Authenticity-Driven Classification of Disaster-Related Tweets

Group: *PhDs in Hogwarts*

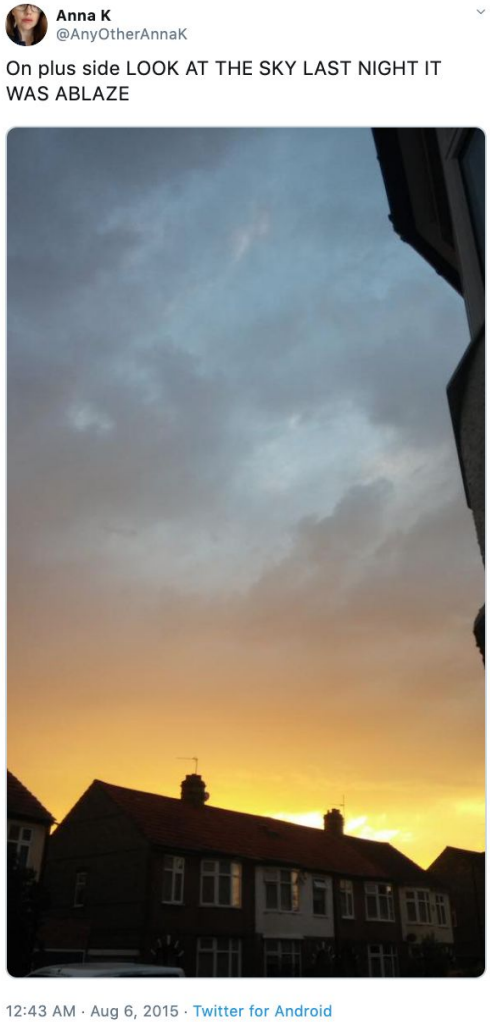
Group Members: *Lulin Yang (luy30@pitt.edu); Xiaoxuan Qin (xiq33@pitt.edu); Wendi Li (wel242@pitt.edu)*

Date: *Apr 21, 2025*

Background

Twitter has emerged as a key communication platform during emergencies. With the widespread use of smartphones, individuals can now report unfolding events in real time. As a result, organizations such as disaster response teams and news outlets are increasingly interested in automatically monitoring Twitter.

However, it's not always straightforward to determine whether a tweet is genuinely reporting a disaster.



Source: Kaggle NLP Getting Started Competition

Dataset (Disaster-Related Tweet from kaggle)

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

id	keyword	text	target
48	ablaze	@bbcmtd Wholesale Markets ablaze http://t.co/lHYXEOHY6C	1
49	ablaze	We always try to bring the heavy. #metal #RT http://t.co/YAo1e0xngw	0
50	ablaze	#AFRICANBAZE: Breaking news:Nigeria flag set ablaze in Aba. http://t.co/2nndBGwyEi	1
52	ablaze	Crying out for more! Set me ablaze	0
53	ablaze	On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE http://t.co/qqsmshaJ3N	0
54	ablaze	@PhDSquares #mufc they've built so much hype around new acquisitions but I doubt they will set the EPL ablaze this season.	0
55	ablaze	INEC Office in Abia Set Ablaze - http://t.co/3ImaomknnA	1
56	ablaze	Barbados #Bridgetown JAMAICA ¬â√õ√í Two cars set ablaze: SANTA CRUZ ¬â√õ√Ì Head of the St Elizabeth Police Superintende... http://t.co/wDUEaj8Q4J	1
57	ablaze	Ablaze for you Lord :D	0
59	ablaze	Check these out: http://t.co/rOI2NSmEJJ http://t.co/3Tj8ZjiN21 http://t.co/YDUiXEfIpE http://t.co/LxTjc87KLS #nsfw	0
61	ablaze	on the outside you're ablaze and alive but you're dead inside	0
62	ablaze	Had an awesome time visiting the CFC head office the ancop site and ablaze. Thanks to Tita Vida for taking care of us ??	0
63	ablaze	SOOOO PUMPED FOR ABLAZE ???? @southridgelife	0
64	ablaze	I wanted to set Chicago ablaze with my preaching... But not my hotel! http://t.co/o9qknbfOFX	0

