National College of Ireland

Project Submission Sheet

Student Name:	Dhwani Sanjay Kariya		
Student ID:	24238180		
Programme:	MSc. In Artificial Intelligence for Business	**Year:**	2025
Module:	Data Analytics for Business		
Lecturer:	John Kelly		
Submission Due Date:	19/12/2025		
Project Title:	Beyond likes and shares: the textual DNA of viral posts on social media platforms like Twitter and Reddit		
Word Count:	3055		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
<u>ALL</u> internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Dhwani Sanjay Kariya
Date:	19/12/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

[Insert Module Name]

[Insert Title of your assignment]

Your Name/Student Number	Course	Date

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Data Analysis for Business:

[Beyond likes and shares: the textual DNA of viral posts]

Author: Dhwani Sanjay Kariya,

MSc in Artificial Intelligence for Business,

National College of Ireland

Abstract

Social media is a very powerful channel of communicating with the masses, marketing, and raising awareness of a brand. However, only a small percentage of the posts publish virally, which in this case refers to the high metrics of engagement. The determination of virality is extraordinary in organisations that aim to create more powerful content and ultimately lessen the marketing costs.

This paper will consider the effects of text-based features, namely lexical choice, stylistic tone, and affective sentiment, on user engagement on Twitter and Reddit, as the two largest platforms. Two real-life datasets were acquired on the Kaggle repository: Twitter US Airline Sentiment which is a collection of 14,640 tweets, and six subreddit collections, which are the result of the Reddit source. A cleaning and consolidation process was followed, and a complete corpus of 17,949 posts was created that was to be analysed further.

Preprocess preliminary work was done to clean up extraneous metadata and to impute some missing values, text fields were standardised, and the schema was aligned across the platforms. RapidMiner AI Studio was used to perform the exploratory data analysis in order to describe the platform composition, sentiment distributions, and most frequently occurring lexical items. The resulting observations shed light on the early differences in behaviour on the platforms and establish some background for the discussion of the major questions of the research in the form of emotional tone, the content modality, and the cross-platform interaction.

Introduction

Social media has taken the spotlight in brand communication, but only a relatively small percentage of the posts get an undue proportion of visibility and engagement. This uncertainty creates efficiencies in organisations: it spends resources on marketing without the assurance of amplification. The practical business question that this project attempts to respond to is: what content characteristics, sentiment, wording, and topic do increase the likelihood of a post going viral, and is it the same characteristics on different platforms like Twitter and Reddit? It is applicable in the field of business decision-making because, according to the previous studies, emotional arousal and content attributes have a powerful impact on sharing behaviour (Berger and Milkman, 2012) ; (Jiménez-Zafra *et al.*, 2021). Moreover, the content diffusion behaviour is influenced by the platform affordances, such as fast reaction introduced by Twitter versus discussion format represented by Reddit (Nevatia *et al.*, no date). Each of these mechanisms will be well-understood to enable marketing teams to create posts that effectively fulfill the requirements of platforms

and appeal to emotions, which will have a greater organic and will decrease the use of paid media.

Objective
The study aims to measure sentiment vs. virality, determine topics/content styles related to increased engagement and compare Twitter and Reddit virality signatures.

Research Questions
1. How does the affective high tone of a post influence its likelihood of attaining viral propagation?
2. What types of content or subject matter show the relatively high engagement levels in varying social media platforms?
3. Do behavioural differences between Twitter and Reddit posts exist regarding their propensity to be viral?

Literature Review

Three interrelated topics include the literature that revolves around (1) affective motivators of sharing, (2) content attributes, i.e. multimedia products and (3) platform-specific mechanics and measurement choices.

Emotional drivers: According to the theoretical background, there is an upward tendency in the probability of sharing due to high-arousal emotions, be it positive or negative. This is supported by empirical studies on Twitter (Jiménez-Zafra et al., 2021) which reveal that variables related to sentiment affect the propensity to retweet using regression analyses, which in turn encourages the use of sentiment variables and the method of analysis followed in this paper.

Textual characteristics and video/audio: The current statement is substantiated by various empirical studies that show that the opportunity to have images, hyperlinks, or other multimedia items creates a powerful effect on the virality (Bruni, Francalanci and Giacomazzi, 2012).

Viral metrics and mechanics of platform: The measurement of the virality is criticised and refined in recent methodological studies (Elmas, Selim and Houssiaux, 2023) such as in the methods that multiply retweet ratios by author follower counts, and introduces transformer-based early-detection models. They also analyse numerous factors that contribute to virality, such as resonance and emotion, timing, and platform optimisation (Nevatia *et al.*, no date). In these studies, the choice of metrics (engineered engagement score and a top-percentile label of virality) and the approach to the modelling strategy in Part 2 are informed.

Methodology

4.1 Selection and extraction of data

To attain cover across platforms, two data sets that are publicly available were chosen:
- **Twitter US Airline Sentiment** (about 12,000 tweets): this contains text of the tweet, label of the airline, and human-labelled sentiment, making it easy to verify the sentiment (manually) as well as analyse it (analysis of human sentiment distribution).

• **Reddit** Top Posts (six subreddits: AskReddit, TodayILearned, Technology, WorldNews, Funny, Memes): bundled together to compose a wide-structure of the topical corpus.

