Social Media Incident Streams

A Text Retrieval Conference (TREC) Challenge

Ian Soboroff, ITL, NIST

Any mention of companies or services in this talk is for information only; it does not imply recommendation or endorsement by NIST.

#PSCR2020
Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately.

Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

* Please note, unless mentioned in reference to a NIST Publication, all information and data presented is preliminary/in-progress and subject to change.
RT @RobDavis_Wx: LANDSLIDE!... road blocked in Costa Rica after M7.6 earthquake twitpic.com/are653 (courtesy: @skasoul) #temblorcr
11:53 AM - Sep 5, 2012 - Twitter Web Client

JR Santos @jr0403
Replying to @DickGordonDG
@ChairmanGordon @philredcross sir until now no rescue at the place i tweeted earlier at mercedes homes3 biñan laguna. Thank you! #RescuePH
2:48 AM - Aug 19, 2013 - Twitter for iPhone

Philippine Red Cross @philredcross - Aug 19, 2013
Replying to @jr0403 @jr0403 @ChairmanGordon may WASAR teams na po kami nakadeploy
1 1 1

JR Santos @jr0403 - Aug 19, 2013
@philredcross @ChairmanGordon thank you po. Nagmamaseg kasi sila ulit kanina.

Mike Allen @MikeAllen_YMM
#ymmfire has breached the hill and coming down towards #hwy63 and Grayling Terrace #ymm #

ALERT LEVEL 4, Marikina River at 18meters. Marikina residents near the river, you guys need to evacuate NOW. #marinangPH #floodPH

RT @dude_funk: #rescuePH Dela Paz, Biñan (?) Laguna. Newly kidney operated Tita currently on rooftop with other family due to flood. Please ...
10:52 PM - Aug 18, 2013 - Twitter for iPhone

(actual tweets from TREC-IS incident collections)
Problem

• People take to Twitter during crises, but no one can monitor it.
  • Hashtags and keywords have high volume, spam, irrelevant information.
  • Flood of hopes and prayers.
• Can computer systems find critical, actionable tweets in this mess?
• Can they get them to the right people in time?
AI systems can be built to filter out noise, identify critical tweets, prioritize them, and route them to the right people. But AI depends on high quality training data.
Text Retrieval Conference

trec.nist.gov

• TREC is an evaluation workshop series started by NIST in 1992.
• TREC features a set of tracks that pose data challenges around different problems in search, information filtering, and information access.
• Each track creates a dataset for the open participant community: universities and industry research labs who sign up to attempt the challenge.
• The community participation process is leveraged to create ground truth and simultaneously measure the effectiveness of participant solutions to the track challenge.
The TREC data revitalized research on information retrieval. Having a standard, widely available, and carefully constructed set of data laid the groundwork for further innovation in the field. The yearly TREC conference fostered collaboration, innovation, and a measured dose of competition (and bragging rights) that led to better information retrieval.

Hal Varian  
Google Chief Economist  
March 4, 2008

This project [the TREC Legal track] can be expected to identify both cost effective and reliable search and information retrieval methodologies and best practice recommendations, which, if adhered to, certainly would support an argument that the party employing them performed a reasonable ESI search, whether for privilege review or other purposes.

Magistrate Judge Paul Grimm  
Victor Stanley v. Creative Pipe

TREC is an annual benchmarking exercise that has become a de facto standard in Information Retrieval evaluation.

Stephen Robertson  
Microsoft  
SIGIR 2007

TREC has proven to be a valuable forum in which IBM Research has contributed to an improved understanding of search, while at the same time the insights obtained by participating in TREC have helped to improve IBM’s products and services.

Alan Marwick, et al.  
IBM chapter of the TREC book  
2005

In other words, for every $1 NIST and its partners invested in TREC, at least $3.35 to $5.07 in benefits accrued to IR researchers…These responses suggest that the benefits of TREC to both private and academic organizations go well beyond those quantified by this study’s economic benefits.

RTI International  
Economic Impact Assessment of NIST’s TREC Program  
December 2010
TREC Incident Streams track

• Started in 2018.
• Provides 33 Twitter datasets from earthquake, wildfire, hurricane, flood, bomb and shooting events.
• Each tweet is labeled to indicate:
  • **Relevance**: Does it contain actionable information?
  • **Categories**: What kind of information does it contain?
  • **Criticality**: How important is it that public safety should see this tweet?
25 categories in 4 major groups:

- Immediate needs
- or, Useful as metadata
- or, Useful post-event
- or, Useful for research

Derived from existing taxonomies as well as research surveys of social media use during emergencies.
A variant of the task focuses on a subset of 12 categories.

For this task, all the “Other” categories are collapsed.

