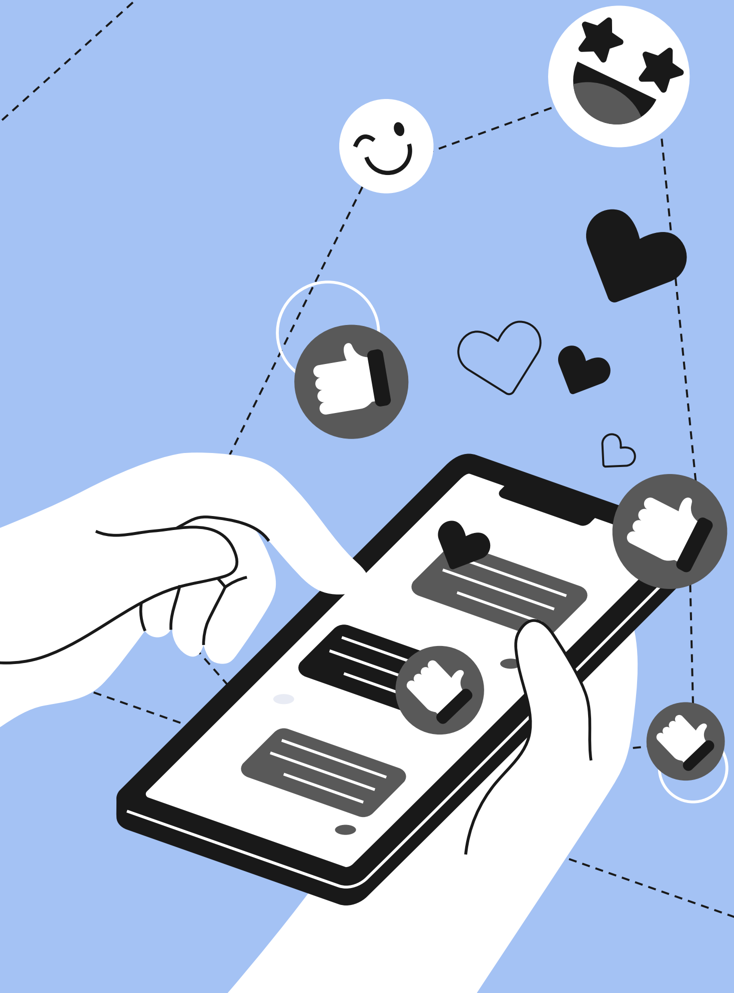


Twitter News Spread Analysis

By: Yash Dani, Waleed Samouh, Malik Samouh, Eyad Mahmoud

Table of contents

Introduction

Overview of the problem

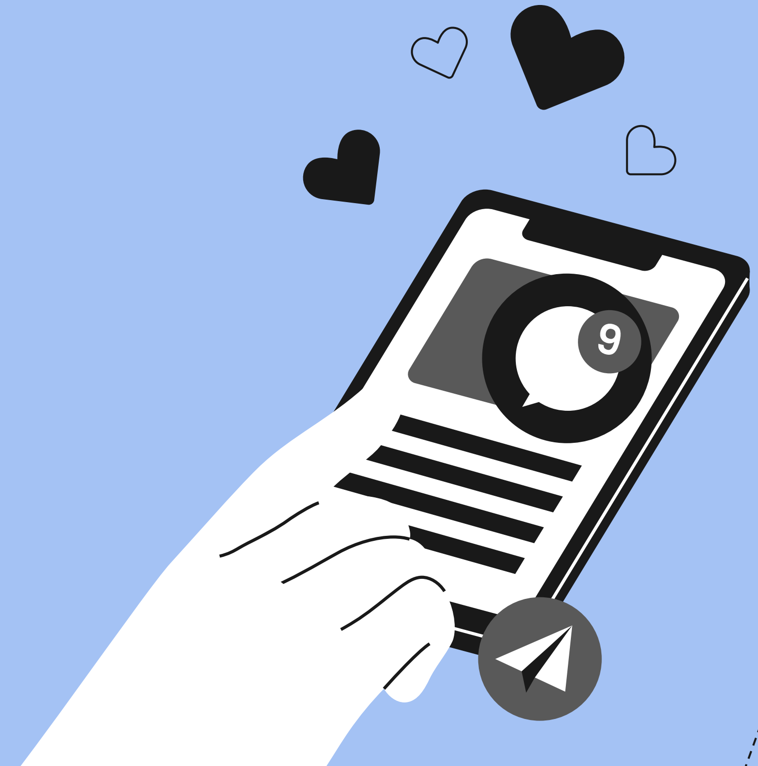
- Does fake news spread faster on twitter?
- How can we mitigate that?

Importance of the study

- Fake news can influence public opinion, political outcomes and social behaviour.

Objective

- To analyze the spread patterns of fake news vs. real news on Twitter.



Problem Definition

Problem 1

Cleaning and labeling the dataset to categorize news into real, fake, rumor unverified, and rumor verified.

Problem 2

Visualizing the network of news propagation.

Problem 3

Analyzing the characteristics and patterns of news spread.

Methodology - Data Collection & Cleaning

Data Source & Collection

- Twitter15 dataset
- Twitter16 dataset
- Twitter API Collection

Data Cleaning

- Removed duplicates and irrelevant entries
- Ensured correct bucketing and labeling of files

