

# CSE-713 Group-Presentation

**Title:** An Efficient Machine Learning Approach to Detect Sentiments from Text Data.

Group Members:

- NAZMUL KARIM TANVIR (23166011)
- TOUFIQUL ALAM SAMS (24166027)
- TASNIM FUYARA CHHOAN (23366035)
- SAMIN YASAR (23273007)


# Table of Contents

01		03		05
Introduction, Background		Data Analysis		References
	02		04	
	Methodology		Result Analysis & Conclusion	

## Introduction:

- We'll delve into the utilization of Naive Bayes and modern techniques like Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) for this purpose.


## Motivation:

- Understanding Human Emotions
  - Enhancing NLP systems
  - Improve Decision Making Abilities
  - Mental Health Analysis
- 

## **Research Problem:**

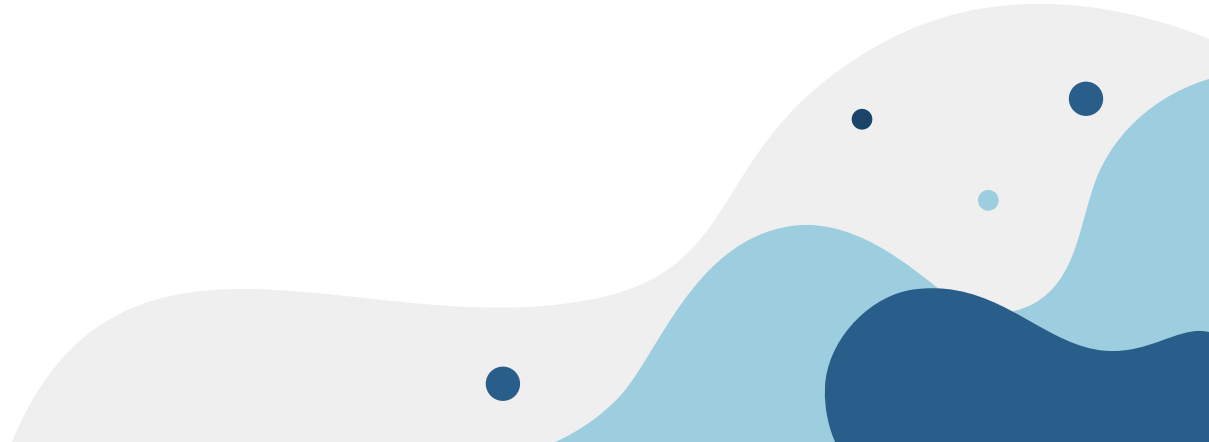
- The study addresses challenges in sentiment analysis, focusing on word score polarity categorization and contextual extraction.
- Contextual understanding in text, especially regarding memes and sarcasm, remains a hurdle.

## **Research Objective:**

- Comprehend sentiment analysis
  - Increase level of mid polarities of the datasets
  - Bias issues
- 

## **Related Work :**

- Notable works include frameworks for news text analysis, sentiment analysis in social media during the Covid-19 pandemic, and models for text summarization employing natural language processing (NLP).



# Sample Reviewed Papers

## Paper 1

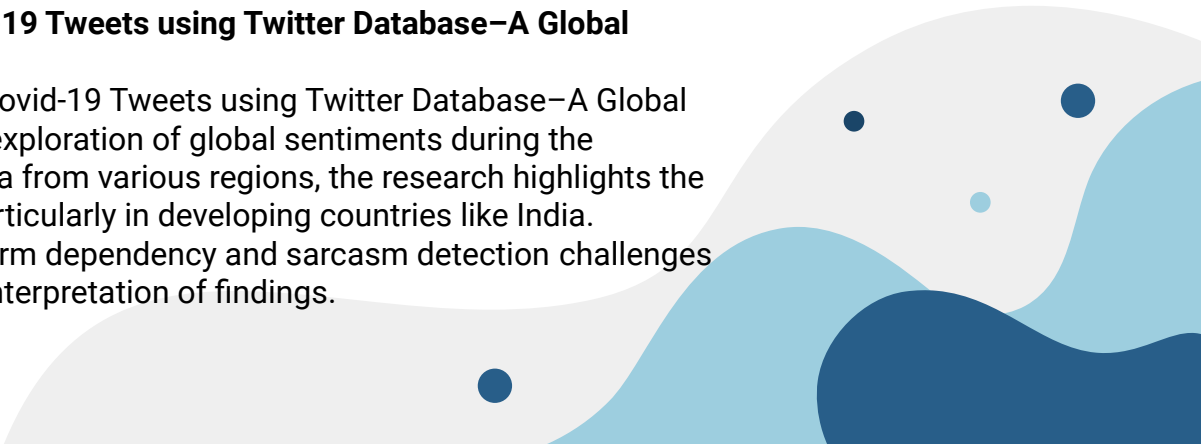
### **Title: News text Analysis using Text Summarization and Sentiment Analysis based on NLP**

