Messages Monitoring Model Capable of Detecting and Deleting Spam in Public Telegram Group Chats

Ahmed Almohammed's
Capstone Project for
MiSK DSI Program

Misk مسك
مؤسسة محمد بن سلمان
Mohammed Bin Salman
Foundation

المهارات
Skills

Contents

- Problem Addressed
- Data Acquisition
- Preprocessing Data
- Model Development
- Model Deployment
- Problems Associated with Current Solution
- Conclusion
- References
- End of Presentation

Problem Addressed

Data Acquisition

uciml

Kaggle

team-ai

Kaggle

@DeshDSingh

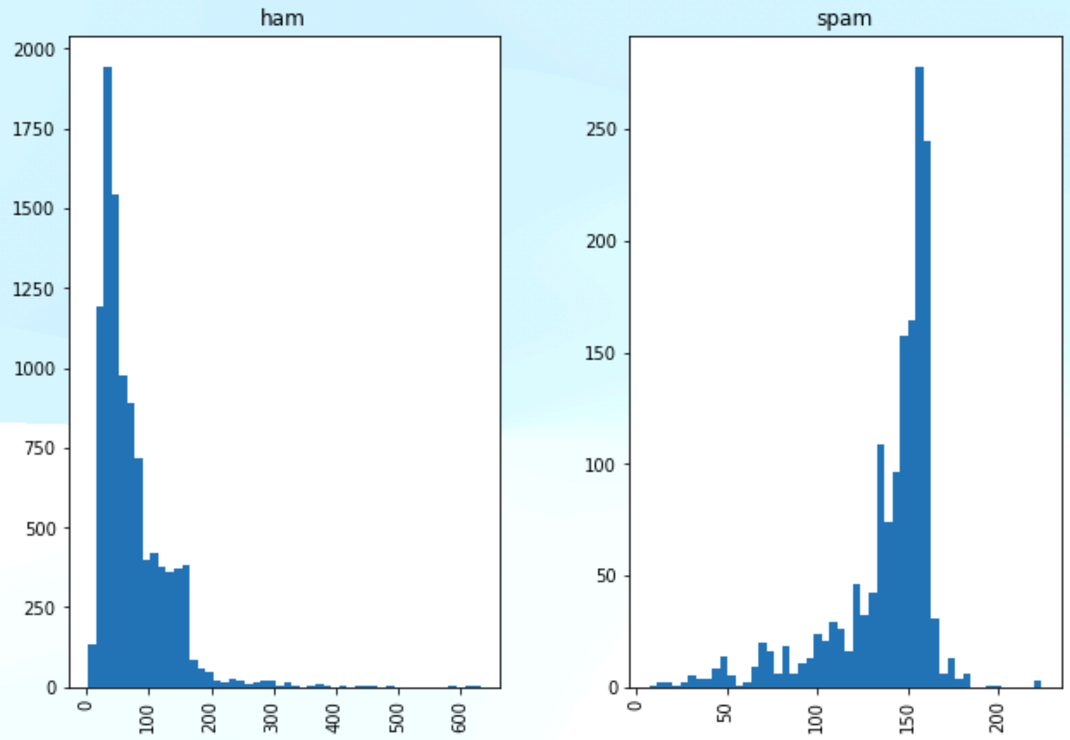
GitHub
Repository

🤗HuggingFace

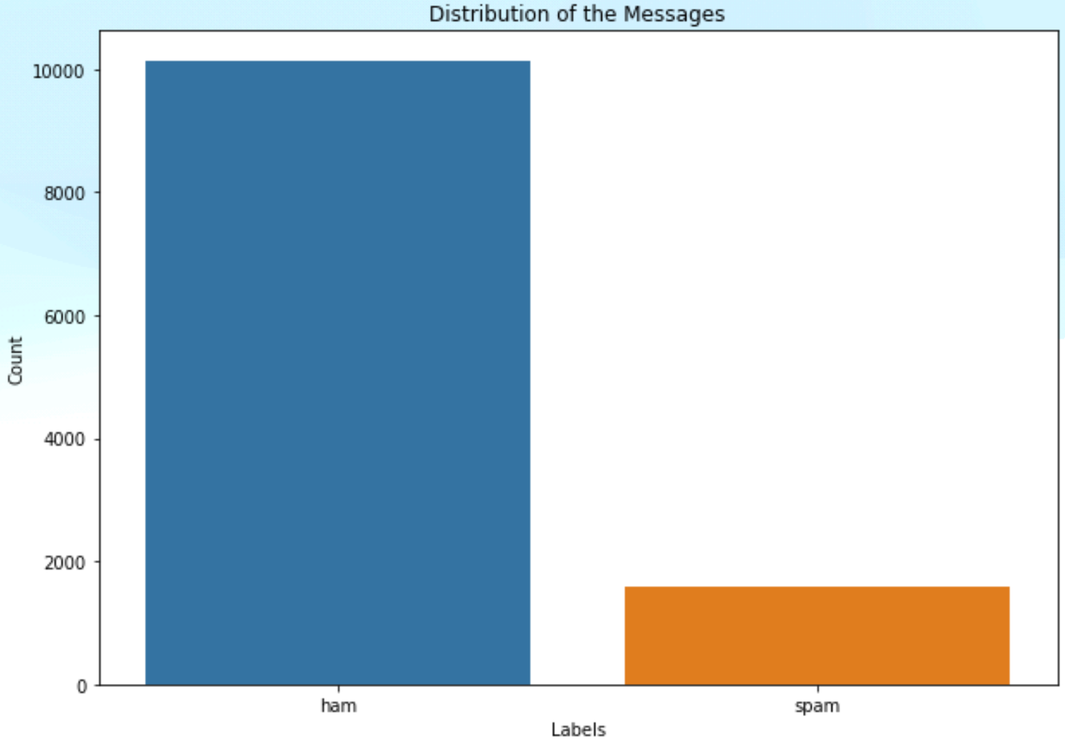
HuggingFace
Datasets

Preprocessing Data

Main Steps Taken in Preprocessing the Data



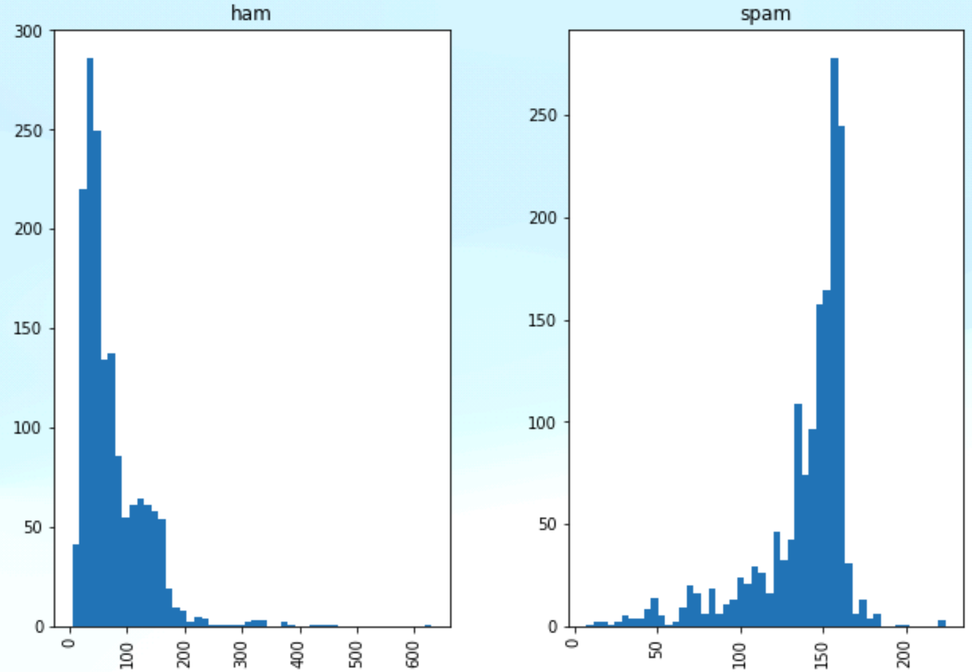
Examining the data

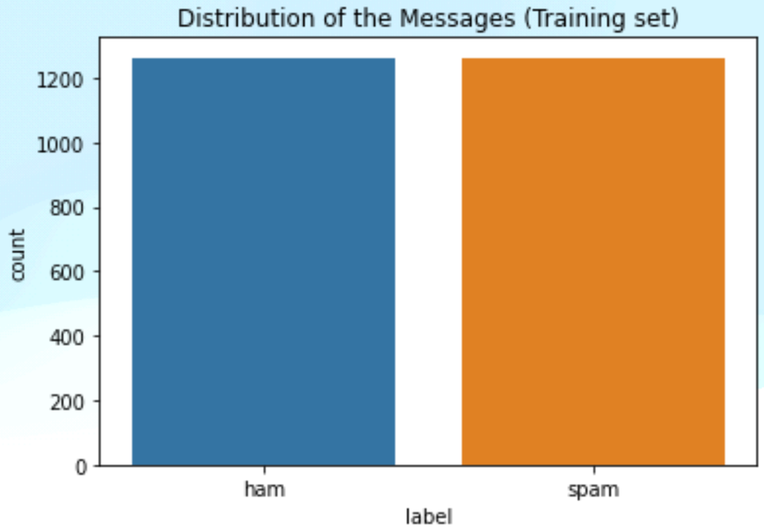


Fixing target imbalance

Main Steps Taken in Preprocessing the Data

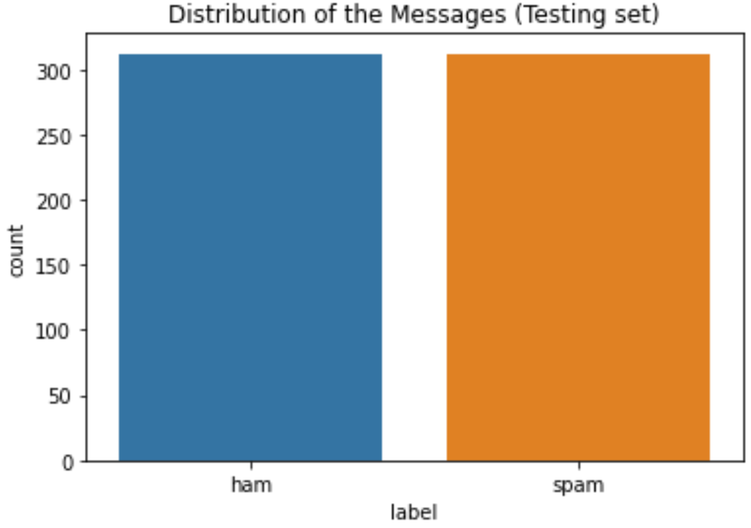
Undersampling





Tokenization

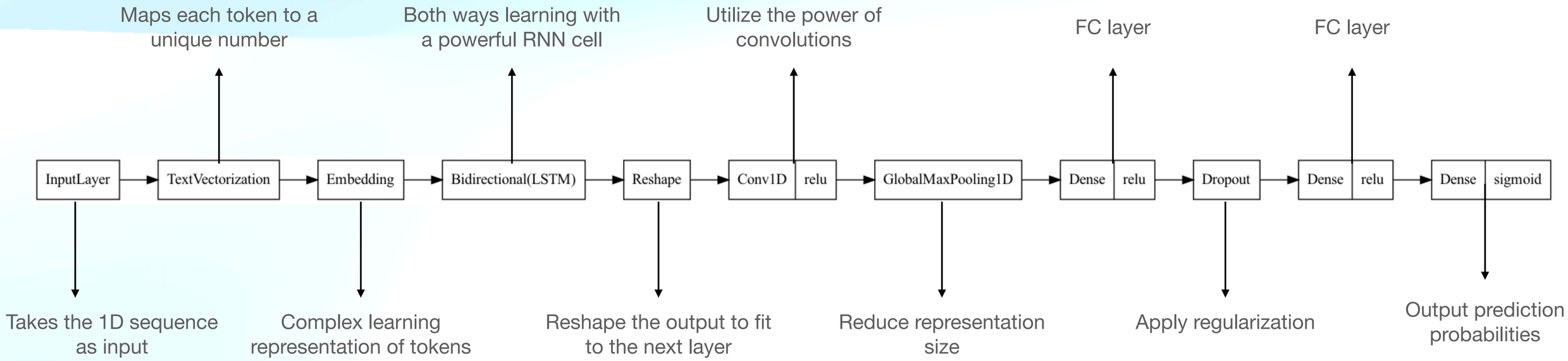
```python
max_tokens = 10000
output_sequence_length = avg_token_length

# create a text vectorization layer
text_vectorizer = TextVectorization(max_tokens=max_tokens,
                                    standardize="lower_and_strip_punctuation",
                                    split="whitespace",
                                    ngrams=None,
                                    output_mode="int",
                                    output_sequence_length=output_sequence_length)
```



