Web Mining Project: Mining Information From Social Media Networks During Crisis Events

Maria Bracque Vendrell - 21800103
Théo Condette - 21803967
Rodrigue Nasr - 22201765
Hoai-Nam Nguyen - 21911359

M2 DSSS & M2 ES - Semester 2

Contents

1 Introduction

In today's digital era, social media platforms have become integral channels for communication, particularly during times of crisis. Among these platforms, Twitter stands as a prominent hub where people worldwide convene to share news, connect, and engage in conversations.

The objective of this project is to analyze and describe part of the social web graph of Twitter using web mining techniques. This graph represents the network of connections between users, encompassing relationships, interactions, and shared content. Our objective is to gain insights into how users connect and communicate, as well as the types of content that drive engagement. From viral tweets to trending hashtags, each element contributes to shaping the ongoing conversation and influencing public discourse.

In the first part of the project, we will create a dashboard specific to various event types (e.g., floods, fires) by analyzing related social sub-graphs. The dashboard will provide key insights such as the intensity of social activity, identifying central users based on different criteria like connectivity and information dissemination, and analyzing the distribution of tweet posts categorized by levels of criticality (Low, Medium, High, Critical).

In the second part, for each event-based graph analyzed, we will construct embeddings for a sample of N users using both the graph structure and the content of their posts. We will outline our approach to sampling users, compute user-user similarities based on these embeddings, and explore if correlations exhibit similar trends across different event types, fostering a comparative discussion.

2 Data Description

The dataset contains real-world Twitter posts shared during past crisis events such as floods, fires, tornadoes, and earthquakes. This dataset was made available by the NIST TREC Incident Stream Initiative organizers, with the primary goal of fostering research and academia in the automated processing of social media streams during emergencies. It facilitates the categorization of information and assistance requests made by citizens to emergency operators. The dataset primarily includes the following high-level data to facilitate emergency response through data processing:

- *Events*: Represent occurrences of crisis events, such as the "2012 Guatemala earthquake" or the "2013 Alberta floods."
- *EventType/Topic*: Corresponds to categories of events, such as "flood," "earthquake," "typhoon," or "bombing."
- *User*: Refers to active Twitter users during the specified event.
- *Tweet*: Denotes a post emitted by a user during a crisis event, where each tweet is associated with a specific event. Tweets can be original posts, retweets, or replies, and may mention other users or contain event-related hashtags (e.g., "nswfires," "abflood"). Each tweet includes a textual content section containing its main message.
- *TweetPriority*: Provides human-labeled information about the criticality level of a tweet, indicating the degree of emergency. Four priority levels are distinguished: Low, Medium, High, and Critical.
- $TweetHigh - LevelCategory$: Offers human-labeled information about the types of information that may be required by emergency response officers across various disasters, including Advice, News, Volunteer, and Search and Rescue.

3 Data Preparation

To prepare the database for our project, we began by reading the data from a graphml file and converting it into a directed graph format using NetworkX. This step ensures compatibility with Python's data manipulation libraries.

After, we created separate datasets for each node type present in the graph: users, events, tweets, and hashtags. For users, we extract relevant information such as user ID, tweet count, likes count, and followers count. Similarly, for events, we selected event ID, event type, and associated topic ID. Tweets include details such as tweet ID, creation time, retweet count, like count, tweet text, priority, and topic ID. Hashtags, on the other hand, are characterized by their occurrences and unique ID.

Further, we decided to enrich the tweet dataset. We computed the number of replies to each tweet, thus quantifying the level of engagement. This involves extracting edges representing replies to tweets and aggregating them to compute the total number of replies for each tweet. This information will help us identify key users and understand their impact on information dissemination during crisis events.

Lastly, we merge the tweet and event datasets based on their common topic ID, facilitating a comprehensive analysis that incorporates both tweet content and event context. This enables us to gain deeper insights into the relationship between tweets and the events they pertain to, thereby enhancing our understanding of the broader social context surrounding crisis events.

