Sentiment Analysis Whatsapp on Android App

DEVANAGA SAPUTRA

01 BUSINESS UNDERSTANDING

This dataset predominantly features user-generated reviews and ratings of the Whatssap Android App, with updates made on a daily basis. It also provides details on the relevancy of the reviews and the dates when they were uploaded. The latest update is on 2024 - 05 - 30. This dataset having 225876 data with 8 column the dataset has been upload on kaggle https://www.kaggle.com/datasets/ashishkumarak/whatssap-reviews-daily-updated

02 BUSINESS QUESTION

- What are the trends in user opinions of the WhatsApp Android app?
- Which models among Multinomial Naive Bayes, Bernoulli Naive Bayes, Decision Tree, Random Forest, Extra Tree Clasifier, and AdaBoost are suitable for predicting user opinions?
- What are the most frequent words or topics that appear in positive and negative reviews?

3. Data Understanding

reviewID: Unique identifyer for each review
userName: Name of user who sent the review
content: Text content of the review
score: Rating given by the user (1 to 5)
thumbsUpCount: Number of likes to the review
reviewCreatedVersion: Version of app when the review was created
at: Timestamp review
appVersion: Version of the app being reviewed

4. Data Peparation

Program Language: R
Packages:
Numpy
Pandas
Matplotlib
Seaborn
Nltk
String
Contractions
Re
NLTK
Wordcloud
Sklearn

4.1. Data Cleansing

Data cleansing starts with cleaning null values, then removing duplicate values in the dataset, as well as removing variables that are out of the topic. Additionally, creating a new variable called Rate_Label, which contains values from 0 to 2, where 0 represents negative, 1 represents neutral, and 2 represents positive sentiment.

- There are only 9 data points with missing values.
- There are 13,725 duplicate data points in the dataset.
- Formation of the Rate_Label is based on the Score given by the user. It is categorized as Negative if the score is between 1-2, Neutral if the score is 3, and Positive if the score is between 4-5.

- Removing NaN values is done to ensure that there is no missing information in the subsequent processes.
- The removal of variables reviewID, userName, reviewCreatedVersion, and appVersion is conducted because those variables are far from the topic of interest.
- The creation of the Rate_Label variable is performed to understand the sentiment trend from users.

After this process, the dataset contains a total of 212,151 data points.

4.2 Feature Engineering

In this process, the creation of the Year variable is based on the year of the review, preprocessing of user reviews to make them easier to process, and the addition of variables indicating the number of words and letters in each review.

