Topic Modeling for Disaster Identification in Tweet Dataset

Using TF-IDF Vectorization and Non-Negative Matrix Factorization to model topics for identification of disaster related content

Megan Arnold
DTSA 5510 Final
June 5th, 2025

Problem Overview

- Social media is full of real-time data. Most of it is noise, some of it is meaningful
- Emergency responders and state officials need to isolate accurate disaster-related information quickly
- Challenge: Labeled social media data is rare
- Challenge: Class imbalance of disaster-related tweet vs non-disaster related tweet is high
- Challenge: Large influx of data means the data processing pipeline must have small overhead.

Dataset Balance:

```
Dataset Balance (Percent of positive samples):

0.1859278803869833


Count of Positive and Negative Samples:

target
0    9256
1    2114
Name: count, dtype: int64
```

Non-Relevant Tweet Examples:

```
176                       Grover Airplane Accident Doctor
216                             “ambulance nurses” ❤❤❤
226                               You must be annihilated!
266                               Thot status: annihilated
275                               Thot Status: Annihilated
287                              i want u like annihilation
303                   ANNIHILATION https://t.co/3QhIwn016i
312                             Annihilation! One a my favs
320                                    annihilation on hulu
330     Euroleague Bet365 µε 8 units https://t.co/btn3KLIumB
```

Project Goals

- Explore tweet dataset to **understand cleaning requirements** and guide model selections
- **Vectorize natural language** data for use in unsupervised machine learning model
    - **TF-IDF** - Term Frequency - Inverse Document Frequency
    - **Embedding** - Semantic Encoding
- Evaluate different vectorization techniques in their ability to pair with the unsupervised topic modeling technique of **Non-Negative Matrix Factorization**
- **Evaluate hyperparameter** tuning outcomes on post-hoc analysis
- Make inference pipeline **recommendations**

Dataset Overview

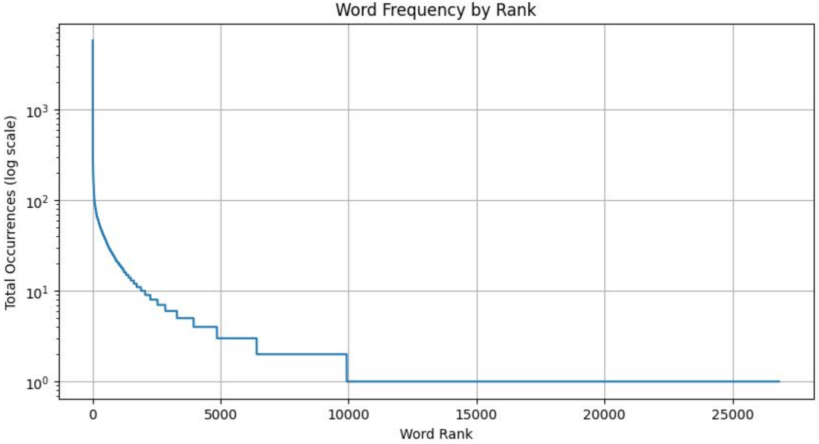
- 11370 tweets, labeled as either **disaster-related(1) or not-disaster related(0)**
- Class **imbalance** of 18.6% of the positive class
- Features:
  - Tweet Text
  - Keyword that triggered the crawler
  - Location
  - Label (only used for post-hoc evaluation)

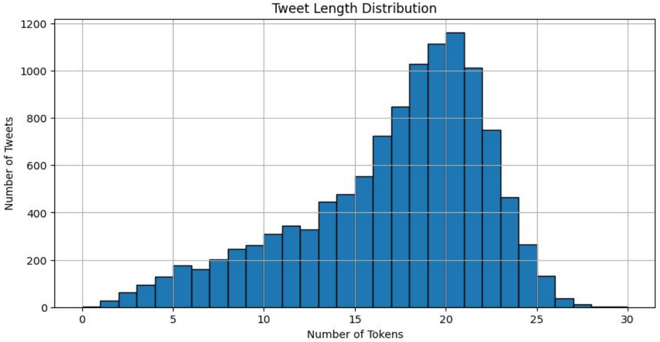
Tweet Examples:

- Terms in A Demon Burning Dark: The Ruined: People who cannot use magic or interact with it without some harm or ac… https://t.co/ZEDawOfuu4
- 📷 Heartfelt appreciation to Prime Minister YAB Tun Dr. wife, YABhg. Tun Dr. Siti Hasmah Mohd Ali fo… https://t.co/YOwUp1BYUP
- #WATCH Former CM Akhilesh Yadav who went to meet injured of Kannauj accident, at a hospital in Chhibramau asks Emergency Med…
- ❤️❤️❤️ he gave us everything... He had a horrible foot infection once so wore one thong… https://t.co/mA9sFl6Shw
- 😁yeah! His new swag is on point 100%, since the accident! Like this is a totally transformed Bob…
- This is cool and all these days I have been doing "git push origin CURRENT_BRANCH_NAME". You know that… https://t.co/mr0YAGEWqj
- #Preorder #newrelease today! 12 witnesses connected to or investigating #THENUTCRACKERCONSPIRACY have died either in a…
- my back and neck are still fucked up from the accident 😡😡😭😭
- RT! Prince Harry just confirmed that his mother's (Princess Diana) death was not an accident! https://t.co/1ADe3uZ3eR
- Note to Democrats: It's not a Muslim ban. Islam is not a race. Soleimani was a terrorist & was exterminated not assas…
- Juwan Johnson/Oregon is one big dude. Looks like a tight end stuck in the receiver group by accident.
- More appearances of the man with the upside-down face. A New Year's Eve party at an Air Force base in 1943 where a man…
- The speeding car rammed into a group of people, who were returning after attending a temple festival of Ayyappan Kavu in Thum…
- My friend (an army) just lost her father in an accident and her mom right now is still unconscious. Please pray for her mo…
- MLINDO THE VOCALIST IN ANOTHER CAR ACCIDENT https://t.co/BXR9rEgAk6
- "There are no greater treasures than the highest human qualities such as compassion, courage and hope. Not even tragic accide…
- Please help our friends in - they have had a non fault accident that's resulted in their vehicle being writte…
- When you hurt your younger sibling by "accident" https://t.co/DAMTEoQtZU
- David Cameron's decision to hold a referendum expressed in the medium of a road traffic accident. https://t.co/XATYHFtXoa

Exploratory Data Analysis

- Most tweets were less than 30 words
- Vocabulary follows **Zipf's Law** in Word Rank Frequency Distribution Chart
- Log-scale term frequency plot showed elbow around 1500 words
- Decisions:
  - Used 1500 words as the number of components for TF-IDF vectorization

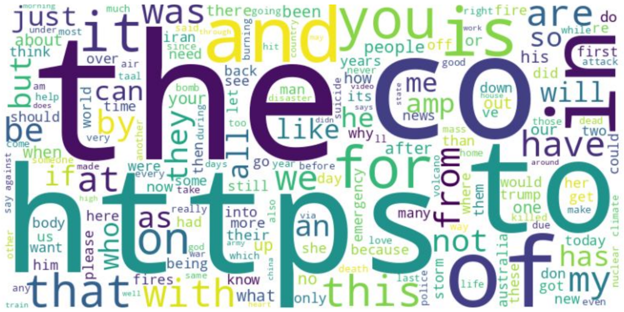




Preprocessing and Cleaning

- **Cleaned** raw tweet data:
    - Removed URLs
    - Removed Emojis
    - Removed Foreign Characters
    - Removed English Stopwords
    - Removed Foreign Characters
    - Removed Numbers
- Dropped **Keyword** and **Location** features
    - Keyword features only indicated the keyword present in the tweet. Using TF-IDF vectorizor encodes this information
    - Location feature wasn't standardized, nor did it provide context on the semantic meaning of the tweet

