Siena's Incident Stream System
SISS

Ting Liu, Sharon Gower Small, Patrick Baumgardner, Lydia Cartwright,
Michael Coppola and Samuil Orlioglu

Siena College Institute for Artificial Intelligence
515 Loudon Road
Loudonville, NY 12211
tliu, ssmall, pj24baum, l23cart, mj04copp, s28orli
@siena.edu

Abstract

This paper discusses our work and participation in the Text Retrieval Conference (TREC) Incident Streams track (IS) of 2021. The mass adoption of mobile internet-enabled devices paired with wide-spread use of social media platforms for communication and coordination has created new ways for the public on-the-ground to contact response services. With the rise of social media, emergency service operators are now expected to monitor those channels and answer questions from the public. However, they do not have adequate tools or manpower to effectively monitor social media, due to the large volume of information posted on these platforms and the need to categorize, cross-reference and verify that information. The TREC Incident Streams (TREC-IS) track aims to provide a base for research to tackle this technology gap. TREC-IS was designed to bring together academia and industry to research technologies to automatically process social media streams during emergency situations with the aim of categorizing information and aid requests made on social media for emergency service operators. Given a corpus of tweets for a set of emergency events, participants are required to categorize each tweet into its information type, i.e. *Request-Search & Rescue, Report-Weather, CallToAction-Volunteer*, etc. and its level of criticality. This paper discusses our work and submission to TREC-IS 2021.

1. Introduction

The Incident Streams Track (Buntain et al., 2020) first run in 2018 is a program in the Text Retrieval Conference (TREC) (Voorhees 2007). TREC is a program co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense and it focuses on supporting research in information retrieval and extraction, and to increase availability of appropriate evaluation techniques. TREC-IS was designed to bring together academia and industry to research technologies to automatically process social media streams during emergency situations with the aim of categorizing information and aid requests made on social media for emergency service operators. We had a team of four undergraduate researchers work for 6 weeks to generate our submissions and explore other ideas that we believed could potentially boost performance for this type of task. We will discuss two aspects of our work: 1) our work with the BERT language model and 2) our evaluation on the importance of various tweet features relative to the priority of the tweet i.e., the time of the tweet vs. time of the event, location of the tweet vs. location of the event, tweet length, etc.

2. Incident Streams Literature Review

Many techniques to create a system that can effectively and efficiently label information types and priority levels have been tried. The Terroir team at the University of Glasgow (Hepburn et al., 2020) used text-based features, examining what distinguishes tweets between priority levels, as well as numerical features including the number of hashtags and the presence of URLs and other media. The team trained on Balanced Random Forest (BRF) and Easy Ensemble (EE) models and found that BRF models performed higher than EE. While the BRF model did well in identifying information types, it did not score high when predicting priority levels. A team at University College Dublin (Wang and Lillis, 2020) worked with multi-task transfer learning, fine-tuning transformer encoder-based models like BERT and sequence-to-sequence transformers like T5. They had two scenarios, one where they used an encoder model and one where they used a sequence-to-sequence model. For each of the scenarios, they trained two prediction models, one for predicting post category, and one for predicting post priority. Their work outperformed other runs in information type classification and in predicting priority levels.

One of our hypotheses was that solutions needed to be event specific to take system up a level in their performance. A system would then just need to identify the event type from the tweet and then apply the appropriate event specific solution in order to predict the information type and priority level. Work has been done in first story detection of events through Twitter and detection of single events. Wang and Goutte work with clustering temporal profiles of hashtags to then input into multivariate change point detection algorithms to find changes in events in Twitter streams (Wang and Goutte., 2020). Their method outperforms others in that it identifies up to 40% of subevents in the datasets tested. A team at University of Edinburgh (Wurzer et al., 2020) used k-term hashing for first story detection of events that operates O(1) per tweet. Rather than comparing a tweet with each that came before it, it can be compared with just one model that combines all previous tweets to greatly increase efficiency. Studies from Radboud University explore the use of estimating future events based on tweet text (Hürriyetoğlu et al., 2014). They parsed tweets for keywords and their variations and predicted the time of the event. Their systems typically had a margin of error of less than ten hours.