- `id` - a unique identifier for each tweet
- `text` - the text of the tweet
- `keyword` - a particular keyword from the tweet (may be blank)
- `target` - denotes whether a tweet is about a real disaster (`1`) or not (`0`)

Research Questions

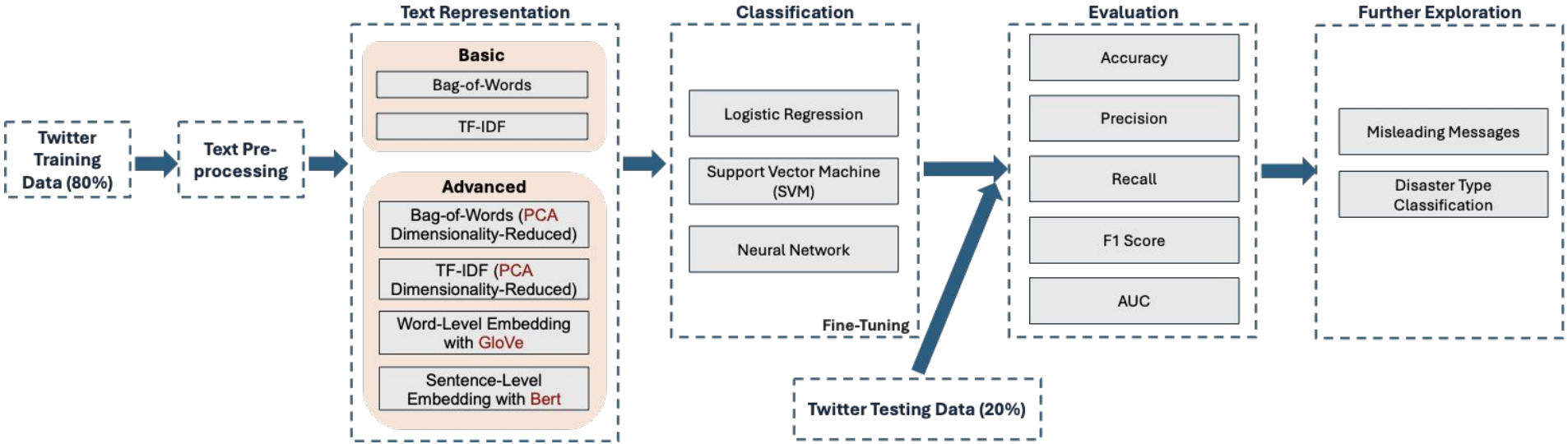
Q1: *Can we detect whether a tweet refers to a real disaster event?*

Q2: *Which text representations and predictive models yield the best detection performance?*

Q3: *What types of messages are most often misclassified?*

Q4: *Can we distinguish between different types of disasters (e.g., natural vs. human-caused)?*

Analytical Workflow



Data pre-processing

Two versions of cleaned texts

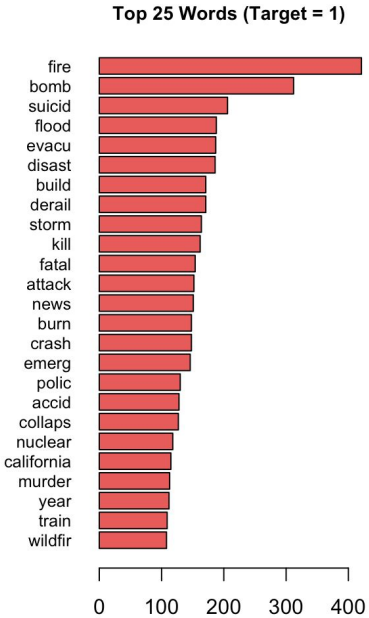
(As a pre-trained language model, BERT can naturally handle punctuation, verb tense, and context — so removing these elements during cleaning is unnecessary and may reduce performance)

