Journal of Universal Computer Science, vol. 27, no. 10 (2021), 1128-1148
submitted: 15/03/2021, accepted: 27/7/2021, appeared: 28/10/2021 CC BY-ND 4.0

Adapting Pre-trained Language Models to Rumor Detection on Twitter

Hamda SLIMI
(ENSI, Manouba University, Manouba,
Laboratory of Computer Science for Industrial Systems (LISI), INSAT, Carthage University,
Tunis, Tunisia,
https://orcid.org/0000-0002-8494-6551, hamda.slimi@ensi-uma.tn)

Ibrahim BOUNHAS
(Laboratory of Computer Science for Industrial Systems (LISI), INSAT, Carthage University,
Tunis, Tunisia,
https://orcid.org/0000-0002-6310-7062, bounhas.ibrahim@gmail.com)

Yahya SLIMANI
(ISAMM, Manouba University, Manouba,
Laboratory of Computer Science for Industrial Systems (LISI), INSAT, Carthage University,
Tunis, Tunisia,
https://orcid.org/0000-0002-4684-3703, yahya.slimani@gmail.com)

Abstract: Fake news has invaded social media platforms where false information is being propagated with malicious intent at a fast pace. These circumstances required the development of solutions to monitor and detect rumor in a timely manner. In this paper, we propose an approach that seeks to detect emerging and unseen rumors on Twitter by adapting a pre-trained language model to the task of rumor detection, namely RoBERTa. A comparison against content-based characteristics has shown the capability of the model to surpass handcrafted features. Experimental results show that our approach outperforms state of the art ones in all metrics and that the fine tuning of RoBERTa led to richer word embeddings that consistently and significantly enhance the precision of rumor recognition.

Keywords: Twitter, Rumor Detection, RoBERTa, Pre-trained Language Models, Fine Tuning
Categories: H.3.3, I.2.7
DOI: 10.3897/jucs.65918

1 Introduction

The evaluation of information credibility relates to multiple areas such as communication, psychology, information science, marketing, and interdisciplinary human-computer interaction efforts (HCI). Each field has analyzed the construct of credibility assessment and its functional meaning using a variety of approaches [Rieh and Danielson, 2007, Bounhas et al., 2015b].

Millennials and post millennials rely mainly and oftentimes solely on online sources to provide them with newsworthy content. However, fake news is trending, which showcases the state of social media credibility. This phenomenon has jeopardized to some extent the 2016 US elections. Facebook CEO Mark Zuckerberg stated that during these elections nearly 126 million Americans saw Russian backed politically oriented content in their Facebook feed. Furthermore, a study conducted by Pew Research Center has shown that 51% of adults in the US consider news shared in social media as inaccurate. The distrust is due in part to the rumors often shared across social media platforms and the absence of gatekeepers to obstruct their propagation.

A rumor is defined according to the merriam-webster [1], as *"talk or opinion widely disseminated with no discernible source"*. In the scope of our study, we define rumor in social media as an information that is presented in a manner that attracts the reader's attention and incites them to share it, although, its content is unverified and its truth value can be questioned.

Such characteristics coupled with the ease of sharing and the absence of safeguards to verify and approve content before it is published, facilitate a fast dissemination of unverified information. In online platforms such as social media, more specifically in Twitter, newsworthy content is been propagated at an overwhelming rate [Kwak et al., 2010]. Amongst these information lies a portion of rumorous content that was published in a malicious intent to alter the public opinion and deceive it [Badawy et al., 2018, Morgan, 2018, Bradshaw and Howard, 2017]. Even the recent pandemic of COVID19 has seen immense propagation of false or unverified information that target substances that may cure patients or far-fetched explanation of the origin of the virus and how it behaves [Tasnim Samia, 2020].

Some text related problems require the model to have a language understanding capability in order to solve them. Rumor detection in Twitter is one of them, where it can be noticed that Natural Language Processing (NLP) can provide powerful tools to distinguish unverified content from trustworthy one. Text classification in NLP has seen various real world implementations whether in fraud [Fisher et al., 2016], bot detection [Gilani et al., 2016] or sentiment analysis [Kanakaraj and Guddeti, 2015, Cenni et al., 2017]. Rumor detection in Twitter is one of the fields were NLP was applied to enable an understanding of rumorous content [Li et al., 2020, Su et al., 2020, Hamidian and Diab, 2019]. Recent advances in the field of vector representations of words has witnessed the introduction of pre-trained language models (PLMs) [Liu et al., 2019, Devlin et al., 2019, Zaib et al., 2020, Yang et al., 2019]. They excelled at text classification tasks and surpassed previous techniques such as Word2Vec [Mikolov et al., 2013] and Glove [Pennington et al., 2014] in sentiment analysis [Sun et al., 2019, Xu et al., 2019], event detection [Chakma et al., 2020] and other fields [Zhu et al., 2019, Wang et al., 2019].

Language models when pre-trained on large scale unlabeled data are able to output embeddings that encompass a universal language representation. These models are suitable and efficient for downstream NLP tasks once they are fine tuned on the target task. In the context of rumor detection, the use of PLMs is still in its early investigations. To the best of our knowledge, only a work by [Han et al., 2019] has explored ELMo PLM [Peters et al., 2018] to augment rumor datasets and improve the performance of rumor detection models. Furthermore, approaches seeking to detect rumor or fake news in other platforms (Reddit, News Outlets, Facebook) are not within the scope of our study [Jwa et al., 2019, Majumder and Das, 2020, Baruah et al., 2020].

The lack of approaches exploring PLM in rumor detection on Twitter and their potential in text classification tasks is the main incentive for our approach. The proposed solution introduced novel tweet processing steps and a set of PLM-based data augmentation techniques that enable efficient fine-tuning of the RoBERTa model [Liu et al., 2019] to the task of rumor detection. The fine-tuned RoBERTa Pre-trained Language

[1] https://www.merriam-webster.com/

Model (RoPLM) can provide rumor-sensitive word embedding for tweets, enabling better performance in recognizing rumor-propagating tweets.

The rest of the paper is organized as follows. Section 2 presents a brief glance into previous approaches of rumor detection in Twitter. In Section 3 we present the proposed approach which is experimented in Section 4 where we present our results and major findings. Finally, Section 5 concludes the paper and suggests some future research directions.

2 Related Works

Mankind has sought credible sources that provide reliable and trustworthy information since the dawn of time. However, distinguishing between rumor-promoting users and trustworthy ones remains a challenging task. Numerous solutions were proposed to evaluate content and source credibility both offline (documents, historical events, etc.) [Bounhas et al., 2015b] and online (social media, websites) [Castillo et al., 2011, Hamdi et al., 2020, Jin et al., 2014, Slimi et al., 2019b, Wu et al., 2016].

Indeed, computer credibility matters when it acts as a knowledge source, a decision aiding tool or report on work that was performed by humans. They also point out that computers lose credibility when they provide information that the user deems as false and they gain credibility when the content is perceived as correct. Thus, a source consistency in providing reliable and trustworthy content is a major factor in determining its credibility [Fogg and Tseng., 1999].

