1-2023

# Is a pretrained model the answer to situational awareness detection on social media?

Siaw Ling LO
*Singapore Management University*, sllo@smu.edu.sg

Kahhe LEE
*Singapore Management University*, kahhe.lee.2019@scis.smu.edu.sg

Yuhao ZHANG
*Singapore Management University*, yuhaozhang@smu.edu.sg

# Is a Pretrained Model the Answer to Situational Awareness Detection on Social Media?


## Abstract

*Social media can be valuable for extracting information about an event or incident on the ground. However, the vast amount of content shared, and the linguistic variants of languages used on social media make it challenging to identify important situational awareness content to aid in decision-making for first responders. In this study, we assess whether pretrained models can be used to address the aforementioned challenges on social media. Various pretrained models, including static word embedding (such as Word2Vec and GloVe) and contextualized word embedding (such as DistilBERT) are studied in detail. According to our findings, a vanilla DistilBERT pretrained language model is insufficient to identify situation awareness information. Fine-tuning by using datasets of various event types and vocabulary extension is essential to adapt a DistilBERT model for real-world situational awareness detection.*



**Keywords:** Pretrained models, situational awareness, BERT, fine tuning, vocabulary extension


## 1. Introduction

With its open and real-time broadcasting nature, social media is often the platform to go to when an incident or a crisis occurs. Social media can contain valuable information from an eyewitness account or be a way to seek assistance. In this study, we focus on identifying social media content that can provide useful information on urban events such as event type (e.g., civil disorder, armed assault etc.), number of injury/casualties, infrastructure damages or weapon uses, for first responders (e.g., police force or paramedics) to manage situations on the ground. We refer to such useful information as situational awareness information. An example of a tweet containing situational awareness information is *"@username After the bus crashed n knocked down the Indian national, they rage to the extent of smashing up the bus, burning ambulance and police cars, how is it not a riot?*" In contrast, *"@username Game tonight. Team is recently on fire! We will win!*" does not contain any situation awareness information that can be useful for first responders even though the term 'fire' is mentioned. Specifically, situational awareness refers to a state of understanding the 'big picture' in time- and safety-critical situations. Identifying situational awareness information is crucial in comprehending the different aspects of an event and possible development in the near future. In recent years, policy-makers and first responders have started to seek such information from Twitter, a popular social media platform, to make timely and informed decisions in a crisis-like event.

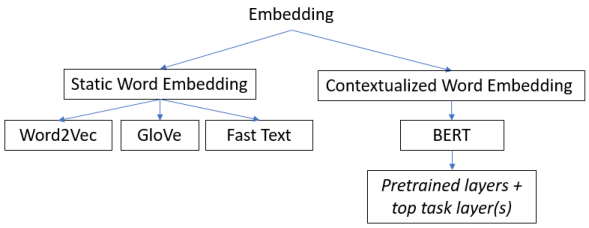


**Figure 1. Embedding methods used in this study**

In view of the vast number of possible scenarios or event types, for example, civil disorder, fire/explosion or armed assault, it can be a challenge to build a model that can generalize to most events and identify relevant situational awareness information from an unseen event. With the recent advancement in natural language processing (NLP), word embedding is now a de facto approach in almost all NLP applications. The promising results of word embedding (Devlin et al., 2019; Mikolov et al., 2013) have opened up many options for building a text classification model. This approach enables quick prototyping of NLP applications that use embedded vectors to encode words that have relevant context. In general, there are two types of word embeddings, static word embedding and contextualized word embedding (as depicted in Figure 1). In this study, we assess various static word embeddings, namely, Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and Fast Text (Bojanowski et al., 2017), because they are commonly used together with a

machine learning method, such as a classifier, for downstream tasks. The pretrained embedding of both Word2Vec and GloVe maps words to a representation vector, while Fast Text uses a subword method that captures the compositional information from the literal character sequences of language. Since the data source of this study is from social media, which is known to contain variants of a formal language, a subword representation of Fast Text may work better than the whole word-embedding methods of Word2Vec and GloVe. The other pretrained model family is based on contextualized word embedding. There are multiple options, such as ELMo (Peters et al., 2018) and GPT (Radford et al., 2018), but this study focuses on the most popular, the BERT model (Devlin et al., 2019). A contextual representation of a word is learned through language modeling, where given a context, a language model predicts the probability of a word occurring in that context. On top of the pretrained model, BERT allows for fine-tuning the trained model to adapt to specific contexts by training top layers via some training data. This is an attractive option since we are exploring whether a vanilla pretrained model (or a pretrained model without any fine-tuning) is sufficient for building a classifier that can identify content with situational awareness.

Even though social media can be a valuable source to extract relevant information for situational awareness, social media content is known to be informal with language variants, misspelled words or just noises. It is not uncommon to mix a few languages and use localized lingo to form a unique language to express emotion, especially in a multicultural environment (Zielinski & Bügel, 2012). This approach is evident as compared with a formal language (e.g., English), the localized vernacular language can better resonate with the community (Asiaone, 2015). Hence, it is not surprising that Twitter, as an informal channel for spreading news and information, is packed with localized or multilingual idioms so that messages can be conveyed more effectively. For example, the following two tweets sent during the 2013 Little India Riot in Singapore contain informal sentences (tweet 1) and localized lingo (tweet 2):

Tweet 1: "*A bus driver accidentally hit a Bangladeshi. His pals got angry, and thus, the riot began. Police car overturned, ambulance set on fire*."

