Real Time Analysis of Accident Related Tweets By Mining Twitter Data

Ameya S. Gamre

University of Texas at Dallas

***Abstract*—In times of accidents, vast amounts of data are generated via computer-mediated communication (CMC) that are difficult to manually cull and organize into a coherent picture. Yet valuable information is broadcast, and can provide useful insight into time- and safety-critical situations if captured and analyzed properly and rapidly. We describe an approach for automatically identifying messages communicated via Twitter that contribute to situational awareness, and explain why it is beneficial for those seeking information during accidents. We will collect Twitter messages and build a classifier to automatically detect messages that may contribute to situational awareness, utilizing a combination of hand annotated and automatically-extracted linguistic features. Our system will categorize tweets that contribute to situational awareness. The results are promising, and have the potential to aid the general public in culling and analyzing information.**

I. Introduction

IN this Digital Age, social media has played a major role in everyday life. It has become a major source of information for people and the only medium where the masses can voice their opinions and concerns. We intend to use this data for real-time analysis of accidents in Dallas. We extend this research by focusing on Twitter communications (tweets) generated during accidents, and show how Natural Language Processing (NLP) techniques contribute to the task of sifting through massive datasets when time is at a premium and safety of people and property is in question. So much information is now broadcast during accidents that it is infeasible for humans to effectively find it, much less organize, make sense of, and act on it. To locate useful information, computational methods must be developed and implemented to augment human efforts at information comprehension and integration. The popular microblogging service Twitter serves as an outlet for many to offer and receive useful information; it provides a way for those experiencing such emergency to gather more, or different, information than they may be able to using mainstream media and other traditional forms of information dissemination. This access provides affected populations with the possibility to make more informed decisions. The challenge, however, is in locating the right information. In addition to broadcasting valuable, actionable information via Twitter during accidents, many also send general information that is void of helpful details, or communicate empathetic, supportive messages that lack tactical information. Tweets that include tactical, actionable information contribute to situational awareness; such tweets include content that demonstrates an awareness of the scope of the crisis as well as specific details about the situation. We offer an approach for automatically locating information that has the potential to contribute to situational awareness in the multitude of tweets broadcast during accidents. Our overarching goal is to help affected populations cull and analyze pertinent information communicated via computer-mediated communication. Our assumption is that immediate, dynamic culling of tweets with information pertaining to situational awareness could be used to inform and update applications aimed at helping members of the public, formal response agencies, aid organizations and concerned outsiders understand and act accordingly during accidents.

II. Previous Work

In the past, several studies have analysed the use of social network and media for event detection. The role of Natural Language Processing in this area is huge as there are large amounts of textual data floating around. Angelica [1] in 2017 showed us how they extracted tweets using the Twitter streaming API, manually labeled the data and split them in training and tests sets, and trained the SVM classifier. The features used were different n-gram ranges. The two classes used were Traffic and Non-traffic.Tweets that were traffic related but did not possess any information about the location of the incident, were labeled as non-traffic related tweets. SVM model gave a very good performance across different n-gram ranges features. Sudha [2] stated how NLP can be used to gauge situational awareness during mass emergency from tweets. Here, they explained what is situational awareness tweet content, objective tweet content, subjective tweet content, register tweet content and personal/impersonal tweet content. The features used here were Unigrams, Bigrams, Part-of-speech tags, subjectivity of tweets, register of tweet and the tone of tweet. Classifiers were implemented to predict subjectivity, register and the tone of the tweet. Brinda [3] used sentiment analysis to analyze demonetization tweets. The steps involve Twitter data collection, preprocessing, feature extraction and the models used. The techniques for feature extraction were the bag of words model, TF-IDF (Term Frequency-Inverse Document Frequency), and N Grams. Naive Bayes Algorithm(NB), Support Vector Machines(SVM) and Logistic Regression(LR) were the models used. It showed that LR gave the best accuracy but

also took longer than NB but lesser than SVM to complete. Sara [4] demonstrates the use of n-grams and POS tags to extract features from tweets. False positives are of greater concern, since they represent noise that could be misleading [5]. To address them, ways to incorporate user feedback; increasing the weight of tweets that have been retweeted; and the effects of using different limits in our machine learning algorithms were studied. Because Twitter has a high degree of redundancy, it is less likely that all tweets that represent the same information and are written in different styles will be misclassified. To measure classifier accuracy, they tested a sample of manually annotated tweets [5]. As a deployed system, it will continuously classify incoming tweets based on models built on data from previous similar events. The most important contribution this study makes to social media research is to demonstrate that using sentiment analysis to learn from customers is likely less effective than humans reading streams of consumer chatter. This result [5] is invaluable for improving social media monitoring practices. Empirical proof that an NLP approach is potentially superior to Sentiment Analysis suggests that efforts to build an information system based on NLP techniques are a worthwhile and beneficial goal. NLP-based software promises the potential to substantially increase the knowledge firms may glean from tapping into customer-to customer exchanges and enhance the effectiveness with which they monitor and respond to customer to-firm communications [5]. [6] experiments with with word features unigram, bigram, trigram and tetragram.