4 Analysis of Social Network Dynamics

In this section, we will analyze the social network dynamics for each event type, like floods and fires. We will work on social sub-graphs and will highlight social activity levels, central users, and tweet topics categorized by criticality levels ("Low", "Medium", "High" and "Critical").

4.1 Intensity of Social Activity

To understand how much people are talking about each event, we will use a multilayer graph representation that shows connections between users.

4.1.1 Hashtag Occurrence

People often use hashtags to discuss specific topics on social media and even more during crisis events. Below, there is a bar plot of the 50 most used hashtags, giving us an idea of what people are talking about.

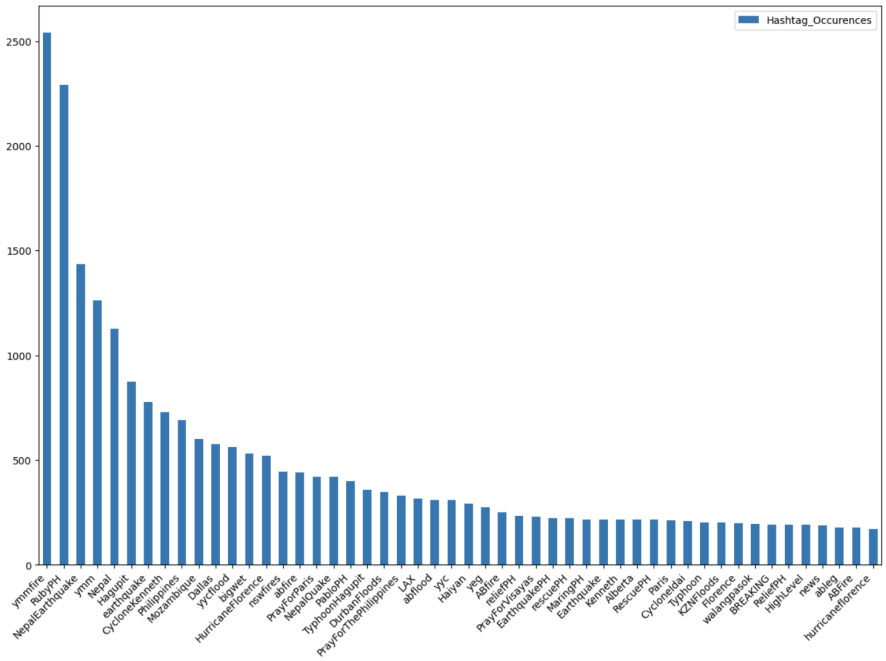

Figure 1: Top 50 Hashtag Occurrences

On this plot, we can see that the five most used hashtags are:

1. ymmfire: the wildfire in northern Alberta (Canada)
2. RubyPH: Typhoon Ruby that hit the Philippines
3. NepalEarthquake: the Nepal earthquake, also known as the Gorkha earthquake
4. ymm
5. Nepal

When looking at which hashtag was used the most within a day from the first occurrence of the event-type wildfire only, we find the following graph:

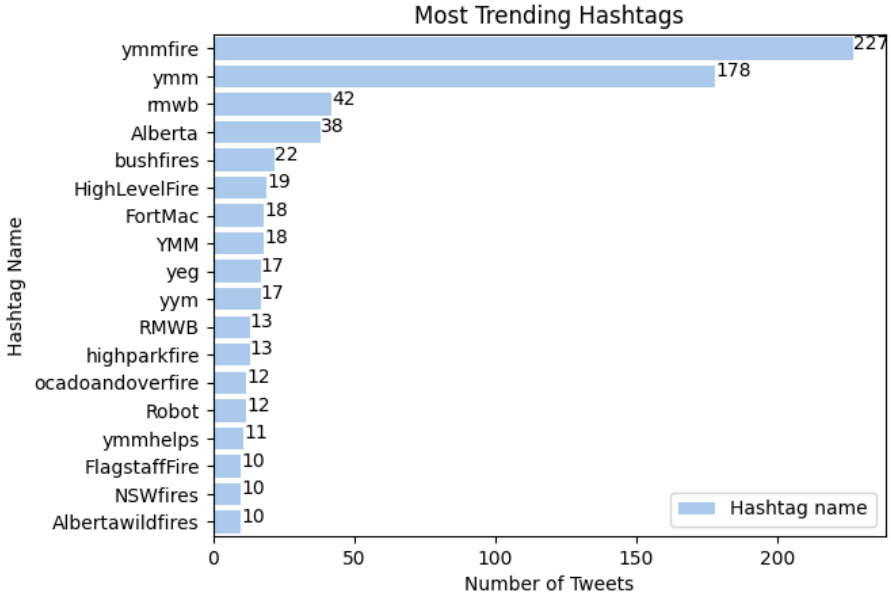

Figure 2: Most Trending Hashtag Within a Day of the First Occurrence

We can see that #ymmfire was used 227 times.

4.1.2 Retweet Counts

Next, we want to plot the top tweets by retweet count for each event type. We are going to group the tweets by event type, and for each group, we will select the top 10 tweets with the highest retweet counts.

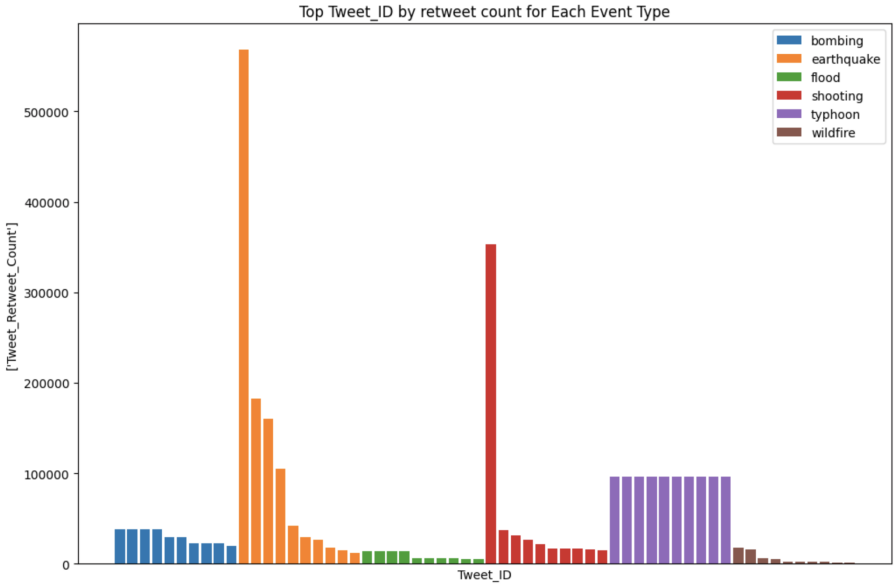

Figure 3: Top 10 Retweets by Event Type

We can see that the type of event that has the most retweets is earthquake.

4.1.3 Tweet Replies

Now, let's analyze the number of tweet replies for each event type and identify the top 10 tweets with the highest reply counts for each event type.