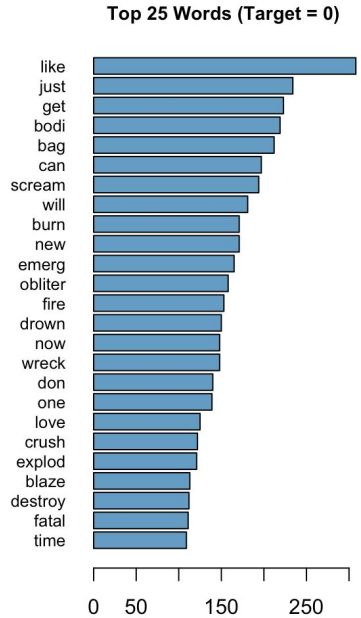
Step	Cleaned Text	Cleaned Text for BERT Embedding
Remove URLs	✅	✅
Lowercasing	✅	✅
Remove punctuation	✅	❌
Remove numbers	✅	❌
Remove stopwords	✅	❌
Remove emojis (non-ASCII)	✅	✅
Stemming	✅	❌
Strip whitespace	✅	✅

Basic Text Representations (Two Methods)

Bag-of-Words (BOW)

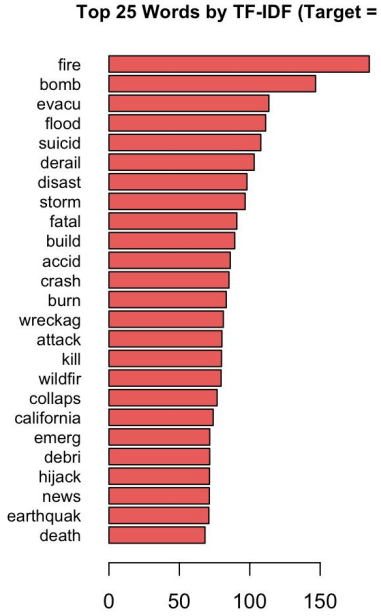
Represents text by word counts, ignoring order and context.

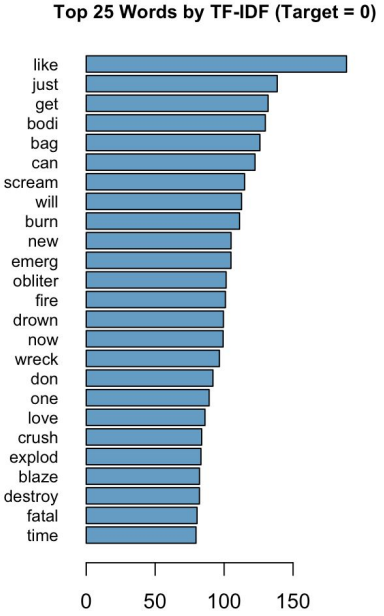




TF-IDF

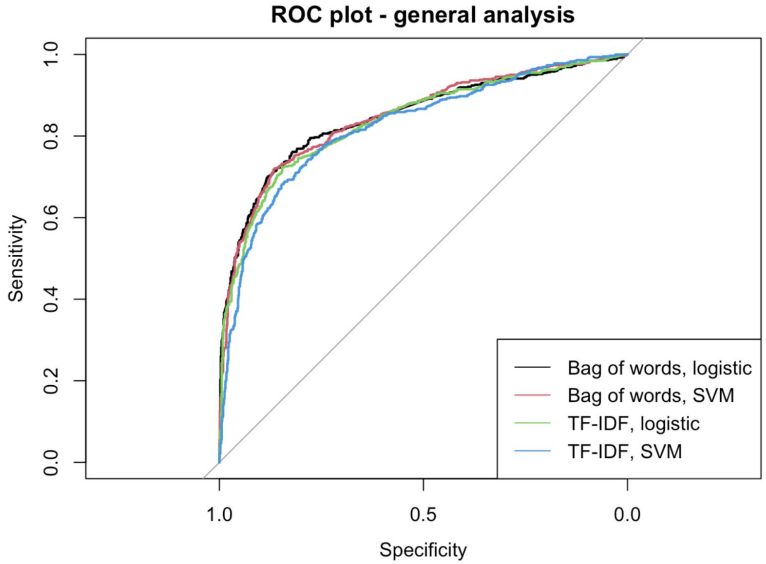
Weighs word importance by frequency and rarity across documents.





Basic Text Representations (Prediction Performance)

Method	Model	Accuracy	Precision	Recall	F1.score	AUC
Bag of Words	Logistic	0.7989488	0.7863591	0.8939567	0.8367129	0.8466318
Bag of Words	SVM	0.8009198	0.8072805	0.8597491	0.8326891	0.8466071
TFIDF	Logistic	0.7825230	0.7660819	0.8962372	0.8260641	0.8373171
TFIDF	SVM	0.7687254	0.7850163	0.8244014	0.8042269	0.8234494



SVM Kernel: Given the inherently high-dimensional sparse nature of our Bag-of-Words and TF-IDF features, we employ a ***linear*** kernel SVM for its computational efficiency and interpretability.