III. Methodology

In this section, we present the approach to retrieve, process and classify accident related tweets.

A. Data Collection

First, old tweets were extracted using a library which allows us to get unlimited tweets. Jefferson [7] made available a library which helps walk around Twitter Official APIs limitation where we cannot get tweets older than a week.

B. Preprocessing

- Tokenization: This step involves splitting text into a set of tokens(words) . Tweets can contain a lot of characters such as emoticons, special characters, hashtags, etc. which can get in the way while we are training a classifier model. Hence punctuation, URLs were removed using regular expressions. Porter stemming algorithm was used.
- Stop word removal: Stop words were removed using the NLTK library.

Example:
Original Tweet:Accident in #Forney on Hwy 80 WB at Clements Dr stop and go traffic back to FM-740 delay of 10 mins #DFWTraffic http:// bit.ly/14TuwzZ
Pre-procesed tweet: Accident Forney Hwy 80 WB Clements Dr stop go traffic back FM740 delay 10 mins DFWTraffic

C. Classification

Here the task is to classify tweets into 'situational' and 'non-situational'. After going through a lot of research papers, it was decided to use Multinomial Naive Bayes for this job.

D. Feature Extraction

We use lexical features like word ngram features. Whether estimating probabilities of next words or of whole sequences, the N-gram model is one of the most important tools in speech and language processing. N-grams are crucial in NLP tasks like part-ofspeech tagging, natural language generation, and word similarity, as well as in applications from authorship identification and sentiment extraction to predictive text input systems for cell phones [8]. This is the motivation for using the N-gram model for feature extraction.
The n-gram ranges used were unigrams, bigrams, trigrams, unigram + bigrams and unigrams + bigrams + trigrams. A feature vector consisting of unigrams will count the frequency (or inverse document frequency) of the word "happy". Conversely, a feature vector consisting of bigrams could give us more details for features like "not happy" or "very happy", which gives us more insights into the general sentiment of a tweet. For this, Sklearn's Count Vectorizer method was used. It converts a collection of text documents to a matrix of token counts. This implementation produces a sparse representation of the counts using scipy.sparse.csr_matrix.
TF-IDF transformation is applied to the feature vector. Sklearn's TfidfTransformer transforms a count matrix to a normalized tf or tf-idf representation. Tf means term-frequency while tf-idf means term-frequency times inverse document-frequency. This is a common term weighting scheme in information retrieval, that has also found good use in document classification. The goal of using tf-idf instead of the raw frequencies of occurrence of a token in a given document is to scale down the impact of tokens that occur very frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus.

IV. Implementation

A. Datasets

After going through a lot of tweets, the accident related keywords were narrowed down to two, 'accident' and 'crash' as words like 'wreck' and 'mishap' were not used often to refer accidents. 1497 tweets were extracted from near the Dallas region that had the accident related keywords and each tweet was labeled situational or non-situational manually. Tweets that had the keywords and a location were labeled situational whereas tweets that had the keywords but lacked a location were labeled non-situational. 997 tweets containing the keywords, were set aside as training set and 500 tweets were set aside for the

test set. A 10-fold cross validation on the training dataset, using different n-gram ranges was performed.

B. Evaluation Metrics

The following four measurements were made in order to evaluate the performance of the classifier: True Positives (TP), False Positives (TP) True Negatives (TN) and False Negatives (FN). True negative and true positive are non-situational and situational related tweets, which were classified correctly as non-situational and situational related, respectively. False negative tweets are those situational tweets that were misclassified as non-situational, whereas false positive tweets are those non-situational tweets that were misclassified as traffic situational. From these values, we can then calculate the following statistical metrics: a) Accuracy is the fraction of the classification that is correct. It's calculated by dividing the correctly classified tweets by the total number of tweets (1).

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

b) Precision of a class is the fraction of correctly classified tweets out of all tweets classified to that class (2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

c) Recall of a class is the fraction of correctly classified tweets out of all tweets that actually belong to that class (3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

d) F1-score is the harmonic mean of precision and recall (4).

$$F1Score = \frac{2PR}{P + R} \tag{4}$$

C. Live Tweets

Finally, there was a section put out for live tweets wherein if we tweet through our phones from near the Dallas region, it would appear on the command line and along with it, would appear its predicted class. The block diagram in Figure 1 shows the flow.

V. Results

Table II shows the results for Precision, Recall, F1 score and Accuracy for different N-gram ranges for Multinomial NB classifier on the validation sets. It can be seen that unigrams had the highest accuracy while trigrams had the worst. Hence the unigrams features are used to evaluate to the classifier on the test data set. The accuracy on the test set came out to be 97.79%.

