Sentiment analysis for tweets

—

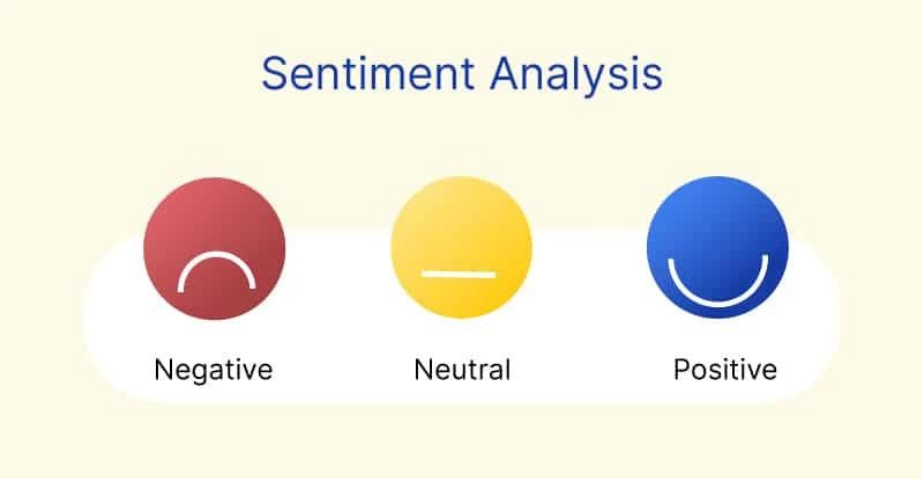
Prepared by: Alaa & Tasneem

Agenda

- Overview
- Challenges
- The dataset
- Data preprocessing
- Data analysis
- System architecture

Overview

- Our project is to create a model for sentiment analysis of tweets that can label tweets to negative, positive or neutral.
- This can be beneficial for many applications:
  - Business can have better management for crises.
  - Business can develop better marketing strategies.
  - Business can have improved customer services.



Challenges

- Finding a dataset.
- Having non alphanumeric characters.
- Selecting the best supervised classification model.
- Choosing the feature extraction model.

The dataset

- Source: Kaggel.
- Link: https://www.kaggle.com/datasets/dunyajasim/twitter-dataset-for-sentiment-analysis.
- Snapshot of data:

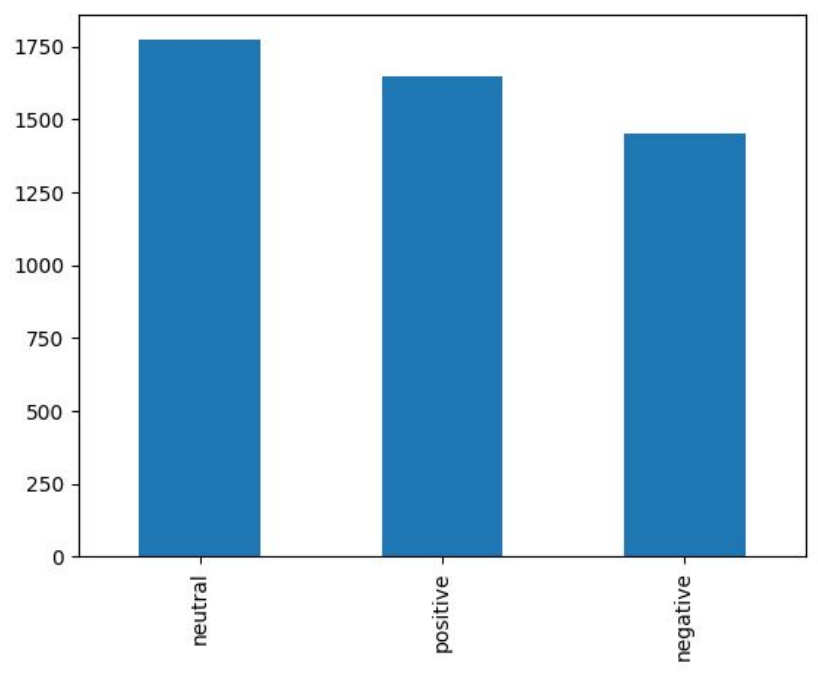
	File Name	Caption	LABEL
0	1.txt	How I feel today #legday #jelly #aching #gym	negative
1	10.txt	@ArrivaTW absolute disgrace two carriages from Bangor half way there standing room only #disgraced	negative
2	100.txt	This is my Valentine's from 1 of my nephews. I am elated; sometimes the little things are the biggest & best things!	positive
3	1000.txt	betterfeelingfilms: RT via Instagram: First day of filming #powerless back in 2011. Can't ¡	neutral
4	1001.txt	Zoe's first love #Rattled @JohnnyHarper15	positive
5	1002.txt	Chaotic Love - giclee print ?65 at #art #love #chaotic #abstract #blue #silver #prints #buy	positive
6	1003.txt	They gna be mad when I reach that goal though. #Rejected the wrong girl ? just getting started & already turn heads.?	negative
7	1004.txt	On day 9.. It's now in my daily routine.. Feeling guuuuurdddd! ? #Aching #PainNoGain #FeelingGood	negative
8	1005.txt	#ANIMALABUSE #TORONTO #PUPPY #TORTURE WE OFFER $1K #REWARD puppy #beaten #bound #burned	neutral
9	1006.txt	Mike will not accept this plastic rose. @wfaamike @wfaachannel8 @wfaagmt #rejected	negative
10	1007.txt	Just ate four cookies. #remorse	negative