These categories are those most likely to be the most useful to public safety personnel during an emergency situation.
• After assigning all pertinent categories, the assessor indicates the tweets **criticality**.

• Does this tweet need immediate attention from emergency personnel, or can it wait?

<table>
<thead>
<tr>
<th>Critical (Notify immediately)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>High (Should be viewed by officer)</td>
<td></td>
</tr>
<tr>
<td>Medium (Can be viewed later)</td>
<td></td>
</tr>
<tr>
<td>Low (Can be safely ignored)</td>
<td></td>
</tr>
</tbody>
</table>
RQ1.3 – What Makes a Tweet Actionable?

Having shown that there is consistently valuable information available on social media during emergencies, the natural question is: what makes a tweet actionable, or at least valuable for a responder to see? This is important for system builders, such that they have some idea what factors their systems should be considering. Our human assessors provide both information type labels and priority labels for each tweet in the different events. However, this does not tell us why these labels were assigned or what information was used to make that determination. As such, we require additional information to answer the above question.

To solve this, we perform a smaller-scale labelling study aimed at identifying common factors that indicate a tweet is actionable. To do so, we select the 170 tweets that were labeled as ‘Critical’ priority by our assessors in 2019-B - forming a set of tweets that we can be confident are valuable. Next, we rendered each tweet using the same assessment interface as the original assessor. Each tweet was viewed in turn, and a new assessor identified features of the tweet that appeared to them to contribute to its action-ability. Note, this new assessor was a computer scientist, and experienced in the construction of automatic systems for this task (i.e. was a participant in previous TREC-IS editions). Hence, the way that they categorized information shown was reflective of factors that they believed an automatic system should consider when categorizing each tweet. After all tweets had been analysed, we aggregated the outcome into three high level information sources (Tweet Text, Linked Content and Author) with a total of 10 sub-categories. These are shown in the top part of Table 4 below.

The right-hand columns of Table 4 report the number of critical tweets (and the proportion) where each type of information was important when determining whether those tweets were actionable. First, as we might expect, the CoRe Paper – Social Media for Disaster Response and Resilience Proceedings of the 17th ISCRAM Conference – Blacksburg, VA, USA May 2020 Amanda Lee Hughes, Fiona McNeill and Christopher Zobel, eds.

![Criticality Distribution](image1.png)  
(a) Criticality Distribution  
![Average Criticality by Information Type](image2.png)  
(b) Average Criticality by Information Type (95% Confidence)
SEVERE TROPICAL STORM TRAMI, KNOWN IN THE PHILIPPINES AS TROPICAL STORM MARING, WAS A TROPICAL CYCLONE THAT BROUGHT HEAVY RAINS TO TAIWAN AND EAST CHINA DURING MID-AUGUST 2013. THE USER IS A RESPONSE OFFICER RESPONSIBLE FOR METRO MANILA, ONE OF THE THREE DEFINED METROPOLITAN AREAS OF THE PHILIPPINES. WIKIPEDIA PAGE
Call for participation

Task definition

Data procurement

Ground truth labeling

Participant experiments

Ground truth assessment

Results evaluation

Results analysis

Yearly TREC Cycle

Currently we crawl Twitter, but next year we will be using their Enterprise API services

