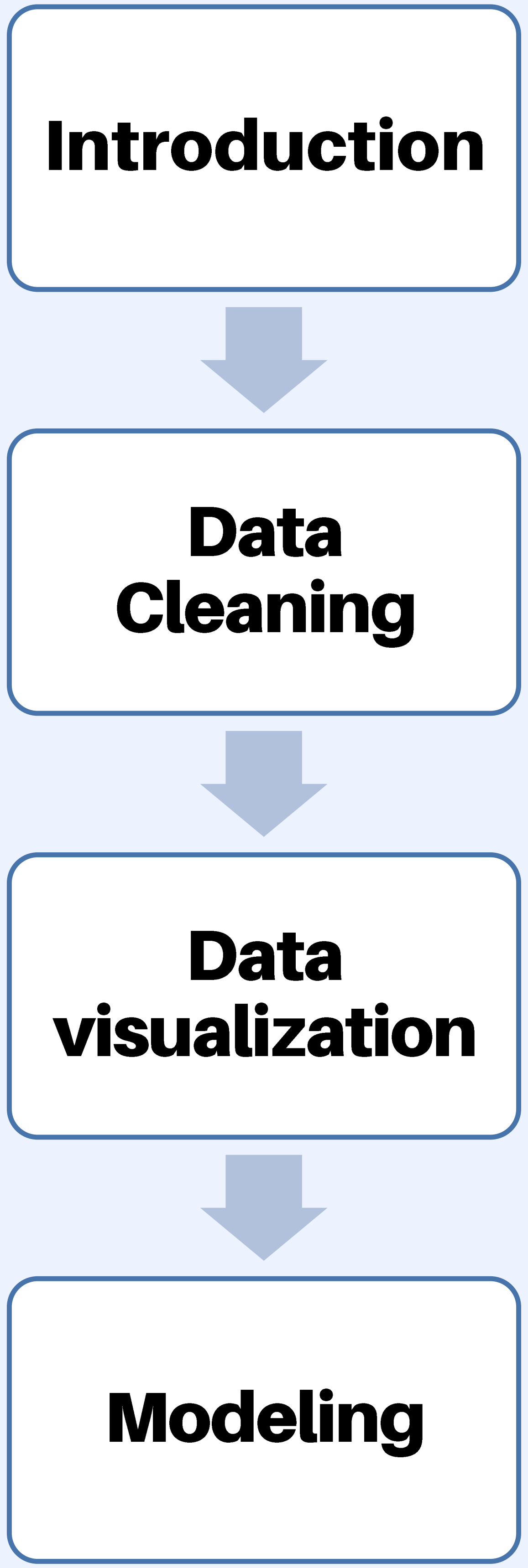
Quora Insincere Questions Classification