3. Track Overview

The goal of TREC-IS was to support emergency response services' efforts to harness the information in social media to respond better to social crisis situations. Participants were provided with a stream of filtered event relevant tweets and an ontology of information types, Table 1 below.

High-Level Information Type	Description	Example Low Level Types
Request-GoodsServices*†	The user is asking for a particular service or physical good.	PsychiatricNeed, Equipment, ShelterNeeded
Request-SearchAndRescue*†‡	The user is requesting a rescue (for themselves or others)	SelfRescue, OtherRescue
Request-InformationWanted†‡	The user is requesting information	PersonsNews, MissingPersons, EventStatus
CallToAction-Volunteer†‡	The user is asking people to volunteer to help the response effort	RegisterNow
CallToAction-Donations	The user is asking people to donate goods/money	DonateMoney, DonateGoods
CallToAction-MovePeople*†‡	The user is asking people to leave an area or go to another area	EvacuateNow, GatherAt
Report-FirstPartyObservation†	The user is giving an eye-witness account	CollapsedStructure, PeopleEvacuating
Report-ThirdPartyObservation	The user is reporting a information from someone else	CollapsedStructure, PeopleEvacuating
Report-Weather	The user is providing a weather report (current or forecast)	Current, Forecast
Report-EmergingThreats*†‡	The user is reporting a potential problem that may cause future loss of life or damage	BuildingsAtRisk, PowerOutage, Looting
Report-MultimediaShare†	The user is sharing images or video	Video, Images, Map
Report-ServiceAvailable*†‡	The user is reporting that someone is providing a service	HospitalOperating, ShelterOffered
Report-Factoid	The user is relating some facts, typically numerical	LandDevastated, InjuriesCount, KilledCount
Report-Official	An official report by a government or public safety representative	OfficialStatement, RegionalWarning, PublicAlert
Report-CleanUp	A report of the clean up after the event	CleanUpAction
Report-Hashtags	Reporting which hashtags correspond to each event	SuggestHashtags
Report-News	The post providing/linking to continuous coverage of the event	NewsHeadline, SelfPromotion
Report-NewSubEvent*†‡	The user is reporting a new occurrence that public safety officers need to respond to.	PeopleTrapped, UnexplodedBombFound
Report-Location†	The post contains information about the user or observation location.	Locations, GPS coordinates
Other-Advice‡	The author is providing some advice to the public	SuggestBestPractices, CallHotline
Other-Sentiment	The post is expressing some sentiment about the event	Sadness, Hope, Wellwishing
Other-Discussion	Users are discussing the event	Causes, Blame, Rumors
Other-Irrelevant	The post is irrelevant, contains no information	Irrelevant
Other-ContextualInformation	The post is generic news, e.g. reporting that the event occurred	NewsHeadline
Other-OriginalEvent	The Responder already knows this information	KnownAlready

* – "Actionable" Information Types, † – Task-2 Information Types, ‡ – COVID-19 Information Types (2020-A only)

Table 1. Ontology High-level Information Types

Participants were required to return an information type label and priority rating for all tweets. TREC supplied us with a labeled set of 91,515 tweets covering 71 different topics that happened between 2011 and 2020. These topics consist of 12 different types of events. The 2021 test collection consisted of 1,532,359 tweets. We will first discuss our work using the BERT language model to predict information types of tweets. We will also discuss our evaluation on the importance of various tweet features relative to the priority of the tweet. One approach analyzed if there was a correlation between the time the tweet was posted versus the time of the event and the priority level of the tweet. A second approach we explored was if the geographic distance between the event location and the location the tweet was posted from affected the priority level of the tweet. Additionally, we evaluated the same for length, the retweet count, the favorite count, the hashtag count, and whether the tweet is a reply and/or contains a mention.

4. Our Approach

Three members (Liu, Orlioglu and Coppola) of our team focused on language models while the other three (Small, Baumgardner and Cartwright) worked on the features relative to priority analysis. In this section we will first discuss our work using the BERT language model to predict information types and tweets.