Tweet 2: *"@username hahaha cannot experience liao! I see the bangla bo panf chance smash the bus window throw dust bin here and there!!*" [English translation: @username hahaha cannot experience anymore! I can see the Bangladeshi (or foreign workers) take the opportunity to smash the bus window and throw the dust bins everywhere.]

In this study, we assess the various pretrained models (Figure 1) and their ability to detect situational awareness information from tweets with localized lingo. In other words, we are interested in analyzing whether pretrained models can be used to build a binary classifier to differentiate tweets with situational awareness information from those without. Most studies focus on finding the best machine learning method or a classifier, but we would like to explore how we can use available NLP resources, such as pretrained models, to build a classifier that can be used in the real world. Even though resources (Olteanu et al., 2014) consisting of labeled tweets and event types are available for detailed study, detecting situational awareness information from an unseen event remains a challenge. Due to the promising result of contextualized word embedding and its well-known transfer learning ability, we would like to study whether a pretrained model can learn the contextual details of situational awareness content and thus generalize to other unseen events. Specifically, can a pretrained model that is trained on one event be used to detect situational awareness information from another event? In addition, most of these resources are curated from various countries and thus may not contain expressions commonly found in a Singaporean multicultural context. Therefore, it is of interest to study localized tweets and assess whether the approach developed can be deployed in real-world applications.

The main contributions of this work can be summarized as follows:

- We have assessed the suitability of various pretrained models in detecting situational awareness content from Twitter and observed that a vanilla pretrained model is insufficient to deliver on the task.
- The comparison between a typical classifier and a BERT sequence classifier reveals the importance of using BERT pretrained embedding with its sequence classifier to achieve the optimal result.
- In view of the localized languages found in tweets, it is essential to fine-tune and implement vocabulary extension to use the pretrained models.
- From our findings, detecting situational awareness information from social media can be a complex task and cannot be addressed using just pretrained models alone. We have observed that a pretrained model that is trained on one dataset cannot be used directly on another dataset to extract situational awareness content, and further testing is essential to ensure that the model can be generalized well for unseen events.

In the next section, we will discuss some related works in situational awareness identification and the use of pretrained models. Next, we outline the methods used and describe our results and findings in detail. We then discuss our observations of the findings and future plans before conclusions are drawn in the last section.

## 2. Literature Review

### 2.1. Situational Awareness Detection

Various methods have been proposed to identify situational awareness content from tweets. These include lexicons, traditional machine learning methods, deep learning methods and knowledge graphs. Olteanu et al. (2014) collected crisis-specific tweets, as these tweets created a lexicon revolving around crisis situations for training and automatically detecting crisis topics on Twitter. The authors' methodology shows great potential and improvement above manually labeled tweets. Ning et al. (Ning et al., 2019) used topic modeling to collect relevant tweets and designed a novel correlative convolutional neural network (CNN) to learn multifaceted features to identify crisis-awareness tweets. Zade et al. (2018) proposed a hybrid approach using a knowledge graph and deep learning to extract crisis-relevant tweets. Specifically, a Babelfy knowledge graph is used to extract relevant features. Even though deep learning approaches, for example, CNNs, perform well, if the training data are small (a few hundred), these approaches underperform compared to traditional machine learning methods, such as naive Bayes and support vector machines (SVMs) (Zhang et al., 2016). In this study, we focus on the effect of various pretrained models in detecting situational awareness information from social media content; these models include pretrained models built from static and contextualized word embeddings. We also pair the pretrained models with various classifiers, for example, SVMs and BERT sequence classifiers, to assess whether the traditional machine learning methods improve classification performance when the training data are small. Ultimately, we are interested in ascertaining whether a vanilla pretrained model is sufficient to detect situational awareness information and, if not, what can be done to adapt the pretrained models to handle localized content.

### 2.2. Use of Pretrained Models in Detecting Situational Awareness

Experimental results (Gonza´lez-Carvajal S, 2020; Ruder et al., 2019) comparing BERT with traditional machine learning models suggest that using BERT (or other models that can benefit from transfer learning) can significantly improve the performance of NLP models on disaster datasets partly due to the few data available for specific events, and transfer learning is promising since embedding models, such as BERT, are first pretrained via a large dataset to learn a wide range of general information about a language, and the small dataset can then be used to fine-tune the model for different NLP tasks. However, the common approach of situational awareness identification (Jain et al., 2019; Ning et al., 2019; Olteanu et al., 2014) is to use multiple crisis event types to train a classifier. Since we are interested in assessing whether a fine-tuned pretrained model can be generalized to predict an unseen event, we decided to design the study to use only specific events.

