ULMFiT for Twitter Fake News Spreader Profiling

Notebook for PAN at CLEF 2020

[1]H. L. Shashirekha, [2]F. Balouchzahi
Department of Computer Science, Mangalore University, Mangalore - 574199, India
[1]hlsrekha@gmail.com, [2]frs_b@yahoo.com

Abstract. 21st century is named as the age of information technologies. Social applications such as Facebook, Twitter, Instagram, etc. have become a quick and huge media for spreading news over the internet. At the same time, the ability for the wide spread of news that is of low quality with intentionally false information is creating havocs causing damage to the extent of losing lives in the society. Such news is termed as fake news and detecting the fake news spreader is drawing more attention these days as fake news can manipulate communities' minds and also social trust. Until date, many studies have been done in this area and most of them are based on Machine Learning and Deep Learning approaches. In this paper, we have proposed a Universal Language Model Fine-Tuning model based on Transfer Learning to detect potential fake news spreaders on Twitter. The proposed model collects wiki text data to train the Language Model to capture general features of the language and this knowledge is transferred to build a classifier using fake news spreaders dataset provided by PAN 2020 to identify the fake news spreader. The results obtained on PAN 2020 fake news dataset are encouraging.

CEUR-WS.org/Vol-2696/paper_126.pdf

1 Introduction

In this era, social media is overwhelming the lives of people and people are sharing various information using different platforms of social media such as Google+, Facebook, WhatsApp and Twitter [1]. The velocity of news spreading on internet is highly increasing due to the availability of various social media platforms and pocket friendly mobile data packs. Social media has become more attractive especially for the younger generation mainly because of the inherent benefits of fast dissemination of information and easy access to the information [2]. At the same time, the ability for the wide spread of news that is of low quality with intentionally false information is creating havocs causing damage to the extent of losing lives in the society [3].

Two major concepts of fake news are veracity and intention. Veracity is about the news that includes some information and the authenticity of that content is possible to be verified as they are. For example, in case of a news about earthquake in Japan, the probability of this news being true is higher but it is a challenge to prove that it is fake or not. Intention refers to the goal of spreader to use false information intentionally to mislead the reader.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-25 September 2020, Thessaloniki, Greece

Fake news is not a new challenge as people have been exposed to propaganda, tabloid news, and satirical reporting since ages. But nowadays, the heavy dependence on the internet, trending stories on social media, new methods of monetizing content, etc., have been found to rely on information without using trustworthy traditional media outlets [4]. Fake news is hazardous since it is spread to manipulate readers' opinions and beliefs [5]. Hence, detecting fake news spreaders becomes very much important in today's scenario and is gaining popularity day by day as users play a key role in creating and sharing incorrect or false information intentionally or accidently [6]. In spite of many systems including automatic detection systems and human based systems, detection of fake news spreaders is still a challenging task [7].

Detecting fake news spreaders in Twitter can be modeled as a typical binary Text Classification (TC) problem that labels a given news spreader as fake or genuine. TC is a Supervised Machine Learning (ML) technique that automatically assigns a label from the predefined set of labels to a given unlabelled input. It has wide applications in various domains, such as target marketing, medical diagnosis, news classification, and document organization [8]. There are several popular approaches for TC in general and for fake news spreader profiling in particular. In this paper, we propose a Universal Language Model Fine-Tuning (ULMFiT) model for fake news spreader detection based on Transfer Learning (TL).

1.1 Transfer Learning

TL is generally known as one of the novel inventions in the field of Deep Learning and Computer Vision. Conventionally, in ML every model is built from the scratch using a specific dataset. However, a model based on TL approach uses the knowledge obtained from building one model called as a source model in building another model called as target model. The former model is called as source task and later the target task. While the source task uses one dataset called as source dataset to build/learn the source learning system or source model, target task uses the knowledge obtained in building the source model along with the target dataset used for fine tuning the target model. For example, the source model can be a Language model (LM) that represents the general features of a language, target model can be TC, source dataset can be Wikipedia text and the target dataset can be fake news [9]. LM is a probability distribution over word sequences in a language and introduces a useful hypothesis space for many other NLP tasks [10]. As the knowledge obtained in building the source model is transferred to build the target model, learning is named as Transfer Learning. Figure 1 illustrates the difference between conventional ML and TL. After the introduction of TL, LM has drawn more attention as it acts as an informative knowledge of a language.

1.2 ULMFiT

ULMFiT is a model based on TL and can be used for many NLP tasks such as TC and NER [9]. It uses the knowledge of LM as source model and then fine tunes the target model using the task-specific data or target dataset. Figure 2 represents architecture of ULMFiT. It includes 3 steps i) pre-training LM using large corpus like Wikipedia to capture the high-level language features and the resultant model is called as pre-trained LM ii) fine-tune the target model using pre-trained LM and task-

specific or target dataset iii) final model which accepts the test/unlabelled data to assign a label.