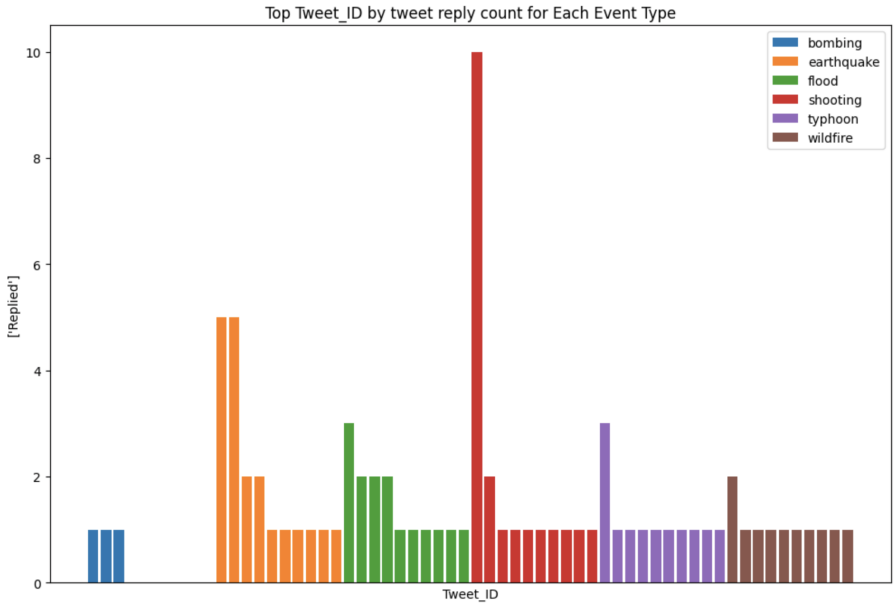

Figure 4: Top 10 Tweets with the Highest Number of Replies by Event Type

We can see that the event type containing the tweet with the most replies is shooting.

4.1.4 Tweet Likes

Following, let's have a look at the top tweets with the highest like counts for each event type. This will give us insights into the most popular and well-received tweets during different types of events.

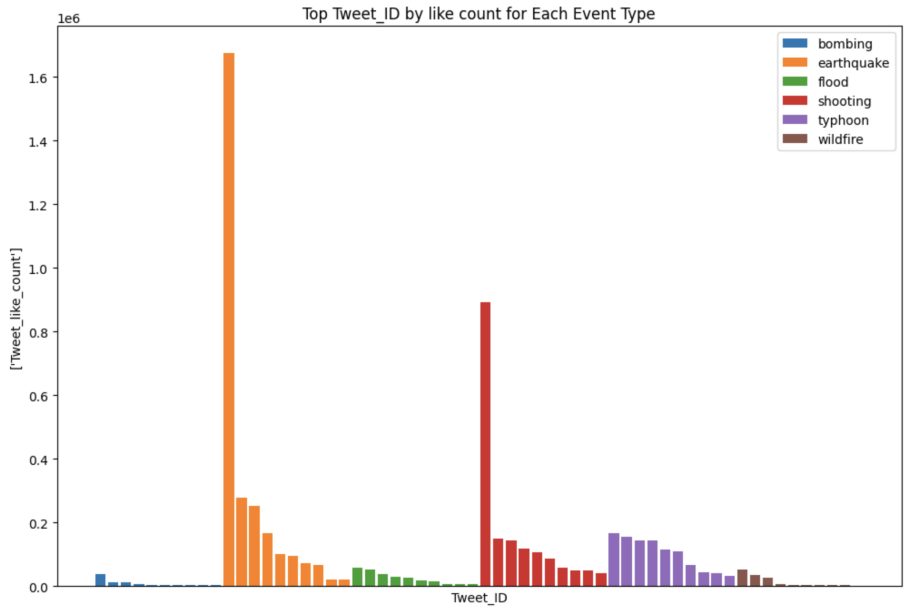

Figure 5: Top 10 Tweets with the Highest Number of Likes by Event Type

Here we notice that the tweet with the highest number of likes is from the event-type earthquake.

