removing hashtags; removing contractions (e.g., convert "won't" to "will" and "not"); removing non-English Tweets if present, removing links (which not only incorporates "https://t.co/", but also "http" and "www"), and removing Emojis. When we looked at the dataset for Subtask 2 and 3, the Tweet was divided into several parts, and each part was present in a separate column. To deal with this, we merged them into one column in the data frame separated by a space. The validation size was set to 0.2 to partition our dataset for the sake of evaluating our model's predictive power according to a set of predefined metrics.

MediaEval'21, December 13-15 2021, Online

4 EXPERIMENTS

4.1 Subtask 1 The objective is to build a multi-class classifier that can flag whether a Tweet promotes, supports or discusses at least one (or many) of the conspiracy theories.

Table 1. Baseline performance on old development set.

Accuracy	Precision	Recall	F1	MCC
73.92%	56.85%	54.46%	54.69%	41.88%

Pre-Processing. Links that contain or start with "https://t.co/" are removed, but links such as the ones beginning with "http" and "www" are still present even after applying the control's normalization. Username handles are also not filtered out.

Data Integration. Combining two datasets requires them to have the same dimensions, as well as consistent and meaningful class labels. We observed that there are some discrepancies between these two datasets, which would impede the flow of the integration process. For this reason, before integration, we began by carefully selecting class labels from fine-grained classification that would make sense in the new dataset. We replaced the class label of tuples that is 1 with 3 and 3 with 1. We also came to a consensus that label 2 is irrelevant in our new context. Thus, we excluded all tuples having this class label. To form a uniform dataset with a uniform number of dimensions, we extracted only the "Tweet" and the "Label" dimensions from the old dataset, finally rendering the previous dataset integrable with the new one (no missing Tweets detected). Before fusion, the number of tuples of the old dataset was

5,946 rows. After integration and removing rows that are labeled as 2, we got a total of 6,769 tuples.

Modeling We chose Logistic Regression as a baseline model because it performed the best in the control experiment [5]. However, we modified it by applying some hyperparameter tuning to adjust to our new, fully integrated dataset. For example, we altered the class weight attribute to 1: 0.1, 2: 0.7, 3: 0.2. We also increased maximal iterations from 2000 to 4000 because there were some instances in which the model did not converge. We kept the same feature extracting technique (CountVectorizer) and utilized the spacy to tokenize our text. In addition, the test size was set to 0.2 to partition our dataset for the sake of evaluating our model's predictive power according to a set of predefined metrics. We utilized the voting classifier to combine several selected models. The selected models were based on similar related works on Tweets [7]. They are SVC, Multinomial NB, Logistic Regression, and Random Forest Classifier. The voting type was set to "hard."

BERT for Tweets "BERT-large was trained on 64 TPU chips for four days at an estimated cost of $7,000." [8]. Selecting a pretrained model for BERT is a crucial step when fitting your model. For instance, we initially used the pretrained model offered by Google (BERT-Large, Uncased) [9]. We ended up with dismal results. As it turns out, the BERT-Large pretrained model was trained and is based on conversational English text. However, we know that the structure and nature of Tweets are very different from those of any other text. For this reason, we searched to find a pretrained model for Tweets, and we stumbled upon BERTweet [10]. We based our code on a similar work that was already done on Kaggle, but for disaster Tweets. This code has utilized BERTweet pretrained model from "Vinai" [11] with some modifications. For example, shifting our class labels (1 to 0, 2 to 1, and 3 to 2) was a requirement for BERTweet to work. We kept the same hyperparameters (5 epochs and batch set size to 8) and changed the num_classes parameter.

4.2 Subtask 2 & 3

Encoding & Decoding Since our data contains multiple target variables, it was beyond the models' inference capabilities of more than one dependent variable at once. So, we came up with an idea to encode every occurrence of a combination of binary target variables into a single target variable. For example, 0,0,0,0,0,0,0,0,0 has 754 occurrences and we encoded it with 0. For the sake of simultaneously reducing the number of class labels and improving generalization, we decided to apply a threshold to remove any rare combination of binaries that has an occurrence less than the threshold. We found that the ideal threshold, 20, would capture the most frequent occurrences. A total of 10 encodings (class labels) were produced after using this threshold. When the model produces a label 3, the decoding process is going to translate it back into 0,0,0,0,0,0,1,0,0. The same experiments were applied to subtasks 2 and 3, except for the data integration, as the class labels of the old dataset are irrelevant in this context.

Subtask 1	MCC Score	Subtask 2	MCC Score	Subtask 3	MCC Score
001	0.106	101	0.0807	201	0.08926
002	0.0784	102	0.0775	202	0.03060
003	0.0995	103	0.0724	203	0.0676

Table 2 MCC for Official Test Runs

5. RESULTS

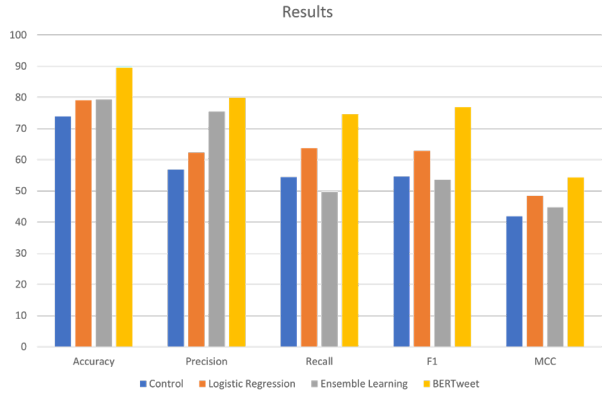

Figure 2 Validation set results for Subtask 1

BERTweet outperforms all models in terms of multiple metrics on validation set for subtask 1, as illustrated in Figure 2.

