Natural Language Processing with Disaster Tweets

Vivian Do

University of San Diego

Master of Science, Applied Data Science

ADS 501

12 September 2022

Contents

Today, every citizen is armed with a smartphone that allows them to document an event as it occurs. These events can then be shared worldwide, in the form of photos, videos, and written text. This phenomenon, known as ‘citizen journalism’, allows journalism to become a more democratic practice that everyone can participate in. Our client, NBCUniversal News Group, hopes to capitalize on the millions of user-generated content (UGC) on social media produced by citizen journalism to gather more information on disasters happening around the nation.

Business Understanding

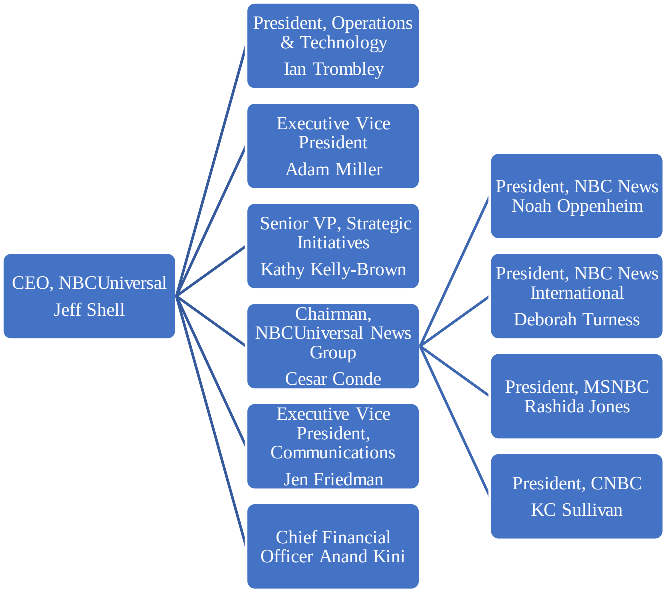
Background

Organization

Figure 1 depicts the leadership of NBCUniversal. NBCUniversal News Group (the client), is the news division of NBCUniversal and comprises of NBC News, NBC News International, MSNBC, and CNBC.

Figure 1

Flowchart of NBCUniversal News Group Leadership



Key persons of the organization include: CEO of NBCUniversal, Jeff Shell, who oversees NBCUniversal and all of its entertainment, news and sports entities; Cesar Conde, the Chairman of NBCUniversal's news division, NBCUniversal News Group. The presidents of each news entity are as follows: Noah Oppenheim for NBC News, Deborah Turness for NBC News International, Rashida Jones for MSNBC, and KC Sullivan for CNBC. The office of the Chief Financial officer, headed by Anand Kini, serves as the internal sponsor for the project. The steering committee consists of the Chairman of NBCUniversal News Group, Cesar Conde, along with the presidents for each news group as mentioned above. The assigned data protection officer (DPO) is Kim Nguyen.

The business units affected include IT, finance, and legal. A successful model will require IT implementation and management to monitor Twitter as a platform. Financial funding will be required to train and test the model and includes the hiring of additional personnel if necessary. The legal team will serve as advisors for issues related to data privacy during the project and as part of monitoring thereafter. All effected business units have been made aware of the ongoing data mining project.

Current Solution

Currently, NBCNews obtains a portion of its stories through its tip lines, where individuals can reach out through various messaging apps (WhatsApp, WeChat, etc), by phone, or by email. Tips are filtered based on relevance, validity, and interest and only a small portion of tips are actually investigated. According to data provided by the client, an average of 100 tips are received per day and only two percent are deemed newsworthy. Verification of a tip requires back and forth communication between the informant and the employee. This process can take days and often leads to dead ends as communication is lost before all of the facts can be verified.

Tip lines are most useful for developing stories, such as when an individual wants to report on unwanted development happening in the community or other unjust events. For emergent events such as natural disasters, tips lines are less valuable due to a long verification step and difficulty maintain in contact with the informant.

In addition to its tip lines, individuals can connect with NBCNews on various social media platforms by tagging its account (@NBCNews) in their posts. Although there is a higher volume of posts, the tags are saturated with different types of tweets, such as personal opinions related to a news story. NBCNews presently employs media interns going through tags and comments on Twitter to look for posts made about disasters happening throughout the world. This current solution is highly accurate, since employees are able to decipher right away if a post is about an actual event, especially if there is a visual aid.

Problem Area

This project is an effort to mitigate some of the challenges of traditional tip lines and automate the process of receiving information. The client hopes to utilize user-generated content on popular social media platforms, specifically Twitter. The problem is that it is not always clear whether an event is actually happening due to the nature of social media. It is common to use slang or speak in a metaphorical manner that is clear to a person, but not to a machine. For example, the description of a ‘disaster’ can be subjective. It can be a catastrophe phenomenon such as an earthquake or multi-car crash or used to describe a wedding not going well (“This wedding is a disaster!”).

Accuracy is vital in the world of news and inaccurate evaluations could lead to severe financial consequences for the client. The client could incur a significant financial loss if money and resources are diverted to places where a disaster did not occur. Additionally, an inaccurate reporting of a real disaster or false alarm could damage the client's reputation and relationship to viewers worldwide. A decreased viewership would result in a loss of advertisement revenue in the millions.

Business Objectives and Success Criteria

According to the Pew Research Center, 23% of adults in the US use Twitter (Auxier & Anderson, 2021). On average, 6,000 Tweets are posted every second, which sums up to 500 million tweets per day ("Twitter Usage Statistics"). The ubiquity and abundance of tweets provide a wealth of information into events happening all around the world. The client wishes to harness the power of social media to gain an edge against their competitors.

The main objective is to programmatically monitor Twitter for natural disasters and accidents happening globally. Tweets can be identified as they are posted for a quicker response to emergent disasters. In addition, communication with the informant would be greatly improved as they can be reached directly through their Twitter account. If they had posted recently, direct messaging through Twitter is much more reliable than through an email that they may not check often.

The success criteria for the business objective has been determined as follows:

(1) Obtain more than eight percent (8%) of stories from Twitter. This is the current statistic for stories obtained through tip lines.

(2) Increase the number of daily viewers by twenty percent (20%). An increased viewership could be leveraged for commercial purposes such as marketing advertisements during intermission.

The hope is that accurate news reporting that incorporates more personal first-hand accounts will attract more viewers. If successful, the business would be able to gather more information on disasters that are happening as they are posted on social media.

Inventory of resources

The base hardware is a 2019 Macbook Air, which will be available at all times throughout the duration of the project. The hardware maintenance schedule has been adjusted as to not conflict with this project.

The dataset used for this project was created by Figure Eight Inc., a Machine Learning and Artificial Intelligence company based in San Francisco and originally shared on their website (https://appen.com/pre-labeled-datasets/). The data was obtained from a Kaggle coding competition titled "Natural Language Processing with Disaster Tweets". It contains both a training set and test set in a comma-separated values (CSV) file. In the training set, each Twitter post (Tweet) has an additional column denoting whether the tweet is about a real disaster (1) or not (0). Exploratory data analysis and visualizations will be performed using Rstudio (Version 2022.07.0) to gain further insight into the data. Binary classification to determine whether a tweet is about a real disaster will be performed using a number of possible algorithms, namely Logistic Regression and k-Nearest Neighbors.

General questions regarding this project can be directed to Ankush Morey, the Senior Data Scientist leading this project. Technical support inquiries should be directed to Michael Mann, the Director of Information Technology. They are both available during normal business hours. If Ankush Morey is not available, please direct questions to Jason, Lead of Data & Analytics.

