Sentiment Analysis

Twitter

Presented by Clivia Kong

BIO

- Experiences & Skills
  - Financial Accountant with multiple commercial industry experiences
  - Financial Data analysis, budgeting and forecasting

- Education
  - Master degree with Accounting and Commerce (MQ)
  - Bachelor degree of Financial Management (NJU)

CONTENT

01 **Business & Data Overview**

02 **Exploratory Data Analysis**

03 **Data Pipeline and Features**

04 **Modelling and Evaluations**

Business Context

Twitter is an open service that’s home to a world of diverse people, perspectives, ideas, and information. - Twitter

Every second, on average, around 6000 tweets are tweeted on Twitter!
Corresponding to over 350,000 tweets per Minute!
500 million tweets per day
200 billion tweets per year

How can we use machine to identify the sentiment of these vast volumes of tweets?

Sentiment Analysis
NLP

Business Context

Sentiment Analysis

Business Value

- ➢ Organize massive amounts of tweets into information in real-time:
  - Be aware of negative reviews about an important product launch before it gets worse
  - Use positive comments to develop new products solve customers’ pain points.
  - Understand the opinions of users about a variety of topics

Data Overview

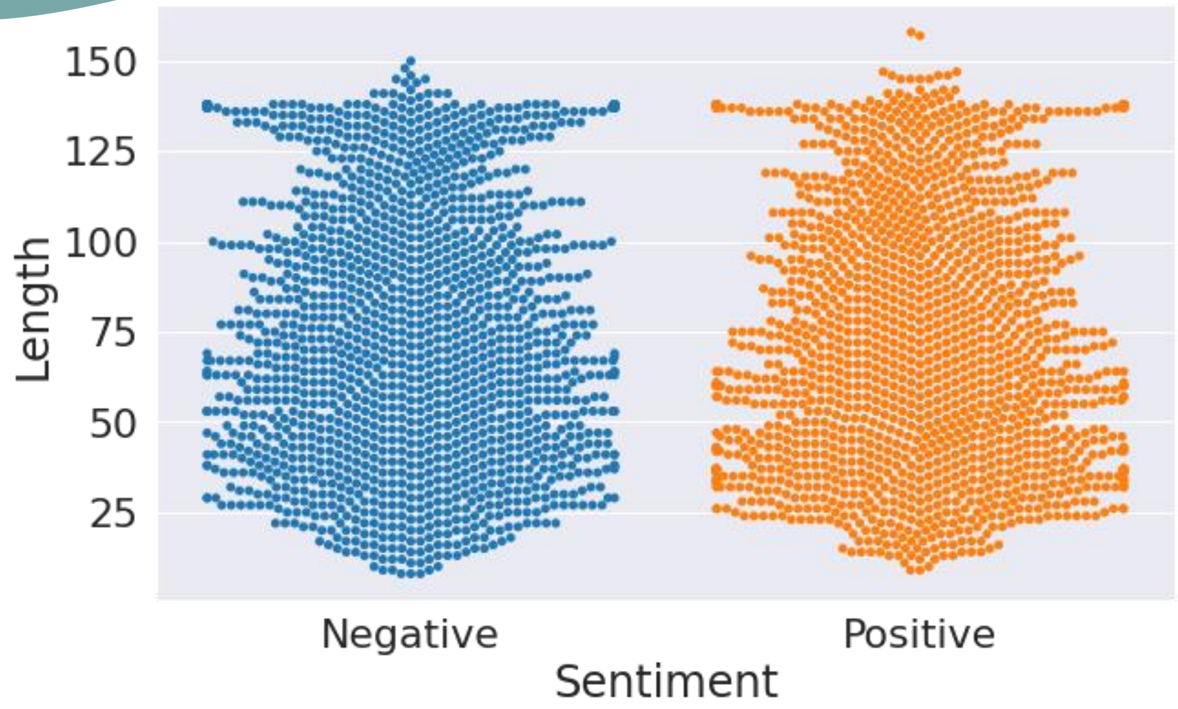
- 3000 tweets sourcing from Kaggle open datasets
- 1500 Negative tweets, 1500 Positive tweets
- Key features: Sentiment label, Tweet Content

