CMTA: COVID-19 Misinformation Multilingual Analysis on Twitter

Raj Pranesh, Mehrdad Farokhnejad, Ambesh Shekhar, Genoveva Vargas-Solar

CMTA: COVID-19 Misinformation Multilingual Analysis on Twitter

Raj Ratn Pranesh[1,*], **Mehrdad Farokhnejad**[2,**],
Ambesh Shekhar[1,*], and **Genoveva Vargas-Solar**[3,‘]

[1]Birla Institute of Technology, Mesra, India
[2]Univ. Grenoble Alpes, CNRS, LIG, Grenoble, France
[3]CNRS, LIRIS-LAFMIA Lyon, France
[*](raj.ratn18, ambesh.sinha)@gmail.com
[**]mehrdad.farokhnejad@univ-grenoble-alpes.fr
[‘]genoveva.vargas-solar@liris.cnrs.fr

Abstract

The internet has actually come to be an essential resource of health knowledge for individuals around the world in the present situation of the coronavirus condition pandemic(COVID-19). During pandemic situations, myths, sensationalism, rumours and misinformation, generated intentionally or unintentionally, spread rapidly through social networks. Twitter is one of these popular social networks people use to share COVID-19 related news, information, and thoughts that reflect their perception and opinion about the pandemic. Evaluation of tweets for recognizing misinformation can create beneficial understanding to review the top quality and also the readability of online information concerning the COVID-19. This paper presents a multilingual COVID-19 related tweet analysis method, CMTA, that uses BERT, a deep learning model for multilingual tweet misinformation detection and classification. CMTA extracts features from multilingual textual data, which is then categorized into specific information classes. Classification is done by a Dense-CNN model trained on tweets manually annotated into information classes (i.e., 'false', 'partly false', 'misleading'). The paper presents an analysis of multilingual tweets from February to June, showing the distribution type of information spread across different languages. To access the performance of the CMTA multilingual model, we performed a comparative analysis of 8 monolingual model and CMTA for the misinformation detection task. The results show that our proposed CMTA model has surpassed various monolingual models which consolidated the fact that through transfer learning a multilingual framework could be developed.

1 Introduction

Since late 2019, the coronavirus disease COVID-19 has spread worldwide to more than 216 countries (Organization et al., 2020). COVID-19 has created a massive impact on multiple sectors including countries economy, government bodies, private companies, media houses and most importantly, affecting the mental and physical health of human beings by tempering their daily routine activities (Torales et al., 2020; Fernandes, 2020).

COVID-19 also has made us realize how well the world is interconnected through the Internet. Social media is a significant conduit where people share their response, thoughts, news, information related to COVID-19, with one in three individuals worldwide participating in social media, with two-thirds of people utilizing it on the Internet (Ortiz-Ospina, 2020). Studies have shown that many people connect to the Internet and social media platforms such as Twitter, Facebook, Whatsapp, Instagram and Reddit every day and utilizing it for getting information/news through them (Matsa and Shearer, 2018) (Hitlin and Olmstead, 2018). Twitter users are known, especially, for posting and exchanging news: almost 60% of Twitter users classify it as excellent or incredibly helpful for sharing preventive health information (Wilford et al., 2018).

Nonetheless, social media is still full of misinformation regarding health. It is difficult to assess the authenticity of health information on the Internet for people with non-medical experience. Precise and reliable dissemination of correct information about the virus that causes a pandemic will help to monitor the spread of the virus and related population anxiety (Sharma et al., 2017). Social media content and misinformation may have intense implications for public opinion and behavior, positively or negatively influencing the viewpoint of those who access it (Brindha et al., 2020; Kouzy et al., 2020).

The WHO director-general stated at the Munich security conference in February 2020, 'we are not only fighting an epidemic; we are fighting an info-

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 270–283
August 5–6, 2021.

demic' (Zarocostas, 2020). It is clear that there is no way of stopping the transmission of COVID-19, so it is necessary to check information on the Internet in order to prevent the panic and disinformation linked to the disease. Seeking accurate and valid information is the biggest challenge with Internet health information (Eysenbach et al., 2002).

Misinformation appears in several ways in the case of COVID-19, such as 'COVID-19 is a biological agent developed by either the US or China','COVID-19 is the potential by-product of Chinese cuisine, such as bat soup amongst other ingredients,'and 'breath-holding self-detection test', unconfirmed home remedies such as vitamin C, urine from animals, turmeric etc. In its worse, this type of misinformation will lead individuals to resort to unsuccessful (and actually directly harmful) remedies, either to overreact (e.g. by hoarding goods) or to underreact quite dangerously (e.g., by deliberately engaging in risky behavior and inadvertently spreading the virus). (Brindha et al., 2020; Pennycook et al., 2020). Unfortunately, the fake news spread faster than the virus (Gallotti et al., 2020).

An online social platform such as Twitter provides particularly fertile ground for the spread of misinformation (Frenkel et al., 2020). Twitter gives direct access to extraordinary content, which may intensify rumors and dubious information (Cinelli et al., 2020). With such a huge amount of human-generated information being exchanged every day, it has attracted Natural Language Processing (NLP) researchers to explore, analyze, and generate valuable insights about people response to COVID-19. People response is analyzed with respect to sentiments and misinformation and malicious information detection.

This paper proposes CMTA, a multilingual tweet analysis and information (misinformation) detection method for understanding both the negative and positive sides of social media during COVID-19 pandemic. CMTA uses Multilingual BERT, trained on 104 multiple languages to derive features from tweets and 1D convolution for finding the correlation between data of hidden states. It also uses a dense layer for linear transformation on contextual embeddings to provide inferential points. Our work helps in providing better results in finding the proximity of being fake. We used manually annotated multilingual COVID-19 related tweets for training deep neural network model in order to detect and identify the type of misinformation present in tweets belonging to different language groups.

