

ChatGPT tweets across industries:

Can we predict the industry
that the tweet is associated
with from the sentiments?

Group 18:

Harnehmatt Kaur, Henry Qin, Raon Kim



Our original data

Total observations



500k

Features

Date

id

content

username

like_count

retweet_count

How & Why we got it:

- From kaggle
- Come on, it's about ChatGPT!

Our modified data

Total observations

30k

Features

Date

id

content

username

like_count

retweet_count

Industry

Sentiment

What we did to the data:

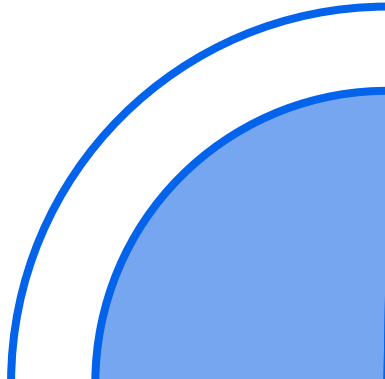
- Added another feature “industries” based on content search
- Filtered the non industry-related tweets
- Added sentiment label to each data using TextBlob



Industries:

- Technology
- Education
- Finance
- Media & Entertainment
- Retail & E-commerce
- Healthcare
- Hospitality & Tourism
- Other

Sentiment score:

- Number between -1 and 1 inclusive.
 - -1 means a negative opinion
 - 1 means a positive opinion
- 

Preprocessing Tweets



Tokenize



Cleaning

Remove URL &
punctuations & stopwords
Convert to lowercase



Lemmatize

WordNetLemmatizer()



**Label with
sentiment scores**

TextBlob library



(This step takes into account the **popularity** of the tweet too, by considering **like and retweet counts**!)

How we added the industry feature

ChatGPT 3.5 ▾



You

Can you give me twenty keywords for each of industries of "Technology", "Entertainment", "Finance", "HealthCare", "Education", "Retail & E-Commerce", "Hospitality & Tourism", "Media & Entertainment"?



ChatGPT

Sure, here are twenty keywords for each industry:

Technology:

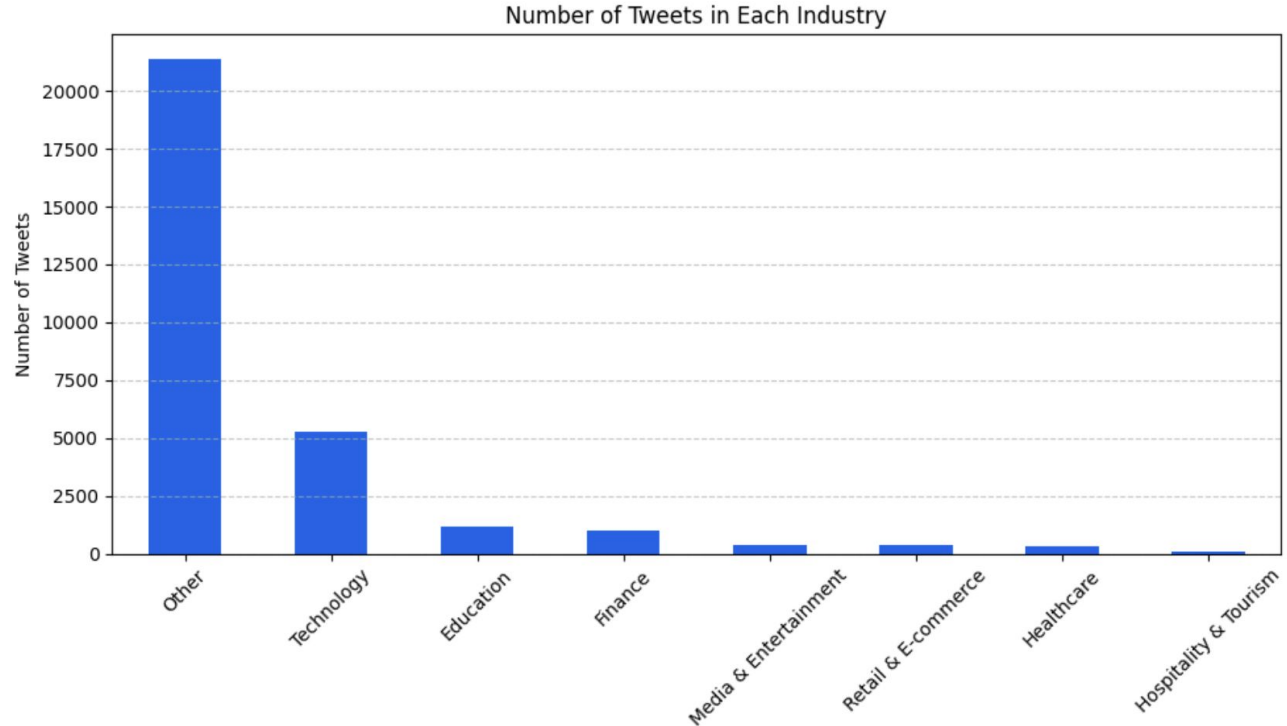
1. Artificial Intelligence
2. Blockchain
3. Cloud Computing
4. Internet of Things (IoT)
5. Cybersecurity
6. Machine Learning
7. Virtual Reality
8. Augmented Reality
9. Big Data
10. Quantum Computing