Fan et al. (2020) proposed a hybrid machine learning pipeline to uncover important and relevant information on disaster events. In particular, the BERT model is used for classifying posts with different humanitarian categories, such as injured people or infrastructure and utility damage. In the paper, the fine-tuned BERT model showed very promising results when compared with other deep learning models. Even though the classification categories are not exactly the same, ours being detecting information at a higher level, i.e., identifying if the tweets contain situational awareness, it is encouraging to know the BERT model is used successfully in a finer-grained classification. However, one observation is that the proposed pipeline is applied to only one case study – Hurricane Harvey in Houston. It remains unknown whether the BERT model can perform equally well if applied to other events. Hence, in this study, we used two case studies and two events, which are not exactly the same, to assess whether the fine-tuned BERT model can be applied to an unseen event.

Jain et al. (2019) compared the performance of BERT with previously developed NLP models on a group of disaster-specific classification benchmarks, and their analysis showed that BERT had no advantage over the other previous word-embedding methods, such as ELMo, Word2Vec, and GloVe. Due to the conflicting results of earlier studies, it is of interest to analyze the effect of pretrained models and recommend an approach that is feasible to apply on a real-world dataset. Since most of the approaches adopted some levels of fine-tuning of the pretrained model to leverage the benefit of transfer learning, this study also includes fine-tuning of pretrained models. In addition to fine-tuning, we assess whether the extension of vocabulary on the pretrained model affects performance. This assessment is used mainly to address the scarce resource localized language that

is commonly found in tweets, but is not represented in most pretrained models.

## 3. Methods

### 3.1. Twitter Data Collection

The data source is extracted from a Singapore tweet source (2013-little-india-riot) (Lim & Achananuparp, 2012) and a multilingual social media study (Lo et al., 2017). The data source contains tweets and metadata related to the 2013 Little India Riot event. This is the worst case of public violence and civil disorder in Singapore in over four decades. The riot erupted in Little India on Sunday night, 8 December 2013 and lasted for approximately two hours. The trigger point was the fatal traffic accident between a foreign worker and a private bus. Some 300 rioters took part in the unrest, with 54 responding officers and 8 civilians injured, while 29 vehicles were damaged, 5 of which were burned. From 8 December 2013 9:00 PM to 13 December 2013 8:00 AM, 183,000 tweets covering the event and the aftermaths were published. These tweets were collected from a set of Twitter users whose profile location is Singapore or geotagged as Singapore in real time by using the Twitter public stream API. To retrieve related tweets, a list of keywords was used to extract the Little India Riot tweets for this study. The keywords include riot, Little India, policeman, accident, fireman, and alcohol. We selected the Little India Riot event because it contains multiple event types, such as civil disorder, traffic incidents, assault, injury/casualty and fire/explosion, which can potentially be used to identify other situational awareness information of the various event types.

### 3.2. Data Analysis and Annotation

With Singapore as a multicultural society, it is of interest to analyze the language covered in the dataset. The langdetect Python library is used for this purpose, and the analysis shows that 91.66% of the dataset collected is English, followed by Malay/Indonesian, which, at 3.45%, represents the second largest portion of the data, and the rest of data are of mixed language at 4.92%. This finding is important to understand if pretrained models, which are trained using English, can be used to handle a dataset with a mixture of languages.

Further analysis via dictionary matching (Lo et al., 2017) detected the presence of Singlish (14.67%), the colloquial Singaporean English that has incorporated elements of some Chinese dialects and the Malay language (Leimgruber, 2011). This dictionary is constructed by consolidating several internet resources and contains 978 unique Singlish terms. The same dictionary is used as the vocabulary extension for the pretrained BERT model in this study and the details will be mentioned in a later section.

Since the dataset collected is unlabeled, the tweets were further filtered to reduce the number of tweets to construct the ground truth datasets. The filtering process consists of two parts – automatically annotating tweets using the occurrence of keywords, followed by stratified random sampling. We utilize word clouds to select a list of keywords that are relevant to urban crisis events such as riot, fire, hurt etc., and make use of Gensim Word2Vec (both Skip-gram and CBOW algorithms) (Mikolov et al., 2013) to derive a list of high-similarity words from the selected keywords. Thereafter, we conducted stratified random sampling based on the time of the tweets, which totaled 2,277 tweets, and conducted the manual annotation process. Two annotators who are familiar with Singlish and the event annotate the content to indicate if the tweet contains any situational awareness information. Annotation of ‘yes’ or ‘no’ was used, depending on the questions below:

- Does the tweet contain any injury/casualty or infrastructure damage information?
- Does the tweet show any urban crisis detail, for example, fire/explosion or weapon, about the event?

### 3.3. Data Preprocessing and Feature Analysis

Unlike text data from news sources, tweets are known to be noisy and often mixed with linguistic variations. It is hence very important to clean up the tweet before any content extraction. In this study, as a form of basic preprocessing, we remove mentions, URL links, emojis, smileys, stop words and additional white spaces from the tweets. Lemmatization is also used to remove inflectional endings and keep the dictionary form of each word. The overall architecture is shown in Figure 2. In addition to the basic preprocessing, feature analysis that assesses the effect of hashtag, digit and punctuation are included. Since we are interested in extracting situational awareness information, one of the special features of tweets, the hashtag, which can contain event information, is kept for feature analysis. In addition, the digit or number mentioned in tweets may carry important details, such as the number of people injured or the location information that can be useful. To better understand the context and meaning of the content, we decide to keep punctuation since it preserves the sentence structure. Feature analysis is followed by two different classification approaches. The first, or the typical classifier, makes use of a static word embedding and a