For experimenting with our method, we used trained models for a systematic analysis of COVID-19 related tweets collected from February to June 2020. The analysis of tweets is done based on the distribution of the type of information present in tweets concerning the language used for writing a tweet. We investigated the presence of false information spread throughout Tweeter by classifying the tweets in three classes: 'false', 'partly false' and 'misleading'. We have provided illustrative statistical representation of our findings and detailed discussion about the insights discovered in our survey. The **motivation** for designing a multilingual method lies behind the need of analyzing not just monolingual tweets but also multilingual tweets by building a single deep learning framework that would be able to understand tweets in multiple languages. That being said, we also analysed the performance of CMTA multilingual BERT framework with respect to 8 monolingual BERT models. The performance score achieved by the multilingual model were very close to that of monolingual models which suggests that utilizing a singular multilingual model for COVID-19 tweet analysis and disinformation categorization is a reliable and robust method.

2 Related Work

The COVID-19 pandemic has resulted in studies investigating the various types of misinformation arising during the COVID-19 crisis (Brennen et al., 2020; Dharawat et al., 2020; Singh et al., 2020; Kouzy et al., 2020). Studies investigate a small subset of claims (Singh et al., 2020) or manually annotate Twitter data (Kouzy et al., 2020). In (Brennen et al., 2020) authors analyse different types of sources for looking for COVID-19 misinformation. Pennycook et al. (Pennycook et al., 2020) introduced an attention-based account of misinformation and observed that people tend to believe false claims about COVID-19 and share false claims when they do not think critically about the accuracy and veracity of the information. Kouzy et al. (Kouzy et al., 2020) annotated about 600 messages containing hashtags about COVID-19, they observed that about one-fourth of messages contain some form of misinformation, and about 17% contain some unverifiable information. With such mis-

information overload, any decision making procedure based on misinformation has a high likelihood of severely impacting people's health (Ingraham and Tignanelli, 2020). The work in (Huang and Carley, 2020) examined the global spread of information related to crucial disinformation stories and "fake news" URLs during the early stages of the global pandemic on Twitter. Their study shows that news agencies, government officials, and individual news reporters send messages that spread widely and play critical roles. Tweets citing URLs for "fake news" and reports of propaganda are more likely than news or government pages shared by regular users and bots.

The work in (Sharma et al., 2020) focused on topic modelling and designed a dashboard to track Twitter's misinformation regarding the COVID-19 pandemic. The dashboard presents a summary of information derived from Twitter posts, including topics, sentiment, false and misleading information shared on social media related to COVID-19. Cinelli et al. (Singh et al., 2020) track (mis)-information flow across 2.7M tweets and compare it with infection rates.They noticed a major Spatio-temporal connection between information flow and new COVID-19 instances, and while there are discussions about myths and connections to low-quality information, their influence is less prominent than other themes specific to the crisis. To find and measure causal relationships between pandemic features (e.g. the number of infections and deaths) and Twitter behaviour and public sentiment, the work in (Gencoglu and Gruber, 2020) introduced the first example of a causal inference method. Their proposed approach has shown that they can efficiently collect epidemiological domain knowledge and identify factors that influence public interest and attention.

The discussion around the COVID-19 pandemic and the government policies was investigated in(Lopez et al., 2020). They used Twitter data in multiple languages from various countries and found common responses to the pandemic and how they differ across time using text mining. Moreover, they presented insights as to how information and misinformation were transmitted via Twitter. Similarly, to demonstrate the epidemiological effect of COVID-19 on press publications in Bogota, Colombia, (Saire and Navarro, 2020) used text mining on Twitter data. They intuitively note a strong correlation between the number of tweets and the number of infected people in the area.

Most of the works described above focus on analysing tweets related to single language such as English. In our work we have designed a single model leveraging multilingual BERT for the analysis of tweets in multiple languages. Furthermore, we used a large data set to train and analyze the tweets. Our aim is to provide a system that will be restricted to any language for analysing social media data.

3 Data preparation

This section discusses the steps involved in the collection of COVID-19 related tweets. For training our misinformation detection deep learning model, we have extracted annotated misinformation data from multiple publicly available open databases. We also collected a very large number of multilingual tweets consisting of over 2 million tweets belonging to eight different languages.

3.1 Training Dataset

In order to train and test our misinformation detection model, we collected the training data from an online fact-checker website called Poynter (Poynter Institute, 2020). Poynter have a specific COVID-19 related misinformation detection program named 'CoronaVirusFacts/DatosCoronaVirus Alliance Database[1]'. This database contains thousands of labelled social media information such as news, posts, claims, articles about COVID-19 which were manually verified and annotated by human volunteers(fact-checkers) from all around the globe. The database gathers all the misinformation related to topics such as COVID-19 cure, detection, the effect on animals, foods, travel, government policies, crime, lockdown.

The misinformation dataset was available in 2 languages- 'English' and 'Spanish'. Since we were training a multilingual BERT model, we crawled through the content of all 2 websites using Beautifulsoup[2], a Python library for scraping information from web pages. We scrape 8471 English language false news/information belonging to nine major classes namely, 'False', 'Partially false', 'Misleading', 'No evidence', 'Four Pinocchios', 'Incorrect', 'Three Pinocchios', 'Two Pinocchios' and 'Mostly False'. For each article we gath-

[1] https://www.poynter.org/covid-19-poynter-resources/

[2] Python module is available at https://pypi.org/project/beautifulsoup4/

Classes	Number of tweets
False (Poynter Institute, 2020) (English)	2,869
Partially False (English)	2,765
Misleading (English)	2,837
False (Spanish)	191
Partially False (Spanish)	161
Misleading (Spanish)	179
False (Alam et al., 2020) (English)	500
Total	9,502

Table 1: Collected Misinformation Dataset

ered the article's title, it's content and the fact checker's misinformation-type label. Similarly, from the Spanish[3] databases we collected 531 misinformation articles respectively. The collected data contains the misinformation published on social media platforms such as Facebook, Twitter, What'sapp, YouTube and were mostly related to political-biased news, scientifically dubious information and conspiracy theories, misleading news and rumors about COVID-19. We also used one more human annotated fact-checked tweet dataset (Alam et al., 2020) available at the public repository[4]. The dataset contained true and false labelled tweets in English and Arabic language. We used only false labelled tweets consisting of 500 English. We compiled (table 1) a total of 9,502 micro-articles distributed across 9 misinformation classes.