4.1 BERT Model implementation

Introduced by Google in 2017, Bidirectional Encoder Representations from Transformers (BERT) has achieved the State-of-the-Art performance in many different areas in natural language process. Wang and Lillis's (2020) system also demonstrated that BERT outperformed other machine learning methods in this task. Therefore, we decided to build our system on top of BERT.

Similar to Wang and Lillis's (2020) work, we created a classifier model that maps a list of classification tokens ([CLSs]) to the certain information types. To predict the priority of an event, we built another model that transforms the output into four priority levels with the distribution calculated through sigmoid function to pinpoint the priority level.

We employed ktrain[1], a lightweight wrapper for the deep learning libraries, like TensorFlow Keras, for our system development because it provides many different options for classification, so that we can try different machine learning techniques to compare.

We used the annotation data, which contains over 91K annotated tweets, from 2018 to 2020B corpus. We only focused on the raw tweet content in this version because of 1. time limit and 2. to create a baseline. From the annotated corpus in json format, we pulled the tweets' content with human annotated information types and priority and converted them into a csv format file, which was utilized to train our two models. The first bar in the two charts in Figure 1 shows the system performance by just using the raw text of annotated tweets.

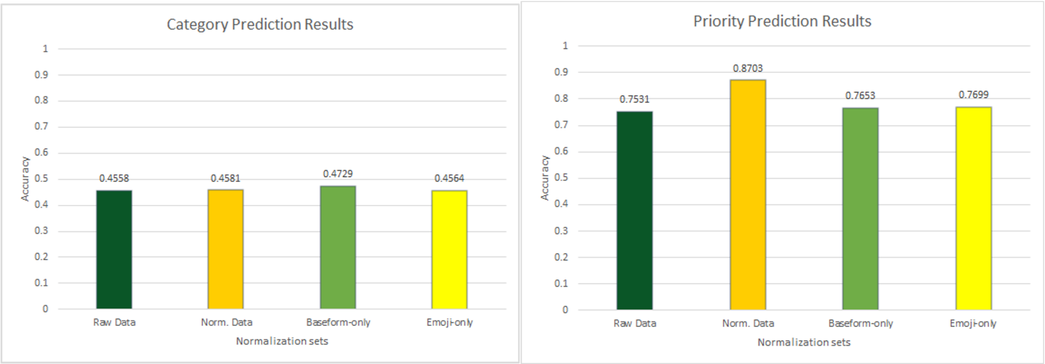

Figure 1. System performance on training set

Run	nDCG@100	Info-Type F1 [Actionable]	Info-Type F1 [All]	Info-Type Accuracy	Priority F1 [Actionable]	Priority F1 [All]	Priority R [Actionable]	Priority R [All]
F	0.4285	0.1861	0.3108	0.8565	0.0864	0.1609	0.0231	0.1099

Table 2. System performance on 2021A data

We only participated in the first task of this year. Table 2 shows the performance of our system running on 2021A corpus. There are still a lot of potentials for our system to make improvements. During the summer, three members of our team worked on the tweet text and the goal is to see if preprocessing of the tweet text can improve the system performance.

The first change is to clean the text and remove the part that won't contribute to the machine learning process. For example, the URL link at the end of each tweet was removed

[1] https://github.com/amaiya/ktrain

during this process. The second improvement was to convert the words[2] in tweet text into their base form. After all words variants are replaced, we hope that the training process can find more anchors and exist anchors will be more emphasized and therefore, the system performance can be boosted.

The third text normalization focuses on the emoji in the text. People usually add emoji in the text to express their emotions, which could be good indicators. However, the emoji in the text is usually treated as punctuations and filtered out. Therefore, we employed two emoji tools[3] to map the emoji into the corresponding words.