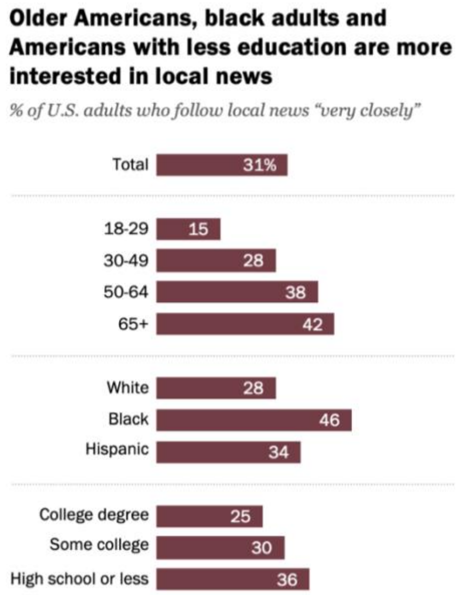
Requirements, assumptions, constraints, and RESOLVEDD Strategy

Requirements

The target group are those residing in the US, aged 18 and older, who follow local news closely. According to a survey conducted by Pew Research Center in 2018, the majority of US adults who watched local news included: adults aged 50+ (80%), Blacks and Hispanics (80%), and those with a high school education (or less) or some college (66%) (Barthel et al, 2019).

Figure 2

Profile of Target Group



Note. Whites and blacks include only non-Hispanics; Hispanics can be of any race. Source: Survey conducted Oct. 15- Nov. 8, 2018. From "Older Americans, Black Adults and Americans with Less Education More Interested in Local News," by Barthel M., Grieco E., and Shearer E, 2019, *Pew Research Center*, https://www.pewresearch.org/journalism/2019/08/14/older-americans-black-adults-and-americans-with-less-education-more-interested-in-local-news/. Copyright 2022 by Pew Research Center.

The project plan, as detailed in the GANTT chart (see Figure 3) should be strictly followed each week to ensure completion by the deadline. Additional legal requirements are in regards to the use of Twitter and its contents. This project assumes that user-generated content on Twitter can be collected in accordance with Twitter's terms of usage and relevant social media laws.

Assumptions

The following assumptions are made regarding the data: (1) Each tweet has been posted by a Twitter user, and not a bot; (2) The associated 'keyword' has been selected from the body of the tweet, and not from other text, such as the username of the user. This can be verified during the data exploration process by cross-referencing the keyword and text of the Tweet; (3) Tweets with a target level of 1 in the training set have been verified and confirmed as real disasters. Similarly, tweets with a target level of 0 in the training set have been verified and confirmed as non-disasters. Additionally, the location associated with each real disaster has been verified and accurate.

Constraints

The project is subjected to budgeting, legal, and ethical constraints. The data mining process may require multiple iterations in the modeling and evaluation phases and will be subjected to budgeting limits and project deadlines. According to Twitter's privacy policy, access of content on Twitter must be through its APIs (application programming interfaces). Usage of content must be in accordance to Twitter's Developer Agreement and Policy, Display Requirements, and Automation Rules. Misuse of Twitter's APIs is "subject[ed] to enforcement action, which can include suspension and termination of access" ("More about the restricted uses of the Twitter APIs"). More information can be found here:

https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases. Ethical considerations regarding data privacy will be discussed and reviewed using the RESOLVEDD Strategy.

RESOLVEDD Strategy

The RESOLVEDD Strategy is used to introduce and address ethical concerns and can be used in business applications or AI development. The nine steps include: Review, Estimate, Solutions, Outcomes, Likely (impact), Values, Evaluate, Decide, and Defend (Vakkuri & Kemell, 2019).

First we will review the facts, including the history, background, and context of the project. One of the cornerstone of research ethics is informed consent. Researchers must explain to the intended subjects the purpose of the research, what the subjects' participation will entail, and any risks involved. Once informed, the intended subjects must give their consent to participate in the research. With the advent of the internet and the widespread use of social media and other publicly available platforms, researchers are able to study human behavior without the explicit consent of the subjects ("Ethics and Data Protection"). This project has been created to derive an algorithm to predict whether a post on Twitter is about a real disaster or not. In a sense, we are conducting research on the linguistic behaviors of individuals who are experiencing a real or metaphorical disaster. All research must be conducted in adherence to Twitter's privacy policy regarding its API (application programming interfaces) usage.

Second, we estimate the conflict or problem. Ethical considerations pertaining to the project involve those of data privacy and protection. Although Twitter's privacy policy allows third party access of Twitter content through its APIs, Twitter users have not explicitly consented to the use of their tweets in this specific application. Data protection is also an issue because we are storing and working with sensitive data. Our dataset contains information regarding the location and text of the tweets, which can be cross-referenced to obtain the user's profile information such as name/pseudonym, username, and profile pictures. Individuals have been identified through seemingly 'anonymous' datasets, so data protection is paramount. The main issue is that even if our research is legally permissible, there are still financial and reputational repercussions should these principles be violated.

Next, we list the main possible solutions and evaluate each solution in terms of outcomes, likely impacts, and values upheld. Our first solution is to seek the informed consent of all individuals whose tweets have been collected in our dataset. This could be done by working with the original creator of the dataset (Figure-Eight) to obtain Twitter usernames associated with each tweet and then contacting them through Twitter direct messaging. This upholds the values of informed consent and allows individuals to opt out of the research if they wanted to. This is the ideal solution but is not the most realistic in terms of time and resources. It is likely that each tweet was created by a different user, so over ten thousand users must be reached. Additionally, we would expect the response rate to be low—some users are suspended, others may not be in use, while some may be unresponsive. We also have to consider the possibility of individuals opting out of the research, which could have further implications in terms of the volume of our dataset. The second solution is to implement company-wide measures for data protection, which serves to protect our dataset and materials associated with the project. This solution could have the greatest financial burden if the client does not already use a data protection service/platform and could make it more difficult to work with the data by adding additional steps.

Lastly, we decide which solution is best and defend it against objectives to its main weaknesses. Based on our evaluation, we decide that the second solution is the best for our situation. We assume that by posting on Twitter, users have agreed to share that content with the general public, other users, and Twitter partners ("Privacy Policy"); therefore, even without their explicit consent, we assume that there is no reasonable expectation of privacy. Although the second solution is more costly, it provides an extra layer of protection for our data subjects and could potentially save the client more money later down the line.

Risks and Contingencies

In the event that the data has poor quality and coverage, new data must be sourced. Data of poor quality means that there are too many null values that cannot be imputed using other means, and where deletion of these rows would result in an incomplete dataset. Data of poor coverage implies that the dataset is mostly homogenous; this is especially problematic in the training set. A model built upon this data would result in a weak predictive power not suitable for everyday predictions. New datasets can be obtained from other sources (e.g the UCI Machine Learning Repository) and other social media platforms can also be considered, such as Facebook or Instagram.

The data mining process is not linear and many processes may be revisited over several times before the project is completed. For this reason, sometimes the final deadline may be postponed if multiple iterations are required. If delays are expected, the client must be informed. Additional funding may be requested and the timeline for completion may be extended, pending approval from the project lead and client.

There is also the risk that our NLP is unable to accurately interpret the meaning of a tweet due to sarcasm. Inaccurate predictions can lead to increased police presence in vulnerable populations that perpetuate racial bias patterns and cause excessive damage to these communities. Past scandals involving bias in social media surveillance models have proven to be

PR nightmares and should be avoided. For example, the Boston Police Department (BPD) surveilled Facebook and Twitter between 2014-2016 for “Islamic extremist terminology” to identify possible terrorist attacks. The ACLU of Massachusetts later found evidence that the BPD had targeted phrases like “#MuslimLivesMatter” and “ummah”, the Arabic word for ‘community’ (Bousquet, 2018).