4.1.5 Retweets and Likes

Another interesting relation to look at is the number of likes and retweets per post. Thus, we plotted the following graph:

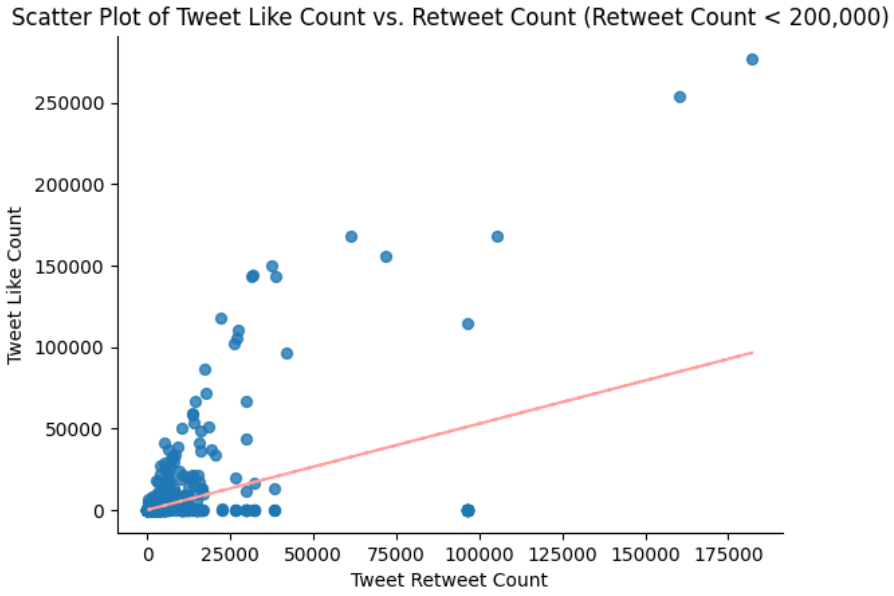

Figure 6: Scatter Plot of Tweet Like Count vs. Retweet Count

By looking at the plot and the trend line, it looks like tweets tend to have more retweets than likes.

4.2 Central Users

Central users play a crucial role in the dynamics of a social network, acting as pivotal figures in connecting, spreading, and gathering information. In our analysis, we define central users as nodes representing individual users, with edges representing interactions (retweets, replies and mentions). By aggregating the frequency of these interactions, we construct a network that highlights the interconnections of users.

For this section, we are going to focus on the wildfire event-type sub-graph. Analysis of other events (Bombing, Earthquake, Flood...) are produced in similar way and results can be checked in the notebook.

We are tasked to examine the central users based on different criterion. Related to their ability to connect users, the degree centrality is chosen because it computes the number of connections each user has. For user's ability to spread information,

we option for betweeness centrality as this measure helps to identify users who act as bridges between different parts of the network and are easily to propagate the information that they have. Finally, to find central users whose ability is information gathering, eigenvector centrality is reasonable choice because this centrality assess the influence of a user based on their connections and the connections of their connections. Users with high eigenvector centrality are connected to other influential users, making them potentially important gatherers of information.

4.2.1 Degree Centrality

The degree centrality of a node is simply its degree (the number of edges it has). The higher the degree, the more central the node is.

For the degree centrality, we found that the top five users with the highest degree centrality are:

Node	Degree Centrality
n85174	67.0
n61678	64.0
n85147	53.0
n87456	50.0
n86625	47.0

Here is a plot showing the top node's connectivity in the graph:

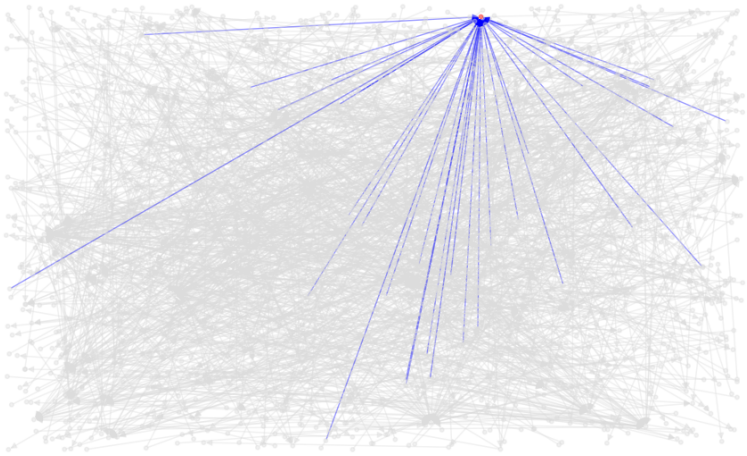

Figure 7: n85174 connectivity graph

4.2.2 Betweeness Centrality

Betweenness centrality measures the extent to which a node lies on paths between other nodes in the network. A node with high betweenness centrality has a large influence on the transfer of information through the network.

