

On Informative Tweet Identification for Tracking Mass Events

Renato Stoffalette João ^a

L3S Research Center, Leibniz University of Hannover, Appelstraße 9A, Hannover, 30167, Germany

Keywords: Machine Learning, Classification, Deep Learning.

Abstract: Twitter has been heavily used as an important channel for communicating and discussing about events in real-time. In such major events, many uninformative tweets are also published rapidly by many users, making it hard to follow the events. In this paper, we address this problem by investigating machine learning methods for automatically identifying informative tweets among those that are relevant to a target event. We examine both traditional approaches with a rich set of handcrafted features and state of the art approaches with automatically learned features. We further propose a hybrid model that leverages both the handcrafted features and the automatically learned ones. Our experiments on several large datasets of real-world events show that the latter approaches significantly outperform the former and our proposed model performs the best, suggesting highly effective mechanisms for tracking mass events.

1 INTRODUCTION

Lately Twitter has become an important channel for communication and information broadcasting. A large number of its users have been using the platform for seeking and sharing the information about events. Particularly, during undesired mass events like natural disasters or terrorist attacks, Twitter users post tweets, share updates, inform other users about current situations, etc. However, in addition to these information, a lot of tweets are merely for discussing and expressing opinions and emotions towards the events, which makes it challenging for professionals involved in crisis management to actually collect relevant information for better understanding the situations and respond more rapidly (Vieweg et al., 2010).


Considering the large volume of tweets published by Twitter users, manual sifting to find useful information is inherently impractical (Meier, 2013). Thus automatic mechanisms for identification of the informative tweets are required to assist not only the average citizen to become aware of the situation but also the professionals to take measures immediately and potentially save lives.

In this work, we investigate the viability of machine learning approaches for developing such an automatic mechanism. We study both traditional ones that use handcrafted features, as well as the state of the art representation learning approach, the BERT-

based models (Devlin et al., 2019), to classify tweets according to their informativeness. We implement a rich set of features for the former, examine different usage of the latter as well the combinations of both. Furthermore, we propose a hybrid model that leverages both the BERT-based models and the handcrafted features. We evaluate all these models on large datasets collected during several natural and man-caused disasters. In summary, we make the following contributions.

- We investigate a rich set of features that include Bag-of-Words, text-based, and user-based features for traditional models, and examine the performance of BERT-based models for the informative tweet classification problem.
- We further propose a hybrid model that combines a BERT-based model with handcrafted features for the problem.
- We conduct comprehensive experiments for evaluating the performance of these diverse models.
- Empirically, we demonstrate that deep BERT-based models outperform the traditional ones for the task without requiring complicated feature engineering, while our proposed model performs the best.

The remaining of this paper is organized as follows. We firstly review the related works in Section 2, then we describe the methods and the features in Section 3. Section 4 describes our experiments, datasets and

^a  <https://orcid.org/0000-0003-4929-4524>

give details about our implementation methods. In Section 4.4 we report the results obtained from our experiments. Finally, we draw some conclusions and point out some future directions in Section 5.

2 RELATED WORK

Social media platforms such as Twitter and Facebook have become valuable communication channels over the years. Twitter enables people to share all kinds of information by posting short text messages, called tweets. Although social media services are full of conversational messages, it is also an environment where users post newsworthy information related to some natural or human-induced disaster. Identifying such information can help not only the ordinary citizen but it can also assist professionals and organizations in coordinating their response for potentially saving lives and diminishing catastrophic losses (Imran et al., 2015).

A number of automated systems have been proposed to extract and classify crisis related information from social media channels, for example Crisis-Tracker (Rogstadius et al., 2013), Twitcident (Abel et al., 2012), AIDR (Imran et al., 2014), among others. For a more complete list of systems, please refer to the survey by Imran et al. (Imran et al., 2015).

Machine learning and natural language processing play an important role when it comes to classifying crisis related tweets automatically, and the approach applied to extract textual features can determine the performance of an automated classifier. Castillo et al. (Castillo et al., 2011) proposed automatic techniques to assess the credibility of tweets related to specific topics or events, using features extracted from user's posting behavior and tweet's text. Verma, et al. (Verma et al., 2011) used Naive Bayes and MaxEnt classifiers to find situational awareness tweets from several crises and Cameron et al. (Cameron et al., 2012) described a platform for emergency situation awareness where they classified interesting tweets using an SVM classifier.

