A Comparative Study of Pre-trained Language Models to Filter Informative Code-mixed Data on Social Media during Disasters

Hossein Salemi
Information Sciences & Technology Department
George Mason University
Fairfax, Virginia, USA
hsalemi@gmu.edu

Yasas Senarath
Information Sciences & Technology Department
George Mason University
Fairfax, Virginia, USA
ywijesu@gmu.edu

Hemant Purohit
Information Sciences & Technology Department
George Mason University
Fairfax, Virginia, USA
hpurohit@gmu.edu

ABSTRACT

Social media can inform response agencies during disasters to help affected people. However, filtering informative messages from social media content is challenging due to the ungrammatical text, out-of-vocabulary words, etc., that limit the context interpretation of messages. Further, there has been limited exploration of the challenge of code-mixing (using words from another language in a given text of one language) in user-generated content during disasters. Hence, we proposed a new code-mixed dataset of tweets related to the 2017 Iran-Iraq Earthquake and annotated them based on their informativeness characteristics. Additionally, we have evaluated the performance of state-of-the-art pre-trained language models: mBERT, RoBERTa, and XLM-R, on the proposed dataset. The results show that mBERT (with F1 score of 72%) overweighs the other models in classifying informative code-mixed messages. Moreover, we analyzed some patterns of exploiting code-mixing by users, which can help future works in developing these models.

Keywords

Code-mixing, Crisis Informatics, Language Model, Multilingual Data

INTRODUCTION

Social media platforms are widely used by the public during natural disasters such as earthquakes, floods, and hurricanes to report on a variety of information (Purohit and Peterson 2020; Imran, Castillo, Diaz, et al. 2015; Vieweg 2012). Type of such information vary from injured or dead people, damaged infrastructures and buildings, requesting help and support by affected people to posting on supportive efforts such as donation and voluntary activities. A wide range of studies have been conducted on using social media for crisis management that demonstrate the potential of these messages to provide actionable information, which, in turn, helps disaster response agencies in accelerating disaster relief efforts (Imran, Castillo, Diaz, et al. 2015; Purohit, Castillo, et al. 2013). However, processing social media content is challenged by a variety of content characteristics involving language ambiguities,

erroneous grammar usage, evolving terminologies, etc., some of which have been addressed in the prior works. For example, parsing noisy and informal content to filter and rank relevant messages, as well as to classify them into different categories (Purohit and Peterson 2020; Imran, Castillo, Diaz, et al. 2015; Reuter et al. 2018). Researchers have developed methods to identify and categorize relevant social media messages across disaster events using the Artificial Intelligence (AI) techniques of Natural Language Processing (NLP) and machine learning (Krishnan, Purohit, et al. 2020b; Ullah et al. 2021; Devaraj et al. 2020). Additionally, researchers have started exploring and introducing pre-trained language models such as BERT, to identify actionable messages (Zhou et al. 2022; Kayi et al. 2020).

Despite the growing use of NLP and machine learning methods in processing social media for crisis management, there are still challenges in analyzing all types of language variations. Multilingual users sometimes use multiple languages in the same utterance text when they post it on social media. This linguistic phenomenon of using multiple languages in a content, which is called code-mixing, can be occurred in inter-sentential, intra-sentential and word-level boundaries (Barman, Wagner, et al. 2016). *Code-switching* is another term which is used interchangeably with *code-mixing*. Although they have some differences in their definition, here we refer code-mixing to any type of mixing two or more languages in one utterance of a social media message. Using multiple languages together, causes code-mixing to present more challenges to NLP solutions for various tasks such as language identification, part of speech tagging, parsing, and translation (Çetinoğlu et al. 2016). Some samples of code-mixed content in the proposed dataset and whether they can be informative or not for rescue and response teams are described in Table 1. Sample 1 and 3 contain informative content, as we define in Data Annotation section, while Sample 2 is not informative, although it is related to the event.

Current state-of-the-art language models in NLP have been mainly trained on monolingual or multilingual corpora, and so they have not been trained for code-mixed content. What can cause processing code-mixed content to be more challenging is that in addition to linguistic aspects of code-mixed content, the socio-linguistic characteristics of users can affect the patterns of code-mixed text (Doğruöz et al. 2021), but current language models only take the linguistic aspect of text into consideration. Moreover, training large language models for this purpose requires providing a big training set of labeled code-mixed data which is time-consuming and effort-intensive. However, massive pre-trained multilingual language models enable us to use their capabilities for modeling code-mixed content by fine-tuning them with small code-mixed dataset (Doğruöz et al. 2021).

In this paper, we have proposed a novel code-mixed dataset of tweets posted during an earthquake disaster, and have evaluated the capability of state-of-the-art multilingual pre-trained language models in filtering informative code-mixed tweets related to the disaster event. The main contributions of our paper are the following:

- To our knowledge, this is the first curated dataset of Persian-English code-mixed data on social media for crisis informatics research.
- Using the data generated on social media during the 2017 Iran-Iraq earthquake event, we curate and label the code-mixed dataset for informativeness of tweets to aid response agencies in helping affected people.
- We evaluate the performance of three state-of-the-art pre-trained language models (mBERT, RoBERTa, and XLM-R) on the classification task for the proposed code-mixed labeled dataset and present novel insights for building efficient classification models of code-mixed data.