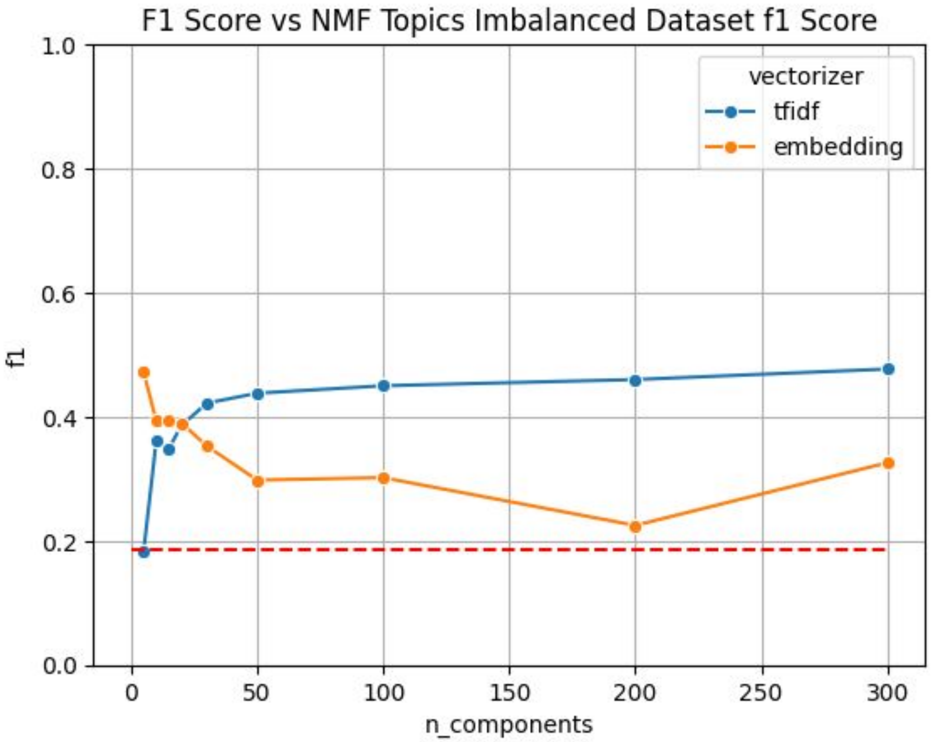
Before Cleaning



After Cleaning

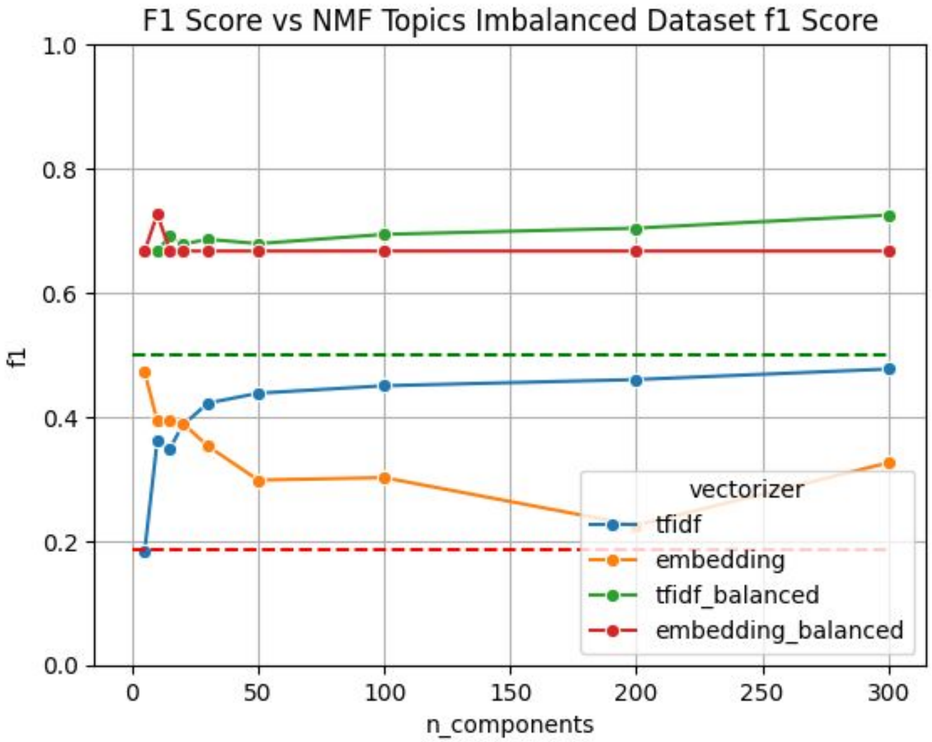
Feature Engineering

- TF-IDF Vectorization
  - TF-IDF vectorization creates a sparse matrix
  - The features are term/ngram level
  - Additive feature matrix
- Embedding
  - Embedding vectorization creates a dense matrix
  - The vectors are the semantic meaning of the text, not the additive components within the data