With the recent advances in natural language processing and the emergence of techniques such as word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) and GloVe (Pennington et al., 2014), deep neural networks have successfully been applied in similar tasks. Caragea et al. (Caragea et al., 2016) for instance, demonstrated that convolutional neural networks outperformed traditional classifiers in tweet classification. Nguyen et al. (Nguyen et al., 2017) also used a convolutional neural network based model to classify crisis-relevant tweets. These results sug-

gest a promising approach for this informative tweet classification task.

3 METHODOLOGY

Identifying informative tweets is a critical task, particularly during catastrophic events. There is however no simple rules that can be applied for the task. We therefore approach the problem of informative tweets identification as a supervised learning problem. In the following subsections, we shall discuss several models for the task. We start with some conventional classification models that make use of features engineered from the tweets as well as the users who posted the tweets. Next, we present the deep learning approaches for the task, and describe our proposed model.

3.1 Traditional Models

Several machine learning approaches have been proposed for the task of automatically detecting crisis-related tweets, for example, Naive Bayes (Li et al., 2018), Support Vector Machines (Caragea et al., 2016), and Random Forests (Kaufhold et al., 2020). Thus, as the baselines, we have trained these traditional classifiers to automatically classify a tweet into either *Informative* or *Not Informative*. Specifically, we have implemented the following models.

- **LOGISTIC REGRESSION (LR)**: a classifier that models the probability of a label based on a set of independent features,
- **DECISION TREE (DT)**: a classifier that successively divides the features space to maximise a given metric (e.g., information gain),
- **RANDOM FOREST (RF)**: a classifier that utilises an ensemble of uncorrelated decision trees,
- **NAIVE BAYES (NB)**: a Gaussian Naive Bayes classifier,
- **MULTILAYER PERCEPTRON (MP)**: a network of linear classifiers, (*perceptrons*) that uses the backpropagation technique to classify the instances, and
- **SUPPORT VECTOR MACHINE (SVM)**: a discriminative classifier formally defined by a separating hyperplane.

All the classifiers deployed in this work were implemented in Python using the machine learning library Scikit-Learn (Pedregosa et al., 2011). The source code of our models implementations is freely available at <https://github.com/renatosjoao/infotweets.git>.

3.1.1 Features

Inspired by previous works, we investigated a set of features based on the tweets' contents as well as on the users who posted the tweets (Acerbo and Rossi, 2017; Graf et al., 2018; Imran et al., 2013; Verma et al., 2011). These features are described as follows.

- Text-based features: the ones that are calculated from the content of a tweet, including
 - n_{chars} : This feature refers to the number of characters a tweet contains.
 - n_{words} : The number of words a tweet contains. After removing symbols and patterns we count the number of words that is present in the tweet.
 - $n_{hashtags}$: The number of occurrences of #hashtags in a tweet. It can indicate the user wants to highlight some specific subject of interest.
 - n_{url} : The number of URLs contained in a tweet.
 - n_{at} : The number of @ tags in the tweet can be an indicator that the user is tagging people to draw their attention.
 - $b_{hashtag}$: Binary valued feature referring to the presence of #hashtags. True if at least one #hashtag is present in the tweet, false otherwise.
 - b_{at} : Binary valued feature referring to the presence of @ tags. True if the tweet contains @ tags, false otherwise.
 - b_{rt} : Binary valued feature referring to a retweeted message. True if the tweet contains retweet patterns, such as rt@, false otherwise.
 - b_{slang} : Binary valued feature referring to slangs in the tweet. True if the tweet contains any slang, false otherwise. Internet abbreviations are examples of text informality, which are representative of conversations. We built a dictionary of slangs from an online slang dictionary¹.
 - b_{url} : Binary valued feature about the presence of URLs. True if at least one URL is present in the tweet, false otherwise.
 - t_{lex} : Tweet lexical diversity refers to the number of unique words divided by the total number of words in the tweet.
 - b_{interj} : Binary valued feature referring to an interjection. True if the tweet contains interjections, false otherwise. We built a dictionary of interjections from an online list of interjections².

