ADVANCED CLASSIFICATION PREDICT

TEAM AE6:

Khuliso Muleka, Sbusiso Phakathi, Shanice Pillay and Seromo Podile

AGENDA

- Introduction
- Data Cleaning and Preprocessing
- Exploratory Data Analysis
- Model Building
- Hyperparameter tuning
- Conclusion

Introduction

Problem statement :

Build a machine learning classification model to predict whether people believe in climate change or not.

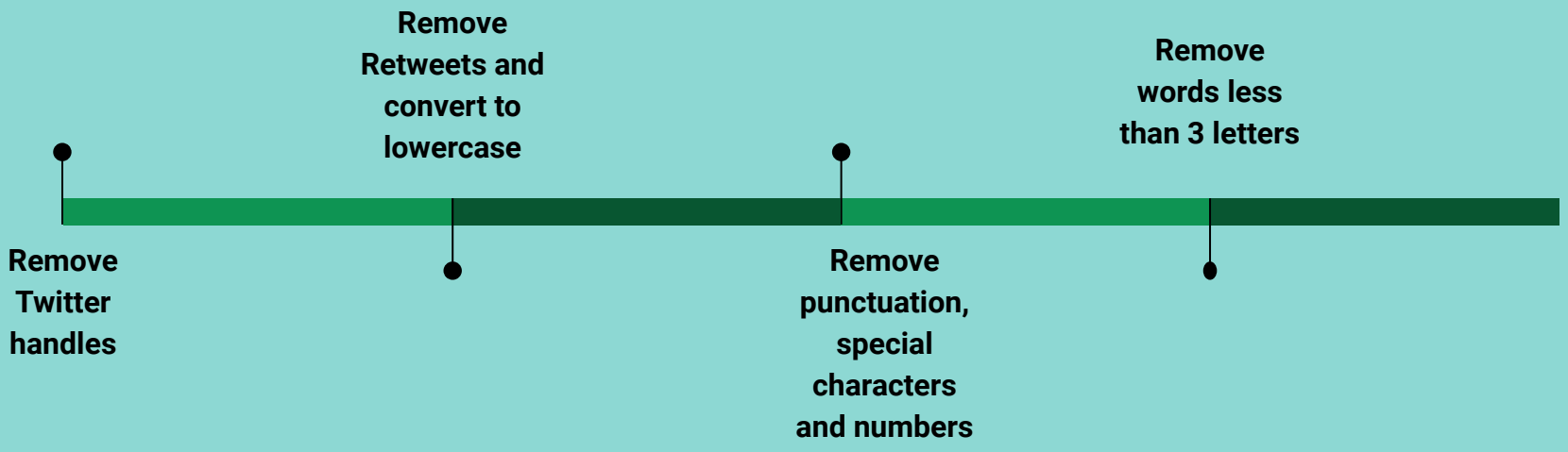
Variables :

- **sentiment: Sentiment of tweet**
- **message: Tweet text**
- **tweetid: Twitter unique id**