Labels

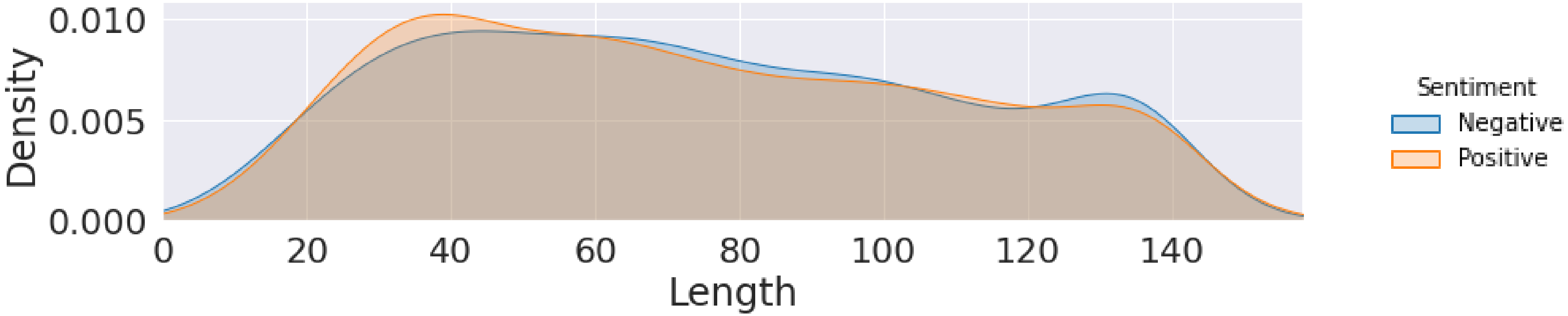
Informal text

Sentiment	Tweet Content
Positive	Nice touch from www.wiggle.co.uk - complimentary bag of Haribo in my delivery of hiking clothes
Negative	Crazy wind today = no birding http://ff.im/1XTTi
Positive	@tracecyrus http://twitpic.com/7horz - This guitar is beautiful
Negative	Need a hug
Positive	and now go shoppppppping ROFL!!!!

EDA

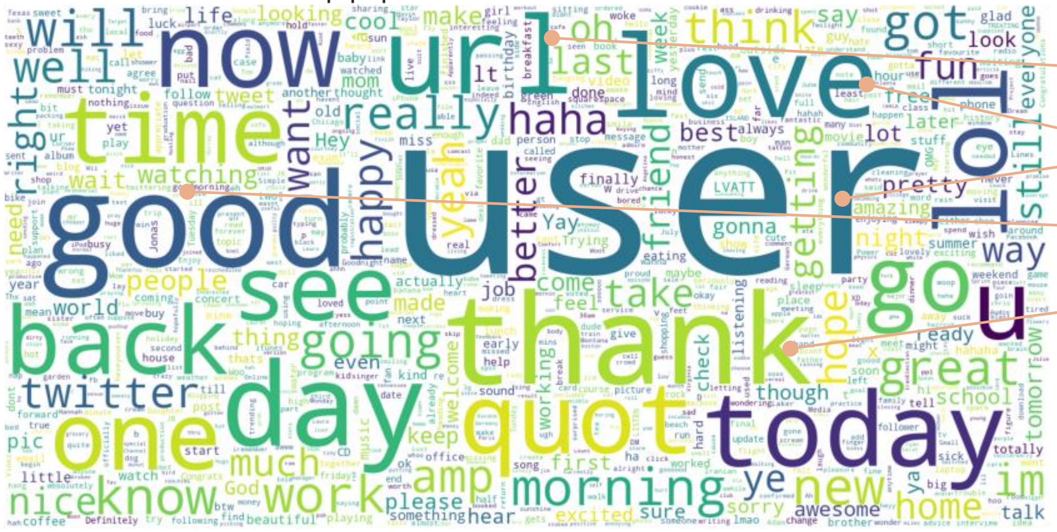


- ◆ Length of positive tweets are similar with negative tweets
- ◆ Positive tweets have more than 150 length text
- ◆ Most of tweets' length are in range from 20 to 140
- ◆ People with positive emotions tends to post around 40 words text
- ◆ People with negative emotions will text more words