Defining misinformation classes: The collected data was unevenly distributed across 9 classes. We put the classes such as 'No evidence', 'Four Pinocchios[5]', 'Incorrect', 'Three Pinocchios[6]', 'Two Pinocchios[7]' and 'Mostly False' under the minority group because of having very few labels. On the other hand, labels like 'False', 'Partially false' and 'Misleading' comprises the majority group as most of the collected articles belongs to this group. In order to structure and distribute the dataset uniformly for training our model, we reformed the dataset by merging the minority group labels into the majority group labels. The classes ('Four Pinocchios' and 'Incorrect') that correspond to completely false information were merged together into the 'False' class. 'Three Pinocchios' and 'Two Pinocchios' were merged together into 'Partially false' class. 'No evidence' and 'Mostly False' were put together with the 'Misleading' class.

Table 6 gives a clear understanding of our training dataset and showcase some misinformation articles present in our training dataset. Column 1 shows the reformed label assigned by us, column 2 shows the original label assigned by the fact-checker, column 3 gives a misinformation example associated with the label present in column 2, and column 4 provides a reasoning given by the fact-checker behind assigning a particular label (column 2) to the misinformation (column 3). For example, if we would look at the entry number '3' in the table 6, the misinformation is about the adverse effect of 5G radiation over the COVID-19 patients. This was labeled 'Incorrect' by the fact-checker. After analysing the fact-checker rating and the explanation given, we labelled it as 'False' misinformation. Entry number '5' talks about the COVID-19 test cost. The explanation given by fact-checker is valid as it is not sure if there is any fee in USA for COVID-19 test or not. So because of the lack of evidence and uncertainty we labelled it as 'Partially false'. Entry number '7' in the table talks about a video showing COVID-19 corpus dumping in the sea. Based on the explanation, the video was coupled with the wrong information to mislead the audience. So it was labelled as 'Misleading' misinformation.

3.2 Inference Dataset

Once we finished training our multilingual tweet misinformation detection model we aimed to use it for predicting and analysing the misinformation spread across all over the social media platforms in multiple languages. In order to do so, we collected around 2,137,106 multilingual tweets consisting of tweets belonging to eight major languages, namely- 'English', 'Spanish', 'Indonesian', 'French', 'Japanese', 'Thai', 'Hindi' and 'German'. We used an ongoing dataset of tweets IDs associated with the novel coronavirus COVID-19 (Chen et al., 2020). Started on January 28, 2020, the current version of dataset contains 212,978,935 tweets divided into groups based on their publishing month. The dataset was collected using multilingual COVID-19 related keywords and contains tweets in more than 30 languages. We used tweepy[8] which is a Python module for accessing twitter API. For our analysis we decided to retrieve the tweets using the tweet IDs of the tweets pub-

[3] https://chequeado.com/latamcoronavirus/

[4] https://github.com/firojalam/COVID-19-tweets-for-check-worthiness

[5] 90%-95% changes of it being false

[6] 70%-75% changes of it being false

[7] 50%-55% changes of it being false

[8] Python module is available at http://www.tweepy.org

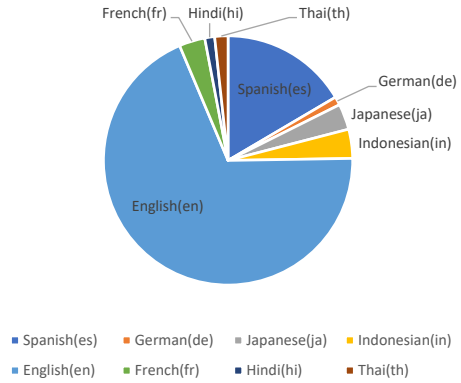

Figure 1: Language-wise Dataset Distribution Pie chart.

Language	ISO	Number of tweets
English	en	1,472,448
Spanish	es	353,294
Indonesian	in	80,764
French	fr	71,722
Japanese	ja	71,418
Thai	th	36,824
Hindi	hi	27,320
German	de	23,316
Sum		2137106

Table 2: Language-wise Dataset Distribution

lished in past 5 months (February, March, April, May and June). Table 2 shows the total number of tweets collected by us and figure 1 shows their distribution across eight different language.

4 The CMTA Method

In this section, we have given a detailed sequential overview of CMTA method design. Both misinformation[9] and disinformation[10], according to the Oxford English Dictionary, are false or misleading information. Misinformation refers to information that is accidentally false and spread without the intent to hurt, whereas disinformation refers to false information that is intentionally produced and shared to cause hurt (Hernon, 1995). Claims do not have to be entirely truthful or incorrect; they can contain a small amount of false or inaccurate information(Shahi and Nandini, 2020). This work uses the general notion of misinformation and makes no distinction between misinformation and disinformation as it is practically difficult to determine one's intention computationally. Figure 2 shows the phases of the analytics pipeline of CMTA with their internal processes. CMTA implements a data science pipeline consisting of four phases: (1) tokenizing, (2) text features extraction, (3) linear transformation, and (4) classification. The first phases (tokenizing, text feature extraction, linear transformation) correspond to a substantial data-preparation process intended to build a multi-lingual vectorized representation of texts. The objective is to achieve a numerical pivot representation of texts agnostic of the language. CMTA classification task uses a dense layer and leads to a trained network model that can be used to classify micro-texts (e.g. tweets) into three misinformation classes: 'false', 'partly false' and 'misleading'.