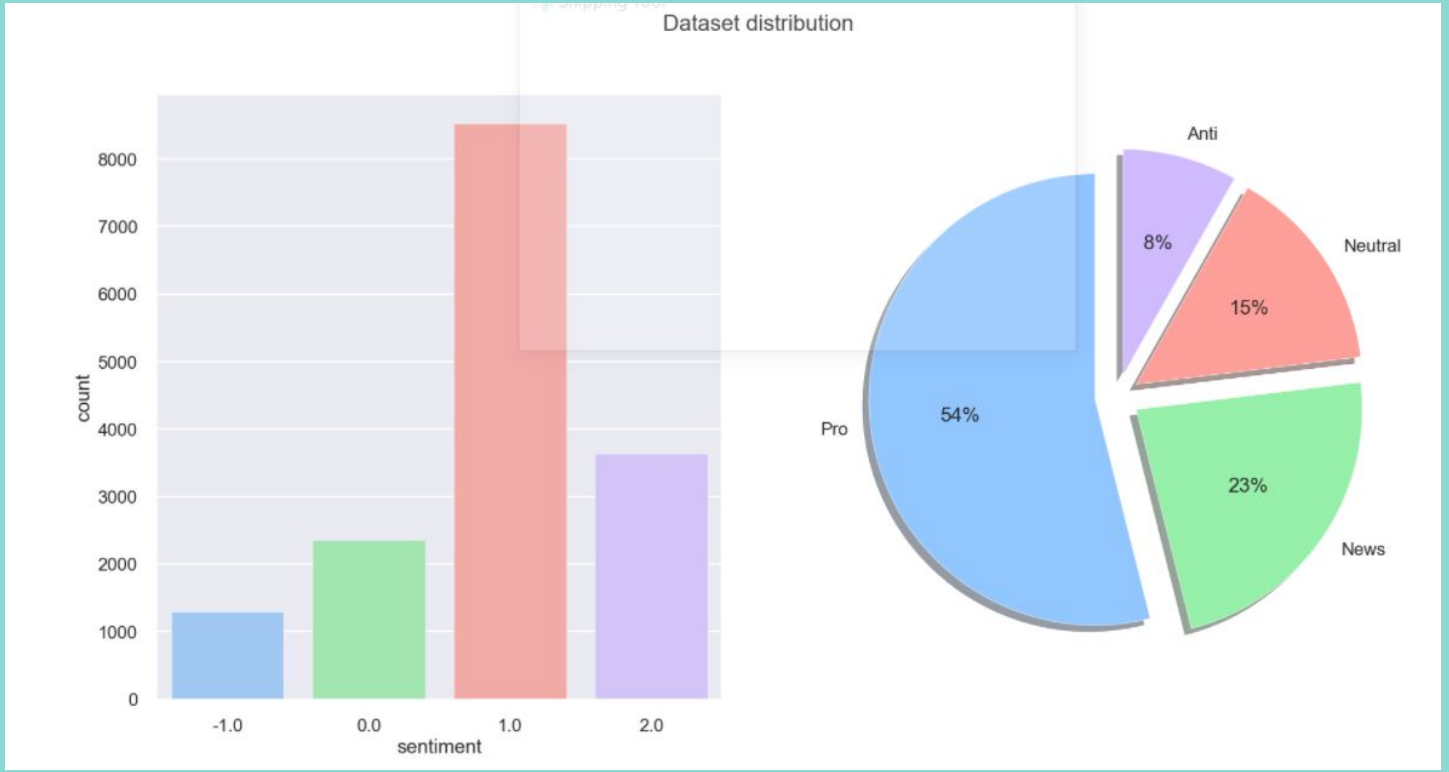
Sentiment:

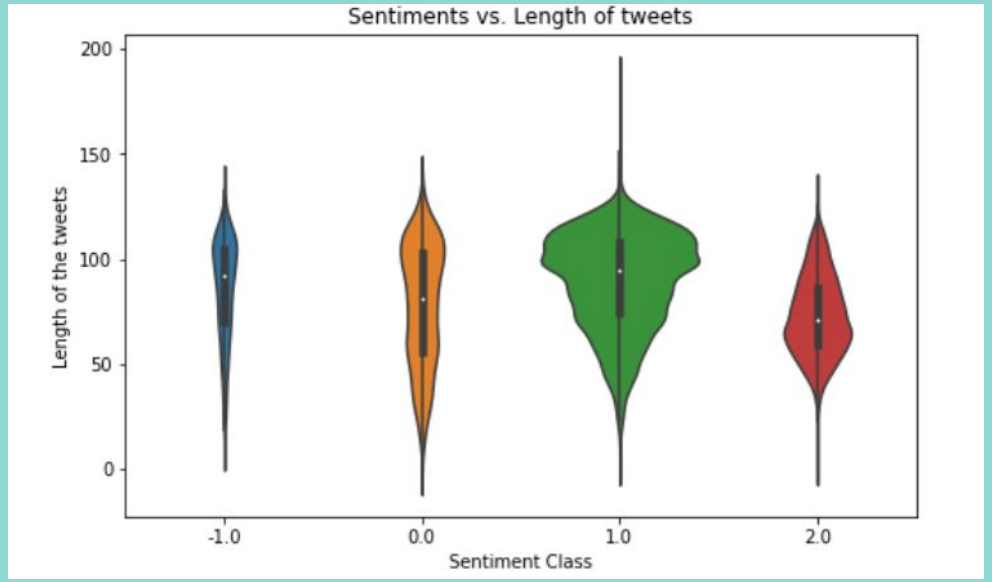
- **2 - News**
- **1 - Pro**
- **0 - Neutral**
- **-1 - Anti**