To test whether the proposed changes can improve the system performance, we created three new datasets,

1. One dataset with cleaning process and base form conversion (base form dataset)
2. One dataset with cleaning process and emoji conversion (emoji dataset)
3. One dataset with all changes (normalized dataset)

Bar 3 in Figure 1 is the system performance using base form dataset and bar 4 shows the performance using emoji dataset. Individually, both processes contribute some improvement to both information type and priority classification. Additionally, using the data pre-processed by both normalization methods boosted the performance (Bar 2 in Figure 1) of priority classification significantly but didn't show much help to improve the performance of information type classification. We're still working to find the reason why this happens and will report it in the official version of the paper.

4.2 Time & Location Analysis

Three members of our team decided to focus on determining if there are features present in the tweets that could be utilized by systems relative to determining the priority of a tweet. We hypothesized that each event type would show different patterns for some of the tweet features. For example, we expected tweets pertaining to ongoing events like floods and earthquakes to exist long after the start time of the event, as those events have longer lasting effects that response personnel should know about. Meanwhile, officers usually fully respond to a shooting shortly after it's over and therefore tweets beyond a certain time threshold would have a lower priority.

Tweet objects have many different components to them. We started by looking at these components and examining the data provided by the IS track, searching for attributes that might be relevant to the priority given to the tweet, and checking for any noticeable patterns. Specifically, we looked at the location of the tweet, the retweet count, the favorite count, the hashtag count, and whether the tweet is a reply and/or contains a mention, length of the tweet, the time/date, and their connection to the priority of the tweet.

In mining data for all of these fields we balanced our data set so that we were not overtraining to the highly abundant low priority set. We understand the controversy of this practice, as tweets during actual events would not be balanced in the usual fashion, but in order to validate our theory as best we could we needed to not overtrain on irrelevant

[2] www.nltk.org/.

[3] http://pypi.org/project/demoji/ and http://emoticonr.com/

tweets. In addition to running machine learning tests with balanced sets across priority fields, we also ran tests staggered upper bounds for each priority, with low priority having the highest upper bound. In this way, there are still more low priority tweets in the tests than any other priority, as well as less critical priority tweets, simulating the actual environment, to an extent.

In our tests, we focused on identifying critical priority tweets: the most urgent level assigned in the data. We used several classifiers to find out what had any effect on the data, including SVM, random forest, gradient boosting, k-nearest neighbors, neural network, linear regression, and logistic regression.

Given that our work is ultimately for event specific solutions, we also worked to find a way to make a system that could sort tweets by their event type so they could then work with the approach that would give the highest results. Here we detail our methods for accomplishing this task and include information on our approach in searching for relevant features, as well as the results they gave when training and testing a model with them.

4.2.1 Event type

Since our goal was to potentially determine if features used in event specific solutions could boost performance, one of our students looked for an efficient way to sort tweets based on event type. The idea of finding trigger words to mark different events was tested by collecting the most common words used in tweets per event type. Before checking for these common words, stop words, URLs and mentions were removed from the post text to avoid skewing the data with irrelevant words and clutter. Once we had a list of trigger words for each event type, words that were specific to individual subevents were removed as they would not be helpful in determining event types of future tweets. These words included locations and specific people that may have been involved in an incident because for instance, if there is a shooting in Texas, the word Texas will not be helpful in identifying the event type shooting for a shooting in Colorado.