Haya Almadhi

Amwaj Hantol

Outlines



Introduction

Data Cleaning

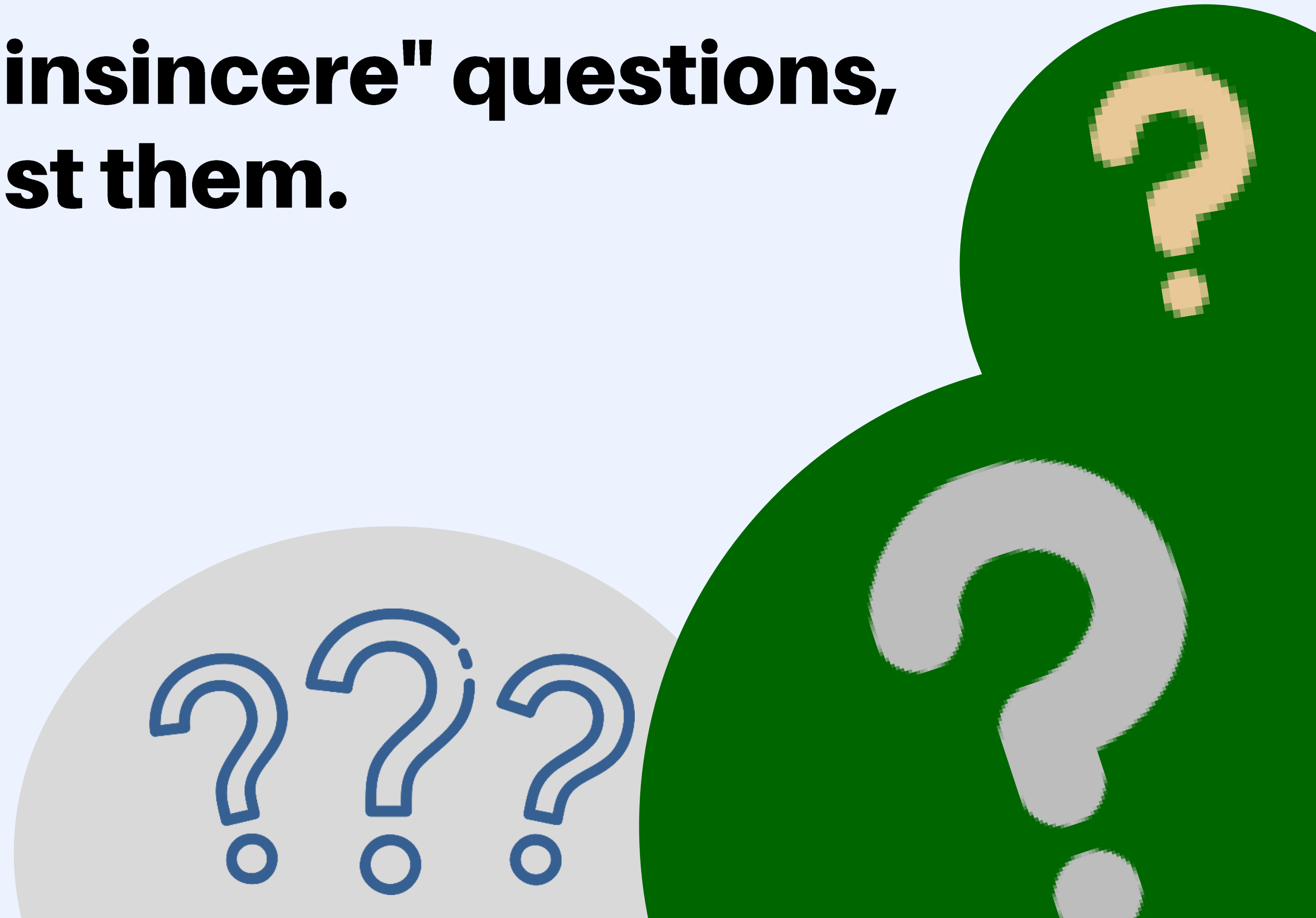
Data visualization

Modeling

Introduction

Quora is a question answer website, anyone can ask question from any domain, which are then answered by others.

In this project we propose to classify these "insincere" questions, then appropriate actions can be taken against them.



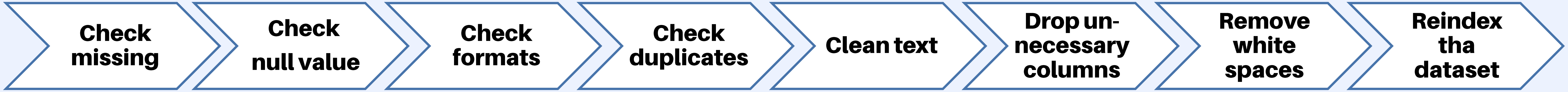
Dataset

Contains approximately 1.3 million rows and 4 columns:
Unnamed: 0 , target , qid and question_text.