The paper introduces a text summarization model using natural language processing (NLP) and sentiment analysis to address information overload. It uses the NLTK library and Python for sentiment analysis, achieving 91.67% accuracy. However, its reliance on a self-generated dataset and specific tools may limit its generalizability and flexibility. Future work aims to enhance the model with advanced tools, NLP techniques, and a recommendation system for broader applicability across different domains and languages.

## Paper 2

### **Title: Sentiment Analysis of Covid-19 Tweets using Twitter Database–A Global Scenario**

The study "Sentiment Analysis of Covid-19 Tweets using Twitter Database–A Global Scenario" offers a comprehensive exploration of global sentiments during the pandemic. By analyzing Twitter data from various regions, the research highlights the prevalence of sadness and fear, particularly in developing countries like India. However, limitations such as platform dependency and sarcasm detection challenges underscore the need for nuanced interpretation of findings.



# Sample Reviewed Papers

## Paper 3

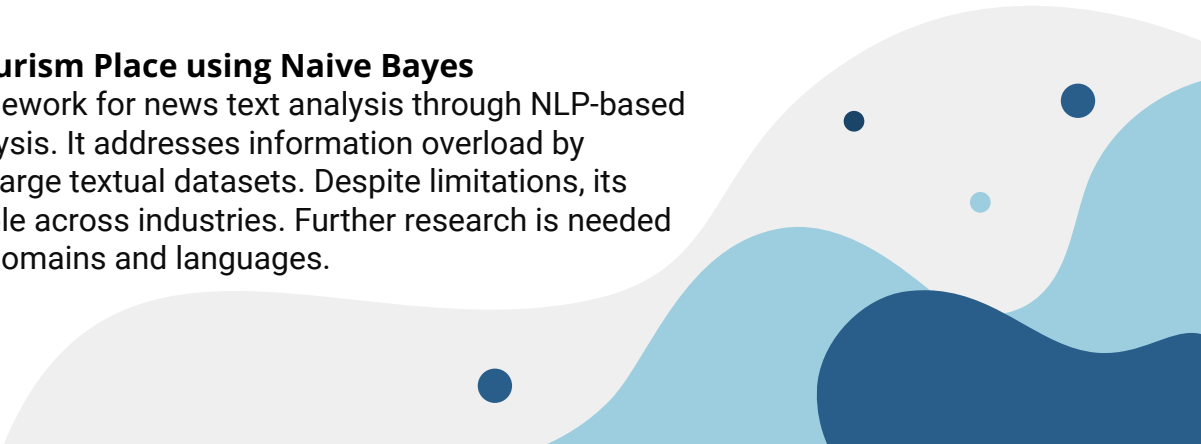
**Title: Sentiment Analysis of Weather-Related Tweets from Cities within Hot Climates**

Through an analysis of weather-related tweets from Phoenix and Singapore, this study identified a pattern of higher pain during temperature rises. Singapore was always negative, although Phoenix's feelings changed with the seasons. These results show how regional opinions are reflected on social media: tweets from Phoenix mimic weather forecasts, and Singaporeans express dissatisfaction. Additionally, the data points to long-term heat repercussions and vulnerability to local climate events.