To restore trust to social media platforms, numerous approaches were proposed by researchers to detect trustworthy content and sources of information. For the scope of our study, we will focus on solutions that were developed for Twitter.

We take a quick and brief glance at some papers that set the basis for credibility evaluation on Twitter. [Castillo et al., 2011] is the most cited work in the field of credibility evaluation in Twitter. They relied on a set of hand-crafted features and a set of classifiers to detect newsworthy tweets and evaluate their credibility. [Gupta et al., 2014] are one of the first to tackle the assessment of tweets credibility in real time. In [Jin et al., 2014], authors approached credibility assessment of tweets from a social and graph-based features angle, where they established a hierarchical network structure that describes event, sub-events and messages levels and explores them through the propagation of credibility values to determine tweet and event veracity.

In recent works, the main focus was to explore either machine learning or deep learning models to detect latent features in tweet textual content and use them to determine if a tweet propagates a rumor or not. This task has been handled in two manners: i) as a four-class problem (unverified/non-rumor/true rumor/false rumor) which seeks to evaluate information credibility ;and, ii) as a binary problem (rumor/non-rumor) which aims to detect rumor content independently of its truth value. In the proposed approach, we adopt the latter.

2.1 Machine Learning-based Approaches

In [Mendoza et al., 2010], authors analyzed the impact of a crisis event on the flow of information in twitter. They focused on the propagation of false rumors. They stated that tweets corresponding to rumors propagate differently than confirmed truth since rumors tend to be questioned more often. [Zou et al., 2015] introduced a generative probabilistic model to enable the real-time prediction of credible events. Their approach predicts the credibility label of an event based only on few tweets without relying on the whole set of tweets describing the event. The model relies on an update function that modifies the value of credibility at each influx of new tweets. Each tweet is characterized by a set of content features and its community interactions. In fact, they consider a feedback on a message (tweet) positive if the total number of retweets and likes is greater than a predefined threshold. They found that the model Precision improved as the number of tweets increased. [Kwon et al., 2017] split the events into various time windows ranging from hours to months. Then, they investigated the impact and consistency of features in detecting rumors over time. They confirmed that structural and temporal features allowed the recognition of rumors in longer time windows. Nevertheless, these features are not available during the early stages of rumor propagation. Therefore, to ascertain the veracity of rumors promptly, the authors suggest using user and linguistic features.

2.2 Deep Learning-based approaches

Deep learning-based solutions have attracted researchers' attention in recent years due to their ability to discern latent features within the data (deep features) and perform well contrary to classical machine learning approaches that usually rely on hand-crafted features. Indeed, fake news and rumor detection were one of those fields where we can distinguish several prominent solutions that often surpass machine learning-based approaches in the task of evaluating information credibility. However, for the scope of our study, we will discuss and solely focus on approaches that address rumor detection on Twitter. Thus, the evaluation of news articles [Ma et al., 2018] and Reddit posts credibility [Gorrell et al., 2019] will not be included. We justify such a selection criterion to the nature of our proposed approach that relies on binary class labeled rumor datasets of tweets.

Usually, stance classification and rumor detection are handled as two separate tasks [Gorrell et al., 2019]. However, [Alkhodair et al., 2020] proposed a model that handles the aforementioned tasks jointly which required the extraction of commonly shared features from the two tasks. Their approach mainly relies on Reccurent Neural Networks (RNN) and handles the rumor detection on the event level instead of the tweet level. RNN architectures were also adopted by [Alkhodair et al., 2020] where they proposed a model that jointly embeds the tweets using word2vec and classifies them as rumorous or not. By evaluating their approach on various datasets they have proven the ability of their model to detect unseen emerging rumors.

In [Ahsan et al., 2019] approach, authors proposed a deep learning model which employs CNN to detect fast-paced rumors by relying on two sets of features; hand-crafted features and tweet embedding. They simulate the flow of breaking news rumor by varying the topics included in each of the training and test sets. The main drawbacks of their approach resides in their evaluation scheme where they opted for a split of 70/30 instead of a cross-validation. They also only evaluated the model performance based on accuracy which is a measure that can be misleading in unbalanced datasets such as Pheme where the model evaluation results are skewed by the majority class.

To the best of our knowledge, pre-trained language models were explored in fake news detection [Zhang et al., 2020, Wu and Chien, 2020, Antoun et al., 2020] but not in the context of rumor detection in Twitter. The difference between these two tasks is that Fake news detection consists of the prediction of the chances of a particular piece of information (news article, reviews, posts, etc.) being intentionally deceptive, while rumor detection tries to distinguish between verified and unverified information (instead of true or false) and the unverified information may turn out to be true or false, or may remain unresolved.

In the proposed approach, we harness the potential of pre-trained language model RoBERTa through the exploration of transfer learning on rumor dataset to adapt the PLM to the task of rumor detection in Twitter. We also aim to answer the following research questions:

- **RQ1:** what are the factors that contribute to a proper representation of the rumor detection problem thus enabling a better performance by the fine tuned models?
- **RQ2:** can rumor sensitive embedding obtained through fine tuned PLM surpass hand crafted features in the task of the detection of emerging and unseen rumor?

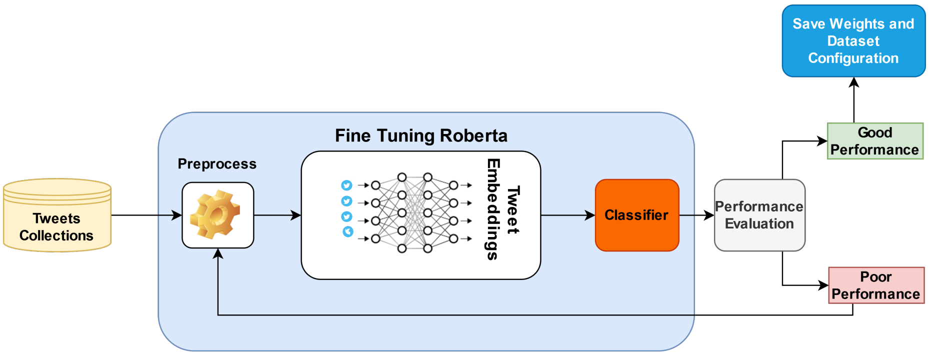

Figure 1: Architecture of the Proposed Approach

3 Proposed Approach

In our proposed approach, we tackle rumor detection in Twitter (TRD). Previous works [Castillo et al., 2011, Gupta et al., 2014, Thakur et al., 2018] have shown that tweet features e.g. number of retweets, number of likes and user features such as followers and followees count, etc.., may contribute to identifying rumor from verified information. However, these approaches were surpassed by deep learning-based approaches that do not rely on hand crafted features but instead resort to discerning hidden latent features from tweets text. In this section, we present an approach that explores the ability of RoBERTa [Liu et al., 2019] in adapting to TRD task when it is exposed and trained on sufficient data describing the task. The proposed approach explores the fine tuned model to output rich and task sensitive word embedding to detect unseen rumor tweets. It is composed of three components as illustrated in Figure 1. A preprocessing component, a fine tuning component and a classification component, where the preprocessing steps are reiterated when the results are not satisfying.