From

Kaggle

Data Cleaning



Check missing

Check null value

Check formats

Check duplicates

Clean text

Drop un-necessary columns

Remove white spaces

Reindex tha dataset

Data Cleaning

Before cleaning:

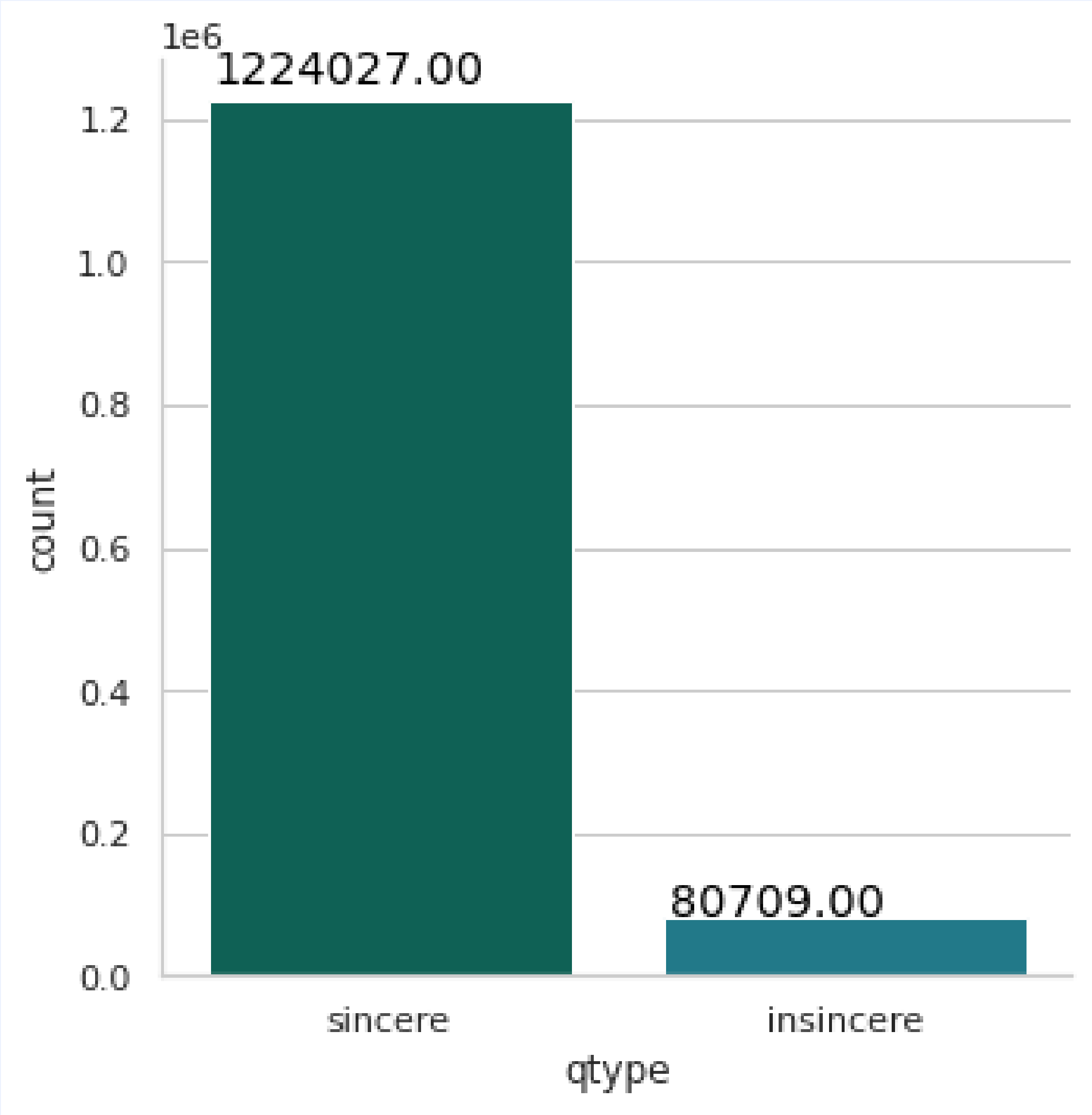
Unnamed: 0	qid	question_text	target
0	00002165[illegible]db923c7e6	how did quebec nationalists see their province...	0

