Text-mining Assignment Submission Cover Sheet

|  |  |
| --- | --- |
| Assignment Title | Twitter Emotions |
| Module | Part 2 |
| Student Name | Kaihua Wen |

### **Data Understanding and Preparation (20 points)**

* **Definition of Problem**

Clearly state the problem definition, what type of data mining task it is, where was the data set sourced from, etc.

Business Understanding is the first phase of the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. It focuses on understanding the project's objectives and requirements from a business perspective. Text mining, also known as text data mining, is the process of deriving high-quality information from text. It involves deriving high-quality information from unstructured text using algorithms and software to extract patterns, trends, and insights. This project focuses on text mining to analyse a dataset sourced from Kaggle: [Emotions Dataset](https://www.kaggle.com/datasets/nelgiriyewithana/emotions)[1]. The dataset comprises English Twitter messages annotated with six fundamental emotions: anger, fear, joy, love, sadness, and surprise. For business understanding, understanding these emotions can provide valuable insights for businesses in various domains, such as customer service, marketing, and product development. The primary objective is to accurately classify tweets into six fundamental emotions. This classification can help businesses in decision-making and strategy formulation by providing a clearer understanding of customer emotions and trends.

* **Data Preparation**

Include details of any data cleaning, transformations, data enrichment, feature engineering, feature reduction, etc

Data Preparation is the third phase. This phase involves transforming raw data into a clean and structured format suitable for analysis. Before EDA and model building, data preprocessing is crucial. Initially, noisy and irrelevant text, such as HTML tags (“<br><\br>”), usernames (“user@name”), URLs, brackets, special characters, and numerical digits, are removed. Using the NLTK package, I tokenize the text, do the lowercasing and remove stop words. Additionally, I apply stemming [2] and lemmatization [3] techniques. Stemming reduces words to their root form by removing suffixes, while lemmatization converts all word variations to their base form. Stemming uses heuristic rules to chop off word endings, making it faster and simpler, as seen in algorithms like the Porter Stemmer. For example, "running," "runner," and "ran" might all be reduced to "run." In contrast, lemmatization considers the morphological analysis of words and their context, using vocabulary and part-of-speech tagging to return the base form or lemma. For instance, "running" and "ran" would be converted to "run," and "better" to "good." This makes lemmatization more accurate but also more complex and resource-intensive.

### **2. Data Visualization and Exploration (20 points)**

* **Data Exploration & Descriptive Analytics**

Include any data insights discovered through initial exploration and descriptive analytics.

Data Understanding is the second phase of the CRISP-DM methodology. It focuses on collecting and understanding the data that will be used for analysis. As above mentioned, the dataset has six classes. The dataset has two features: text and label which also includes 410k data. Initially, finding the amount of each class is crucial. As depicted in Figure 1(a), the category counts are presented. It is observed that the majority of tweets express joy, followed by sadness, with surprise being the least frequent emotion. And Figure 1(b), is the distribution of categories. Joy accounts for 33.8% of the pie, followed by sadness, anger, fear and love at 29.1%, 13.8%, 11.4% and 8.3% respectively. The smallest proportion of it is surprise, accounting for only 3.6%. Next, I tend to talk about the details of each text sentence, discovering the text length and word cloud. We can see in Figure 2, the maximum of text length is round 300 and most of sentences’ word are between 50 and 100. A few sentences have over 200 words.

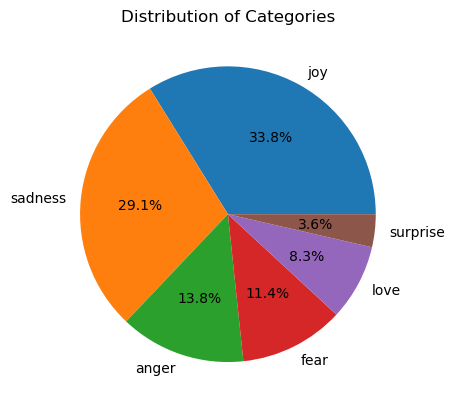
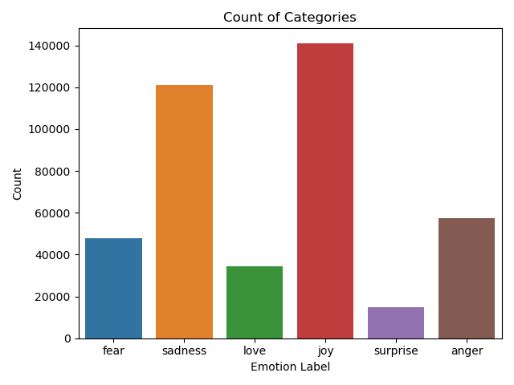