Text tokenization Given a multilingual textual dataset consisting of sentences, CMTA uses the BERT multilingual tokeniser to generate tokens that BERT's embedding layer will further process. CMTA uses MBERT[11] to extract contextual features, namely word and sentence embedding vectors, from text data [12]. In the subsequent CMTA phases that use NLP models, these vectors are used as feature inputs with several advantages. (M)BERT embeddings are word representations that are dynamically informed by the words around them, meaning that the same word's embeddings will change in (M)BERT depending on its related words within two different sentences.

For the non-expert reader, the tokenization process is based on a WordPiece model. It greedily creates a fixed-size vocabulary of individual characters, subwords, and words that best fit a language data (e.g. English) [13]. Each token in a tokenized text must be associated with the sentence's index: sentence 0 (a series of 0s) or sentence 1 (a series of 1s). After breaking the text into tokens, a sentence must be converted from a list of strings to a list of vocabulary indices. The tokenisation result is used

[9] https://www.oed.com/view/Entry/119699?redirectedFrom=misinformation

[10] https://www.oed.com/view/Entry/54579?redirectedFrom=disinformation

[11] https://github.com/google-research/bert/blob/master/multilingual.md

[12] Embeddings are helpful for keyword/search expansion, semantic search and information retrieval. They help accurately retrieve results matching a keyword query intent and contextual meaning, even in the absence of keyword or phrase overlap.

[13] This vocabulary contains whole words, subwords occurring at the front of a word or in isolation (e.g., "em" as in the word "embeddings" is assigned the same vector as the standalone sequence of characters "em" as in "go get em"), subwords not at the front of a word, which are preceded by '##' to denote this case, and individual characters (?)

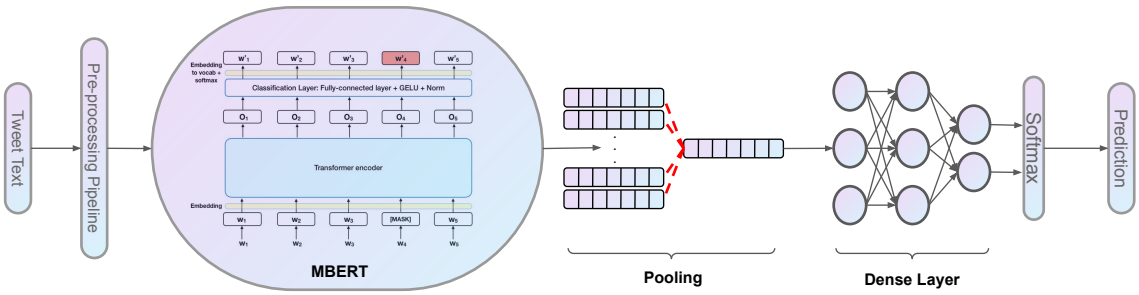

Figure 2: A detailed structure of CMTA architecture.

as input to apply BERT that produces two outputs, one pooled output with contextual embeddings and hidden-states of each layer. The complete set of hidden states for this model are stored in a structure containing four elements: the layer number (13 layers) [14], the batch number (number of sentences submitted to the model), the word / token number in a sentence, the hidden unit/feature number (768 features) [15].

In the case of CMTA, the tokenisation is more complex because it is done for sentences written in different languages. Therefore, it relies on the MBERT model that has been trained for this purpose.

Feature Extraction Phase is intended to exploit the information of hidden-layers produced due to applying BERT to the tokenisation phase result. The objective is to get individual vectors for each token and convert them into a single vector representation of the whole sentence. For each token of our input, we have 13 separate vectors, each of length 768. Thus, to get the individual vectors, it is necessary to combine some of the layer vectors. The challenge is to determine which layer or combination of layers provides the best representation.

Linear convolution The hidden states from the 12th layer are processed in this phase, applying linear convolution and pooling to get correlation among tokens. We apply a three-layer 1D convolution over the hidden states with consecutive pooling layers. The final convolutional layer's output is passed through a global average pooling layer to get a final sentence representation. This representation holds the relation between contextual embeddings of individual tokens in the sentence.

Classification A linear layer is connected to the model in the end for the CMTA classification task.

This classification layer outputs a Softmax value of vector, depending on the output, the index of the highest value in the vector represents the label for the given sequence: 'false', 'partly false' and 'misleading'.

5 Experiment

5.1 Dataset Proprocessing

In data preprocessing, we performed cleaning and structuring of the training and inference dataset. The collected dataset contained lots of unnecessary noises and components such as emojis, symbols, numeric values, hyperlinks to websites and username mentions which were needed to be removed. Since our dataset was multilingual, we had to be very careful while preprocessing as we did not wanted to lose any valuable information. We used simple regular expressions to remove URLs, special characters or symbols, blank rows, re-tweets, user mentions but we did not removed the hashtags from the data. As hashtags might contain useful information. For example in the sentence- 'Wear mask to protect yourself from #COVID-19 #corona', only '#' symbol was removed during the preprocessing(e.g. 'Wear mask to protect yourself from COVID19 corona'). We removed stop words using NLTK[16], a Python library for natural language processing. NLTK supports multiple languages except few languages such as Hindi and Thai in our case. For preprocessing Hindi dataset we used CLTK(Classical Lan-

[14] It is 13 because the first element is the input embeddings, the rest is the outputs of each of BERT's 12 layers.

[15] That is 219,648 unique values to represent our one sentence!