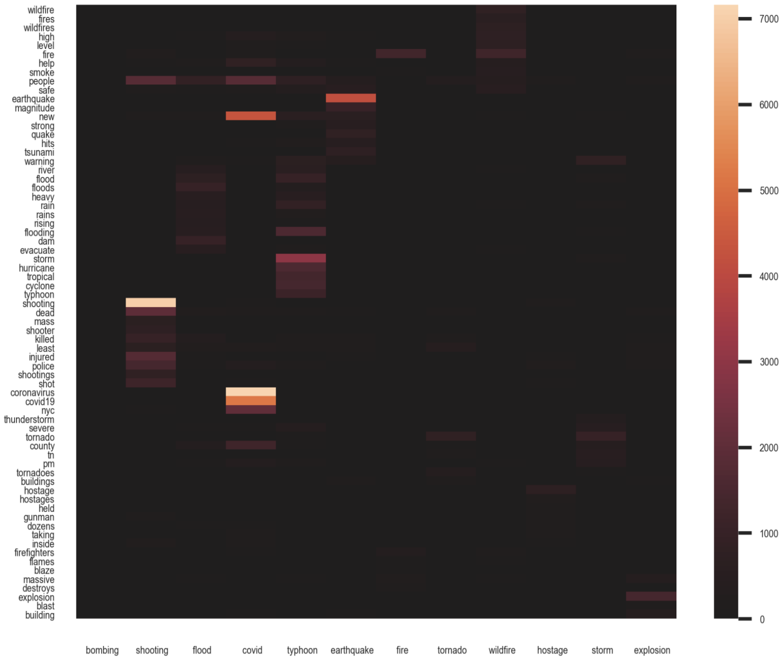

Figure 2: Heat map of trigger words and event types found in 2018-2020 data

Figure 2 above shows the results of our work described above looking for trigger words. This figure shows the commonly used words found in the tweet streams, in relation to the number of times a particular word was used in a tweet by event type. Words such as “shooting”, “dead”, “mass” and “shooter” were more abundant in tweets given the event type shooting, whereas words such as “storm”, “flooding”, and “floods” were more common in tweets with the event type typhoon.

4.2.2 Augmenting the Annotated data

TREC's supplied event annotations are not tagged with dates or longitude and latitude locations, while all tweets have a timestamp and many have location information. The time and date of the start of each event had to be researched and then manually added to the annotations. Date and time of day could be found for shootings, earthquakes, tornados, bombings, explosions, hostage situations, and some fires. Typhoons, storms, floods, wildfires, and the pandemic could only be as accurate as the first day the event happened, midnight UTC. Some tweets appeared before the stated time. Most of those were irrelevant, but some had a higher priority even though they were found to be not directly related to the main event. These would have to be considered in a real scenario where a system would need to sort thousands of live tweets, so they were left in the tests. For our location testing we selected events that were centered around identifiable narrower locations, i.e. shootings, bombings, and some of the floods and fires. Those events that were over a larger geographic area were not selected for testing at this time.

4.2.3 Length

When considering length of tweet. We had to exclude tweets prior to December of 2017. Prior to this date the limit on characters in a tweet was 140. Twitter increased the character limit to 240. With twice as many characters available, results would surely be different for any events going forward. We did not consider the older tweets as they were made with the older limit in mind, and any new tweets would only be restricted by the newer larger one. This results in a smaller dataset, but the results would likely be more pertinent to our purpose.

4.2.4 Results

The results of the elapsed time between the event and the posting of each individual tweet showed promising results when graphed. The largest numbers are different depending on the event type, but there is clearly an inverse relationship between the elapsed time and the priority of each tweet. COVID tweets were excluded for this, as its time span was too broad to be relevant for this test. Otherwise, the latest tweets overall are only about two months after the start of the event. Shootings and earthquakes had ample sample data and showed good trends as expected. Additionally, floods and typhoons exemplified the trend just as well. Most of the other event types either did not show the same trend or did not have enough topics and tweets to justify showing any trend on their own. With regular balancing of the data, most of the events with substantial data had a priority prediction accuracy of between .4 and .5 with most classifiers, again with the exception of neural nets and linear regression. Unexpectedly, floods produced accuracy results above .6, and just above .7 with some classifiers when using the staggered balanced data. This was unexpected because floods do not have a timestamp including any hours or minutes. It's possible the wider distribution of dates in critical tweets is responsible for there being better results.