Paper Organization: In the rest of this paper, the related works in exploiting social media content for disaster management by using machine learning methods and language models are described and the efforts for dealing with code-mixed data are reviewed. After that, we propose our methodology for collecting and filtering our proposed dataset, which is followed by defining our guidelines for annotating the dataset and describing the process of fine-tuning the pre-defined language models. The experiments and results are presented in the next section, and then we conduct the prediction error analysis to reveal the shortcomings of the models in the classification of code-mixed content. We also analyze different patterns of code-mixed content which have been used by users in our dataset to provide directions for modeling code-mixed content by language models in future works.

RELATED WORK

Collecting and storing a huge amount of social media content for processing is extensively challenging, specially during disaster situations in which response time is very critical. Social stream analytics systems, such as CitizenHelper (Pandey and Purohit 2018) and AIDR (Imran, Castillo, Lucas, et al. 2014) which have been designed for disaster informatics applications can be useful for this purpose. SMDRM (Lorini, Salamon, et al. 2021) is

Table 1. Code-mixed samples of the dataset. In translation, the English part of the sample is highlighted in bold. The column labeled "Info." indicates whether a tweet is informative or not.

ID	Sample	Info.	Reason
1	Right now. Our helping group in Kermanshah provenance and in earthquake areas توزیع کمک های مردمی... **Translation: Right now. Our helping group in Kermanshah provenance and in earthquake areas** Distribution of public aid ...	YES	This tweet is informative since it can help disaster management agencies to identify the regions in which public donations have been distributed, and so they should focus on the other regions for more efficient distribution of aid.
2	They pulled this baby out from under the debris after 3 days! خنده ت چیکار کنم؟ مگه میشه با این خنده ی تو گریه نکرد ؟ #Iranearthquake #کرمانشاه #زلزله_کرمانشاه **Translation: They pulled this baby out from under the debris after 3 days!** What should I do with your laugh? Is it possible not to cry with your laugh? #Kermanshah_Earthquake #Kermanshah **#Iran-earthquake**	NO	In this tweet, the user has reported the rescue teams' effort in rescuing someone from under the debris and sympathized with the rescued person, so it cannot be considered as informative content for rescue and response teams.
3	توزیع اقلام اهدایی در میان زلزله‌زده‌ها پس از نیازسنجی اولیه @ Kermanshah, Iran **Translation:** Distribution of donated items among affected people after initial needs assessment @ **Kermanshah, Iran**	YES	This tweet, in addition to having information about the distribution of donated items (like Sample 1), reveals the necessity of considering code-mixed content, as while the tweet is in Persian, the location has been mentioned in English

another operational platform which has been developed to analyze extracted social media content for disaster risk management. In practice, for analyzing text and images retrieved from social media, while image analysis methods can be used for situational awareness and damage assessment processes (Rufolo et al. 2021; Lorini, Rufolo, et al. 2022), NLP techniques and machine learning methods have provided automated solutions for the classification, prioritization, and summarization of time-sensitive and actionable content of social media during disasters (Purohit and Peterson 2020). In (Caragea et al. 2016), a convolutional neural networks (CNN) model was used to identify informative messages in disaster events, while (Kruspe 2019) exploited few-shot method for filtering disaster-related tweets. Unsupervised domain adaptation technique is another method which was used in (Krishnan, Purohit, et al. 2020b) to train filtering model for new crisis events based on data observed in past crisis events. Also, (Spiliopoulou et al. 2020) utilized adversarial neural network to improve filtering performance by removing event-specific biases from the trained model. Moreover, massive language models such as XLM-R have been used in recent works to classify crisis-related tweets (Krishnan, Purohit, et al. 2020a).

Table 2. An overview of widely used language models trained on multiple languages.

Model	Description	Number of Languages
mBERT (Devlin et al. 2019)	A multi-lingual language model that is trained with the objective of masked language modeling and next sentence prediction using data from Wikipedia.	104
XLM-R (Conneau et al. 2019)	A multi-lingual language model trained using masked language modeling on 2.5 TB of newly created and cleaned CommonCrawl data.	100

Pre-trained language models based on transformers (Vaswani et al. 2017) have significantly improved performance for many natural language processing tasks. BERT (Devlin et al. 2019) has shown to perform well in multitude of tasks in English language. However, since it is only pre-trained on a large English corpus it is not generalizable for fine-tuning on other language tasks. Therefore, the authors additionally proposed a multilingual version of BERT (mBERT) that is trained on corpora containing documents of 104 languages. Moreover, it was found that the training process used by BERT was not optimized, and (Liu et al. 2019) proposed an alternative robust version called *RoBERTa*. Additionally, (Conneau et al. 2019) proposed an extension to RoBERTa called XLM-R that supports multiple languages by training on a CommonCrawl corpus of one hundred languages. These transformer based multi-lingual models have seen to perform well on existing multilingual dataset tasks. The Table 2 provides a summary of the multilingual language models previously discussed.

The experiment in (Pires et al. 2019) showed that fine-tuning mBERT with code-mixed data can result a good performance in Part of Speech Tagging task on code-mixed content. This observation was confirmed by GLUECoS (Khanuja et al. 2020), an evaluation benchmark for code-mixing on NLP tasks including Language Identification (LID), Part of Speech (POS) tagging, Named Entity Recognition (NER), Sentiment Analysis, Question Answering and a code-switched Natural Language Inference task. The experiments demonstrated that a modified version of mBERT that had been fine-tuned on a mixture of synthetically generated code-switched data and real code-switched data outperformed cross-lingual embedding techniques for most datasets.