3.1 Tweets Preprocessing

On the one hand, tweets at their raw state are noisy [Kumar and Harish, 2018]. The use of colloquialism and the presence of linguistic noise, emojis, links, and hashtags hinders the

Name	# tweets	Class Distribution (%)		# topics	Annotation Type
		Rumor	Non Rumor		
Pheme [Zubiaga et al., 2016b]	6425	37.39	62.61	9	Tweet Level
Twitter 15(Tw15) [Ma et al., 2018]	499	48.90	51.10	N/A	Tweet Level
Twitter 16 (Tw16) [Ma et al., 2018]	336	45.83	54.17	N/A	Tweet Level
CredBank [Mitra and Gilbert, 2015]	14052	6.33	93.67	6	Topic Level

Table 1: Description of datasets used in the fine tuning of RoPLM

ability to represent tweets consistently. On the other hand, pre-trained language models require text that is clear of noise to output a representative word embedding. Thus, we apply a set of preprocessing steps to tweets to enable and promote for a richer word embedding. Preprocessing steps are as follows:

- **Tweet Cleaning:** removing stop words, link, hashtags, four letter words, and also replacing emojis with their corresponding word e.g. smiley emoji is replaced by the word happy.
- **Labels Unification (LU):** topic level annotation refers to dataset where tweets of the same topic have the same label (rumor/non-rumor). Whereas tweet level annotation refers to datasets where each tweet has its own label, independently from the topic that it belongs to. When combining datasets that adhere to different annotation schemes (tweet level, topic level), a unification of labels is applied. By selecting a single random tweet from each topic of the dataset (topic level labelling) it leads to a single tweet per a topic. This step is needed because tweets of the same topic should not necessarily have the same label (rumor/non-rumor).
- **Tuning Class Distribution(TCD):** In anomaly, rumor and fraud detection fields, the target class is usually the minority class. Such class distribution hinders the ability of the model to provide precise results due to the lack of sufficient instances describing the target class [Chawla et al., 2003]. By tuning class distribution we attempt to remedy this issue. We either impose a balanced class distribution or make the rumor class the majority one. The main driver for this preprocessing is to determine the impact of class distribution on the fine tuning of RoBERTa [Liu et al., 2019] model and on the classification task.

3.2 Tweets Embedding Using RoBERTa

Machine learning classifiers require a numerical representation of text data in order to derive insight from the information they encompass. Indeed, several approaches can represent text in a numerical fashion like Word2vec [Mikolov et al., 2013], Glove [Pennington et al., 2014], and other feature learning approaches. However, they output a representation which fails at modeling polysemous words and syntactic structures. This context-independent word embeddings can hinder the performance of machine learning classifiers.

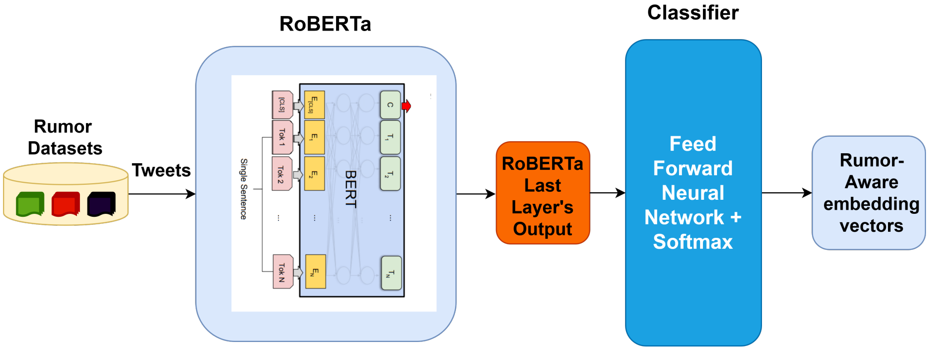

Figure 2: Process of Extracting Rumor Sensitive Embedding from Tweets

Pre-trained Language Models is an emerging new set of approaches that seeks to overcome the shortcomings of previous feature learning approaches. Pre-trained language models were trained extensively on large unlabeled text datasets on high-end machines equipped with GPUs and TPUs to account for the memory space that these models require. Accordingly, besides relying on a transformer based multi-layer network, they come to provide a concise and context-sensitive representation of the words reflecting hence a decent (or a deep semantic) understanding of the language.

Usually, word representations such as Word2Vec [Mikolov et al., 2013] are learned through a generalized context and do not provide task-specific information in their embedding. However, the fine-tuning aspect of a language model allows the user to adapt these word embeddings/representations by training on the task-specific dataset which yields embeddings that are context-sensitive and task-specific.

Robustly optimized BERT approach (RoBERTa) [Liu et al., 2019] is a successor pre-trained language model of BERT. Similar to BERT [Devlin et al., 2019], it is based on 12 transformer layers that are trained on a single task instead of two. Whereas BERT is trained on both the masked word and the next sentence prediction tasks RoBERTa was solely trained on the masked word prediction task. However, the latter was improved by varying the candidate word for masking across epochs. Furthermore, the RoBERTa model was trained on 160 GB of text data whereas BERT was trained only on 16 GB of data. Such a difference in the size of the training corpus as well as the training scenario induced an enhancement in the quality of the representation vectors and their sensitivity towards the context thus leading to a richer word embedding.

Lately, pre-trained language models such as RoBERTa [Liu et al., 2019], BERT [Devlin et al., 2019], and DistilBERT [Zaib et al., 2020] attracted the researchers' attention from various fields due to their ability at transfer learning where a swift procedure of fine-tuning can allow the model to take advantage of the learned model at a low computational cost. Thus, they can adapt to NLP problems and provide predictions on previously unknown tasks such as sentiment analysis, question answering, and event nugget detection.

Our main goal is to adapt RoBERTa model to the task of rumor detection by the means of transfer learning. Transfer learning [Pan and Yang, 2009] at its core consists of the ability of a model that was trained for a general purpose task to be tuned for a specific task when it is exposed and trained on sufficient data describing the problem. To this end, we modify the architecture of the RoBERTa model by adding a linear layer on top of the pooled output for tweet classification. The resulting architecture (RoBERTa classification model) is fine tuned on rumor annotated datasets (see Figure 2).

Simple as it may appear, fine tuning on a downstream task like rumor detection has its hardness. The main obstacles that we encountered during tuning of the RoBERTa model on rumor detection task are: the scarcity of high quality datasets, the target class being the minority class and the variance in annotation schemes across datasets. Such constraints impose further preprocessing steps on the data to render it uniform in all datasets. Nevertheless, this preprocessing can reduce the size of these datasets where in some cases tweets are deleted to balance the class distribution. Therefore, we refer to artificial augmentation of data. This augmentation is achieved through novel approaches based on PLM which generate new samples that vary from the original ones.

3.3 Fine Tuning RoPLM for TRD

The art of fine tuning a model requires sufficient knowledge about the studied domain and the Pre-trained Language Model (PLM) and access to domain-specific data. In this section, we will present the process that was employed to fine tune RoBERTa for the task of rumor detection on Twitter.

