A Fine-Grained Analysis of Misinformation in COVID-19 Tweets

Sumit Kumar
Birla Institute of Technology
Mesra, India
sumit.atlancey@gmail.com

Raj Ratn Pranesh
Birla Institute of Technology
Mesra, India
raj.ratn18@gmail.com

Kathleen M. Carley
Carnegie Mellon University
Pittsburgh PA, USA
kathleen.carley@cs.cmu.edu

Abstract

In the past few years, there has been expeditious growth in usage of social media platforms and blogging websites which has passed 3.8 billion marks of active users that use text as a prominent means for interactive communication. A fraction of users spread misinformation on social media. As Twitter has 330 million monthly active users, researchers have been using it as a source of data for misinformation identification (Huang and Carley, 2020). In this paper, We have proposed a Twitter dataset for fine-grained classification. Our dataset is consist of 1970 manually annotated tweets and is categorized into 4 misinformation classes, i.e, "Irrelevant", "Conspiracy", "True Information", and "False Information" on the basis of response erupted during COVID-19. In this work, we also generated useful insights on our dataset and performed a systematic analysis of various language models, namely, *RNN* (BiLSTM, LSTM), *CNN* (TextCNN), *BERT*, *ROBERTA* and *ALBERT* for the classification task on our dataset. Through our work, we aim at contributing to the substantial efforts of the research community for the identification and mitigation of misinformation on the internet.

1 Introduction

The global outbreak of COVID-19 resulted in a crisis all around the world. The pandemic originated from Wuhan, China, and rapidly spread all over the world causing casualties and affected human life drastically. By November 1, 2020, the total case count crossed the bar of 46.2 Million, and 1.2 Million people have lost their lives worldwide. As novel coronavirus continues to spread, it has a massive impact on several sectors for instance country economy, public and private sectors, government bodies, and above all affecting the mental and physical health of the people by tempering their daily routines. In spite of this, people expressed their thoughts, news, opinions, and information related to COVID-19 across several social media platforms such as Twitter, Facebook, Whatsapp, Instagram, and Reddit. We have attempted to track this panic response through Twitter. An enormously large number of tweets were tweeted during the outbreak of novel coronavirus. The tweets of our interests are discussed conspiracy theories related to the disease, true prevention and cures, and fake prevention and cures. Studies have shown that numerous people connect to the internet and social media platforms every day to gather information/news through them. As in (Matsa and Shearer, 2018) a great amount of human-generated information being exchanged every day, it has attracted researchers to explore, analyze, and generate valuable insights about people reaction to COVID-19 to identify and mitigate through misinformation detection. In this work, we have manually collected and annotated 1970 user-generated COVID-19 tweets into 4 classes of misinformation. We have also performed a comparative analysis of various existing language models for the misinformation classification task.

The four key **motivation** in this paper are:

- With the advancement of technology, digital news is more widely exposed to users globally and contributes to the increment of spreading hoaxes and disinformation online.
- COVID-19 being one of the largest recent pandemic, the spread of misinformation related to COVID-19 can make the current situation worse and eventually adversely affect the overall functioning of daily human life.
- Hence, it is very crucial for us to develop a misinformation detection system that would

analyze the content (semantic approach) of user-generated data such as tweet and assign a most probable misinformation class to it.

- Developing a misinformation classification system would help in the identification and filtration of information distributed all over the social media platforms and therefore, in the future automatic misinformation systems can be deployed for monitoring and mitigation of misinformation.

Few **Challenges** that we faced are:

- The most challenging task is to determine the type of misinformation in a given Tweet which could be difficult for humans to trace and distinguish themselves, let alone machines.

- Covid-19 is a recent occurrence. So, due to the lack of standardized datasets on the internet, we have to manually annotate to prepare data for training.

2 Related Work

In the past few years, Research has been done in the field of natural language processing which involves analyzing and identification of misinformation in textual representations.