There have been increasing interest in proposing language models and technologies for modeling code-mixed content in recent years. These works have concentrated on a wide range of NLP tasks, including Part of Speech Tagging (Barman, Wagner, et al. 2016; Vyas et al. 2014), Language Identification (Barman, Das, et al. 2014; Dowlagar and Mamidi 2021; Rani et al. 2022), Named Entity Recognition (V. Singh et al. 2018), Sentiment Analysis (Chakravarthi et al. 2020), and Translation (Gautam et al. 2021). Furthermore, from the perspective of dealing with code-mixed content, some works simplified the complicated structure of code-mixed data by transforming the code-mixed content into a monolingual script by back-transliteration (Dowlagar and Mamidi 2021; Gautam et al. 2021) or translation (Bhoi et al. 2020). (Baral et al. 2022 has also used translation and transliteration representations of code-mixed content to develop a new loss function which can enable BERT model to be trained on code-mixed data. However, the performance of these techniques extensively depends on the accuracy of transliteration and translation methods (Ghosh et al. 2018). Adapter-based fine tuning is another method which has been used in (Rathnayake et al. 2022) to fine-tune pre-trained language models on code-mixed content by adding language adapters to the models' architecture. Moreover, Transfer Learning approaches have been widely used to exploit pre-trained language models for analyzing code-mixed data (Krishnan, Anastasopoulos, et al. 2021; Aguilar and Solorio 2020; Lu et al. 2022). Code-switching can also be used in cross-lingual Transfer Learning, although there are some limitations such as grammatical inconsistency in using this approach (Feng et al. 2022). Also, a survey of computational techniques for modeling code-mixed data in NLP and speech has been presented in (Sitaram et al. 2019).

METHODOLOGY

In this section, we describe our methodology for collecting and creating our dataset, in addition to defining our terminology for filtering and annotating the dataset. Then, the process of fine-tuning state-of-the-art pre-trained language models which are used in this work is explained.

Dataset Description

In order to evaluate the performance of state-of-the-art pre-trained language models in text classification task on code-mixed text, we have created a code-mixed dataset containing 1758 tweets related to 2017 Iran-Iraq earthquake. All tweets in the proposed datasets have been checked to contain code-mixed Persian-English content. Additionally, they have been annotated manually to represent whether they have informative content to help disaster response teams during disaster situation or not.

Data Collection

In order to create our dataset, we have queried Twitter API to retrieve tweets related to 2017 Iran-Iraq Earthquake in the period of November 12, 2017 to December 12, 2017, based on a list of keywords related to the event. In this project we have used TweetKit[1], a python Twitter client focused mainly on Twitter API for academic research, for retrieving tweets. Since our goal was to retrieve Persian-English code-mixed tweets, we had used two lists of

[1] https://github.com/ysenarath/tweetkit

Table 3. The number of tweets in the collected and filtered dataset

Dataset Name	Collected Dataset	Extended Dataset	Code-mixed	code-mixed Dataset	Cleaned Dataset
Persian Dataset	367091	592725	1654	3962	1758
English Dataset	189778	328758	2308		

keywords in Persian and English. This enabled us to collect two datasets containing Persian and English tweets. With this procedure, we created our collected Persian Dataset with 367091 tweets and collected English Dataset with 189778 tweets.

Data Filtering and Cleaning

To prepare the collected dataset for manual annotation, we filtered the dataset with the following steps:

1. Since each tweet object retrieved from Twitter API may contain some tweets in replying to the main tweet and these replies may convey informative content, we extended our datasets by extracting and adding these reply tweets to the collected datasets.
2. After that, in order to find code-mixed tweets in the Persian Dataset which are related to the event, we utilized an English dictionary which contains English words related to the event to filter the extended Persian Dataset and create a code-mixed set of tweets. The same procedure was applied on extended English dataset by using a Persian dictionary to create another code-mixed set of tweets. Finally, we merged these to generated code-mixed set of tweets to create the final code-mixed dataset.
3. Since retweets do not add any informative content they were removed from the dataset. Additionally, irrelevant tweets to the event and tweets which were not in Persian and English were removed from the dataset manually.
4. Since it is not permitted to use space in hashtags, multiple-word hashtags in English are commonly written in camel-case format, such as "#iranEarthquake", but since there is no upper-case alphabet in Persian, multiple-word hashtags are made by separating its words by "_", such as "#زلزله_کرمانشاه". Therefore, we replaced every "_" with a space in the text.
5. We cleaned the tweets by converting the text to lowercase, removing emojis, apostrophes, mentions, URL links, and punctuation in English and Persian.
6. According to our observation of tweets, using English hashtags in Persian text and using Persian hashtags in English text were common patterns in code-mixed tweets. Also, it was common that users used hashtags to emphasis on the resources that affected people need. Therefore, we kept hashtags by removing "#", since they added informative content to the tweets.
7. After cleaning tweets, it was possible to have some tweets with a same text. These duplicates were removed from the final dataset.

