DISASTER MANAGEMENT WITH TWITTER USIG NLP

Sanket Kesharwani
Department of Networking and Communications
SRM Institute of Science And Technology
Chennai, India

Shlok Gupta
Department of Networking and Communications
SRM Institute of Science And Technology
Chennai, India

Dr. V.R Balasaraswati
Department of Networking and Communications
SRM Institute of Science And Technology
Chennai, India

I. ABSTRACT

Microblogging platforms such as Twitter provide active communication channels during mass convergence and emergency events such as earthquakes, typhoons. During the sudden onset of a crisis situation, affected people post useful information on Twitter that can be used for situational awareness and other humanitarian disaster response efforts, if processed timely and effectively. Processing social media information pose multiple challenges such as parsing noisy, brief and informal messages, learning information categories from the incoming stream of messages and classifying them into different classes among others. One of the basic necessities of many of these tasks is the availability of data, in particular human-annotated data. In this paper, we present human-annotated Twitter corpora collected during 19 different crises that took place between 2013 and 2015. To demonstrate the utility of the annotations, we train machine learning classifiers. Moreover, we publish first largest word2vec word embeddings trained on 52 million crisis-related tweets. To deal with tweets language issues, we present human-annotated normalized lexical resources for different lexical variations.

Keywords: - Natural Language Processing, Twitter, Disaster Response, Supervised Learning

II. INTRODUCTION

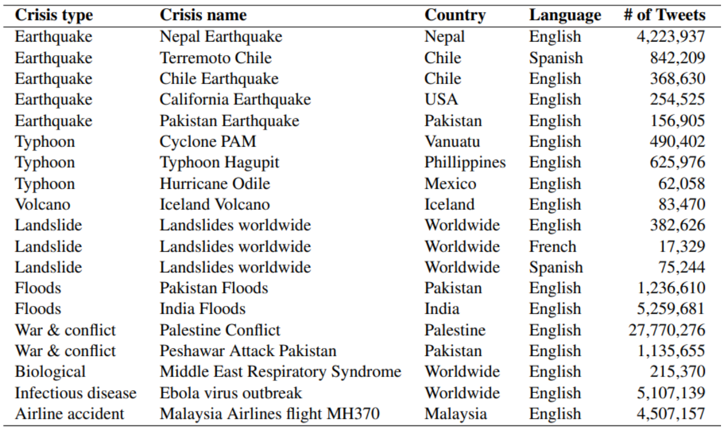
Twitter has been extensively used as an active communication channel, especially during mass convergence events such as natural disasters like earthquakes, floods, typhons. During the onset of a crisis, a variety of information is posted in real-time by affected people; by people who are in need of help (e.g., food, shelter, medical assistance, etc.) or by people who are willing to donate or offer volunteering services. Moreover, humanitarian and formal crisis response organizations such as government agencies, public health care NGOs, and military are tasked with responsibilities to save lives, reach people who need help, etc. Situation-sensitive requirements arise during such events and formal disaster response agencies look for actionable and tactical information in real-time to effectively estimate early damage assessment, and to launch relief efforts accordingly. Recent studies have shown the importance of social media messages to enhance situational awareness and also indicate that these messages contain significant actionable and tactical information (Cameron et al., 2012; Imran et al., 2013; Purohit et al., 2013). Many Natural-Language Processing (NLP) techniques such as automatic summarization, information classification, named-entity recognition, information extraction can be used to process such social media messages (Bontcheva et al., 2013; Imran et al., 2015). However, many social media messages are very brief, informal, and often contain slangs, typo graphical errors, abbreviations, and incorrect grammar (Han et al., 2013). These issues degrade the performance of many NLP techniques when used down the processing pipeline (Ritter et al., 2010; Foster et al., 2011). We provide human annotations (volunteers and crowd-sourced workers) of two types. First, the tweets are annotated with a set of categories such as displaced people, financial needs, infrastructure, etc. These annotation schemes were built using input taken from formal crisis response agencies such as United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA). Second, the tweets are annotated to identify (OOV) terms, such as slangs, places names, abbreviations, misspellings, etc. and their corrections and normalized forms. This dataset can form the basis for research in text classification for short messages and for research on normalizing informal language. Creating large corpora for training supervised machine learning models is hard because it requires time and money that may not be available. However, since our dataset was used for disaster relief efforts, volunteers were willing to annotate it; this work can now be leveraged to improve text classification and language processing tasks. Our work provides annotations for around 50,000 thousand messages, which is a significant corpus, that will enable research into applied machine learning and consequently benefit the disaster relief (and other) research communities. Our dataset has been collected from various countries and during various times of the year. This diversity would make it an interesting dataset that if used would be a foil to solutions that only work for specific language "dialects", e.g., American English and would fail or suffer from degradation of quality if applied to variations, such as Indian English. Our work shows that when a dataset is used for a real application, we could obtain larger number of annotations than otherwise. These can then be used to improve text processing as a byproduct. The annotated data is also used to train machine-learning classifiers. In this case, we use three well-known learning algorithms: Naive Bayes, Random Forest, and Support Vector Machines (SVM). We remark that these classifiers arXiv:1605.05894v2 [cs.CL] 31 May 2016 are useful for formal crisis response organizations as well as for the research community to build more effective computational methods (Pak and Paroubek, 2010; Imran et al., 2015) on top. We also train word2vec word embeddings from all 52 million messages and make them available to research community