Fine tuning a PLM requires domain specific data. Therefore, we referred to four standard datasets as shown in Table 1. Both BERT and RoBERTa PLMs were tested. In our experiments we fine tune the RoBERTa model for a number of epochs set initially to 20; we manually stop the training when no progress is made upon the F1-score of the validation set within 5 epochs. As for the batch size and the maximum sequence length, they are fixed at 32 and 256, respectively. For the training of our model, we apply the AdamW optimizer having a learning rate of 25e-6, a weight decay equal to 0.01, a gradient clipping set to 1 and a warmup of 10%. Each of the dataset configurations mentioned previously is fed to the RoPLM model during the fine tuning stage where it adjusts its weights according to tweet labels and the words used within each class (rumor/non rumor). Once the weights are adjusted, the test set is provided to the fine tuned RoPLM, which outputs a rumor-sensitive embedding of its tweets. The tweet embedding vector is represented on 768 dimensions. This process allows the RoPLM to discern the hidden latent features that describe each class of tweets. We investigate the fine-tuning of the RoBERTa model on rumor datasets allowing rumor detection on twitter. Further details about tweets preprocessing and augmentation will be provided in the next section.

3.4 Tweets Classification

Our main goal is the detection of emerging and unseen rumor on Twitter. Although, it is largely dependent on the quality of the embeddings, the important role that the choice of classifier holds can not be ignored. To this end, we choose three distinct classifiers, namely Decision Trees, Random Forest and Support Vector Classifier.

The fine tuning component is the main driver of our approach and to evaluate its quality we feed the embeddings that it outputs to a set of classifiers to determine the ability of these embeddings at representing the rumor class. Furthermore, an evaluation of the embeddings also implies the evaluation of the dataset configuration and the preprocessing steps that were applied. It provides us with insights that enable us to improve mainly the label unification and the tuning of class distribution.

As we can notice in Figure 1, the classifier component determines whether the preprocessing should be reiterated or not. In fact, the quality of the embedding are determined by a subset of scores that the classifier yield. These scores are the micro F1 score, Precision and Recall.

4 Experiments and Results

In this section, we will present the datasets that were used during the fine tuning and the testing of RoBERTa Pre-trained Language Model (RoPLM). We will also discuss the various parameters that impact classifier performance and may lead to a better fine tuning of RoPLM for TRD.

4.1 Dataset Description

In this section, we enumerate gold standard datasets in the domain of rumor detection on Twitter. The class distribution and annotation scheme on each of these datasets are displayed in Table 1. CredBank contains over 60 million tweets but in our approach we take a sample of 14K tweets from the overall dataset. We also combine Twitter 15 and Twitter 16 into one single dataset and we refer to it as Tw1516. As for Pheme [Zubiaga et al., 2016b] it is a version of Pheme with 9 topics. Since five of these topics are in the test set [Zubiaga et al., 2016a], we can not include them in the fine tuning stage. Therefore, they are removed and only the remaining four topics are kept. The resulting dataset contains 624 tweets and we will refer to it as Pheme4.

Topic	# tweets	# Non-rumor	# Rumor
Charlie Hebdo	2079	1621	458
Ferguson	1143	859	284
GermanWings	469	231	238
Ottawa Shooting	890	420	470
Sydney Siege	1221	699	522
Total	5802	3830	1972

Table 2: Test datasets content description

To evaluate the performance of our approach, we use Pheme dataset [Zubiaga et al., 2016a] as it provides tweet level rumor annotations, which is coherent with our perception of the rumor detection task. The dataset is also publicly available here [2]. It contains 5802 tweets describing five breaking news events that occurred in 2015. Each tweet is classified as rumor or non-rumor. For the context of our evaluation, we consider rumors to be non-credible and annotated as (0) and non-rumors as credible and annotated as (1). The content of the dataset is presented in detail in Table 2.

[2] https://figshare.com/articles/Pheme_dataset_of_rumours_and_non-rumours/4010619

In order to determine the pertinence and drawbacks of our proposed approach, a set of comparative baselines local and external are established. Local baselines cover features that we developed such as user and content features. Whereas, external baselines refer to state of the art approaches that rely on Word2Vec or unique approaches to detect rumors on Twitter.

4.2 Comparative Baselines

From the works tackling the rumor detection task, we select a portion as baselines [Alkhodair et al., 2020, Ajao et al., 2018, Ajao et al., 2019] for our proposed approach based on the following criteria. First, the data source should be Twitter since the proposed approach explores only tweet datasets. Second, we select works that have achieved state of the art (SOTA) results on Pheme dataset [Zubiaga et al., 2016a] since it is the gold standard dataset used to experiment our approach. Finally, details about their evaluation protocol are provided to enable a reproduction of the evaluation scenario.

Based on the aforementioned criteria, only [Alkhodair et al., 2020] paper satisfied all of them especially that the authors adopted a 5-fold cross validation which is identical to our evaluation protocol. As for [Ajao et al., 2019], although the authors explored the Pheme [Zubiaga et al., 2016a] dataset and attained competitive results, they did not provide details about their evaluation protocol. Therefore, we omit their work and do not consider it as a baseline.

Apart from external baselines, we also establish a user and content based approach for rumor detection. The latter is used to determine the capacity of tweet embedding obtained from RoPLM TRD at surpassing hand crafted features. In the next section, we provide details about the process of user and content features extraction and show the techniques used in scaling the features.

4.2.1 User Features Extraction

Twitter has substantial information about its users which may portray their behaviour. Furthermore, Twitter provides its users with a variety of ways to interact with the tweets they encounter in their feed. A user may retweet, reply and like a tweet. We believe that these interactions provide valuable information about users' opinion on the tweets they interacted with. Thus, for each user in the dataset, we crawled the basic features that twitter provides about him: his account creation date, the number of tweets he authored, the number of followers and followees,etc. A full list of user features is provided in Table 3. Previous approaches have shown that the inclusion of user features in the assessment of tweet credibility results in a more precise evaluation.

- **Ratio of likes to number of followers**: this value shows the percentage of user followers that liked (Favorited) his tweets to the overall number of his followers. Where disparity is a sign of non-valuable or non appreciated content e.g: a user with 20k followers and an average of 100 likes per tweet illustrates that his content is not well received by his followers.

- **Ratio of retweets to number of followers**: this metric depicts the percentage of followers that retweeted his tweets to the overall number of his followers. It allows us to determine to which degree does the user followers deem his content to be reliable and worthy of sharing.

- **Ratio liked to likes**: this ratio represents the disparity between a user engagement in other tweets and his followers' engagement with his tweets.

In a previous paper [Slimi et al., 2019c], we validated the pertinence of the aforementioned features in user credibility evaluation.

4.2.2 Content Features Extraction

Basic features are provided by twitter and the raw values are considered, e.g. number of likes, user mentions, URLs contained within a tweet and tweet textual content. Then, hand crafted features are created through the combination of multiple basic features or the ressortment to exterior libraries.