Figure 1. (a) The count of categories and (b) the distribution of categories

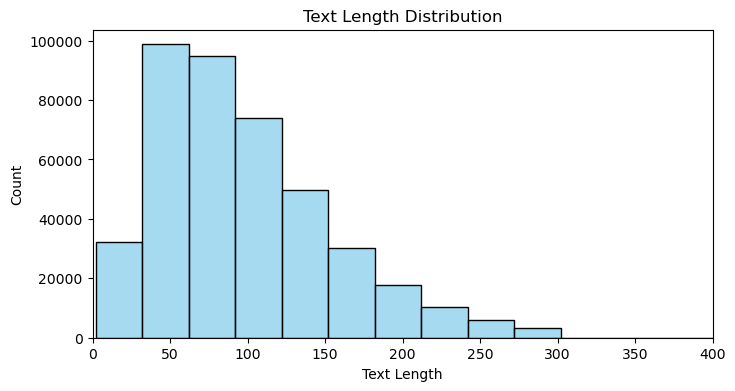


Figure 2. The distribution of text length

To determine the frequency of words appearing in different classes, word clouds were generated for each category as shown in Figure 3. In each sub-plot, it is apparent to obverse different keywords in the category. For example, “scare” and “fear” in fear class, “hate” and “depress” in sadness class, “love” in love class, etc. The t-SNE [13] visualization of the about 400k tweets reveals significant overlap among the data points, suggesting that the tweets across different emotional categories share substantial similarities in their textual features. This overlap implies that the features used for classification may not be sufficiently discriminative to separate the emotions clearly.

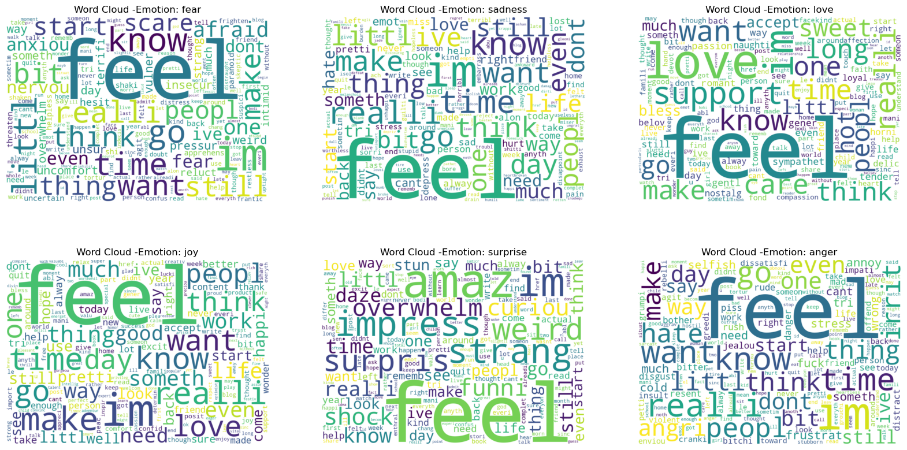


Figure 3. The word cloud

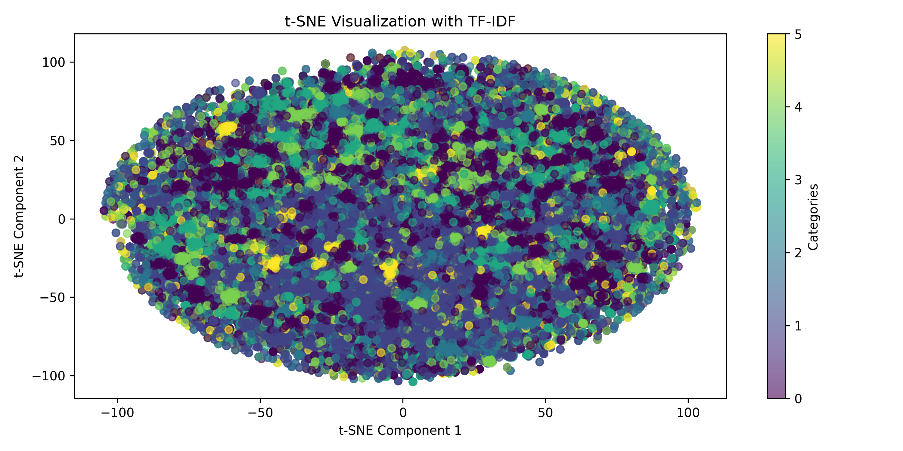


Figure 4. T-SNE visualization with TF-IDF

* **Identification of the most important variables**

Explain the process and methods used to identify the most influential variables in the data set.

Feature importance showcases the most relevant and influential features in making predictions or understanding patterns in the data. **Random Forest (RF)** [4]  is used to showcase the feature's importance mainly. This chart shows the impact of various features on the prediction accuracy of a random forest model for Twitter sentiment classification. Feature Dim279 is the most important, followed by Dim230 and Dim266, indicating these features significantly contribute to the model's performance.