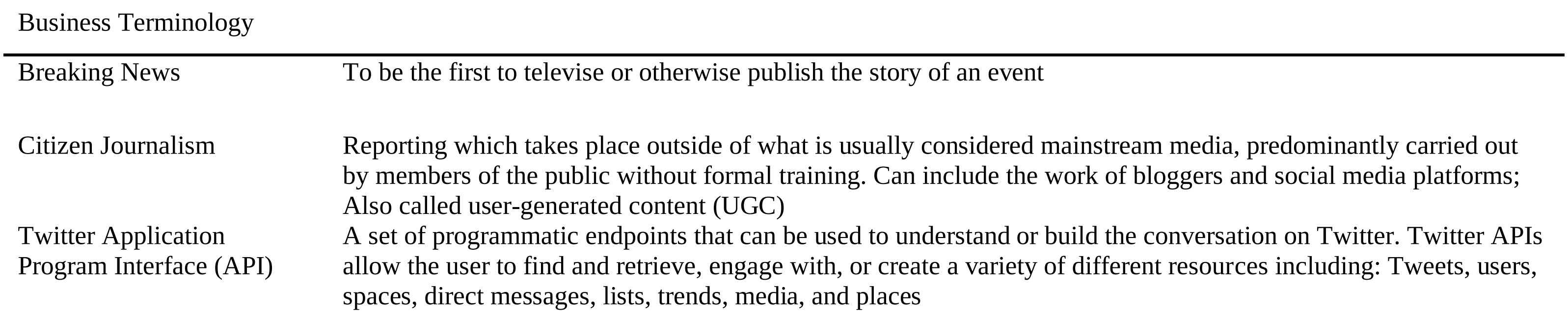
According to Sarsam et al. (2020), there are several challenges to sarcasm detection on Twitter including: (a) the added ambiguity from the limited character count of 280 characters; (b) informal uses of the language such as slang and abbreviations; and (c) the lack of a predefined structure. Sarcasm detection APIs (such as the one by Komprehend.io) can be used in conjunction with our model to best interpret tweets for its full sentiment.

This project deals with sensitive data and there are risks associated with data privacy. To mitigate a total loss of our data other any project materials in the event of a hack or other disruption, the data will be stored in a secure location, such as a virtual data room (VDR). In addition, two-factor authentication will be required and employees will only be able to access the content while connected to a company network.

Terminology

Table 1

Business Terminology



Business Terminology

Breaking News	To be the first to televise or otherwise publish the story of an event
Citizen Journalism	Reporting which takes place outside of what is usually considered mainstream media, predominantly carried out by members of the public without formal training. Can include the work of bloggers and social media platforms; Also called user-generated content (UGC)
Twitter Application Program Interface (API)	A set of programmatic endpoints that can be used to understand or build the conversation on Twitter. Twitter APIs allow the user to find and retrieve, engage with, or create a variety of different resources including: Tweets, users, spaces, direct messages, lists, trends, media, and places

Table 2

Data Mining Terminology

Data Mining Terminology	
Predictive Analytics	The use of statistics and modeling techniques to make predictions about future outcomes and performance. This project is based on the process of predictive analytics to process future Tweets.
Machine Learning	A branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. We will use machine learning to programmatically monitor Twitter for disasters as they occur.
Supervised Learning	The machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. We will utilize supervised learning algorithms based on the training dataset.
Binary classification	A supervised learning algorithm that categorizes new observations into one of two classes and is usually denoted as (1) or (0). In our project, we will be classifying tweets based on whether they are about a real disaster (1) or not (0).
Algorithm	A series of instructions telling a computer how to transform a set of facts about the world into useful information. Our models will be based on machine learning algorithms.
Training set	A portion of our actual dataset that is fed into the machine learning model to discover and learn from. Our training set consists of Tweets with known outcomes (whether or not they are about real disasters). We will be using this portion of the data to predict which Tweets in the future are about real disasters.
Test set	The portion of the dataset used to evaluate the performance and progress of the algorithm built using the training set. Our test set is the second CSV file containing Tweets that we will be classifying.
Natural Language Processing (NLP)	The branch of computer science concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. It enables computers to process human language and 'understand' its full meaning, complete wth the speaker or writer's intent and sentiment. The desired goal of this project is to create a natural language processing (NLP) model that can tell the difference in a Tweet about a literal or metaphorical disasters.
Stemming	Stemming is a natural language processing technique that lowers inflection in words to their root forms, hence aiding in the preprocessing of text, words, and documents for text normalization. We use stemming to com
Exploratory Data Analysis	The process utilized by data scientists to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods. For this project, we will perform exploratory data analysis on the training set to gain further insight into the types of Tweets that are about a real disaster.
CSV File	A text file with values separated by commas that has a specific format that allows the data to be saved in a table structured format. Our dataset (training and test set) is saved as two separate CSV files.

Data Mining Goals and Success Criteria

The data mining problem type is predictive, with a goal of predicting which tweets are about a real disaster and which ones aren’t. To achieve this goal, a model utilizing natural language processing (NLP) is needed to judge the contents of a tweet to understand its intent and sentiment. To predict which tweets are about a real disaster or not, we consider the following predictive analytics solutions:

(a) User prediction—Identify which Twitter users are active citizen journalists/bloggers and most likely to have accurate information regarding a disaster or event

(b) Tweet prediction—Predict which tweets are most likely to be about a real disaster based on the content of the Tweet

(c) Location prediction—Consider which locations are associated with specific disasters and make predictions based on the proximity of the location of the Tweet to the disaster mentioned. For example, those located close to the San Andreas Fault (such as California) will experience high frequencies of earthquakes versus those in the Midwest will experience higher frequencies of tornados. For general accidents, consider urban areas with higher population densities.

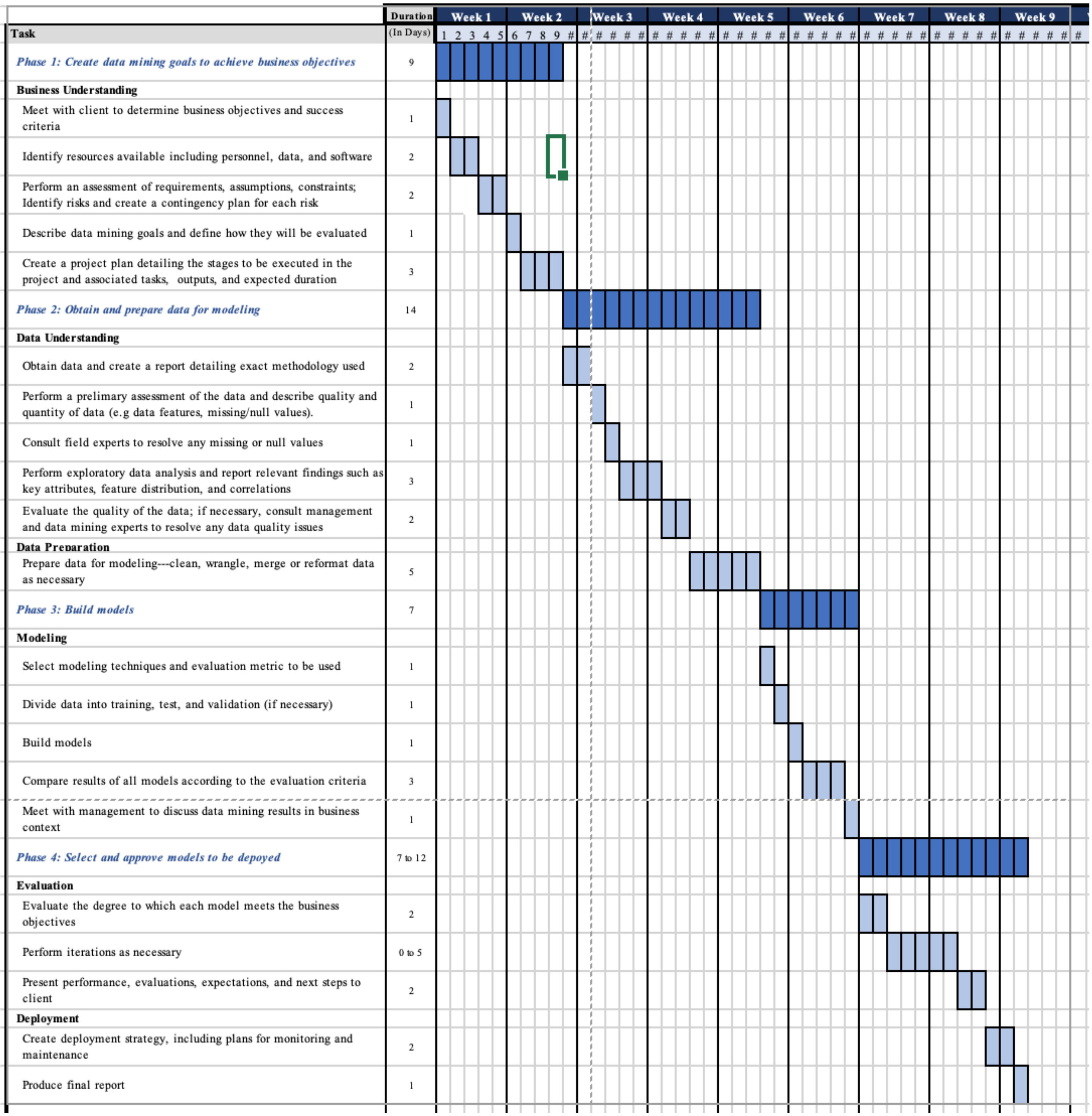
Each proposed solution requires different data, so the availability of the data will give further insight into which method will work best moving forward.