The final dataset resulted from our filtering and cleaning procedure contains 1758 Persian-English code-mixed tweets. Figure 1 demonstrates our procedure for collecting and filtering our dataset from Twitter API. Furthermore, the size of dataset in different steps of our procedure are shown in Table 3.

Data Annotation

In the proposed dataset, tweets have been annotated and categorized into *Informative* or *Non-Informative* based on the following definition:

Informative: In order to create a baseline for annotating our dataset, we consider a tweet is informative, if it conveys any information which can help disaster response teams in acceleration and enhancement of their procedures. Based on the guidelines in (Alam et al. 2018; Purohit and Peterson 2020), the tweet is considered as *Informative* in this paper, if it reports one or more of the following items:

- Cautions, advice, and warnings
- Injured, dead, or affected people

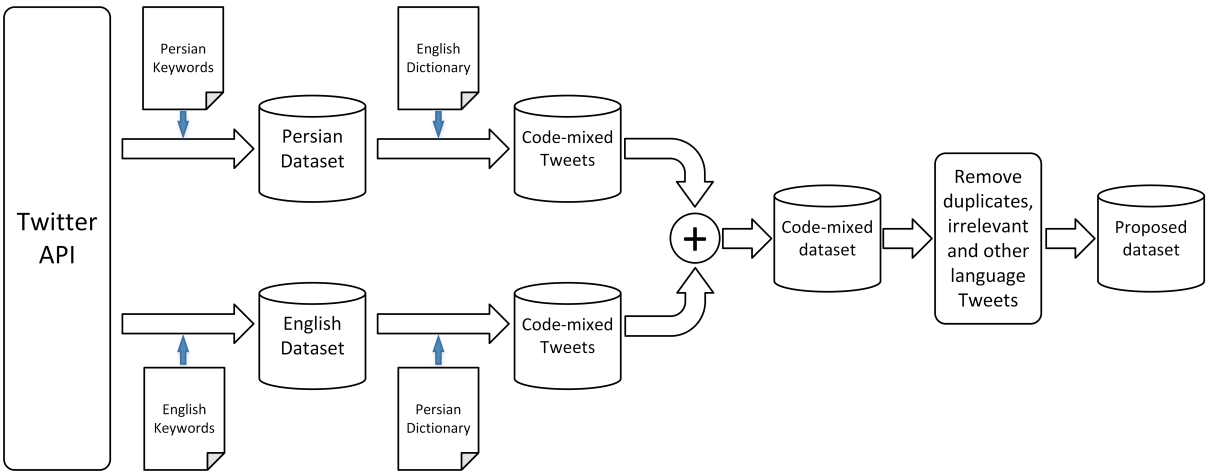

Figure 1. Data collection process of proposed dataset

- Resources need by affected people
- Rescue and volunteering activities, and obstacles against them
- Donation request or efforts, and obstacles against them
- Damaged houses, buildings, infrastructures, and monuments
- Blocked roads and connections

Non-Informative: In this dataset, in the filtering step all tweets that are not related to the event have been removed. So, all tweets which are not considered as *Informative*, are considered as *Non-Informative*.

These guidelines have been followed by one annotator in the annotation process. The annotated dataset contains 1758 code-mixed Persian-English tweets among which 709 tweets have been annotated to be *Informative* and the remaining categorized to *Non-Informative* class.

Fine-Tuning pre-trained language models

In recent years, state-of-the-art pre-trained language models have been widely used in different NLP tasks, such as text classification. Fine-tuning these models can enable us to modify their weights according to the context of our dataset while we can use their powerful capabilities in language modeling tasks such as text classification. The fine-tuning process helps the model to learn about the task specific patterns of the la nguage. We created classification models for filtering informative code-mixed content on social media by fine-tuning transformer based pre-trained language models.

The transformer is an encoder–decoder model architecture that adopts the attention mechanism for attending to the salient portions of the input. The encoder and decoder are composed of multiple scaled dot-product attention units called attention-heads in each layer (Multi-head attention) to attend to information from different representations (Vaswani et al. 2017). This proposed model architecture is used for creating a large pre-trained language model by learning on a large corpus in a semi-supervised fashion using objectives such as masked language modeling and next sentence prediction (Devlin et al. 2019).

We first identified several pre-trained language models that use transformer deep learning architecture (refer to the experiments section for information on identified models). Then we extended the language model identified previously by adding a fully-connected (output) layer after the language model. We used the encoding provided by the $[cls]$ token in the language model as the input for the output layer. Finally, we utilized the Adam optimizer with weight decay regularization to train the extended model using the labeled data.

EXPERIMENTS AND RESULTS

To evaluate how these models can help us in classifying and filtering informative code-mixed tweets, we conducted experiments with three pre-trained language models: 1) RoBERTa (Liu et al. 2019), 2) XLM-R (Conneau et al. 2019) and 3) mBERT (BERT multilingual base model) (Devlin et al. 2019).

Table 4. Performance of different fine-tuned models discussed in Results and Analysis Section. The value formatting is in the form $mean \pm std\%$.

	Accuracy	Recall	Precision	F1-Score
RoBERTa	76 ± 3	50 ± 12	84 ± 5	62 ± 9
XLM-R	77 ± 8	60 ± 30	72 ± 26	63 ± 26
mBERT	81 ± 8	72 ± 26	71 ± 25	72 ± 25

RoBERTa: A robust transformers model pretrained on a large corpus of English data using the masked language modeling (MLM) objective.