Data Preprocessing

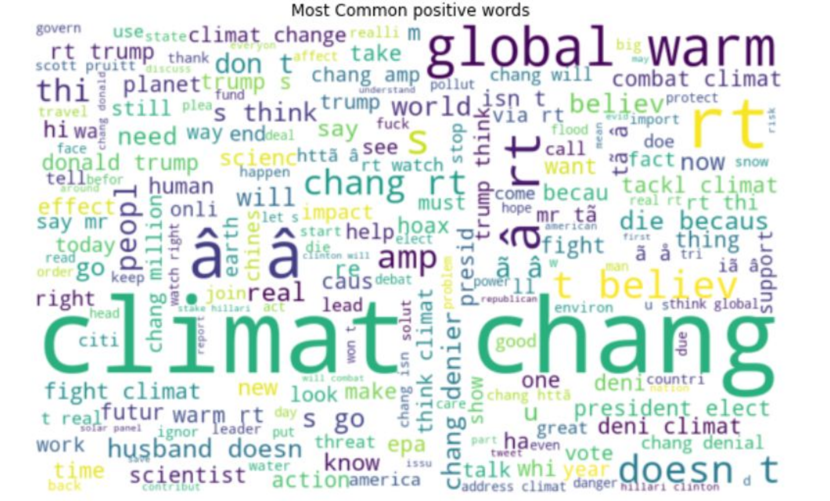


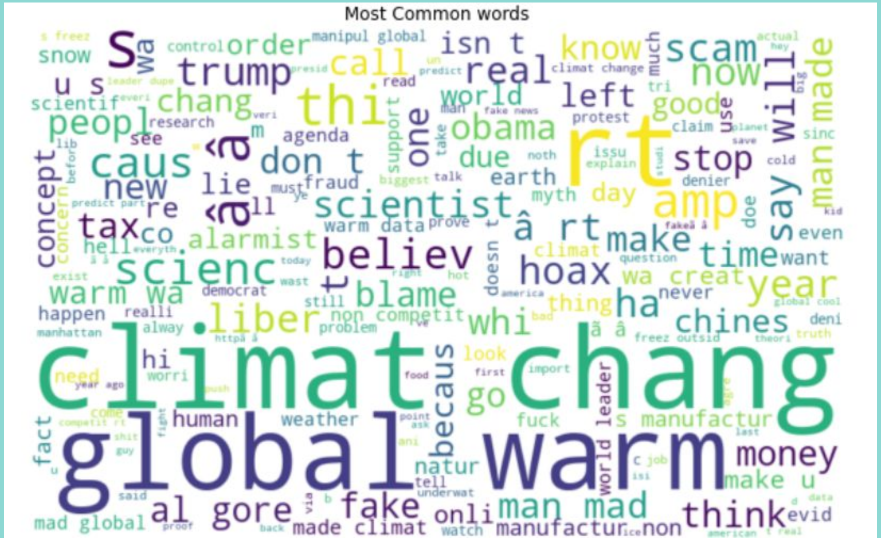
Exploratory Data Analysis



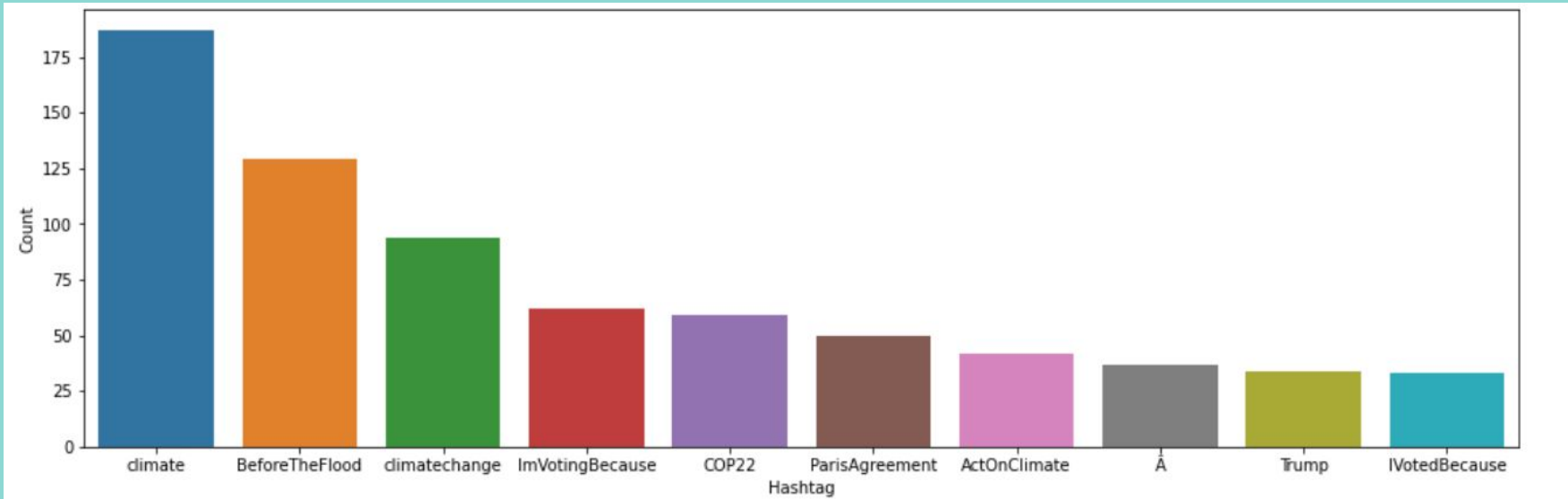


sentiment	count	mean	std	min	25%	50%	75%	max
-1.0	1296.0	86.058642	24.148120	11.0	70.0	92.0	105.0	133.0
0.0	2353.0	78.153421	28.750226	0.0	56.0	81.0	103.0	137.0
1.0	8530.0	89.730832	22.953587	0.0	74.0	95.0	108.0	189.0
2.0	3640.0	72.850549	19.070909	0.0	59.0	71.0	86.0	133.0