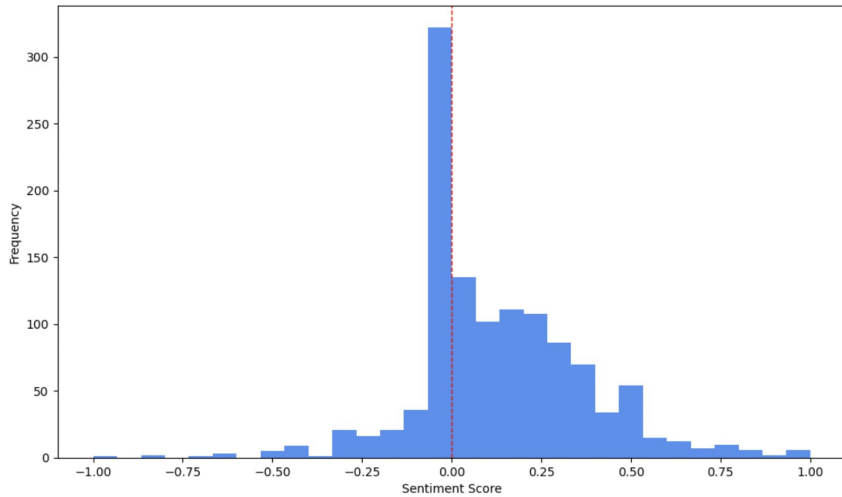
Number of Tweets in Each Industry

Other	21370
Technology	5288
Education	1196
Finance	1013
Media & Entertainment	394
Retail & E-commerce	348
Healthcare	319
Hospitality & Tourism	72