III. DATA COLLECTION

Collected crisis-related messages from Twitter posted during 8 different crises that took place from 2017 to 2020. Fig 3.1 shows the list of crisis events along with their names, crisis type (e.g., earthquake, flood), countries where they took place, and the number of tweets each crisis contains. To collect these messages, we used AIDR (Artificial Intelligence for Disaster Response) platform [10]. AIDR is an open-source platform to collect and classify Twitter messages during the onset of a humanitarian crisis. AIDR has been used by UN OCHA during many major disasters such as Nepal Earthquake, Typhoon Hagupit. AIDR provides different convenient ways to collect messages from Twitter using the Twitter's streaming API. There are several data collection strategy. For example, collecting tweets that contain some keywords and are specifically from a particular geographical area/region/city (e.g., Chennai). The detailed data collection strategies used to collect the datasets shown in Fig 3.1 are included in each dataset folder

IV. DATA ANNOTATION

a) Injured or dead people: Reports of casualties and/or injured people due to the crisis
b) Missing, trapped, or found people: Reports and/or questions about missing or found people
c) Displaced people and evacuations: People who have relocated due to the crisis, even for a short time (includes evacuations)
d) Infrastructure and utilities damage: Reports of damaged buildings, roads, bridges, or utilities/services interrupted or restored
e) Donation needs or offers or volunteering services: Reports of urgent needs or donations of shelter and/or supplies such as food, water, clothing, money, medical supplies or blood; and volunteering services
f) Caution and advice: Reports of warnings issued or lifted, guidance and tips
g) Sympathy and emotional support: Prayers, thoughts, and emotional support
h) Other useful information: Other useful information that helps understand the situation Not related or irrelevant: Unrelated to the situation or irrelevant
i) Not related or irrelevant: Unrelated to the situation or irrelevant



Crisis type	Crisis name	Country	Language	# of Tweets
Earthquake	Nepal Earthquake	Nepal	English	4,223,937
Earthquake	Terremoto Chile	Chile	Spanish	842,209
Earthquake	Chile Earthquake	Chile	English	368,630
Earthquake	California Earthquake	USA	English	254,525
Earthquake	Pakistan Earthquake	Pakistan	English	156,905
Typhoon	Cyclone PAM	Vanuatu	English	490,402
Typhoon	Typhoon Hagupit	Phillippines	English	625,976
Typhoon	Hurricane Odile	Mexico	English	62,058
Volcano	Iceland Volcano	Iceland	English	83,470
Landslide	Landslides worldwide	Worldwide	English	382,626
Landslide	Landslides worldwide	Worldwide	French	17,329
Landslide	Landslides worldwide	Worldwide	Spanish	75,244
Floods	Pakistan Floods	Pakistan	English	1,236,610
Floods	India Floods	India	English	5,259,681
War & conflict	Palestine Conflict	Palestine	English	27,770,276
War & conflict	Peshawar Attack Pakistan	Pakistan	English	1,135,655
Biological	Middle East Respiratory Syndrome	Worldwide	English	215,370
Infectious disease	Ebola virus outbreak	Worldwide	English	5,107,139
Airline accident	Malaysia Airlines flight MH370	Malaysia	English	4,507,157

Fig 3.1 – Crises datasets details including crisis type,

name, language of messages, country, Number of tweets

V. CLASSIFICATION OF MESSAGES

To make sense of huge amounts of Twitter messages posted during crises, we consider a basic operation, that is, the automatic categorization of messages into the categories of interest. This is a multiclass categorization problem in which instances are categorized into one of several classes. Specifically, we aim at learning a predictor h : X → Y, where X is the set of messages and Y is a finite set of categories. For this purpose, we use three well-known learning algorithms i.e., Naive Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF).

5.1 Preprocessing and feature extraction

The following Preprocessing steps are performed. First, stop-words, URLs, and usermentions are removed from the Twitter messages. Previous studies found these two features outperform when used for similar tasks. Them stemming takes place, it is the process that chops off the ends of words in the hope of achieving goal correctly most of the time and often includes the removal of derivational affixes. Then finally lemmatization takes place. It usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base and dictionary form of a word.

VI. TWITTER TEXT NORMALIZATION

6.1. Language issues in Twitter messages

The quality—in terms of readability, grammar, sentence structure etc.—of Twitter messages vary significantly. Typically, Twitter messages are brief, informal, noisy, unstructured, and often contain misspellings and grammatical mistakes. Moreover, due to Twitter's 140 character limit restriction, Twitter users intentionally shorten words by using abbreviations, acronyms, slangs, and sometimes words without spaces. We divide these lexical variations into the following five categories:

a) Typos/misspellings: e.g. earthquak (earthquake), missin (missing), ovrcme (overcome)
b) Single-word abbreviation/slangs: e.g. pls (please), srsly (seriously), govt (government), msg (message)

c) Multi-word abbreviation/slangs: e.g. imo (in my opinion), I'm (I am), brb (be right back)

d) Phonetics substitutions: e.g. 2morrow (tomorrow), 4ever (forever), 4g8 (forget), w8 (wait)

e) Words without spaces: e.g. pray for Nepal (pray for Nepal), we help (we help), we need shelter (we need shelter)

VII. IDENTIFICATION OF OOV WORDS

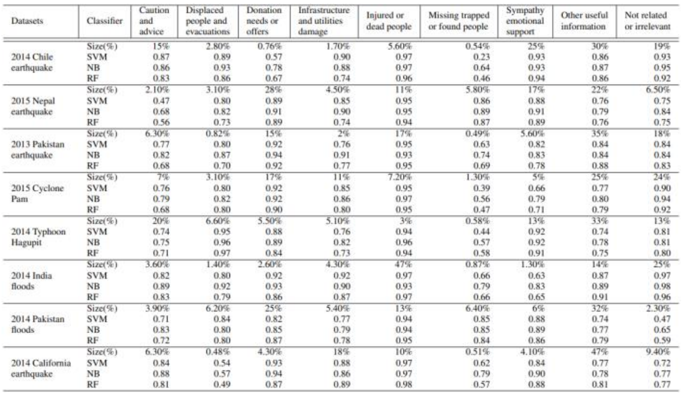
To identify candidate OOV words that require normalization, first build initial vocabularies consisting of lexical variations mentioned in the previous section. Then use a dictionary available on the web to normalize abbreviations, chat shortcuts, and slang. Also use the SCOWL (Spell Checker Oriented Word Lists) aspell English dictionary 6 that consists of 349,554 English words. The SCOWL dictionary is suitable for English spell checkers for most of English dialects. Although, the SCOWL dictionary contains places names (e.g., names of countries and famous cities), after testing it on Nepal Earthquake data, the problem with that is its coverage is not complete and a large number of cities/towns of Nepal are missing. To overcome this issue, use the MaxMind 7 world cities database that consists of 3,173,959 cities. Using the above resources, try to find OOV words in the dataset. However, Large number of OOVs consist of misspelled. Words for which a correct form can be obtained using one edit-distance change (i.e., by performing one insertion, deletion, or substitution operation). For this purpose, train a language model using lists of most frequent words from Wiktionary the British National Corpus and words in our SCOWL dictionary