Advanced Text Representations (Four Methods)

Dimensionality Reduction

Bag-of-Words (PCA Dimensionality-Reduced)

- *Top 50 principal components selected via PCA*

TF-IDF (PCA Dimensionality-Reduced)

- *Top 50 principal components selected via PCA*

Semantic Embeddings

Word-Level Embedding with GloVe

- *Each word represented by a 200-dimensional GloVe vector*
- *For each tweet, taking the average of its word vectors*

Sentence-Level Embedding with Bert

- *Each sentence (tweet) is encoded as 384-dimensional Vector through model "all-MiniLM-L6-v2"*

Advanced Text Representations (Four Methods)

Dimensionality Reduction

Bag-of-Words (PCA Dimensionality-Reduced)

- *Top 50 principal components selected via PCA*

TF-IDF (PCA Dimensionality-Reduced)

- *Top 50 principal components selected via PCA*

Semantic Embeddings

Word-Level Embedding with GloVe

- *Each word represented by a 200-dimensional GloVe vector*
- *For each tweet, taking the average of its word vectors*

Sentence-Level Embedding with Bert

- *Each sentence (tweet) is encoded as 384-dimensional Vector through model "all-MiniLM-L6-v2"*

Advanced Text Representations (Four Methods)

Dimensionality Reduction

Bag-of-Words (PCA Dimensionality-Reduced)

- *Top 50 principal components selected via PCA*

TF-IDF (PCA Dimensionality-Reduced)

- *Top 50 principal components selected via PCA*

Semantic Embeddings

Word-Level Embedding with GloVe

- *Each word represented by a 200-dimensional GloVe vector*
- *For each tweet, taking the average of its word vectors*

Sentence-Level Embedding with Bert

- *Each sentence (tweet) is encoded as 384-dimensional Vector through model "all-MiniLM-L6-v2"*

Advanced Text Representations (Prediction Performance)

Method	Model	Accuracy	Precision	Recall	F1.score	AUC
Bag of Words	Logistic	0.7220762	0.7229862	0.8392246	0.7767810	0.7664192
Bag of Words	SVM	0.7168200	0.7169261	0.8403649	0.7737533	0.7634174
Bag of Words	Neural network	0.7726675	0.7600000	0.6775194	0.7163934	0.8128283
TFIDF	Logistic	0.7128778	0.7075472	0.8551881	0.7743934	0.7780303
TFIDF	SVM	0.7168200	0.7169261	0.8403649	0.7737533	0.7634174
TFIDF	Neural network	0.7614980	0.7491166	0.6573643	0.7002477	0.8116668
Word-Embedding (GloVe)	Logistic	0.7095926	0.7267987	0.7947548	0.7592593	0.7673906
Word-Embedding (GloVe)	SVM	0.6977661	0.6762468	0.9122007	0.7766990	0.7644525
Word-Embedding (GloVe)	Neural network	0.7253614	0.6857610	0.6496124	0.6671975	0.7811708
Sentence-Embedding (BERT)	Logistic	0.8147175	0.8316611	0.8506271	0.8410372	0.8647106
Sentence-Embedding (BERT)	SVM	0.7943495	0.7993631	0.8586089	0.8279274	0.8615338
Sentence-Embedding (BERT)	Neural network	0.8153745	0.8013245	0.7503876	0.7750200	0.8694996

Overall

- *Reliable overall performance (almost all of accuracy, F1 score and AUC are over 0.7)*
- *Most of recall outperforms precision and accuracy, indicating strong disaster detection*
- *Lower precision suggests frequent false positives across models*