## Paper 4

**Title: Sentiment Analysis on Tourism Place using Naive Bayes**

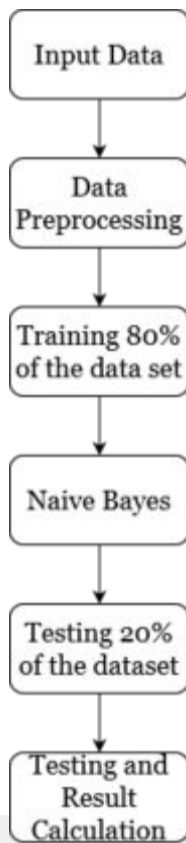
The paper introduces a robust framework for news text analysis through NLP-based summarization and sentiment analysis. It addresses information overload by efficiently extracting insights from large textual datasets. Despite limitations, its accuracy and utility render it valuable across industries. Further research is needed to enhance its capabilities across domains and languages.



- **Workflow and Methodology**

## Naive Bayes Model

- $P(A|B) = \frac{P(B|A)}{P(B)}$
- Multinomial Naive Bayes
- Complement Naive Bayes



### Some key features for NB

- text to numerical conversion
- random state=42
- 5-label to 3-label conversion



- **Workflow and Methodology (cont.)**

## ➤Multinomial Naive Bayes

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w|c) = \frac{\text{count}(w,c) + 1}{\text{count}(c) + |V|}$$

- suitable for classification with discrete features (e.g., word counts )
- represent the frequencies of certain events
- assumes independency of features in a class

## ➤Complement Naive Bayes

$$\text{NB} \rightarrow \operatorname{argmax}_y p(y) \prod p(w_i | y)^{f_i}$$

$$\text{CNB} \rightarrow \operatorname{argmin}_y p(y) \prod \frac{1}{p(w_i | \hat{y})^{f_i}}$$

- addresses the issue of imbalanced class distributions
- gives higher weight to features that are highly predictive for the minority class but less frequent in the majority classes.

# Data analysis

## Dataset - 10,000 Spotify review

Time_submitted	review	sentiment	Total_thumbsup
07-09-22 15:00	Great music service, the audio is high quality and t	5	2
07-09-22 14:21	Please ignore previous negative rating. This app is	5	1
07-09-22 13:27	This pop-up "Get the best Spotify experience on A	4	0
07-09-22 13:26	Really buggy and terrible to use as of recently	1	1
07-09-22 13:20	Dear Spotify why do I get songs that I didn't put o	1	1
07-09-22 13:20	The player controls sometimes disappear for no r	3	7
07-09-22 13:19	I love the selection and the lyrics are provided wit	5	0
07-09-22 13:17	Still extremely slow when changing storage to ext	3	16
07-09-22 13:16	It's a great app and the best mp3 music app I have	5	0
07-09-22 13:11	I'm deleting this app, for the following reasons: Th	1	318
07-09-22 13:11	Love Spotify, and usually this app is the best, but a	2	1
07-09-22 13:10	Can't play Spotify when on WiFi	1	1

# Data analysis

## About the Dataset

- Scraping Spotify reviews on Google Play Store

- This dataset contains reviews of Spotify App from 1/1/2022 - 7/9/2022 collected from Google Play Store

- Total Row: 61594 rows

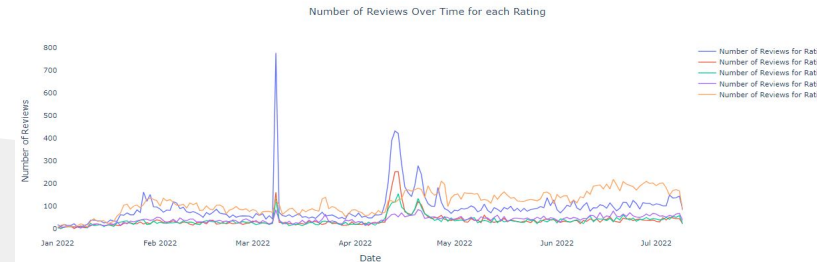
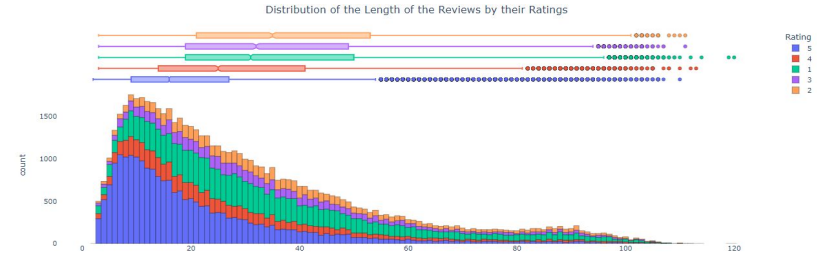
- Total 5 column