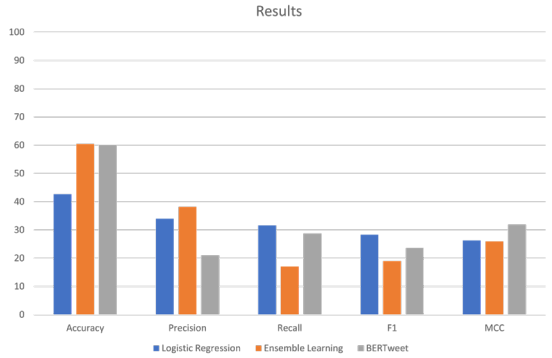

Figure 3 Validation set results for Subtask 2

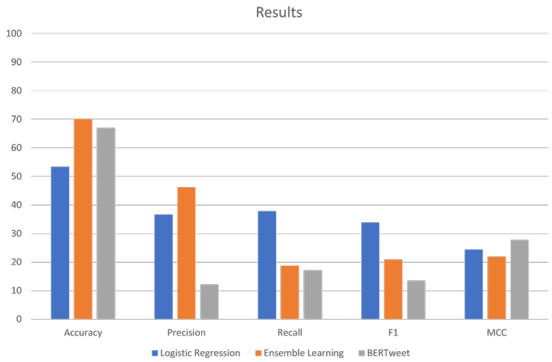

Figure 4 Validation set results for Subtask 3

Logistic regression has the best recall and F1 for Subtasks 2 and 3, where BERTweet has the comparable score and highest MCC, as illustrated in Figures 3 and 4. Table 2 summarizes return results on the test set.

CONCLUSION Tweet content normalization techniques improve the predictive power of the pipeline. BERTweet was significantly better at predicting the subtask 1 data with MCC 0.106. The new Normalizations + Logistic Regression performed the best in both subtasks 2 and 3.

REFERENCES

[1] Claire Wardle, Hossein Derakhshan, INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making, DGI(2017)09, Avenue de l'Europe F - 67075 Strasbourg Cedex, France: Council of Europe, 2017.

[2] Kate Taylor. "A viral rumor that McDonald's uses ground worm filler in burgers has been debunked" Business Insider. https://www.businessinsider.com/debunked-mcdonalds-uses-worm-filler-2016-1#:~:text=A%20viral%20rumor%20that%20McDonald's,in%20burgers%20has%20been%20debunked&text=Robert%20Galbraith%2FReuters%20If%20you,worry%20%E2%80%94%20it%20is%20completely%20false (accessed September 19, 2021).

[3] Emily Stewart. "America's growing fake news problem, in one chart" Vox. https://www.vox.com/policy-and-politics/2020/12/22/22195488/fake-news-social-media-2020 (accessed September 19, 2021).

[4] Cartier Stennis. "Defining what makes Twitter's audience unique" Twitter Blog. https://blog.twitter.com/en_us/topics/insights/2018/defining-what-makes-twitters-audience-unique (accessed September 19, 2021).

[5] Andrew Magill, Lia Nogueira De Moura, Maria Tomasso, Mirna Elizondo, Jelena Tešić. "Enriching Content Analysis of Tweets using Community Discovery Graph Analysis", MediaEval 2020 workshop paper.

[6] Konstantin Pogorelov, and Daniel Thilo Schroeder, and Stefan Brenner, and Johannes Langguth. FakeNews: Corona Virus and Conspiracies Multimedia Analysis Subtask at MediaEval 2021. Proc. of the MediaEval 2021 Workshop, Online, 13-15 December 2021.

[7] Ankit, & Saleena, Nabizath. (2018). An Ensemble Classification System for Twitter Sentiment Analysis. Procedia Computer Science. 132. 937-946. 10.1016/j.procs.2018.05.109.

[8] Roy Schwartz, Jesse Dodge, Noah A. Smith, Oren Etzioni. "GreenAI". https://dl.acm.org/doi/fullHtml/10.1145/3381831 (accessed October 20, 2021).

[9] Jacob Devlin. "Google Research / BERT". Github. https://github.com/google-research/bert (accessed October 20, 2021).

[10] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. "BERTweet: A pre-trained language model for English Tweets". Aclanthology. https://aclanthology.org/2020.emnlp-demos.2.pdf. (accessed October 20, 2021).

[11] Matthias Bachfischer. "Disaster Tweets - BERTweet". GitHub, https://www.kaggle.com/matthiasbachfischer/disaster-tweets-bertweet (accessed October 20, 2021).

[12] Konstantin Pogorelov, and Daniel Thilo Schroeder, and Petra Filkuková, and Stefan Brenner, and Johannes Langguth. WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets. Proc. of the 2021 Workshop on Open Challenges in Online Social Networks, pp. 21-25. 2021.