The data mining goals are as follows: (a) Identify which keywords or phrases are associated with a real disaster and (b) Develop a model to predict which tweets are associated with a real disaster (1) or not (0). The model should produce a prediction score and a threshold that converts this score into one of the target levels. The model will be assessed using a receiver operating characteristic index (ROC Index), with a value of above 0.7 to indicate a strong model (Kelleher, 2020).

Project plan/ Order of tasks

Figure 3

Gantt Chart



Task	Duration (In Days)
Phase 1: Create data mining goals to achieve business objectives	9
Business Understanding	
Meet with client to determine business objectives and success criteria	1
Identify resources available including personnel, data, and software	2
Perform an assessment of requirements, assumptions, constraints; Identify risks and create a contingency plan for each risk	2
Describe data mining goals and define how they will be evaluated	1
Create a project plan detailing the stages to be executed in the project and associated tasks, outputs, and expected duration	3
Phase 2: Obtain and prepare data for modeling	14
Data Understanding	
Obtain data and create a report detailing exact methodology used	2
Perform a prelimary assessment of the data and describe quality and quantity of data (e.g data features, missing/null values).	1
Consult field experts to resolve any missing or null values	1
Perform exploratory data analysis and report relevant findings such as key attributes, feature distribution, and correlations	3
Evaluate the quality of the data; if necessary, consult management and data mining experts to resolve any data quality issues	2
Data Preparation	
Prepare data for modeling---clean, wrangle, merge or reformat data as necessary	5
Phase 3: Build models	7
Modeling	
Select modeling techniques and evaluation metric to be used	1
Divide data into training, test, and validation (if necessary)	1
Build models	1
Compare results of all models according to the evaluation criteria	3
Meet with management to discuss data mining results in business context	1
Phase 4: Select and approve models to be depoyed	7 to 12
Evaluation	
Evaluate the degree to which each model meets the business objectives	2
Perform iterations as necessary	0 to 5
Present performance, evaluations, expectations, and next steps to client	2
Deployment	
Create deployment strategy, including plans for monitoring and maintenance	2
Produce final report	1

Data Understanding

Initial Data Collection Report

We will be using the training set (train.csv) and test set (test.csv) files. Table 3 shows the available fields for our dataset.

Table 3

Dataset Fields

Fields	
id	a unique identifier for each tweet
text	the text of the tweet
location	the location the tweet was sent from (may be blank)
keyword	a particular keyword from the text (may be blank)
target	**in train.csv only,** this denotes whether a tweet is about a real disaster (1) or not (0)

Note. From Kaggle "Natural Language Processing with Disaster Tweets" Competition Overview

https://www.kaggle.com/competitions/nlp-getting-started/overview

In the training set, there are 7613 Tweets, of which there are 61(0.8%) with missing keywords 2533 (33%) with missing locations. It may be possible to manually infer missing keywords for the 61 tweets from the 'text' field; however, this would not be plausible for tweets with missing locations. Tweets can be made globally by anyone with a device connected to the internet and it would be extremely difficult to pinpoint the Tweet to one location. 'Location' field will have to be investigated further to determine if it is correlated with the 'target' field and thus useful in the model. It may be possible to remove 'location' altogether. Although we would lose a parameter for the final model, it is a better option than removing 33% of all tweets.

In the test set, there are 3263 Tweets, of which there are 26 (0.8%) with missing keywords and 1105 (33%) with missing locations. Missing values can be extrapolated in the same way as those in the training set mentioned above.

Both datasets contain free text entries in the following fields: ‘text’, ‘location’, and ‘keyword’. In the training set, there are 3341 unique locations, 221 unique keywords, and 7503 unique tweets. In the test set, there are 1602 unique locations, 221 unique keywords, and 3243 unique tweets. It is necessary to resolve for misspellings as these posts were made on a social media platform and there are bound to be mistakes if we consider that many of these tweets are made during an actual event.

Based on these preliminary findings, data quality and quantity appears to be sufficient to build a feasible model.

Data Description Report

‘id’ is a numerical field ranging from 1 to 10873 in the training set and from 0 to 10875 in the test set. Each tweet has a unique id but the numbers are not continuous (some numbers are skipped). The training set contains an additional field labeled ‘target’ that contains two target levels. A (1) indicates that a tweet is about a real disaster while (0) indicates that a tweet is not about a real disaster. Both training set and test set contain duplicate tweets under different ‘id’ fields, found by matching the contents of the tweet using the ‘text’ field. There are 110 duplicates in the training set, while the test set has 20. For instances of duplicates, all duplicates will be deleted except for the first entry (the instance with the smallest ‘id’ value). Figure 4 and 5 depict the word clouds for keywords in the training and test set, respectively (Wordclouds were generated using the stylecloud package in Python). Keywords are written in a size relative to their frequency, with the largest words shown the biggest.