After cleaning:

clean_qtext	question_text
quebec nationalist see provinc nation	sincere

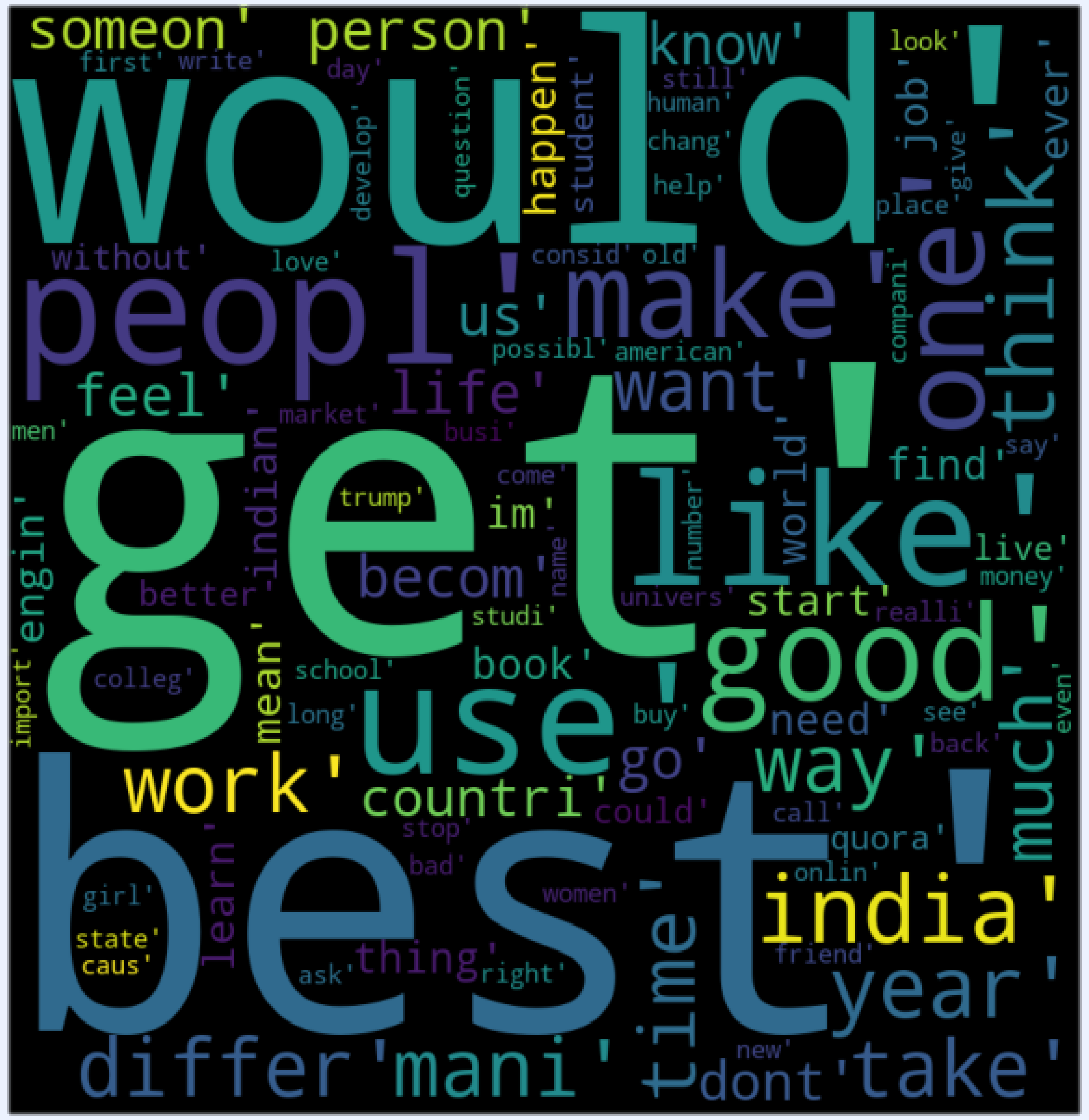
Data visualization

