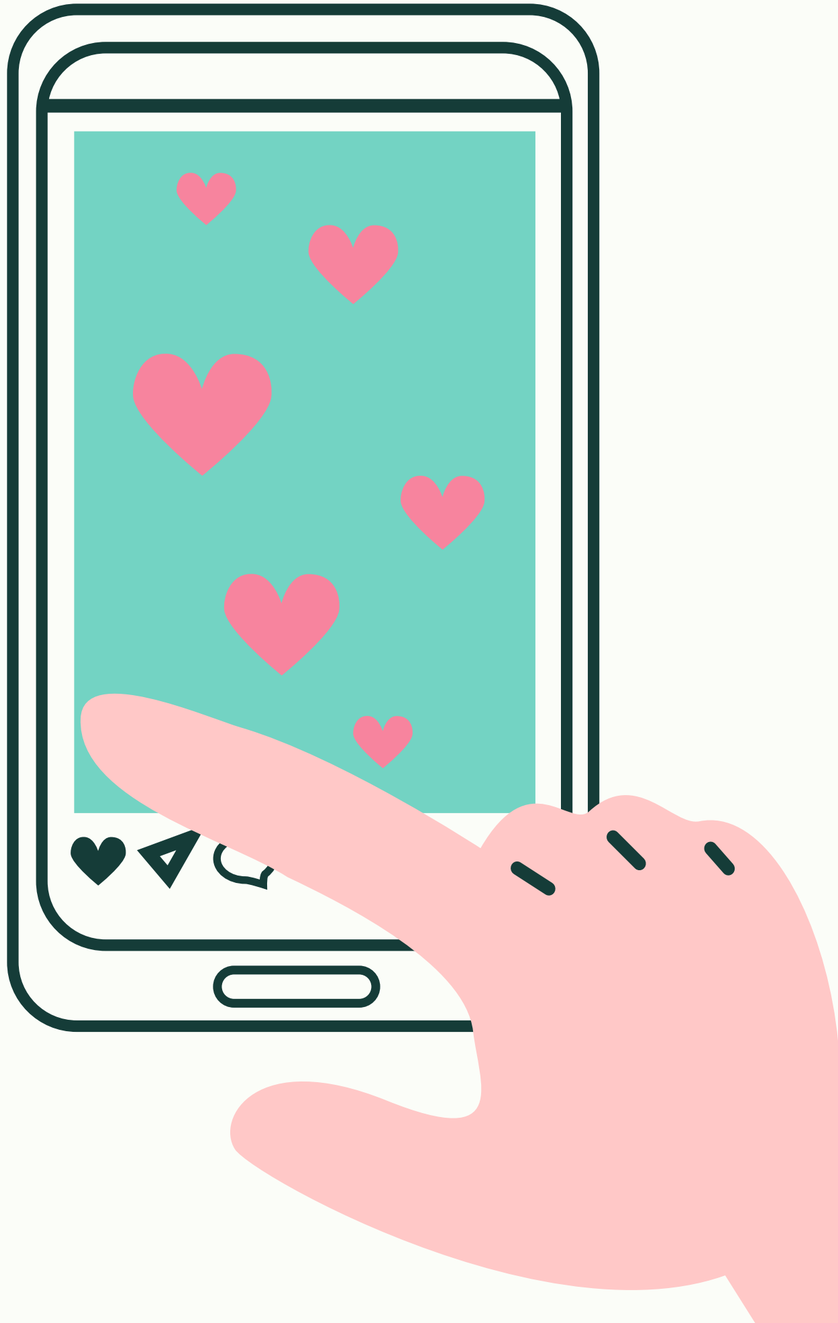
BEYOND LIKES AND SHARES

The Textual DNA of Viral Social Media Posts



Dhwani Sanjay Kariya
MSc Artificial Intelligence for Business
Data Analytics for Business
National College of Ireland

Business Problem

Key Question:
Which features of the text are the best contributors to virality?

Very low percentage of online posts that become viral.

Content creation often supported by speculation as opposed to being systematically practiced.

High costs of marketing often accompany poor penetration of the unpaid audience.

OBJECTIVES

Analyse Sentiment Vs Virality

Identify Engagement Traits

Compare Twitter Vs Reddit

Research Questions

Research Question 1

Does emotional tone affect Virality?

Research Question 2

What content types gain higher engaement?

Research Question 3

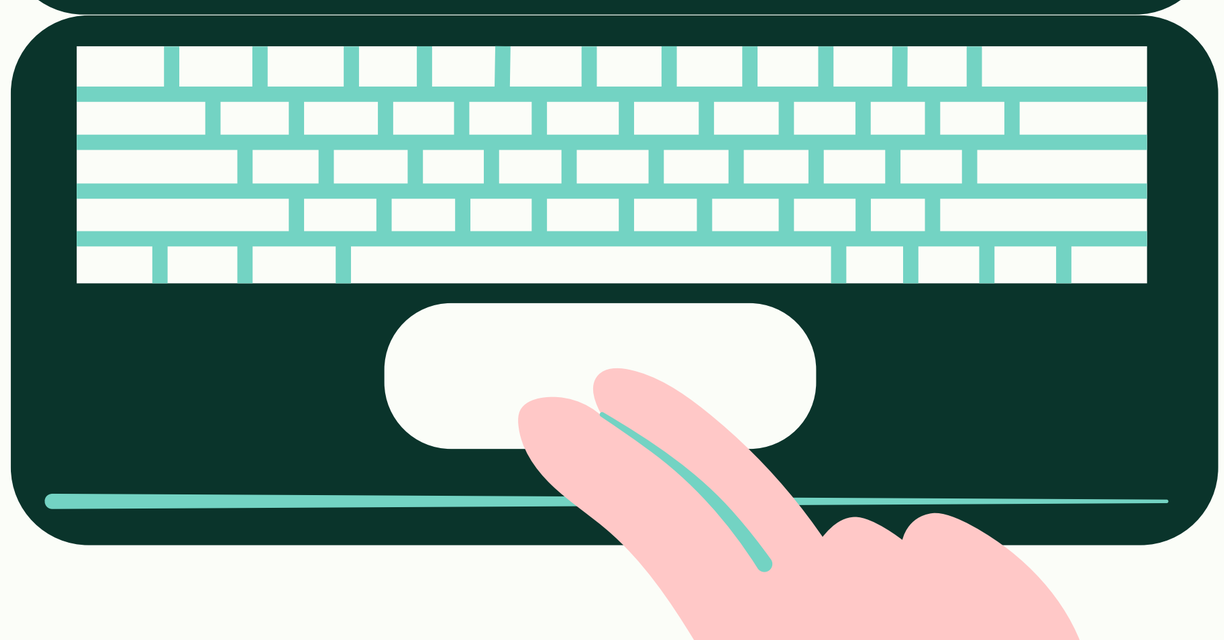
Are there platform differences?

DATA *used*

Datasets from Kaggle

- **Twitter US Airline Sentiment - Around 14,000 tweets**
- **Reddit posts (6 subreddits) - around 6,000 posts**

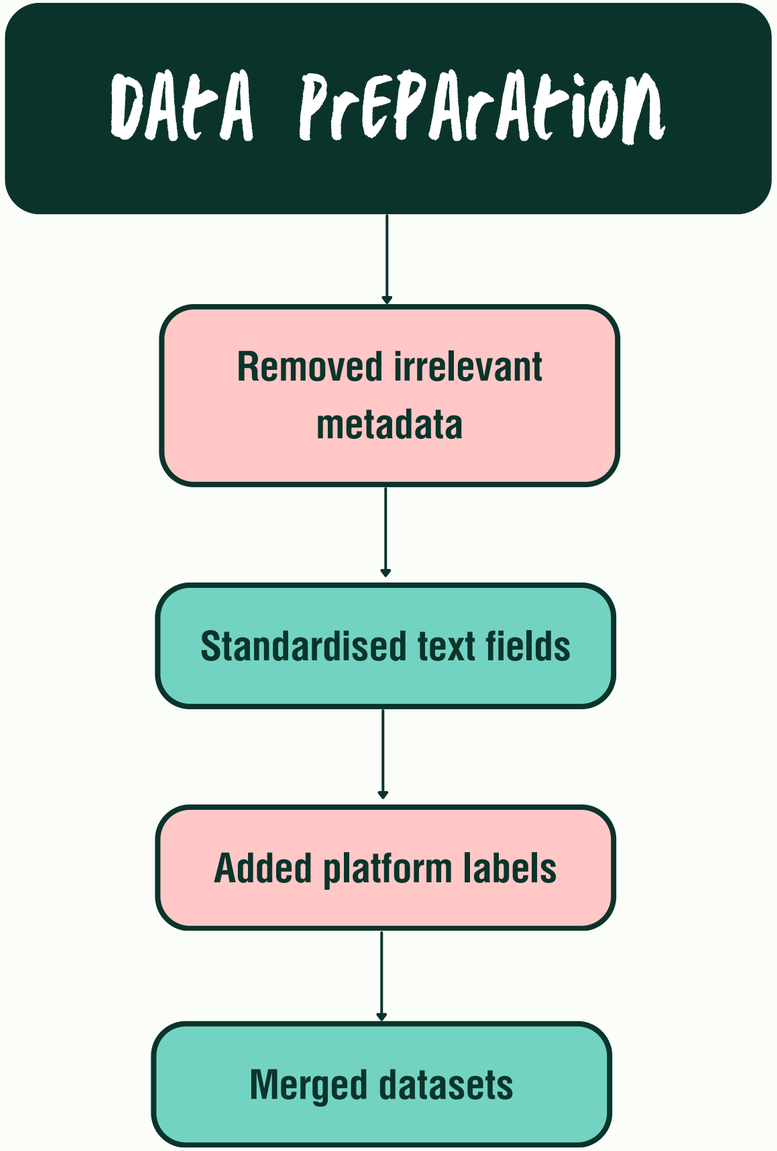
Final dataset of 20,571 posts (text only)



Data Preparation & Eda

Exploratory Data Analysis & Visualisation tool:

RapidMiner AI Studio



Volume

Twitter posts dominate the dataset posts by volume

Sentiment

Airline tweets skew strongly negative

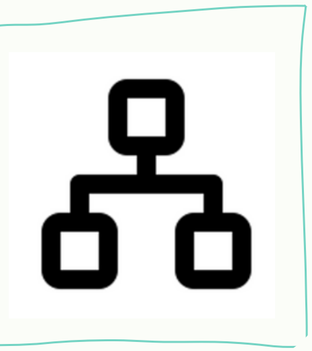
Content

Reddit shows diverse content styles

Frequency

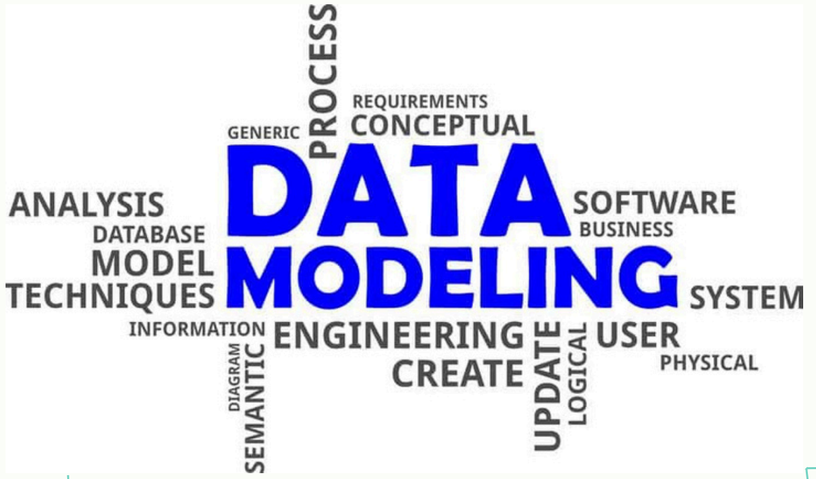
Frequent words confirm clean text

MODELLING Approach

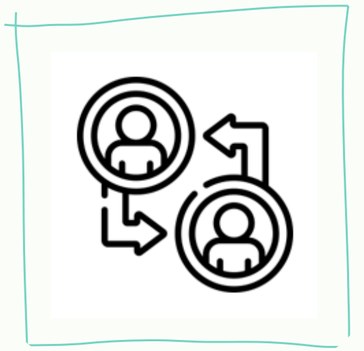


Viral Vs Non Viral

Binary Classification



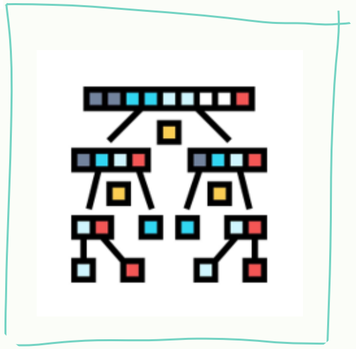
Top 5% by Engagement proxy



Virality Definition

Baseline model

Logistic Regression

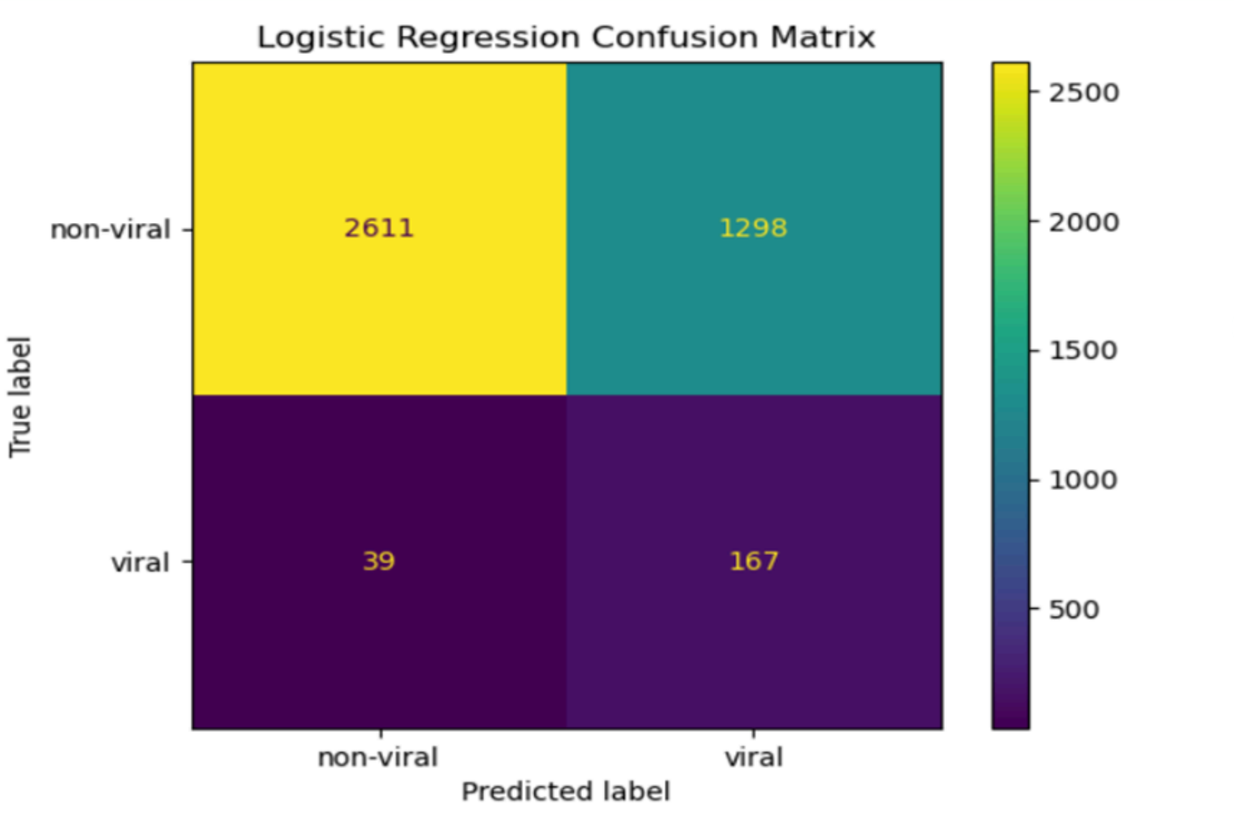


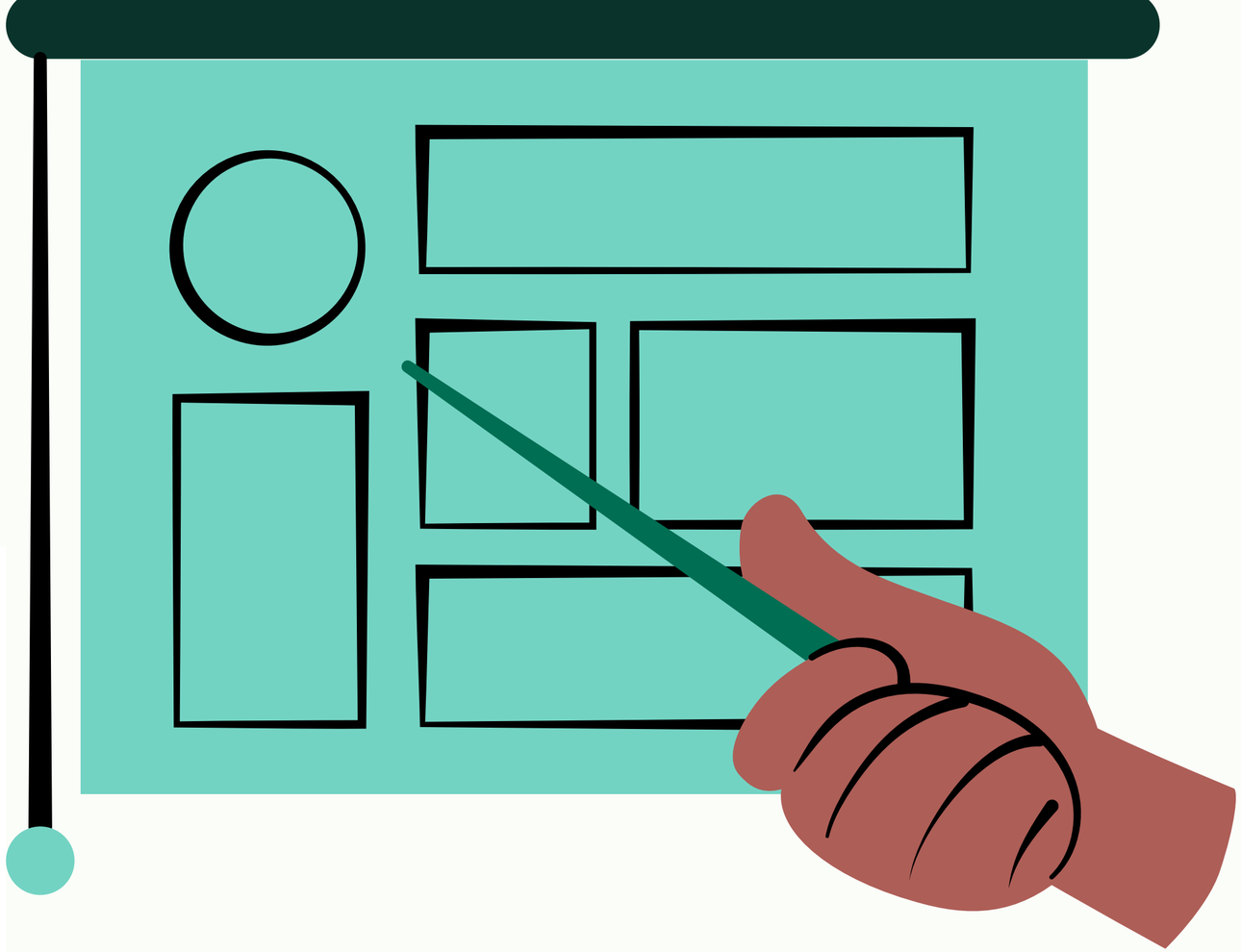
Non-linear model

Random Forest

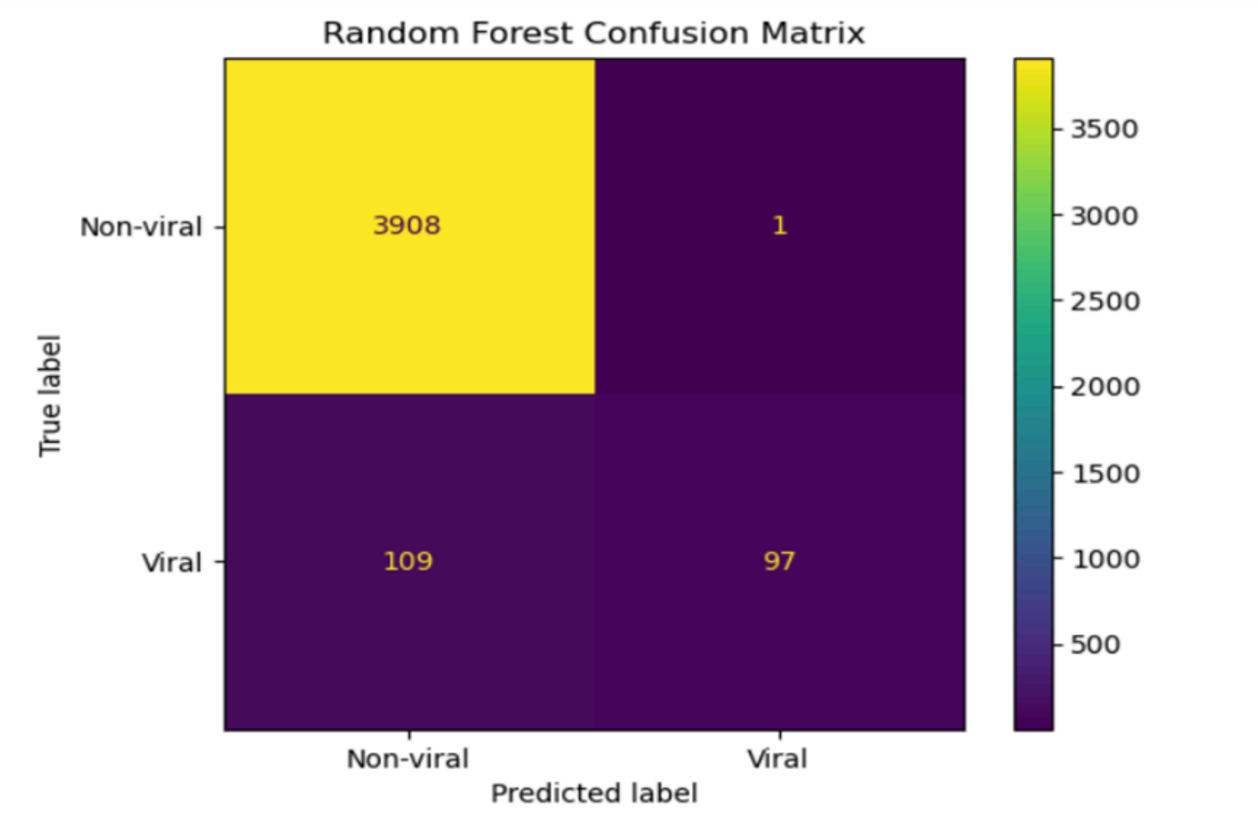
Logistic Regression Results

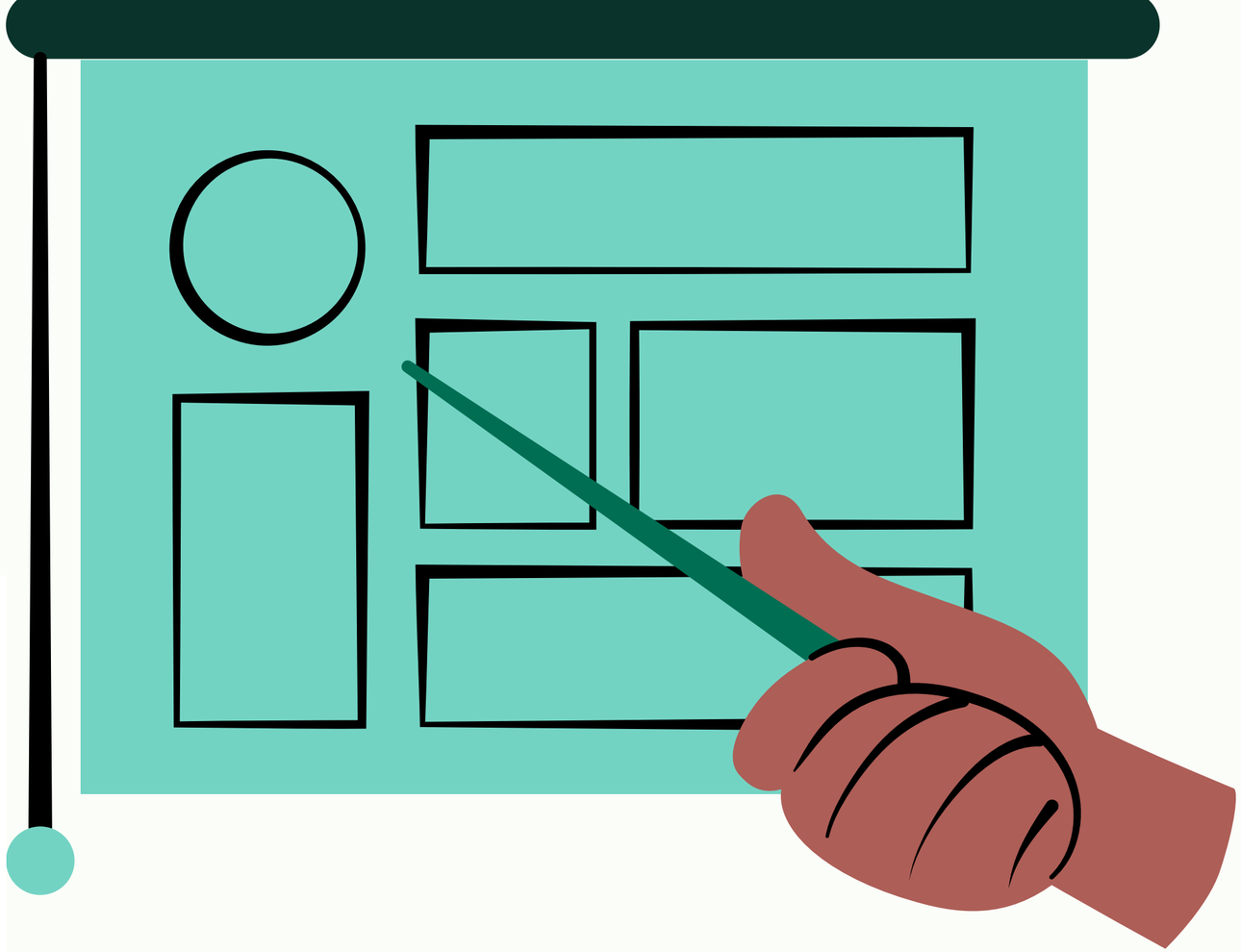
- **Strong** at non-viral post detection
- **Weak** at identifying viral posts
- Linear features alone are **not** sufficient





- **Showed** high ROC-AUC ~ 0.99
- **Better** at viral posts detection
- **Covers** the non-linear interactions





RANDOM FOREST
Results

Business Insights & Ethics

Insights

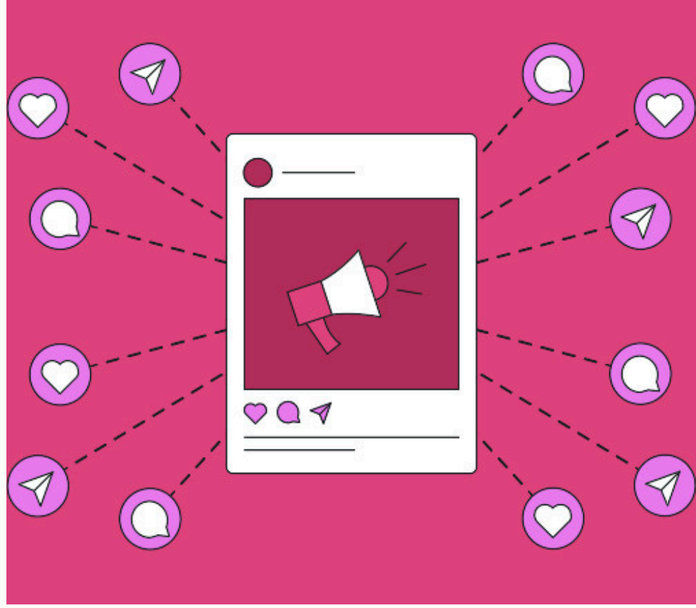
- Emotionally charged content has been found to be more diffuse.
- Long and wordy posts show better performance.
- Reddit tends to favor content that is being motivated by textual factors.

Ethical Considerations

- Platform Biasness
- Misclassifying sentiment due to algorithmic models
- Learning More Risk of amplifying negativity by means of algorithmic reinforcement

Project Conclusion

- Textual characteristics have significant viral effects
- Linear models are poor in comparison to non-linear models.
- Results favor a data-driven method when designing content.



Answering the research questions

RQ1: The strength of emotions is also discovered to slightly improve the virality of content.

RQ2: Content with a higher length of content, expensiveness, and sentiment density will be more likely to attain high engagement levels.

RQ3: Reddit seems to be of a distortion of encouraging the intensive discourse, and Twitter presents the pattern of engagement, owing majorly to the negative connotation.

THANK YOU