Figure 4

Wordcloud of Keywords in Training Set

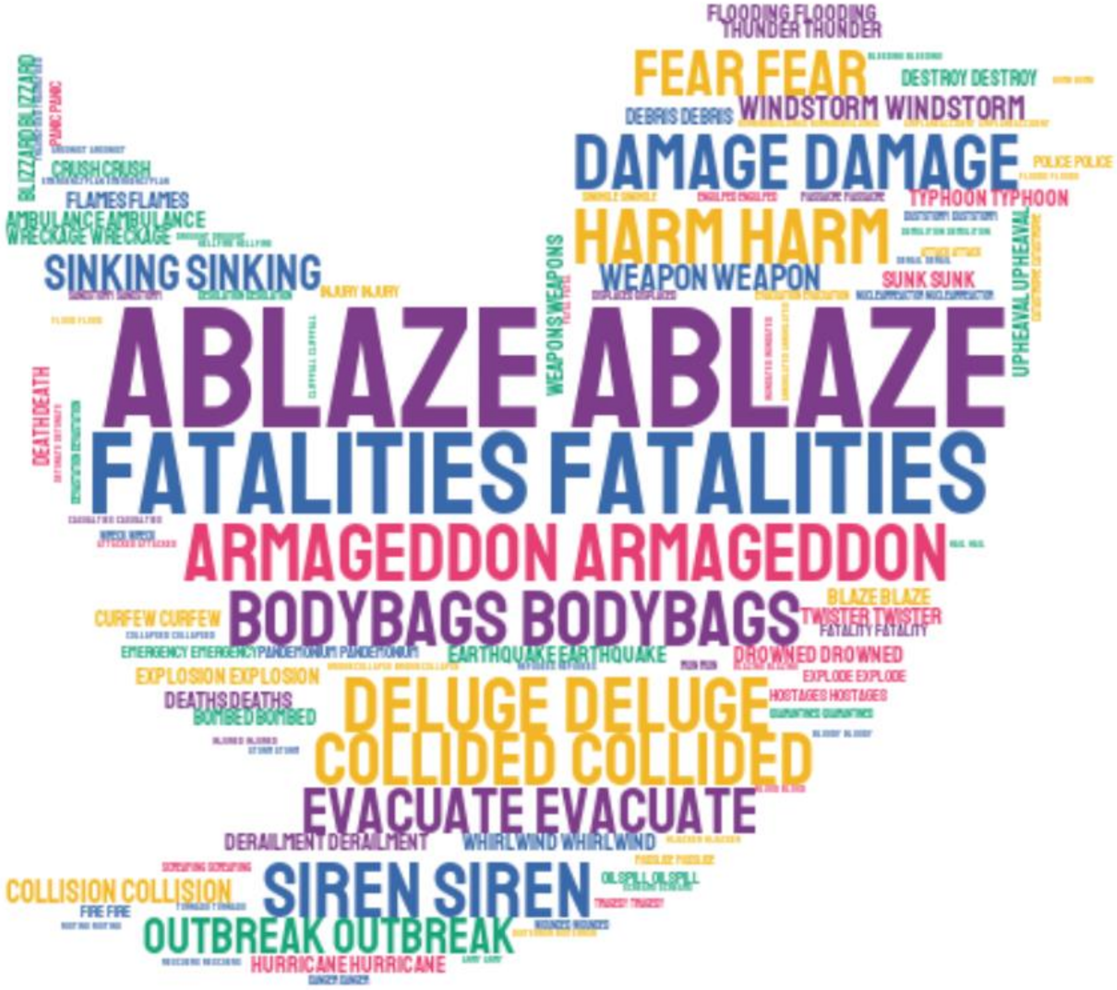
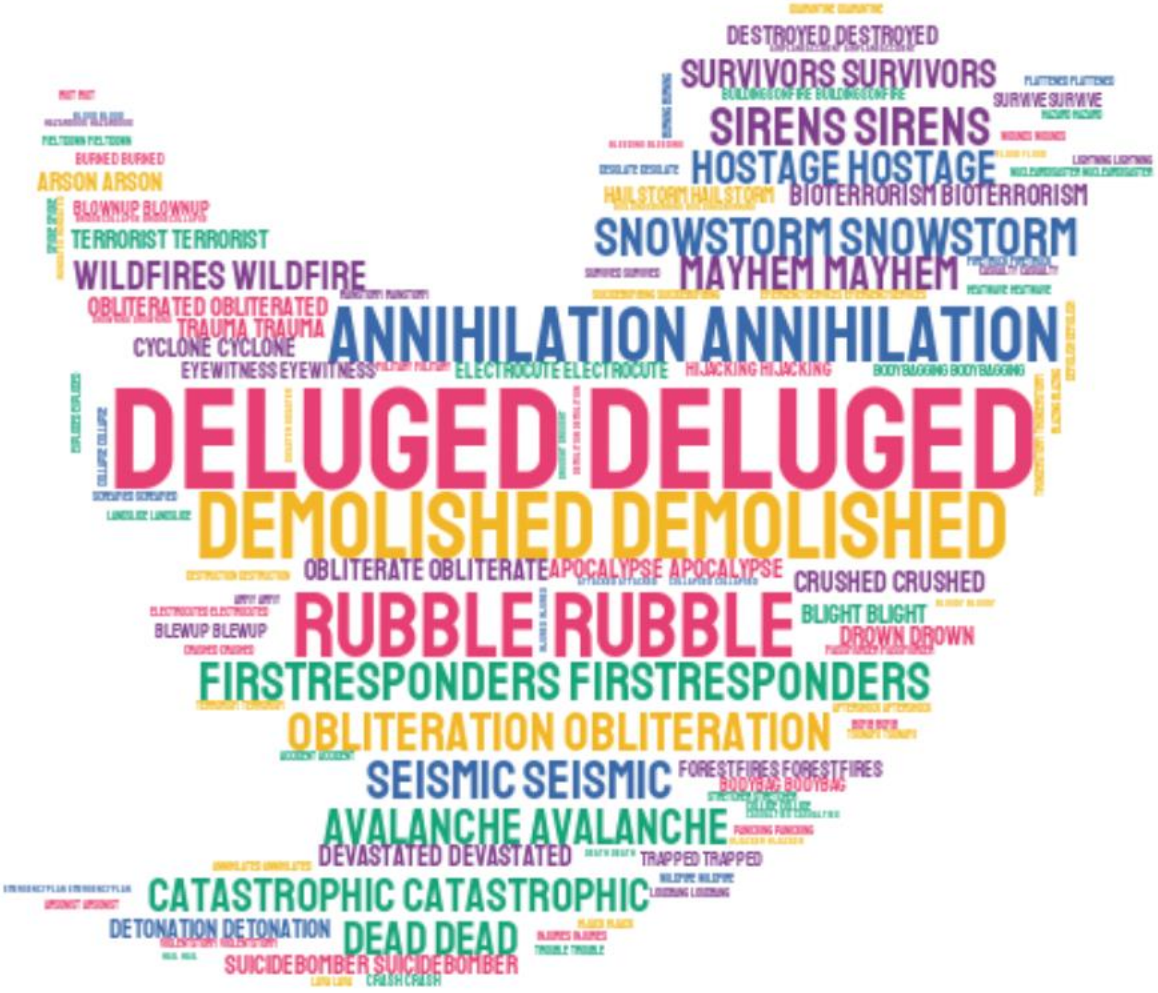