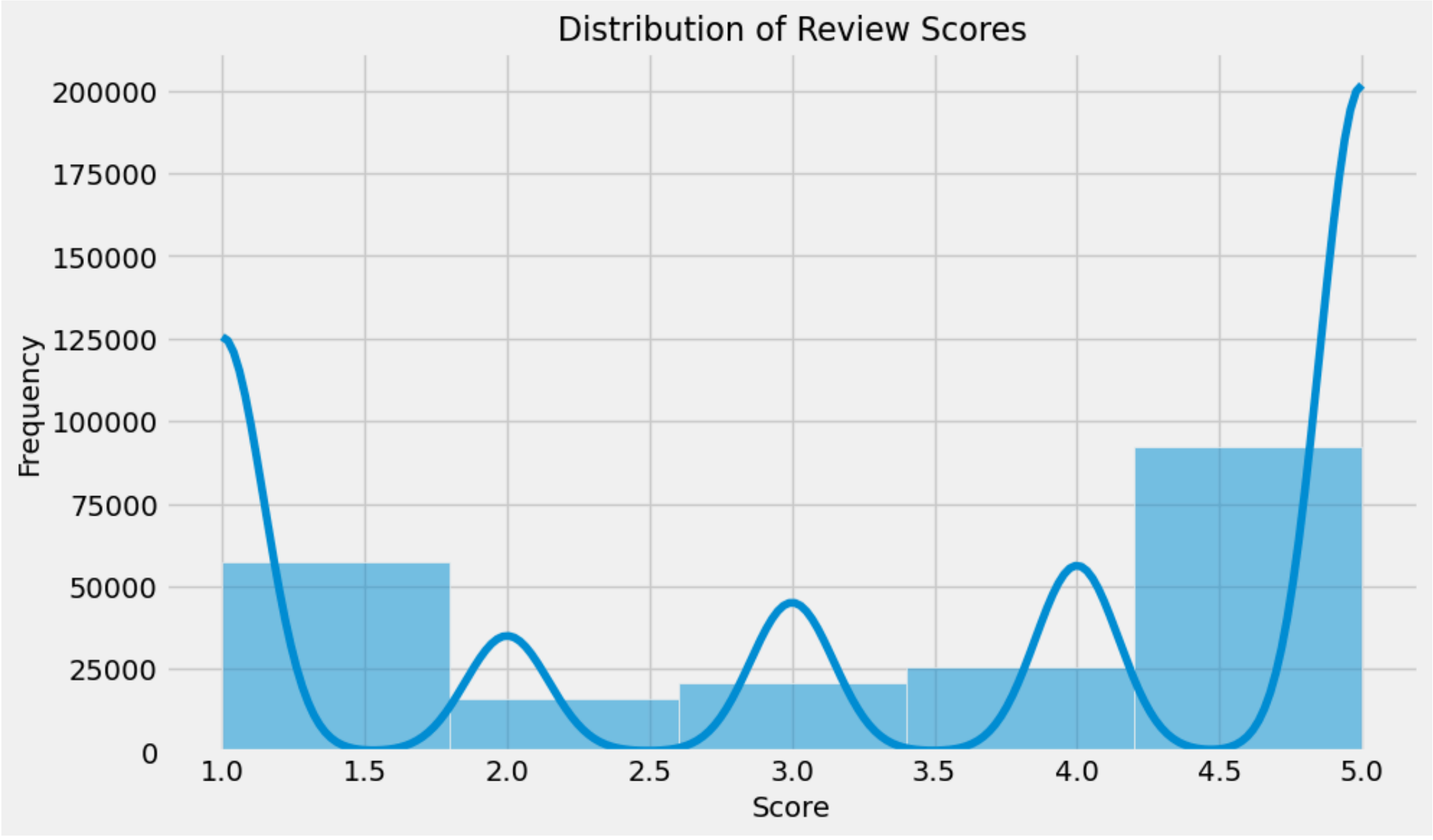
- The Year variable is needed to facilitate the exploration of trends in positive or negative reviews and to simplify the machine's learning process for the model creation later on.

- The creation of the Word and Char variables is used to assist the model in understanding the relationships between variables.