Most Common positive words
climat chang
global warm
rt
â â
â rt
t believ
doesn t
chang rt
husband doesn
president elect
deni climat
combat climat
tackl climat
die becaus
donald trump
fight climat


Most Common words
climat chang
global warm
rt
thi
s
â â
trump
scienc
believ
al gore
fake
man mad
think
amp
scientist
hoax
scam
obama
liber
blame

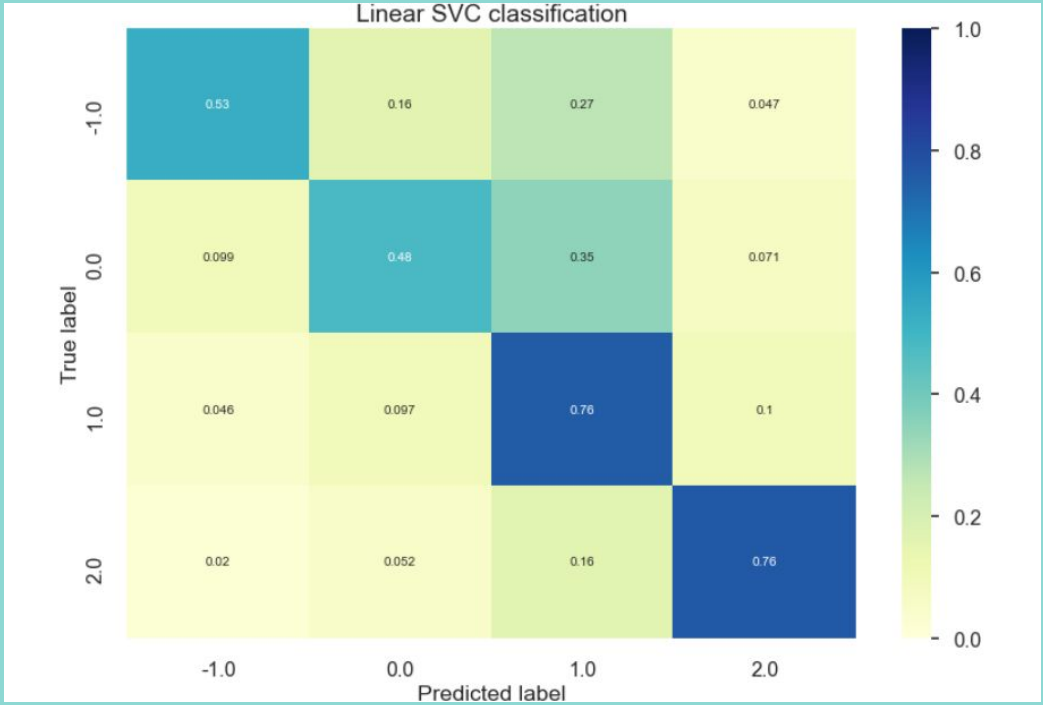
Bar graph showing top 10 hashtags in the Pro Sentiment



Model Selection and Evaluation

1. Random Forest
2. Naive Bayes
3. K Nearest Neighbors
4. Logistic regression
5. Linear Support vector classifier

Best performing model: Linear SVC



	precision	recall	f1-score	support
-1.0	0.52	0.53	0.52	278
0.0	0.45	0.48	0.47	425
1.0	0.80	0.76	0.78	1755
2.0	0.71	0.76	0.74	706
accuracy			0.70	3164
macro avg	0.62	0.63	0.63	3164
weighted avg	0.71	0.70	0.70	3164

accuracy 0.7019595448798989
f1_score 0.7040492608468198

Hyperparameter tuning

Model Selection - Linear SVC model

```
Best cross-validation score: 0.74
Best parameters:  {'clf__C': 1, 'tfidf__ngram_range': (1, 2), 'tfidf__use_idf': True}
F1 score improved by 7.0 %
Old f1_score 0.7040492608468198
New F1 score 0.7559060165399226
```

Conclusion

- Best performing model: Linear SVC Model
- Unbalanced data - changed results slightly.
- Sentiment analysis - Pro Tweets are longer in length, Anti sentiment is fairly centered around politics, mostly mentioning the Trump both neutral and news sentiment have high interaction.
- Improvements - implement deep learning neural network models in the future.

Thank you!

Let us know if you have questions or clarifications.