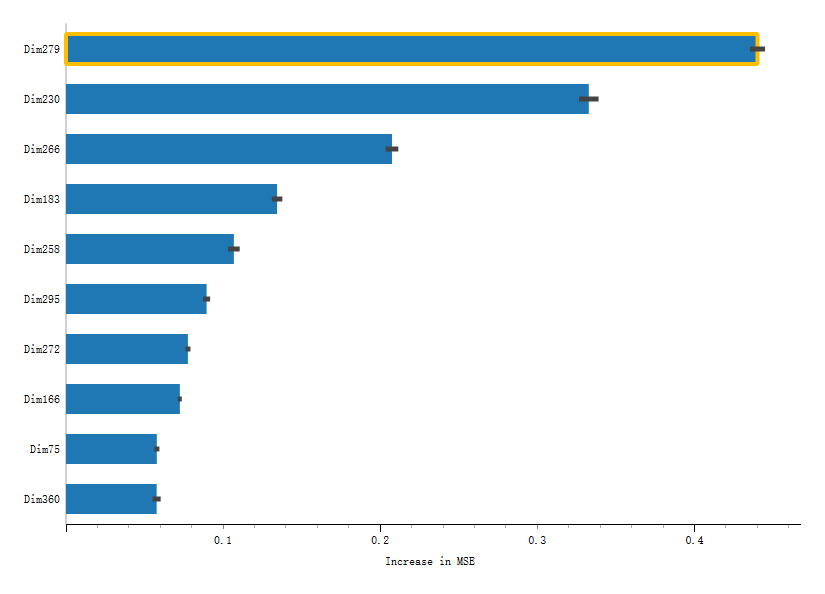


Figure 5. The feature importance in RF

### **3. Model Building and Evaluation (20 points)**

* **Details of Algorithms & Configurations**

Provide a detailed account of the algorithms used in the modeling process, including any specific configurations and parameters.

Modelling is the fourth phase of the CRISP-DM methodology. It involves selecting and applying various modelling techniques to the pre-processed data to create predictive or descriptive models. In this assignment, I implement 3 techniques (**Bags of Words** [5], **TF-IDF** [6] and **Word2Vec** [7]) and 5 machine learning models, **LogR** (Logistic Regression) [8], **Decision Tree** (**DT**) [9], **MultinomialNB** (**MNB**) [10], **Light Gradient Boosting Machine** (**LGB**) [11] and **XGBoosting** (**XGB**) [12]. To consider the **LGB** expects input features to be in 'float32/64' format, so all 3 techniques set “*dtype = float64*”. In Bags of Word and TF-IDF, my configuration is “*Norm = l2*”, “*max\_df* = 1” and “*min\_df = 1*” to specify the regularization and document frequency. In Word2Vec, I load the Google pretrain model to convert text from word to vector, which covers about 3 million words and phrases.

Among these machine models, In **LR**, The penalty was set to l2, the solver to lbfgs, and the multiclass option to auto. In **DT**, the criterion was set as Gini. In **MNB**, **LGB** and **XGB**, there is a default configuration. It is worth mentioning that setting “*probability=True*” in **XGB** to calculate AUC metrics successfully. Additionally, the data was split into training (70%) and testing sets (30%).

* **Model Performance Metrics & Evaluation of Results**

Present the performance metrics used to evaluate the models. Compare the results, highlighting the strengths and weaknesses of each model.

Modelling is the fourth phase of the CRISP-DM methodology. It involves selecting and applying various modelling techniques to the pre-processed data to create predictive or descriptive models. This section will discuss 15 models utilizing 3 techniques, accompanied by the visualization of their model performance metrics and evaluation results., including *Accuracy*, *Recall*, *Precision*, *AUC* and *time* [14]. Initially, using Bags of Words, Figure 6 shows that **LGB** has the best performance among these 5 models in terms of *Accuracy*, *Recall* and *Precision*, achieving 0.886 in *Accuracy*, 0.871 in *Recall* and 0.835 in *Precision*. The second-best model is **XGB**, with scores close to whose scores are close to **LGB**. On the other hand, **DT** performs the worst on average. Additionally, Table 1 reveals that **XGB** has the highest score in AUC of 0.986 while **MNB** has the shortest runtime, taking less than 1 second. Notably, the **DT** model not only has the worst AUC of 0.871 but also takes a considerable amount of time to run.

Figure 9 shows the confusion matrix for these models using the Bag of Words technique. It reveals that the **LGB** model is the best at predicting the six classes overall. **MNB** and **LR** models are particularly good at correctly predicting the 'sadness' label. **XGB** excels at predicting the 'surprise' label, whereas **MNB** struggles with this label, achieving only a 0.35 accuracy.