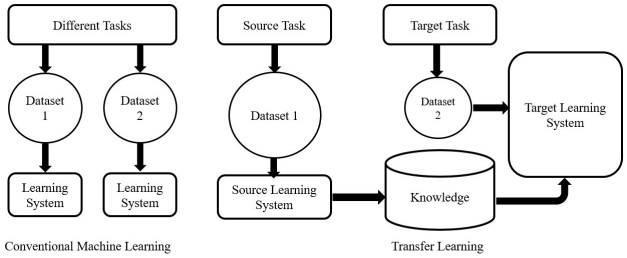

Figure 1. Conventional Machine Learning versus Transfer Learning

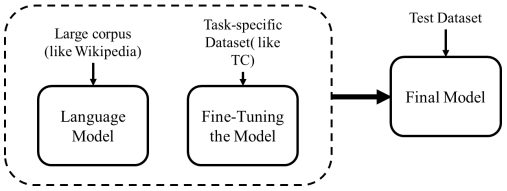

Figure 2. Architecture of ULMFiT

The advantage of TL is, when a given dataset is too small to train a learning model the knowledge obtained in a pre-trained LM on a source dataset can be transferred to the target task, resulting in the improvement of the target model even when the source and target datasets have different distributions or features [9] [11][12].

The rest of the paper is organized as follows. Section 2 gives the related work followed by the proposed methodology in section 3. While section 4 describes the experiments and results, section 5 gives the conclusion of the paper.

2 Related Works

In spite of the availability of many automated tools and techniques for the detection of fake news spreaders, it is still a challenging task. Some of the relevant works are mentioned below:

An Artificial Neural Network model for Language Identification task for Indian native Languages namely Tamil, Hindi, Kannada, Malayalam, Bengali and Telugu written in Roman script has been explored by Hamada et. al. [1]. The data sets used in task are collection of comments from different regional newspapers and Facebook pages. They obtained an accuracy score of 35.30 %. The same authors also obtained accuracies of 47.60% and 47.30% respectively in another work using ensemble classifier made of multinomial Bayes, SVM and random forest tree [13]. Francisco et. al. [14] proposed Low Dimensionality Representation (LDR) for language variety identification and has applied LDR to the age and gender identification task at the

PAN Lab at CLEF. The results they obtained are competitive with the best performing teams in the author profiling task.

Shu et. al. [2] constructs a real-world dataset by measuring users trust level of "experienced[1]" and "native[2]" users on fake news. They have performed a comparative analysis of explicit and implicit profile features between these user groups, which reveals their potential to differentiate fake news. Shu et. al. [3] have explored the fake news problem from a data mining perspective, including feature extraction and model construction and have reviewed different approaches for fake news detection. Bilal et al. [5] presents an approach based on a combination of emotional information from documents using a deep learning network. The authors used one dataset including trusted news (real news) created from English Gig word corpus and another dataset is a collection of news from seven different unreliable news sites as false news and have reported an F1 score of 96%. A Bot detection approach using behavioral and other informal cues is proposed by Andrew et. al. [15]. They have used random forest classifier and a gradient boosting classifier and also applied a hyper parameter optimization on over 476 million revisions that has been collected from Wikipedia articles. They have reported the model performance as 88% precision and 60% recall.

EmoCred model based on LSTM neural network proposed by Anastasia et. al. [16] incorporates emotional signals to differentiate between credible and non-credible claims. It accepts word embeddings as input from claims and a vector of emotional signals. The authors used Politifact[3] that contain the text of the claims, the speaker, and the credit rating of each claim. Six different credibility ratings: true, mostly true, half true, mostly false, false, and pants-on-fire has been combined into two classes as true and false and obtained 61.7% F1 score for generating the emotional signals. “DeClarE” is an automated end-to-end neural network model proposed by Kashyap et. al. [17]. They capture signals from external evidence articles and model joint interactions between various factors like the context of a claim, the language of reporting articles, and the trustworthiness of their sources. Their model was evaluated on Snopes[4], Politifact [5], and a SemEval Twitter rumor dataset and obtained F1 scores of 79% and 68% for Snopes and Politifact respectively and a macro accuracy score of 57% for SemEval dataset.