Time_submitted	Review	Rating	Total_thumbsup	Reply
2022-07-09 15:00:00	Great music service, the audio is high quality...	5	2	NaN
2022-07-09 14:21:22	Please ignore previous negative rating. This a...	5	1	NaN
2022-07-09 13:27:32	This pop-up "Get the best Spotify experience o...	4	0	NaN
2022-07-09 13:26:45	Really buggy and terrible to use as of recently	1	1	NaN
2022-07-09 13:20:49	Dear Spotify why do I get songs that I didn't ...	1	1	NaN
2022-07-09 13:20:20	The player controls sometimes disappear for no...	3	7	NaN
2022-07-09 13:19:21	I love the selection and the lyrics are provid...	5	0	NaN
2022-07-09 13:17:22	Still extremely slow when changing storage to ...	3	16	NaN
2022-07-09 13:16:49	It's a great app and the best mp3 music app I ...	5	0	NaN
2022-07-09 13:11:32	I'm deleting this app, for the following reaso...	1	318	NaN

# Data analysis

## Exploratory Data Analysis

- Peaks at 1 and 5 points indicate extreme sentiments.
- Higher-rated reviews have shorter text (Figure 1).
- Review peaks on March 8, 2022, and April 11-23, 2022 (Figure 2), suggest increased feedback.
- During peaks (Figure 3), lower ratings dominate, hinting at app issues.
- Subsequently, 5-point reviews increase, while 1-point reviews decline (Figure 4), suggesting satisfaction improvements.
- Deeper analysis needed to enhance app features and user experience, potentially improving satisfaction and retention.



## Data analysis

## Data Pre-processing:

- Lowercased all text.
- Expanded contractions.
- Removed numbers, punctuation, emojis, non-Latin characters, and words less than 2 characters.
- Tokenized text and removed stopwords.
- Lemmatized text.
- Additional Pre-Processing

### Top used 100 Words before Text Cleaning

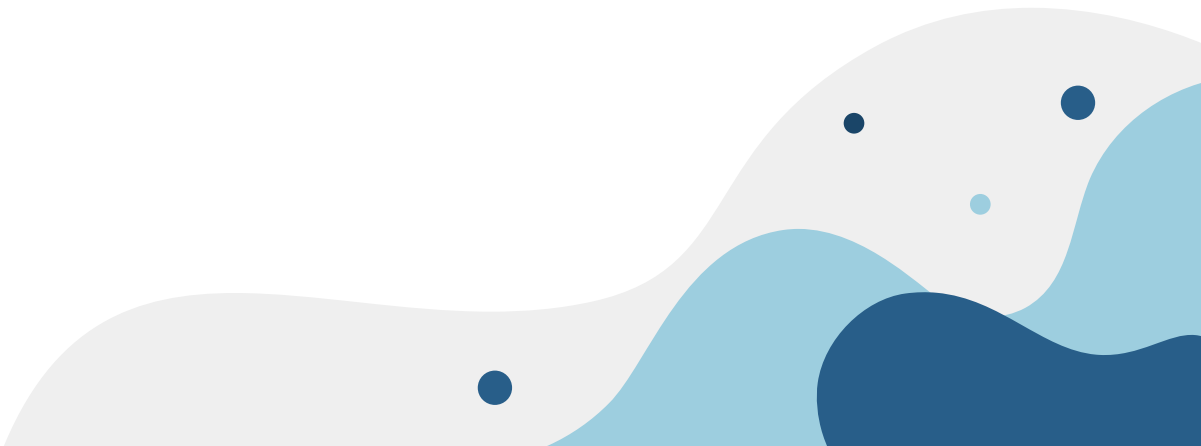


### Top used 100 Words after Text Cleaning



# Result Analysis

- We Conducted research on sentiment analysis using various algorithms, focusing on precision, recall, and F1-score.
- Identified imbalances in sentiment class distribution, particularly in neutral and mid-pole sentiments.
- Implemented Oversampling to address imbalance, ensuring equal representation for each sentiment class during model training.
- Utilized bias handling techniques to improve model accuracy and fairness across all sentiment classes.