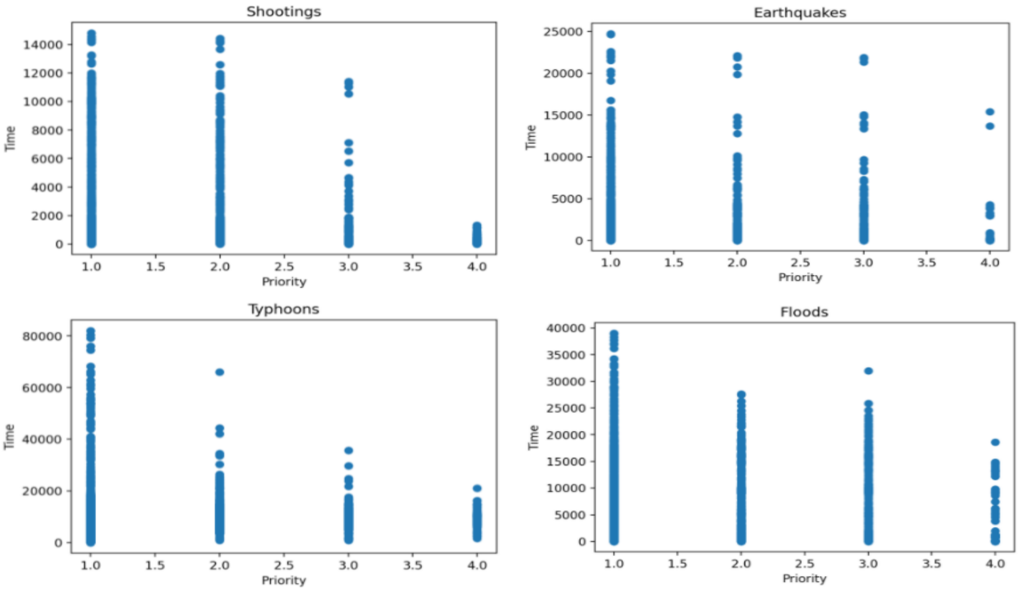

Figure 3: Graphs showing distribution of elapsed time across the four priority levels (whole numbers). Four event types with the best results. Tweets from before the stated time of event are excluded to better show the trend.

We expected higher priority tweets to be longer (up to the maximum 240), as they'd contain more pertinent information. This wasn't the case as most critical tweets were actually usually at about 150 characters or less for some events, with the average length at 116. The users are likely trying to be quick and concise when getting information out. However, there are several outliers to this, making it a rather weak rule of thumb. Low, medium, and high priorities all have a more evenly distributed series of lengths. Our machine learning tests did not provide interesting results. The accuracy of most classifiers fell between .3 and .5 with the exception of the neural net, linear regression, and logistic regression, which all scored even lower. Overall, there likely isn't much of a future in exploration of this feature for the IS track.

When testing our features, retweet count, favorite count, hashtag count, and whether the tweet is a reply and/or contains a mention did not give us very strong results with accuracy scores ranging from .25 to .35. However, we did find differences between event types. For instance, on average, priority level 2 wildfire event tweets have about 5 favorites, while tweets of the same priority with the event type earthquake have around 80 favorites on average, and explosions have 40 on average (Figure 4). Discrepancies in the metadata such as these encouraged our thought that event specific solutions could be advantageous.

When testing our trigger words against tweets to search for event types, we got good results with accuracy scores between .68 and .75 for event identification. Event types were also analyzed for tweets from 2018 to 2020 with trigger words from the same years. In this data 12 event types were present: bombing, shooting, flood, COVID, typhoon, earthquake, fire, tornado, wildfire, hostage, storm, and explosion. Event type was also tested against tweets from just 2020 with trigger words from 2018 and 2019. Here only 5 event types were present: shooting, flood, typhoon, earthquake, and wildfire. Both tests gave similar results, with the second test giving numbers just slightly lower.

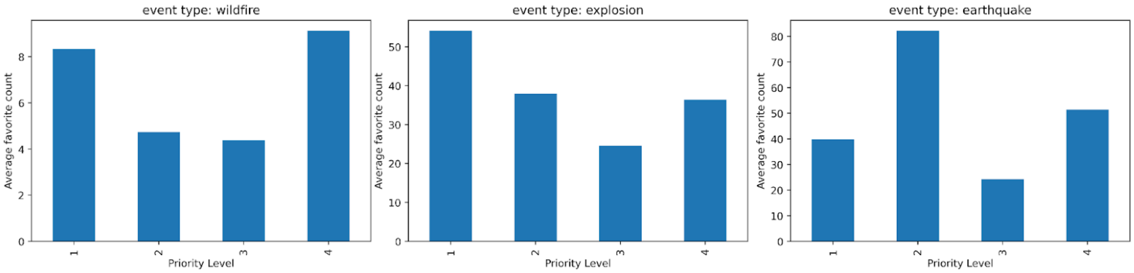

Figure 4: Comparing favorite count between event types.