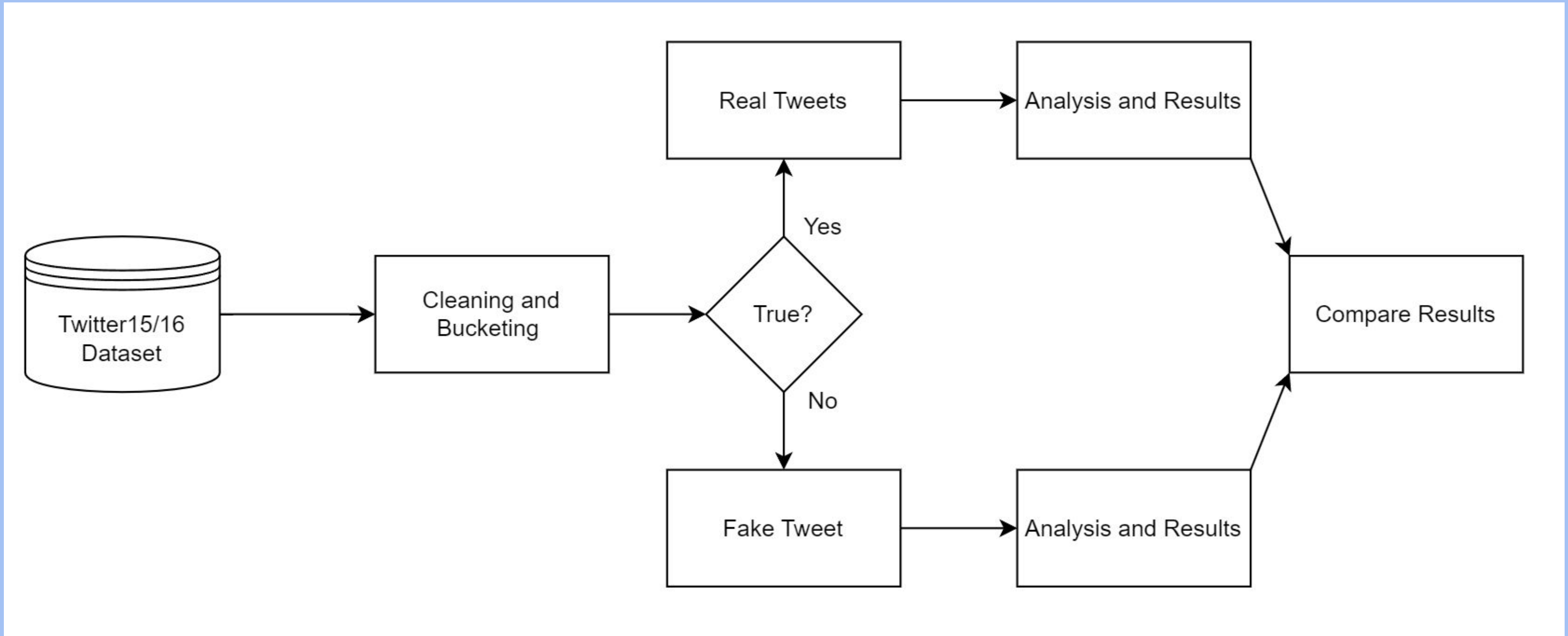
Data Categorization

- Categories: real, fake, rumor unverified, rumor verified

Scripts and Tools

- Pandas
- Numpy
- Networkx
- Matplotlib

Framework


Twitter15/16
Dataset
Cleaning and
Bucketing
True?
Yes
No
Real Tweets
Fake Tweet
Analysis and Results
Analysis and Results
Compare Results

Methodology - Network Construction

G(N, E) Directed Graph:

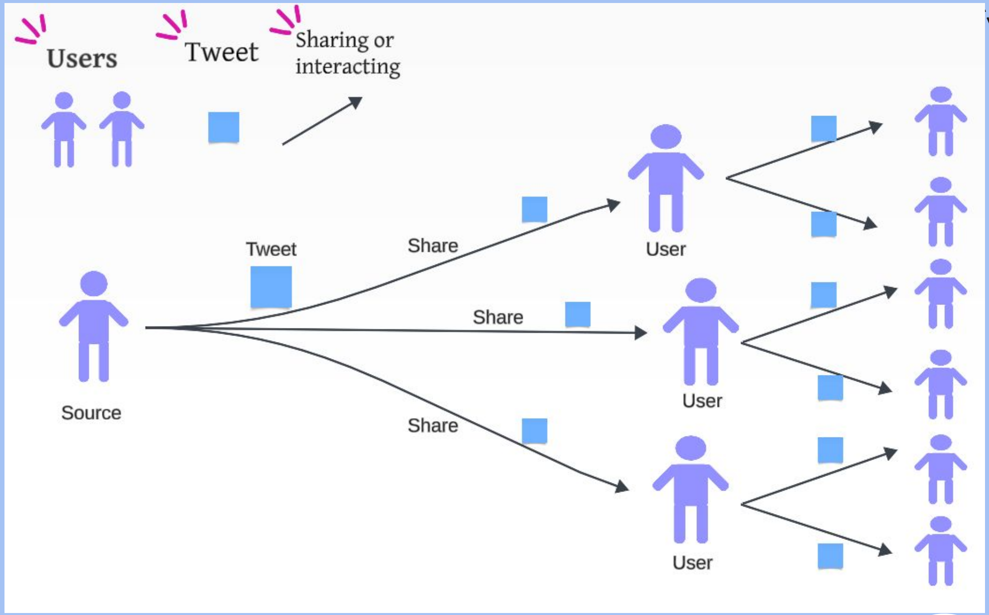
- N: set of nodes, node is a user on twitter
- E: Set of edges, edge is user interaction with a tweet

Source:

- User that tweeted the initial tweet

Tweet:

- Graph per one tweet



Methodology - Network Visualization

Initial Visualization:

- Plotted general graph for individual files

Propagation and Cascading Effects:

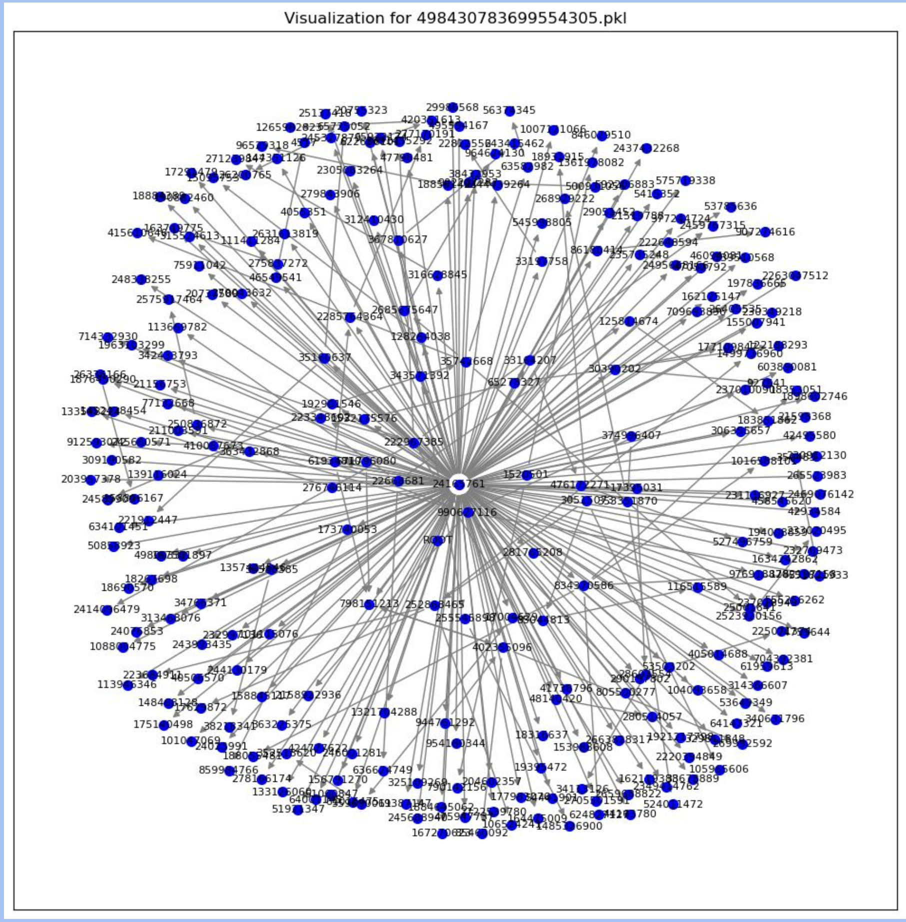
- Identified and analyzed

Top Chains Identification:

- Found longest chains in Twitter15 and Twitter16 datasets

Graph Plotting:

- Visualized top files with longest chains
- Created general and single chain graphs for both real and fake news



Methodology- Statistical Analysis

Avg Chain Length & Tree Depth

Calculated the average length of propagation chains.
Analyzed the average depth of the propagation trees.

Avg Number of Nodes & Edges

Determined the average number of nodes in the network.
Computed the average number of edges connecting the nodes.

Cascade Size

Measured the average size of cascades

Propagation Delay

Evaluated the average delay in news propagation

Reaction Time

Assessed the average time it takes for reactions to spread

Centrality Metrics

Calculated average betweenness and closeness centrality for nodes

Experiments and Results- Analysis of Longest Chains

Longest Chains Identification:

- Used Python scripts to identify the longest chains

Top Files Analysis:

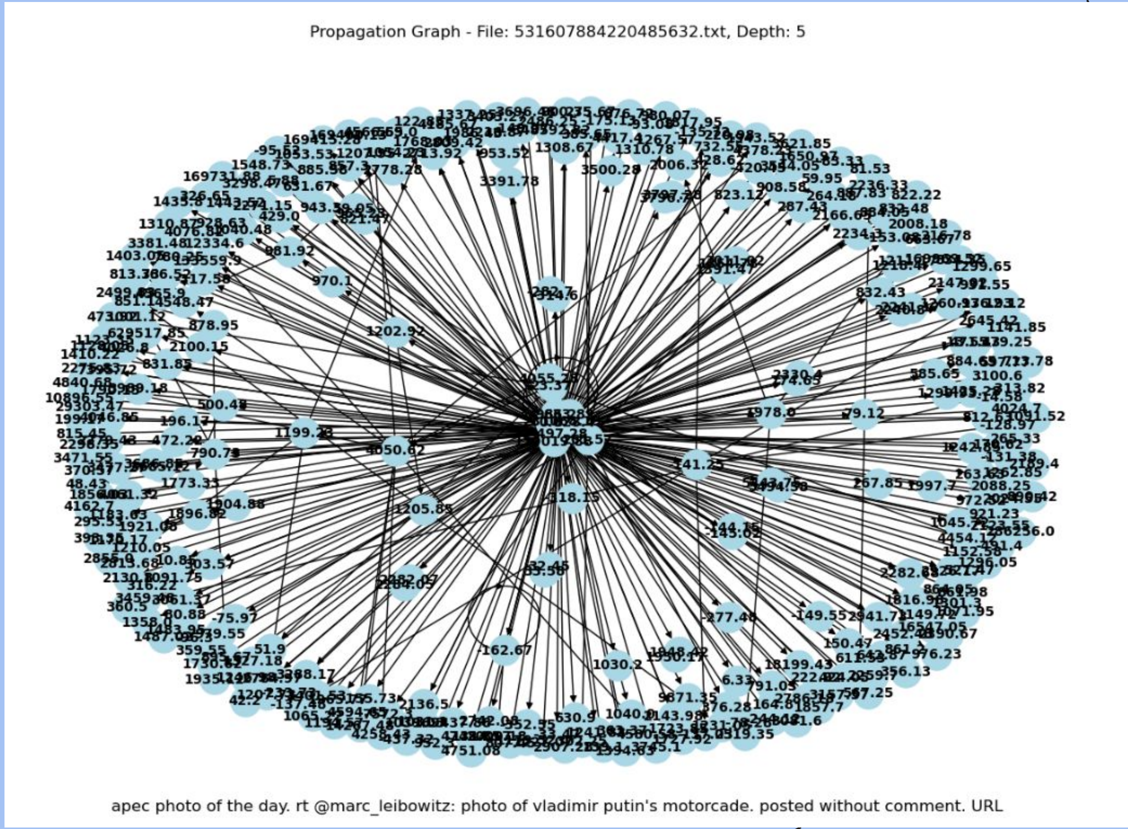
- Analyzed top 3 longest chains in real and fake news for Twitter15 and Twitter16 datasets

Comparison:

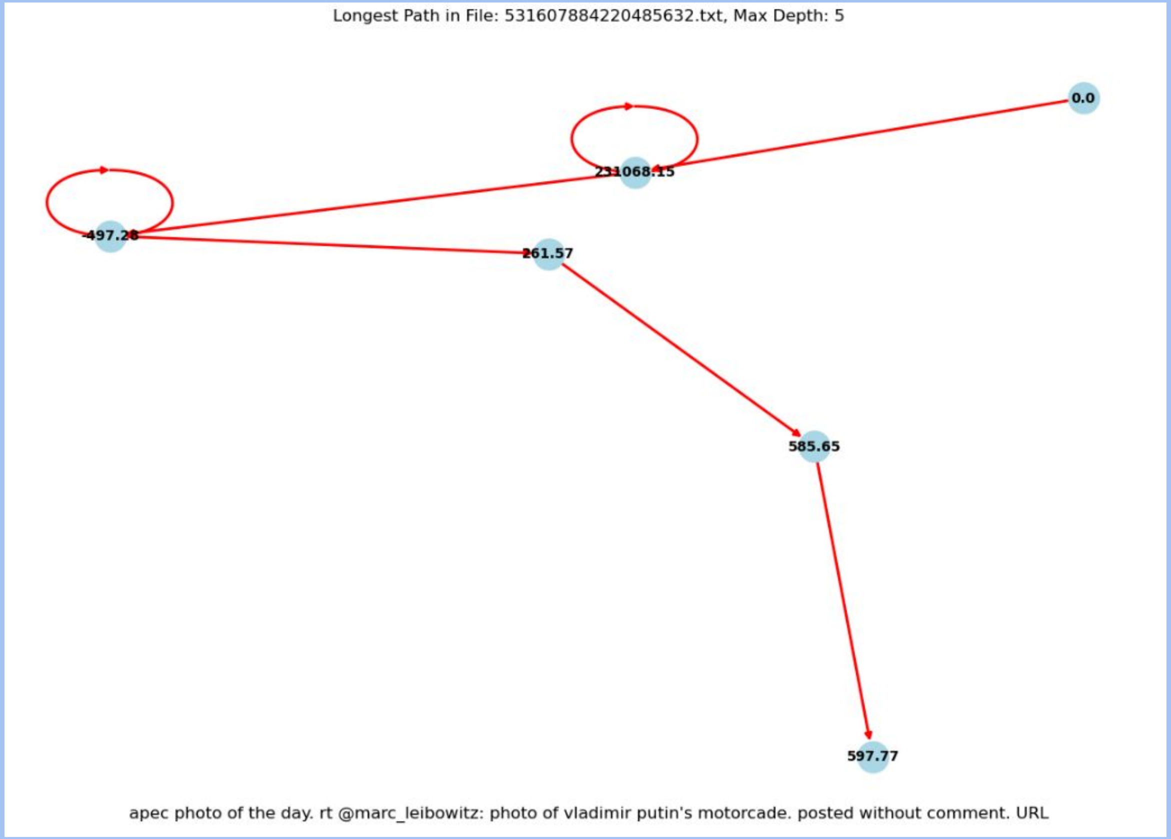
- Compared chain lengths and propagation patterns between real and fake news