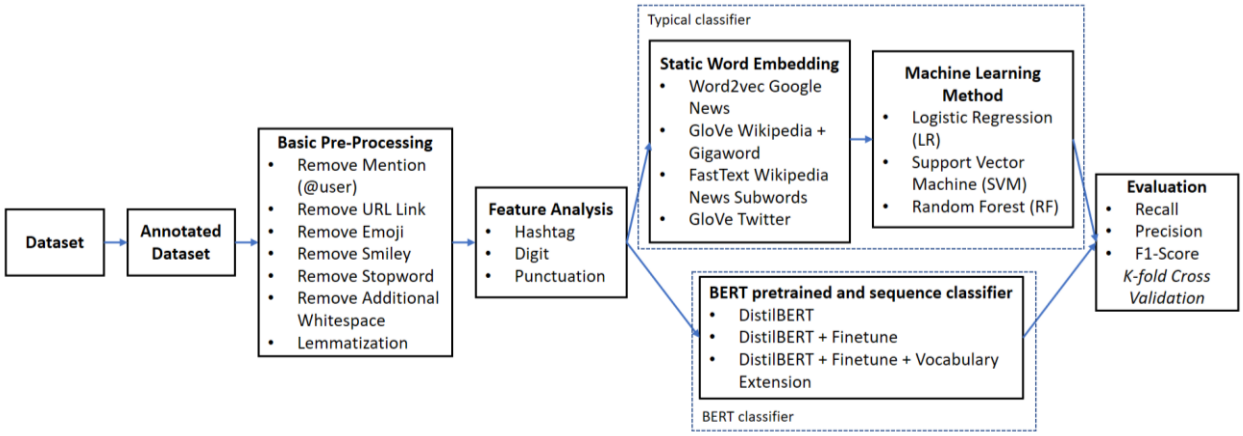


**Figure 2. The Overall Architecture**

machine learning method; the second, or the BERT classifier, uses the popular BERT (Devlin et al., 2019) pretrained approach. We adopted k-fold cross-validation and the evaluation is based on recall, precision and the F1 score. A detailed explanation of each component is as follows.

Even though the BERT classifier is known to work with raw data (without any preprocessing) (Devlin et al., 2019), we have included the preprocessing and feature analysis for all the classifiers (including the BERT classifier) so that we can be consistent with the comparison.

### 3.4. Static Word Embedding

Various pretrained word-embedding models that are trained using large corpora, such as Wikipedia or news corpora, have been released for many NLP downstream tasks. Two of the popular static pretrained models are GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013). GloVe, which stands "Global Vectors for word representation", is trained on Wikipedia 2014 and consists of 1.9 million vocabularies. Various embedding dimensions of 50, 100, 200 and 300 are available. The Word2Vec pretrained model includes word vectors of 3 million words and phrases that are trained on approximately 100 billion words from a Google news dataset and stored using a 300-dimension embedding. In addition, since both Word2Vec Google News and GloVe Wiki use formal language and this study uses tweets, additional embedding, FastText Wiki (Bojanowski et al., 2017) and GloVe Twitter are adopted to include embedding models that are pretrained with informal languages, thereby providing a more balanced evaluation. FastText uses character-based embedding and allows rare words, which can be common in tweets, to be represented appropriately. In addition to GloVe Wiki, GloVe Twitter is also used to analyze whether it is a better representation for the dataset. Specifically, the four static embeddings used in this study are word2vec-google-news-300, glove-wiki-gigaword-300, glove-twitter-200, and fasttext-wiki-news-subwords-300 distributions from Gensim.

### 3.5. Machine Learning (ML) Method

After the tweets have been represented by a word-embedding method, it is common to use an ML model for classification. In this study, we have adopted common ML methods that have performed well in NLP research. Specifically, three methods are selected. They are the support vector machine (SVM), logistic regression (LR) and the random forest (RF) from scikit-learn implementation. Default parameters are adopted in all testing unless explicitly stated. Due to the high dimensionality of NLP representation, the number of estimators for RF is set to 500 instead of the default 100. In our preliminary study, naïve Bayes did not perform well and hence was excluded from this study.

### 3.6. BERT Pretrained and Sequence Classifier

Currently, one of the most popular state-of-the-art models in the NLP domain is BERT (Devlin et al., 2019), which stands for "Bidirectional Encoder Representations from Transformers" and is trained on an unlabeled, plain text corpus from BooksCorpus (800 M words) and English Wikipedia (2,500 M words). BERT is designed to read in both directions to achieve better performance in terms of masked language modeling and next sentence prediction. In contrast to the traditional method of static word embeddings – GloVe and Word2Vec, which map every word to a fixed vector, BERT allows for each word to be defined by their surrounding context

through a bidirectional self-attention mechanism. This ability can be an added advantage to understanding informal tweets.