We are early on in our analysis comparing the priority of tweets relative to the distance between the event and the tweet. The following 19 events were included in our analysis relative to distance from the event location and its effects on priority:

tennesseeDerecho2020
daytonOhioShooting2020
coloradoStemShooting2019
sanFransicsoPierFire2020
elPasoWalrmartShooting2020
southAfricaFloods2019
texasAMCommerceShooting2020
whaleyBridgeDamCollapse2019
edenvilleDamFailure2020
houstonExplosion2020
nepalEarthquake2015
brooklynBlockPartyShooting2020
virraMallHostageSituation2020
keralaFloods2020
gilroygarlicShooting2020
flSchoolShooting2018
shootingDallas2017
hurricaneBarry2020
baltimoreFlashFlood2020

We manually tagged the events with their longitude and latitude as seen in the augmented event below:

```
<top>
<num>62</num>
<dataset>tennesseeDerecho2020</dataset>
<title>Tennessee derecho</title>
<type>storm</type>
<url>https://www.washingtonpost.com/weather/2020/05/04/deadly-derecho-slammed-
nashville-with-70-mph-winds-sunday-snapping-trees-knocking-out-power/</url>
<narr>Intense windstorm caused damage across Tennessee
</narr>
<long>
-86.779068
</long>
<lat>
36.166340
</lat>
</top>
```

We were able to then easily automatically compute the distance between the event and the tweet in miles using either the contents of the *Place* feature or the *Location* feature of tweets that contained that data. While tweets that had information in those fields was sparse, we saw a fairly clear trend between distance and priority, Figure 5. Work is ongoing to explore this further and to determine how this holds across different events.

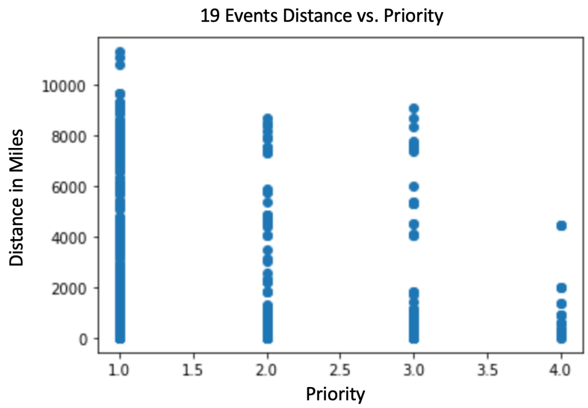

Figure 5: Distance relative to Priority of tweets

References

Buntain, Cody, Richard McCreadie and Ian Soboroff. 2020. Incident Streams 2020: TREC-IS in the Time of COVID-19. In Proceedings of The Twenty-Ninth Text REtrieval Conference, Gaithersburg, Maryland, November 2020.

Hepburn, Alexander J. and Richard McCreadie, 2020. *University of Glasgow Terrier Team (uogTr) at the TREC 2020 Incident Streams Track*. University of Glasgow, UK.

Hürriyetoğlu, Ali, Nelleke Oostdijk, and Antal van den Bosch, 2014. *Estimating Time to Event from Tweets Using Temporal Expressions.* Radboud University Nijmegen, The Netherlands.

Wang, Congcong and David Lillis, 2020. *Multi-task transfer learning for finding actionable information from crisis-related messages on social media*. University College Dublin.

Wang, Yunli and Cyril Goutte, 2017. *Detecting Changes in Twitter Streams using Temporal Clusters of Hashtags.* NRC Canada, Ottawa.

Wurzer, Dominik, Victor Lavrenko, and Miles Osborne, 2015. *Twitter-scale New Event Detection via K-term Hashing.* University of Edinburgh.

Voorhees, Ellen M. 2007. Overview of TREC 2007. In Proceedings of The Sixteenth Text Retrieval Conference (TREC 2007).