¹<https://www.lifewire.com/urban-internet-slang-dictionary-3486341>

²<https://www.vidarholen.net/contents/interjections/>

- bow : Bag-of-Words features. Real-valued vectors are calculated with TF×IDF of the words and Twitter posts from each corpus for a finite number of words from the vocabulary.

- User-based features: the ones that are calculated from the user who posted the tweet, including
 - b_{usr} : Binary valued feature representing whether the user account is verified. True if the user has a verified account, false otherwise.
 - $n_{followers}$: This feature represents the number of followers the user who posted the tweet has. Since this number may vary considerably we calculated it as $\log_{10}(n_{followers} + 1)$.
 - $n_{followees}$: Number of accounts the user who posted the tweet follows, calculated as $\log_{10}(n_{followees} + 1)$.
 - n_{tweets} : This feature represents the total number of tweets posted by the user. There can be the case where the user has not posted many tweets as well as there can be cases of influential users who post messages more frequently, thus we calculate this feature as $\log_{10}(n_{tweets} + 1)$.

3.2 Deep Learning Approaches

We now discuss deep learning based approaches that are widely used in recent works (Nguyen et al., 2017; Neppalli et al., 2018).

3.2.1 Word Embedding Methods

The traditional models such as the *Bag-of-Words* do not capture well the meaning of the words and consider each word as a separate feature. Word embeddings have been proposed and widely used neural models that map words into real number vectors such that similar words are closer to each other in a higher dimensional space. The word embeddings captures the semantical and syntactical information of words taking into consideration the surrounding context.

In this work, we examine the following typical word embedding methods:

- **Word2vec**: (Mikolov et al., 2013b) is one famous method of neural words embeddings initially proposed in two variants: (i) a Bag-of-Words model that predicts the current word based on the context words, and (ii) a skip-gram model that predicts surrounding words given the current word.
- **GloVe**: is an extension to the Word2vec method for efficiently learning word vectors, proposed by (Pennington et al., 2014) which uses global corpus statistics for words representations and learns

the embeddings by dimensionality reduction of the co-occurrence count matrix.

- **Fasttext:** (Bojanowski et al., 2016) is an extension to the skip-gram model from the original Word2vec model which takes into account sub-word information, i.e. it learns representations for character n-grams, and represents words as the sum of the n-gram vectors. The idea is to capture morphological characteristics of words.

We make use of the pre-trained word vectors of the above models^{3,4,5}. The feature vector of each tweet is then determined by taking the average of all embedding vectors of its words.

3.2.2 Text Embedding Methods

Generalized from word embeddings, text embedding methods compute a vector for each group of words taken collectively as a single unit, e.g., a sentence, a paragraph, or the whole document. In this work, we examine a typical method for text embedding, namely Doc2vec, and state-of-the-art ones, namely BERT-based models.

Doc2vec: generates efficient and high quality distributed vectors of a complete document (Mikolov et al., 2013b). The main objective of Doc2Vec is to convert the sentence (or paragraph) into a vector. It is a generalization on the Word2vec model.

BERT: is a model developed on a multi-layer bidirectional Transformer encoder (Vaswani et al., 2017; Devlin et al., 2019). It makes use of an attention mechanism that learns contextual relations between words in texts. In its generic format, the Transformer includes two separate mechanisms, an encoder that reads input text and a decoder that produces the task prediction. The encoder is composed of a stack of multiple layers, and each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. The decoder is also composed of a stack of multiple identical layers with the addition of a third sub-layer, which performs multi-head attention over the output of the encoder stack. One key component of the Transformer encoder is the multi head self-attention layer, i.e. a function that can be formulated as querying a dictionary with key-value pairs.

The most straightforward usage of BERT is to employ it as a blackbox for feature engineering. This is the combination of the default BERT model and conventional classifiers. The final hidden state of the first

word ([CLS]) from BERT is the encoded sentence representation and it is input to conventional classifiers for the predictions task.