  - Embedding dimensions don’t directly relate to a word or phrase, but a semantic meaning



Model and Hyperparameter Tuning

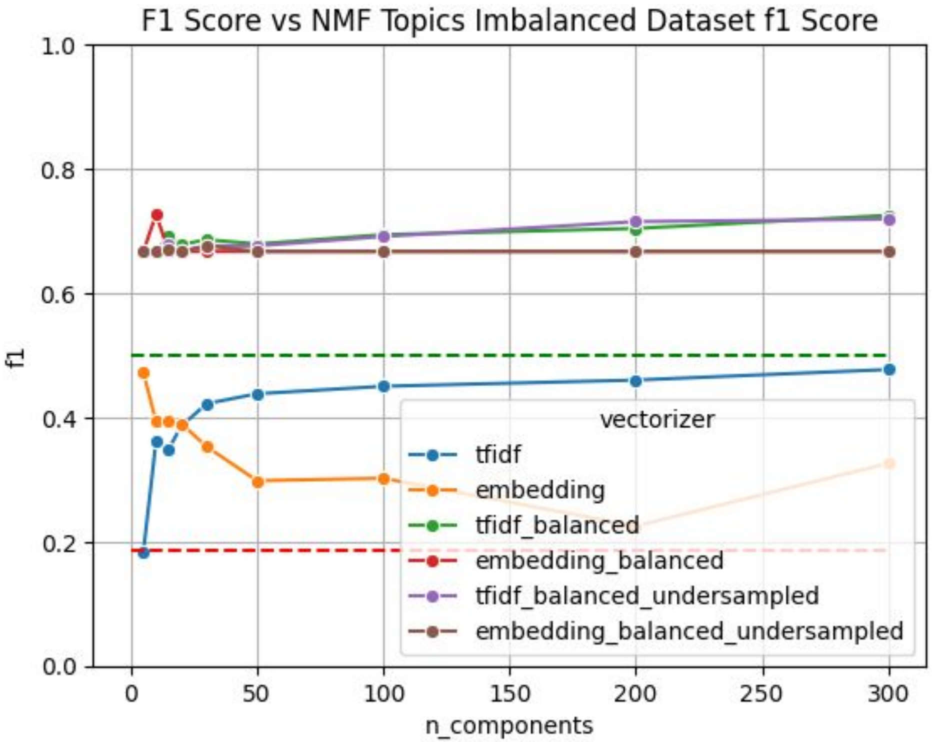
- Non-Negative Matrix Factorization
  - Tuned Parameters:
    - n_topics
- TF-IDF Vectorization Tuning
  - Tuned Parameters:
    - max_df
    - min_df
    - ngram range
    - n_components
    - token pattern
    - stop words