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

guage Toolkit) [17] which supports Hindi stop words. For removing Thai stop words from Thai tweets, we used PyThaiNLP (Wannaphong Phatthiyaphaibun, 2016). The emojis were removed using their unicodes. For training our model we divided the dataset into training, validation and testing dataset in the ratio of 80%/10%10% respectively. The final count for train, validation and test dataset was 7,602, 950, 950.

5.2 Model Setup and Training

Training Setting We fine-tuned the Sequence Classifier from HuggingFace based on the parameters as specified in (Devlin et al., 2018). Thus, we set a batch size of 32, learning rate 1e-4, with Adam Weight Decay as the optimizer. We run the model for training for 10 epochs. Then, we save the model weights of the transformer. These will be helpful for the further training.

Hyperparameters' Setting Table 3 lists every hyperparameter for training and testing our model. All the calculations and selection of hyperparmaters are done based on tests and for the best output from the model. After performing several iterations on distinct sets of hyper-parameters, based on the analysis of the model's performance, we adopted the one showing promising results on our dataset.

Parameters	Value
Pool Size of Average Pooling	8
Pool Size of Max Pooling	8
Dropout Probability	0.36
Number of Dense layers	4
Text Length	128
Batch Size	32
Epochs	10
Optimizer	Adam
Learning Rate	1×10^{-4}

Table 3: Hyper-parameters for training

5.3 Results assessment

This section discusses the performance our multilingual model over the test data. On the test dataset, our model was able to achieve an accuracy(%) of **82.17** and F_1(%) of **82.54**. The precision and recall reported by the model were **82.07** and **82.30** respectively. Table 5 shows model's prediction over few examples from the test dataset along with their actual label. As we shown in the table, the model prediction in case of entry number '1', '2', '3' and '4' our model was able to predict the correct the label. But in case of entry number '5' the label predicted by our model was 'False' whereas the actual label is 'Misleading'. If we would look at the misinformation at the entry number '5' which is a Spanish text- 'El medicamento contra piojos sirve como tratamiento contra Covid-19.' and who's English translation would be- ". This misinformation claims about a COVID-19 medicine and since this could be 'false' and 'misleading' misinformation at the same time, our model predicted it as a 'false' misinformation rather than 'misleading'.

[17] https://docs.cltk.org/en/latest/index.html

6 Multilingual Misinformation Analysis

In this section, we provide a detailed analysis misinformation distribution across the multilingual tweets. We used our trained multilingual model to predict and categorize the misinformation type present in tweets. We conducted our sequential misinformation analysis on a collection of over 2 million multilingual tweets. Our survey studied and analyzed the distribution of COVID-19 misinformation across eight major languages, (i.e. 'English', 'Spanish', 'Indonesian', 'French', 'Japanese', 'Thai', 'Hindi' and 'German') for five months (i.e. February, March, April, May and June). Figure 4 shows the month-wise distribution of misinformation types for each language. Table 4 presents a detailed count of misinformation classes across all the languages. In the figure 6, we could observe that for February, March and June months our model predicted large number of tweets as 'False', followed by 'Misleading' which is second largest and the number of 'Partially false' was the least. For the tweets generated during the month of April and May, our model discovered that the number of 'Partially false' tweets are more than 'Misleading' tweets and 'False' tweets were again in majority. Figure **??** parallelly showcase the overall(all 5 months together) spread of misinformation types across each language. We could clearly see that German tweets have the highest number of 'Misleading' tweets whereas French have the least. Spanish tweets beats other language's tweets by becoming the language with largest source of 'False' misinformation. Germany generated the least number of 'False' tweets. Hindi tweets tends to have the highest number of 'Partially false' tweets whereas

Lingo	February			March			April		
	Misinformation			Misinformation			Misinformation		
	False	Partially False	Misleading	False	Partially False	Misleading	False	Partially False	Misleading
Spanish	58346	6653	13740	67956	10913	8826	34125	5437	3604
German	517	581	2505	862	1438	3043	584	892	2664
Japanese	1920	3079	5245	448	692	2650	1635	2850	5840
Indonesian	11157	3226	1951	12573	4336	1582	9073	3367	1273
English	88369	62747	76640	92428	96571	105143	77368	74947	63473
French	4464	3472	1155	12024	10270	1670	6650	5300	763
Hindi	500	870	202	756	909	348	2211	2868	705
Thai	1950	1074	2780	6036	736	7678	2263	554	2917

Lingo	May			June		
	Misinformation			Misinformation		
	False	Partially False	Misleading	False	Partially False	Misleading
Spanish	57821	8214	7107	54965	8828	6759
German	1076	1426	4430	616	657	2028
Japanese	8984	12324	18125	1741	2496	3389
Indonesian	12695	4574	1805	9114	3038	1000
English	140494	128326	119391	135172	101896	109483
French	8475	7667	842	4952	3535	483
Hindi	4560	6057	1343	2501	2739	751
Thai	2825	470	1830	2103	486	3122

Table 4: Language-wise predicted misinformation labels of tweets

Test Data	Actual Label	Prediction	Accuracy(/)
Dr. Megha Vyas from Pune, India died due to COVID-19 while treating COVID patients.	False	False	
El plátano bloquea "la entrada celular del COVID-19"	False	False	
Asymptomatic people are very rarely contagious, said the WHO.	Partially False	Partially False	
Patanjali Coronil drops can help cure coronavirus.	Misleading	Misleading	
El medicamento contra piojos sirve como tratamiento contra Covid-19.	Misleading	False	

Table 5: Misinformation data examples along with model's prediction and actual label

Thai have the least of all. Following more specific observation made with respect to the languages:

- English: The misinformation distribution for English data, indicates that there is a majority of **False** tweets during the five months, whereas the distribution of **Misleading** labelled data is slightly less than as compared to **False** labelled data. **Partially False** labelled tweets are moderately distributed, as in month April we can see that there is a greater number with respect to other months.