For the betweenness centrality, we found that the top five users with the highest betweenness centrality are:

Node	Betweenness Centrality
n67357	5.22e-05
n61678	2.51e-05
n85230	1.17e-05
n85208	9.45e-06
n85195	9.28e-06

Here is a plot showing the top node's connectivity in the graph:

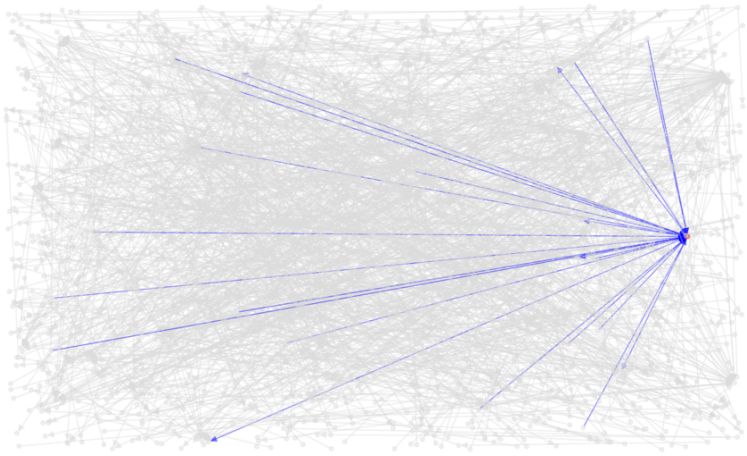

Figure 8: n67357 connectivity graph

4.2.3 Page Rank

PageRank is an algorithm used by Google Search to rank web pages in their search engine results. It assigns a numerical weighting to each element of a hyperlinked set of documents, to measure its relative importance within the set. In our context, we adapt PageRank to measure the importance of users in our social network.

For the page rank, we found that the top five users with the highest page rank are:

Node	Page Rank
n61678	0.0187
n63257	0.0163
n85147	0.0146
n87456	0.0103
n85174	0.0086

Here is a plot showing the top node's connectivity in the graph:

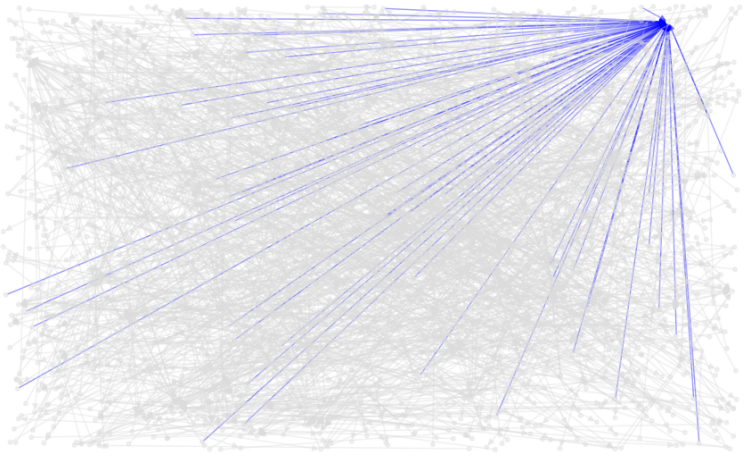

Figure 9: n61678 connectivity graph

4.3 Distribution of the tweet posts and their main topics wrt to levels of criticality

In this section, we explore the distribution of tweet posts and their associated main topics across various levels of criticality. The analysis focuses on categorizing tweet posts based on their levels of criticality and event types.