Insights:

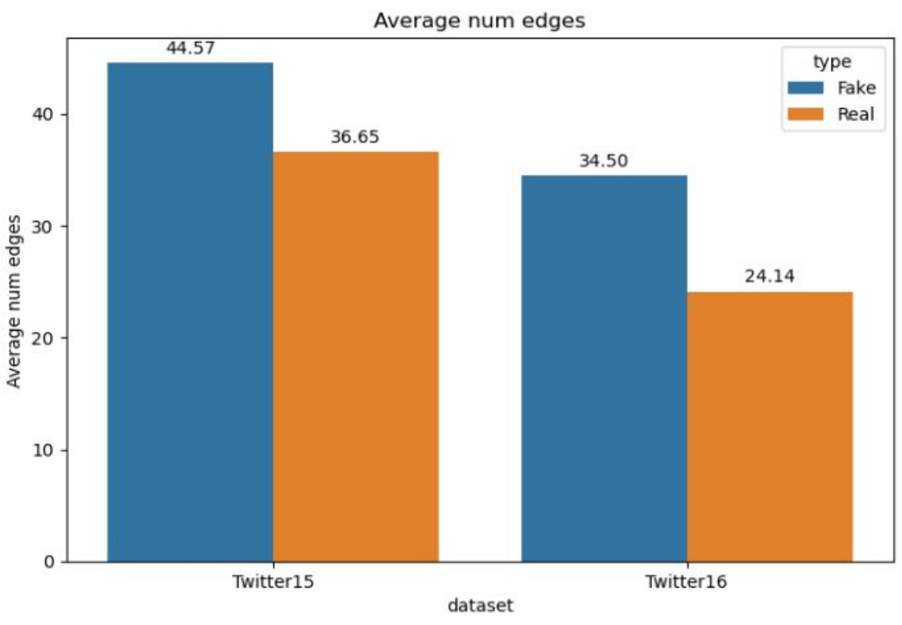
- Observed differences in spread dynamics and depth of propagation