3 Methodology

An overview of the proposed fake news spreader detection model is described in Figure 3. The model constructed using the state-of-the-art ULMFiT architecture developed by Howard et. al. [10] consists of pre-training the LM and then fine-tuning the fake news spreader detection model by using the pre-trained LM and fake news spreader dataset provided by PAN2020. Two separate models are constructed to detect the fake news given in English and Spanish. Inspired by Stephen et. al. [18],

[1] Users who are able to recognize fake news items like false
[2] Users who are more likely to believe fake news
[3] It is a fact-checking website where the credibility of different claims is investigated.
[4] www.snopes.com
[5] www.politifact.com

LM and Target classifier are created using text.models module from fastai library. This module implements the encoder for an ASGD Weight-Dropped LSTM (AWD-LSTM) which can be plugged in with a decoder to create an LM and also with some classifying layers to create a text classifier.

AWD-LSTM is a regular LSTM to which several regularization and optimization techniques are applied and built layer by layer by grabbing a PyTorch neural network model [9]. Its architecture as described by Howard and Ruder [10] consists of a word embedding of size 400, 3 layers and 1150 hidden activations per layer. The AWD-LSTM has been dominating the state-of-the-art language modeling and many studies on word-level models incorporate AWD-LSTMs. It also has shown noticeable results on character-level models [18].

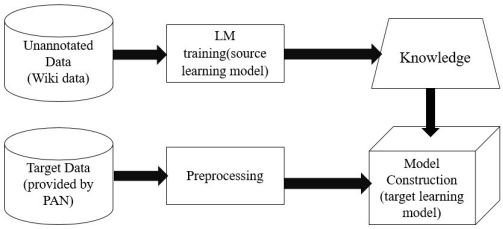

Figure 3. Overview of ULMFiT for Twitter fake news spreader profiling

3.1 Training LM (Source Learning Model)

LM also called as source learning model is trained on the source data collected from English/Spanish Wikipedia. Source data set usually is an unannotated data set that contains general domain texts to train LM to obtain general features like grammar of the language. A sufficiently large English/Spanish text data are collected from Wikipedia to create a source dataset of English/Spanish language respectively and LM is trained to learn the general features of the language. Wikipedia articles that were available in the month of January 2020 are collected in xml format and then the sentences are extracted from the raw text using WikiExtractor[6] module. Once the source model completes its learning the knowledge thus learned is used to build the target task of fake news spreader detection. The knowledge can also be saved for future use for other English/Spanish NLP applications. Details of source dataset for both the languages are given in Table 1.

3.2 Target Model

The target model is created using the knowledge obtained from LM followed by fine-tuning the model using the target dataset. The pre-trained LM is used to train target task data for various cycles to fine-tune the knowledge based on target task**.** Target dataset is the labeled data used for classification tasks which is provided by PAN for registered users only. The dataset consists of 300 XML files in a folder per language (English, Spanish) [19]. Each folder contains:

[6] https://github.com/attardi/wikiextractor

- An XML file per author (Twitter user) consisting of 100 tweets each and the name of the XML file corresponds to the unique author id.
- A truth.txt file with the list of authors and ground truth.

The details of the dataset provided by PAN are given in Table 2. Target data is preprocessed and then used for fine-tuning the classification task. Preprocessing involves tokenization, removing punctuations and stop words, lemmatization and removing other unwanted characters. Emojis are small images used to express emotion and are useful in text analysis [13]. Hence, they are converted to respective words or phrases and those words or phrases are treated similar to content bearing words.

Table 1. Details of source dataset

Language	*No. Articles*	*No. Sentences*	*No. Words*
English	63341	2050239	68011619
Spanish	68490	1531438	64530355

Table 2. Details of target datasets provided by PAN

Language	*No. of Authors*	*No. of tweets per author*	*No. of class 0 data*	*No. of class 1 data*
English	100	300	150	150
Spanish	100	300	150	150

4 Experimental results

As per PAN 2020 rules for submitting software in Virtual Machine (VM), learning model has to be first constructed locally and saved followed by loading the model in PAN VM and finally submitting the model through TIRA Integrated Research Architecture submission system [20]. ULMFiT model is created using Google Colab[7] as it requires GPU and higher RAM size in learning cycles.

The proposed model was evaluated through PAN submission system and the performance of model was made available by the task moderator. Model's runtime reported by PAN is 00:35:48 (hh:mm:ss). Almost half of this time is spent on loading the model using fastai library and rest for predictions. Details of results obtained by the proposed model are given in Table 3. The proposed model resulted with 64% accuracy for Spanish and 62% for English language data.