- Spanish: From the distribution graph, Spanish tweets have greater frequency of **False** labelled tweets, whereas the **Misleading** tweets and **Partially False** tweets shows almost same number of tweet across the five months.

- German: There was a surge of **Misleading** labelled tweets during the month February, and the count remained the same throughout the five months. There was also an increase in **Partially False** tweets in March but it decreased in successive months, leading to minor **False** labelled tweets.

- Japanese: In the graph of language wise-distribution4, it can be seem that on an average throughout the five months, approx 20% of Japanese tweets are labelled **False**, similarly approx 30% of the Japanese tweets are labelled **Partially False**, leading to the majority of 50% data are labelled as **Misleading**. We can also see that there was a huge increase in **Misleading** tweets in March, tweeted in

Japanese language.

- Indonesian: In our distribution for Indonesian tweets approximately 10% of tweets are labelled as **Misleading** and in contrary there is a large distribution of **False** labelled tweets. Approximately 34% of the data in Indonesian dialect is labelled as **Partially False** throughout the five months.

- French: Figure4 shows the misinformation distribution across all of the five months in the French tweets. The largest majority of the tweets were classified as **False** misinformation. Among **Partially false** and **Misleading**, the least number of tweets were labelled as **Misleading**.

- Hindi: The frequency of Hindi tweets is low in the dataset used in our experiment. Yet, our model can predict or label Hindi tweets. Tweets in Hindi have low numbers of **Misleading** tweets, whereas the **Partially False** tweets class has a great frequency. **False** labelled tweets are slightly low compared to **Partially False** tweets in this dialect.

- Thai: The distribution of Thai tweets, shows that our model prediction is majorly oriented towards the **Misleading** tweets. The distribution of **Misleading** labelled tweets it the greatest among the labelled classes, in contrast to **Partially False** tweets. **False** labelled tweets are comparatively moderate in this language.

7 Conclusion and Future Work

In this paper, we presented a BERT based multilingual model for analysing COVID-19 related multilingual tweets. We performed a detailed systematic survey for detecting disinformation spread on the social media platform- Twitter. We were able to detect misinformation distribution across eight major languages and presented a quantified magnitude of misinformation distributed across different languages in last 5 months. We also demonstrated that our single multilingual CMTA framework performed significantly well as compared to the monolingual misinformation detection models. We strongly believe that our model can help in filtration of misinformation and factual data present in multiple languages during the pandemic.

In future, we aim at collecting more annotated training data and performing analysis of a larger multilingual dataset to gain deeper understanding. We aim at improving our model's robustness and contextual understanding for better performance in the classification task. Since analysis was done on a limited dataset the results cannot be generalised. We hope that through our work researchers could gain more deeper insights about misinformation spread across major languages and hence utilizing the information in building more reliable social media platform.

References

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, and Preslav Nakov. 2020. Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society.

J Scott Brennen, Felix Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. Types, sources, and claims of covid-19 misinformation. *Reuters Institute*, 7.

Ms D Brindha, R Jayaseelan, and S Kadeswara. 2020. Social media reigned by information or misinformation about covid-19: a phenomenological study.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004*.

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*.

Arkin R Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. 2020. Drink bleach or do what now? covid-hera: A dataset for risk-informed health decision making in the presence of covid19 misinformation.

Gunther Eysenbach, John Powell, Oliver Kuss, and Eun-Ryoung Sa. 2002. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *Jama*, 287(20):2691–2700.

Nuno Fernandes. 2020. Economic effects of coronavirus outbreak (covid-19) on the world economy. *Available at SSRN 3557504*.

Sheera Frenkel, Davey Alba, and Raymond Zhong. 2020. Surge of virus misinformation stumps facebook and twitter. *The New York Times*.

Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. 2020. Assessing the risks of" infodemics" in response to covid-19 epidemics. *arXiv preprint arXiv:2004.03997*.

Oguzhan Gencoglu and Mathias Gruber. 2020. Causal modeling of twitter activity during covid-19. *arXiv preprint arXiv:2005.07952*.

Peter Hernon. 1995. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government information quarterly*, 12(2):133–139.

P Hitlin and K Olmstead. 2018. The science people see on social media. pew research center.

Binxuan Huang and Kathleen M Carley. 2020. Disinformation and misinformation on twitter during the novel coronavirus outbreak. *arXiv preprint arXiv:2006.04278*.

Nicholas E Ingraham and Christopher J Tignanelli. 2020. Fact versus science fiction: fighting coronavirus disease 2019 requires the wisdom to know the difference. *Critical Care Explorations*, 2(4).

Yohei Kikuta. 2019. Bert pretrained model trained on japanese wikipedia articles. https://github.com/yoheikikuta/bert-japanese.

Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W Akl, and Khalil Baddour. 2020. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3).

Christian E Lopez, Malolan Vasu, and Caleb Gallemore. 2020. Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset. *arXiv preprint arXiv:2003.10359*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Katerina Eva Matsa and Elisa Shearer. 2018. News use across social media platforms 2018— pew research center. *Journalism and Media*.

World Health Organization et al. 2020. Coronavirus disease 2019 (covid-19): situation report, 188.

Esteban Ortiz-Ospina. 2020. The rise of social media. Technical report.

Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780.

2020 Poynter Institute. 2020. The international fact-checking network.

Josimar E Chire Saire and Roberto C Navarro. 2020. What is the people posting about symptoms related to coronavirus in bogota, colombia? *arXiv preprint arXiv:2003.11159*.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*.

Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, and Yan Liu. 2020. Covid-19 on social media: Analyzing misinformation in twitter conversations. *arXiv preprint arXiv:2003.12309*.

Megha Sharma, Kapil Yadav, Nitika Yadav, and Keith C Ferdinand. 2017. Zika virus pandemic—analysis of facebook as a social media health information platform. *American journal of infection control*, 45(3):301–302.

Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*.

