Team Strategy :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| DASHANG | NEVINA | KARAN | SIMRAN | VISHAKHA |
| Removal special character or stop words | Remove RT | Functions for creating another column for hashtag | Swear words column (low priority) | Stemming |
| Standardization and gramatization for words | Remove @mention | Text cleaning regex - "@\S+|https?:\S+|http?:\S|[^A-Za-z0-9]+" | Undersampling/Oversampling check for train | hashtag (based on frequency) |
| Treating emoticons | Seeing items bought or sold (Low priority) | Try Tensorflow | Changing all the tweets into lowercase | word freq based word cloud (normal) |
| can add subjectivity : 0 for fact; opinion has high value | Changing all the tweets into lowercase | Swear words column (low priority) | hashtag (based on frequency) |  |
| **Models:** | **Models:** | **Models:** | **Models:** | **Models:** |

Positive/Negative wordcloud

Things to do:

Remove sxsw hashtag bar chart

Make funnel chart

Compound vader sentiment freq distribution

How to represent tweet data

**Steps :**

Functions for creating another column for hashtag - KARAN

Text cleaning regex - "@\S+|https?:\S+|http?:\S|[^A-Za-z0-9]+" - karan

Try Tensorflow - KARAN

Remove @mention - NEVINA

Remove RT - NEVINA

Removal special character or stop words - DASHANG

Swear words column (low priority) - KARAN/ SIMRAN

Undersampling/Oversampling check for train? - SIMRAN

Look for missing values. -ALL (binary column for RT or No)

Seeing items bought or sold (Low priority) - NEVINA - POS tagging

Stemming - VISHAKHA

Changing all the tweets into lowercase - SIMRAN/NEVINA

Standardization and gramatization for words - ???(use library)

Treating emoticons - DASHANG

---(optional) can add subjectivity : 0 for fact; opinion has high value

EDA :

**Wordcloud**

1) hashtag (based on frequency) - SIMRAN/VISHAKHA

2) word freq based word cloud (normal) - VISHAKHA

**For presentation :**

1.increase the spectrum of sentiment. Like adding extreme positive or extreme negative using swear words.

2. Identify stakeholders

**Models:**

Naive Bayes

SVM

Random Forest Classifier

Grid search CV/ Randomized search CV

XG Boost

Decision tree classifier

Logistic regression