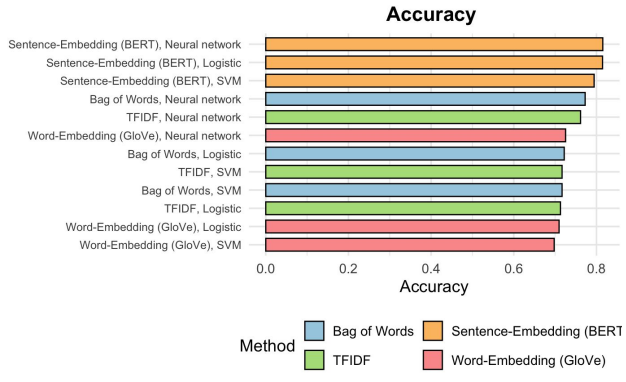
Advanced Text Representations (Prediction Performance)

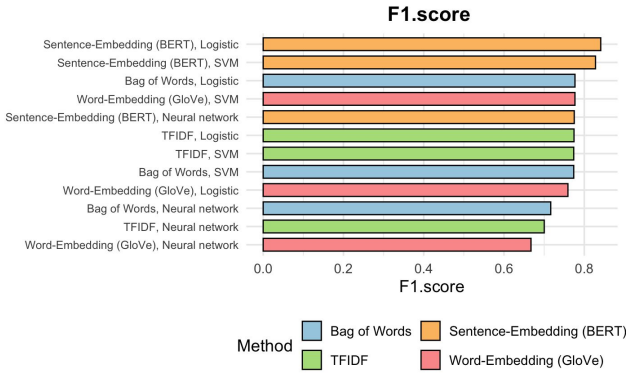
Method	Model	Accuracy	Precision	Recall	F1.score	AUC
Bag of Words	Logistic	0.7220762	0.7229862	0.8392246	0.7767810	0.7664192
Bag of Words	SVM	0.7168200	0.7169261	0.8403649	0.7737533	0.7634174
Bag of Words	Neural network	0.7726675	0.7600000	0.6775194	0.7163934	0.8128283
TFIDF	Logistic	0.7128778	0.7075472	0.8551881	0.7743934	0.7780303
TFIDF	SVM	0.7168200	0.7169261	0.8403649	0.7737533	0.7634174
TFIDF	Neural network	0.7614980	0.7491166	0.6573643	0.7002477	0.8116668
Word-Embedding (GloVe)	Logistic	0.7095926	0.7267987	0.7947548	0.7592593	0.7673906
Word-Embedding (GloVe)	SVM	0.6977661	0.6762468	0.9122007	0.7766990	0.7644525
Word-Embedding (GloVe)	Neural network	0.7253614	0.6857610	0.6496124	0.6671975	0.7811708
Sentence-Embedding (BERT)	Logistic	0.8147175	0.8316611	0.8506271	0.8410372	0.8647106
Sentence-Embedding (BERT)	SVM	0.7943495	0.7993631	0.8586089	0.8279274	0.8615338
Sentence-Embedding (BERT)	Neural network	0.8153745	0.8013245	0.7503876	0.7750200	0.8694996

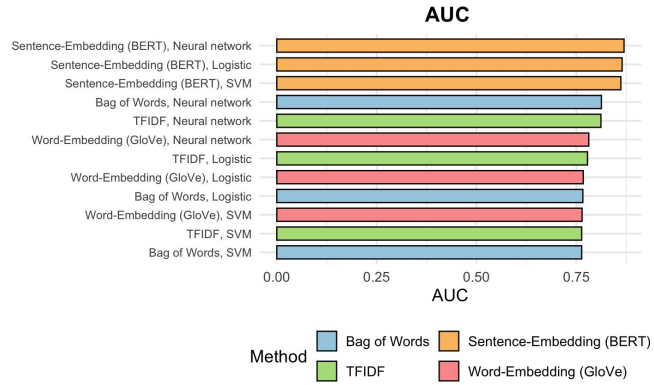
Overall

- *Reliable overall performance (almost all of accuracy, F1 score and AUC are over 0.7)*
- *Most of recall outperforms precision and accuracy, indicating strong disaster detection*
- *Lower precision suggests frequent false positives across models*

Advanced Text Representations (Prediction Performance)