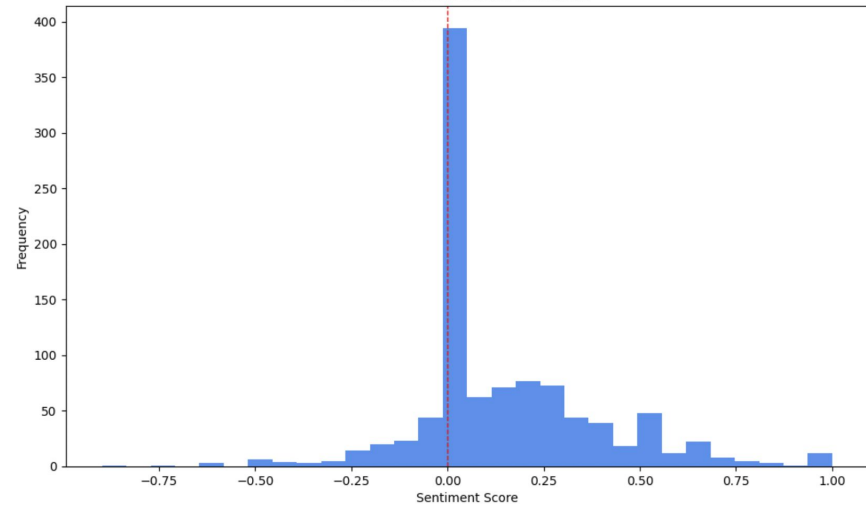


Sentiment Distribution in Industries

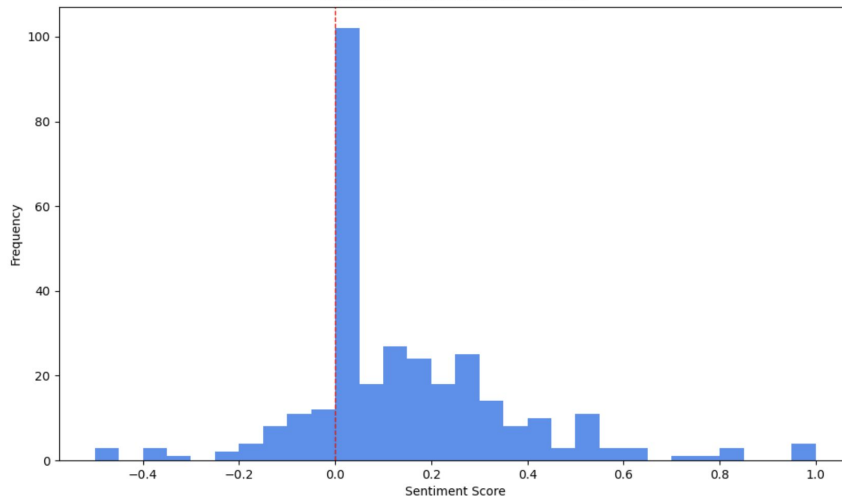
Sentiment Distribution in Education



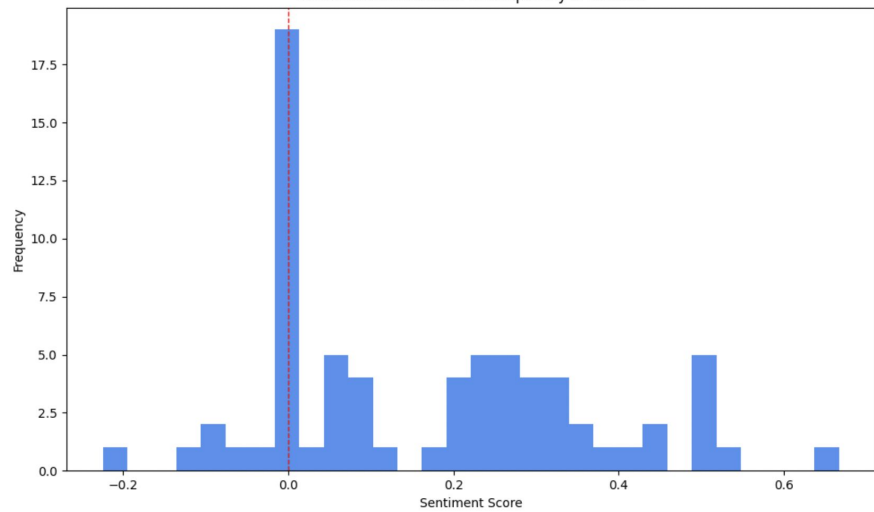
Sentiment Distribution in Finance