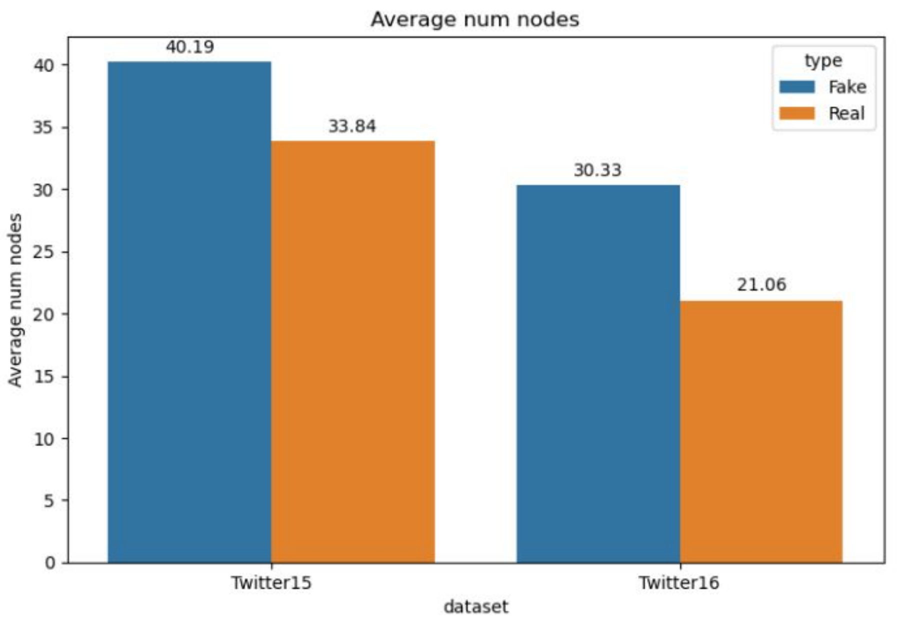


Longest Chain

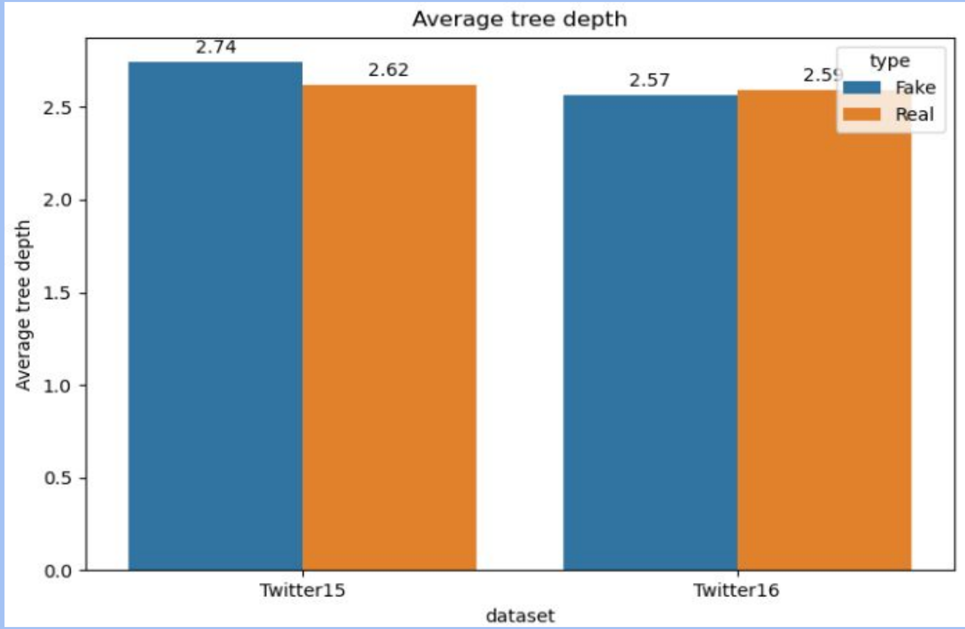


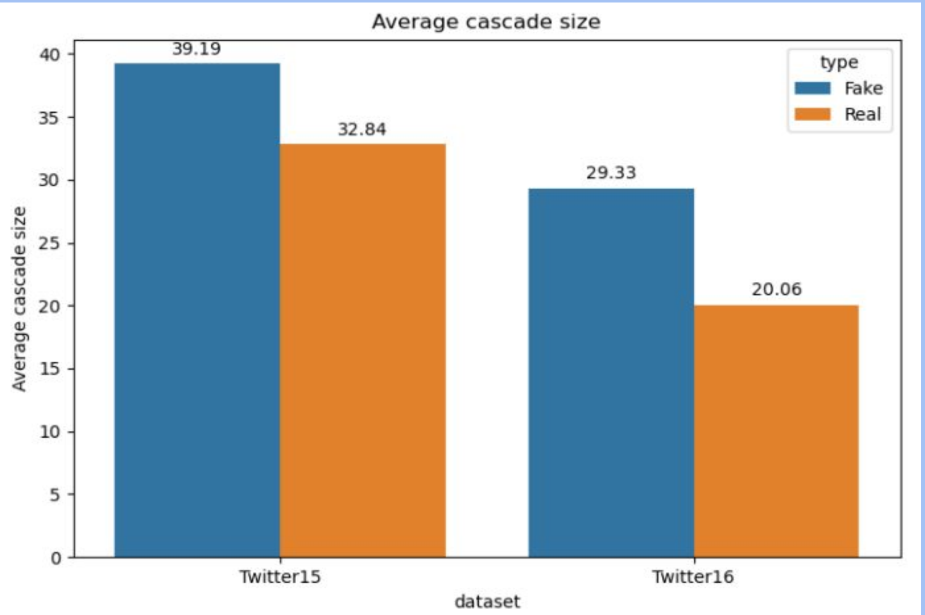
Experiments and Results - Statistical Analysis



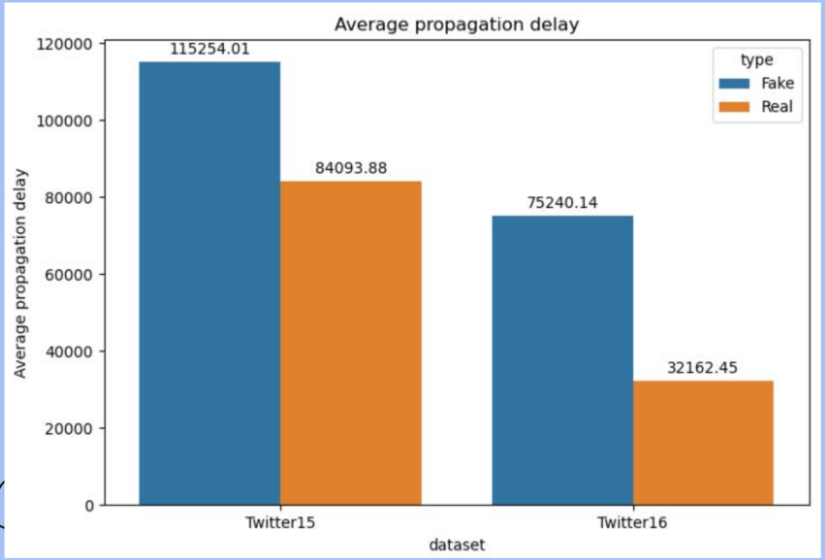


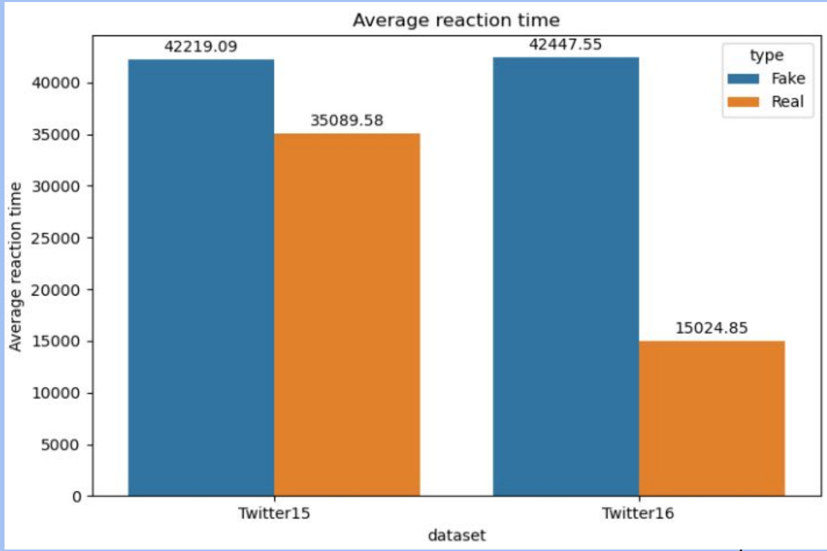
Experiments and Results – Statistical Analysis