[7] https://colab.research.google.com/

Table 3. Performance of the proposed model

Language	*Accuracy (%)*
English	62 %
Spanish	64 %

6 Conclusion

This paper presents ULMFiT model for profiling fake tweet spreaders based on Transfer Learning approach. The proposed model is initially trained on a general domain English/Spanish data collected from Wikipedia to build an LM model, and then the acquired knowledge is transferred to build the fake news spreader detection task as the target model. The model resulted with 64% accuracy for Spanish and 62% for English language data. Further, the data collected from Wikipedia and LM can be used for any other English/Spanish NLP task.

References

1. Nayel Hamada A., and H. L. Shashirekha. "Mangalore University INLI@ FIRE2018: Artificial Neural Network and Ensemble based Models for INLI". In FIRE (Working Notes), pp. 110-118, 2018.
2. Shu Kai, Suhang Wang, and Huan Liu. "Understanding User Profiles on Social Media for Fake News Detection". In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 430-435, 2018.
3. Shu Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. "Fake News Detection on Social Media: A Data Mining Perspective". ACM SIGKDD Explorations Newsletter 19, No. 1, pp. 22-36, 2017.
4. Haber Morey. "The Real Risks of Fake News". Risk Management 64, no. 3, pp.10-12, 2017.
5. Ghanem Bilal, Paolo Rosso, and Francisco Rangel. "An Emotional Analysis of False Information in Social Media and News Articles". ACM Transactions on Internet Technology (TOIT) 20, no. 2, pp. 1-18, 2020.
6. Giachanou Anastasia, Esteban A. Ríssola, Bilal Ghanem, Fabio Crestani, and Paolo Rosso. "The Role of Personality and Linguistic Patterns in Discriminating Between Fake News Spreaders and Fact Checkers". In International Conference on Applications of Natural Language to Information Systems, Springer, Cham, pp. 181-192, 2020.
7. Vo Nguyen, and Kyumin Lee. "Learning from Fact-checkers: Analysis and Generation of Fact-checking Language". In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 335-344, 2019.
8. Aggarwal Charu C., and Cheng Xiang Zhai. "A Survey of Text Classification Algorithms. In Mining Text Data". Springer pp. 163-222, Boston, MA, 2012.
9. Faltl Sandra, Michael Schimpke and Constantin Hackober. "Ulmfit: State-Of-The-Art in Text Analysis". Seminar Information Systems (WS18/19), 2019.

10. Howard Jeremy, and Sebastian Ruder. “Universal Language Model Fine-Tuning for Text Classification”, arXiv preprint arXiv: 1801.06146, 2018.
11. Semwal Tushar, Promod Yenigalla, Gaurav Mathur, and Shivashankar B. Nair. “A Practitioners Guide to Transfer Learning for Text Classification Using Convolution Neural Networks”. In Proceedings of the 2018 Society for Industrial and Applied Mathematics (SIAM) International Conference on Data Mining, pp. 513-521, 2018.
12. Pan Sinno Jialin, James T. Kwok, and Qiang Yang, “Transfer Learning via Dimensionality Reduction”, In AAAI, vol. 8, pp. 677-682, 2008.
13. Nayel, Hamada A., and H. L. Shashirekha. “Mangalore-University@ INLI-FIRE-2017: Indian Native Language Identification using Support Vector Machines and Ensemble approach”. In FIRE (Working Notes), pp. 106-109, 2017.
14. Rangel Francisco, Marc Franco-Salvador, and Paolo Rosso. “A Low Dimensionality Representation for Language Variety Identification”. In International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Cham, pp. 156-169, 2016.
15. Hall Andrew, Loren Terveen, and Aaron Halfaker. “Bot Detection on Wikidata Using Behavioral and Other Informal Cues”. Proceedings of the ACM on Human-Computer Interaction, Vol. 2, No. CSCW, Article 64, November 2018.
16. Giachanou Anastasia, Paolo Rosso, and Fabio Crestani. “Leveraging Emotional Signals for Credibility Detection”. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 877-880, 2019.
17. Popat Kashyap, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. “DeClarE: Debunking Fake News and False Claims Using Evidence-Aware Deep Learning”. arXiv preprint arXiv: 1809.06416, 2018.
18. Merity Stephen, Nitish Shirish Keskar, and Richard Socher. “Regularizing and Optimizing LSTM Language Models”. arXiv preprint arXiv: 1708.02182, 2017.
19. Rangel F., Giachanou A., Ghanem B., and Rosso P. “Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter”. In: L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol (eds.) CLEF 2020 Labs and Workshops, Notebook Papers, CEUR Workshop Proceedings, 2020.
20. Potthast Martin, Tim Gollub, Matti Wiegmann, and Benno Stein. “TIRA Integrated Research Architecture”. In Information Retrieval Evaluation in a Changing World, Springer, Cham, pp. 123-160, 2019.