5. Exploratory Data Analysis

5.1 Distibution of Review Score

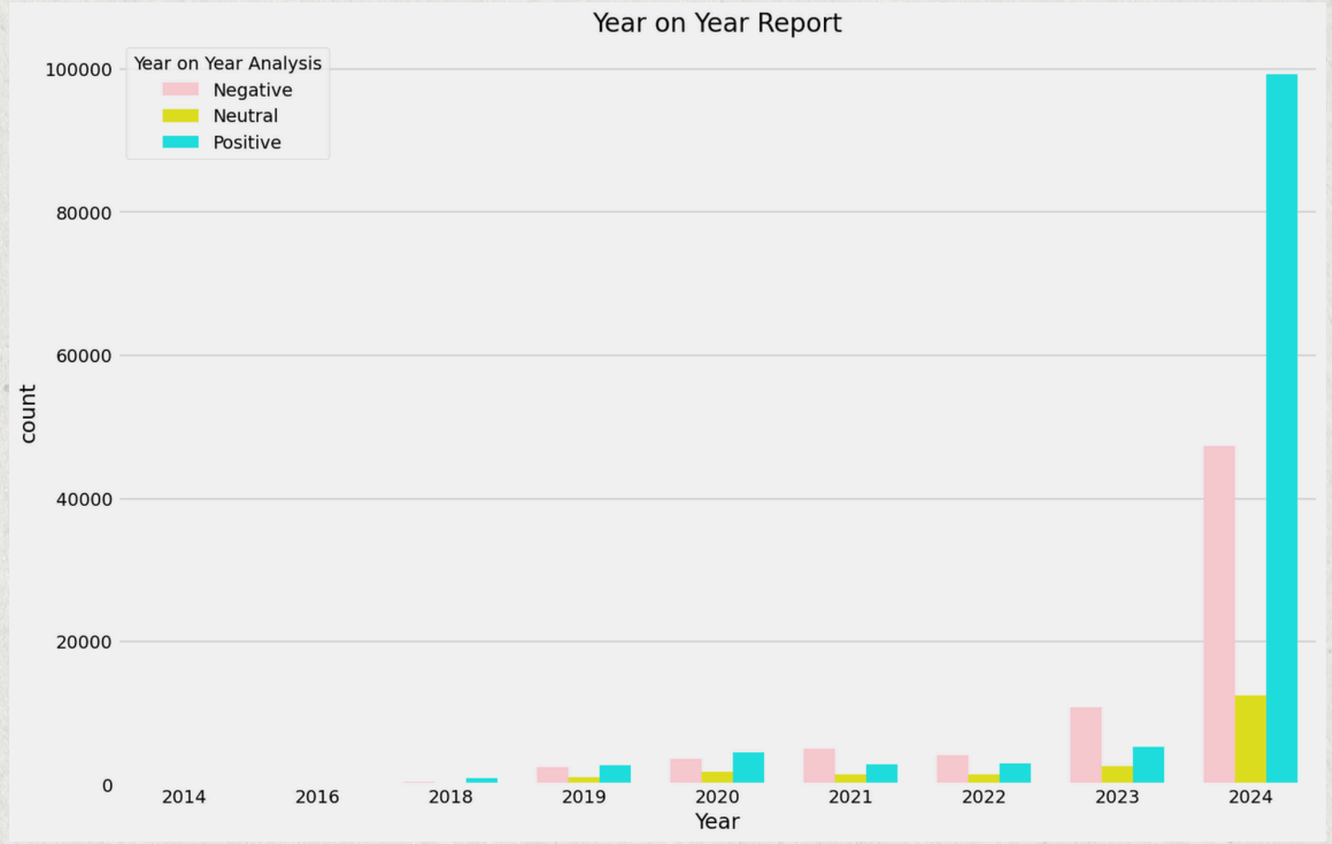
The distribution of review scores indicates how users perceive the WhatsApp application. A large number of reviews with a score of 5 suggest that the majority of users are highly satisfied with the application and have provided positive feedback. However, the presence of reviews with a score of 1 suggests that there are also users who are dissatisfied with the application and have provided negative feedback. This distribution highlights the variability in user opinions, ranging from very positive to very negative, regarding their experiences with the WhatsApp application.