The original BERT model is pre-trained in a general domain corpus. Thus, for a text classification task in a specific domain, the data distribution may be different. In this way in order to obtain improved results, we need to further train BERT on a domain specific data. There are a couple of ways to further train BERT on a domain specific corpus. The first one is to train the entire pre-trained model on the new corpus and feed the output into a softmax function. In this way, the error is back propagated throughout the entire model's architecture and the weights are updated for this domain specific corpus. Another method is to train some of BERT's layers while freezing others, or we can freeze all the layers and attach extra neural network layers and train this new model where only the weights of the attached layers will be updated. These are so called fine tuning procedures, and in this work we will be fine tuning BERT, by encoding Twitter sentences with the BERT encoder and running more training iterations and backpropagating the error throughout the entire model.

3.3 Our Proposed Model

We now describe a hybrid model, called BERT_{Hyb}, that combines both the handcrafted features with the ones learned by BERT. BERT_{Hyb} model feeds a vector of handcrafted features from the tweet through a linear layer, and also feeds the vector produced by BERT for the first token (CLS) of the tweet through another linear layer. The outputs of these two layers are concatenated and fed through a third linear layer, whose output is subsequently fed through a softmax layer to produce the prediction whether a tweet is *Informative* or *Not Informative*.

4 EXPERIMENTS

We now present our experiments to empirically evaluate the methods presented above. In the following subsections, we shall describe the datasets, define the evaluation metrics, the experiment settings, and report the results.

4.1 Datasets

We use the following datasets to evaluate the models.

- **CRISISLEXT26.** (Olteanu et al., 2015) This is a dataset of tweets collected during twenty six large crisis events in 2012 and 2013, with about 1,000

³Word2vec: <https://code.google.com/archive/p/word2vec/>

⁴GloVe: <https://nlp.stanford.edu/projects/glove/>

⁵Fasttext: <https://fasttext.cc/>

tweets labeled per crisis for informativeness, information type, and source.

- **CRISISLEXT6.** (Olteanu et al., 2014) This dataset includes English tweets posted during six large events in 2012 and 2013, with about 60.000 tweets labeled by relatedness as *On-topic* or *Off-topic* with each event. We assume the tweets labeled as *On-topic* being the *Informative* tweets and *Off-topic* being *Not Informative* respectively.
- **CRISISMMD.** (Alam et al., 2018) CrisisMMD is a dataset that contains tweets with both text and image contents. There are 16,000 tweets that were collected from seven events that took place in 2017 in five countries.
- **COVID.** (Nguyen et al., 2020) This dataset consists of 10K English Tweets collected during the Covid pandemic. It is split into training set with 3303 *Informative* tweets and 3697 *Uninformative* tweets, and a validation set with 472 and 528 *Informative* and *Uninformative* tweets respectively.

In their original form, the above datasets provide only tweets' content together with their ids and labels. To calculate the user based features we crawl from Twitter the full information of all the tweets. However, some tweets are no longer available. We thus create a version of each dataset that consists of the subset of tweets that we can crawl full information from Twitter. These versions are COVID and COVID_{SUBSET}, CRISISLEXT6 and CRISISLEXT6_{SUBSET}, CRISISLEXT26 and CRISISLEXT26_{SUBSET}, CRISISMMD and CRISISMMD_{SUBSET} respectively. The basic statistics of all the datasets and their subsets are shown in Tables 1 and 2 respectively.

4.2 Evaluation Metrics

To evaluate the informative tweets classification task we employ the following performance metrics. **Precision (P)**: the fraction of the correctly classified instances among the instances assigned to the class. **Recall (R)**: the fraction of the correctly classified instances among all instances of the class and **F-score (F1)**: the harmonic mean of precision and recall. In this work we compute the metrics independently for each class and then take the average, i.e. Macro Precision, Macro Recall and Macro F-score.

4.3 Experiment Settings

We normalized all characters in the tweets to their lower-cased forms followed by the removal of punctuation and non ASCII characters as well as non English

words, then we calculated the text-based features and user-based features. The Bag-of-Words feature was calculated for the entire corpus of tweets, however in our experiments we only calculated it for words appearing at least 5 times in the entire corpus and up to a limit of 10000 times. The words with length less than two characters were also pruned.