Even though BERT is popular and has achieved promising results, training a BERT model can be resource intensive and requires substantial computing power. With the aim of eventually deploying the model as a real-world application, we adopt a more optimized and lightweight model from the BERT family, DistilBERT (Sanh et al., 2019). DistilBERT is trained with the same corpus as the BERT model and is reported to be 40% smaller and 60% faster than BERT. DistilBERT uses knowledge distillation compression techniques where the compact model is trained to reproduce the same behavior of a larger model, thus allowing DistilBERT to retain 97% of BERT performance. Henceforth, we make use of the DistilBERT model to evaluate the dataset by utilizing additional techniques, i.e., fine-tuning and vocabulary extension, to assess the effectiveness of a pretrained model on the classification of a localized dataset.

We adopted HuggingFace's DistilBertForSequenceClassification (HuggingFace, n.d.) in this study. The pretrained model used is DistilBERT-base-uncased, and the training parameters are Batch Size: 8; Loss Functions: CrossEntropyLoss; Learning Rate: 1e-05; Optimizer: AdamW and Adam Epsilon: 1e-8

### 3.7. Fine-Tuning and Vocabulary Extension

Fine-tuning a pretrained model essentially updates the pretrained weights based on the new dataset that was fed into the model to adapt the model to a domain-specific dataset. During the training phase, each data point traverses through all neurons from the first to the last layer by using a forward pass method, followed by a backward pass to adjust the weights and biases to optimize the cost function. This process is done through several epochs of training while keeping track of the training and validation loss to facilitate saving the best model, which is determined by the lowest validation loss.

Off-the-shelf word-embedding models are not trained with datasets that contain Singlish text. Henceforth, our study makes use of the vocabulary extension approach, i.e., the out-of-vocabulary technique, as an additional evaluation to extend the vocabulary of the pretrained embedding model to incorporate Singlish text. To facilitate this process, we made use of the Singlish dictionary mentioned earlier. Since some of the Singlish terms are meaningful only in N-gram format, we generated a list of unique N-grams, where N ranges from 1 to 3. For example, '*goyang kaki*' [meaning 'relax' but the literal English translation is 'shake leg'] and '*confirm plus chop*' [meaning 'absolutely', but the literal English translation is 'confirm and stamp'] can have different meanings if the N-gram is not considered. To ensure that each of these N-grams is relevant to the dataset and to address the potential misspelled words or phrases, we conducted another round of filtering by using Jaccard similarity to match each N-gram against every tweet and retain only the found Singlish term with a similarity score of 0.7 and above. Finally, we add these Singlish terms into the pretrained model vocabulary and initiate a resizing of the token embeddings before fine-tuning the model.

### 3.8. Evaluation

Two methods are used in this study. 1) Repeated K-fold cross-validation and 2) grid search K-fold cross-validation. Repeated K-fold cross-validation is adopted in feature analysis since it is more important to assess the impact of the features rather than the model performance. The whole dataset is used to minimize the influence of different distributions of performance scores due to different dataset splits. The results are extracted from fivefold cross-validation with three repeats to address the noisy estimation of model performance. Grid search K-fold cross-validation is used to identify the best parameters for the ML method. The source data are split into two parts – 80% training and 20% testing. The results reported are based on the best model after a fivefold grid search of the optimal parameters for the ML methods on the training data. In other words, a fivefold cross-validation is performed on the training data together with a range of parameter values to identify an optimal model for the testing data. Considering that the total number of Little India source data is 2,277 records, 1821 records (or 80%) are split into five portions with validation data having 365 records. The testing data are 456 records (or 20%). An explicit random seed is set for reproducibility of the results.

Typical accuracy metrics used for statistical analysis of binary classification consider the true positive and true negative and have known issues in terms of reflecting the performance of a classifier (Sokolova et al., 2006). Therefore, we used the F-measure or F1 score as the metric when assessing the performance of the various approaches proposed. The F1 score is the harmonic mean of both precision and recall, where precision is defined as the ratio of true positives found from the predicted positives, while recall is the ratio of true positives identified from the actual positives.

## 4. Results

### 4.1. Feature Analysis

Since there are multiple features to analyze, namely, digit, hashtag, and punctuation, this study analyses the combinations of the following:

1) Four static word-embedding models (as specified in Figure 2) - Word2vec Google News (w2vGoogleNews), GloVe Wikipedia + Gigaword (gloveWikiGiga), FastText Wikipedia News Subwords (fastTextWikiNews) and GloVe Twitter (gloveTwitter).
2) ML methods – SVM, RF and LR

Each feature is removed individually and collectively to assess the impact on the performance of different word-embedding and ML methods. Due to the large number of combinations, Table 1 displays the results concerning top 10 F1 scores from the various configurations with the three ML methods used in this study.