Previous approaches settle for the presence or absence of a URL in the tweet and consider it as a binary feature [Castillo et al., 2011]. The proposed feature evaluates the credibility of a URL within a certain topic by considering its frequency as shown in formula (1). A popular URL within a topic is ought to be trustworthy since it has been shared by multiple users. Each user has a URL score based on the average score of URL of his tweets.

For each $URL_i \in URL_T$, where: $URL_T = \{URL_1,\ URL_2, ... \ URL_{n_T}\}$ a set of n URLs of topic T. We use formula (1) to compute URL frequency within a topic T.

$$Freq(URL_i) = \frac{1}{n_T} \sum_{j=1}^{n_T} [URL_i = URL_j] \tag{1}$$

Each tweet Tw contains a set of URL_{Tw} of m URLs. To compute Tweet URL score we refer to formula , 2):

$$Score(URL_{Tw}) = \frac{1}{m_{Tw}} \sum_{i=1}^{m_{Tw}} Freq(URL_i) \tag{2}$$

The impact of the URL feature in tweet credibility evaluation has been discussed and studied thoroughly in a previous paper [Slimi et al., 2019a].

4.2.3 Features Scaling

The process of feature extraction using various methods can result in features that vary in scale. In our proposed approach, we refer to RoBERTa, and hand crafted features to obtain a vector representation of each tweet. Thus, inducing substantial differences in scale between features. In fact, the features that were obtained vary in magnitude and range as shown in Table 3.

These factors impact the precision of the model since most machine learning (ML) approaches resort in their computations to the euclidean distance between two data points. Also, in tree-based ML approaches such as Random forest and Decision trees, the feature importance is impacted by the range of the values where the feature with the highest order of magnitude may rank higher in feature importance even if it does not have a positive impact on the results. These misinterpretation are due to the disparities in features scales which requires the use of a features scaling technique.

In [Singh et al., 2015], it was shown that feature scaling has an impact on classifier performance. Furthermore, minmax scaler was proven to have a positive effect on classification results. Thus, we apply it as a scaling strategy for the proposed approach.

Feature	Min	Max
Nbr favorite	0.000	149783.00
LengthL	17.000	152.00
LengthW	3.000	31.00
URLScore	0.002	0.06
AccountAge	9.000	3197.00
Ratio ff	0.090	417479.00
AvgStats	0.020	402.32
RtPercFoll	0.0	39.55
FvPerFoll	0.0	10.44
FvtedPerFvCot	0.0	96201.00

Table 3: The various features range

The preprocessing module under sklearn in python enables a custom range for features scaling. However, we choose the default range which is [0-1]. MinMax scaler is defined by the formula (3).

$$x_i^{'} = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \tag{3}$$

To obtain proper context and task sensitive embedding, RoPLM requires sufficient data about the task in question. Since rumor detection datasets are small, a data augmentation step is needed. In the upcoming section, we detail how we augmented the datasets using a variety of PLMs.

4.3 Data Augmentation

The resulting datasets lack in both size and variety within each class of tweets. As a remedy to these deficiencies, we resort to data augmentation which is one of the suitable techniques for our situation. Data augmentation is defined by [Van Dyk and Meng, 2001] as methods for constructing iterative optimization or sampling algorithms via the introduction of unobserved data or latent variables. It offers us the ability to increase the number of tweets of each class (rumor/non-rumor) without the need to collect newer tweets. Our datasets (set of tweets) can be augmented through a panoply of techniques. We refer to pre-trained language models [Kumar et al., 2020] in our data augmentation task for two main reasons; i) it offers a flexibility to develop the size of the dataset in a swift manner that does not require going through the whole dataset at once but rather a tweet at a time; ii) when generating a new sample from the original tweet it alters the words while preserving the meaning of the source tweet.

The augmentation process is as follows: the tweet is fed to a PLM, which produces a newer version of the input tweet by either altering some words (substitution) or by inserting new ones without changing the meaning or the semantic value [Kumar et al., 2020]. The generated tweet inherits the label of the original tweet, which preserves the original class distribution of the datasets. This process is repeated four times, and for each time, a distinct PLM is used. The PLMs that were deployed in this process are: BERT (insert, substitute), RoBERTa (substitute), DistilBERT (substitute). The augmented datasets should facilitate the fine tuning process and thus allowing for a better understanding of the rumor detection task.

4.4 Experimental Settings

We performed a 5-fold cross-validation using a set of classifiers; Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT) to evaluate the performance of the proposed approach. Only results for the best performing classifier were shown.

The benefits of cross validation resides in two aspects; the evaluation of new unseen tweets through the test fold and it's ability at reducing the chance of over-fitting the model. The set of metrics that were used to evaluate the performance are the micro values of: F1-Score, Precision and Recall. Per class evaluation was performed to show how well each fine tuning strategy or a set of features is able to recognize rumor from non-rumor tweets.

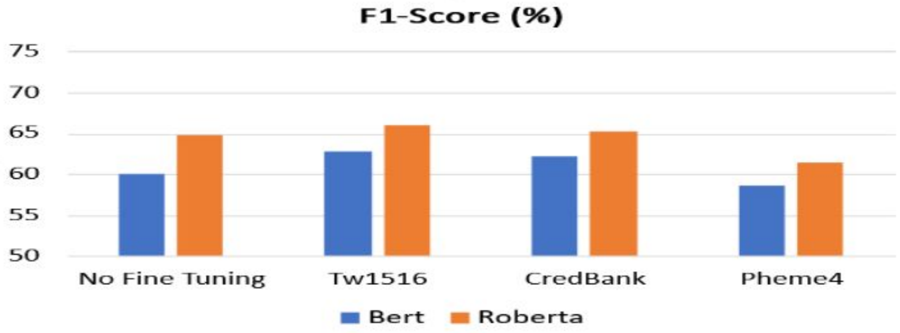

Figure 3: BERT vs RoBERTa

4.5 Results and Discussion

In this section, only the results of the best performing classifier are displayed. Through the obtained results, we seek to assert three constructs of our approach.

1. Which pre-trained language model performs better in the task of rumor detection on Twitter?
2. Does the data augmentation improve the quality of the embedding output by RoPLM for TRD?
3. What dataset configuration yields the best results and why?

When we refer to a combination of two or more datasets we use the sign (+). So Tw15+16 means that we concatenated datasets Twitter 15 and Twitter 16.