In parallel we then tokenized the sentences and encoded the tokens using the BERT encoder. Each dataset is randomly split into 10 mutually exclusive subsets and 10-fold cross validation was used to measure the performance of the models. For the conventional classifiers we used the implementation from the scikit-learn tool (Pedregosa et al., 2011) and all the algorithms were set to use the default parameter values. As regards BERT fine tuning, we used the stochastic gradient descent optimizer with a learning rate of 0.001, momentum 0.9 and ran the training process for 20 epochs. We set the batch size to 16 and limited the BERT sentence encoding to the maximum length of 80. In this work the BERT models were built based on the pytorch-pretrained-BERT repository <https://github.com/huggingface/pytorch-pretrained-BERT>.

4.4 Results

We show the results in terms of macro average F-score. Table 3 shows the performance of the implemented models on all the datasets used in this work. The two best results obtained in each dataset is highlighted in bold face. Only the COVID and CRISISMMD datasets were split into training and validation sets by default, however to make it fair and comparable across all the datasets and approaches we performed 10-fold cross validation with the entire datasets (combined training and validation sets).

In the first six rows we show the classification performance of conventional classifiers using the hand-crafted features proposed in this work. For the full datasets it is only possible to calculate the Twitter-based features, as the user-based features are strongly dependent on the complete tweet information, and since we had to crawl the Twitter platform to obtain the complete information, we realised that many tweets had been deleted.

We noticed the performance of the classifiers varies on a per dataset basis and classifiers performed differently on each of the datasets. For the COVID dataset we observed the LOGISTIC REGRESSION classifier performed the best with Macro F1 of 57.07, while for CRISISLEXT6 datasets and CRISISLEXT26 MLP showed the best score 75.56 and 68.10, respectively. And for CRISISMMD, RANDOM FOREST outperformed the other classifiers with

Table 1: Complete datasets classes distributions.

<i>Dataset</i>	<i>#Informative</i>	<i>#Not Informative</i>	<i>Total</i>
COVID	3,772	4,221	7,993
CRISISLEX6	32,461	27,620	60,081
CRISISLEX26	16,849	7,731	24,580
CRISISMMD	11,509	4,549	16,058

Table 2: Subsets classes distribution.

<i>Dataset</i>	<i>#Informative</i>	<i>#Not Informative</i>	<i>Total</i>
COVID _{SUBSET}	3,378	3,816	7,194
CRISISLEX6 _{SUBSET}	20,568	17,422	37,990
CRISISLEX26 _{SUBSET}	11,023	4,442	15,465
CRISISMMD _{SUBSET}	9,343	3,443	12,786

a score of 55.85.

The following six rows show the classification performance using Bag-of-Words as input features. Here again we noticed the performance of the classifiers varies on a per dataset basis, however we observed considerable performance improvement across all datasets which demonstrates that the bag-of-words is a stronger features encoding method than the handcrafted features approach only.

In the following six rows we show the results of the classification task using a combination of the handcrafted features with the Bag-of-Words features. It is interesting to observe that for the majority of the classifiers this combination does not produce improved results over the COVID and the CRISISLEX6 datasets. Only NAIVE BAYES demonstrated considerable improvement over the previous approach for the COVID dataset. However, all the classifiers demonstrated improvement in the CRISISLEX26 dataset when compared to using the Bag-of-Words only approach, and for the CRISISMMD dataset again only NAIVE BAYES demonstrated improvement when compared to the previous approach.

The next six rows show the results of the conventional classifiers using Fasttext word embeddings. For the COVID and CRISISLEX6 datasets, MLP produced the best results, while for the CRISISLEX26 and for the CRISISMMD datasets, LOGISTIC REGRESSION demonstrated the best macro F-score. In the following six rows we can see the classification results using GloVe word embeddings. The performance results observed from the classifiers using this embedding technique seem to be similar to the Fasttext word embeddings varying not too much across datasets.

In the following six rows we show the performance results of one approach in which we use the conventional classifiers using BERT encoded features

combined with the handcrafted features. We have not noticed improvements using this approach of combining BERT word embeddings with handcrafted features on the COVID and CRISISMMD datasets, however we observed some improvements in the CRISISLEX6 and CRISISLEX26 dataset for the majority of the classifiers.