VI. Output and Screenshots

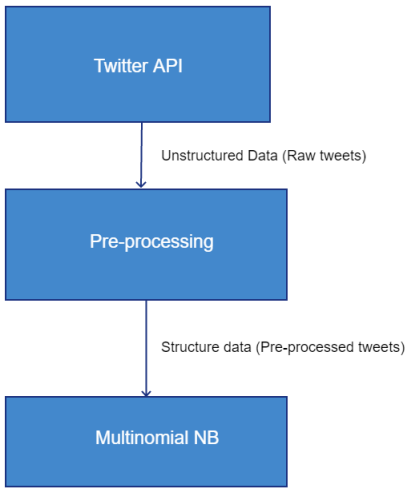

Fig. 1. Block Diagram

Fig. 2. Situational Tweet

TABLE I
VALIDATION WITH DIFFERENT N-GRAM RANGES

N-gram	Class	Precision	Recall	F1 Score	Accuracy
Unigram	Non-Situational	99%	96%	97%	96.57%
	Situational	99%	100%	99%	
Bigram	Non-Situational	100%	89%	94%	85.34%
	Situational	97%	100%	99%	
Trigram	Non-Situational	100%	76%	86%	81.32%
	Situational	95%	100%	97%	
Unigram + Bigram	Non-Situational	100%	96%	98%	94.87%
	Situational	99%	100%	100%	
Unigram + Bigram + Trigram	Non-Situational	100%	94%	97%	92.165%
	Situational	99%	100%	99%	

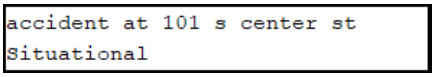

Fig. 3. Classified : Situational

Fig. 4. Non Situational Tweet

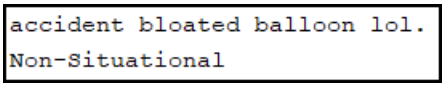

Fig. 5. Classified : Non-Situational

Fig. 6. An example of a False Positive Tweet

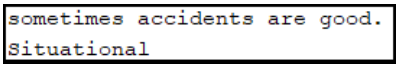

Fig. 7. An example of a False Positive(i.e.non-situational tweet classified as situational)

As we see in the above example, there are scenarios where non-situational tweets would be classified as situational.

VII. Conclusion

We have learnt and implemented the following techniques: data streaming techniques, feature extraction techniques and classification techniques. In this work, experiments were performed to find the best ngram model for analysis of accident related information from tweets using word level features features. Empirical evaluations are carried on the test set using multinomial naive bayes classifier in combination with different ngram level features. From the results it can be concluded that the unigram model is better than the other word based features. The unigram feature gave the best accuracy obtained as 96.57% for classification. As a part of future work, other machine learning algorithms with other types of features such as syntactic and semantic and a combination of both can be used. While Multinomial model has been popular in terms of text based nlp techniques, the SVM model has also proven to be strong and it will be tried out next.
By maintaining a database of locations, street addresses and the exact locations can be extracted from situational tweets. We can keep track of the number of accidents at a particular area. If the number goes higher than a threshold, that particular area would be termed accident prone. The government agencies as well as several NGO's will have a direct way to keep an eye on accident prone regions whereas the general public will know which route to avoid at a given time. Since the data is crowd-sourced, the entire responsibility of data is shared among users. The major benefits of this project include but not limited to: Low cost and crowd-sourced solution for a problem almost everyone faces in their lifetime, Saves human life and resources, Quick and real time response during times of emergency, No need of expensive hardware, sensors, trackers, cameras, etc.

References

[1] S. A. Georgakis Panagiotis, Petalas Yannis, "Incident detection using data from social media," in *IEEE 20th International Conference on Intelligent Transportation Systems*, 2017.
[2] W. J. C. Sudha Verma, Sarah Vieweg *et al.*, "Natural language processing to the rescue?: Extracting 'situational awareness' tweets during mass emergency," in *Proc. IEEE of the Fifth Int. AAAI Conf. on Weblogs and Social Media*, 2012.
[3] M. P. Brinda Hegde, Nagashree H S, "Sentiment analysis of twitter data: A machine learning approach to analyse demonetization tweets," in *International Research Journal of Engineering and Technology(IRJET)*, 2018.
[4] A. A. Sara Rosenthal and K. McKeown, "Sentiment detection of sentences and subjective phrases in social media," in *Proceedings of the 8th International Workshop on Semantic Evaluation*, 2014, pp. 198–202.
[5] Larson and Keri, "The impact of natural language processing based textual analysis of social media interactions on decision making," in *Proc. of the 21st European Conf. on Information Systems*, 2012.
[6] P. V. R. T. JHANSI RANI, K. ANURADHA, "Sentiment classification on twitter data using word n gram model," in *International Journal of Technology and Engineering Science*, 2016.
[7] J. Henrique, "Get old tweets programatically."
[8] J. H. M. Daniel Jurafsky, *Speech and Language Processing, 2nd Edition*, 2009.

TABLE II
VALIDATION WITH DIFFERENT N-GRAM RANGES

Class	Precision	Recall	F1 Score	Accuracy
Non-Situational	99%	96%	97%	96.57%
Situational	99%	100%	99%	