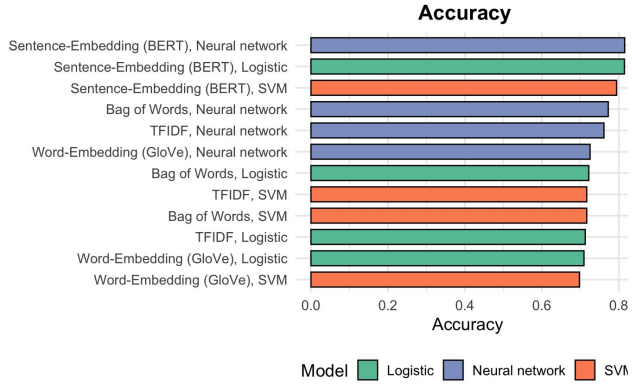


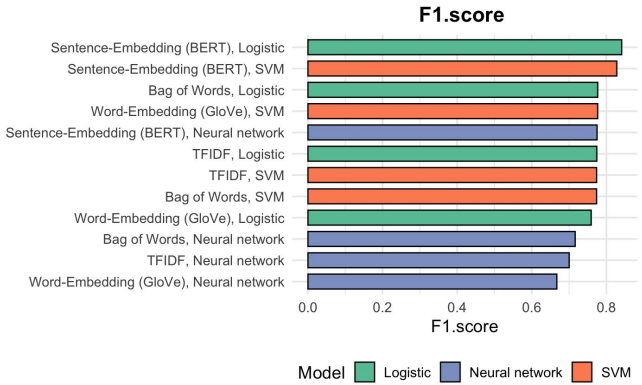


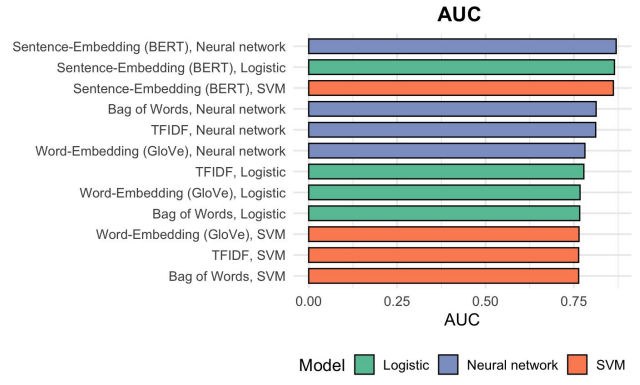
Text representation method selection

- *BERT embeddings consistently perform best, especially in F1 and AUC.*
- *GloVe underperforms, likely due to lack of context.*
- *TF-IDF and Bag of Words remain competitive with simpler models.*

Advanced Text Representations (Prediction Performance)







Model Selection

- *Logistic regression performs consistently well, especially with BERT.*
- *Neural networks do better with richer features like BERT and TF-IDF, but struggle with weaker ones like GloVe.*
- *SVM has high recall but lower precision.*

Q1: *Can we detect whether a tweet refers to a real disaster event?*

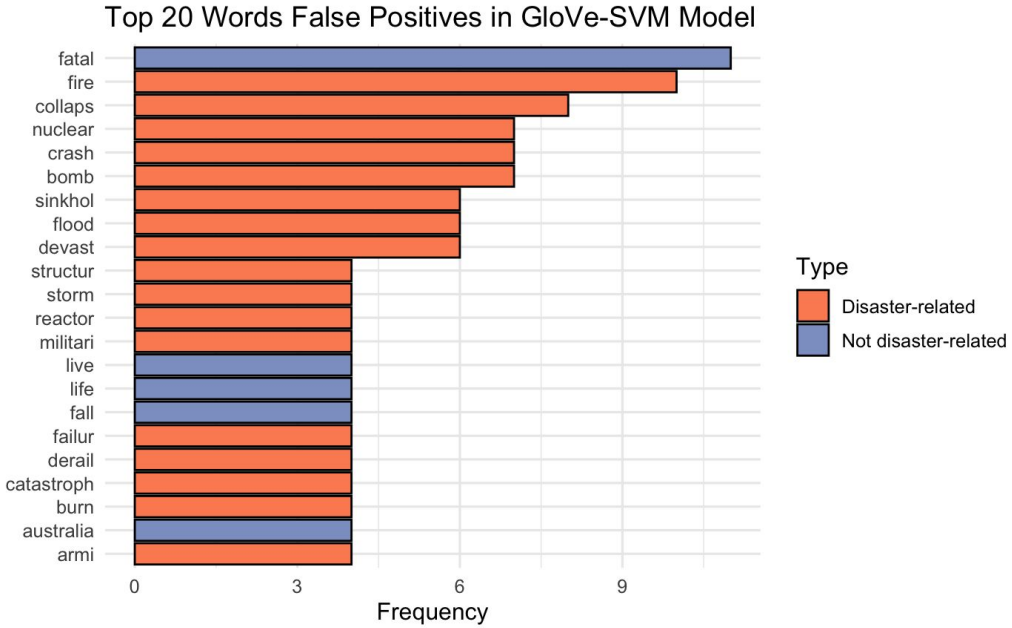
Yes. The predictive models achieve strong overall performance, with accuracy, F1-score, and AUC mostly above 0.7 across all method–model combinations. High recall values (often > 0.8) indicate that the models are effective at identifying disaster-related tweets with low miss rates.