We can label a small number of tweets here …

... but we can label many more relevant tweets based on participant outputs

November 18-20, 2020

trec.nist.gov/pubs.html
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>nyu.fast.multi</td>
<td>0.04</td>
<td>0.01 0.05 0.05</td>
<td>0.05 0.16 0.09 0.26</td>
<td>0.42 0.38 0.22</td>
<td>0.08 0.54 0.22</td>
<td>0.05 0.25 0.05</td>
<td>0.55 0.37 0.05</td>
<td>0.08 0.31</td>
<td>0.91 0.13</td>
<td>0.03 0.11</td>
<td>0.03 0.11</td>
<td>0.02 0.08</td>
<td>0.08 0.11</td>
<td>0.02 0.08</td>
</tr>
<tr>
<td>nyu.base.multi</td>
<td>0.05</td>
<td>0.11 0.12 0.12</td>
<td>0.16 0.09 0.03 0.34</td>
<td>0.01 0.15 0.13</td>
<td>0.25 0.18 0.03</td>
<td>0.15 0.3 0.05</td>
<td>0.45 0.13 0.03</td>
<td>0.02 0.06</td>
<td>0.00 0.18</td>
<td>0.18 0.27</td>
<td>0.26 0.26</td>
<td>0.03 0.32</td>
<td>0.08 0.18</td>
<td>0.02 0.08</td>
</tr>
<tr>
<td>nyu.fast.sing</td>
<td>0.04</td>
<td>0.04 0.07 0.07</td>
<td>0.19 0.01 0.09 0.08</td>
<td>0.03 0.17 0.11</td>
<td>0.43 0.33 0.15</td>
<td>0.01 0.01 0.23</td>
<td>0.25 0.18 0.03</td>
<td>0.06 0.14</td>
<td>0.00 0.18</td>
<td>0.00 0.18</td>
<td>0.00 0.18</td>
<td>0.00 0.18</td>
<td>0.00 0.18</td>
<td>0.00 0.18</td>
</tr>
<tr>
<td>run1_baseline</td>
<td>0.02</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.03 0.03 0.06</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
</tr>
<tr>
<td>run2_negative</td>
<td>0.02</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
</tr>
<tr>
<td>run3.inn</td>
<td>0.02</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
</tr>
<tr>
<td>run4_all</td>
<td>0.02</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
</tr>
<tr>
<td>UCDbaseline</td>
<td>0.08</td>
<td>0.01 0.04 0.06 0.06</td>
<td>0.17 0.09 0.15 0.31</td>
<td>0.58 0.22 0.05</td>
<td>0.48 0.31 0.65</td>
<td>0.22 0.06 0.53</td>
<td>0.35 0.06 0.28</td>
<td>0.01 0.1</td>
<td>0.00 0.16</td>
<td>0.00 0.16</td>
<td>0.00 0.16</td>
<td>0.00 0.16</td>
<td>0.00 0.16</td>
<td>0.00 0.16</td>
</tr>
<tr>
<td>UCDcncelmo</td>
<td>0.03</td>
<td>0.06 0.03 0.03</td>
<td>0.26 0.13 0.17 0.27</td>
<td>0.3 0.13 0.19 0.26</td>
<td>0.26 0.25 0.12</td>
<td>0.01 0.21 0.06</td>
<td>0.09 0.01 0.21</td>
<td>0.06 0.01 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
</tr>
<tr>
<td>UCDblstmlalpha</td>
<td>0.08</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
</tr>
<tr>
<td>UCDblstmbeta</td>
<td>0.09</td>
<td>0.07 0.18 0.18</td>
<td>0.18 0.09 0.04 0.37</td>
<td>0.11 0.25 0.04</td>
<td>0.12 0.27 0.35</td>
<td>0.15 0.03 0.15</td>
<td>0.15 0.26 0.05</td>
<td>0.04 0.34 0.08</td>
<td>0.01 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
</tr>
<tr>
<td>UPB-BERT</td>
<td>0.08</td>
<td>0.02 0.22 0.05 0.05</td>
<td>0.24 0.14 0.17 0.44</td>
<td>0.01 0.24 0.09</td>
<td>0.48 0.28 0.5</td>
<td>0.24 0.15 0.41</td>
<td>0.06 0.35 0.01</td>
<td>0.08 0.14 0.08</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
</tr>
<tr>
<td>UPB-FOCAL</td>
<td>0.14</td>
<td>0.03 0.13 0.09 0.09</td>
<td>0.36 0.13 0.19 0.45</td>
<td>0.01 0.16 0.09</td>
<td>0.46 0.28 0.59</td>
<td>0.19 0.16 0.38</td>
<td>0.33 0.04 0.31</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
</tr>
</tbody>
</table>

F1 scores for various models and datasets.
Comparison of approaches

(a) Learning Paradigm Groups

(b) Text Featurization Groups

CONCLUSIONS

In this paper we have provided an overview of the new 2019 editions (2019-A and 2019-B) of the TREC Incident Streams (TREC-IS). TREC-IS is a standardization initiative that develops test collections and evaluation methodologies for automatic and semi-automatic filtering approaches that identify and categorize information and aid-requests made on social media during crisis situations. It also incorporates re-occurring data challenges in which researchers/developers can participate, enabling comparison of state-of-the-art systems for the task. Indeed, over two years and three editions, TREC-IS has manually annotated tweet streams for 33 emergency events, comprising 35,000 tweets and producing over 125,000 labels.

Furthermore, this paper provides analysis of both the labeled tweets and participating systems to TREC-IS in 2019. It provides insights into both what information is actionable and critical for crisis responders, as well as what automated techniques perform well in identifying high-priority, actionable information during times of crisis. From this analysis, we have shown that high priority information on social media tends to be either calls for aid, warnings about new sub-events or threats, evacuation information and reports of services coming back online, which is consistent with analysis of TREC-IS 2018. Furthermore, we showed that overall the volume of high or critical
Outcomes

- 33 emergency events collected and annotated.
- 11 teams participated in the 2018 challenge, 10 teams in 2019, expecting about the same for 2020.
- Papers about the effort published in ISCRAM 2019 and 2020.
- All event datasets freely available from trecis.org.
- Papers by teams at trec.nist.gov under Publications → Proceedings.
THANK YOU