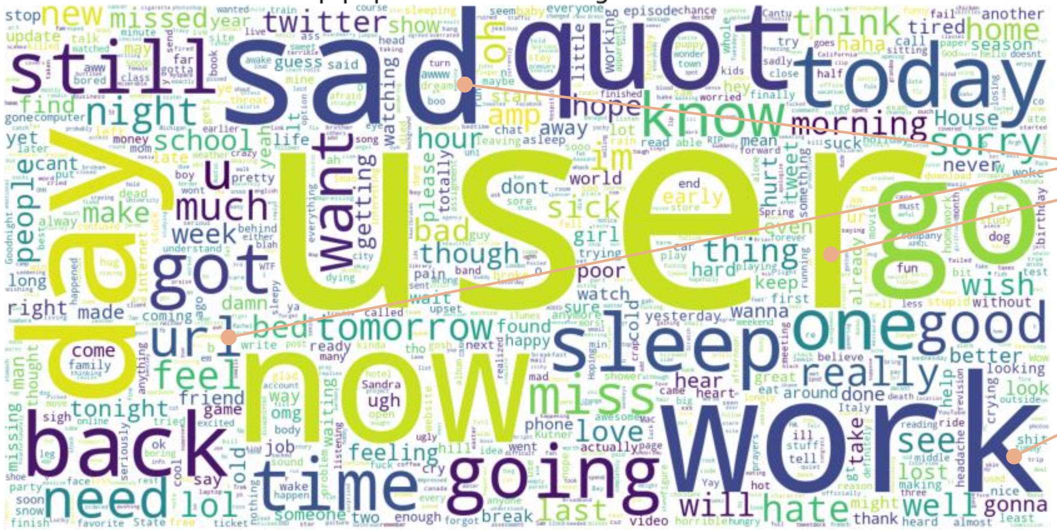


EDA

Top popular words in Positive tweets



Top popular words in Negative tweets



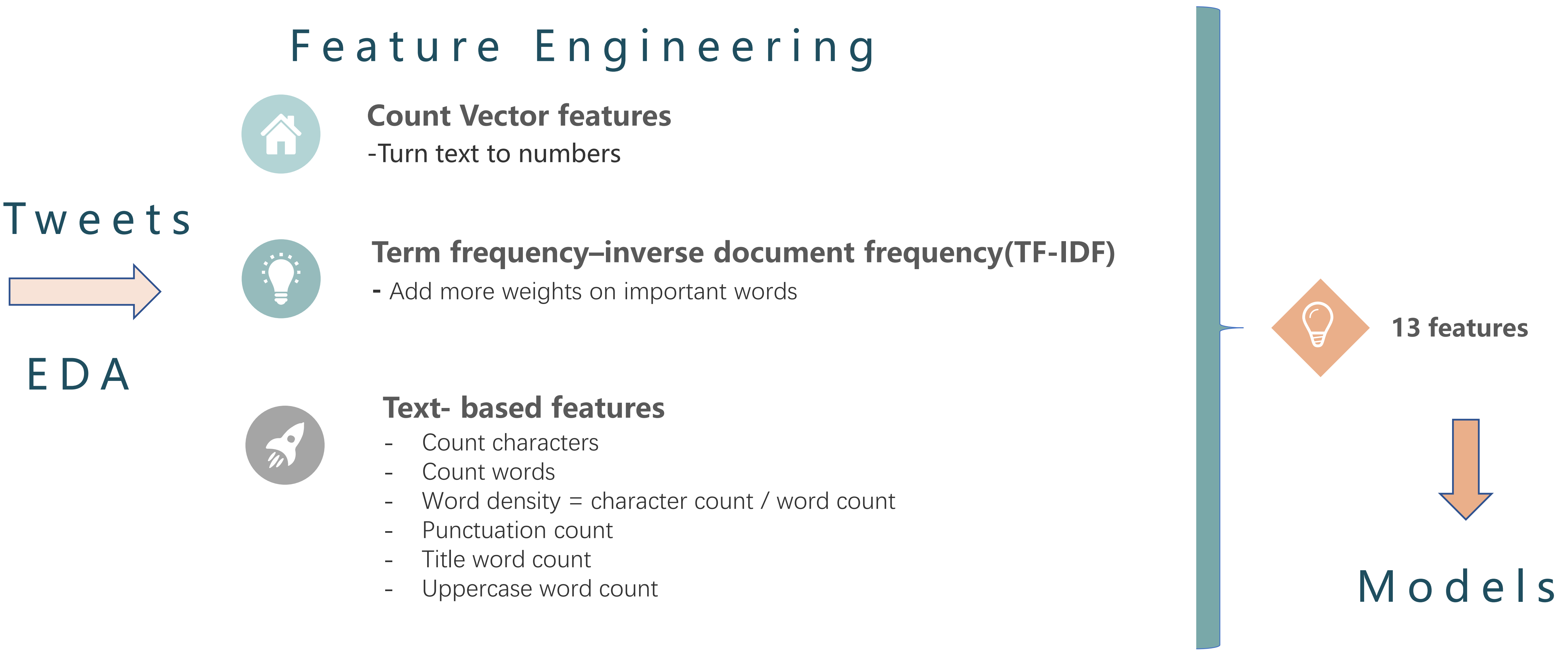
Positive Tweets

- ✓ Like to @ friends or other users or share links to express positive emotions
- ✓ **Positive feelings: Thank, love, good**

Negative Tweets

- □ Like to @ more friends or users or rarely share links to express negative emotions
- □ **Negative feelings: Work, day, sad**

Data Pipeline



Tweets

EDA

Feature Engineering

Count Vector features

-Turn text to numbers

Term frequency–inverse document frequency(TF-IDF)

- Add more weights on important words

Text- based features

- Count characters
- Count words
- Word density = character count / word count
- Punctuation count
- Title word count
- Uppercase word count

13 features

Models

Model Valuation

Naïve Bayes

Classification Models Accuracy	Count Vectors	Word Level TF-IDF	N-Gram TF-IDF	CharLevel TF-IDF	Average Score
Logistic Regression	0.711667	0.700000	0.645000	0.698333	0.685556
Naïve Bayes	0.730000	0.715000	0.658333	0.721667	0.701111
Support Vector Machine	0.690000	0.696667	0.645000	0.683333	0.677222
Random Forest	0.700000	0.726667	0.616667	0.701667	0.677778
Gradient Boosting	0.681667	0.690000	0.608333	0.688333	0.660000

- ✓ Naïve Bayes has highest average score
- ✓ Compare with ‘count vectors’ training set with ‘character level TF-IDF’ training set, I select TF-IDF because TF-IDF gives us a way to associate critical words with a number that represents who relevant this word is in whole sentence