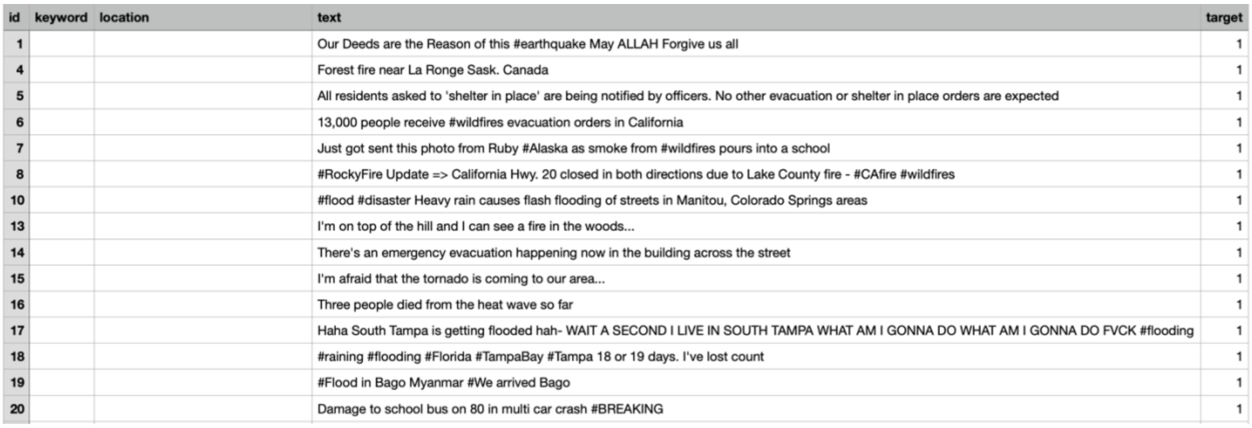
Figure 5

Wordcloud of Keywords in Test Set



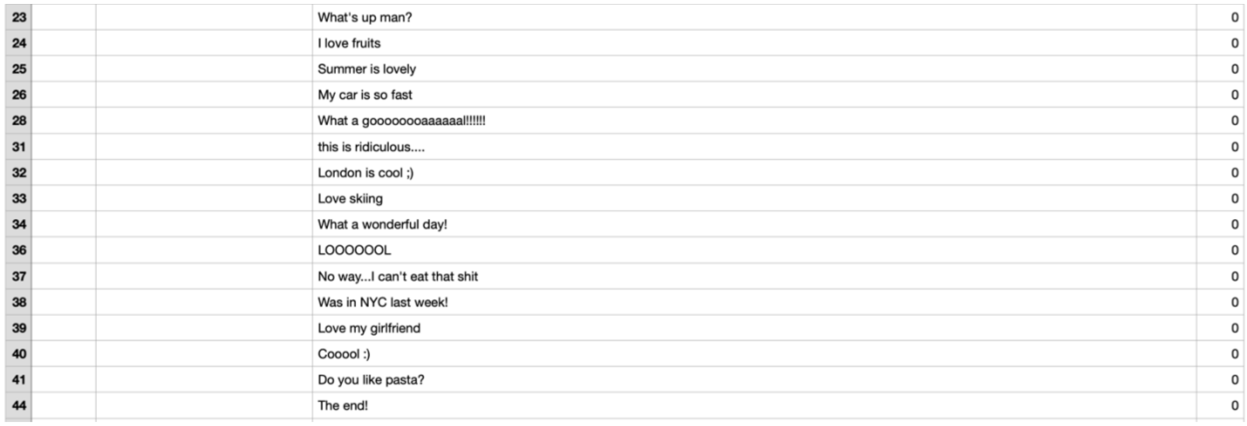
Null values for 'keyword'

In the training set, there 61 tweets with missing keywords. For the following tweets, the keywords can be imputed by matching keywords from other tweets with the content of the tweet itself. Some tweets use more than one keyword, so there must be a method to decide which category to impute (for example, id 10 contains both ‘flood’ and ‘disaster).



id	keyword	location	text	target
1			Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all	1
4			Forest fire near La Ronge Sask. Canada	1
5			All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected	1
6			13,000 people receive #wildfires evacuation orders in California	1
7			Just got sent this photo from Ruby #Alaska as smoke from #wildfires pours into a school	1
8			#RockyFire Update => California Hwy. 20 closed in both directions due to Lake County fire - #CAfire #wildfires	1
10			#flood #disaster Heavy rain causes flash flooding of streets in Manitou, Colorado Springs areas	1
13			I'm on top of the hill and I can see a fire in the woods...	1
14			There's an emergency evacuation happening now in the building across the street	1
15			I'm afraid that the tornado is coming to our area...	1
16			Three people died from the heat wave so far	1
17			Haha South Tampa is getting flooded hah- WAIT A SECOND I LIVE IN SOUTH TAMPA WHAT AM I GONNA DO WHAT AM I GONNA DO FVCK #flooding	1
18			#raining #flooding #Florida #TampaBay #Tampa 18 or 19 days. I've lost count	1
19			#Flood in Bago Myanmar #We arrived Bago	1
20			Damage to school bus on 80 in multi car crash #BREAKING	1

The following tweets do not mention a disaster, real or otherwise. They also do not have a value for location or keyword.



23			What's up man?	0
24			I love fruits	0
25			Summer is lovely	0
26			My car is so fast	0
28			What a goooooooaaaaaal!!!!!!	0
31			this is ridiculous....	0
32			London is cool ;)	0
33			Love skiing	0
34			What a wonderful day!	0
36			LOOOOOOL	0
37			No way...I can't eat that shit	0
38			Was in NYC last week!	0
39			Love my girlfriend	0
40			Cooool :)	0
41			Do you like pasta?	0
44			The end!	0

Data Exploration Report

Data Cleaning

Before the data can be explored, several findings in the data description report need to be addressed. First, duplicates from both the training set (110) and test set (20) were deleted. For the training set, missing keywords for tweet id 1-20 were imputed using keyword categories found in the text. For tweets with more than one keyword (id 7, 10, 20), the more specific keyword is chosen (e.g. id 10 contains both “flood” and “disaster”, and “flood” is chosen as the imputed value). The completed rows, with the chosen keyword in the text bolded, are as follows:

id	keyword	location	text	target
1	earthquake		Our Deeds are the Reason of this #**earthquake** May ALLAH Forgive us all	1
4	forest%20fire		**Forest fire** near La Ronge Sask. Canada	1
5	evacuation		All residents asked to 'shelter in place' are being notified by officers. No other **evacuation** or shelter in place orders are expected	1
6	wild%20fires		13,000 people receive #**wildfires** evacuation orders in California	1
7	wild%20fires		Just got sent this photo from Ruby #Alaska as smoke from #**wildfires** pours into a school	1
8	wild%20fires		#RockyFire Update => California Hwy. 20 closed in both directions due to Lake County fire - #CAfire #**wildfires**	1
10	flood		#flood #**disaster** Heavy rain causes flash **flooding** of streets in Manitou, Colorado Springs areas	1
13	fire		I'm on top of the hill and I can see a **fire** in the woods...	1
14	evacuation		There's an emergency **evacuation** happening now in the building across the street	1
15	tornado		I'm afraid that the **tornado** is coming to our area...	1
16	heat%20wave		Three people died from the **heat wave** so far	1
17	flooding		Haha South Tampa is getting **flooded** hah- WAIT A SECOND I LIVE IN SOUTH TAMPA WHAT AM I GONNA DO WHAT AM I GONNA DO FVCK #**flooding**	1
18	flooding		#raining #**flooding** #Florida #TampaBay #Tampa 18 or 19 days. I've lost count	1
19	flood		#**Flood** in Bago Myanmar #We arrived Bago	1
20	crash		Damage to school bus on 80 in multi car **crash** #BREAKING	1

Tweets id 23-44 are deleted since they do not mention any key words and where the sentiment is clearly positive. The training set now has 7487 tweets, and the test set has 3242 tweets.

Hypothesis

The null hypothesis is that a tweet cannot be characterized as being about a real disaster or not by any of the tweet characteristic; that is, the target level (1 or 0) cannot be determined by any of the tweet fields (‘id’, ‘keyword’, ‘location’, or ‘text’). Exploratory data analysis will be conducted in order to determine if any characteristics can be used to predict target level.

Data Exploration

For further analysis of the body of the tweets, an additional column for the character count is added to the dataframe. Figure 6 shows the distribution of character counts across all tweets for the training set. Most tweets are under 140 characters, which was Twitter’s character limit prior to November 8th , 2017 (Boot et al., 2019).