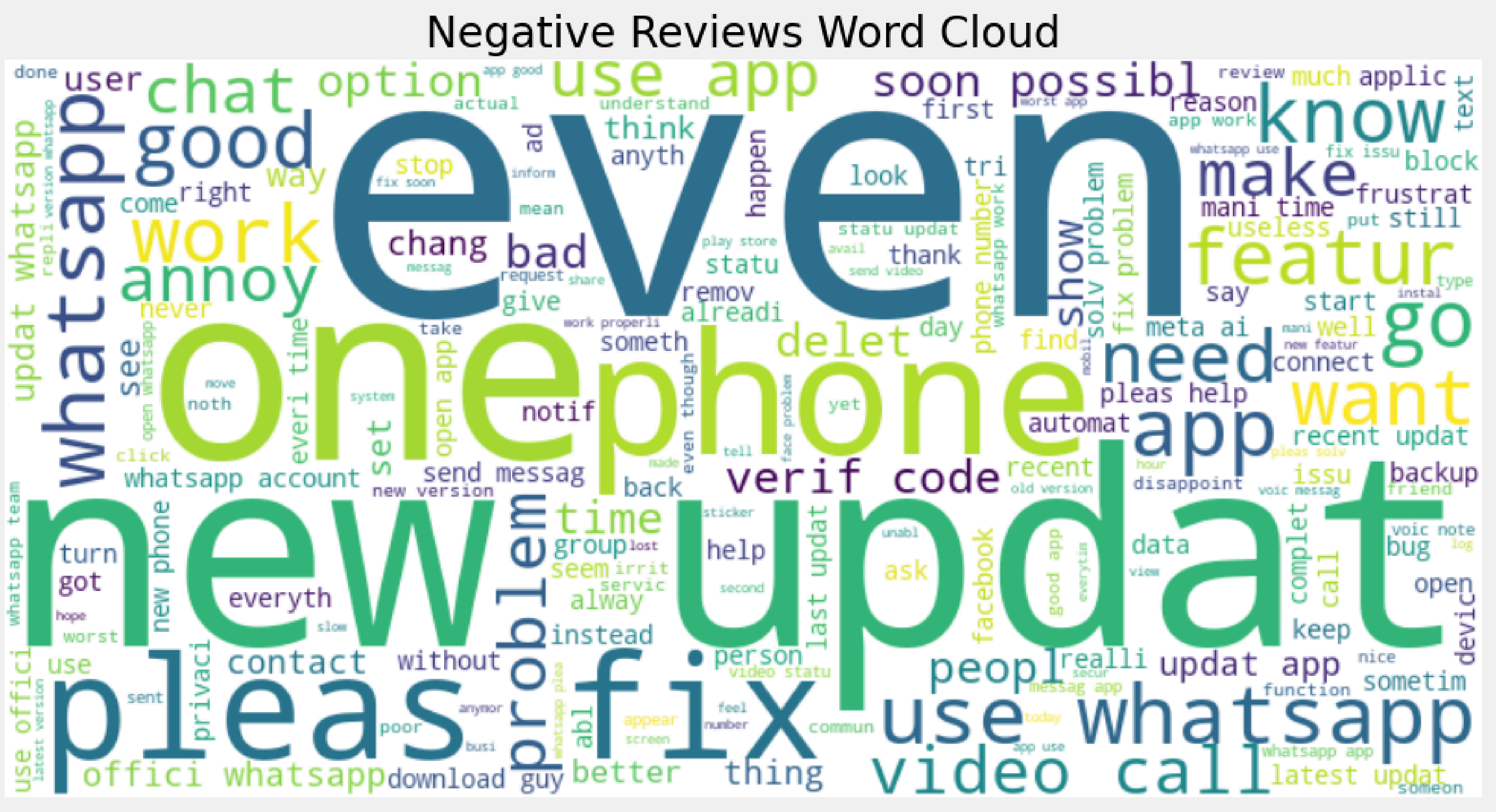
5.2. Year on Year Report

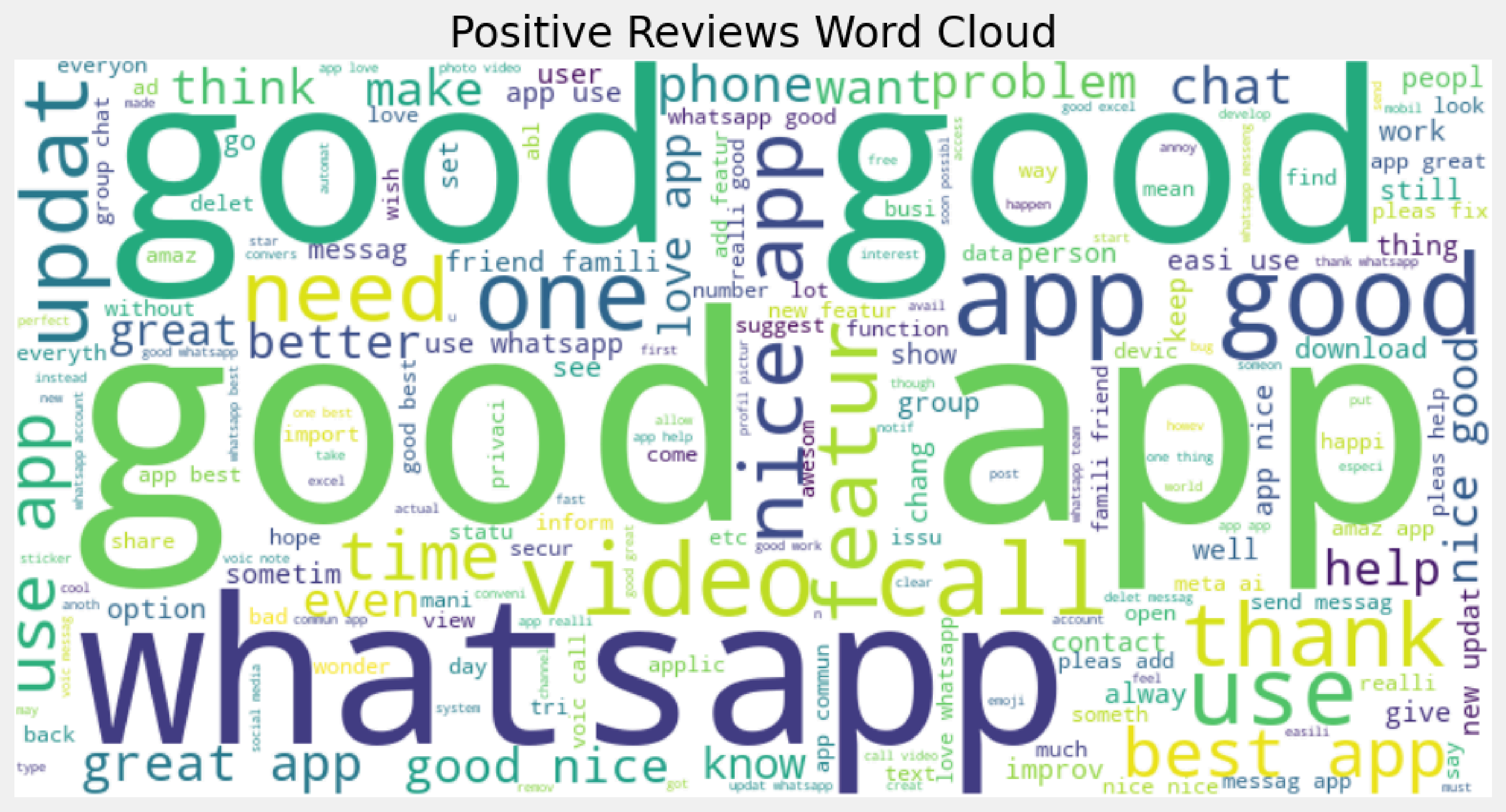
- This indicates that in 2024, users began to trust the performance of the WhatsApp application more. This trust may have been influenced by updates and features such as increased group chat capacity and additional features for group video calls and group calls.