(Huang and Carley, 2020) made use of Twitter data for a detailed analysis of Tweets mentioning "fake news" URLs and disinformation story-lines that are most likely to be spread by regular users. They also showed that unlike real news and normal tweets, tweets containing URLs pointing to "fake news" sites are most likely to be retweeted within the source country and so are less likely to spread internationally. Along with, They utilized machine learning systems to predict users' latent attributes, such as their locations and political orientations.

(Beskow and Carley, 2019) presented a methodology named twitter_sim ABM designed for exploring the explicit actions users make in Twitter which captures the varied actions of malicious agents like bots/tolls. They showed the use of this model in exploring the emerging behavior of specific disinformation maneuvers. Also, They validated some of the key variables in the model from empirical Twitter data.

(Wu et al., 2019) deliberately tried to discuss some key points such as misinformation detection from text classification. They utilized feature engineering methods and sources information using the dataset, they successfully give a detailed explanation for the identification of misinformation spreaders and how they propagate the information.

The above-summarised work talks about various methods for misinformation identification and detection using machine learning model. These papers performed the misinformation classification task on a limited data with very general classes. While our paper focuses on various deep learning language models that were trained on our proposed dataset with fine-grained misinformation classes.

3 Dataset

This section will explain the dataset generation process and description of the dataset that we proposed in this paper. We condense the approach for collecting and Pre-processing the user-generated dataset through the tweets to come up with a final dataset. We have summarised the features of the dataset through some examples in Table 3, along with the data annotation schemes and guidelines.

3.1 Data Collection

We crawled through Twitter data using the Tweepy[1] which is a Python library for accessing Twitter Application Programming Interface (API[2]), and collected a sample of tweets.

To extract the required tweets, we build a set of keywords related to the usage of hashtags in the semantic sentence (e.g., #Covid-19, #Coronavirus) in both lowercase and uppercase. The following keywords such as `covid19`, `quarantine`, `quarantinelife`, `publichealth`, `pandemic`, `terrorism`, `BioWeapon`, `immune`, `5G`, `wuhan`, `wuhancoronavirus`, `conspiracytheory` were used to collect tweets. Hence, The final collected dataset contains 1970 tweets.

3.2 Data Annotation

The gathered data were annotated by two human annotators of linguistic background and proficiency in English using the categories mentioned in Table 2. The categories were chosen based on the frequency of the occurrences of Irrelevant text, Conspiracy text, True Information text, and False Information text associated with tweets.

A general description of each class is given below.

[1] https://www.tweepy.org/

[2] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

Class	tweet count
Irrelevant	768
Conspiracy	468
True Information	462
False Information	272
Total	1970

Table 1: Label associated with each class

Name	Annotation Class
Mention of Irrelevant text	0
Mention of Conspiracy text	1
Mention of True Information text	2
Mention of False Information text	3

Table 2: Label associated with each class

Mention of Irrelevant text: A tweet shall be classified as irrelevant if it may or may not mention COVID-19 or SARS-Cov-2, but if it cannot be classified in any of the other categories below.

Mention of Conspiracy text: A tweet shall be classified as a conspiracy if it endorses a conspiracy story. Some examples of conspiracy themes related to COVID-19 include:

- It is a bio-weapon.
- Electromagnetic fields and the introduction of 5G wireless technologies led to COVID-19 outbreaks.
- It leaked from the Wuhan Labs or Wuhan Institute of Virology in China.
- It was predicted by Dean Koontz.

Mention of True Information text: A tweet shall be classified as true information if it endorses a method of treatment to ease the pain (rest and sleep, keep warm, drink plenty of liquids, etc.), and if it explicitly endorses a method of prevention, and if any of the following conditions are met:

- The information has been verified by the World Health Organization (WHO) site or by the Center of Disease Control and Prevention (CDC) site.
- The information is supported by a peer-reviewed scientific journal that appears in Ulrich's Global Serials Directory as both "Active" and "Refereed/Peer-reviewed".
- Tweet links directly to the news stories which correctly cites a peer-reviewed journal article.