The data were obtained at Kaggle, the textual field of the datasets was normalised and later the complete dataset was compiled into one CSV file, called combined_cleaned_dataset.csv, which provides a foundation of the exploratory data analysis and representation. The cross-platform setup allows making a comparative study of text-based drivers despite the fact that multimodal drivers (images or videos) which are described by the literature do not appear in this purely textual study.

4.2 Data cleaning and preprocessing

This step is to clean and preprocess the data. The entire cleaning and preprocessing was done in a Jupyter notebook, which transformed the raw Twitter and Reddit data into a coherent and analysis-friendly format.

1. Selecting Relevant Columns
- Reddit: Only title and subreddit were retained, most other fields (e.g. selftext, author, created_utc) were either meaningless or largely vacuous.
- Twitter: Text, airline, and airline sentiment were only retained. The following chosen areas are directly connected to the content, topic, and sentiment of the post being investigated - the three key variables to investigate the trends related to virality.

2. Standardising Text Fields
In Reddit, the column title is used whereas in Twitter, it is done by text. In order to bring them together, the title page of the Reddit website was changed to the text, and both websites formed a single column of texts.

3. Adding a Platform Label
The platform level differences could be analysed later by introducing a new column source (such as the reddit and twitter).

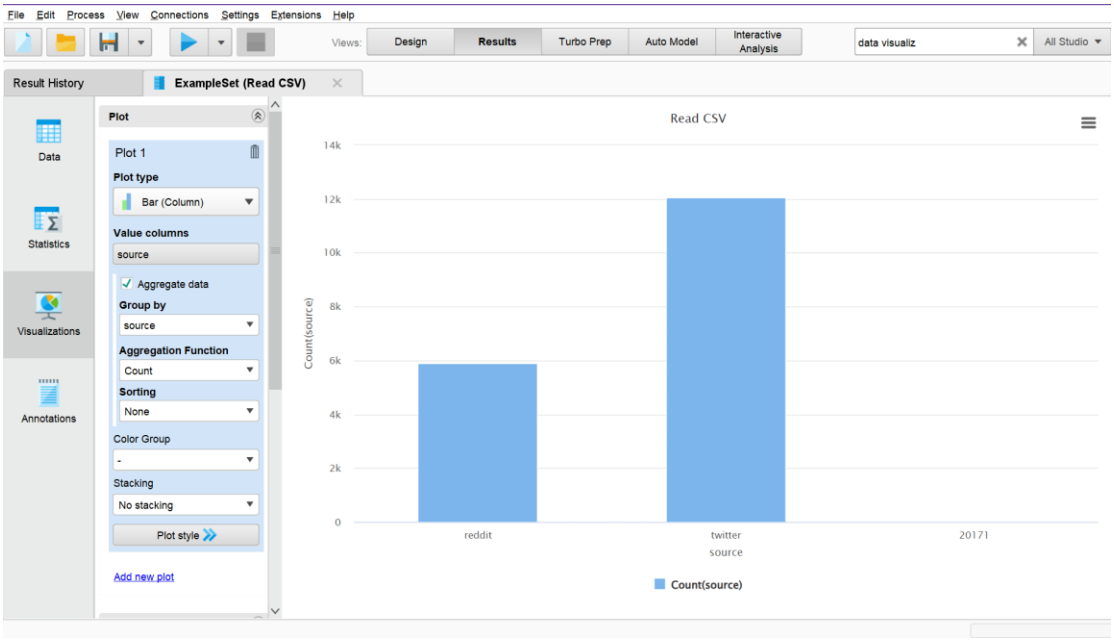
4. Handling Missing Values
In Reddit, the missing values were in some metadata fields, whereas the text field was still extant, as only the helpful columns were chosen. The retained columns also did not have missing values in Twitter. As such, imputation was not necessary.

5. Merging the Datasets
A combination of the two cleaned datasets was performed with the aid of concatenation making a total of 17,949 posts. This gives enough depth to statistical and sentiment-based exploration.

6. Exporting the visualisation
The resulting cleaned data were stored as the combined_cleaned_dataset.csv and loaded into RapidMiner to do some EDA.

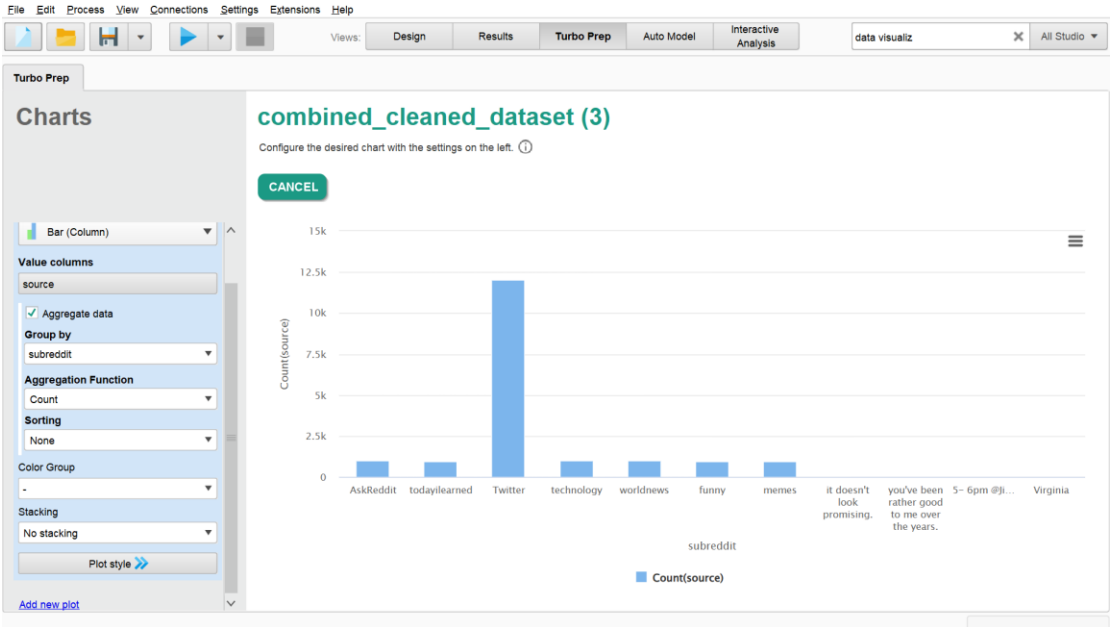
4.3 Exploratory Data Analysis (EDA)
RapidMiner AI Studio was used to undertake Exploratory Data Analysis (EDA). The fundamental EDA plots and cursory points of interpretation consist of:

Plot 1: Plot Distribution (Twitter vs Reddit)



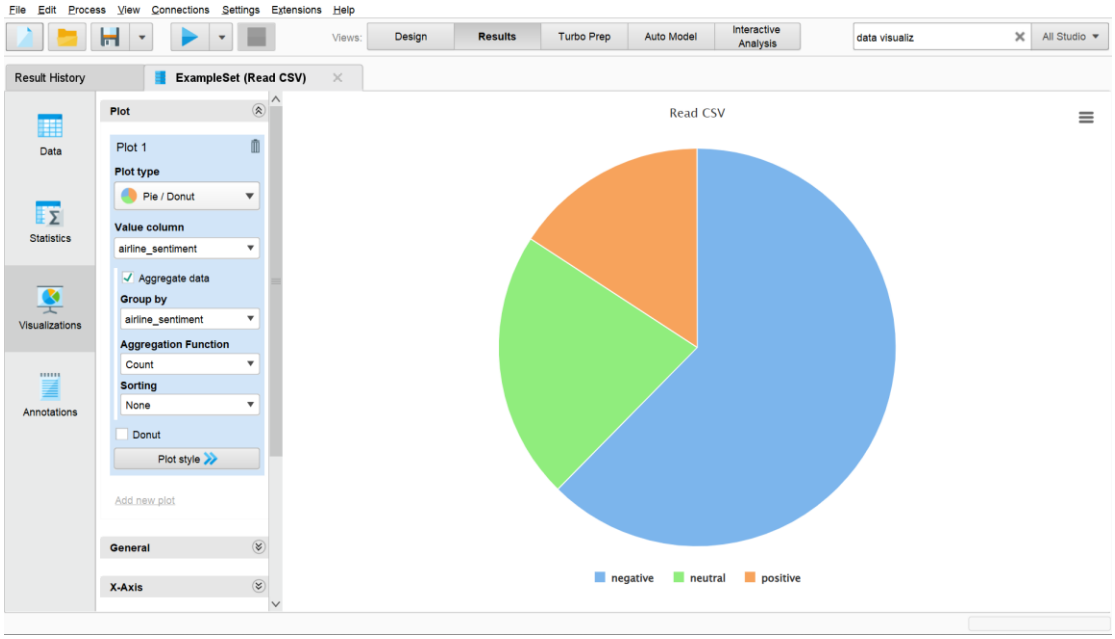
The combined data shows that the Twitter data is contributing more records than the Reddit ones, although the airline data on Twitter is quite compact but complete. This imbalance counts because using aggregated statistics, eg sentiment scores and frequency counts of words, is going to fall towards the platform with the higher volumetric number.

Plot 2: Reddit subreddit distribution



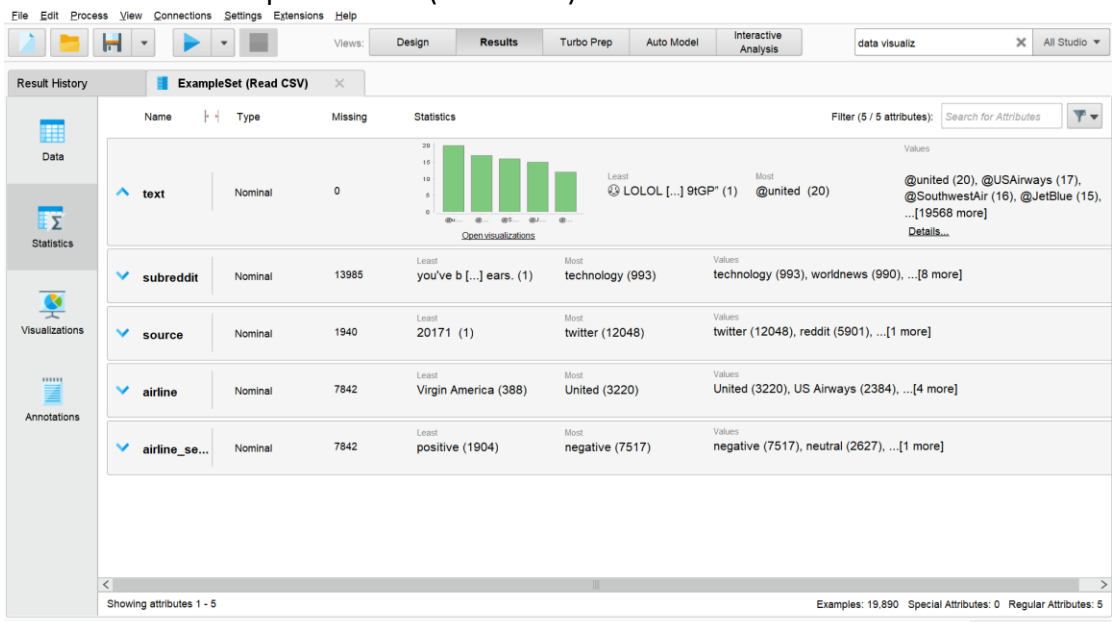
A bar chart is used to show the number of posts made by each of the chosen Reddit subreddits. Communities like AskReddit, memes, worldnews, funny display moderate traffic. The values labelled as Twitter do not have any subreddit.

Plot 3 -Twitter sentiment distribution.



As the visualization shows, the probability of Twitter airline posts being retweeted in a specific context (negative) is higher, hence the need to include sentiment as a predictive in Part-2 modelling.

Plot 4 -the most frequent words (combined).



Frequent word lists include conversational tokens (people, think, time), indicating that the text cleaning process was developed and offered a guide to downstream topic modelling.

4.4 Analytical Techniques and Modelling Methods

This phase is done after cleaning, engineering features, and exploratory analysis processes performed during Part 1 because after cleaning and analysing the data, the next step is to apply predictive analytics to single out the textual variables that drive social media virality.

An analytical problem was posed in the form of binary classification, in which posts were tagged with viral or non-viral posts. A post was considered viral when it ranked within the top 5% of the observations that were sorted by an engineered gauge of engagement based upon sentiment magnitude and text length. The approach allows

virality to be modelled without those platform-specific measures of engagement (likes or shares).

In Jupyter Notebook, there were two Python operations to create two supervised machine-learning models:

• Logistic Regression, chosen as a baseline model because of such characteristics as interpretability and a need to be used in text-based classification.

• The random forest Classifier, which was selected with the aim of identifying non-linear interactions and relationships among textual-based features and sentiment-based features.

Feature Representation TF-IDF was applied to textual data as it was being turned into data vectors (TF-IdF) unigrams and bigrams, in order to identify both singular and short phrases. Besides TF-IDF features, a few meta- features that had been engineered were added like sentiment (VADER compound) score, word count, character length, hashtag count, mention count and exclamation count and presence of a question mark.

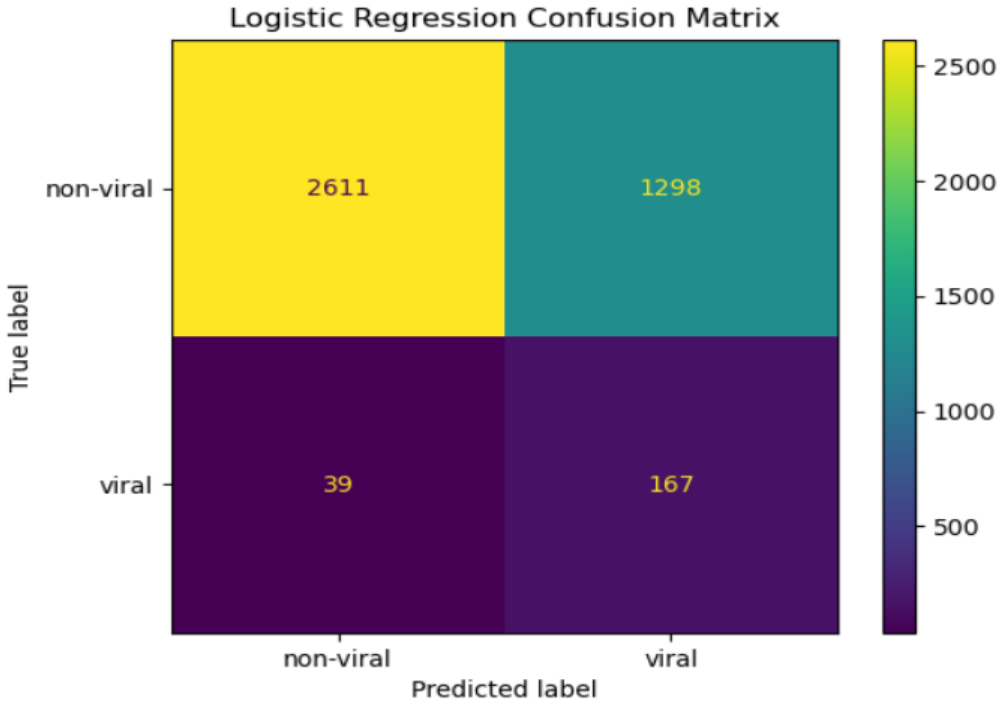
The final feature matrix was a combination of sparse TF-IDF features and dense meta-features, which is a complete representation of both semantic and structural properties of each post. Train Test Split and Evaluation Strategy. To maintain the viral and non-viral post proportion in the dataset, it was divided into 80 per cent train dataset and 20 per cent test dataset, with the sampling being stratified. There were several measurements of model performance, accuracy, precision, recall, F1-score, ROC–AUC and confusion matrices. Considering the high level of imbalance in the classes (about 5 % viral posts) accuracy was not applied as the main performance measure. Rather, recall, F1-score, and ROC-AUC were given priority because they give a more valuable measure when there is an imbalance in the conditions.

Analysis and Results

5.1 Logistic Regression Results

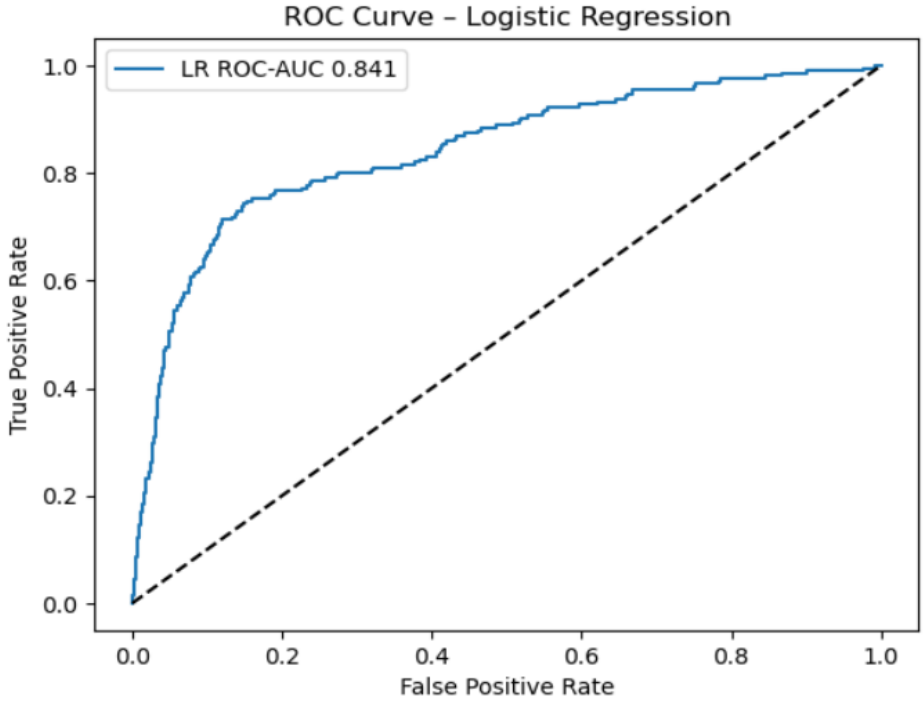
The logistic regression was used as a control model to determine whether there were any linear relationships between text features and popularity.

Plot 5 -LR Confusion Matrix



The above matrix shows that the model correctly classified most non-viral posts but was unable to identify viral posts and therefore its number of false negatives is great. Such behaviour is indicative of few viral posts, and the constraints within linear decision boundaries to model nonlinear engagement dynamics.

Plot 6 -LR ROC Curve



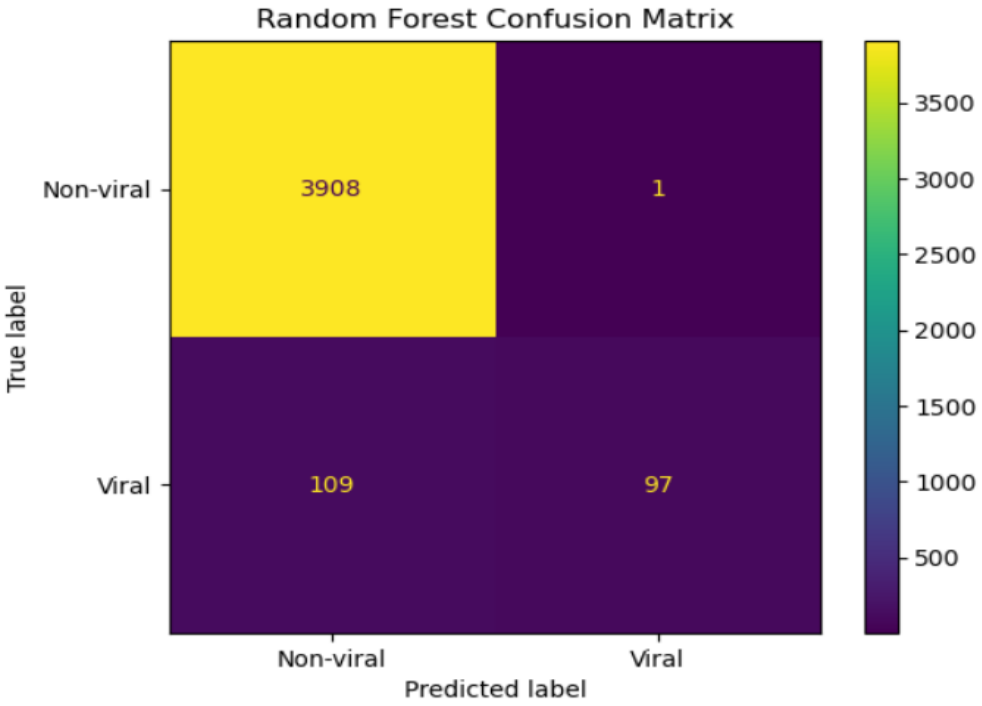
The above curve has an average degree of discriminative power, which means that phenomenon like sentiment and text length do impact the virality but would not be adequate factors in explaining the viral phenomenon.

Noted result: The identified outcome implies that simple textual and sentiment-based indicators do contribute to the virality; however, they do not always have a linear effect on engagement.

5.2 Random Forest Results

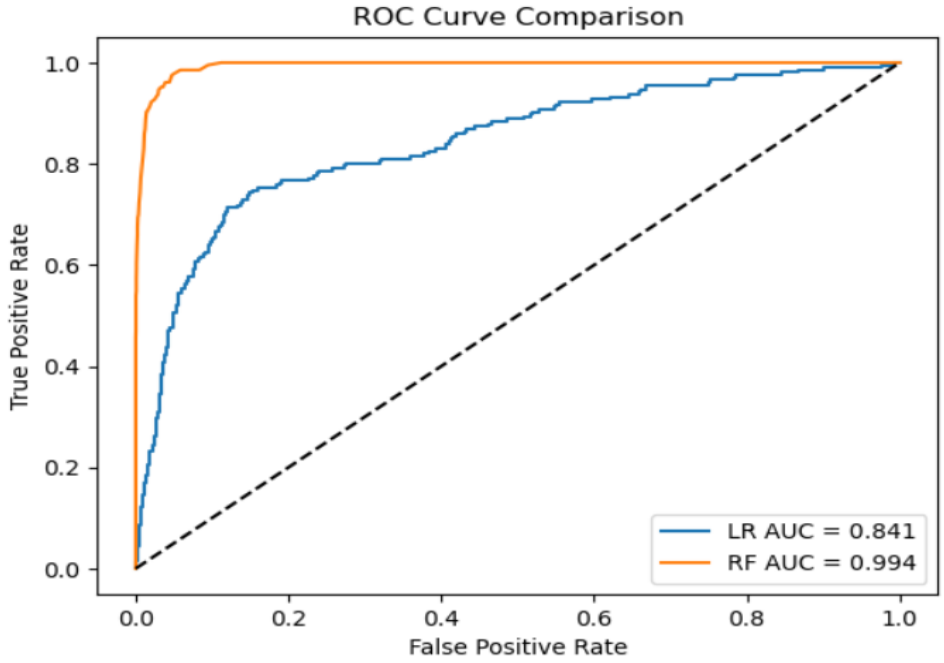
Random Forest model was better than the Logistic Regression in all the evaluation measures.

Plot 7 -RF Confusion Matrix



The above matrix shows that there are almost perfect classifications of non-viral posts and better issues of viral posts.

Plot 8 -ROC -AUC



The ROC-AOC award is also very high, which means that the ranking is highly achieved and the separation between viral and non-viral information is successful.

The analysis of feature importance gives the findings that the most significant predictors are: sentiment intensity, word count and character length, attaining punctuations (exclamation marks and questions) and emotionally charged words.

Noted result: Non-linear interplay between the emotional tone of content and its structural content is largely stimulated by virality and can thus be summarised by ensemble algorithms like the random forest.

5.3 Platform-Level Comparison

Comparing the proportion of virality on digital platforms suggests that the posts in Reddit have a higher tendency to be considered viral than posts in Twitter. The reason behind this deviation would be recklessly generous policy of Reddit in terms of longer discussion-driven content, compared to restrained participation in Twitter, which would be predetermined by extrinsic factors like the arrangement of followers networks and multimedia add-ons.

Noted result: Attributes that are based on text only give a better predictive metric of virality on Reddit, compared to Twitter.

Discussion

The result of the findings is that textual characteristics (emotional intensity, content length, and use of expressive language) play an important role in the virality of social media. The fact that the superior performance of the Random Forest model was reached supports the idea that it operates based on non-linear interdependencies as opposed to linear relationships between virality. However, there exist a number of limitations that should be mentioned. The first one is that the engagement metrics were determined through the proxy as opposed to the real interaction records. Second, the mechanisms that are platform-specific, including algorithmic curation, the topology of follower network, and multimedia content, were not analysed. Third, the work assumes that the text is the key factor of engagement, whereas this assumption might not be universally true on different social media websites. Nevertheless, the findings are strong within the text-focused analytics boundaries and have a lot of parallels with the available literature on emotional content and online interactions.

6.1 Ethical Consciousness and Bias Awareness

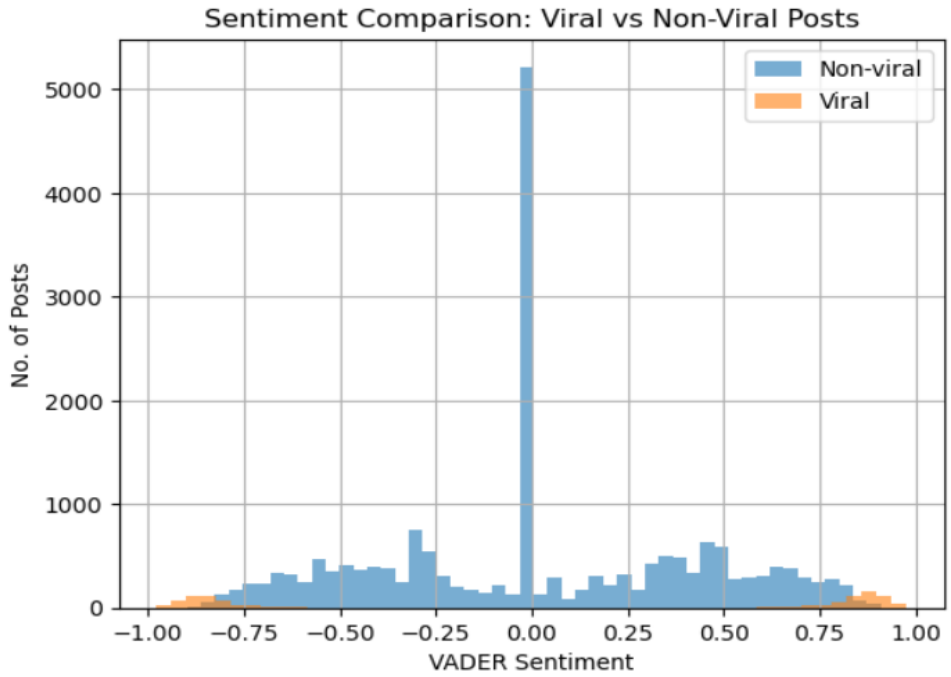
Various ethical and bias-related factors are identified in this investigation. The data is provided by Reddit and Twitter, whose audiences are not the ones that are representative of the general population. In turn, the linguistic and behavioural biases that are platform-specific can be presented. The sentiment analysis tool was done using the VADER lexicon; it works with informal English language, though this tool can fail to uncover sarcasm, cultural expression or subtle context. Moreover, since there are no standard metrics of engagement between platforms, a proxy of virality at the level of sentiment magnitude and text length was used. Even though it is appropriate in the exploratory analysis, such simplification can over-represent emotionally extreme or longer posts. Lastly, predictive models of virality can be very dangerous to propagate sensational or a negative kind of content in case of misuse. Trusting to this, this model aims at generating analytic sensitivity and not content optimisation directly thus it is aligned to ethical AI practices that embrace bias education and engaging features of usage (Mehrabi *et al.*, 2022).

Conclusion

This paper analysed the textual predictors of viral content on social media by pooling together Twitter and Reddit data, specifically the sentiment and content characteristics and cross-platform behaviour.

Research Question 1: How does the affective high tone of a post influence its likelihood of attaining viral propagation?
The evaluation shows that there is a positive correlation of emotional intensity with post virality. Exploratory statistics indicate a further presence of significantly higher absolute VADER sentiment scores in virus posts as opposed to non-viral posts and the trend cuts across both media. The central tendency of viral posting is biased towards positive and negative ends of sentiment comparison plots, implying that more emotive content is better placed to attract the attention of the audience.

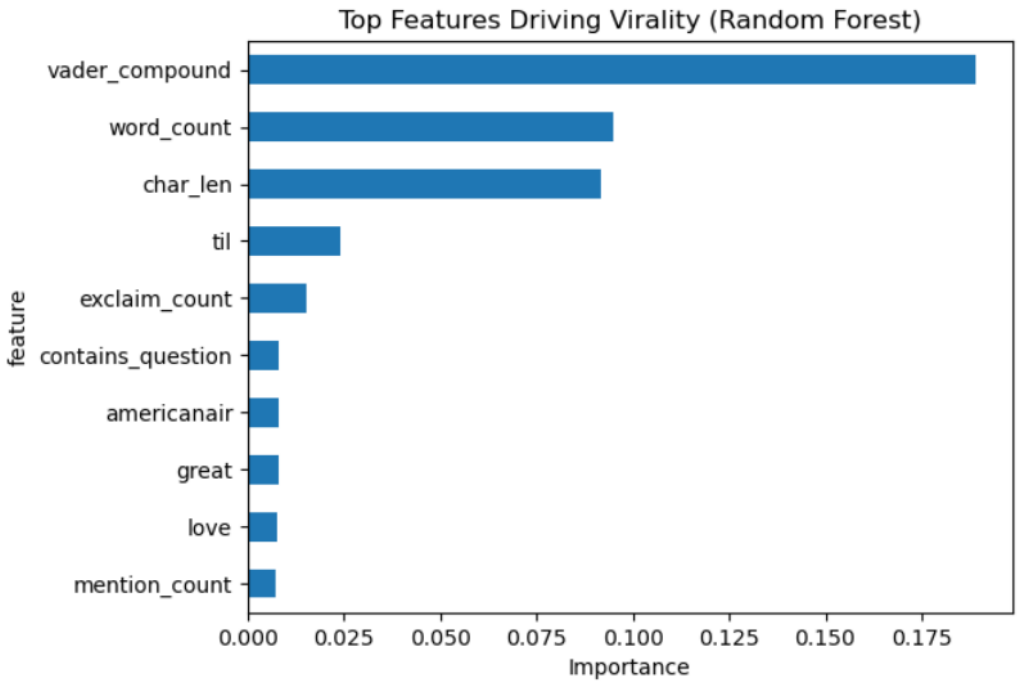
Plot 9 -Sentiment Comparison



This relationship is supported by predictive modelling, and the impression qualities are the key point of classification accuracy.

Research Question 2: What types of content or subject matter show the relatively high engagement levels in varying social media platforms?
The data reveal that post length and expressive cues are important determinants which are structural features. Relative comparisons of the posts on the basis of their viral and non-viral posts, descriptive statistics and comparison of groups indicate that, compared to the other posts, viral posts are longer and more expressive and, in many cases, contain neither punctuation or, at least, interrogatives.

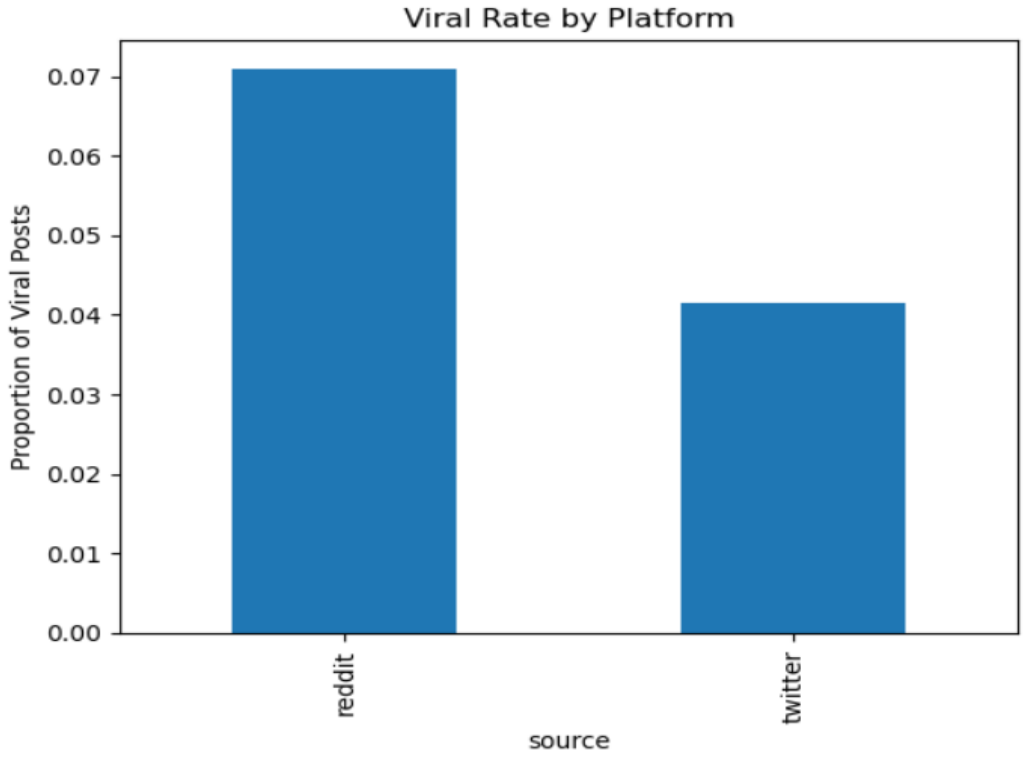
Plot 10 -Features driving the virality of Posts



Combined, these properties have a beneficial effect on TF-IDF-based text representations based models (which opens the possibility that both semantic content and its form of expression have effects on virality).

Research Question 3: Do behavioural differences between Twitter and Reddit posts exist regarding their propensity to be viral?
Platform-level examination gives unique behavioural patterns. Posts to Twitter have a more negative tone of voice particularly when it comes to the airlines as compared to posts to Reddit where the communities have a high amount of thematic and stylistic diversity. Regardless of these differences, predictive patterns are universal across platforms which denotes that the essential text based drivers of virality cut across a niche in social media platforms. The prevalence of the viral material of the Reddit indicates that the longer textual forms of substance could indicate a high level of virality with the short-textual forms of virality indicated in tweets, which supports the validity of a text-only computational method towards the Reddit data.

Plot 11 -Virality by Platform



Altogether, it is possible to note that the Random Forest model is better than the Logistic Regression due to the non-linear drivers of virality. In the business sense, the findings show that the content that contains emotional depth and well-structured organisation has a higher probability of going viral. Such learnings can be used to create better, more competent and data-driven content strategies that consider ethical and platform-specific limitations.

7.1 Data Analytics Artefact

The analytical artwork as a result of doing so in this study is a predictive virality classifier, written in Python, in a Jupyter Notebook. The artefact includes an extensive data analytics pipeline, i.e., data preprocessing, feature engineering, TF -IDF text-vectorisation, and trained machine-learning models. The last random forest model was retained as a reusable object and could be implemented on new and unseen social media posts to determine the likelihood of viral spread on the basis of text characteristics only. The artefact denotes the use of evaluation measures, such as confusion matrices and ROC curves, as the basis of transparency and interpretability. This artefact aids in the pre-publication evaluation of the content and allows evidence-based decision-making when forming a strategy in social media.

References

Berger, J. and Milkman, K.L. (2012) "What Makes Online Content Viral?," Journal of Marketing Research, 49(2), pp. 192–205. Available at: https://doi.org/10.1509/jmr.10.0353.

Bruni, L., Francalanci, C. and Giacomazzi, P. (2012) "The Role of Multimedia Content in Determining the Virality of Social Media Information," Information, 3(3), pp. 278–289. Available at: https://doi.org/10.3390/info3030278.

Elmas, T., Selim, S. and Houssiaux, C. (2023) "Measuring and Detecting Virality on Social Media: The Case of Twitter's Viral Tweets Topic," in Companion Proceedings of the ACM Web Conference 2023, pp. 314–317. Available at: https://doi.org/10.1145/3543873.3587373.

Jiménez-Zafra, S.M. et al. (2021) "How do sentiments affect virality on Twitter?," Royal Society Open Science, 8(4), p. rsos.201756, 201756. Available at: https://doi.org/10.1098/rsos.201756.

Mehrabi, N. et al. (2022) "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys, 54(6), pp. 1–35. Available at: https://doi.org/10.1145/3457607.

Nevatia, A. et al. (no date) "Exploring the Key Determinants of Social Media Content Creation: Analyzing Their Influence on Content Virality."