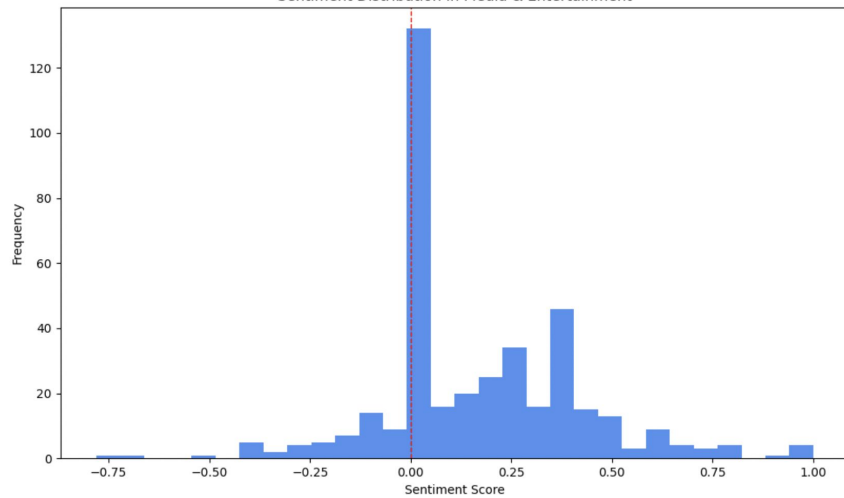
Sentiment Distribution in Healthcare



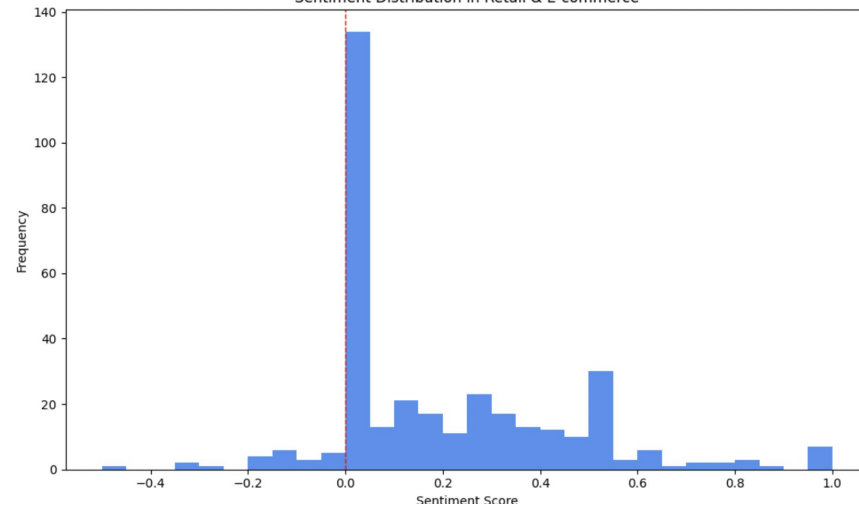
Sentiment Distribution in Hospitality & Tourism



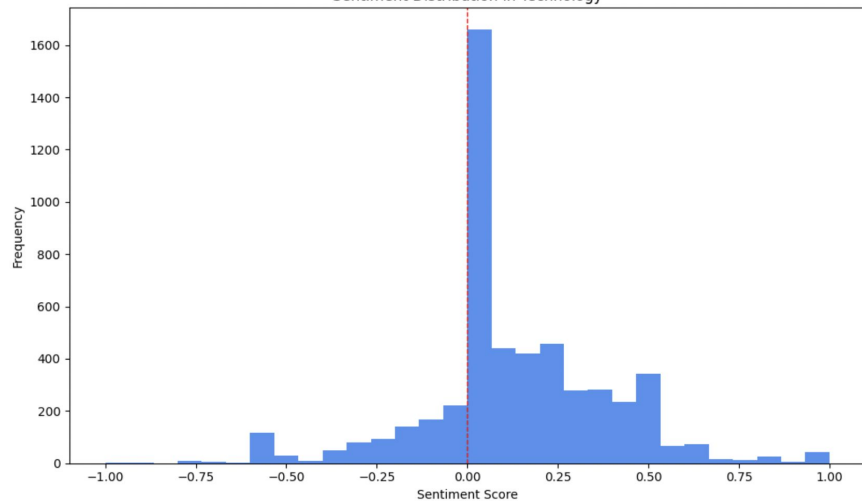
Sentiment Distribution in Media & Entertainment