Finally in the last row we show the results of our proposed approach BERT_{Hyb}. Our model outperforms all the previously cited methods across all datasets used in this work. For COVID dataset it produced a macro F-score of 84.41 which is 2.5 percentage points improvement over the best result from previous approaches (LR using Bag-of-Words features). For CRISISLEX6 we observed 95.96 macro F-score, for CRISISLEX26 we obtained 79.09 macro F-score, which is the highest improvement (7 percentage points over SVM using handcrafted features combined with Bag-of-Words) and for CRISISMMD our model produced 77.66 macro F-score.

There are some reasons that can explain why our hybrid model performs much better than other models tested in this paper. The first one is the fact that BERT encoder uses a contextual representation in which it processes words in relation to all the other words in the sequence, rather than one by one separately, and the second reason is the fact that we ran several training iterations while adjusting weights, and using different optimization functions to minimise the training loss.

We also evaluated the proposed approach in the subsets of the original datasets. As mentioned before these subsets were created so we could also calculate features related to the user who posted the message. We noticed again that the handcrafted features alone did not produce satisfactory results. The best observed macro F-scores varied between 55.49 for the CRISISMMD_{SUBSET} using a NAIVE BAYES

Table 3: Models performance on the original datasets.

FEATURES	MODELS	COVID	CRISISLEXT6	CRISISLEXT26	CRISISMMD
		MacroF1	MacroF1	MacroF1	MacroF1
HANDCRAFTED	LR	57.07(+/- 0.02)	75.09(+/- 0.14)	64.60(+/- 0.05)	48.91(+/- 0.02)
	DT	51.99(+/- 0.02)	72.39(+/- 0.14)	61.82(+/- 0.05)	55.63(+/- 0.02)
	RF	54.14(+/- 0.02)	74.05(+/- 0.14)	64.14(+/- 0.05)	55.85(+/- 0.02)
	NB	42.79(+/- 0.02)	72.51(+/- 0.14)	65.79(+/- 0.05)	50.60(+/- 0.02)
	MLP	49.84(+/- 0.02)	75.56(+/- 0.14)	68.10(+/- 0.05)	48.42(+/- 0.02)
	SVM	56.11(+/- 0.02)	75.41(+/- 0.14)	65.53(+/- 0.05)	49.81(+/- 0.02)
BAG-OF-WORDS	LR	81.90 (+/- 0.04)	92.90(+/- 0.09)	66.46(+/- 0.17)	72.68(+/- 0.03)
	DT	75.13(+/- 0.04)	91.42(+/- 0.09)	53.04(+/- 0.17)	68.97(+/- 0.03)
	RF	81.06(+/- 0.04)	93.51 (+/- 0.09)	62.59(+/- 0.17)	73.21(+/- 0.03)
	NB	66.75(+/- 0.04)	80.35(+/- 0.09)	57.09(+/- 0.17)	47.56(+/- 0.03)
	MLP	75.23(+/- 0.04)	91.74(+/- 0.09)	63.39(+/- 0.17)	71.48(+/- 0.03)
	SVM	81.38(+/- 0.04)	93.21(+/- 0.09)	65.01(+/- 0.17)	66.00(+/- 0.03)
HANDCRAFTED + BoW	LR	78.29(+/- 0.05)	83.58(+/- 0.12)	69.70(+/- 0.12)	65.12(+/- 0.03)
	DT	74.68(+/- 0.05)	90.80(+/- 0.12)	61.26(+/- 0.12)	66.24(+/- 0.03)
	RF	80.47(+/- 0.05)	93.28(+/- 0.12)	66.55(+/- 0.12)	70.61(+/- 0.03)
	NB	71.56(+/- 0.05)	79.00(+/- 0.12)	60.83(+/- 0.12)	57.53(+/- 0.03)