Mention of False Information text: A tweet shall be classified as false information if the content endorses a cure and any of the following conditions are met:

- The information cannot be verified by the World Health Organization (WHO) site or by the Center of Disease Control and Prevention (CDC) site.
- The information is not supported by a peer-reviewed scientific journal that appears in Ulrich's Global Serials Directory as both "Active" and "Refereed/Peer-reviewed".
- The information calls out or makes fun of a fake cure, fake prevention, fake treatment, or a conspiracy theory.

3.3 Dataset Preprocessing

Before conducting the analysis and experiments, we preprocessed tweets by firstly converting them to lowercase representation. We also made the tweets free from any unnecessary elements such as username, mentions, links, retweets. We used NLTK[3], a Python module for text processing that removed the English stopwords and performed lemmatization of tweets.

4 Methodology

In this section, we have sequentially discussed the architecture of various classification models used in our experiment. We used *LSTM*,

[3]https://www.nltk.org/

Class	Tweet Data
Irrelevant	Morning everyone. I hope you all have a relaxing Sunday!\n#StayAtHomeSaveLives #coronavirus #covid19 #SundayThoughts
Irrelevant	Did you know that 66% of U.S. workers are working remotely full-time as a result of the #coronavirus pandemic.
Conspiracy	Many people across the globe lost lives due to #coronavirus spreaded by #China. It is not a disease, it is a biological weapon created by China to destroy and rule the world. Why there are no cases in China, while being the most populated country across the globe
Conspiracy	Covid 19 is a genetically modified virus. It is a bioweapon that came from a bio lab in China and has nothing to do with eating meat. It did not come from eating bat soup and it is spread by respiratory droplets NOT what you eat.
True Information	The French Ministry of Health has issued a press release to notify all citizens that cocaine does NOT protect against COVID-19.
True Information	You took it to protect against malaria. There are NO trials that show it is effective against #COVID19
False Information	Gargle with Listerine. Preferably the alcohol Listerine. It will kill the covid-19 in your throat and mouth.
False Information	Use the space heater or blow dryer and breathe the heat in through your nose. It will kill the covid-19 in your nostrils. Use a pinch of salt in boiling water breathe in.

Table 3: Tweet with their respective misinformation class

Bi-LSTM, *TextCNN*, *CNN-RNN*, *RNN-CNN*, *Distil_Bert*, $BERT_{base}$, $BERT_{large}$, *Distil_RoBERTa*, $RoBERTa_{base}$, $RoBERTa_{large}$, $Albert_{base_V2}$, $Albert_{large_V2}$.

4.1 CNN

In this subsection, we described the Convolution Neural Networks (Fukushima, 1988) for classification and also outlines the methods for text classification specifically. Convolutional neural networks are multistage trainable neural network architectures developed for classification tasks (LeCun et al., 1998). Each stage contains different layers as summarized below:

- **Embedding Layer:** The function of an embedding layer is to transform the text inputs into a form that can be used by the CNN model. Here, each word of a text document is transformed into a dense vector of fixed size.
- **Convolutional Layer:** A Convolutional layer comprises of several kernel matrices that perform the convolution mathematical operation on their input and process an output matrix of features upon the addition of a bias value.
- **Pooling Layer:** A pooling layer performs dimensionality reduction of the input feature vectors. It uses sub-sampling to the output of the convolutional layer matrices combing neighboring elements. we have used the max-pooling function for the pooling.
- **Fully Connected Layer:** : A classic fully connected neural network layer is connected to the Pooling layers via a Dropout layer in order to prevent overfitting. The softmax activation function is used for defining the final output of this layer. The following objective function is commonly used in the task:

$$E_w = \frac{1}{n} \sum_{p=1}^{P} \sum_{j=1}^{N_l} (o_{j,p}^{L} - o_{j,p})^2 \qquad (1)$$

where P is the number of patterns, $o_{j,p}^{L}$ is the output of j^{th} neuron that belongs to L^{th} layer, N_{l} is the number of neurons in output of L^{th} layer, $y_{j,p}$ is the desirable target of j^{th} neuron of pattern p and y_{i} is the output associated with an input vector x_{i} to the CNN.

In order to minimize the cost function E_w , we use Adam Optimizer (Kingma and Ba, 2014).