**Table 1. Feature analysis using various ML methods and word embedding** (selecting only the configurations of top 10 F1 score and ranked by F1)

| ML | Word embedding | Features removed | Precision | Recall | F1 |
|---|---|---|---|---|---|
| SVM | gloveWikiGiga | None | 0.844 | 0.739 | **0.784** |
| SVM | gloveTwitter | Hash | 0.853 | 0.722 | 0.78 |
| SVM | gloveTwitter | None | 0.849 | 0.724 | 0.78 |
| SVM | w2vGoogleNews | Digit | 0.853 | 0.725 | 0.779 |
| SVM | gloveWikiGiga | Punc | 0.79 | **0.777** | 0.779 |
| SVM | w2vGoogleNews | None | 0.795 | 0.765 | 0.774 |
| LR | w2vGoogleNews | Digit | 0.837 | 0.728 | 0.774 |
| LR | w2vGoogleNews | All | 0.842 | 0.717 | 0.77 |
| SVM | fastTextWikiNews | All | 0.798 | 0.748 | 0.767 |
| RF | gloveTwitter | Punc + Digit | **0.866** | 0.684 | 0.76 |

As Table 1 shows, SVM obviously outperforms LR and RF in terms of recall and the F1 score. RF, on the other hand, outperforms SVM and LR in terms of the precision score, although RF scores mostly the lowest in recall. When comparing the static embedding models in terms of the results from all the SVM models, gloveWikiGiga and gloveTwitter are the better performing static embedding models. Even though the top 5 SVM configurations have similar performance, SVM with gloveWikiGiga as the word-embedding model is the top performer with none of the features removed. Due to the real-time broadcasting nature of social media, it enables early detection and thus high recall is preferred to capture all the relevant situational awareness content. In view that the configuration of SVM model using gloveWikiGiga as embedding achieves the highest recall, the punctuation feature is removed for all subsequent models used. Surprisingly, fastTextWikiNews does not perform as well even though it is chosen with the assumption that it may encode tweets better.

**Table 2. Performance analysis using different ML methods on Little India testing data** (using gloveWikiGiga as the embedding)

| ML methods | Precision | Recall | F1 |
|---|---|---|---|
| SVM | 0.821 | **0.727** | **0.771** |
| LR | 0.821 | **0.727** | **0.771** |
| RF | **0.874** | 0.683 | 0.767 |

As observed in Table 2, SVM and LR perform well in predicting test data of the same event (i.e., Little India riot). Even though RF has a very high precision, in our context, high recall is more important since it is essential to capture as much relevant content from the test data.

### 4.2. BERT Classifiers

BERT is often used as a pretrained embedding and a classifier together since most implementations include a sequence classifier. In this study, we are interested in assessing whether the vanilla version of a pretrained model is suitable for situational awareness detection on tweets. Therefore, we design the following setup to evaluate the BERT classifiers:

1) DistilBERT pretrained model + sequence classifier (vanilla mode, i.e., no fine-tuning)
2) DistilBERT pretrained model + sequence classifier + fine-tune
3) DistilBERT pretrained model + sequence classifier + fine-tune + vocabulary extension

The first is essentially the vanilla version of the BERT classifier. The second is a fine-tuned version with source data in the hope that the model will learn a more relevant context according to the annotated data. The third is an approach to address the localized language variations and Singlish that was detected in the source data.

For the vanilla pretrained model, our aim is to assess if the model can separate the tweets into two groups without knowing the task beforehand. Based on the result shown in Table 3, it is obvious that with

an F1 score of 0.467, the model is as good as a random guess in identifying situational awareness information from tweets. However, the performance of the pretrained model achieves a significant improvement with fine-tuning by using the annotated data. The result is further improved with vocabulary extension. It is evident that there is a need to adopt a further step to build a more promising model if the source data are not well represented by the content in the pretrained model. It is observed that DistilBERT pretrained + Fine-Tune + Vocab Ext performed the best (in terms of recall and the F1 score) among all the models tested (including the typical classifier approach in Table 2).

**Table 3. Results of DistilBERT on Little India testing data**

| | Precision | Recall | F1 |
|---|---|---|---|
| Pretrained | 0.418 | 0.530 | 0.467 |
| Pretrained + Fine-Tune | 0.728 | 0.820 | 0.771 |
| Pretrained + Fine-Tune + Vocab Ext | **0.745** | **0.847** | **0.793** |

Considering that SVM has consistently performed well on the typical classifier approach (both Tables 1 and 2), the embedding from DistilBERT pretrained + Fine-Tune + Vocab Ext is extracted to work with the SVM classifier. In other words, instead of the sequence classifier, the DistilBERT embedding (after fine-tuning and vocabulary extension) is paired with SVM. Table 4 shows the result and the need to pair BERT embedding with the DistilBERT sequence classifier to achieve better performance. This is likely due to the training of neural network and thus allowing the sequence classifier to leverage the contextualized embedding of DistilBERT.

**Table 4. Performance analysis using the embedding from DistilBERT pretrained + Fine-Tune + Vocab Ext (from Table 3) with two classifiers**

| | Precision | Recall | F1 |
|---|---|---|---|
| SVM classifier | 0.790 | 0.759 | 0.774 |
| DistilBERT sequence classifier (Table 3) | 0.745 | 0.847 | **0.793** |

## 4.3. Analysis on an Unseen Dataset

It is of interest to further analyze whether the results observed in Table 3 are reproducible in another event of a slightly different nature. An unseen dataset is extracted from the same Singapore tweet source (Lim & Achananuparp, 2012). The earlier 8 December 2013 Little India Riot event had incidents of traffic accidents, civil disorder (with assault incidents) and fire/explosion, while the new unseen dataset, the 25 June 2013 Orchard Cineleisure Slashing event, was mainly an armed assault incident. However, both contain situational awareness information that can be important for first responders and police forces. A total of 448 records were extracted for the Orchard slashing dataset. In view of the smaller dataset, we reserved 135 records (or 30%) as the testing data and retained 313 records (or 70%) as the training data.