	MLP	75.28(+/- 0.05)	91.51(+/- 0.12)	63.96(+/- 0.12)	69.58(+/- 0.03)
	SVM	75.05(+/- 0.05)	92.96(+/- 0.12)	72.09 (+/- 0.12)	66.18(+/- 0.03)
FASTTEXT	LR	77.60(+/- 0.04)	89.29(+/- 0.08)	71.30(+/- 0.10)	74.09(+/- 0.02)
	DT	64.42(+/- 0.04)	79.26(+/- 0.08)	60.85(+/- 0.10)	63.74(+/- 0.02)
	RF	76.16(+/- 0.04)	88.62(+/- 0.08)	69.83(+/- 0.10)	71.54(+/- 0.02)
	NB	74.73(+/- 0.04)	77.18(+/- 0.08)	63.41(+/- 0.10)	66.89(+/- 0.02)
	MLP	80.01(+/- 0.04)	91.28(+/- 0.08)	67.81(+/- 0.10)	74.00(+/- 0.02)
	SVM	76.43(+/- 0.04)	89.46(+/- 0.08)	70.14(+/- 0.10)	71.40(+/- 0.02)
GLOVE	LR	79.68(+/- 0.04)	86.82(+/- 0.11)	70.40(+/- 0.09)	74.59(+/- 0.02)
	DT	66.76(+/- 0.04)	77.49(+/- 0.11)	60.05(+/- 0.09)	63.54(+/- 0.02)
	RF	77.80(+/- 0.04)	87.36(+/- 0.11)	66.27(+/- 0.09)	72.41(+/- 0.02)
	NB	76.29(+/- 0.04)	81.72(+/- 0.11)	61.30(+/- 0.09)	72.87(+/- 0.02)
	MLP	79.03(+/- 0.04)	87.96(+/- 0.11)	66.21(+/- 0.09)	72.38(+/- 0.02)
	SVM	80.05(+/- 0.04)	89.20(+/- 0.11)	71.75(+/- 0.09)	75.15 (+/- 0.02)
BERT	LR	77.83(+/- 0.03)	90.62(+/- 0.09)	70.41(+/- 0.10)	74.80(+/- 0.03)
	DT	62.19(+/- 0.03)	77.84(+/- 0.09)	60.98(+/- 0.10)	62.84(+/- 0.03)
	RF	74.11(+/- 0.03)	87.51(+/- 0.09)	69.11(+/- 0.10)	70.76(+/- 0.03)
	NB	71.34(+/- 0.03)	77.69(+/- 0.09)	67.59(+/- 0.10)	70.41(+/- 0.03)
	MLP	77.08(+/- 0.03)	89.75(+/- 0.09)	66.54(+/- 0.10)	72.21(+/- 0.03)
	SVM	78.08(+/- 0.03)	91.50(+/- 0.09)	70.53(+/- 0.10)	75.14(+/- 0.03)
HANDCRAFTED + BERT	BERT _{Hyb}	84.41 (+/- 0.01)	95.96 (+/- 0.03)	79.09 (+/- 0.04)	77.66 (+/- 0.01)

classifier and 78.58 for the CRISISLEXT6_{SUBSET} using RANDOM FOREST classifier. However when we used the Bag-of-Words model as input features, the classifiers produced considerably better results for COVID_{SUBSET} and CRISISLEXT6_{SUBSET} datasets in all cases, but for the CRISISLEXT26_{SUBSET} and CRISISMMD_{SUBSET} there were some classifiers that performed better using only the handcrafted features, for example for the CRISISLEXT26_{SUBSET} the RANDOM FOREST model produced a macro F-score of 66.80, while using the Bag-of-Words model it produced only 52.06. The combination of the handcrafted features and Bag-of-Words shows improvement for all datasets only when using the NAIVE BAYES classifier when compared to the Bag-of-Words model, while when compared to the sole handcrafted features the classifiers produce better results in all cases for the COVID_{SUBSET} and CRISISLEXT6_{SUBSET} datasets and the majority of cases in CRISISLEXT26_{SUBSET} and CRISISMMD_{SUBSET} with the exception of the NAIVE BAYES classifier.

Using the Fasttext, GloVe and BERT embeddings as input features to the conventional classifiers showed considerable improvements across all datasets, especially when using LOGISTIC REGRESSION as base classifier, however this was not a pattern observed when using different classification methods.