XLM-R: A multilingual version of RoBERTa that is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages including English and Persian.

mBERT: A pre-trained transformers model using content in 104 languages on Wikipedia with masked language modeling (MLM) and next sentence prediction (NSP) objective.

XLM-R and mBERT have achieved state-of-the-art results on XTREME, a multilingual and multi-task benchmark for cross-lingual transfer learning evaluation (Hu et al. 2020), and some recent works have addressed their capabilities in modeling code-mixed data (Santy et al. 2021; Winata et al. 2021). Thus, in this paper, we have selected XLM-R and mBERT for evaluation, since our dataset contains code-mixed content and these models are pre-trained on multiple languages, so we anticipated that they can model our data better than monolingual models. Additionally, RoBERTa has been selected to compare the performance of a monolingual pre-trained language model with the other multilingual models.

Experiment setup

We conducted several experiments to understand the baseline/benchmark performance of the state-of-the-art deep-learning based model architectures identified p reviously. We took the proposed dataset as the input to the process of training and testing the model performance. We conducted 10-fold cross validation for measuring the performance of each model. We used the parameters proposed in previous studies for fine-tuning (3 epochs, learning rate of $5e^{-5}$, Adam optimizer with weight decay)(Devlin et al. 2019).

For measure the performance of the fine-tuned models, we used precision, recall, F1-score, and a ccuracy. The precision is the ratio of the true positives to total number of true predictions (i.e., sum of true positives and false positives). The recall is the ratio between true positives and the sum of true positives and false negatives. The F1-score is calculated by taking the harmonic mean of the precision and recall. The accuracy metric indicates the percentage of correct predictions (sum of true positives and true negatives).

Results and Analysis

Classification Performance

We present the performance of the fine-tuned language models in Table 4 . As results demonstrates, mBERT model with F1-score of 72% outperforms other language models in all evaluation measures, except precision in which RoBERTa have the higher performance. Furthermore, XLM-R shows a better F1-score than RoBERTa with 63% and 62% respectively. These results are consistent with previous studies that have demonstrated the superior performance of XLM-R and mBERT in multilingual tasks compared to language models trained on a single-language corpus. Since XLM-R is the multilingual version of RoBERTa, we can confirm that multilingual pre-trained language models outperform monolingual ones in modeling code-mixed data. The task-specific factors could be responsible for the better performance of mBERT based model over the XLM-R based model.

Prediction Error Analysis

Analyzing how the language models operate in prediction in our experiments can reveal new insights about the requirements for developing new language models for modeling code-mixed data. For this purpose, we have extracted and analyzed some examples of our testing dataset in which language models have not been able to classify correctly. They are shown in Table 5.

Sample 1 shows the inability of monolingual models in predicting code-mixed content, where the monolingual language model, RoBERTa , failed in classifying informative tweet, while the multilingual language models mBERT

Table 5. Examples of prediction error by the language models. In translation, the English part of the sample highlighted in bold.

ID	Sample	Informative	Prediction		
			mBERT	RoBERTa	XLM-R
1	سایت زلزله شناسی world earthquaks اعلام کرده پوسته زمین در منطقه زاگرس بشدت فعال و احتمال زلزله خیلی قوی در ۴۸ ساعت بعدی در این منطقه بالاست شدت زلزله بقدری زیاداست که ممکن است گسلهای مجاور را فعال کند **Translation:** The seismological site, **world earthquakes** announced that the earth's crust in the Zagros region is extensively active and the possibility of a very strong earthquake in the next 48 hours is high in this region The intensity of the earthquake is so high that it may activate the nearby faults	YES	YES	NO	YES
2	حجم خرابی حاصل از زلزله بالاست... خدا کند بیشتر از این نباشد earthquake **Translation:** The amount of damage caused by the earthquake is high... God forbid it be more than this **earthquake**	YES	NO	NO	NO
3	زلزله خدایا تا کی در این باغ قرار است درخت ها بر روی مردمش فرود آیند دوستای خوبم اگه همجوار کرمانشاه هستید به کمک مردم بشتابید earthquake earthquakeiran **Translation:** Earthquake Oh my God How long are the trees going to fall on its people in this garden My good friends, if you are near Kermanshah, hurry to help the people **earthquake earthquakeiran**	YES	NO	NO	NO
4	کمک به زلزله earthquake iran prayforiran **Translation:** Help to earthquake **earthquake iran prayforiran**	YES	YES	NO	NO
5	فیلم هوایی از شهرستان سرپل ذهاب قبل از زلزله kermanshahpeople **Translation:** Aerial video of Sarpol-e Zahab City before the earthquake **kermanshahpeople**	NO	YES	NO	NO
6	we just hope every single donated fund will reach the victims of the earthquake mr. ali daei we count on u godblessuall . iran فقط امیدواریم آقای علی دایی عزیزکه تمام اقلام ومبالغ اهدایی بنیازمندان زلزله کرمانشاه برسد مارویتان تماما حساب میکنیم دست حق بهمراهتان **Translation: Persian and English parts have the same meaning**	YES	YES	NO	NO
7	خبرنگار اجتماعی ایسنا: من را آورده دم خانه‌ش و می‌گوید دوروزه هیچکس اینجا نیامده موادغذایی نداریم حتی آبم نداریم... زلزله earthquake **Translation:** ISNA social reporter: He/She has brought me to his/her house and said it is two days that no one has come here we do not have edibles we do not even have water... earthquake **earthquake**	YES	NO	NO	NO

and XLM-R operated well. In Sample 2, while the tweet is Informative, no model could classify it right. It seems that this misclassification originates from the fact that the tweet is short and does not have enough information such as location. In sample 3, there is some non-informative sentences in addition to the informative part, which caused all models to predict wrong class. This can address the necessity of new intention methods for considering informative parts of text by transformers in future works.