Data preprocessing

- Dropping unneeded columns of data.
- Cleanse the data by removing unwanted text (like mentions, links, and numbers).
- Punctuation removal (including filtering non-alphanumeric characters if needed)
- Case folding.
- Lemmatization.
- Tokenization.
- Remove stop words.
- Split hashtags from tweets.
- Coding the data label into numbers (1 for pos, 0 for neutral, -1 for neg)

Data Analysis

Categorizing data by label (Positive, negative and neutral)



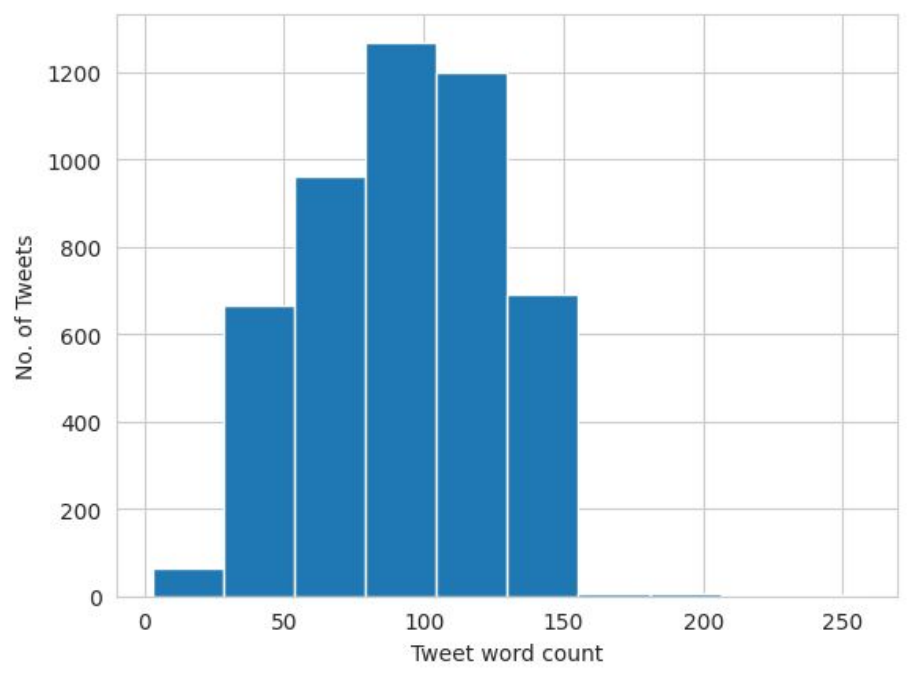

```
[ ]  #Number of Samples in each class

     df['LABEL'].value_counts()

     neutral     1771
     positive    1646
     negative    1452
     Name: LABEL, dtype: int64
```

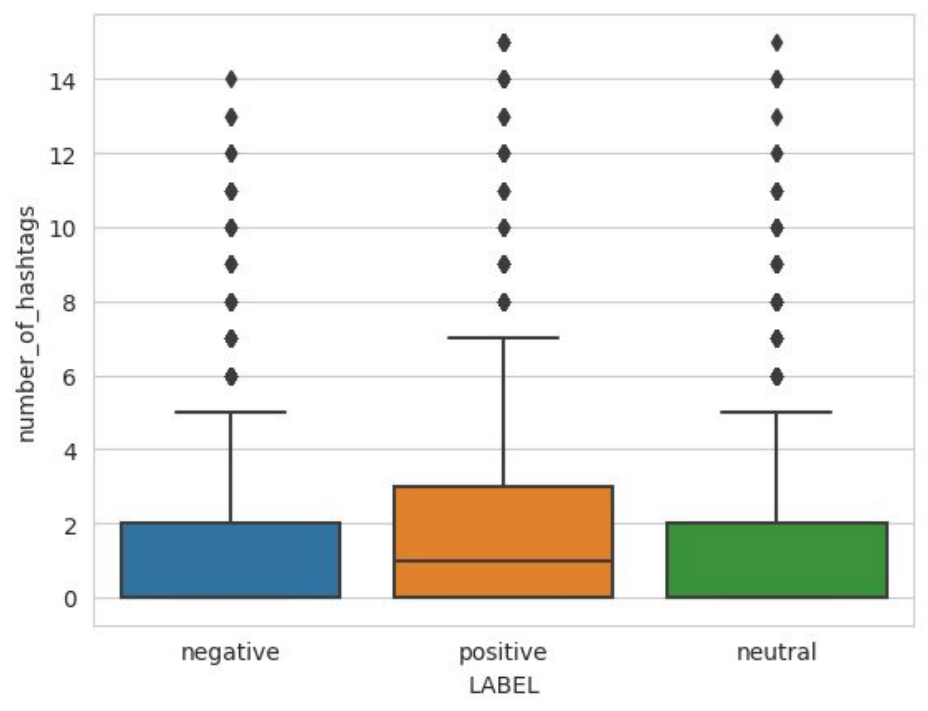
Data Analysis

Visualization for word count in tweets



Data Analysis

Number of hashtags per label (After splitting hashtags from tweets)



System architecture

1. Data Preprocessing (Cleaning data)
2. Split data (75% Training and 25% Testing)
3. Feature selection
4. Model fitting (Supervised classification algorithm)
5. Model performance (testing)
6. Predictions (sentiment classifier)
7. Evaluation

Thank You

Any questions?