Which the most questions sincere or insincere

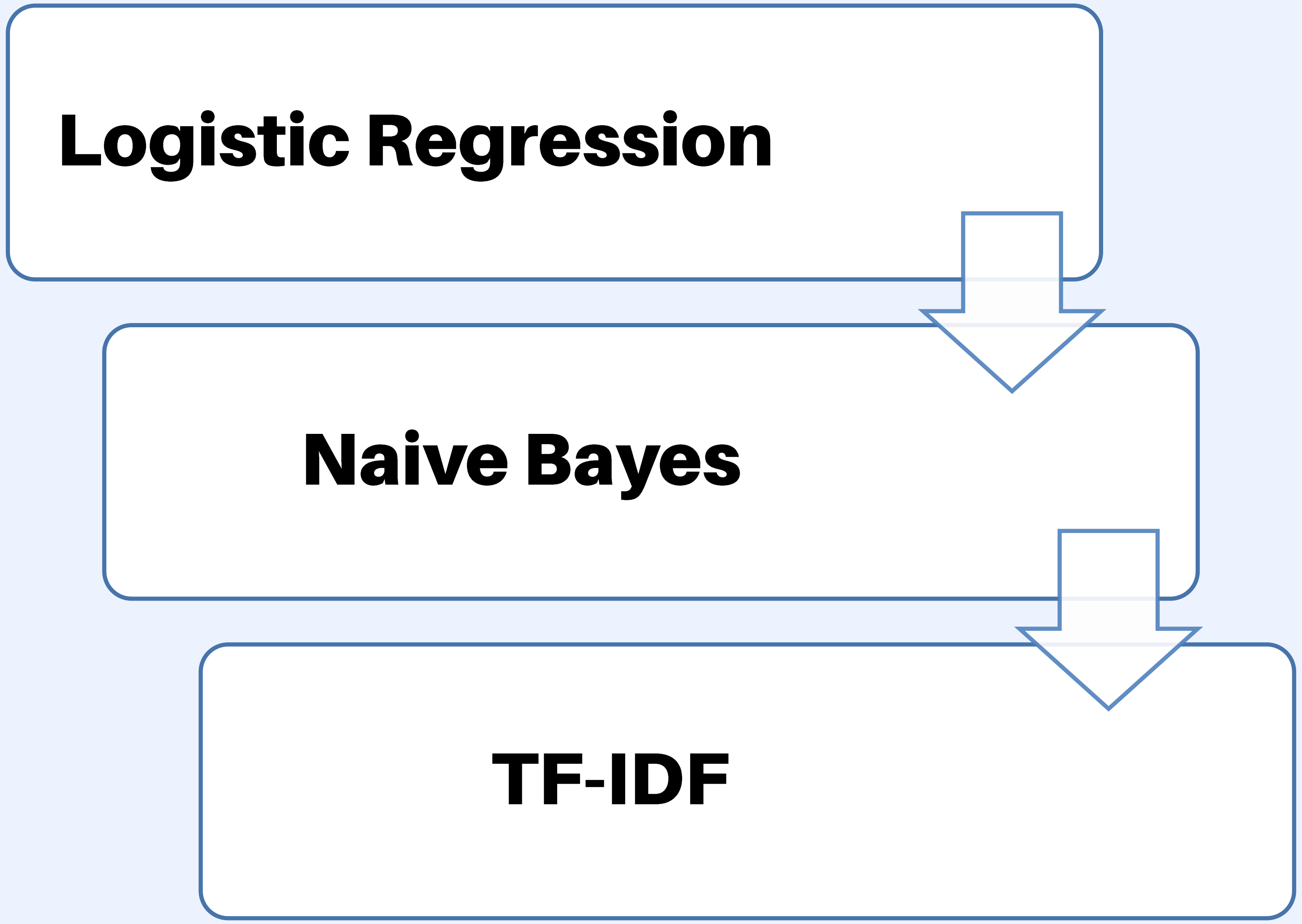


Data visualization

What is the most frequent words ?