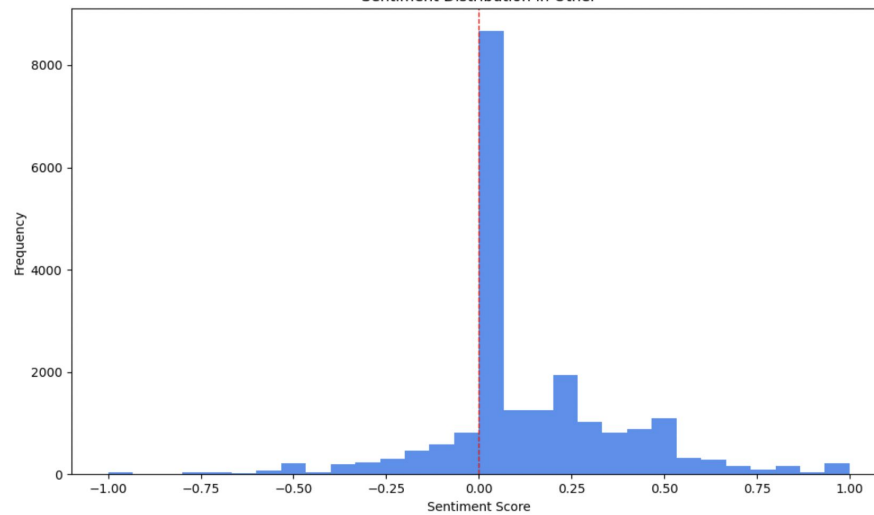
Sentiment Distribution in Retail & E-commerce