To begin with, we extracted tweet data from the graph, filtering out non-tweet nodes and extracting relevant information such as tweet ID, creation timestamp, criticality level, topic, favorite count, and text content. This data was merged with information on event types to provide additional context.

After filtering out tweets with unknown criticality levels, we calculated the count of tweets for each combination of criticality level and event type. This step allowed us to understand how tweet posts are distributed across different levels of criticality and types of events.

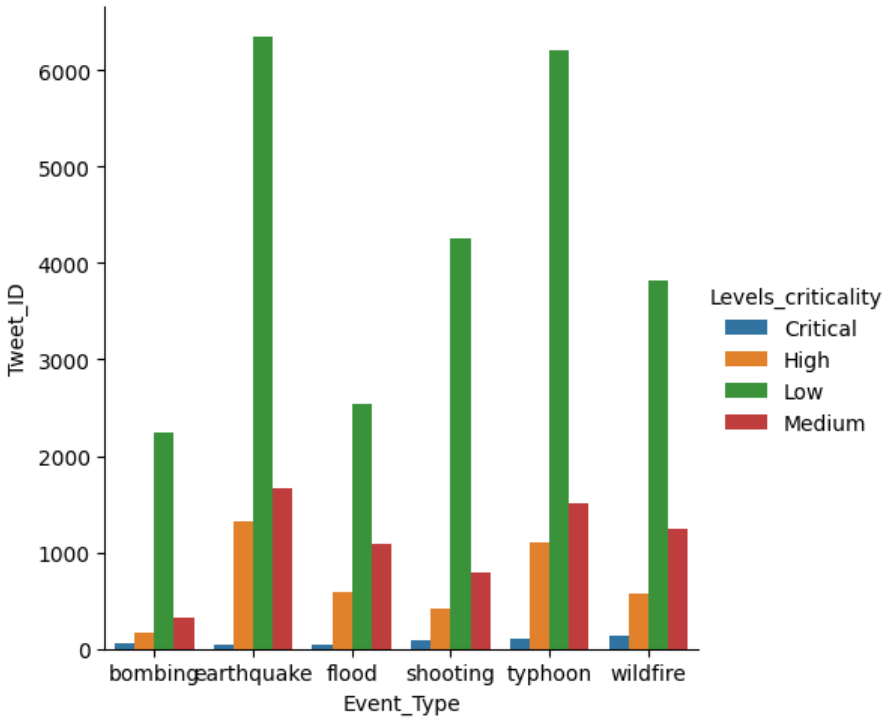

Figure 10: Distribution of Tweet Posts with Respect to Levels of Criticality

The resulting visualization, shown in the bar plot, provides insights into the distribution of tweet posts across different levels of criticality for each event type. Notably, we observed that the category with the highest number of low-level criticality tweets is 'earthquake' and the category with the highest number of critical-level criticality tweets is 'wildfire', suggesting a heightened level of urgency or importance associated with tweets related to wildfire events.

5 User Embeddings and Similarity Analysis

In this section, we present our analysis about user embeddings and similarities through Wildfire event sub-graph. User embeddings are compact numerical representations of users, offer insights into the underlying dynamics and patterns within these networks. Our analysis centers on constructing these embeddings by considering two fundamental aspects: the structure of the network and the textual content shared by users. We are interested in seeing if those similarities are the same, thus checking if similar users in terms of graph connectivity post similar content. Like for previous section, analysis of other events (Bombing, Earthquake, Flood...) are produced in similar way and results can be checked in the notebook.

Our exploration begins with a focus on:

1. Sampling Users: taking randomly a sample of users and building embeddings based on graph structure and contents

2. Computation of User-User similarities: Based on each embedding, compute User-User similarities.

3. Correlation Analysis: Compare the similarities using some correlation metrics to conclude on the presence of a structure-content relation.