VIII. NORMALIZATION OF OOV WORDS

To normalize the identified OOV words, use the CrowdFlower crowdsourcing platform. A crowdsourcing task in this case consists of a Twitter message that contains one or more OOV words and a set of instructions. The workers were asked to read the instructions and examples carefully before providing an answer. A worker reads the given message and provides a correct OOV tag (i.e., slang/abbreviation/acronym, a location name, an organization name, a misspelled word, or a person name). If an OOV is a misspelled word, the worker also provides its corrected form.

XI. RESULT

The current implementation is able to take disaster tweets from twitter, Classify the tweets with respect to displaced people, Donation Need, Infrastructure Damage, Displaced People, Sympathy Messages, Other Relevant and Irreverent Messages. Also, the system is able to extract useful data from tweets e.g., Location, Disaster Type. Used 3 different algorithms to compare accuracy between the classification model. Will provide two sets of annotations related to topic categorization of the tweets and tagging out-of-vocabulary words and their normalizations. These resources will help improve automatic natural language processing of crisis related messages and eventually be useful for humanitarian organizations.



Datasets	Classifier	Caution and advice	Displaced people and evacuations	Donation needs or offers	Infrastructure and utilities damage	Injured or dead people	Missing trapped or found people	Sympathy emotional support	Other useful information	Not related or irrelevant
2014 Chile earthquake	Size(%)	15%	2.80%	0.76%	1.70%	5.60%	0.54%	25%	30%	19%
	SVM	0.87	0.89	0.57	0.90	0.97	0.23	0.93	0.86	0.93
	NB	0.86	0.93	0.78	0.88	0.97	0.64	0.93	0.87	0.95
	RF	0.83	0.86	0.67	0.74	0.96	0.46	0.94	0.86	0.92
2015 Nepal earthquake	Size(%)	2.10%	3.10%	28%	4.50%	11%	5.80%	17%	22%	6.50%
	SVM	0.47	0.80	0.89	0.85	0.95	0.86	0.88	0.76	0.75
	NB	0.68	0.82	0.91	0.90	0.95	0.89	0.91	0.79	0.84
	RF	0.56	0.73	0.89	0.74	0.94	0.87	0.89	0.76	0.75
2013 Pakistan earthquake	Size(%)	6.30%	0.82%	15%	2%	17%	0.49%	5.60%	35%	18%
	SVM	0.77	0.80	0.92	0.76	0.95	0.63	0.82	0.84	0.84
	NB	0.82	0.87	0.94	0.91	0.93	0.74	0.83	0.84	0.84
	RF	0.68	0.70	0.92	0.77	0.95	0.69	0.78	0.88	0.83
2015 Cyclone Pam	Size(%)	7%	3.10%	17%	11%	7.20%	1.30%	5%	25%	24%
	SVM	0.76	0.80	0.92	0.85	0.95	0.39	0.66	0.77	0.90
	NB	0.79	0.82	0.92	0.86	0.97	0.56	0.79	0.80	0.94
	RF	0.68	0.80	0.90	0.80	0.95	0.47	0.71	0.79	0.92
2014 Typhoon Hagupit	Size(%)	20%	6.60%	5.50%	5.10%	3%	0.58%	13%	33%	13%
	SVM	0.74	0.95	0.88	0.76	0.94	0.44	0.92	0.74	0.81
	NB	0.75	0.96	0.89	0.82	0.96	0.57	0.92	0.78	0.81
	RF	0.71	0.97	0.84	0.73	0.94	0.58	0.91	0.75	0.80
2014 India floods	Size(%)	3.60%	1.40%	2.60%	4.30%	47%	0.87%	1.30%	14%	25%
	SVM	0.82	0.80	0.92	0.92	0.97	0.66	0.63	0.87	0.97
	NB	0.89	0.92	0.93	0.90	0.93	0.79	0.83	0.89	0.98
	RF	0.83	0.79	0.86	0.87	0.97	0.66	0.65	0.91	0.96
2014 Pakistan floods	Size(%)	3.90%	6.20%	25%	5.40%	13%	6.40%	6%	32%	2.30%
	SVM	0.71	0.84	0.82	0.77	0.94	0.85	0.88	0.74	0.47
	NB	0.83	0.80	0.85	0.79	0.94	0.85	0.89	0.77	0.65
	RF	0.72	0.80	0.87	0.78	0.95	0.84	0.86	0.79	0.59
2014 California earthquake	Size(%)	6.30%	0.48%	4.30%	18%	10%	0.51%	4.10%	47%	9.40%
	SVM	0.84	0.54	0.93	0.88	0.97	0.62	0.84	0.77	0.72
	NB	0.88	0.57	0.94	0.86	0.97	0.79	0.90	0.78	0.77
	RF	0.81	0.49	0.87	0.89	0.98	0.57	0.88	0.81	0.77