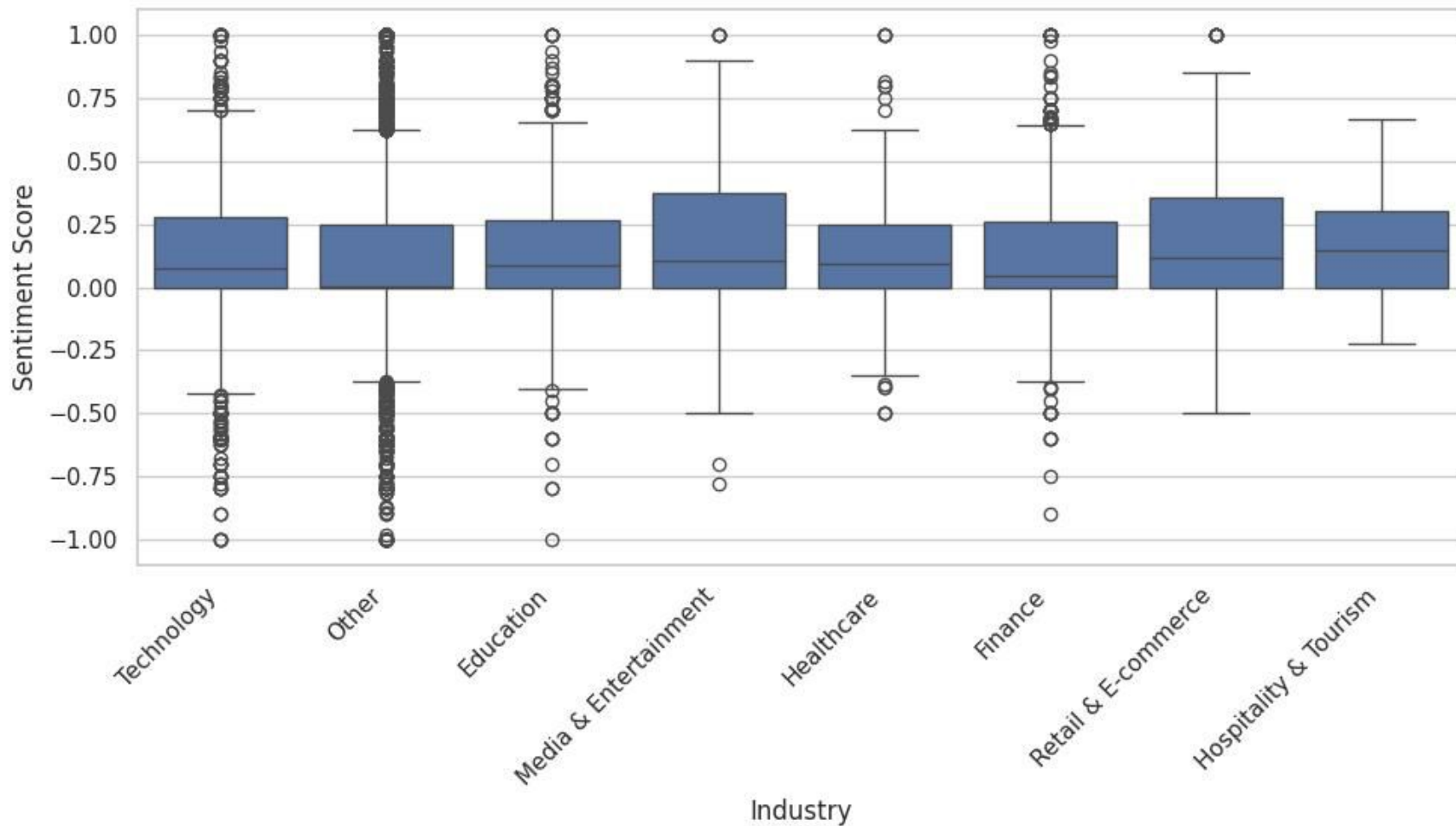
Sentiment Distribution in Technology




Sentiment Distribution in Other



Distribution of Sentiment Scores Across Industries





Machine Learning Model

Our Prediction Model:

Predict Industry using Sentiment

Train & Test Set

- 80/20 Train-Test Split

Naïve-Bayes Model

- Predicts the **industry** (Y - target variable) using **sentiment scores** (X - feature) of tweets
- Calculates the **conditional probability** of each industry category given these sentiment scores using **Bayes' theorem**

Evaluation

Test Accuracy: 0.7103333333333334
Precision: 0.5045734444444445
Recall: 0.7103333333333334
F1-score: 0.5900293639966219

See! We did well!

Tweets	Sentiments Score	Observed	Predicted
This is a metaphor for the limited perception of reality that many people experience and is used in #TheMatrix to add depth and complexity to the film's themes	0.214	Media & Entertainment	Media & Entertainment
Bro used chat gpt for maths 🧠🧠 anyway bro check this out $5 + 1 \times 10 = 15$	0.0	Other	Other
Check out how this new AI quickly answers questions from your PDFs. Perfect for students, researchers, and other curious minds.	0.250	Education	Education
Why Did Elon Musk Sound the Alarm About the Dangers of AI And ChatGPT	-0.100	Technology	Technology

See... We did... well?

Tweets	Sentiments Score	Observed	Predicted
#chatgpt describing cyberpunk in the matrix as a "Dystopian Future: The Matrix takes place in a future where humanity has been enslaved by machines, and the world has become a dark and oppressive place". #bcm325	-0.050	Media & Entertainment	Other
@LeagueOfLegends @riotgames @LoLDev #ChatGPT This can be fun. https://t.co/wLnFzHjJfT	0.3	Media & Entertainment	Other
News: Researchers from @USCMingHsiehEE have developed a new type of #chip that is believed to provide the greatest precision in memory to date. ...	0.412	Technology	Other

Strengths

- Large dataset
 - > Enough tweets to collect tweets with various opinion from various topics
 - > Higher accuracy
- Integrates sentiment analysis, keyword-based industry categorization, and user engagement metric.

Weaknesses & Future Improvements

- We had to evaluate “industry” and “sentiments score” with not a 100% accurate prediction models
 - > Use more accurate model or label them manually
- The sentiments score alone might be too weak of a predictor
 - > Consider using more independent variables
- The dataset is about chatgpt, which have a lot of tweets on technology, compared to other industries



**Thank you
for listening!**



Click “Excellent” for every metric!