5.1 Sampling Users

To build user embeddings and analyze similarities, we first need to sample a subset of users from the event-based graph. We employ a random sampling approach to ensure the representativeness of the sampled users, where we take 20 percent of the total number of users from the wildfire graph we did previously.

Next, from those selected users, we build two embeddings:

1. A graph structure embedding: We create embeddings based on graph structure using “Node2Vec”, which is an effective technique for learning low-dimensional

representations of nodes in a graph while preserving the structural properties of the network.

2. User content embedding: We leverage text content from user posts to create embedded content. We use the TF-IDF vectorization method, which takes into account Term Frequency, measuring how frequently a term appears in the tweet, and Inverse Document Frequency, measuring how important is that term.

Note that to build User-content embeddings, a cleaning of the tweet texts was done by converting text to lowercase, removing punctuation and new line characters, tokenizing sentences, lemmatization, stemming and finally converting back tokenized text to text. Remarque that we do not process stopwords because after removal, the collection of tweets is empty.

5.2 Similarities

Now that we have the 2 types of embedding, we will calculate User-User similarity for each of those embeddings. To measure it, we will use the cosine similarity and apply it to each of them.

The cosine similarity is defined as:

$$Sim_{cos}(u,v) = \frac{\sum_{j=1}^{K} ve_{uj} \times ve_{vj}}{||ve_u|| \times ||ve_v||}$$

with $ve_u(v_{u1}, ..., v_{uK})$ being user u vector profile over K features and $||ve_u||$ being the norm of user u vector profile.

We finally obtain for each type of embedding a 341×341 similarity matrix.

5.3 Correlation Analysis

With this similarity matrix, we will now compute Pearson's correlation coefficient.

```
np.corrcoef(content_similarity.flatten(), graph_similarity.flatten())[0, 1]
```

For the wildfire subgraph, we get a value of approximately 0.12.

This means that there is a positive correlation between content similarity and graph similarity, indicating a weak positive linear relationship between these two factors. The correlation is not very strong, suggesting that user-based similarities in graph structure embeddings and content embeddings are somewhat related but not strongly correlated.

Let us now test if the correlation is statistically significant, using Spearman's rank correlation with a hypothesis testing:

```
spearman_corr, spearman_p_value =
spearmanr(graph_similarity.flatten(), content_similarity.flatten())


print("Spearman's correlation coefficient:", spearman_corr)
print("p-value:", spearman_p_value)
```

We get a positive coefficient (approximately 0.2) with a very low p-value (lower than 0.01 percent). Then we reject the null hypothesis and therefore conclude that a significant correlation exists. However, its value is very low, indicating that the relationship between user-to-user similarity and user content is positive but weak.

The realisation of this analysis has the same conclusion for most of the events : there is a positive and significant correlation between user-user graph similarity and user-user content similarity. Which means that in general close users tend to post same contents in case of emergency. However, we found an exception for shooting events. We get a Pearson's correlation coefficient of around 0.1 and a significant Spearman's rank correlation of -0.01 which is very low. This means that there is a weakly positive linear relationship (due to Pearson's) which is not monotonic (due to Spearman's). This might be due to influential observations that have an impact on Pearson's but not on Spearman's which is based on the rank.

Thus, we conclude that in general, some similar users in terms of graph connectivity may post similar content during emergency.

6 Conclusion

In conclusion, our study has provided insights into how people use Twitter during crises. We found that certain topics and users are more prominent during these events, indicating their importance in spreading information and engaging with others. In addition, by analyzing tweet distribution based on the level of urgency and examining key users within the network, we gained a deeper understanding of how information flows and who plays pivotal roles in shaping discussions during crises.

Moreover, our analysis of user embeddings revealed that users with similar network connections tend to share similar content. This suggests that social network structure influences the types of information shared within the community.

Overall, our findings highlight the critical role of social media in crisis communication and decision-making. Understanding these dynamics can inform more effective strategies for managing and responding to crises in the future.