Julio Torales, Marcelo O'Higgins, João Mauricio Castaldelli-Maia, and Antonio Ventriglio. 2020. The outbreak of covid-19 coronavirus and its impact on global mental health. *International Journal of Social Psychiatry*, page 0020764020915212.

Charin Polpanumas Arthit Suriyawongkul Lalita Lowphansirikul Pattarawat Chormai Wannaphong Phatthiyaphaibun, Korakot Chaovavanich. 2016. PyThaiNLP: Thai Natural Language Processing in Python.

Justin Wilford, Kathryn Osann, and Lari Wenzel. 2018. Social media use among parents of young childhood cancer survivors. *Journal of Oncology Navigation & Survivorship*, 9(1).

John Zarocostas. 2020. How to fight an infodemic. *The Lancet*, 395(10225):676.

A Appendix

A.1 CMTA vs Monolingual BERT Models

In this section, we have presented a comparative performance study of various monolingual BERT models with respect to our proposed multilingual

CMTA model for the misinformation detection task. We investigated eight monolingual BERT model[18], namely, 'English', 'Spanish', 'French', 'Germann', 'Japanese', 'Hindi' 'Thai[19]' and 'Indonesian'.

Data Processing: We utilized the same 9,502 tweets distributed across 3 misinformation classes for training the monolingual models. Since our dataset was consist of tweets in English and Spanish language; we translated the tweets into eight languages for training each of the eight monolingual model. We used Google Translator API[20] for converting the tweets into a particular language.

Experiment and Result: We experimented the multi-lingual data with their respective linguistic based BERT models. We set the model training parameters same as the CMTA model, and preprocessed the data as stated previously. Each of the monolingual model was fine-tuned for 10 ephocs with batch size of 32. using the classification dataset of their respective language. EnglishBERT scored an F1-score of 77.9% on the English tweets, with recall rate of 74.18%. This possible reason could be that it is heavily trained on English Corpus. From huggingface's model library we got SpanishBERT. The model scored an F1-score of 76.2% with recall rate of 72.02% and precision 80.9%. For French tweets we used CamemBERT(Martin et al., 2019) from huggingface. The CamemBERT scored an F1-score of 76.32%, with recall rate of 71.45% and precision 81.91%. GermanBERT showed a significant results on German-basesd tweets. It had a precision of 80.61% with recall rate of 71.43%, resulting to an F1-score of 75.74%. JapaneseBERT derived from the paper (Kikuta, 2019), is 79.56% precise on Japanese tweets with recall rate of 65.36% and F1-score of 71.76%. HindiBERT model had an F1-score of 71.95%, 79.56% precise with recall rate 65.68%. ThaiBERT scored an F1-score of 72.11%, being 79.11% precise with recall rate 66.25% IndonesianBERT is 78.96% precise, recall rate of 65.66%, resulting to an F1-score of 71.69%. Based on the experiment results, we can strongly suggest that the multilingual CMTA model was able to generalize smoothly on the dataset and it's performance was equivalent to the monolingual models.

ModelsMetrics	Precision	Recall	F1-score
EnglishBERT	82.03	74.18	77.90
SpanishBERT	80.9	72.02	76.20
CamemBERT	81.91	71.45	76.32
GermanBERT	80.61	71.43	75.74
JapaneseBERT	79.56	65.36	71.76
HindiBERT	79.56	65.68	71.95
ThaiBERT	79.11	66.25	72.11
IndonesianBERT	78.96	65.66	71.69
CMTA	81.52	74.40	77.79

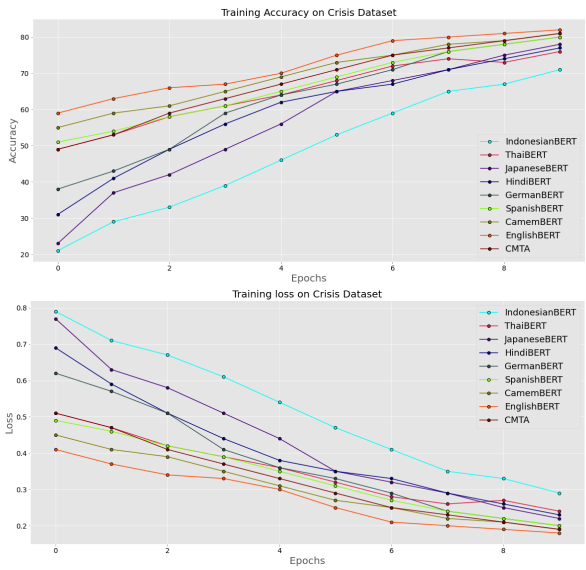

Figure 3: Training Accuracy(Upper) and Training loss(Lower)

[18]Pretrained model from https://huggingface.co/models
[19]ThaiBERT from https://github.com/ThAIKeras/bert
[20]Please refer https://cloud.google.com/translate/docs

Our Rating	IFCN(Poynter) Rating	Misinformation	Explanation
False	False	The border between France and Belgium will be closed.	French and Belgian authorities denied it.
	Four pinocchios	Trump's effort to blame Obama for sluggish coronavirus testing.	There was no "Obama rule," just draft guidance that never took effect and was withdrawn before President Trump took office.
	Inaccurate	Elisa Granato, the first volunteer in the first Europe human tria of a COVID-19 vaccine, has died.	Elisa Granato, the first volunteer in the first Europe human trial of a COVID-19 vaccine, has died.
Partially False	Partially False	Media shows a Florida beach full of people while it's empty.	The different videos were not shot at the same time. The beaches are empty when they are closed.
	Two Pinocchios	The bill for a coronavirus test in the US is $3.000	The CDC is not making people pay the test by now.
	Partly False	Salty and sour foods cause the "body of the COVID-19 virus" to explode and dissolve.	"Consuming fruit juices or gargling with warm water and salt does not protect or kill COVID-19," the World Health Organization Philippines told VERA Files.
Misleading	Misleading	A clip from Mexico depicts the dumping of coronavirus patients corpses into the sea.	Misbar's investigation of the video revealed that it does not depict the dumping of coronavirus patients corpses in Mexico, but rather paratroopers landing from a Russian MI 26 helicopter.
	No Evidence	Media uses photos of puppets on patient stretchers to scare then public.	There is no evidence that any media outlet used this photo for their reporting about COVID-19. Its origin is unclear, maybe it was shot in Mexico and shows a medical training session.
	Mostly False	Coronavirus does not affect people with 'O+' blood type.	The post claiming coronavirus does not affect people with 'O+' blood type is misleading.