There is no sentence in Sample 4 , which has been formed by some hashtags (likely to describe the photo in the tweet), but its tokens contain informative content. The mBERT model, unlike the other methods, was able to classify this sample. This can imply that mBERT relies on some specific token in prediction which can in turn make it vulnerable to resulting false positive in some samples, like Sample 5 where the tweet states the situation of the affected region before the earthquake explicitly, but the tweet is predicted as informative by mBERT incorrectly. Nevertheless, The capability of mBERT in predicting long code-mixed text is presented in Sample 6 where the tweet includes two long distance with the same meaning in both English and Persian. As it is shown, just mBERT predict this sample precisely.

In Sample 7, we observed another inability of the language models which has been likely caused due to informal and colloquial writing style of Persian text. In this tweet, the formal Persian sentence "آب هم نداریم", which means "We do not even have water" and it is informative part of the text, has been written in colloquial style "آبم نداریم". Moreover, it is possible in Persian to write multi-part words in different ways: using space character, half-space character, and no space between multiple parts. This can cause some issues in tokenizing the text by language models. For instance, in Sample 7, the word "مواد غذایی", which means "edibles" and can be informative as a need of affected people, has been written with no space between its parts "موادغذایی". Therefore, this inconsistency in Persian writing style can cause the models not to use these informative content, so that all models failed in classifying this sample.

Analysis of Corpus

The result showed that fine-tuning the pre-trained language models which have been used in this paper can have a relatively good performance on classifying informative code-mixed content, since they have been trained on multiple languages corpora and this enables them to model different languages of a code-mixed text. However, to achieve developing more accurate language models for this purpose, we need to take the characteristics of exploiting code-mixing by users into account. Therefore, in this section, we explain some patterns which have been followed by users in generating code-mixed content in our dataset. These patterns can be helpful for designing and developing more accurate models in future works. These patterns can be categorized as the following:

1. **The tweet in one language with hashtags in the other language:** It was observed that this was a common pattern in code-mixed samples in which the tweet written in one language contains hashtags in the other language, such as Samples 2, 3, 5, 7 in Table 5. In this case, although the main language forms the most tokens of the text, the embedded hashtags may provide more informative content to the language models since hashtags are usually selected to highlight the main purpose of the text or to reference a specific subject such as earthquake event.

2. **Using hashtags in both languages with no comment:** Sometimes, user post their tweets with no comment and just use hashtags. This pattern, which are commonly used for describing the photo shared in the tweet, may originate from the fact that the user has posted in a hurry or the photo contains self-descriptive information such as images of text. Sample 4 in Table 5 shows such a tweet.

3. **Translation of text in both languages:** In this pattern, the user translates their content in English to convey their intent to the international collocutors who do not know original language (Persian in our dataset) as it is shown in Sample 6 in Table 5. In Such tweets the amount of code-mixing of both languages are approximately balanced.

4. **The text is in one language, summarized to the other language:** The intent of the tweet can be similar to the translation pattern, but one part is just the summarization or description of the other part (e.g. Sample 1 in Table 1), and so two parts' tokens are not commonly balanced.

5. **Inserting English words in Persian text:** This is called insertional code-switching in which some words from one language are inserted into the morphosyntactic frame of the other language (Winford 2003). This pattern is common among bilingual users, specially when the words are entity names or have no equivalent word in the base language. For instance, in Sample 1 in Table 5 the name of the website in English (world earthquakes) has been inserted in the Persian text.

6. **Inter-sentential code-switching:** In this case, users switch to another language between sentences of the text. This can have similar style to those of patterns 3 and 4, but the sentences of the text may have different semantics.

CONCLUSION AND FUTURE WORKS

In this paper we proposed a novel dataset of Persian-English code-mixed tweets which were annotated to represent whether they are informative for response teams in a disaster event or not. This dataset enabled us to fine-tune state-of-the-art pre-trained language models and evaluate their performance in classifying informative code-mixed tweets. The result confirmed that multilingual language models are considerably able to model code-mixed data, and among them mBERT had the best performance.

To the best of our knowledge the proposed dataset is the first dataset for informative code-mixed tweets in disaster situations and can be used in future works. However, in order to achieve more accurate models we need to develop new language models with considering the following directions in future:

Considering socio-linguistic aspects of code-mixing: The way in which users generate code-mixed data depends on different social factors, such as user demographic, language proficiency of users in both languages, and the intent of users in generating code-mixed data. We described our observation about different patterns which users had followed in generating code-mixed data in our dataset. These patterns can be considered as guidelines for developing more precise models in future works.