Class Imbalance

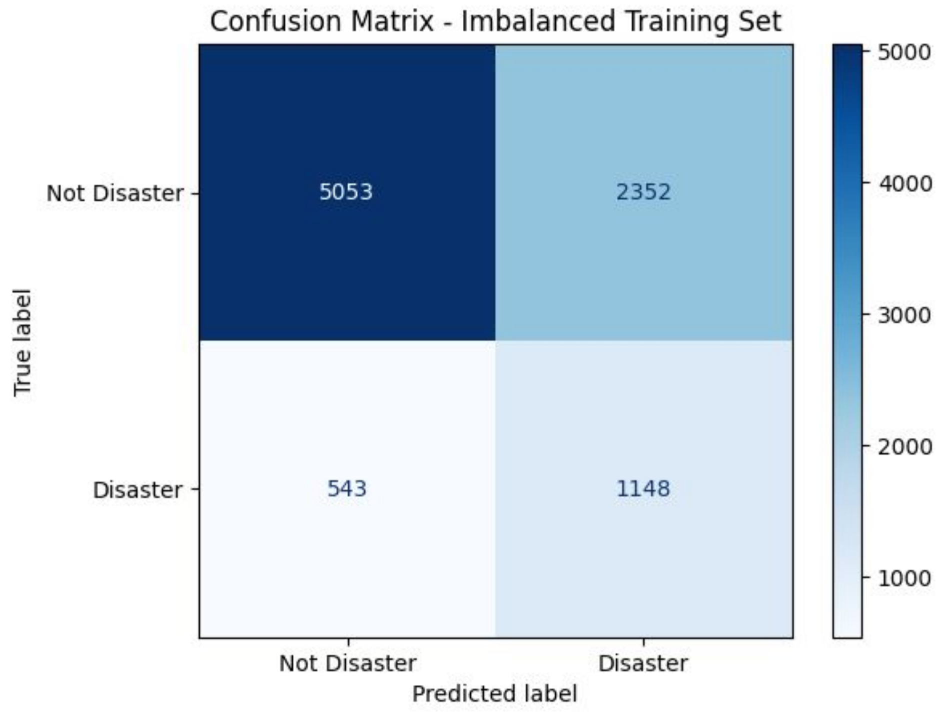
- Original Dataset
  - 18.6% positive class
- Undersample Negative Class
- Oversample Positive Class

Method	F1 Score Baseline	F1 Score Model	F1 Percent Improvement	F1 Score Improvement
Unbalanced	18.6	44.5	139%	25.9
Oversampled	50.0	68	36%	18
Undersampled	50.0	68	36%	18



Final Model Results

- Final Mode:
  - TF-IDF
    - min_df=5
    - max_df=0.75
    - bigrams
  - NMF
    - n_topics = 50
- Evaluation Results
  - Imbalanced Dataset
    - F1 Score = 0.445
    - Baseline = 0.186
  - Balanced Dataset
    - F1 Score = 0.68
    - Baseline = 0.5



Recommendations, Use Cases, and Next Steps

- Pipeline Steps:
  - Create TF-IDF Feature vectors in real time
  - Use NMF to determine most likely topic
  - Preprocessing and Inference is quick and can be performed in batches to reduce the resources per tweet
- Benefits:
  - Doesn't require labeled data
  - Fast Inference won't add to the pipeline lag
  - Can be embedded in current data ingestion pipelines

- Next Steps:
  - Investigate feasibility on larger dataset
  - Augment TF-IDF and NMF pipeline with embedding and similarity clustering pipeline for ensemble model
  - Augment with anomaly detection to identify locations that are showing higher than expected rates of disaster-topic tweets. This would help solve the problem of false positives.

Conclusion

- Unsupervised learning can extract meaningful information from non-labeled, messy datasets
- NMF + TF-IDF work well together to create interpretable topics from natural language datasets
- Future work:
    - Apply similarity clustering with embedding vectors for ensemble models