# Result Analysis

## Imbalance Handling in Dataset:

- Addressed dataset imbalances causing ML models to develop bias towards dominant classes.
- Skewed accuracy scores were observed due to models predominantly predicting the majority class.
- Implemented two techniques: class-weight computation and random oversampling, to address imbalance.
- Both techniques resulted in close accuracy outcomes, with Complement-Naive Bayes showing higher class weights for oversampling.

Accuracy of the Model: 77.08%

```
classification report:
              precision    recall  f1-score   support

     1         0.72      0.88      0.79       4954
     2         0.27      0.05      0.09       1377
     3         0.85      0.85      0.85       5988

 accuracy          0.77       12319
 macro avg         0.61      0.59      0.58       12319
 weighted avg      0.73      0.77      0.74       12319
```

# Result Analysis

MultinomialNB Naive Bayes prediction score for 5labels : 0.6124685445247179

Accuracy of the Model: 61.25%

classification report:

	precision	recall	f1-score	support
1	0.57	0.85	0.68	3531
2	0.29	0.11	0.16	1424
3	0.29	0.10	0.15	1377
4	0.39	0.29	0.33	1568
5	0.78	0.86	0.82	4419
accuracy			0.61	12319
macro avg	0.46	0.44	0.43	12319
weighted avg	0.56	0.61	0.57	12319



# Limitation & conclusion

## Limitations:

1. **Interpretational Challenges:** Sentiment analysis faces difficulties in accurately interpreting complex emotions, sarcasm, or cultural nuances within textual expressions, potentially leading to misinterpretations.
2. **Model Constraints:** While Naive Bayes was employed, its simplistic assumptions may limit its ability to capture the full spectrum of human emotions, suggesting the need for more sophisticated models for comprehensive sentiment analysis.

## Conclusion:

1. **Continuous Evolution:** Despite current limitations, ongoing exploration of advanced techniques like LSTM and Tf-Idf promises to enhance the accuracy and robustness of sentiment analysis methods.
2. **Knowledge Dissemination:** Through efforts to refine research papers and share insights, the project aims to contribute valuable resources for future sentiment analysis research, fostering progress and innovation in the field.

## ● References

- Mishra, A., Sahay, A., Pandey, M. A., & Routaray, S. S. (2023). News text analysis using text summarization and sentiment analysis based on NLP. In 2023 3rd International Conference on Smart Data Intelligence (ICSMDI) (pp. 28-31). Trichy, India.  
<https://doi.org/10.1109/ICSMDI57622.2023.00014>
- Tejaswini, Z. , Rajeswari, K. (2022). Sentiment Analysis of Covid-19 Tweets using Twitter Database–A Global Scenario.  
<https://ieeexplore.ieee.org/document/9989000>
- Dzyuban, Y., Ching, G., Yik, S., Tan, A., Crank, P., Banerjee, S., Pek, R. and Chow, W. (2022). Sentiment Analysis of Weather-Related Tweets from Cities within Hot Climates. *Weather, Climate, and Society* 14(4) pp. 1133-1145. Available at:  
<https://journals.ametsoc.org/view/journals/wcas/14/4/WCAS-D-21-0159.1.xml> . [Accessed 19 Apr 2024].
- A. R. Atmadja, A. Rahmawati, C. N. Alam, P. Dauni and Y. Saputra. (2023) . Sentiment Analysis on Tourism Place using Naive Bayes .  
<https://ieeexplore.ieee.org/document/10366891>

**Thank you**  
**Any Question ?**