Considering linguistic aspects of code-mixing: code-mixed text can be characterized with several linguistic metrics which have been discussed in (Srivastava and M. Singh 2021), such as Code-Mixing Index (CMI), Multilingual Index (MI), Point Switch Average (PS Avg), etc. Consideration of these metrics can make a better representation of code-mixed data for modeling.

Dataset improvement for training: Although pre-trained language models used in this paper can be fine-tuned with small datasets, providing the larger dataset can improve the performance of these models. Additionally, data augmentation techniques can be utilized for improving the size and quality of the dataset which is use for fine-tuning the models.

ACKNOWLEDGMENT

This work was supported in part by award # N-3Q22-002 by the Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D innovation, and workforce development (for more information about CCI, visit cyberinitiative.org). This project was supported by resources provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631).

REFERENCES

Aguilar, G. and Solorio, T. (July 2020). "From English to Code-Switching: Transfer Learning with Strong Morphological Clues". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8033–8044.

Alam, F., Ofli, F., and Imran, M. (June 2018). "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters". In: *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*. USA.

Baral, A., Baksy, A., Sarkar, A., D, D., and Joshi, A. M. (2022). "CalBERT - Code-Mixed Adaptive Language Representations Using BERT". In: *Proceedings of the AAAI 2022 Spring Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence (AAAI-MAKE 2022), Stanford University, Palo Alto, California, USA, March 21-23, 2022*. Ed. by A. Martin, K. Hinkelmann, H. Fill, A. Gerber, D. Lenat, R. Stolle, and F. van Harmelen. Vol. 3121. CEUR Workshop Proceedings. CEUR-WS.org.

Barman, U., Das, A., Wagner, J., and Foster, J. (Oct. 2014). "Code Mixing: A Challenge for Language Identification in the Language of Social Media". In: *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Doha, Qatar: Association for Computational Linguistics, pp. 13–23.

Barman, U., Wagner, J., and Foster, J. (Nov. 2016). "Part-of-speech Tagging of Code-mixed Social Media Content: Pipeline, Stacking and Joint Modelling". In: *Proceedings of the Second Workshop on Computational Approaches to Code Switching*. Austin, Texas: Association for Computational Linguistics, pp. 30–39.

Bhoi, A., Pujari, S. P., and Balabantaray, R. C. (2020). "A deep learning-based social media text analysis framework for disaster resource management". In: *Social Network Analysis and Mining* 10.1, pp. 1–14.

Caragea, C., Silvescu, A., and Tapia, A. H. (2016). "Identifying informative messages in disaster events using convolutional neural networks". In: *International conference on information systems for crisis response and management*, pp. 137–147.

Çetinoğlu, Ö., Schulz, S., and Vu, N. T. (Nov. 2016). "Challenges of Computational Processing of Code-Switching". In: *Proceedings of the Second Workshop on Computational Approaches to Code Switching*. Austin, Texas: Association for Computational Linguistics, pp. 1–11.

Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., and McCrae, J. P. (May 2020). "Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text". English. In: *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. Marseille, France: European Language Resources association, pp. 202–210.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Annual Meeting of the Association for Computational Linguistics*.

Devaraj, A., Murthy, D., and Dontula, A. (2020). "Machine-learning methods for identifying social media-based requests for urgent help during hurricanes". In: *International Journal of Disaster Risk Reduction* 51, p. 101757.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Doğruöz, A. S., Sitaram, S., Bullock, B. E., and Toribio, A. J. (Aug. 2021). "A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 1654–1666.

Dowlagar, S. and Mamidi, R. (Sept. 2021). "A Pre-trained Transformer and CNN Model with Joint Language ID and Part-of-Speech Tagging for Code-Mixed Social-Media Text". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd., pp. 367–374.

Feng, Y., Li, F., and Koehn, P. (Dec. 2022). "Toward the Limitation of Code-Switching in Cross-Lingual Transfer". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5966–5971.

Gautam, D., Kodali, P., Gupta, K., Goel, A., Shrivastava, M., and Kumaraguru, P. (June 2021). "CoMeT: Towards Code-Mixed Translation Using Parallel Monolingual Sentences". In: *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. Online: Association for Computational Linguistics, pp. 47–55.

Ghosh, S., Ghosh, K., Ganguly, D., Chakraborty, T., Jones, G. J. F., Moens, M.-F., and Imran, M. (Aug. 2018). "Exploitation of Social Media for Emergency Relief and Preparedness: Recent Research and Trends". In: *Information Systems Frontiers* 20.5, pp. 901–907.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (13–18 Jul 2020). "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4411–4421.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (June 2015). "Processing Social Media Messages in Mass Emergency: A Survey". In: *ACM Comput. Surv.* 47.4.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). "AIDR: Artificial Intelligence for Disaster Response". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14 Companion. Seoul, Korea: Association for Computing Machinery, pp. 159–162.

Kayi, E. S., Nan, L., Qu, B., Diab, M., and Mckeown, K. (2020). "Detecting Urgency Status of Crisis Tweets: A Transfer Learning Approach for Low Resource Languages". In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4693–4703.

Khanuja, S., Dandapat, S., Srinivasan, A., Sitaram, S., and Choudhury, M. (July 2020). "GLUECoS: An Evaluation Benchmark for Code-Switched NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 3575–3585.