Model Application

Naïve Bayes

➢ I love data science	Positive
➢ This event is not pleasant	Negative
➢ No one likes rainy day	Negative
➢ I'd really truly love going out in this weather!	Positive
➢ Wish you all the best	Positive
➢ May the Luck be with you	Positive

Conclusions and Next Steps

Conclusions

- ✓ Model scores are similar in different models
- ✓ Naïve Bayes have highest score
- ✓ SVM doesn’t perform well

Next Steps

- □ Training more data
- □ More features can be fit into to seek potential improvements
- □ Using emoji to boost sentiment analysis
- □ Applying Bert Model to include connectivity between words

Questions?

Thanks for your watching!

References

- https://primer.ai/business-solutions/what-is-nlp-and-why-do-you-need-it/
- https://www.dsayce.com/social-media/tweets-day/#:~:text=Every%20second%2C%20on%20average%2C%20around%206%2C000%20tweets%20are%20tweeted%20on,200%20billion%20tweets%20per%20year.
- https://blogs.sas.com/content/hiddeninsights/2018/07/16/role-emojis-sentiment-analysis/
- https://www.researchgate.net/profile/Seyed-Ali-Bahrainian/publication/262211692_Sentiment_Analysis_and_Summarization_of_Twitter_Data/links/5f9a2c3f92851c14bcf082e0/Sentiment-Analysis-and-Summarization-of-Twitter-Data.pdf
- https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/

- Data source: https://www.kaggle.com/datasets/kazanova/sentiment140/code?datasetId=2477&searchQuery=begin