Modeling



Logistic Regression

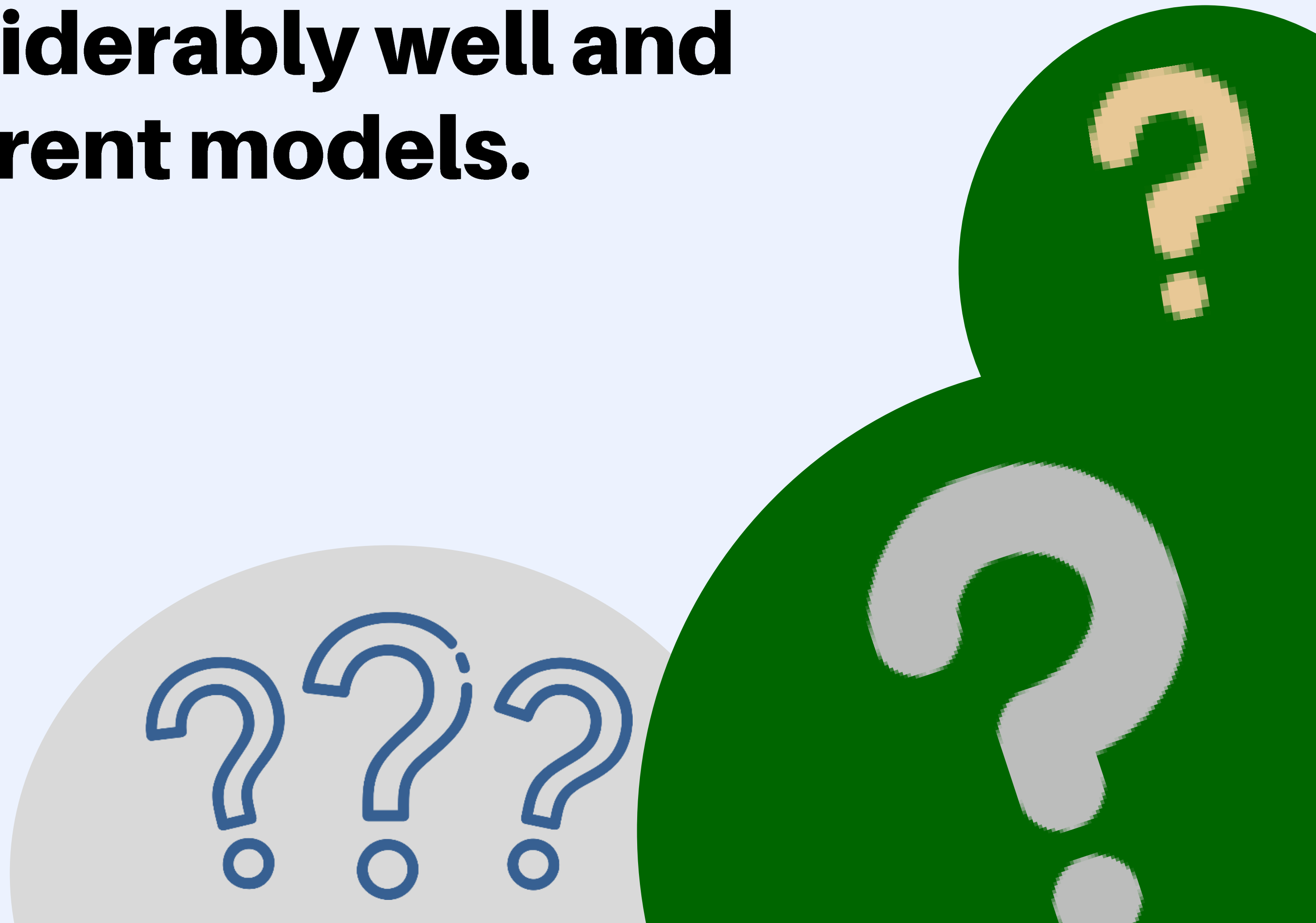
Naive Bayes

TF-IDF

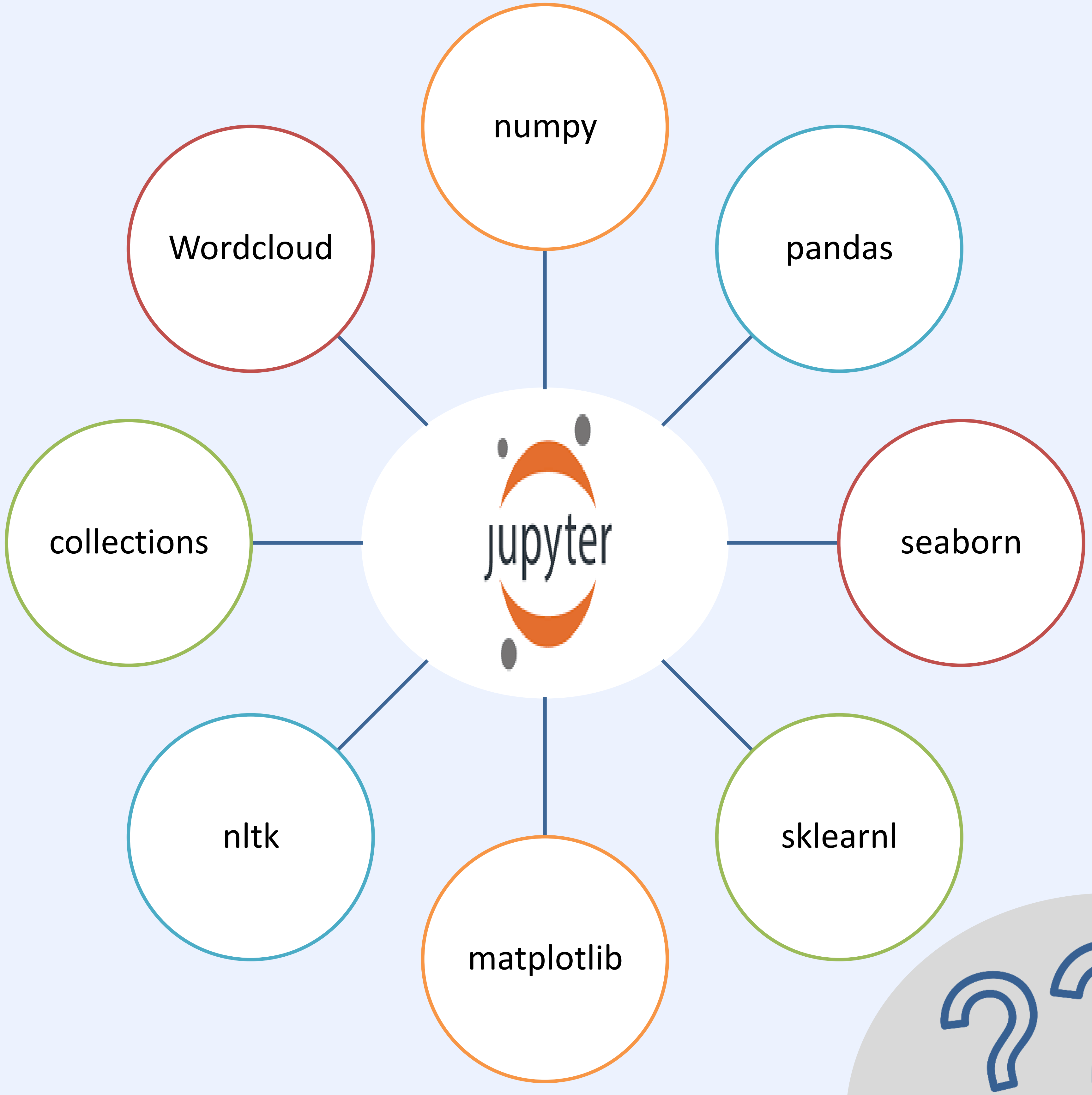
Conclusion

	LogReg1	LogReg2	NB1	NB2	LR1-TFIDF	LR2-TFIDF	NB1-TFIDF	NB2-TFIDF
Accuracy	0.942	0.942	0.939	0.934	0.941	0.941	0.940	0.934
Precision	0.948	0.948	0.954	0.955	0.947	0.947	0.941	0.955
Recall	0.993	0.992	0.983	0.976	0.993	0.993	0.998	0.976
F1 Score	0.970	0.970	0.968	0.965	0.969	0.969	0.969	0.965

According to various models, it can be stated that Logistic regression perform the best Other models also perform considerably well and there is no marginal difference between the different models.