Krishnan, J., Anastasopoulos, A., Purohit, H., and Rangwala, H. (Nov. 2021). "Multilingual Code-Switching for Zero-Shot Cross-Lingual Intent Prediction and Slot Filling". In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 211–223.

Krishnan, J., Purohit, H., and Rangwala, H. (2020a). "Attention Realignment and Pseudo-Labelling for Interpretable Cross-Lingual Classification of Crisis Tweets". In: *Proceedings of the Workshop on Knowledge-infused Mining and Learning (KDD-KiML 2020)*.

Krishnan, J., Purohit, H., and Rangwala, H. (2020b). "Unsupervised and Interpretable Domain Adaptation to Rapidly Filter Tweets for Emergency Services". In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 409–416.

Kruspe, A. (2019). "Few-shot tweet detection in emerging disaster events". In: *arXiv preprint arXiv:1910.02290*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *ArXiv* abs/1907.11692.

Lorini, V., Rufolo, P., and Castillo, C. (2022). "Venice Was Flooding ... One Tweet at a Time". In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2, p. 382.

Lorini, V., Salamon, P., and Castillo, C. (Apr. 2021). "SMDRM - Social Media for Disaster Risk Management". In: *EGU General Assembly Conference Abstracts*. EGU General Assembly Conference Abstracts, EGU21-15012, EGU21–15012.

Lu, S.-E., Lu, B.-H., Lu, C.-Y., and Tsai, R. T.-H. (Dec. 2022). "Exploring Methods for Building Dialects-Mandarin Code-Mixing Corpora: A Case Study in Taiwanese Hokkien". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 6287–6305.

Pandey, R. and Purohit, H. (2018). "CitizenHelper-Adaptive: Expert-Augmented Streaming Analytics System for Emergency Services and Humanitarian Organizations". In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 630–633.

Pires, T., Schlinger, E., and Garrette, D. (July 2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001.

Purohit, H., Castillo, C., Diaz, F., Sheth, A., and Meier, P. (Dec. 2013). "Emergency-relief coordination on social media: Automatically matching resource requests and offers". In: *First Monday*.

Purohit, H. and Peterson, S. (2020). "Social media mining for disaster management and community resilience". In: *Big Data in Emergency Management: Exploitation Techniques for Social and Mobile Data*. Springer, pp. 93–107.

Rani, P., McCrae, J. P., and Fransen, T. (June 2022). "MHE: Code-Mixed Corpora for Similar Language Identification". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3425–3433.

Rathnayake, H., Sumanapala, J., Rukshani, R., and Ranathunga, S. (July 2022). "Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification". In: *Knowledge and Information Systems* 64.7, pp. 1937–1966.

Reuter, C., Hughes, A. L., and Kaufhold, M.-A. (2018). "Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research". In: *International Journal of Human–Computer Interaction* 34.4, pp. 280–294. eprint: https://doi.org/10.1080/10447318.2018.1427832.

Rufolo, P., Muraro, D., and Lorini, V. (2021). In: KJ-NA-30632-EN-N (online).

Santy, S., Srinivasan, A., and Choudhury, M. (Apr. 2021). "BERTologiCoMix: How does Code-Mixing interact with Multilingual BERT?" In: *Proceedings of the Second Workshop on Domain Adaptation for NLP*. Kyiv, Ukraine: Association for Computational Linguistics, pp. 111–121.

Singh, V., Vijay, D., Akhtar, S. S., and Shrivastava, M. (July 2018). "Named Entity Recognition for Hindi-English Code-Mixed Social Media Text". In: *Proceedings of the Seventh Named Entities Workshop*. Melbourne, Australia: Association for Computational Linguistics, pp. 27–35.

Sitaram, S., Chandu, K. R., Rallabandi, S. K., and Black, A. W. (2019). *A Survey of Code-switched Speech and Language Processing*.

Spiliopoulou, E., Maza, S. M., Hovy, E., and Hauptmann, A. (2020). "Event-related bias removal for real-time disaster events". In: *arXiv preprint arXiv:2011.00681*.

Srivastava, V. and Singh, M. (June 2021). "Challenges and Limitations with the Metrics Measuring the Complexity of Code-Mixed Text". In: *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. Online: Association for Computational Linguistics, pp. 6–14.

Ullah, I., Khan, S., Imran, M., and Lee, Y.-K. (2021). "RweetMiner: Automatic identification and categorization of help requests on twitter during disasters". In: *Expert Systems with Applications* 176, p. 114787.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Vieweg, S. E. (2012). "Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications". PhD thesis. University of Colorado at Boulder.

Vyas, Y., Gella, S., Sharma, J., Bali, K., and Choudhury, M. (Oct. 2014). "POS Tagging of English-Hindi Code-Mixed Social Media Content". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 974–979.

Winata, G. I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., and Fung, P. (June 2021). "Are Multilingual Models Effective in Code-Switching?" In: *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. Online: Association for Computational Linguistics, pp. 142–153.

Winford, D. (2003). "Code switching: Linguistic aspects". In: *An introduction to contact linguistics* 126, p. 167.

Zhou, B., Zou, L., Mostafavi, A., Lin, B., Yang, M., Gharaibeh, N., Cai, H., Abedin, J., and Mandal, D. (2022). "VictimFinder: Harvesting rescue requests in disaster response from social media with BERT". In: *Computers, Environment and Urban Systems* 95, p. 101824.