**Table 5. Results of DistilBERT on Orchard slashing unseen testing data**

| | Precision | Recall | F1 |
|---|---|---|---|
| Pretrained | 0.474 | 0.403 | 0.435 |
| Pretrained + Fine-Tune | 0.802 | 0.782 | 0.792 |
| Pretrained + Fine-Tune + Vocab Ext | **0.821** | **0.821** | **0.821** |

As presented in Table 5, the same trend of Table 3 is observed and hence confirms that a vanilla BERT pretrained model should not be used, but instead, further enhancement, specifically fine-tuning and vocabulary extension, will significantly improve the result.

**Table 6. Performance analysis using different ML methods on Orchard slashing unseen testing data**

| ML methods | Precision | Recall | F1 |
|---|---|---|---|
| SVM | **0.820** | 0.746 | 0.781 |
| LR | 0.816 | 0.731 | 0.772 |
| RF | 0.809 | **0.784** | **0.784** |

As the results of Tables 5 and 6 show, adopting a contextualized word-embedding model, such as BERT with fine-tuning and vocabulary extension, has an obvious benefit when handling tweets. The enhanced DistilBERT pretrained model clearly outperformed all other models.

## 4.4. Are the Pretrained Models Generalizable?

Since the pretrained BERT models are known to capture the context of the sentences and should be superior to keyword matching alone, which can be specific in different events, it is of interest to analyze whether the model trained by the Little India dataset can be used to identify situational awareness tweets from Orchard slashing events and vice versa.

**Table 7. Analysis of DistilBERT classifier to generalize when train and test are using different datasets**

| Little India dataset | | | Orchard slashing dataset | | |
|---|---|---|---|---|---|
| Orchard slashing dataset | | | Little India dataset | | |
| Precision | Recall | F1 | Precision | Recall | F1 |
| 0.500 | 0.067 | 0.118 | 0.583 | 0.015 | 0.030 |

As Table 7 shows, when the DistilBERT model is trained with a different dataset from the test dataset (first row as train and second row as test), the model performs extremely poorly, thereby indicating that the model cannot generalize well. This result is likely because the trained models only recognize the context involved in a specific event and cannot detect the situational awareness information from another unseen event. This finding is crucial because most of the research papers (Fan et al., 2020, Jain et al., 2019) have applied their pretrained model on the same dataset and achieved promising results. We have shown that the very promising model does not work as well when applied to another real-world event. This finding highlights that situational awareness detection classification is more complex than other binary classifiers (such as sentiment analysis), and a fine-tuned pretrained model, even with vocabulary extension of localized language, cannot identify the situational awareness content of a different event.

## 5. Discussion

One benefit of using pretrained embeddings is the ability to build a classifier or other AI models without the need to find large text corpora to preprocess and train with appropriate settings to capture the necessary context. Another benefit is the saving on the high computation cost and long training time to access a good quality embedding on a similar domain of the data source. However, it is important to highlight that most of the available pretrained models are trained using English and general content, and these models may not be suitable for capturing localized content (as commonly found on social media). Tables 3 and 5 clearly show that a vanilla pretrained model should not be used directly and that enhancements, such as fine-tuning and vocabulary extension, are essential.

In this study, we have used preprocessed tweets in all the models to have a consistent comparison. However, since BERT is known to work with non-processed dataset, we conducted additional study for fine-tuned version of DistilBERT. In other words, two new models are trained with the original tweets with no preprocessing done and the results are presented in Table 8. Comparing the effect of using preprocessing and no preprocessing, it is observed that incorporating preprocessing in data preparation achieving a better result. This is likely due to the informal nature of tweets with localized context that may not be well encoded by DistilBERT. Interestingly, we also observed from Table 8 the positive effect of vocabulary extension, which is consistent with the observation of using preprocessed data. Specifically, Fine-Tune+Vocab Ext (No preprocessing) (F1=0.777) model has better performance than Fine-Tune (No preprocessing) (F1=0.755).

**Table 8. Effect of preprocessing on DistilBERT models**

| DistilBERT models | Precision | Recall | F1 |
|---|---|---|---|
| Fine-Tune (No Preprocessing) | 0.641 | 0.918 | 0.755 |
| Fine-Tune | 0.728 | 0.820 | **0.771** |
| Fine-Tune + Vocab Ext (No Preprocessing) | 0.814 | 0.743 | 0.777 |
| Fine-Tune + Vocab Ext | 0.745 | 0.847 | **0.793** |

One limitation of the current study is using just one variant of BERT models, specifically, the DistilBERT. It is of interest to compare the findings with the full BERT model or adopting pre-trained model trained with social media sources, such as BERTweet, to assess if other BERT models may be more suitable in identifying situational awareness information from tweets.