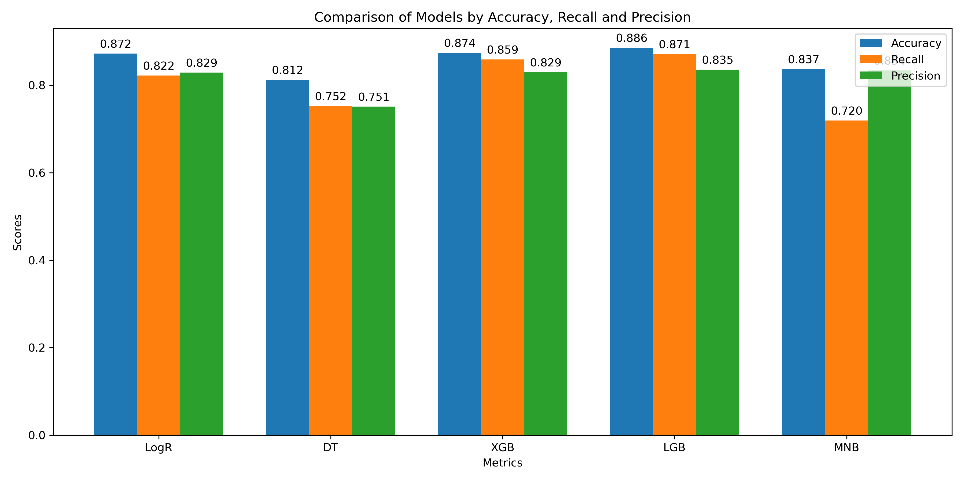


Figure 6. The comparison of metrics in Bags of Word

Table 1. The comparison of AUC and Time

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LogR | DT | XGB | LGB | MNB |
| AUC | 0.984 | 0.871 | **0.986** | 0.990 | 0.968 |
| Time(s) | 16.180 | 155.094 | 17.650 | 12.966 | **0.111** |

Then using TF-IDF technique, Figure 7 shows a similar result with Bags of Words. **LGB** still has best performance achieving 0.884, 0.866 and 0.834 respectively. On the other hands, **MNB** has the worst *Recall* scores, getting 0.538. Additionally, Table 2 reveals that the average runtime of TF-IDF is longer than Bags of Words. As we mentioned above, the best model under criterion of AUC and time is still **XGB** and **MNB**.

Next, Figure 8 and Table 3 shows the performance metrics and result of the Word2Vec technique. Surprisingly, the results are distinctly different from those of TF-IDF and Bag of Words. All models perform poorly using the Word2Vec technique, with the highest Precision score being 0.618, achieved by **LGB**. Moreover, most of the scores are lower than 0.600, indicating that Word2Vec is not suitable for this emotion classification task. In Table 3, **XGB** still has the highest AUC score, and **DT** consistently takes the longest runtime.

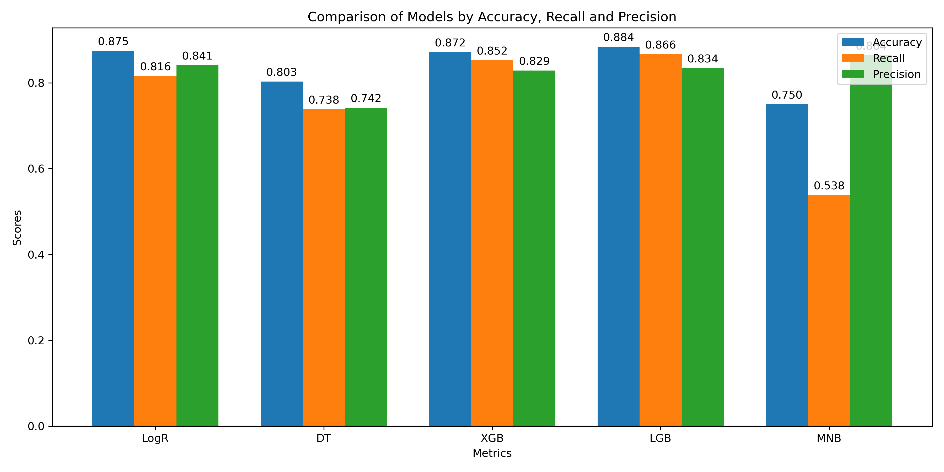


Figure 7. The comparison of metrics in TF-IDF

Table 2. The comparison of AUC and Time

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LogR | DT | XGB | LGB | MNB |
| AUC | 0.986 | 0.866 | **0.985** | 0.990 | 0.966 |
| Time(s) | 16.196 | 182.401 | 143.959 | 44.493 | **0.113** |