First, we tackle the issue of the suitable pre-trained language model for the task of rumor detection. In this step, both BERT and RoBERTa are tested across various contexts to determine the consistency of performance. Primarily, BERT and RoBERTa are tested in their raw state without additional fine tuning. Then, three datasets configurations are used in fine tuning RoPLM namely, CREDBANK, Pheme4 and Tw15+Tw16. As shown in Figure 3, RoBERTa outperforms BERT across all datasets with an average of 3%

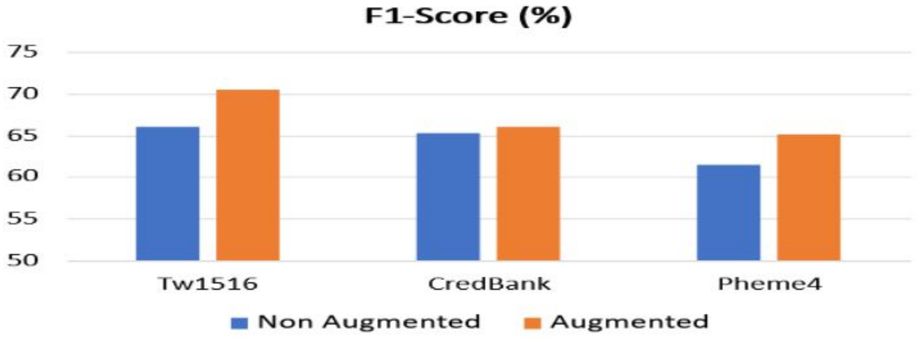

Figure 4: Aug Vs Non Aug datasets

increase in performance. Thus, we adopt RoBERTa as the main pre-trained language model for the task of rumor detection.

To determine if data augmentation may impact the performance of RoBERTa, we test the PLM on three distinct datasets namely, Tw1516, CredBank, Pheme4. Results are displayed in Figure 4. This figure shows that augmented versions of the datasets surpass the non augmented ones. On average the performance of RoBERTa was increased by 3% across various datasets.

In an overall analysis of both Figures 3 and 4, we can notice that the RoBERTa pre trained language model reached an F1 score of 70.4% and 64.8% respectively, with and without fine tuning. This showcases that the choice of dataset, the data augmentation strategy and PLM were suitable for the task of rumor detection.

To further improve the performance of RoPLM, we combine various augmented datasets since more data implies better performance, as it was established by the previous results. Full combination balanced contains Tw1516+Pheme4 with balanced class distribution whereas in full combination refined we added the rumors present in CredBank dataset. In Table 4, we display the results while showing class level metrics since they are necessary to evaluate the ability of the model to recognize rumor.

It can clearly be noticed that recall and Precision for non rumor class are consistently high across various configurations which means that the embedding obtained via a fine tuned RoBERTa model has learned to represent this class accurately. Furthermore, the non rumor class samples are larger than rumor ones across almost all rumor detection datasets (cf Table 1) which promotes for a better Precision and recall for non rumor class. However, our objective is to detect rumor tweets, so misclassifying class rumor samples is a bigger concern than misclassifying non rumor ones.

We take a closer look at the target class scores namely Precision and Recall. In the context of rumor detection, rumor class recall is the determinant metric for the performance of the proposed approach because having false negatives in this class is costly. Classifying a rumor as non rumor can lead to the propagation of unverified and possibly false content. Based on Table 4, we can notice that the highest recall (54.6%) is obtained by ”Full Combination Refined” followed by ”Full Combination Balanced” which obtained (48.4%).

The two least performing combinations are the ones where CredBank was used. We attribute the detrimental effect of CredBank on the results to its class imbalance where rumor only represents 6.3% of the overall dataset (cf Table 1). In fact, in CredBank +

Dataset	Class	Metric (%)			
		Precision	Recall	F1	F1 All Classes (Micro)
Tw1516+CredBank	Rumor	46.9	43.6	41.9	62.2
	Non Rumor	72.3	71.9	70.9	
Tw1516+Pheme4	Rumor	56.8	47.7	50.0	70.3
	Non Rumor	76.2	82.1	78.3	
CredBank+Pheme4	Rumor	29.1	00.8	01.5	65.9
	Non Rumor	66.1	**99.6**	79.4	
Full Combination Balanced	Rumor	59.0	48.4	52.5	71.4
	Non Rumor	76.2	83.3	79.4	
Full Combination Refined	Rumor	**61.8**	**54.6**	**56.3**	73.2
	Non Rumor	**78.8**	82.7	**80.1**	

Table 4: Results of the best performing configurations

Pheme4, the model was unable to recognize rumor class. However, for the non rumor class it reached the best recall (99.6%).

We attribute the increase in performance of full combination refined over the other configurations to the size of the dataset, the number of rumor tweets present in the dataset and the class distribution been balanced.

To further evaluate the ability of the fine tuned RoBERTa model at adapting to rumor detection task, we take a closer look at the micro averaged F1 scores. In Table 4, we can notice that the least performing configuration is Tw1516+CredBank due to CredBank severe class imbalance. A comparison between this score and that of the non fine tuned RoBERTa (see Figure 4) shows that it became less efficient after the fine tuning process. Which showcases that a misrepresentation of task data used in the fine tuning can be detrimental to the model performance.

Achieving 73% F1 score on a data that has not been processed during fine tuning ascertains that PLMs are suitable for the task of TRD. With sufficient data and variance within classes, our proposed approach RoPLM TRD can achieve competitive results on the task of rumor detection. Furthermore, such an ability allows the model to adapt to new emerging and unseen rumor which is the main goal of our research. Our proposed approach promotes for a robust model and rich embedding compared to previous rumor detection approaches. The fact that it is fine tuned on datasets that are completely different from the tested one provides it with the ability to adapt to unseen rumor efficiently and provide representative word embeddings yielding consistent results.

These scores are satisfactory but they still do not surpass SOTA results where [Alkhodair et al., 2020] achieved 79.5% F1 score which is 6.3% higher than our best performing configuration. We believe that the embedding from RoPLM TRD are rich and context sensitive and we justify the gap in F1 score between our approach and the baseline can be to the fact that [Alkhodair et al., 2020] trained their word2vec model directly on Pheme dataset whereas ours did not interact with the Pheme dataset in the fine tuning process. Furthermore, the class imbalance within the Pheme dataset prevents it from benefiting from the obtained contextual word embedding that RoPLM provides. To this end, a final processing of the data is applied by adopting SMOTE [Chawla et al., 2003] an oversampling technique that generates new samples of the minority class until the class distribution is balanced.

Fine tuning Setting	Class	Metrics (%)			All Classes (Micro) (%)
		Precision	Recall	F1	F1
Content And User Features	Rumor	54.9	34.5	39.3	67.2
	Non Rumor	71.8	83.6	76.4	
Content Features Filtered	Rumor	60.5	45.6	51.3	72.4
	Non Rumor	75.0	83.6	78.9	
RoPLM for TRD	Rumor	**79.2**	**83.9**	**80.9**	**80.6**
	Non Rumor	**83.6**	76.7	79.2	
Baseline [Alkhodair et al., 2020]	Rumor	72.8	70.6	71.6	79.5
	Non Rumor	83.3	84.7	83.9	

Table 5: Comparison of feature sets

To illustrate the performance of our proposed approach, we compare it with two baselines that we developed and [Alkhodair et al., 2020] paper that holds the SOTA results in Twitter rumor detection using Pheme. The comparison is shown in Table 5, it can be noticed that content features achieve better results than the combination of content and user features. Which shows that user features in the context of rumor detection can reduce the ability of the model to detect rumor tweets. Furthermore, a comparison of Full Combination Refined results before and after oversampling shows that the model F1 score increased by 6.8% after the oversampling was applied. Thus, the proposed approach outperforms [Alkhodair et al., 2020] work. In fact, we surpass their results substantially in the rumor class, specifically in the key metric Recall where we achieve 83.9% against 70.6%.