- From 2019 to 2023, the period marked by the pandemic and post-pandemic era, quarantine measures were implemented, requiring work, school, and other activities to be conducted online. WhatsApp was one of the applications significantly impacted during this time, with a surge in user numbers. However, during these years, the features provided did not adequately meet users' needs. Limitations included the maximum group chat size, video calls restricted to 2 to 4 participants, and group calls also being limited. These constraints contributed to users giving low ratings or negative reviews for the WhatsApp application.



5.3. Workcloud Review





- Negative review keywords that stand out include 'New Update', 'Fix', 'Feature', 'Work', 'Solve Problem', 'Fix Problem', 'Chat', 'Option', and 'Video Call'.
- Positive review keywords that stand out include 'good', 'nice', 'improve', 'video call', 'feature', and 'update'.

6. Modelling

The model formation process will be carried out using the following models:

- MultinomialNB (mnc)
- BernoulliNB (bnc)
- DecisionTreeClassifier (dtc)
- RandomForestClassifier (rfc)
- ExtraTreesClassifier (etc)
- AdaBoostClassifier (abc)

These models will be evaluated based on their accuracy and precision scores.

6. Model Evaluation

]:

	Model	Accuracy	Precision
0	MultinomialNB	(MultinomialNB, 0.753706191519951)	(MultinomialNB, 0.6958515554367312)
1	BernoulliNB	(BernoulliNB, 0.6931108439982088)	(BernoulliNB, 0.6668038261054319)
2	DecisionTreeClassifier	(DecisionTreeClassifier, 0.6792523981239247)	(DecisionTreeClassifier, 0.6567159874788875)
3	RandomForestClassifier	(RandomForestClassifier, 0.7520328077494167)	(RandomForestClassifier, 0.694502203750418)
4	ExtraTreesClassifier	(ExtraTreesClassifier, 0.7516321383959085)	(ExtraTreesClassifier, 0.6985781462944427)
5	AdaBoostClassifier	(AdaBoostClassifier, 0.7197671403992553)	(AdaBoostClassifier, 0.6726901489614825)

The model selection was based on the highest accuracy and precision scores. Therefore, the model that was found to be the most suitable for representing this sentiment analysis is the Multinomial Naive Bayes model, with an accuracy of 75.37% and a precision of 69.58%. The Random Forest and Extra Trees Classifiers also proved to be quite suitable for this sentiment analysis case.

THANK YOU

Devanaga Saputra

+62815-4674-3628

devanagas7@gmail.com

https://github.com/DeppahNs

Jakarta Barat