Tweet Stream Analysis for Flood Time Estimation

Ali Hürriyetoğlu, Nelleke Oostdijk, Antal van den Bosch, Jurjen Wagemaker

Centre for Language Studies, Radboud University

Information Retrieval for Information Sciences (INFINITI), WP1, Adaptive Information Extraction over Time

Task Definition

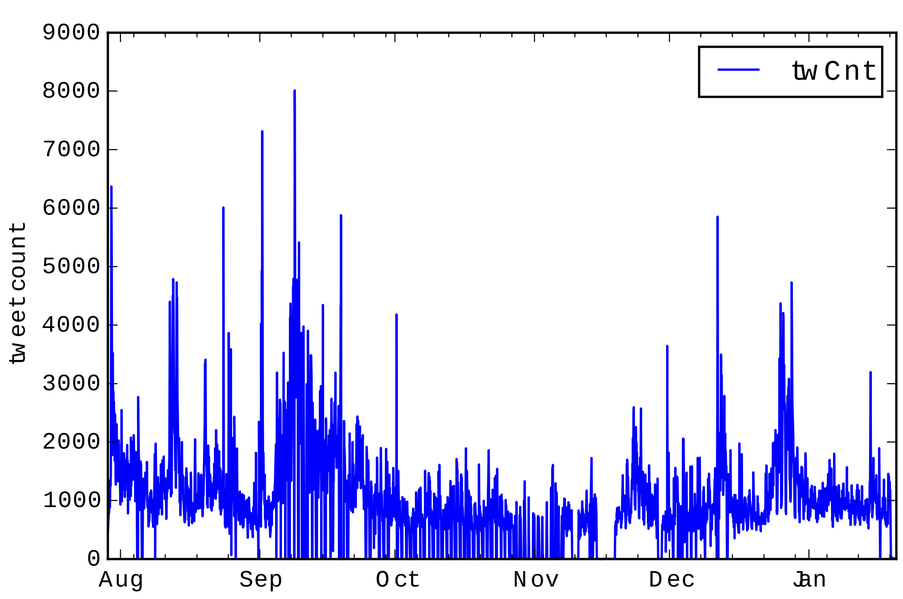
Given a stream of tweets about an impending flood, identify the set of tweets relevant for estimating the remaining time to event (TTE)/

Motivation

The stream contains quite some noise and we need a method that removes irrelevant tweets and allows us to focus on information that can hint at the TTE.

1. Creating a preliminary tweet set

- ✓ Collect tweets by using event-related key terms: flood, floods, flooding, inundation, inundations, landslide, landslides.
- ✓ Eliminate place, organization, and person names to focus on core event information.
- ✓ Eliminate tweets that do not contain the most frequent bigrams to reduce the inherent key term ambiguity.



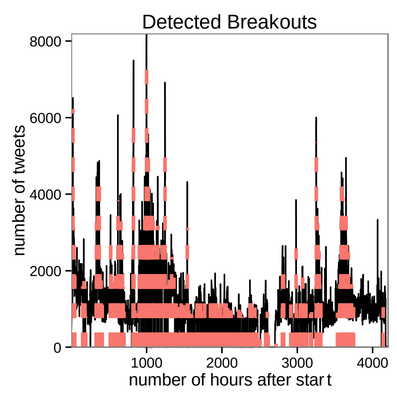
Temporal distribution of the tweets in the data set

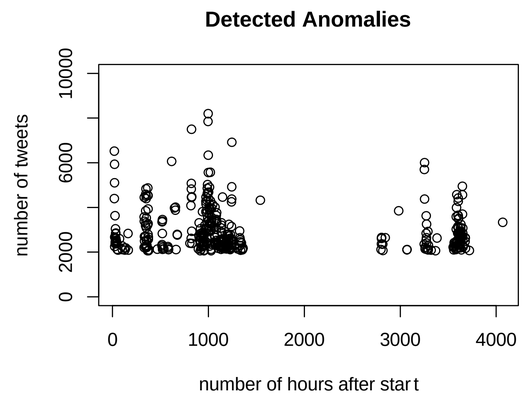
2. Identifying relevant time frames

- ✓ Analyze the tweet count distribution to detect time frames that show peaks which reflect both a breakout and an anomaly.
- ✓ Cluster the peaks using tweet set characteristics:
  - ✓ The ratio between the tweet count and the tweets that have a link, a mention, a hashtag, a photo, or coordinates; number of countries and number of users involved; number of tweets sent from mobile devices
  - ✓ The follower ratio of the users that are present in each peak
  - ✓ The type/token ratio of the text in each peak

2.1 Results

- ✓ Number of peaks was not big enough to identify peak types other than news articles.
- ✓ Used features did not take the key term ambiguity into account.
- ✓ Outlier peaks were identified and eliminated: “flood timeline” and ”flood insurance”.



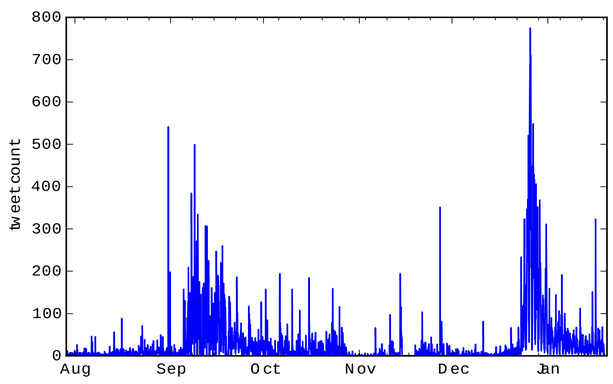


3. Identifying relevant information groups

- ✓ Cluster tweets by using only bigrams whose tokens are not in the stopword list.
- ✓ Eliminate retweets as they do not provide new information and prone to be used for overemphasizing an information group.
- ✓ Tag the clusters: official warnings, user observations and comments about a flood disaster, reference to news articles, historical event information, irrelevant.

3.1 Results

- ✓ Short tweet text yields focused small clusters around specific bigrams
- ✓ Key term ambiguity results in big clusters that contain various information groups.
- ✓ Tweets that belongs to irrelevant information groups were removed from the tweet sets that constitute peaks (i.e. %14 of the peak tweets that are not retweets or contain at least one used feature).



Occurrence pattern of the bigram “flood victims”

4. Conclusion

- ✓ Combination of textual and contextual features of a tweet can reduce the noise in the stream.
- ✓ Excluding place and time from the text of a tweet makes resolving the key term ambiguity difficult.

5. Future Research

- ✓ Include the place, time and named entity information into analysis.
- ✓ Classify peaks into classes, that are induced from the tweet clustering to determine relevant time frames.
- ✓ Create a time to event estimation model based on cleaned text.

Notes

- ✓ Clustering was done with K-Means algorithm that is implemented in scikit-learn, using euclidean distance, k=10 and k=30 for peak and tweet clusterıng respectively.
- ✓ Anomaly and Breakout detection were done with Twitter’s Breakout Detection and Anomaly Detection packages.

Acknowledgements

This research was made possible by the Dutch national program COMMIT and Floodtags.com in terms of funding and providing tweet ids respectively.

a.hurriyetoglu@let.ru.nl
n.oostdijk@let.ru.nl
a.vandenbosch@let.ru.nl
info@floodtags.com

@COMMIT_nl, @Adnext_Commit, @hurrial, @antalvdb, @jurjenwagemaker

Radboud University Nijmegen