Q2: *Which text representations and predictive models yield the best detection performance?*

Sentence-level BERT embeddings combined with logistic regression consistently deliver the best results, achieving the highest F1-score (0.84) and AUC (0.86). Reduced TF-IDF representations also perform well, especially with neural networks. In contrast, GloVe embeddings underperform across models, likely due to limited contextual information.

Further Exploration 1: *What types of messages are most often misclassified? (Q3)*

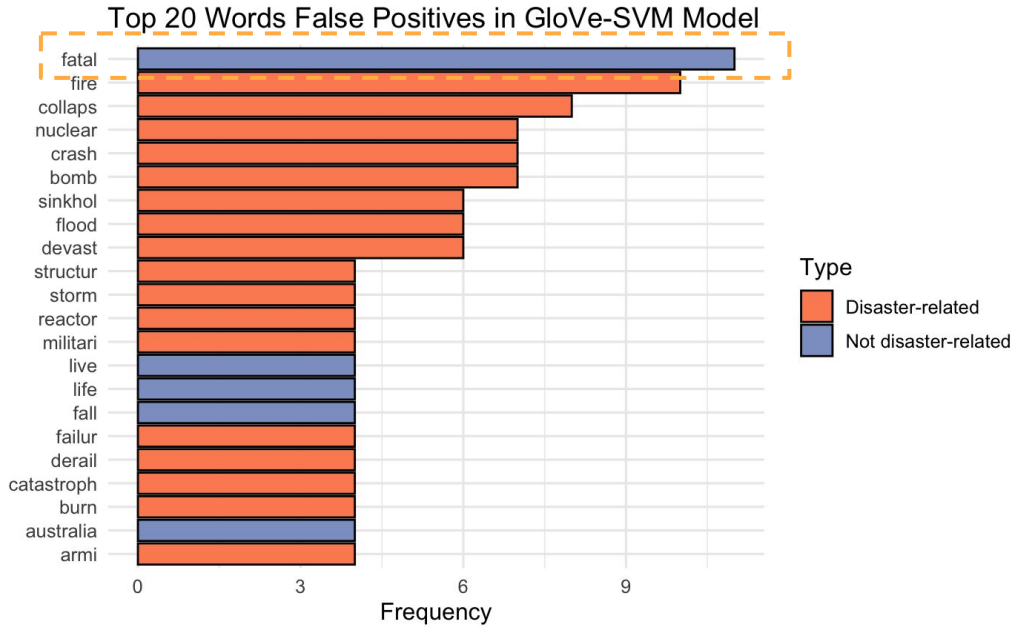
GloVe-SVM selected as the example for error analysis due to high recall (0.91) and low precision (0.67), indicating frequent false positives.



- *Among top 20 false-positive words, 15 are disaster-related, 5 words are not directly related to disaster.*
- *Noisy co-occurrence: Non-disaster words may frequently co-occur with disaster words in training data, leading the model to wrongly associate them with the disaster class.*

Further Exploration 1: *What types of messages are most often misclassified? (Q3)*

GloVe-SVM selected as the example for error analysis due to high recall (0.91) and low precision (0.67), indicating frequent false positives.