Our hybrid model BERT_{Hyb} produced the best performance result for almost all the dataset with the exception of the CRISISLEXT6_{SUBSET}, however the

difference is marginal. The best observed macro F-score is shown when using the Bag-of-Words features model using RANDOM FOREST as base classifier (93.22), while our hybrid approach produced a score of 93.05. In the COVID_{SUBSET} our model showed 84.64 macro F-score which is 2.3 percentage points improvement over the second best result (Bag-of-Words and LR = 82.35). Our model showed 76.68 and 76.54 macro F-score for the CRISISLEXT26_{SUBSET} and CRISISMMD_{SUBSET} datasets respectively. These two datasets seem to be the two datasets where the performance of the models were lower than 80%. Further investigation and a more in depth analysis is required as there is still some room for improvements.

5 CONCLUSIONS

Social media has drawn attention from different sectors of society and the information available during catastrophic events is extremely useful for both the ordinary citizen and the professionals involved in humanitarian purposes, however there is an overload of information that requires an automated filtering method for real time processing of relevant content.

In this work we designed a set of handcrafted features from both the Twitter posts and the users who posted a tweet, and showed experimentally the performance of six conventional classifiers on the infor-

mative tweet classification task. We also trained classifiers with several word embeddings, namely, FastText, GloVe and BERT, as input features. Moreover, we showed that our proposed deep neural model BERT_{Hyb} is more effective in identifying informative tweets as compared to conventional classifiers in different crisis related corpus from Twitter.

As future works we intend to further investigate different deep learning models combinations and implement a complete pipeline where the tweets are crawled and classified in real time based on crisis related trending topics.

REFERENCES

- Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., and Tao, K. (2012). Semantics+ filtering+ search= twitcident. exploring information in social web streams. In *HT'12*.
- Acerbo, F. S. and Rossi, C. (2017). Filtering informative tweets during emergencies: a machine learning approach. In *Proceedings of the First CoNEXT Workshop on ICT Tools for Emergency Networks and Disaster Relief*.
- Alam, F., Ofli, F., and Imran, M. (2018). Crisismmd: Multimodal twitter datasets from natural disasters. In *ICWSM*.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- Cameron, M. A., Power, R., Robinson, B., and Yin, J. (2012). Emergency situation awareness from twitter for crisis management. In *WWW*.
- Caragea, C., Silvescu, A., and Tapia, A. H. (2016). Identifying informative messages in disaster events using convolutional neural networks. In *ISCRAM*.
- Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. In *WWW*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.
- Graf, D., Retschitzegger, W., Schwinger, W., Pröll, B., and Kapsammer, E. (2018). Cross-domain informativeness classification for disaster situations. In *MEDES*.
- Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *CSUR*.
- Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). Aidr: Artificial intelligence for disaster response. In *WWW*.
- Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). Extracting information nuggets from disaster-related messages in social media. In *Iscram*.
- Kaufhold, M.-A., Bayer, M., and Reuter, C. (2020). Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *IPM*, 57(1).
- Li, H., Caragea, D., Caragea, C., and Herndon, N. (2018). Disaster response aided by tweet classification with a domain adaptation approach. *JCCM*.
- Meier, P. (2013). Crisis maps: Harnessing the power of big data to deliver humanitarian assistance. *Forbes Magazine*.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *NIPS*.
- Neppalli, V. K., Caragea, C., and Caragea, D. (2018). Deep neural networks versus naive bayes classifiers for identifying informative tweets during disasters. In *ISCRAM*.
- Nguyen, D. Q., Vu, T., Rahimi, A., Dao, M. H., Nguyen, L. T., and Doan, L. (2020). WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text*.
- Nguyen, D. T., Al-Mannai, K., Joty, S. R., Sajjad, H., Imran, M., and Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. *ICWSM*.
- Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*.
- Olteanu, A., Vieweg, S., and Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. In *CSCW*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *JMLR*.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*.
- Rogstadius, J., Vukovic, M., Teixeira, C. A., Kostakos, V., Karapanos, E., and Laredo, J. A. (2013). Crisis-tracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*.
- Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., Schram, A., and Anderson, K. M. (2011). Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In *ICWSM*.
- Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *SIGCHI*.