Tools



jupyter

numpy

pandas

seaborn

sklearnl

matplotlib

nltk

collections

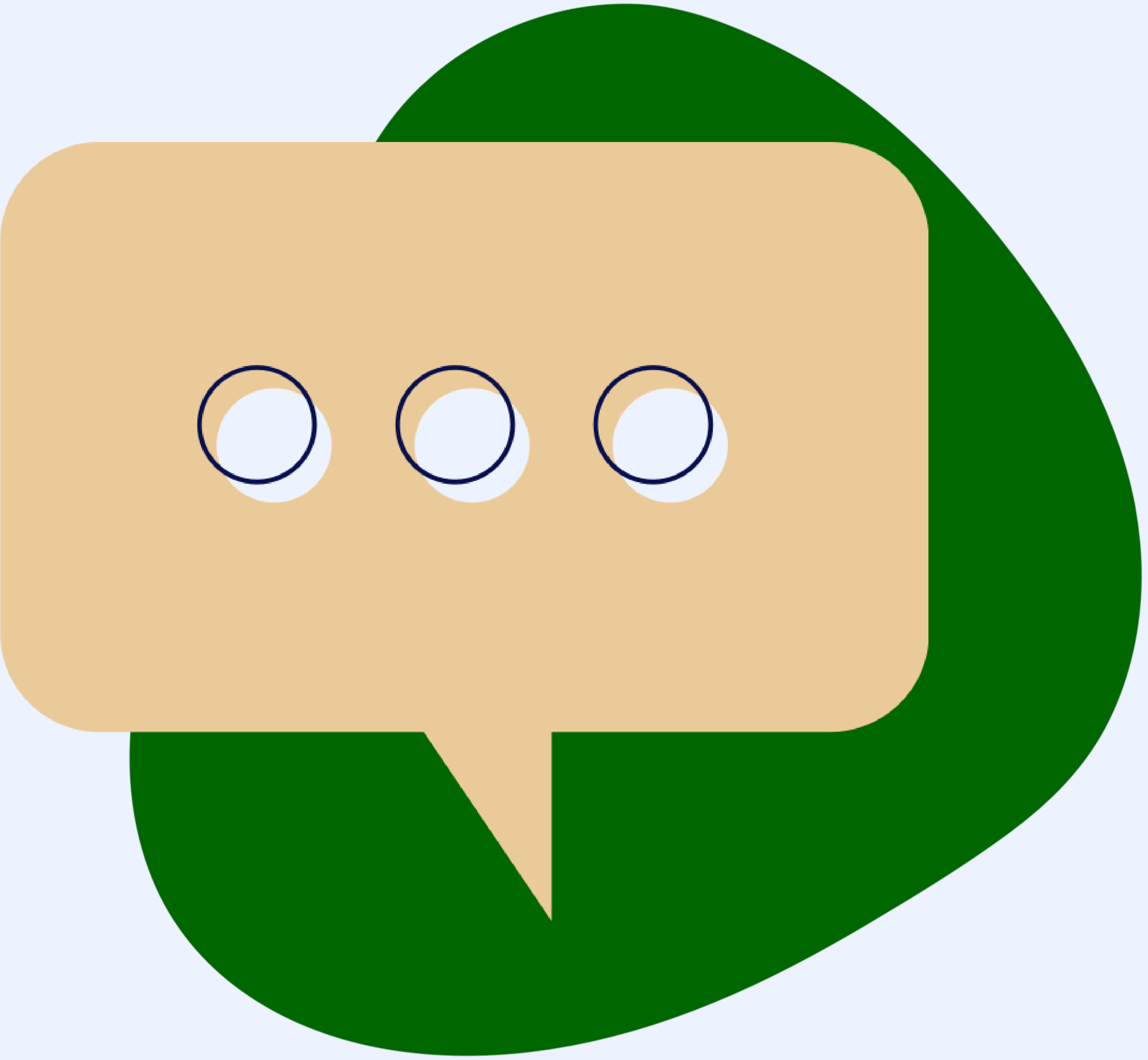
Wordcloud

Future Work

For future implementation, this project can be implemented using various deep learning models, LSTM, ...etc.

The problem of data imbalance can be solved by first training the data with equal number of sincere and insincere questions.

We will include more words for modeling.



Thank you