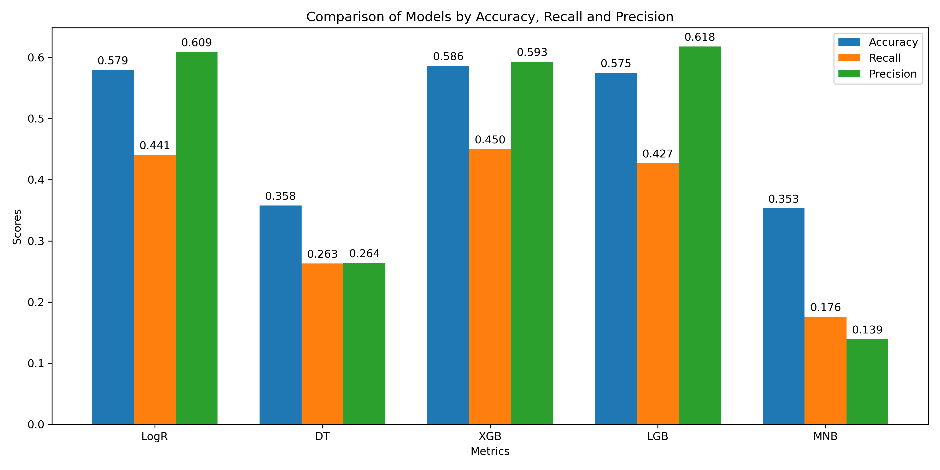


Figure 8. The comparison of metrics in Word2Vec

Table 3. The comparison of AUC and Time

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LogR | DT | XGB | LGB | MNB |
| AUC | 0.822 | 0.566 | **0.834** | 0.826 | 0.719 |
| Time(s) | 24.605 | 149.365 | 82.818 | 66.111 | **1.186** |

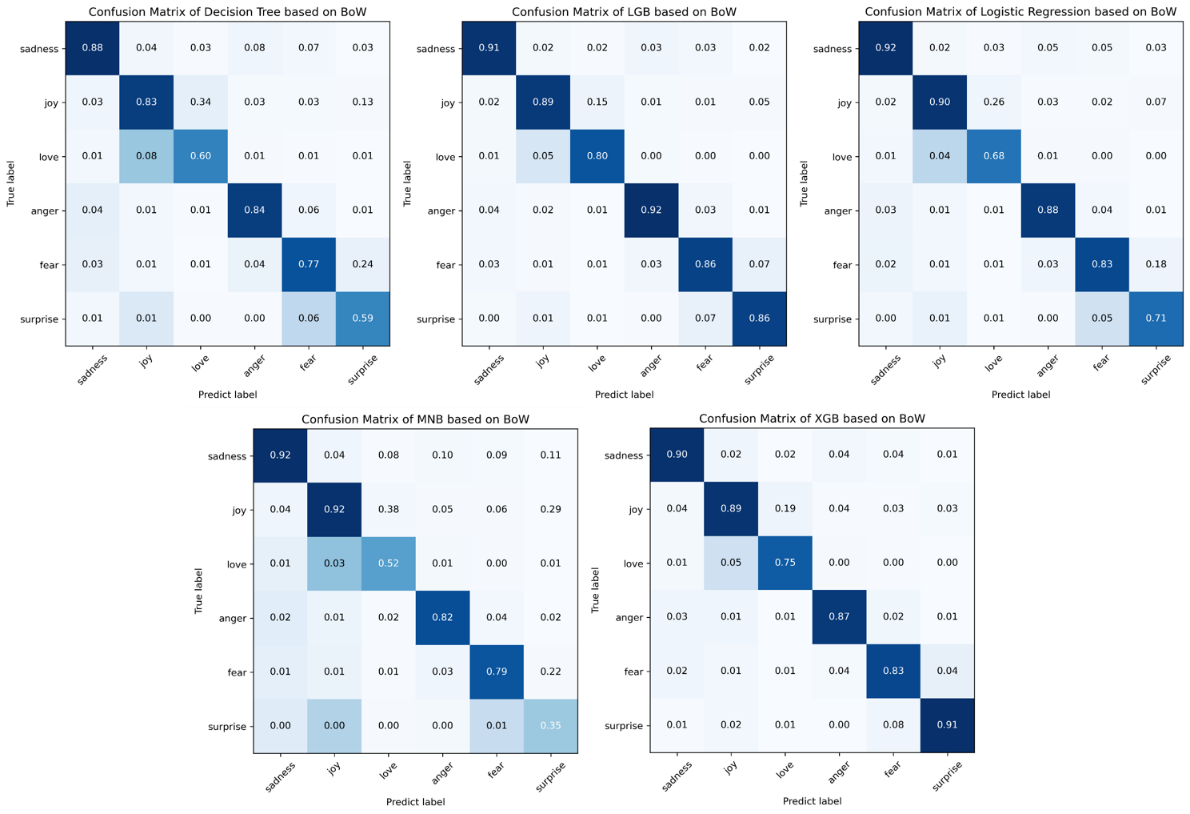


Figure 9. The comparison of metrics in Bags of Word

### **4. Advanced Techniques and Analysis (20 points)**

* **Advanced text-mining techniques applied to the dataset**

Describe any advanced techniques beyond basic modeling applied to the dataset.

In this section, advanced text mining techniques are implemented to discover the inner relationships within the dataset, including **Topic Modelling** (LDA)[15], **Kmeans**[16] and **hierarchical clustering**[17]. Applying topic modelling techniques to identify topics in the dataset is an initial thing to discover semantic patterns portrayed by a text corpus via LDA. Figure 10 shows the top 15 highest frequency of key words about different topic. According to these key words, finding that the emotion key words also appear in each topic. For example, “Love” appears in Topic 1, “Sad” appear in Topic 2 etc. In Figure 11, The plot, which uses principal component analysis (PCA) for dimensionality reduction, reveals significant overlap among clusters, particularly in the central region. The clusters exhibit varying densities and distributions, with some being more concentrated and others more spread out, indicating the complex and intertwined nature of the data points within this dataset.