Figure 6

Histogram of Character Count per Tweet in Training Set

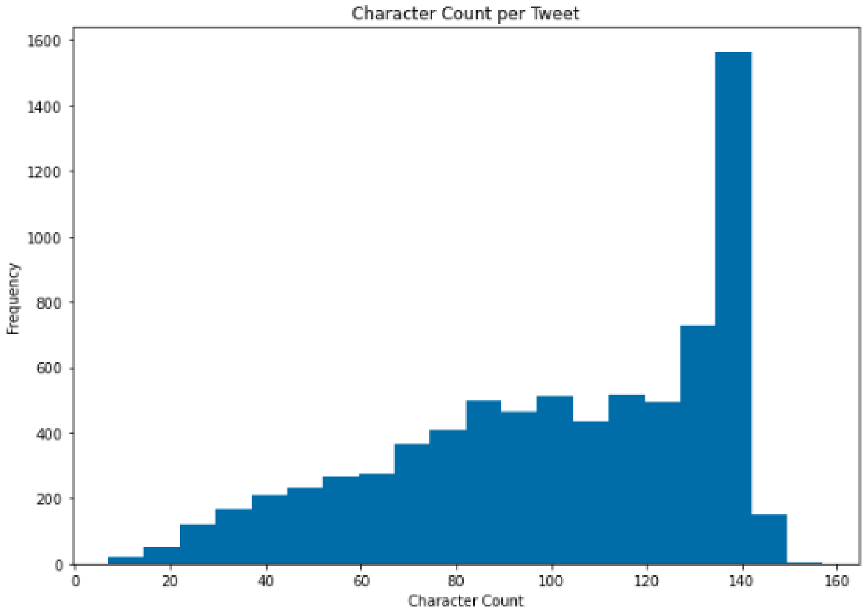
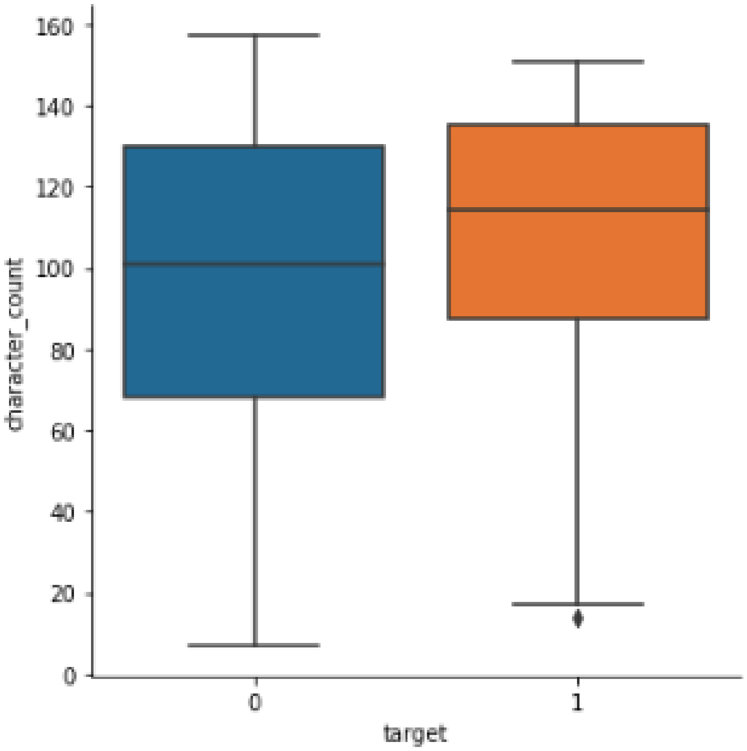


Figure 7 depicts the distribution of the character counts for real disasters and non-real disasters as a boxplot. The level 1 target (indicating real disasters) has a greater median character count and less spread than level 0 target.

Figure 7

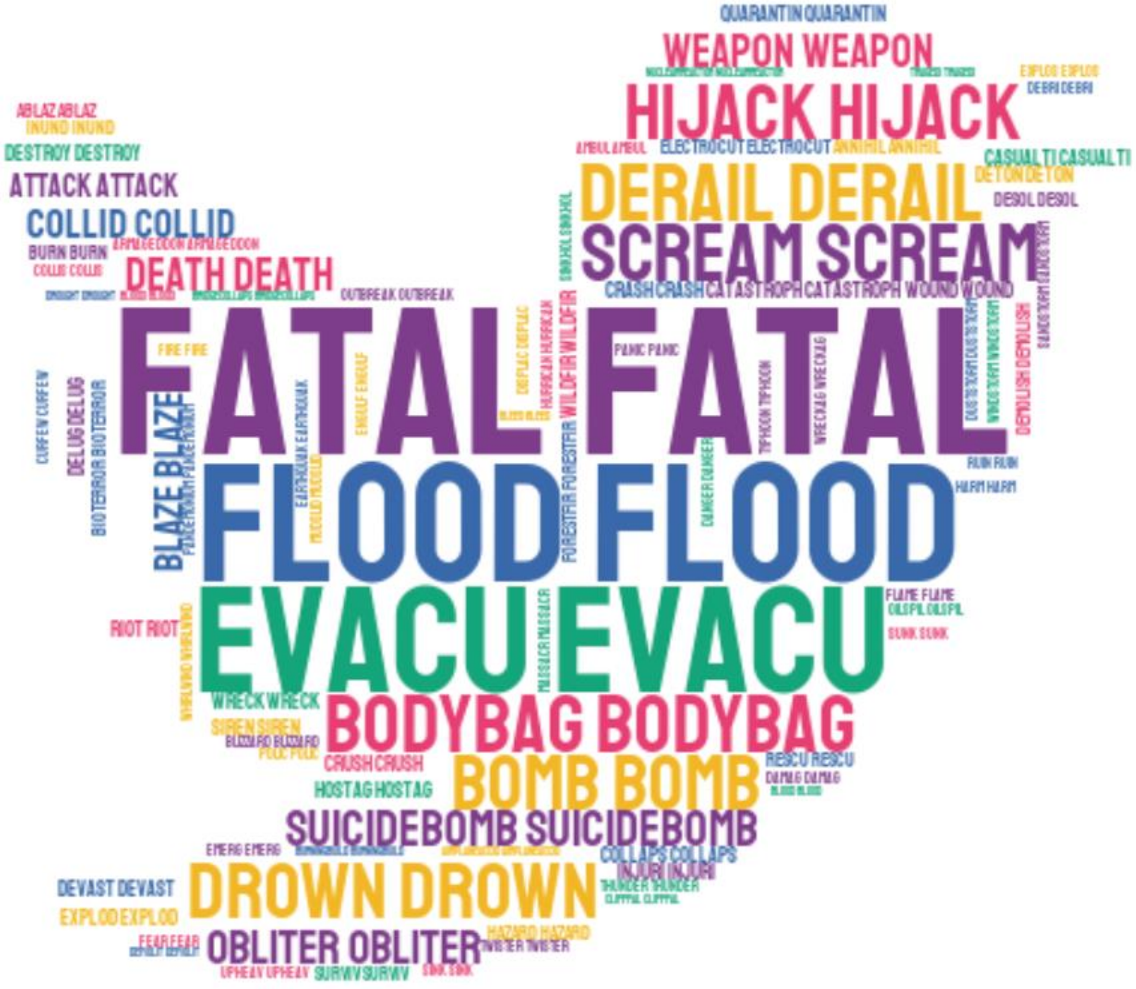
Boxplot of Character Count for each Target Level in the Training Set



Compounded keywords containing a space encoded with “%20” were removed, leaving the full word. Next, stems of keywords were added as another column in the dataframe. Stemming is a natural language processing technique of extracting the root forms of words. This was done using PorterStemmer from the Natural Language Toolkit (nltk). Stemming of the keywords effectively reduced the number of categories from 221 to 165 by combining many variations of the same disaster. For example, ‘flood’, ‘flooding’, ‘floods’ can all be grouped together using the stem ‘flood’. Figure 8 depicts the wordcloud for stems used in the training set.

Figure 8

Wordcloud of Keyword-Stems in Training Set



Out of 165 unique stems in the training set, the most commonly used were: fatal (*N=117)*, evacu (*N=111)*, flood (*N=111)*, bodybag (*N=107)*, derail (*N=106)*, scream (*N=105)*, drown (*N*=104), bomb (*N*=100), hijack (*N*=98), and suicidebomb (*N*=97). Table 4 and Figure 9 depict the 10 most commonly used stems in the training set and the proportion of usage for both target levels. Stems with the highest percentage in target level 1 are: suicidebomb (74.23%, n=72), flood (73.87%, n=82), and bomb (73%, n=73); stems with the highest percentage in target level 0 are: bodybag (82.24%, n=88), scream (80.00%, n=84), and fatal (63.25%, n=74). "Hijack-" is used almost equally in both instances, with a slight preference for non-disasters (51.02%, n=50).