4.2 RNN

Recurrent neural networks (RNN) have been used to produce promising results on different tasks, along with language model and speech recognition

(Kombrink et al., 2011; Graves and Schmidhuber, 2005). An RNN predicts the current output conditioned on long-distance features by keeping a memory based on previous information.

An input layer represents features at time t. Onehot vectors for words, dense vector features such as word embeddings, or sparse features usually represent an input layer. An input layer has the same dimensionality as feature size. An output layer represents a probability distribution over labels at time t and also has the same dimensionality as the size of the labels. Compared to the feedforward network, an RNN holds a relation between the previous hidden state and the current hidden state. This relation is made through the recurrent layer, which is designed to store history information. The following equation is used to calculate the values in the hidden, and output layers:

$$\mathbf{h}(t) = f(\mathbf{U}\mathbf{x}(t) + \mathbf{W}\mathbf{h}(t-1)) \quad (2)$$

$$\mathbf{y}(t) = g(\mathbf{V}\mathbf{h}(t)) \quad (3)$$

where U, W, and V are the connection weights to be computed during training, and f(z) and g(z) are sigmoid and activation functions as given below:

$$f(z) = \frac{1}{1+e^{-z}} \quad (4)$$

$$g(z_m) = \frac{e^{z_m}}{\sum_k e_k^z} \quad (5)$$

For the purpose of sequence tagging, we used Long Short Term Memory (LSTM) and Bidirectional Long Term Short Memory (Bi-LSTM) as in (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Graves et al., 2013).

LSTM networks use purpose-built memory cells to update the hidden layer values. Therefore, they may perform better at finding and utilizing long-range dependencies in the data, unlike a standard RNN. The following equation implements the LSTM model:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (6)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (7)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (8)$$

$$h_t = o_t \tanh(c_t) \quad (9)$$

For a given time, both past and future input features can be accessed in the sequence tagging task. Therefore, we can also utilize a bidirectional LSTM network (Bi-LSTM) as proposed by the author (Graves et al., 2013).

4.3 CNN-RNN

For designing a CNN-RNN classification model, we stacked the CNN layer over the RNN layer. Firstly, the embedding layer of dimension 300 is utilized to generate a vector representation of tweet text using GloVe (Pennington et al., 2014). The embeddings are then supplied into a spatial dropout layer (p = 0.4), followed by a convolution layer and a max-pooling layer (pool size = 2). The polled output is then supplied into a BiLSTM layer followed by a global max polling layer. Finally, the obtained polled output is then passed into a dense layer with 256-dimension (with Relu activation(Nair and Hinton, 2010)), and then the prediction is made using a d layer with softmax activation.

4.4 RNN-CNN

For designing an RNN-CNN classification model, we employed a single BiLSTM layer over the top of a 1D-CNN layer in the following way. The generated embeddings from the embedding layer after applying spatial dropout (p = 0.4) were supplied into a BiLSTM with rnn units = 300, followed by a convolution layer with dimension = 64. Over the obtained output, the global average pool and global max pool values were extracted and concatenated to produce sentence-level embedding which is then passed to a dense layer with Relu activation(Nair and Hinton, 2010). Finally, a hidden layer of size 4 with softmax function is used to make final predictions.

4.5 Transformer Language Models

Large pretrained language models such as BERT, RoBERTa, ALBERT have gained a lot of popularity in the field of NLP. These language models have been shown to learn remarkably well on downstream tasks including machine translation, question-answering, text classification, and summarization.