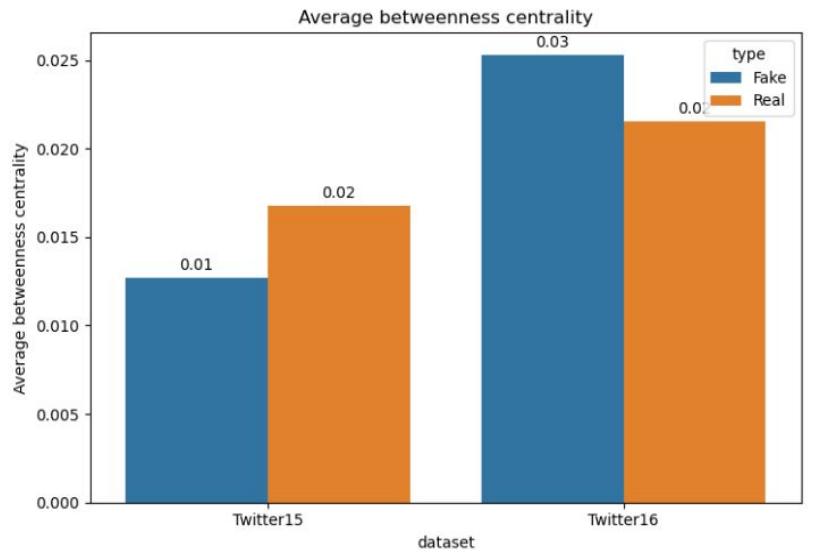


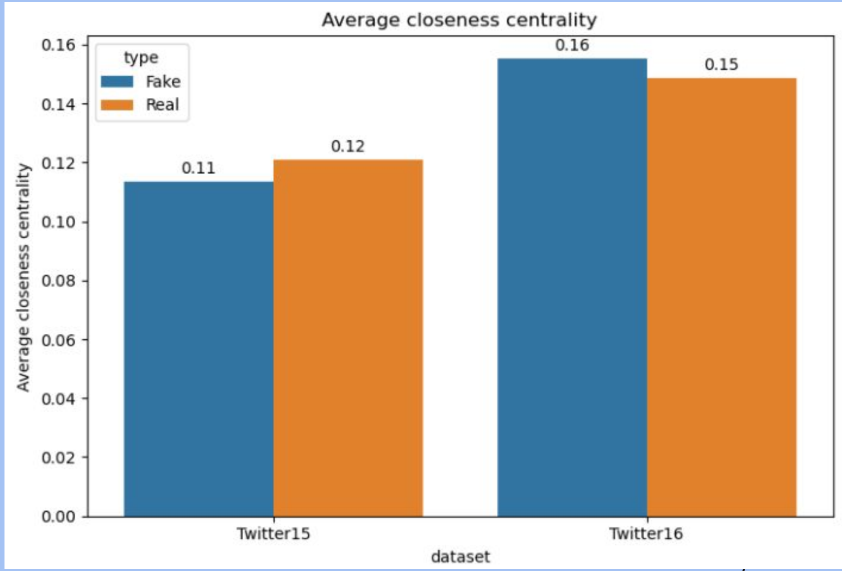
Experiments and Results - Statistical Analysis



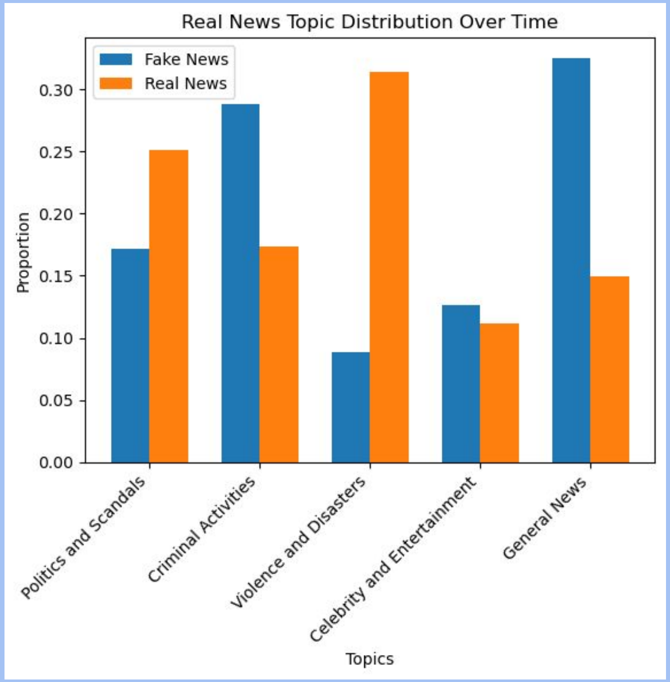


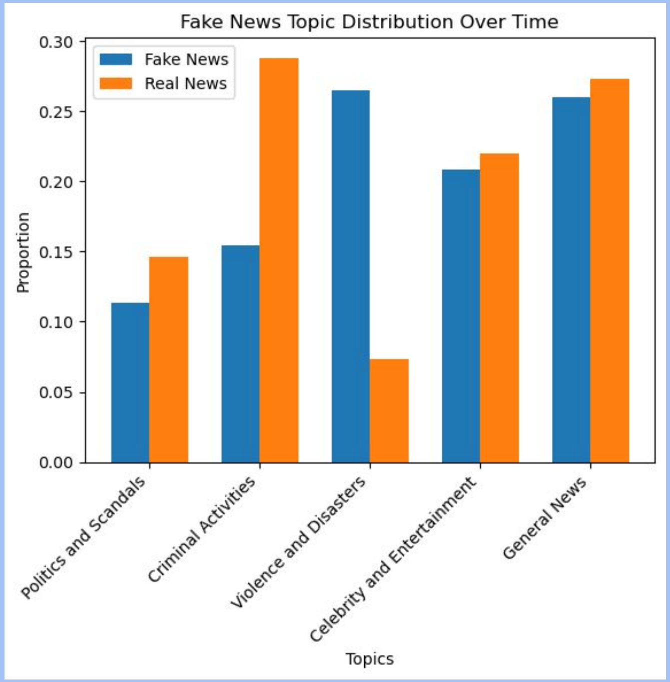
Experiments and Results - Statistical Analysis





Experiments and Results - Category Analysis





Topic Modelling Techniques

Purpose

- Analyze sentiment and emotion in tweets.

Steps

1. **Load Tweets:** Load tweet data from files.
2. **Perform Sentiment Analysis:** Use transformers pipeline for sentiment analysis.
3. **Perform Emotion Detection:** Use pre-trained model for emotion detection..

Key Techniques

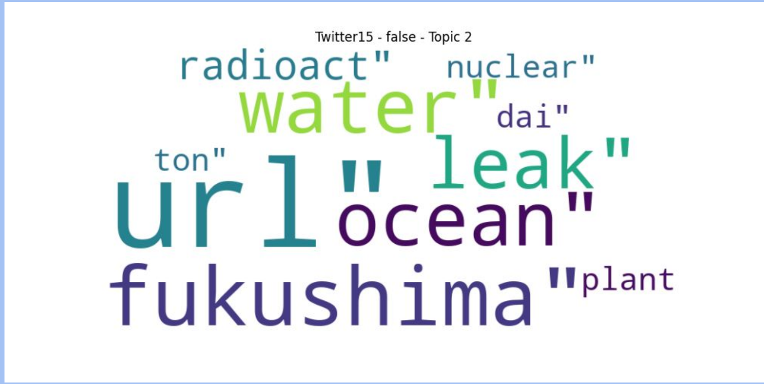
- Sentiment analysis with Hugging Face transformers.
- Emotion detection with DistilRoBERTa model..

Visualizations

- WordCloud of extracted topics and their weights.

Example

- Topic 1: “url”, “watch”, “anonym”
- Topic 2: “url”, “shoot”, “mass”



Sentiment and Emotion Analysis

Purpose

- Discover underlying topics in tweet datasets

Steps

1. **Preprocess tweets:** Clean and prepare tweet text data.
2. **Create Dictionary and Corpus:** Use Gensim to create dictionary and corpus from processed tweets.
3. **Perform LDA:** Apply Latent Dirichlet Allocation (LDA) to identify topics.

Key Techniques

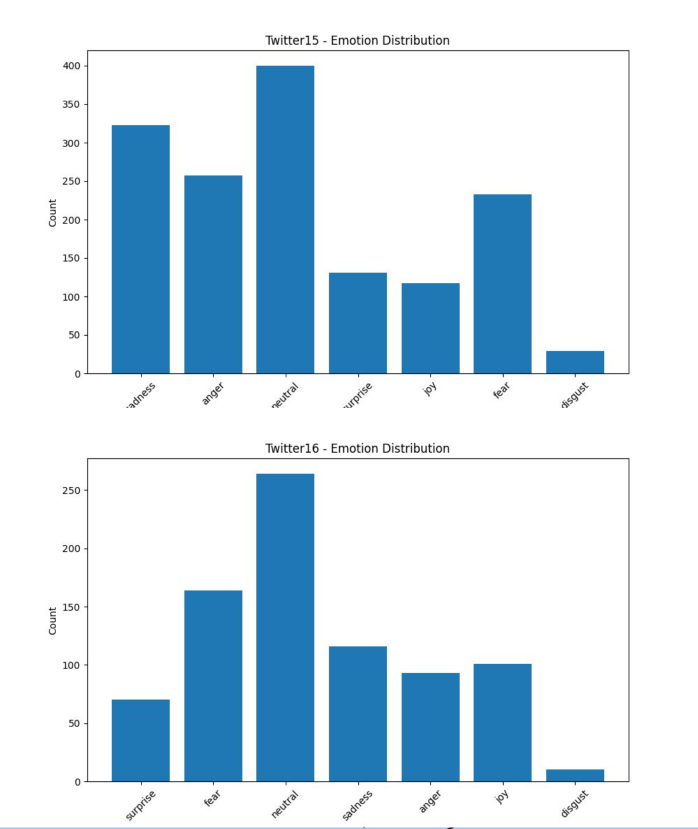
- Text preprocessing with Gensim.
- LDA for topic extraction.

Visualizations

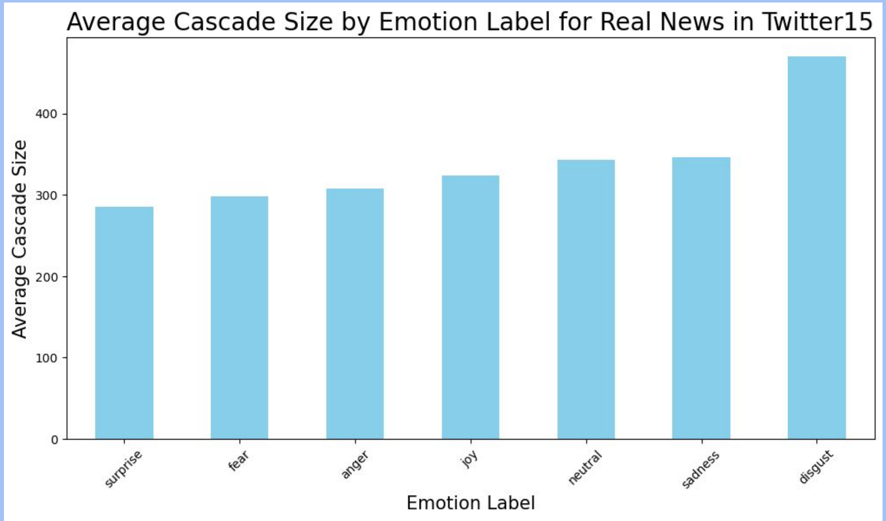
- Example of emotion distribution for twitter 15 and 16

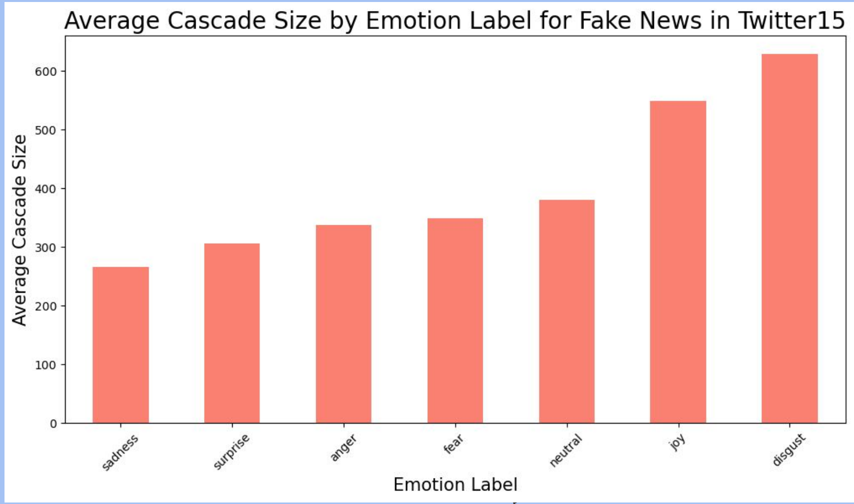
Example

- Tweet 1: Positive sentiment, Sadness emotion
- Tweet 1: Positive sentiment, Sadness emotion



Experiments and Results – Sentimental Analysis





Cascade Triggering Analysis (Random Model)

Steps

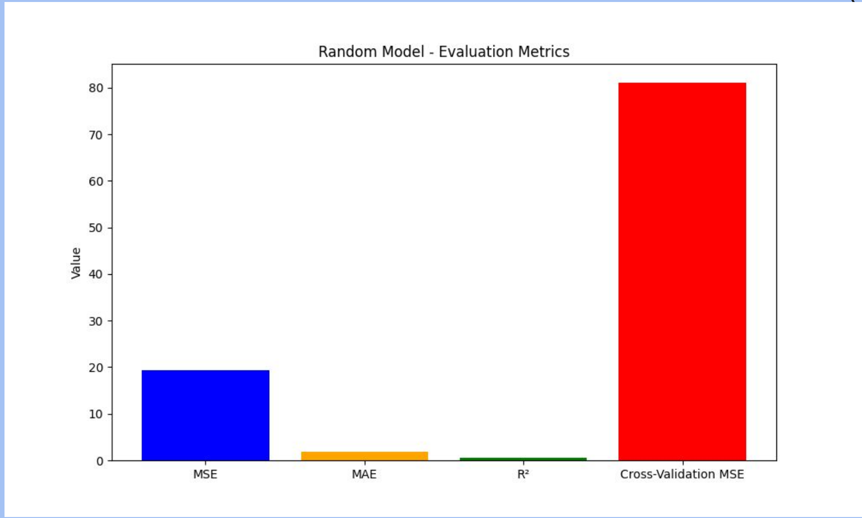
1. **Load Data:** Load sentiment, emotion, and graph data.
2. **Extract Features:** Extract graph and content features.
3. **Train Model:** Use Random Forest Regressor.
4. **Evaluate Model:** Calculate MSE, MAE, R^2, and cross-validation MSE.

Techniques Used

- Random Forest Regressor for prediction.
- Evaluation metrics: MSE, MAE, R^2.

Visualizations

- Metrics comparison for regular model.



Cascade Triggering Analysis (Advanced Model)

Steps

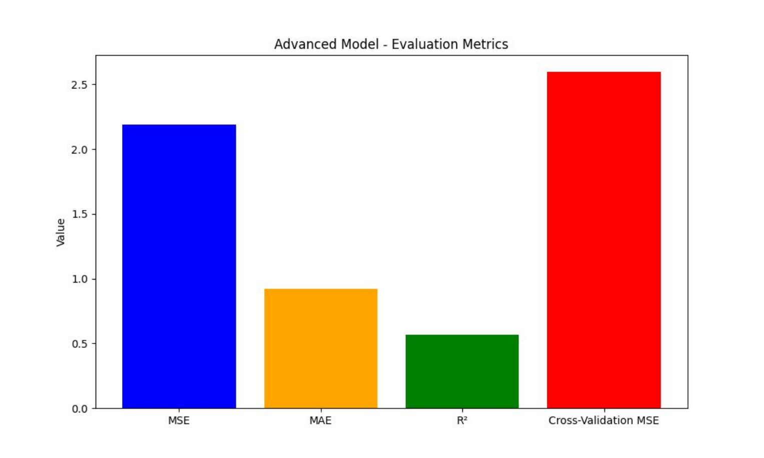
1. **Load Data:** Load sentiment, emotion, and graph data.
2. **Extract Features:** Extract advanced features including temporal, user, and advanced NLP-based content features.
3. **Train Model:** Use Ridge Regression and advanced ensemble methods.
4. **Evaluate Model:** Calculate MSE, MAE, R^2, and cross-validation MSE.

Techniques Used

- Advanced feature engineering (temporal, user, content features).
- Ensemble methods and Ridge Regression for prediction.
- Evaluation metrics: MSE, MAE, R^2.

Visualizations

- Metrics comparison for advanced model.



Comparative Analysis

Metrics Comparison

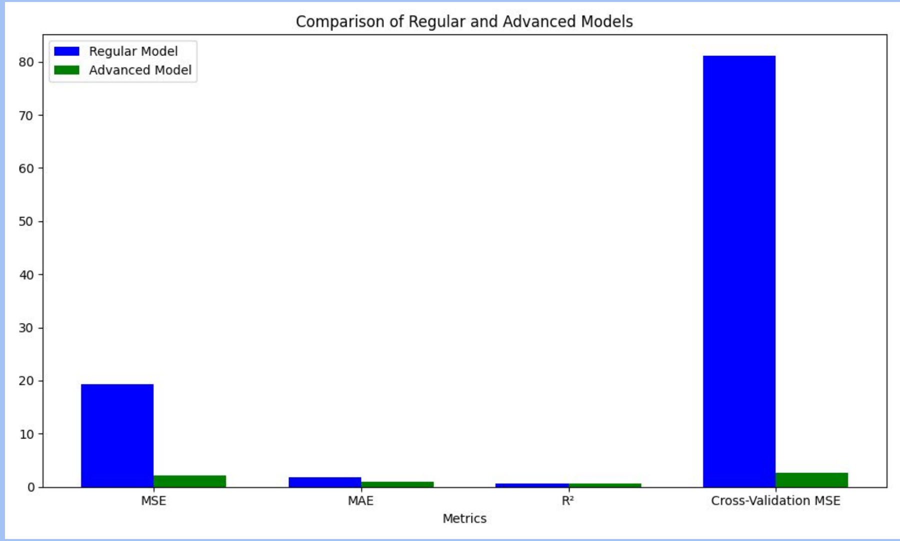
- **MSE:** Regular - 19.3034, Advanced - 2.1905
- **MAE:** Regular - 1.8647, Advanced - 0.9219
- **R^2 Score:** Regular - 0.5829, Advanced - 0.5672
- **Cross-Validation MSE:** Regular - 81.0579, Advanced - 2.5965

Feature Importance Comparison

- **Regular Model:** Balanced distribution of feature importances.
- **Advanced Model:** Highly dominant feature with varying importance for others.

Conclusion

- The advanced model demonstrates significantly improved predictive accuracy and generalization capabilities.



Conclusion

Project Summary

We analyzed tweets to understand sentiment, emotion, and topics, and assessed how information spreads through social networks.

Key Findings

- **Sentiment & Emotion Analysis**: The emotional tone of tweets influences their spread.
- **Topic Modeling**: Identified key topics that drive discussions.
- **Graph Analysis**: Examined tweet propagation networks to extract structural features.
- **Cascade Triggering**: Advanced models predict tweet spread more accurately.

Insights

- **Fake News**: Spreads rapidly due to emotional engagement and specific topics.
- **Accurate Information**: Requires less emotional content but can be impactful with the right topics.

Future Directions

- **Enhance Analysis**: Incorporate more advanced features and models.
- **Broaden Scope**: Apply techniques to other social media platforms.
- **Improve Visualizations**: For better data interpretation and decision-making.

Thank you!

Questions?