IIT-BHU In TREC 2018 Incidents Stream Track

Harshit Mehrotra
Department Of Computer Science and Engineering
IIT (BHU) Varanasi
India
harshit.mehrotra.cse15@iitbhu.ac.in

Dr. Sukomal Pal
Department Of Computer Science and Engineering
IIT (BHU) Varanasi
India
spal.cse@iitbhu.ac.in

ABSTRACT

This paper presents details of the work done by the team of IIT (BHU) Varanasi for the Incidents Stream track in TREC 2018. The task involved classifying tweets posted during a disaster into a number of categories, which are useful for relief work purposes at such a time. The data given was in the form of tweets from one earthquake, tornado, wildfire, flood, shooting and bombing incident.

KEYWORDS

Information retrieval, tweet classification

ACM Reference Format:
Harshit Mehrotra and Dr. Sukomal Pal. 2018. IIT-BHU In TREC 2018 Incidents Stream Track. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 TASK AND DATA

As social media has increased in popularity and reach, its uses have gone far and beyond just staying in touch with friends and acquaintances. Recently, people have been taking to social media platforms like Twitter and Facebook to convey useful and critical information during times of natural disaster. This information in the form of first person accounts, news updates, need and availability of resources is very useful if brought to the notice of the concerned authorities for relief work. The incidents stream track focussed on curating such categorized feeds from tweets. Given 25 categories related to such focus work, we had to categorize tweets from earthquakes, floods, tornados, wildfires, bombing and shooting incidents. Annotated tweets data from one incident of each type was given.

2 METHODOLOGY

The given data provides with every tweet ID, its belonging category, and terms that indicate the relationship. Firstly all train and test tweets are pre-processed by lower-casing, stopword removal and stemming (using Porter stemmer). A category wise frequency distribution is made from the indicator terms as well as the actual content of the tweet. This gives us for all the 6 types of incidents frequently occurring terms for each category. The problem arises with categories which do not have training tweets in any kind of incident. In these cases, frequently used words are obtained from experience of other similar shared tasks involving natural disaster related tweets. Now these frequently appearing words / phrases are used to create queries which are then matched with the test tweet. The matching is carried out using the Lucene library. The category whose query gives the maximum score when matched with a test tweet is attributed to that particular tweet.

The queries used for some categories of two types of incidents are given below as an example:

Flood:
Report-Weather : rain OR pour OR storm OR forecast
Report-EmergingThreats : clos OR collaps OR submerg OR damag OR destroy OR leak
Report-ServiceAvailable : team OR aid OR send OR sent OR provid OR hospit OR avail
Report-Factoid : unaccount OR kill OR miss OR dead OR (worth AND damag) OR injur OR injuri

Tornado:
Report-Weather : rain OR wind OR forecast OR landfal OR heavi OR strong
Report-EmergingThreats : landslid OR trigger
Report-ServiceAvailable : team OR aid OR send OR sent OR provid OR hospit OR avail
Report-Factoid : (eye AND locat) OR dead OR injur OR injuri

As it can be seen above, there are some categories for which queries can be generic and some for which they will be incident specific. The former observation has been used to create queries for categories which have no or very few training tweets from some incident.

3 POSSIBLE IMPROVEMENTS

Query expansion can be used to some extent in order to incorporate synonyms and similar words if there is evident support from sufficient data, which was absent in this case. Nevertheless, it can be considered as an augmentation to the methodology.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Conference'17, July 2017, Washington, DC, USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn