TWITTER TREND ANALYSIS AND

HASHTAG RECOMMENDATION

SPRING 2023

Prepared for

Associate Professor Walaa H. Elashmawi

Eng. Ziad Elgayar

Twitter Trend Analysis and Hashtag Recommendation

This document is a detailed document from A-Z for our "CSE 472: Artificial Intelligence" semester project which we chose to be implementing a Twitter Trend Analysis and Hashtag Recommendation system.

This Project is Brough to You By

Mina Shawket Shaker	*20P9476*
Sarah Sherif Mohamed	*20P2202*
Marita Osama Raouf	*20P2455*
Malak Mohamed Helmy	*20P2434*
Ahmed Hossam Sakr	*20P1009*

We would like to give a special thanks to Associate PROF. "Walaa H. El Ashmawi", and ENG. "Ziad Elgayar" for always responding to our messages and queries during the whole period of the project and never ignoring us.

You can find the whole project files in the following repo:

https://github.com/mina58/Hashtag-Recommender

Table of Contents

Introduction and Background 4

Objectives 5

System Overview 6

Methods 7

- *Data Gathering* 7
- *Data Preprocessing* 7
- *Machine Learning Model* 8

Results and Evaluation 10

- *Output Samples* 11
- *Confusion Matrices* 13
- *Model Report* 15

Introduction and Background

The project is being undertaken to address the opportunity of increasing visibility and engagement on Twitter for businesses and individuals. Social media platforms like Twitter offer a vast audience for sharing content, but it can be challenging to identify relevant trends and generate effective hashtags to reach the target audience. By leveraging AI and ML techniques and data gathering, the project aims to analyze Twitter trends and recommend appropriate hashtags, providing a solution to the problem of low visibility and engagement on the platform.

The practical application of the Twitter Trend Analysis and Hashtag recommendation project lies in its ability to help businesses and individuals enhance their social media presence and reach on Twitter. In the real world, businesses across various industries rely on social media platforms to connect with their target audience, promote their products or services, and build brand awareness. By analyzing Twitter trends and generating relevant hashtags, the project can offer valuable insights into the current interests and discussions happening on the platform.

For instance, a fashion retailer can use the project's recommendations to identify trending fashion styles, celebrity fashion events, or seasonal fashion trends. They can then align their content, such as tweets showcasing new arrivals or fashion tips, with these trends and include the recommended hashtags to increase the visibility of their tweets. This strategic approach can help the retailer attract a larger audience, drive engagement with their brand, and potentially increase sales.

Similarly, individuals such as content creators, influencers, or professionals can leverage the project's capabilities to align their content with trending topics and generate relevant hashtags. This can help them increase their online visibility, attract a wider audience, and establish themselves as thought leaders or experts in their respective fields.

By providing data-driven insights and automated hashtag generation, the project empowers businesses and individuals to optimize their Twitter strategies and maximize their reach in the real world, ultimately contributing to increased visibility, engagement, and potential business growth.

Objectives

1. *Develop a system to gather data on Twitter trends and associated hashtags in real-time.*
2. *Preprocess and clean the data to remove noise and irrelevant information.*
3. *Apply machine learning techniques to identify patterns in the Twitter trends and associated hashtags.*
4. *Develop a recommendation engine that suggests relevant hashtags based on the identified patterns.*
5. *Evaluate the performance of the recommendation engine in terms of accuracy and relevance of the recommended hashtags.*
6. *Provide a user-friendly interface for users to interact with the system and view the recommended hashtags.*

System Overview

1. ***Data gathering****: The data gathering phase involves using Twitter's API to gather data on trending topics and associated hashtags. This data is then stored in a CSV File.*

2. ***Data preprocessing****: The data preprocessing phase involves cleaning and preprocessing the data to remove noise and irrelevant information. This phase could involve techniques such as lemmatization, stemming, and stop word removal. The preprocessed data is then stored in the CSV file.*

3. ***Machine learning****: In this phase, machine learning algorithms are applied to the preprocessed data to identify patterns in the trends and associated hashtags. This phase could involve unsupervised learning techniques such as clustering or topic modeling. The machine learning model is then trained on the preprocessed data.*

4. ***Hashtag recommendation engine****: The machine learning model developed in phase 3 is used to generate hashtag recommendations based on the identified patterns. The recommendations are relevant to the topic being discussed and increase the user's visibility and engagement on Twitter. The recommended hashtags are then stored in a separate database.*

5. ***User interface development****: A user-friendly interface is developed for users to interact with the system and view the recommended hashtags. This could involve developing a web-based application or integrating the system with existing social media management tools.*

6. ***Integration****: The data gathered in phase 1, preprocessed in phase 2, and the recommended hashtags generated in phase 4 are integrated into the user interface developed in phase 5.*

7. ***Evaluation****: The performance of the system is evaluated in terms of accuracy and relevance of the recommended hashtags. This involves testing the system on real-world Twitter data and comparing the recommended hashtags to manually generated ones.*

Methods

Data Gathering

***Twitter API**: The project utilizes the Twitter API to gather real-time data on trending topics and associated hashtags. The API provides access to a wide range of Twitter data, including trends, tweets, and metadata.*

Data Preprocessing

***Tokenization**: The collected tweet data is tokenized, where individual words or tokens are extracted from the text. This process breaks down the text into smaller units, facilitating further analysis and processing.*

***Stopword Removal**: Stopwords, such as common words like "the," "and," or "is," which carry little semantic meaning, are removed from the tokenized data. This helps reduce noise and focuses on more meaningful terms.*

***Stemming & Lemmatization**: The project applies stemming and lemmatization techniques to reduce words to their base or root form. This normalization step helps consolidate variations of words and improves consistency in the dataset.*

***Removal of Special Characters and URLs**: Any special characters, punctuation marks, or URLs present in the tweet data are eliminated. This ensures that only relevant textual content is retained for further analysis.*

***Handling Emojis and Emoticons**: Techniques are employed to handle emojis and emoticons in the tweet data. This could involve converting them to textual representations or mapping them to relevant categories.*

***Handling Hashtags and Mentions**: The project handles hashtags and user mentions by either removing them or preserving them separately for potential analysis. This allows for focused analysis on the text content while considering the contextual information provided by hashtags and mentions.*

***Filtering Irrelevant Information**: Any irrelevant information, such as retweets, duplicate tweets, or non-English tweets, is filtered out to ensure the quality and relevance of the dataset.*

***Data Normalization**: The preprocessed data may undergo further normalization steps, such as lowercasing all text, eliminating case sensitivity, and standardizing the data for consistent analysis.*

Machine Learning Model

The model must satisfy two constraints. First, we need to recommend a hashtag that is relevant to the given tweet and second the recommended hashtag must be currently trending. For that reason, we need to gather data every time interval and after each gathering, we need to train the model from the start to be able to only recommend currently trending hashtags. We automated this process (gathering and training) to gather tweets every 6 hours and to train the model afterwards. This way the model is always up to date.

Our model is RCNN (Recurrent Convolution Neural Network) which combines the strengths of both convolution and recurrent neural networks. Here are the layers of the model:

***Embedding Layer**: This layer takes the input words and maps them to continuous vectors called word embeddings. It helps capture the semantic meaning of words and their relationships within the context of the task at hand.*

***Conv1D Layer**: This layer applies 1-dimensional convolutions to the input sequence of word embeddings. It slides a small window (kernel) over the sequence and performs element-wise multiplications and summations, capturing local patterns or features. The activation function 'relu' (Rectified Linear Unit) introduces non-linearity to the output of the convolutions.*

***AveragePooling1D Layer**: This layer performs average pooling on the convolutions' output. It reduces the dimensionality of the feature maps while retaining the most important information.*

Average pooling computes the average value within a specific window size, further condensing the representation.

***Dropout Layer**: Dropout is a regularization technique used to prevent overfitting. It randomly sets a fraction of input units to 0 at each update during training, which helps in reducing the model's reliance on specific features and improves its generalization capability.*

***Bidirectional LSTM Layer**: Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that can capture long-term dependencies in sequential data. The bidirectional aspect means that two separate LSTM layers are used—one processing the input sequence in the original order and the other processing it in the reverse order. This allows the model to capture information from both past and future contexts.*

***Dropout Layer**: Another dropout layer is applied after the LSTM layer to further prevent overfitting and enhance the model's robustness.*

***Dense Layer**: This fully connected layer maps the output of the previous layers to the desired number of output classes. The activation function used here is 'softmax', which produces a probability distribution over the output classes, indicating the likelihood of each class.*

Results and Evaluation

1. ***Data Gathering****: The success of this phase can be evaluated based on the quantity and quality of the Twitter data gathered. Indicators could include the number of tweets collected, the variety of topics covered, and the accuracy of the data.*

2. ***Data Preprocessing****: The success of this phase can be evaluated based on the quality of the preprocessed data. Indicators could include the reduction in noise and irrelevant information, the accuracy of lemmatization and stemming, and the completeness of the data after preprocessing.*

3. ***Machine Learning****: The success of this phase can be evaluated based on the effectiveness of ML techniques used, and the choice of the model architecture.*

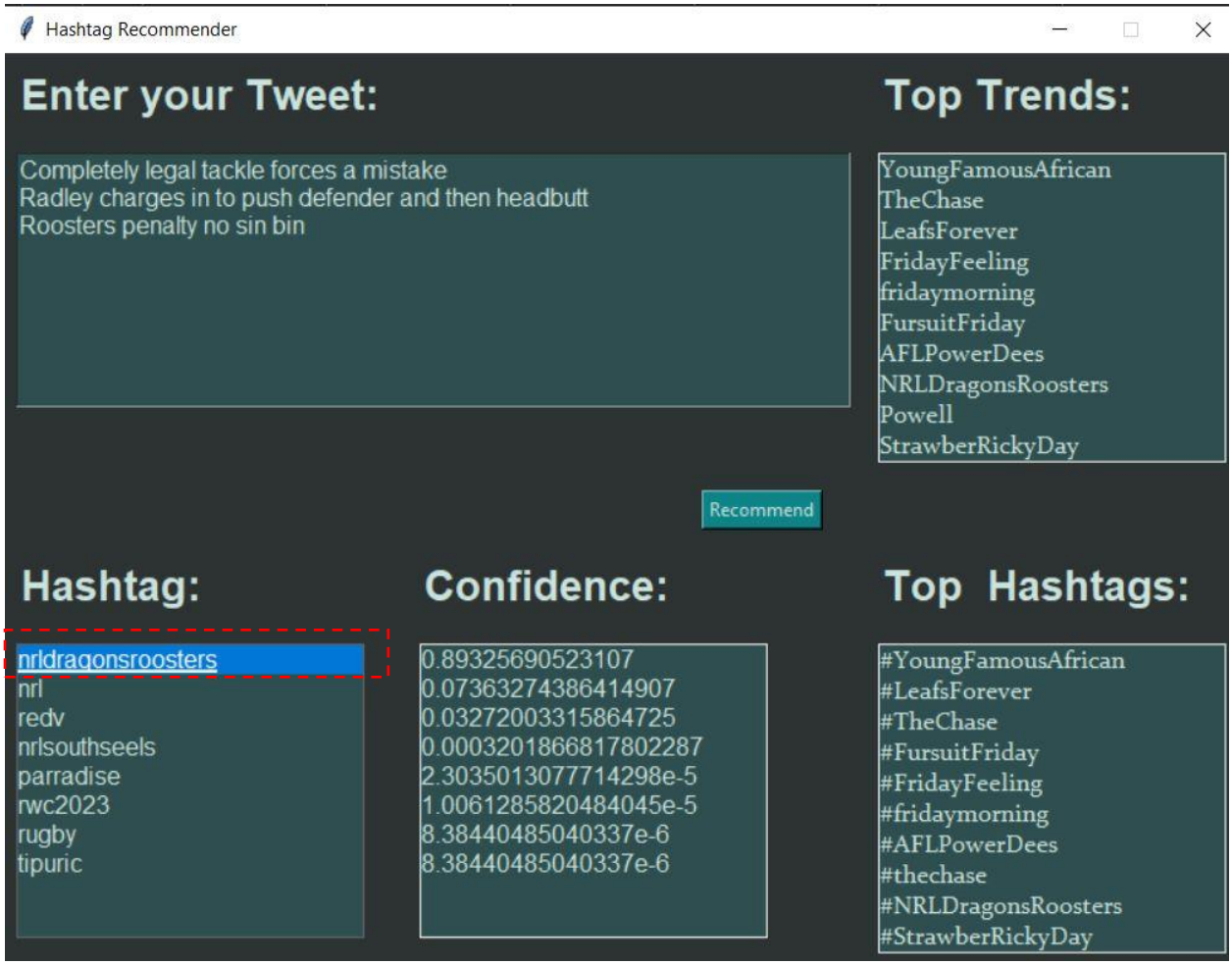
4. ***Hashtag Recommendation Engine****: The success of this phase can be evaluated based on the relevance and diversity of the hashtag recommendations generated. Indicators could include the number of relevant hashtags recommended per trend, the variety of hashtags recommended, and the ability of the engine to recommend niche or emerging hashtags.*

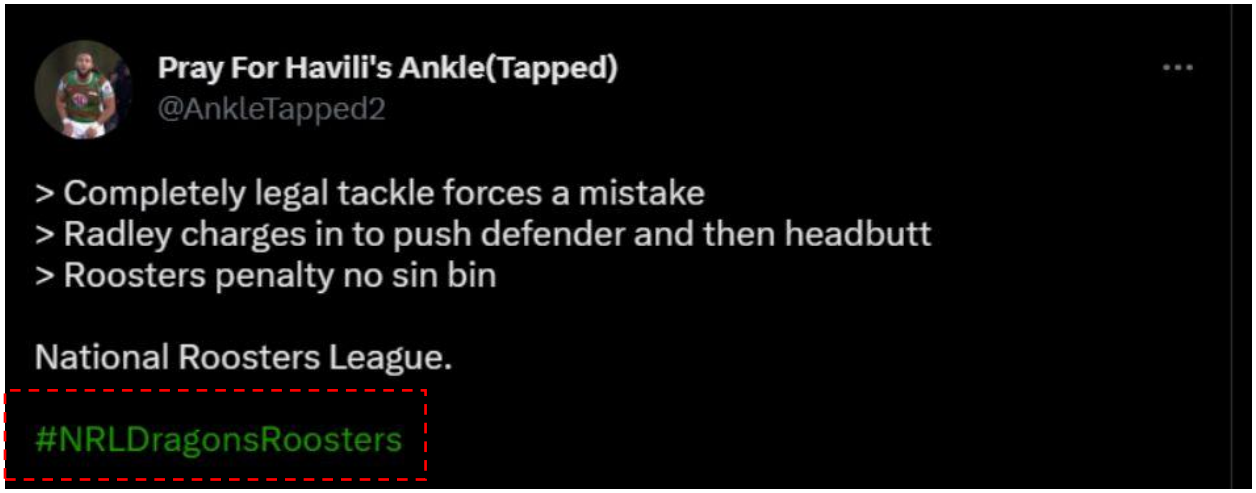
5. ***User Interface Development****: The success of this phase can be evaluated based on the usability and user-friendliness of the developed interface. Indicators could include the ease of use of the interface.*

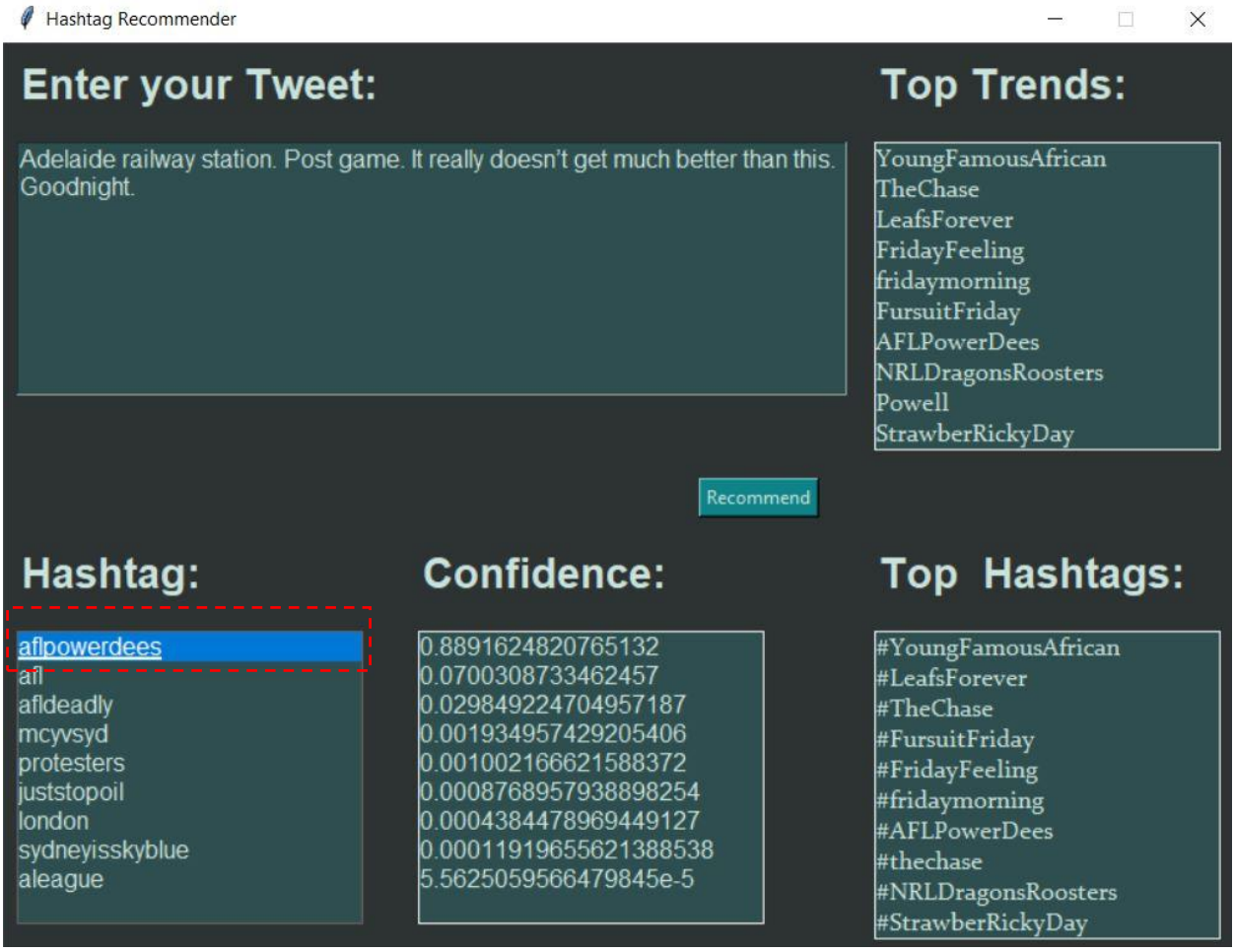
6. ***Integration****: The success of this phase can be evaluated based on the seamless integration of the data and hashtag recommendations into the user interface. Indicators could include the completeness of the data integration, the accuracy of the data displayed, and the speed of the data retrieval.*

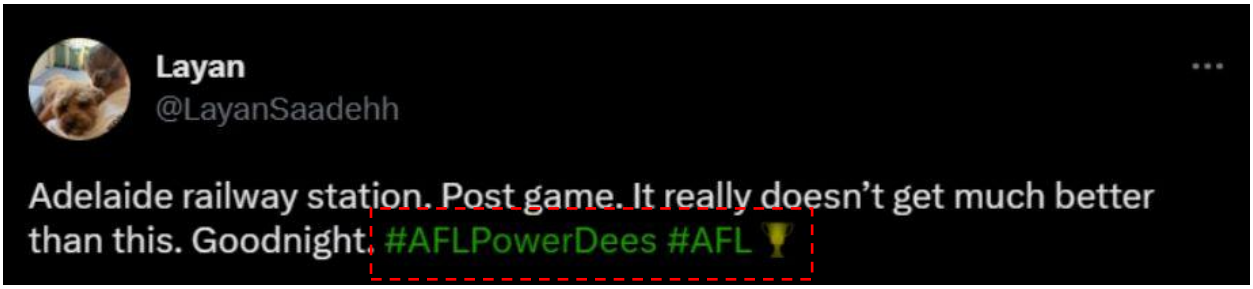
7. ***Evaluation****: The success of this phase can be evaluated based on the accuracy and relevance of the hashtag recommendations generated compared to manually generated ones. Indicators could include the precision and recall of the recommended hashtags, the relevance of the recommended hashtags to the original trend, and the novelty of the recommended hashtags.*

Output Samples

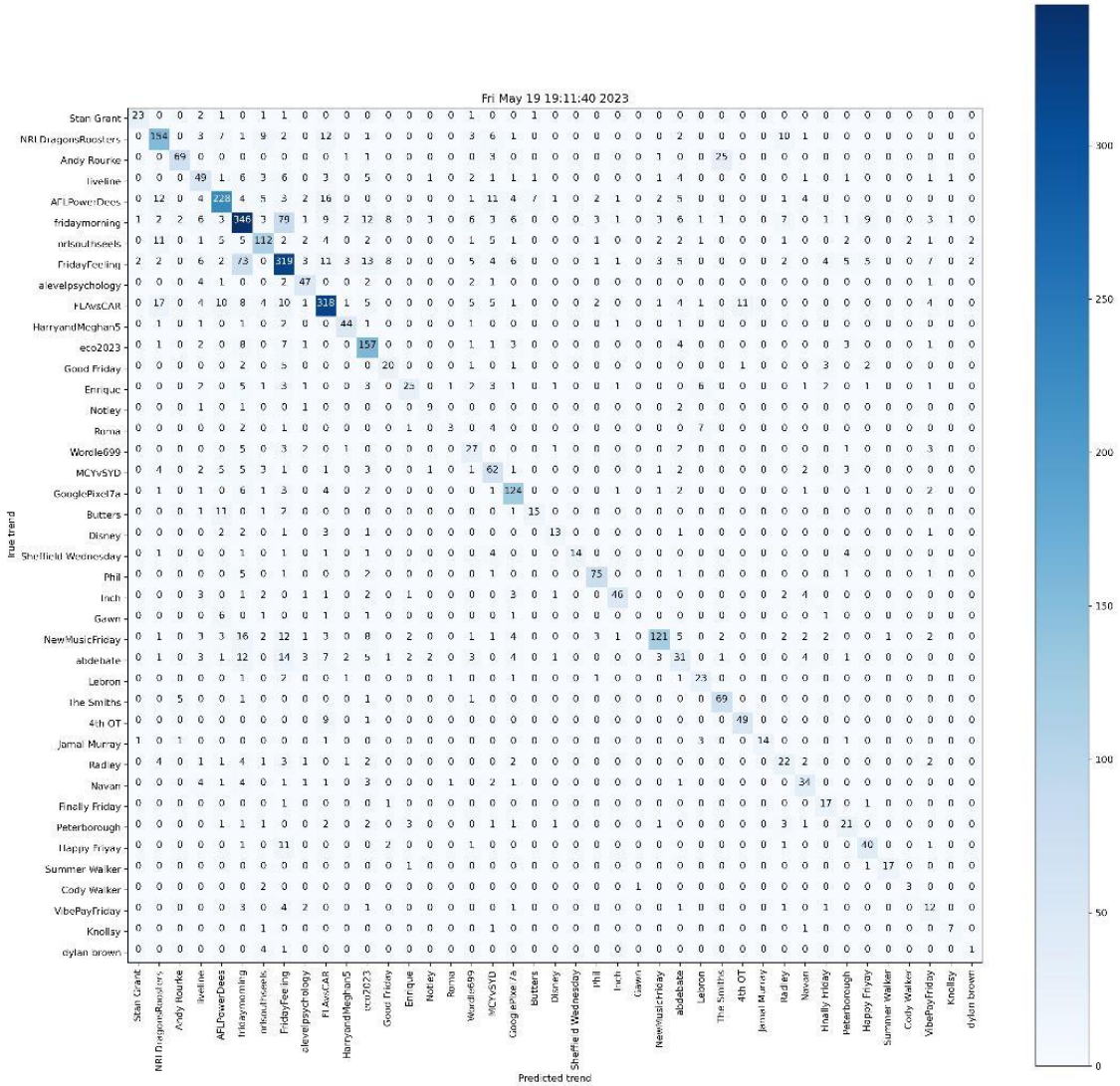


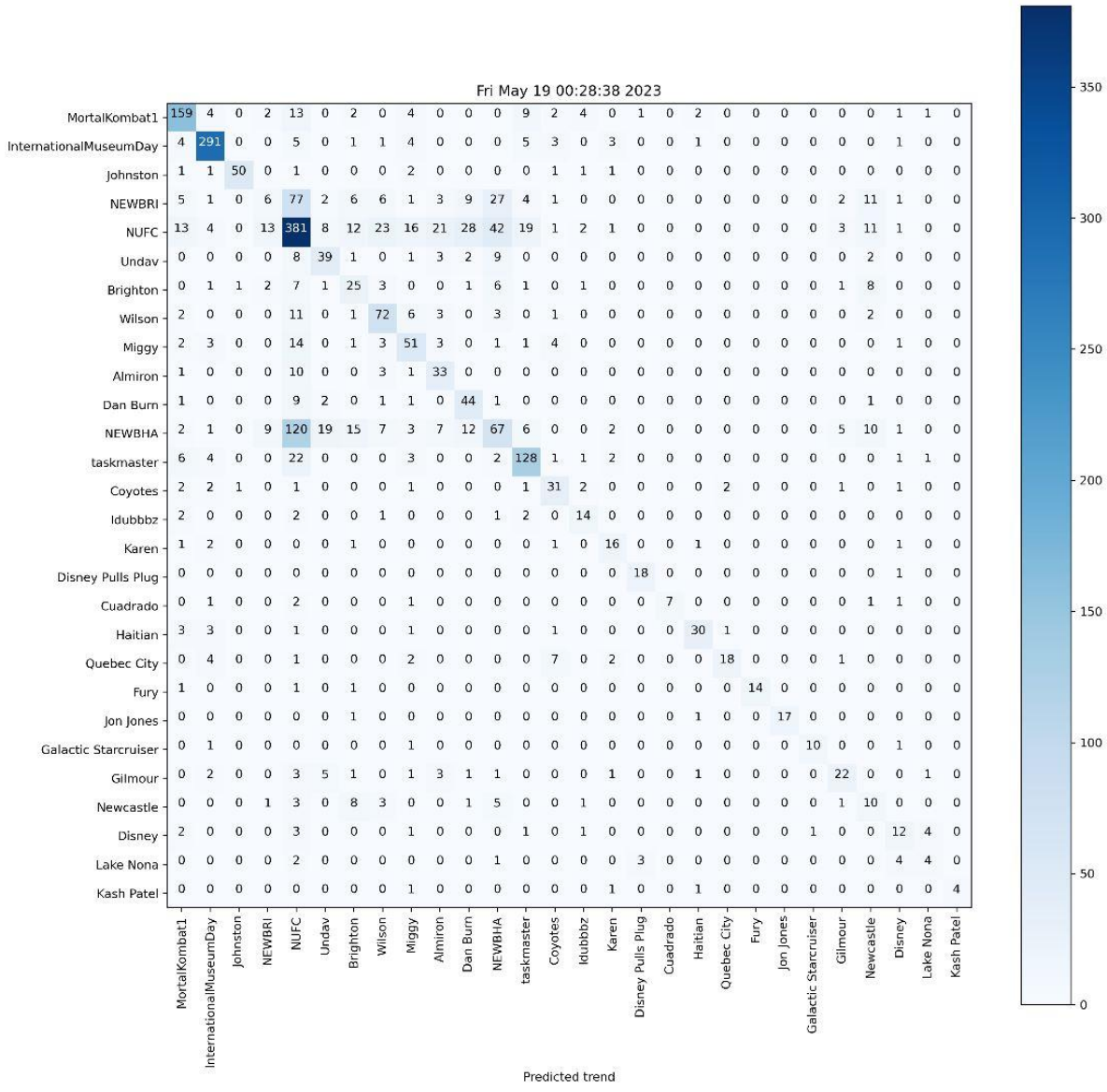


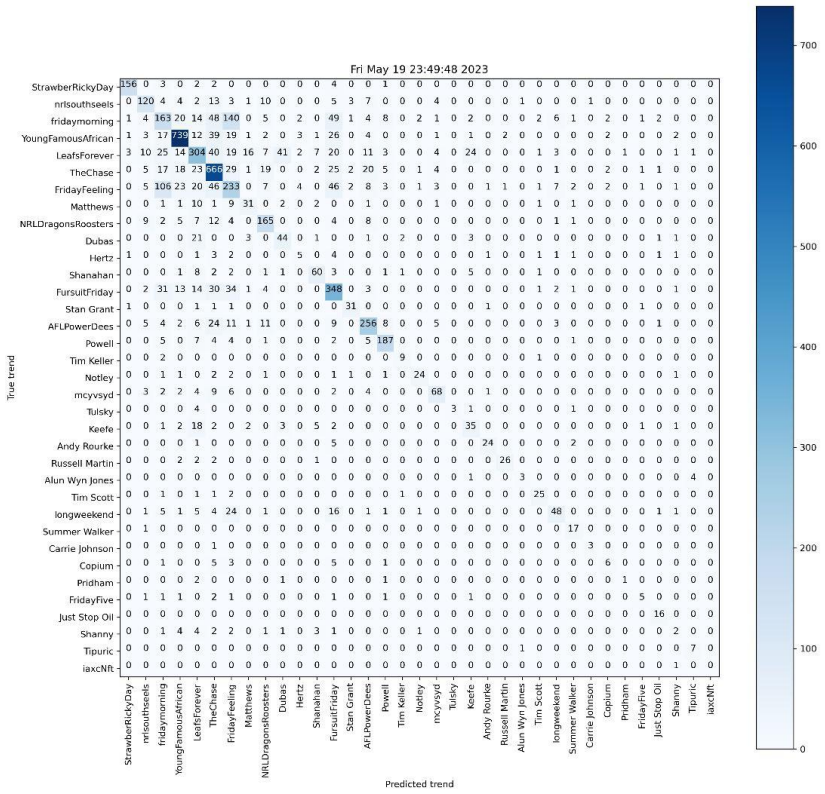

Hashtag Recommender
Enter your Tweet:
Adelaide railway station. Post game. It really doesn't get much better than this. Goodnight.
Top Trends:
YoungFamousAfrican
TheChase
LeafsForever
FridayFeeling
fridaymorning
FursuitFriday
AFLPowerDees
NRLDragonsRoosters
Powell
StrawberRickyDay
Recommend
Hashtag:
aflpowerdees
afl
afldeadly
mcyvsyd
protesters
juststopoil
london
sydneyisskyblue
aleague
Confidence:
0.8891624820765132
0.0700308733462457
0.029849224704957187
0.001934957429205406
0.001002166621588372
0.0008768957938898254
0.0004384478969449127
0.00011919655621388538
5.5625059566479845e-5
Top Hashtags:
#YoungFamousAfrican
#LeafsForever
#TheChase
#FursuitFriday
#FridayFeeling
#fridaymorning
#AFLPowerDees
#thechase
#NRLDragonsRoosters
#StrawberRickyDay


Layan
@LayanSaadehh
Adelaide railway station. Post game. It really doesn't get much better than this. Goodnight. #AFLPowerDees #AFL

Confusion Matrices




Fri May 19 00:28:38 2023
MortalKombat1 159 4 0 2 13 0 2 0 4 0 0 0 9 2 4 0 1 0 2 0 0 0 0 0 0 1 1 0
InternationalMuseumDay 4 291 0 0 5 0 1 1 4 0 0 0 5 3 0 3 0 0 1 0 0 0 0 0 0 1 0 0
Johnston 1 1 50 0 1 0 0 0 2 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
NEWBRI 5 1 0 6 77 2 6 6 1 3 9 27 4 1 0 0 0 0 0 0 0 0 0 2 11 1 0 0
NUFC 13 4 0 13 381 8 12 23 16 21 28 42 19 1 2 1 0 0 0 0 0 0 0 3 11 1 0 0
Undav 0 0 0 0 8 39 1 0 1 3 2 9 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0
Brighton 0 1 1 2 7 1 25 3 0 0 1 6 1 0 1 0 0 0 0 0 0 0 0 1 8 0 0 0
Wilson 2 0 0 0 11 0 1 72 6 3 0 3 0 1 0 0 0 0 0 0 0 0 0 0 2 0 0 0
Miggy 2 3 0 0 14 0 1 3 51 3 0 1 1 4 0 0 0 0 0 0 0 0 0 0 0 1 0 0
Almiron 1 0 0 0 10 0 0 3 1 33 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Dan Burn 1 0 0 0 9 2 0 1 1 0 44 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
NEWBHA 2 1 0 9 120 19 15 7 3 7 12 67 6 0 0 2 0 0 0 0 0 0 0 5 10 1 0 0
taskmaster 6 4 0 0 22 0 0 0 3 0 0 2 128 1 1 2 0 0 0 0 0 0 0 0 0 1 1 0
Coyotes 2 2 1 0 1 0 0 0 1 0 0 0 1 31 2 0 0 0 0 2 0 0 0 1 0 1 0 0
Idubbbz 2 0 0 0 2 0 0 1 0 0 0 1 2 0 14 0 0 0 0 0 0 0 0 0 0 0 0 0
Karen 1 2 0 0 0 0 1 0 0 0 0 0 0 1 0 16 0 0 1 0 0 0 0 0 0 1 0 0
Disney Pulls Plug 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 18 0 0 0 0 0 0 0 0 1 0 0
Cuadrado 0 1 0 0 2 0 0 0 1 0 0 0 0 0 0 0 0 7 0 0 0 0 0 0 1 1 0 0
Haitian 3 3 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 30 1 0 0 0 0 0 0 0 0
Quebec City 0 4 0 0 1 0 0 0 2 0 0 0 0 7 0 2 0 0 0 18 0 0 0 1 0 0 0 0
Fury 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 14 0 0 0 0 0 0 0
Jon Jones 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 17 0 0 0 0 0 0
Galactic Starcruiser 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 10 0 0 1 0 0
Gilmour 0 2 0 0 3 5 1 0 1 3 1 1 0 0 0 1 0 0 1 0 0 0 0 22 0 0 1 0
Newcastle 0 0 0 1 3 0 8 3 0 0 1 5 0 0 1 0 0 0 0 0 0 0 0 1 10 0 0 0
Disney 2 0 0 0 3 0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 12 4 0
Lake Nona 0 0 0 0 2 0 0 0 0 0 0 1 0 0 0 0 3 0 0 0 0 0 0 0 0 4 4 0
Kash Patel 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 4
MortalKombat1
InternationalMuseumDay
Johnston
NEWBRI
NUFC
Undav
Brighton
Wilson
Miggy
Almiron
Dan Burn
NEWBHA
taskmaster
Coyotes
Idubbbz
Karen
Disney Pulls Plug
Cuadrado
Haitian
Quebec City
Fury
Jon Jones
Galactic Starcruiser
Gilmour
Newcastle
Disney
Lake Nona
Kash Patel
Predicted trend
350
300
250
200
150
100
50
0



Model Report

time	accuracy	precision	recall	f1_score	loss	number of tweets	number of trends
Thu May 18 23:25:44 2023	0.867088608	0.868346211	0.867088608	0.863828923	0.122293845	1580	8
Thu May 18 23:30:06 2023	0.895068206	0.893390712	0.895068206	0.892983806	0.100014448	4763	11
Fri May 19 00:10:36 2023	0.569709127	0.563525651	0.569709127	0.539528001	0.699936509	9966	20
Fri May 19 00:28:40 2023	0.609217661	0.595856125	0.609217661	0.592426364	0.692135632	12906	28
Fri May 19 00:48:19 2023	0.627269929	0.624353983	0.627269929	0.619414092	0.643591404	16245	31
Fri May 19 02:21:39 2023	0.658035034	0.65615496	0.658035034	0.65226227	0.545067847	19691	46
Fri May 19 11:32:09 2023	0.593176972	0.628680082	0.593176972	0.600162009	0.602556109	11723	21
Fri May 19 12:10:18 2023	0.78998779	0.807311039	0.78998779	0.794234629	0.291736871	4092	5
Fri May 19 13:56:10 2023	0.532599119	0.548110862	0.532599119	0.527000219	0.683966041	11349	19
Fri May 19 18:11:13 2023	0.685294118	0.690761937	0.685294118	0.681409823	0.434730262	20398	41
Fri May 19 19:01:37 2023	0.685539216	0.694756475	0.685539216	0.683798367	0.43078357	20398	41
Fri May 19 19:11:43 2023	0.681372549	0.691024591	0.681372549	0.680930412	0.433737904	20398	41
Fri May 19 21:11:02 2023	0.708880715	0.706224041	0.708880715	0.706093491	0.312657952	19028	20
Fri May 19 23:57:46 2023	0.669463381	0.667508419	0.669463381	0.665560389	0.343975365	28601	35
Sat May 20 03:10:55 2023	0.952586207	0.950297381	0.952586207	0.951069635	0.036683943	16237	8
Sat May 20 06:14:12 2023	0.784614751	0.784689756	0.784614751	0.782010322	0.351340085	121477	34
Sat May 20 09:15:47 2023	0.800033184	0.802769792	0.800033184	0.798943877	0.344557375	150671	46