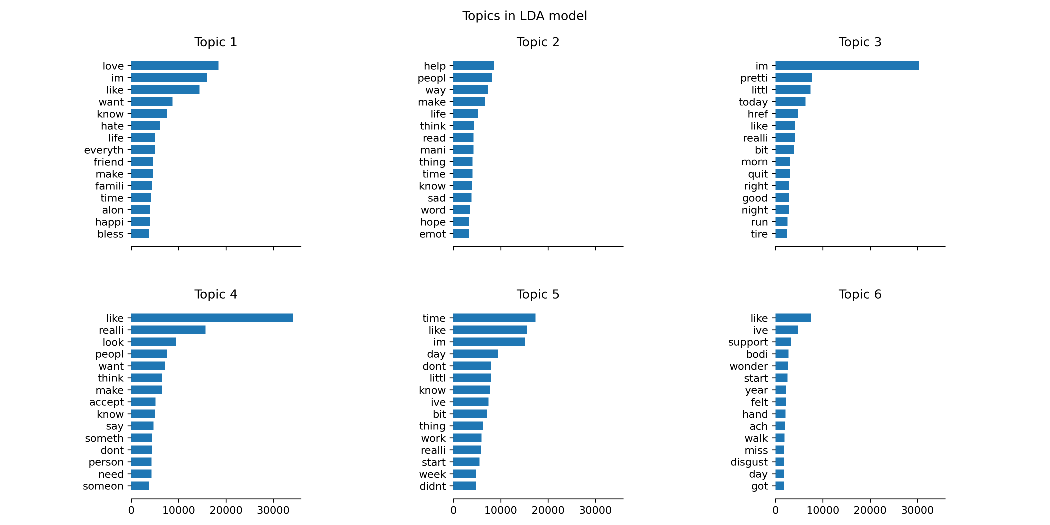


Figure 10. The top 15 key words of each topic in topic modelling

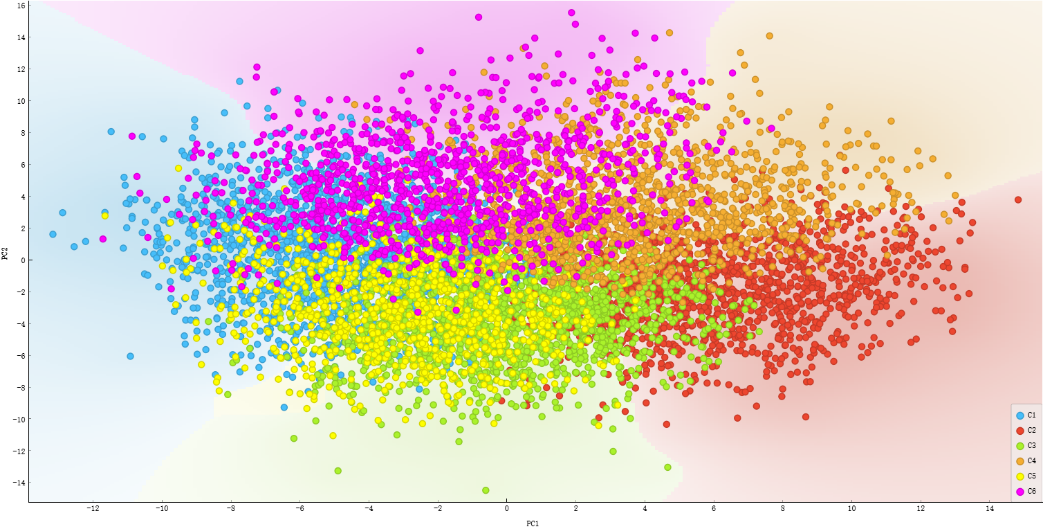


Figure 11. The result of Kmeans

Figure 12 shows the hierarchical clustering applied to the reduced dataset. The data are roughly divided into two clusters: positive (joy, surprise, and love) and negative (sadness, anger, and fear). Additionally, each data point is combined into clusters at each step, as indicated by distinct colours.