4.6 Major Findings

From the aforementioned experiments and results we notice that the fine tuning of a pre trained language model for twitter rumor detection requires sufficient data describing the class rumor. Also, the lack of large datasets can be remedied through the adoption of a PLM-based data augmentation strategy. Furthermore, the annotation scheme that a dataset uses can impact the fine tuning process. Finally, tweet level annotations provide semantically coherent and unambiguous labeling of tweets whereas topic level annotation are more suitable for event credibility evaluation.

5 Conclusion

Twitter as a social platform has altered the manner and speed at which newsworthy information is shared. This induced a sheer amount of data shared at a high velocity rendering the establishment of predefined safeguards that would preemptively prevent the publication of rumor unfeasible. Thus, instead of a prior prevention of rumor sharing we aim at an early detection of rumor. In this work, the proposed approach detects unseen emerging rumor without prior knowledge or interaction with it. It relies on harnessing the potential of pre-trained language models in the rumor-aware representation of tweets text and the augmentation of rumor datasets.

RoBERTa is adapted to the task of rumor detection by fine tuning it on gold standard datasets. Our proposed approach is then tested using machine learning classifiers on datasets that were not present during the fine tuning stage. Such a scenario enables us to test the ability of our model to detect new unseen rumor. Results show that our approach prevails in rumor detection across various topics. A comparison of our approach against baseline approaches that were directly trained and tested on the aimed dataset (Pheme), show that RoPLM for TRD surpasses them across all metrics. The main drawback of PLM based approaches is that they are highly dependent on the quality of data that describes the task. Thus, an inadequate choice of data will result in a misunderstanding of the task during fine tuning stage.

In future works, we consider tackling the detection of sources of fake news in social media where we explore a combination of word embedding and graph features to distinguish non trustworthy users.

References

[Ahsan et al., 2019] Mohammad Ahsan, Madhu Kumari, and TP Sharma. Detection of context-varying rumors on twitter through deep learning. Int. J. Adv. Sci. Technol, 28(1):45–58, 2019.

[Ajao et al., 2019] O. Ajao, D. Bhowmik, and S. Zargari. Sentiment aware fake news detection on online social networks. In ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, May 12-17, pages 2507–2511, 2019.

[Ajao et al., 2018] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. Fake news identification on twitter with hybrid cnn and rnn models. In Proceedings of the 9th International Conference on Social Media and Society, Copenhagen, Denmark, July 18-20, 2018, SMSociety ’18, pages 226—230. ACM, 2018.

[Alkhodair et al., 2020] Sarah A Alkhodair, Steven HH Ding, Benjamin CM Fung, and Junqiang Liu. Detecting breaking news rumors of emerging topics in social media. Information Processing & Management, 57, 2):102018, 2020.

[Antoun et al., 2020] Wissam Antoun, Fady Baly, Rim Achour, Amir Hussein, and Hazem Hajj. State of the art models for fake news detection tasks. In 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), Doha, Qatar, February 2-5, 2020, pages 519–524. IEEE, 2020.

[Badawy et al., 2018] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28-31 Aug, pages 258–265. IEEE, 2018.

[Baruah et al., 2020] Arup Baruah, K Das, F Barbhuiya, and Kuntal Dey. Automatic detection of fake news spreaders using bert. In Conference and Labs of the evaluation Forum (CLEF), Thessaloniki, Greece, September 22-25, 2020.

[Bounhas et al., 2015b] Ibrahim Bounhas, Bilel Elayeb, Fabrice Evrard, and Yahya Slimani. Information reliability evaluation: From arabic storytelling to computer sciences. Journal on Computing and Cultural Heritage, 8 (3):1–33, 2015b.

[Bradshaw and Howard, 2017] Samantha Bradshaw and Philip Howard. Troops, trolls and troublemakers: A global inventory of organized social media manipulation (working paper no. 2017.12). (p. 37). project on computational propaganda Oxford, UK. 2017.

[Castillo et al., 2011] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, March 28 - April 1, WWW ’11, page 675—684. ACM, 2011.

[Cenni et al., 2017] Daniele Cenni, Paolo Nesi, Gianni Pantaleo, and Imad Zaza. Twitter vigilance: a multi-user platform for cross-domain twitter data analytics, NLP and sentiment analysis. In 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, CA, USA, August 4-8, pages 1–8. IEEE, 2017.

[Chakma et al., 2020] Kunal Chakma, Steve Durairaj Swamy, Amitava Das, and Swapan Debbarma. 5w1h-based semantic segmentation of tweets for event detection using bert. In International Conference on Machine Learning, Image Processing, Network Security and Data Sciences, Silchar, India, July 30-31, 2020, pages 57–72. Springer, 2020.

[Chawla et al., 2003] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In 7th European Conference on Principles and Practise of Knowledge discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003, pages 107–119. Springer, 2003.

[Devlin et al., 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Minnesota, USA, Jun 2-7 , 2019, Volume 1, pages 4171–4186. ACL, 2019.

[Fisher et al., 2016] Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. Intelligent Systems in Accounting, Finance and Management, 23(3):157–214, 2016.

[Fogg and Tseng., 1999] B. J. Fogg and Hsiang Tseng. The elements of computer credibility. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Pennsylvania, USA, May 15-20,1999, CHI '99, pages 80—87. ACM, 1999.

[Gilani et al., 2016] Zafar Gilani, Liang Wang, Jon Crowcroft, Mario Almeida, and Reza Farahbakhsh. Stweeler: A framework for twitter bot analysis. In Proceedings of the 25th International Conference Companion on World Wide Web, Montreal, Canada, May 11-15, 2016, pages 37–38, 2016.

[Gorrell et al., 2019] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minnesota, USA, Jun 6-7, pages 845–854. ACL, 2019.

[Gupta et al., 2014] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In Luca Maria Aiello and Daniel McFarland, editors, Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, pages 228–243, Cham, 2014. Springer International Publishing.

[Hamdi et al., 2020] Tarek Hamdi, Hamda Slimi, Ibrahim Bounhas, and Yahya Slimani. A hybrid approach for fake news detection in twitter based on user features and graph embedding. In Dang Van Hung and Meenakshi D'Souza, editors, Distributed Computing and Internet Technology - 16th International Conference, ICDCIT 2020, Bhubaneswar, India, January 9-12, 2020, Proceedings, volume 11969 of LNCS, pages 266–280. Springer, 2020.

[Hamidian and Diab, 2019] Sardar Hamidian and Mona Diab. Gwu NLP at semeval-2019 task 7: Hybrid pipeline for rumour veracity and stance classification on social media. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minnesota, USA, pages 1115–1119, Jun 2019.

[Han et al., 2019] Sooji Han, Jie Gao, and Fabio Ciravegna. Neural language model based training data augmentation for weakly supervised early rumor detection. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Vancouver, Canada, August 27-30, 2019, pages 105–112, 2019.