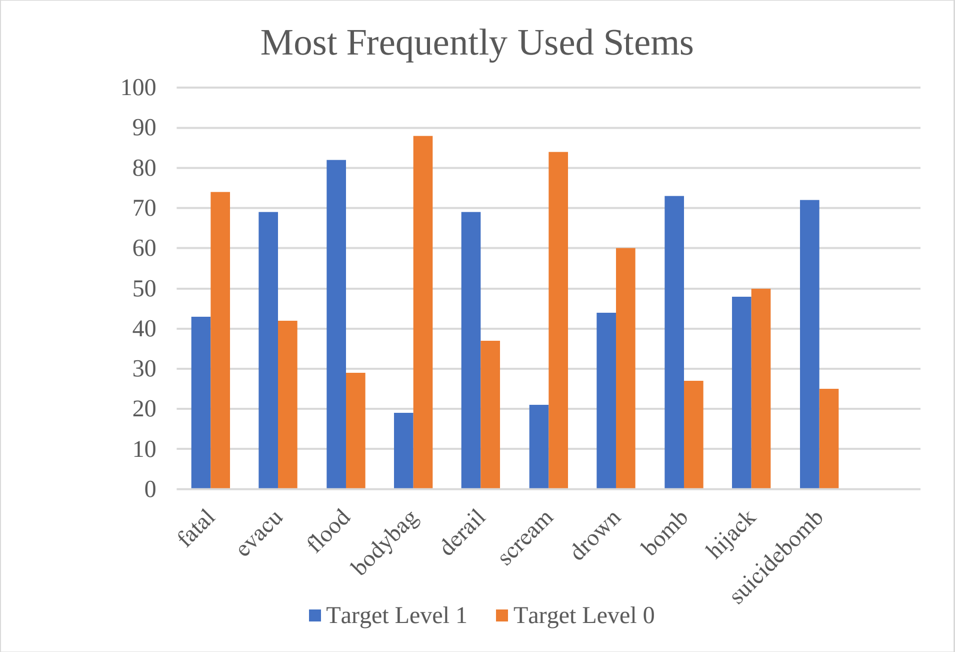
Table 4

Ten Most Commonly Used Stems in Training Set

Stem	Frequency	Target Level 1		Target Level 0	
	N	*n*	%	*n*	%
fatal	117	43	36.75%	74	63.25%
evacu	111	69	62.16%	42	37.84%
flood	111	82	73.87%	29	26.13%
bodybag	107	19	17.76%	88	82.24%
derail	106	69	65.09%	37	34.91%
scream	105	21	20.00%	84	80.00%
drown	104	44	42.31%	60	57.69%
bomb	100	73	73.00%	27	27.00%
hijack	98	48	48.98%	50	51.02%
suicidebomb	97	72	74.23%	25	25.77%

Figure 9

Barchart of Ten Most Commonly Used Stems



The most frequently used stems for real disasters were (with percentages of real disasters compared to the total number of tweets): flood (n = 82, 73.87%), bomb (n= 73, 73.00%), suicidebomb (n= 72, 74.23%), derail (n= 69, 65.09%), evacu (n= 69, 62.16%), and death (n=57, 77.03%). The most frequently used stems for non-real disasters were (with percentages of non-real disasters compared to the total number of tweets): bodybag (n=88, 82.24%), scream (n=84, 80.00%), fatal (n=74, 63.25%), blaze (n=67, 93.06%), and drown (n=60, 57.69%).

Table 5 shows the top 5 stems with the highest percentage of usage for real disasters and non-disasters. Tweets using “oilspil”, “structuralfailur”, “rainstorm”, “naturaldisast”, or “bushfir” were associated with a real disaster over 80% of the time, with “oilspil” corresponding to a real disaster 97.30% of the time. On the other hand, tweets using stems “rubbl”, “blownup”, “blewup”, “trauma”, or “blaze” had the highest rate of correspondence with metaphorical disasters. The higher percentages for stems associated with target level 0 suggests that a working model may have better accuracy predicting non-disasters than real disasters.

Table 5

Top 5 Stems with the Highest Percentage of Usage for Real Disasters and Non-Disasters

Stem	Frequency *N*	Target Level 1 *n* (%)
oilspil	37	36 (97.30%)
structuralfailur	31	28 (90.32%)
rainstorm	34	30 (88.24%)
naturaldisast	34	29 (85.29%)
bushfir	25	20 (80.00%)
Stem	Frequency *N*	Target Level 0 *n* (%)
rubbl	28	27 (96.43%)
blownup	33	31 (93.94%)
blewup	33	31 (93.94%)
trauma	31	29 (93.55%)
blaze	72	67 (93.06%)

Data Quality Report

Our training set has 4289 (57.29%) tweets with a target level of 0 and 3198 (42.71%) tweets with a target level of 1. Although there is a slight bias for non-real disasters, it is not enough to cause classification accuracy. While missing keywords have been resolved, 2533 (33%) tweets are still missing locations. We can label these tweets with an '*Unknown'* location and proceed with data analysis rather than deleting location as a field altogether. Figure 4 showed that some tweets are over 140 characters. These tweets will have to be further investigated to make sure that the count is correct. It may be necessary to determine when the dataset was collected, and whether this was before or after the character limit change in 2019.

References

Automation Rules. (2017). *Twitter*. https://help.twitter.com/en/rules-and-policies/twitter-automation

Auxier, B. & Anderson, M. (2021). Social Media Use in 2021. *Pew Research Center*. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Barthel M., Grieco E., Shearer E. (2019). Older Americans, Black Adults and Americans With Less Education More Interested In Local News. *Pew Research Center.* https://www.pewresearch.org/journalism/2019/08/14/older-americans-black-adults-and-americans-with-less-education-more-interested-in-local-news/

Boot, A.B., Tjong Kim Sang, E., Dijkstra, K. et al. (2019). How character limit affects language usage in tweets. *Palgrave Commun* 5 (76). https://doi.org/10.1057/s41599-019-0280-3

Bousquet, C. (2018). Mining Social Media Data for Policing, the Ethical Way. *Harvard University Data-Smart City Solutions.* https://datasmart.ash.harvard.edu/news/article/mining-social-media-data-policing-ethical-way

Developer Agreement and Policy. (2020). *Twitter. https://developer.twitter.com/en/developer-terms/agreement-and-policy*

Display Requirements: Tweets. *Twitter. https://developer.twitter.com/en/developer-terms/display-requirements*

Ethics and Data Protection. *European Commission* https://ec.europa.eu/info/sites/default/files/5._h2020_ethics_and_data_protection.pdf

More about the restricted uses of the Twitter APIs. *Twitter.* https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.

Natural Language Processing with Disaster Tweets. *Kaggle.* https://www.kaggle.com/competitions/nlp-getting-started/overview

Kelleher, J. Mac Namee, B. & D'Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked, Examples, and Case Studies* (2nd ed.). Massachusetts Institute of Technology.

Leadership. *NBCUniversal*. https://www.nbcuniversal.com/leadership

Privacy Policy. *Twitter.* https://twitter.com/en/privacy

Sarsam S.M., Al-Samarraie H., Alzahrani A.I., & Wright B. (2020). Sarcasm detection using machine learning algorithms in Twitter: A systemic review. *Sage Journals* 62 (5). https://doi.org/10.1177/1470785320921779

Vakkuri, V., Kemell, KK. (2019). Implementing AI Ethics in Practice: An Empirical Evaluation of the RESOLVEDD Strategy. In: Hyrynsalmi, S., Suoranta, M., Nguyen-Duc, A., Tyrväinen, P., Abrahamsson, P. (eds) Software Business. ICSOB 2019. Lecture Notes in Business Information Processing, vol 370. Springer, Cham. https://doi.org/10.1007/978-3-030-33742-1_21

Twitter Usage Statistics. *Internet Live Stats*. https://www.internetlivestats.com/twitter-statistics/