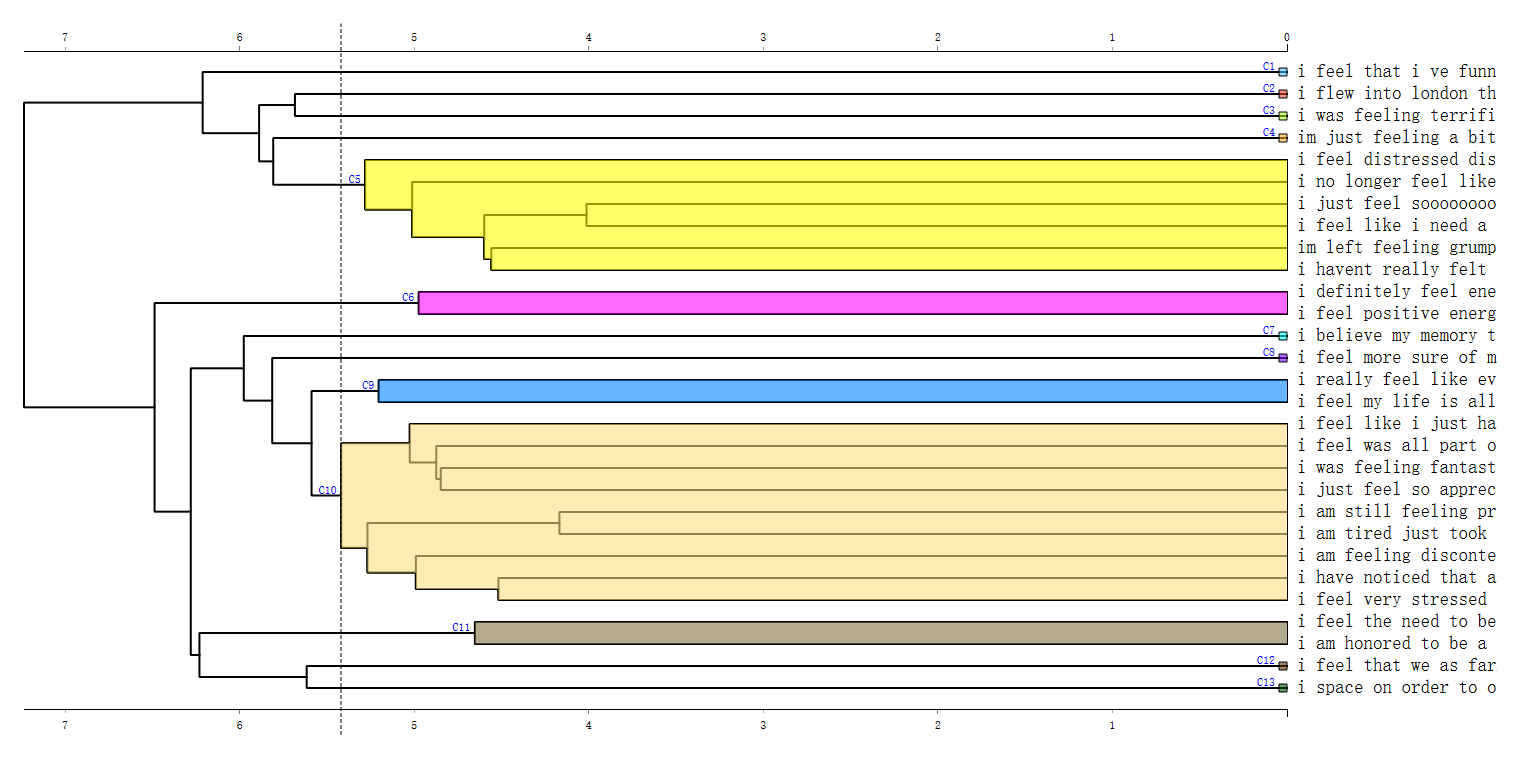


Figure 12. The result of hierarchical clustering

* **Comparison with other Research & Reflections**

Compare your results to at least three other researchers (maximum of five) who used the same data set. What lessons did you learn from doing this? How can your work be improved? Did you include any improvements in your work and what impact did it have?

After searching in Google scholar and Kaggle website, I found 3 different methods with same dataset, [Emotion Analysis and Model](https://www.kaggle.com/code/nordszamora/emotion-analysis-and-model#3)-Model)[18], [Naïve Bayes](https://www.kaggle.com/code/gauthamvijayaraj/na-ve-bayes-92-accuracy)[19], [Emotions Analysis | GRU](https://www.kaggle.com/code/abdmental01/emotions-analysis-gru-94#Model-Building)[20]. They represent different post preprocessing, pre-processing and advanced, GPU needed model. Firstly, the researcher implements **GridSearchCV**[21] to find the best parameter in best model choice, achieving the accuracy of 0.85 in **MNB**. Second, the researcher also implements **MNB** model, while using different pre-processing method. They convert classes from 6 to 3, roughly divided into positive label (fear and surprise), neutral label (joy and love) and negative label (sadness and anger) achieving accuracy of 0.92 in **MNB**. The last one, **Bidirectional Gated Recurrent Units (GRUs)**[22] are essential in natural language processing and sequence modelling tasks due to their ability to capture bidirectional dependencies in sequential data. After modelling building, fitting and visualizing, it achieves 0.93 in accuracy. It advanced and significant performance while the backwards is GPU-need and huge workload to execute.

After comparing results with those of other researchers, a decision was made to enhance preprocessing methods. Initially, special characters and noisy text were removed, and stemming and lemmatization were applied. However, the current approach involves omitting stemming and lemmatization, and re-running the Bag of Words-based models. This change led to significant improvements in the performance metrics and AUC of the five models. For instance, the accuracy of the **DT** model increased from 0.812 to 0.853, and the **LGB** model achieved an accuracy of 0.908.