Fig 9.1 Tweets Classification Using Different Algorithms

IX. REFERENCES

1. *Meera.R.Nair1, G.R.Ramya1, P.Bagavathi Sivakumar1 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India*

2. *Nurdeni, Deden Ade, Budi, Indra, Yunita, Ariana Journal of Management Information & Decision Sciences*
3. https://www.kaggle.com/competitions/nlp-getting-started
4. *Muhammad Imran , Prasenjit Mitra , Carlos Castillo (Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Message) Conference of Information and Knowledge Management (CIKM)-MAY 2016*
5. *Zahra Ashktora, Christopher Brown, Manojit Nand, Carnegie Mellon (Mining Twitter To Inform Disaster Response) 11th International ISCRAM Conference-2007 MAY 2007*
6. *Hathairat Ketmaneechairat and Maleerat Maliyaem (Natural Language Processing For Disaster Management Using Conditional Random Fields) Journal of Advances in Information Technology-2020 AUG 2020*
7. *Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, Hassan Sajjad, Prasenjit Mitra (Applications of Online Deep Learning for Crisis Response Using Social Media Information) Conference of Information and Knowledge Management (CIKM)-2016, OCT 2016*
8. *nlp.standford.edu*
9. *Jayashree Domala, Manmohan Dogra, Vinit Masrani, Dwayne Fernandes, Kevin D'souza, Delicia Fernandes, Tejal Carvalho (Automated Identification Of Disaster News For Crisis Management Using Machine Learning And Natural Language Processing) IEEE JAN 202*
10. *Robin Peters, de Albuquerque (Investigating Images As Indicators For Relevant Social Media Messages In Disaster Management) ISCRAM Conference JAN 2015*
11. *Khondhaker Al Momin, H M Imran Kays, Arif Mohaimin Sadri (Identifying Crisis Response Communities In Online Social Networks For Compound Disasters: The Case Of Hurricane Laura And Covid-19) Conference of Information and Knowledge Management (CIKM)-MAY 2021*
12. *Saptarshi Ghosh, Kripabandhu Ghosh, Debasis Ganguly, Tanmoy Chakraborty, & Gareth J. (Exploitation Of Social Media For Emergency Relief And Preparedness: Recent Research And Trends) SPRINGER-2018 AUG 2018*
13. *Cameron, M. A., Power, R., Robinson, B., and Yin, J. (2018). Emergency situation awareness from twitter for crisis management. In Proc. of the 21st international conference companion on World Wide Web, pages 695– 698.*
14. *Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., and Aswani, N. (2016). Twitie: An open-source information extraction pipeline for microblog text. In RANLP, pages 83–90*
15. *Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR), 47(4):67*
16. *Han, B., Cook, P., and Baldwin, T. (2019). Lexical normalization for social media text. ACM Transactions on Intelligent Systems and Technology (TIST), 4(1):5*
17. AIRD Documentation