Table 6: Misinformation Dataset

Figure 4: Month-wise Disinformation Distribution in Languages.

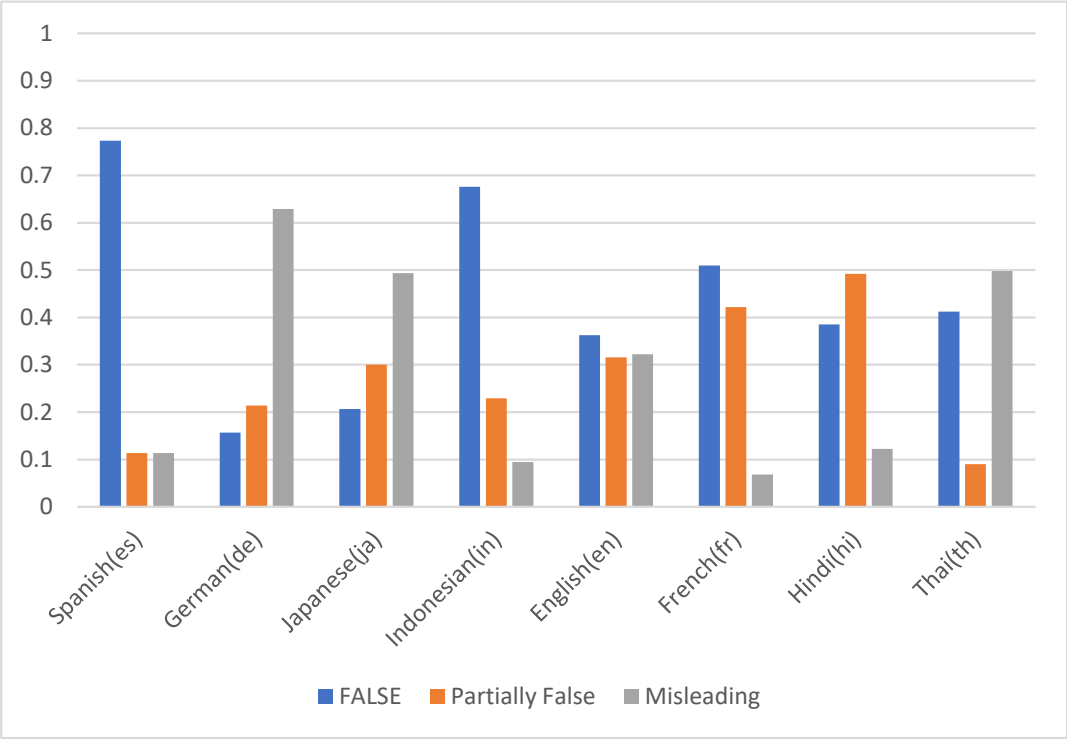

Figure 5: Language-wise Disinformation Distribution.

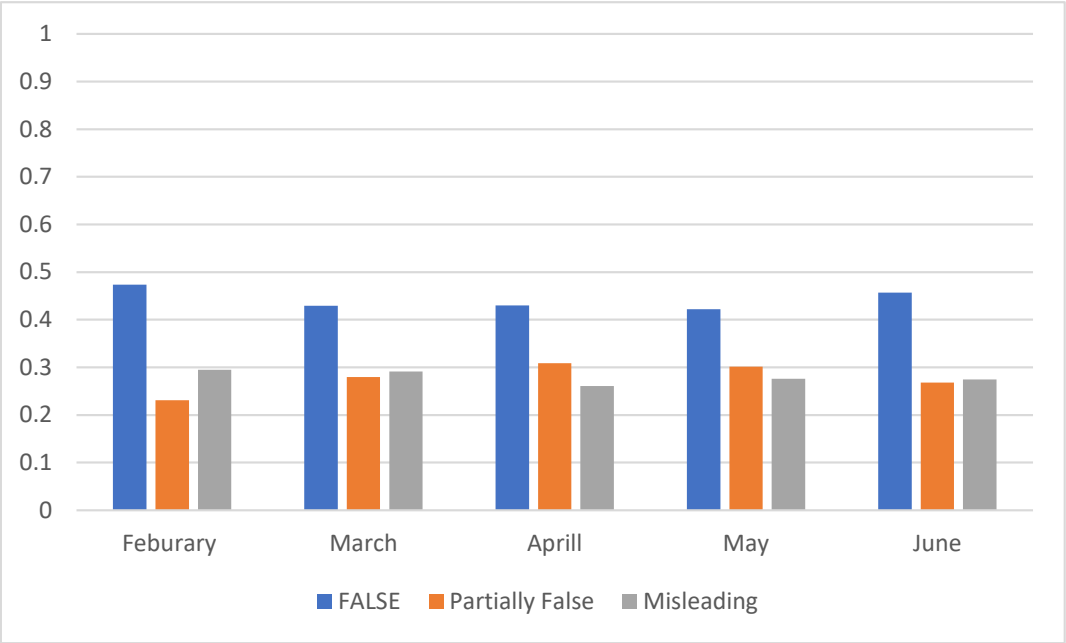

Figure 6: Month-wise Disinformation Distribution.