### **5. List of Insights, Recommendations, and Future Works (20 points)**

* **Conclusions of the Study**

Summarise the key findings and conclusions derived from your analysis.

The analysis of the Kaggle Emotions Dataset, which comprises 410k English Twitter messages annotated with six fundamental emotions, yielded significant insights. The emotion distribution revealed that joy was the most prevalent, followed by sadness, anger, fear, love, and surprise. Keyword analysis identified distinct terms for each emotion, and t-SNE visualization indicated considerable overlap among emotional categories, suggesting similarities in their textual features. Model performance metrics highlighted that the LGB model outperformed others, achieving an accuracy of 0.884 with Bags of Words and improving to 0.908 after refining the preprocessing steps. Conversely, models based on the Word2Vec technique performed poorly, barely reaching 0.50 accuracy. Topic modelling results reflected the emotions accurately through distinct keywords in each topic, while K-means clustering revealed overlapping emotions, consistent with the t-SNE findings

* **Action Plan for Business Strategy Enhancement**

Provide specific recommendations based on the data analysis, including a detailed action plan on how these recommendations can be applied to improve the business strategy.

Based on the data analysis, several actionable recommendations can be implemented to enhance business strategy. For customer service, sentiment analysis tools should be used to monitor and respond to customer feedback in real-time, especially focusing on negative sentiments to improve satisfaction. In marketing, leveraging emotional keywords can help tailor messages that resonate with the target audience, creating positive brand associations. For product development, emotion-driven insights can guide the creation of features that address customer pain points, ensuring continuous improvement based on real-time feedback.

* **Consideration of Social Implications**

Discuss the social implications of the proposed actions, ensuring that the recommendations align with ethical standards and contribute positively to the broader community.

Implementing these recommendations must consider the social implications to ensure ethical standards and positive community impact. Privacy and data protection are paramount, necessitating strict measures to handle customer data. Transparency in how customer feedback is used will build trust, while providing support resources for customers expressing negative emotions will contribute to their emotional well-being. These considerations will help maintain ethical practices and foster a positive relationship with the customer base.

* **Future Work**

Suggest potential areas for future research and development that can build upon the findings of this study.

Future research should explore more advanced feature engineering techniques, such as word embeddings and deep learning, to improve the discriminative power of emotion classification. Expanding the analysis to include multiple languages can provide insights into emotional expressions across different cultures. Developing real-time sentiment analysis systems will enable continuous monitoring and response to customer emotions. Additionally, longitudinal studies can track changes in sentiment over time, and integrating sentiment analysis with other data sources can offer a comprehensive view of customer sentiment across various touchpoints. These steps will enhance the ability to analyse and respond to customer emotions, driving better business outcomes and satisfaction.

### **References (if possible with APA format)**

Use one of the commonly used References and Citation formats.

[1]Nidula Elgiriyewithana. (2024). Emotions [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/7563141

[2] Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.

[3] Plisson, J., Lavrac, N., & Mladenic, D. (2004, October). A rule based approach to word lemmatization. In Proceedings of IS (Vol. 3, pp. 83-86). sn.

[4] Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222.

[5] Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning (pp. 977-984).

[6] Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48).

[7] Goldberg, Y., Levy, O., Rong, X., Perozzi, B., Al-Rfou, R., Skiena, S., & Dean, J. (2003). word2vec parameter learning explained continuous bag-of-word model. In Proceedings of the International Conference on Learning Representations (ICLR 2013). https://doi. org/10.1162/jmlr (pp. 4-5).\

[8] Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). Logistic regression (p. 536). New York: Springer-Verlag.

[9] Su, J., & Zhang, H. (2006, July). A fast decision tree learning algorithm. In Aaai (Vol. 6, pp. 500-505).

[10] Xu, S., Li, Y., & Wang, Z. (2017). Bayesian multinomial Naïve Bayes classifier to text classification. In Advanced Multimedia and Ubiquitous Engineering: MUE/FutureTech 2017 11 (pp. 347-352). Springer Singapore.

[11] Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., & Zeng, W. (2019). Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agricultural water management, 225, 105758.

[12] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[13] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[14] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.

[15] Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data, 6(1), 1-18.

[16] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892.

[17] Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), 86-97.

[18] <https://www.kaggle.com/code/nordszamora/emotion-analysis-and-model#1)-Get-started>

[19] <https://www.kaggle.com/code/gauthamvijayaraj/na-ve-bayes-92-accuracy>

[20] <https://www.kaggle.com/code/abdmental01/emotions-analysis-gru-94#Model-Building>

[21] Ranjan, G. S. K., Verma, A. K., & Radhika, S. (2019, March). K-nearest neighbors and grid search cv based real time fault monitoring system for industries. In 2019 IEEE 5th international conference for convergence in technology (I2CT) (pp. 1-5). IEEE.

[22] Zhao, R., Wang, D., Yan, R., Mao, K., Shen, F., & Wang, J. (2017). Machine health monitoring using local feature-based gated recurrent unit networks. IEEE Transactions on Industrial Electronics, 65(2), 1539-1548.