Even though the result from this study depicts the potential of a BERT classifier, this study did not consider specific computing resources, such as a graphical processing unit (GPU), which are often necessary to achieve a reasonable processing time. The typical classifier approach using the more resource-friendly embeddings, i.e., gloveWikiGiga or w2vGoogle News, performs comparably with traditional machine learning methods, such as SVM. It is thus not necessary to employ the latest or the state-of-the-art methods if there is a resource constraint. However, both approaches, i.e., both the typical classifier and the BERT classifier, suffer from the inability to generalize. Further analysis on SVM on training and testing using different datasets has a similar trend, as shown in Table 7. In other words, even though the model is trained to differentiate if the tweet contains information on situational awareness, the model works well mainly on seen events and is not as promising on unseen events. A new DistilBERT model combining both Little India and Orchard slashing datasets achieved F1 score of 0.797 and a recall of 0.803, which shows the need to include all types of events to train a more robust model. Even so, there is no guarantee that the model can perform well in detecting an event that has never happened. Preliminary study using crisis datasets containing multiple event types such as CrisisLex26 also shown similar results. The following new approaches may likely create a model with better generalization capability: 1) sequence tagging (Yu et al., 2015), which focuses on identifying elements with situational awareness details instead of encoding the events; and 2) expanding the training data by using distant

supervision by external knowledge bases (Alrashdi & O'Keefe, 2020), thus increasing the linguistic variations and improving the recall on unseen events. Our future work includes fine-tuning with various possible urban events for better generalization, extracting location information for data filtering and more precise information extraction, and exploring sequence tagging approach rather than a binary classification approach to identify potential multiple event types from a single tweet.

## 6. Conclusion

Is a pretrained model the answer for situational awareness detection on social media? One obvious answer is not the vanilla BERT pretrained model. This study has demonstrated the need to perform fine-tuning and vocabulary extension on a pretrained BERT model before the model can be used on localized social media content. It is especially so since the content of social media, or in our study, tweets, does not follow the standard structure and vocabulary of English that most pretrained models are built. In addition, we have observed that a pretrained model that is trained on one dataset cannot be generalized to extract situational awareness tweets from an unseen event. It is thus essential to perform further testing on a promising classifier that is built using a pretrained model (be it from static or contextualized word embeddings) to ensure that the model can generalize well for various events before it is deployed for any real-world application.

## 7. References


Alrashdi, R., & O'Keefe, S. (2020). Automatic Labeling of Tweets for Crisis Response Using Distant Supervision. *WWW 2020*, pp. 418-425

Asiaone. (2015). Officials use Singlish, dialects to reach out. https://www.asiaone.com/singapore/officials-use-singlish-dialects-reach-out

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the ACL, 5*, pp. 135-146

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019*, pp. 4171–4186.

Fan, C., Wu, F. and Mostafavi, A., (2020). A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access*, *8*, pp.10478-10490.

Gonza´lez-Carvajal S, G.-M. E. (2020). *Comparing BERT against traditional machine learning text classification*. https://arxiv.org/abs/2005.13012

HuggingFace (n.d.) https://huggingface.co/docs/transformers/model_doc/distilbert#transformers.DistilBertForSequenceClassification

Jain, P., Ross, R., & Schoen-Phelan, B. (2019). Estimating distributed representation performance in disaster-related social media classification. *Proceedings of ASONAM 2019*, pp. 723-727

Leimgruber, J. R. E. (2011). Teaching and Learning Guide for: Singapore English. *Language and Linguistics Compass*, *1*, pp. 47–62.

Lim, E. P., & Achananuparp, P. (2012). Palanteer: A search engine for community generated microblogging data. *Lecture Notes in Computer Science 7634*, pp. 239-248

Lo, S. L., Chiong, R., & Cornforth, D. (2017). An unsupervised multilingual approach for online social media topic identification. *Expert Systems with Applications*, *81*, pp. 282-298

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*, pp. 3111-3119

Ning, X., Yao, L., Benatallah, B., Zhang, Y., Sheng, Q. Z., & Kanhere, S. S. (2019). Source-aware crisis-relevant tweet identification and key information summarization. *ACM Transactions on Internet Technology*, *19*(3).

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). CrisisLex: A lexicon for collecting and filtering Microblogged communications in crises. *Proceedings of ICWSM 2014*, pp. 376-385

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014*, pp. 1532-1543

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *NAACL HLT 2018*, pp. 2227-2237

Radford, A., Narasimhan, T., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training.

Ruder, S., Peters, M., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing tutorial. *NAACL HLT 2019*, pp. 15-18

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *arXiv*.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *AAAI Workshop - Technical Report*, *WS-06-06*.

Yu, Z., Huang, W., & Xu, K. (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging*. arxiv:1508.01991

Zade, H., Shah, K., Rangarajan, V., Kshirsagar, P., Imran, M., & Starbird, K. (2018). From situational awareness to actionability: Towards improving the utility of social media data for crisis response. *Proceedings of the ACM on Human-Computer Interaction*, pp. 1-18

Zhang, G., Kato, J., Wang, Y., & Mase, K. (2016). How to initialize the CNN for small datasets: Extracting discriminative filters from pre-trained model. *ACPR 2015*. https://doi.org/10.1109/ACPR.2015.7486549

Zielinski, A., & Bügel, U. (2012). Multilingual analysis of twitter news in support of mass emergency events. *ISCRAM 2012*, pp. 77-85