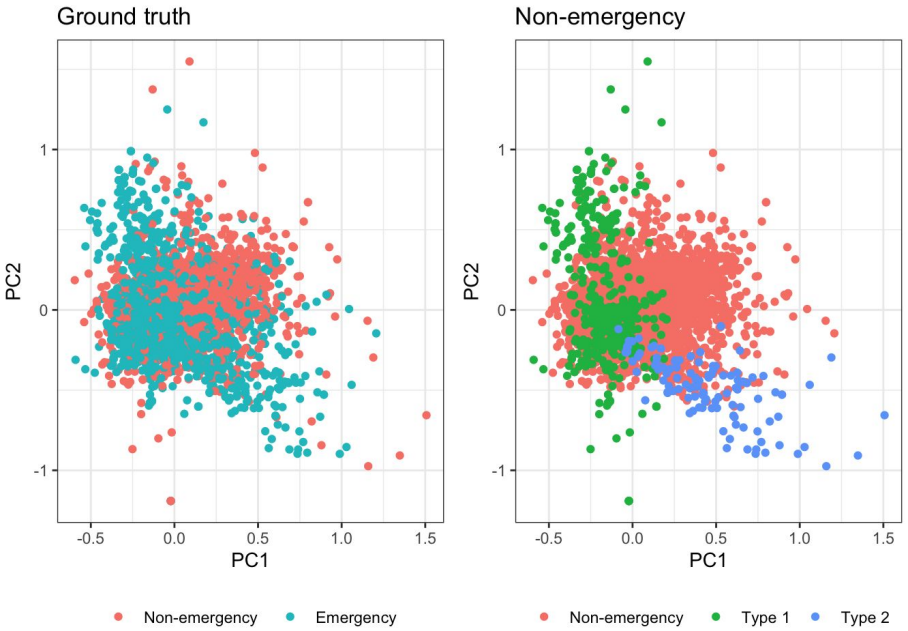


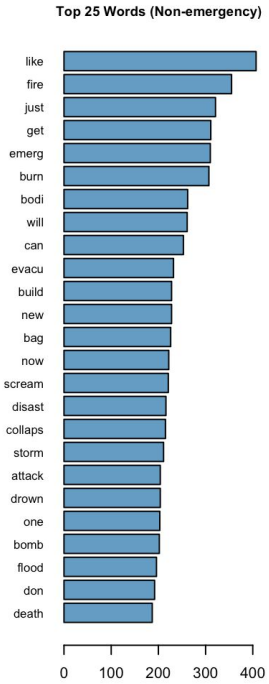
Importance of considering contextual meaning

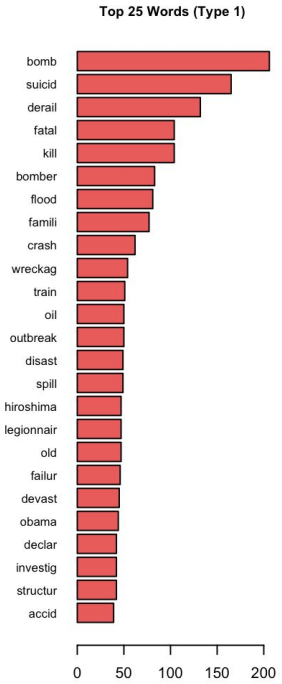
"Fatal" frequently appears in false positives — though it connotes severity or death, it does not always indicate a disaster event. In real-world usage, people often say: "a fatal error in the code" ,"fatal attraction"...

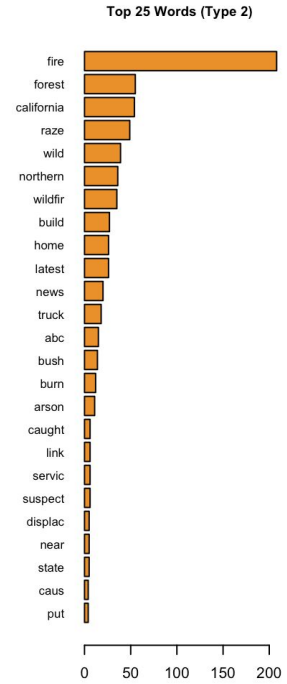
Further Exploration 2: *Can we distinguish between different types of disasters? (Q4)* ***Yes***

GloVe-SVM is selected as the example for distinguishing analysis









Human-caused **Natural**

Future Work

- ★ Fine-tune classification models to further improve the performance
- ★ Explore more advanced or ensemble classification methods
- ★ Enhance misinformation detection with external knowledge sources
- ★ Investigate temporal and geographic features of tweets, integrating metadata