[Jin et al., 2014] Z. Jin, J. Cao, Y. Jiang, and Y. Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In 2014 IEEE International Conference on Data Mining , Shenzhen, China, Dec 14-17, pages 230–239, 2014.

[Jwa et al., 2019] Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). Applied Sciences, 9(19):40–62, 2019.

[Kanakaraj and Guddeti, 2015] Monisha Kanakaraj and Ram Mohana Reddy Guddeti. Nlp based sentiment analysis on twitter data using ensemble classifiers. In 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), Chennai, India, March 26-28, 2015, pages 1–5. IEEE, 2015.

[Kumar and Harish, 2018] HM Keerthi Kumar and BS Harish. Classification of short text using various preprocessing techniques: An empirical evaluation. In Recent Findings in Intelligent Computing Techniques, pages 19–30. Springer, 2018.

[Kumar et al., 2020] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. arXiv, pages arXiv–2003, 2020.

[Kwak et al., 2010] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web, North Carolina, USA, April 26-30, 2010, pages 591–600, 2010.

[Kwon et al., 2017] Kwon, Sejeong and Cha, Meeyoung and Jung, Kyomin. Rumor detection over varying time windows. PloS one, 12(1):e0168344, 2017.

[Li et al., 2020] Zongmin Li, Qi Zhang, Yuhong Wang, and Shihang Wang. Social media rumor refuter feature analysis and crowd identification based on xgboost and NLP. Applied Sciences, 10(14):4711, 2020.

[Liu et al., 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[Ma et al., 2018] Jing Ma, Wei Gao, and Kam-Fai Wong. Rumor detection on twitter with tree-structured recursive neural networks. In The 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, July 15-20. ACL, 2018.

[Majumder and Das, 2020] Soumayan Bandhu Majumder and Dipankar Das. Detecting fake news spreaders on twitter using universal sentence encoder. In Conference and Labs of the evaluation Forum (CLEF), Thessaloniki, Greece, September 22-25, 2020.

[Mendoza et al., 2010] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In Proceedings of the First Workshop on Social Media Analytics, Washington D.C., USA, July 25, 2010, SOMA '10, page 71—79. ACM, 2010.

[Mikolov et al., 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26: 27th Conference on NIPS 2013, Lake Tahoe, Nevada, United States, December 5-8, 2013, pages 3111–3119, 2013.

[Mitra and Gilbert, 2015] Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, Oxford, UK, May 26-29, 2015, pages 258–267, 2015.

[Morgan, 2018] Susan Morgan. Fake news, disinformation, manipulation and online tactics to undermine democracy. Journal of Cyber Policy, 3(1):39–43, 2018.

[Pan and Yang, 2009] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10):1345–1359, 2009.

[Pennington et al., 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar, October 25-29, 2014, pages 1532–1543, 2014.

[Peters et al., 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, Proceedings of the 2018 Conference of the North American Chapter of the ACL: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 2227–2237. ACL, 2018.

[Rieh and Danielson, 2007] Soo Young Rieh and David R. Danielson. Credibility: A multidisciplinary framework. Annu. Rev. Inf. Sci. Technol., 41(1):307–364, 2007.

[Singh et al., 2015] Bikesh Kumar Singh, Kesari Verma, and AS Thoke. Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification. International Journal of Computer Applications, 116(19):11–15, 2015.

[Slimi et al., 2019a] Hamd Slimi, Ibrahim Bounhas, and Yahya Slimani. Url-based tweet credibility evaluation. In Proceedings of the 16th CS/IEEE International Conference on Computer Systems and Applications (AICSSA), Abu Dhabi, UAE, November 3-7, 2019, 2019a.

[Slimi et al., 2019b] Hamda Slimi, Ibrahim Bounhas, and Yahya Slimani. L'éxploitation des techniques de régression pour l'évaluation de la crédibilité des tweets. In Extraction et Gestion des connaissances, EGC 2019, Metz, France, January 21-25, 2019, volume E-35 of RNTI, pages 327–332. Éditions RNTI, 2019b.

[Slimi et al., 2019c] Hamda Slimi, Ibrahim Bounhas, and Yahya Slimani. Twitter users credibility evaluation based on social graph impression. In Proceedings of the 16th International Conference on Applied Computing (AC), Cagliari, Italy, November 7-9, 2019, 2019c.

[Su et al., 2020] Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. Motivations, methods and metrics of misinformation detection: An NLP perspective. Natural Language Processing Research, 1, 2):1–13, 2020.

[Sun et al., 2019] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1, pages 380–385. ACL, 2019.

[Tasnim Samia, 2020] Mazumder Hoimonty Tasnim Samia, Hossain Md Mahbub. Impact of rumors and misinformation on covid-19 in social media. J Prev Med Public Health, 53(3):171–174, 2020.

[Thakur et al., 2018] Hardeo Kumar Thakur, Anand Gupta, Ayushi Bhardwaj, and Devanshi Verma. Rumor detection on twitter using a supervised machine learning framework. International Journal of Information Retrieval Research (IJIRR), 8(3):1–13, 2018.

[Van Dyk and Meng, 2001] David A Van Dyk and Xiao-Li Meng. The art of data augmentation. Journal of Computational and Graphical Statistics, 10(1):1–50, 2001.

[Wang et al., 2019] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage bert: A globally normalized bert model for open-domain question answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 5881–5885. ACL, 2019.

[Wu and Chien, 2020] Shih-Hung Wu and Sheng-Lun Chien. A bert based two-stage fake news spreaders profiling system. In Conference and Labs of the evaluation Forum (CLEF), Thessaloniki, Greece, September 22-25, 2020.

[Wu et al., 2016] Shu Wu, Qiang Liu, Yong Liu, Liang Wang, and Tieniu Tan. Information credibility evaluation on social media. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Arizona, USA, February 12–17, AAAI'16, 2016.

[Xu et al., 2019] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, pages 2324–2335.

[Yang et al., 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, December 8-14, 2019, pages 5754–5764, 2019.

[Zaib et al., 2020] Munazza Zaib, Quan Z. Sheng, and Wei Emma Zhang. A short survey of pre-trained language models for conversational ai-a new age in NLP. In Proceedings of the Australasian Computer Science Week Multiconference, Melbourne, Australia, February 3-7, 2020, ACSW '20. ACM, 2020.

[Zhang et al., 2020] Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In Thoracic Image Analysis - Second International Workshop, TIA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings (IJCNN), pages 1–8. IEEE, 2020.

[Zhu et al., 2019] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. Incorporating bert into neural machine translation. In International Conference on Learning Representations, 2019.

[Zou et al., 2015] J. Zou, F. Fekri, and S. W. McLaughlin. Mining streaming tweets for real-time event credibility prediction in twitter. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, August 25-28, 2015, pages 1586–1589, 2015.

[Zubiaga et al., 2016a] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Learning reporting dynamics during breaking news for rumour detection in social media. Arxiv reprint, 2016a.

[Zubiaga et al., 2016b] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one, 11(3):e0150989, 2016b.