In our work, we have utilized (i) three variant of BERT(Devlin et al., 2018) model: *Distil_Bert*, $BERT_{base}$, $BERT_{large}$, (i) three vari-

Score / Models	Precision	Recall	F1
LSTM	70.78	63.46	66.87
Bi-LSTM	70.67	65.23	67.87
TextCNN	70.33	66.26	65.74
CNN-RNN	71.84	70.98	71.40
RNN-CNN	70.15	67.41	68.75
AlBERT-large-v2	58.75	58.25	58.00
AlBERT-base-v2	55.20	54.90	55.23
BERT-large	67.00	66.75	69.00
BERT-base	66.5	66.25	68.00
Distil-BERT	66.2	66.00	65.00
RoBERTa-base	77.5	75.9	74.00
Distil-RoBERTa	71.00	71.25	71.00
RoBERTa-large	73.75	73.5	76.00

Table 4: Performance score (in %) of various models

ant of RoBERTa(Liu et al., 2019) model: *Distil_RoBERTa*, $RoBERTa_{base}$, $RoBERTa_{large}$ and tho variant of ALBERT(Lan et al., 2019): $Albert_{base_V2}$, $Albert_{large_V2}$. Transformers are contextualized word presentation model, pretrained using bidirectional transformers (Vaswani et al., 2017). Basically, each model utilizes the work for predicting the next sentence and thus, learns the embeddings with a larger context. The transformer architectures such as $BERT$, $RoBERTa$, and $ALBERT$ were needed to be fine-tuned for the misinformation classification task. We use a pool of labeled training examples for fine-tuning BERT for misinformation detection task using the balanced set of proposed annotated data. We performed the fine-tuning of each pretrained language model by building a custom classification head on top of the models. The classification head was consist of a dropout layer(p=0.05) followed by a linear layer(size = 768) with Mish(Misra, 2019) activation function followed by an another dropout layer and a final linear layer(size = 768). The averaged pool of sequential output from 12 encoding layers of used as the custom classifier head's input.

5 Result and Discussion

In this section, we have summarised the result obtained in our experiment and also discussed the performance of various classification models on our dataset. As seen in the table 4, the RoBERTa-large performs better than the rest of the models having an F1-score of 76% with the precision of 73.75% and recall of 73.5% on the dataset. Since RoBERTa-large is trained on a bigger corpus compared with the training datasets of other models, and the1 optimized hyperparameters make this model more suitable for the task. Distil-Roberta model had a precision of 71%, this makes this model more favorable in detecting true informative tweets. BERT-large and BERT-base had a comparative better result on the given dataset, making these models suitable for the task. As for the Distil-BERT, due to fewer parameters, it was hard for this model to infer in favour of this task. As for the Albert-large, it outperformed its smaller version of the model by 2.77% F1-score. Moving onto the untrained networks, the CNN-RNN model comparatively performed better than the other networks, this is because of the CNN architecture, as it is able to highlight the calculated value using kernels, whereas for the RNN-CNN model it surpassed the TextCNN model by 3.01%. LSTM and Bi-LSTM model was also great for the task as they scored 66.87% and 67.87% F1-score respectively.

6 Conclusion

In this paper, we have presented a study of COVID-19 misinformation in tweets. We manually created a dataset consisting of 1970 tweets annotated in 4 classes of misinformation. We utilized various deep learning language models such as RNNs, CNN, BERT, RoBERTa, ALBERT, and performed a comparative analysis of these models for the misinformation type classification task. In our study, we found that the larger pretrained model RoBERTa performed better than the other models. We strongly believe that our model can help infiltration of misinformation data present on the internet as well as to understand public perceptions during the pandemic. In the future, we aim at collecting more annotated training data for improving our model's robustness and contextual understanding for better performance in the classification task. Task such as discovering topics and extracting keywords from multilingual tweets would be interesting.

References

David M Beskow and Kathleen M Carley. 2019. Agent based simulation of bot disinformation maneuvers in twitter. In *2019 Winter Simulation Conference (WSC)*, pages 750–761. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and

Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kunihiko Fukushima. 1988. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Binxuan Huang and Kathleen M. Carley. 2020. Disinformation and misinformation on twitter during the novel coronavirus outbreak.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Stefan Kombrink, Tomáš Mikolov, Martin Karafiát, and Lukáš Burget. 2011. Recurrent neural network based language modeling in meeting recognition. In *Twelfth annual conference of the international speech communication association*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

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

Katerina Eva Matsa and Elisa Shearer. 2018. News use across social media platforms 2018. *Pew Research Center*, 10.

Diganta Misra. 2019. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90.