Model Development

The Optimal Model

Model	Accuracy	F1-Score	Custom Data
Model 0	0.9599	0.9599	0.7500
Model 1	0.9599	0.9598	0.5000
Model 2	0.9775	0.9775	0.7500
Model 3	0.9855	0.9855	0.7500
Model 4	0.9727	0.9727	0.7500
Model 5	0.9839	0.9839	0.5000
Model 6	0.9743	0.9743	1.000



Model Deployment

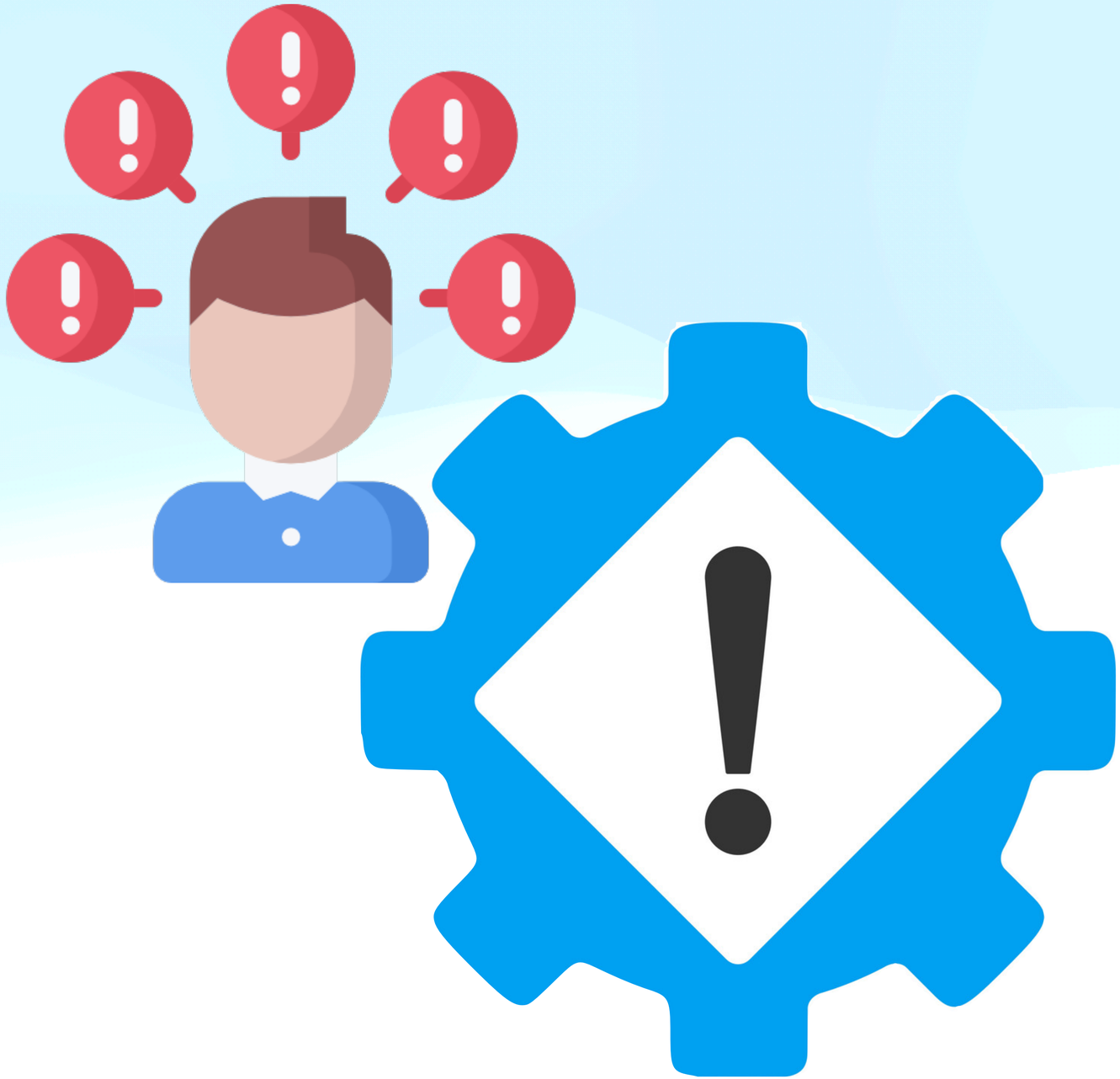
Telegram Bot

- Built a Telegram Bot through the power of `pyrogram` package
- Loaded the model in the Bot script to implement it.
- The script gets the messages from the chat, passes it through the model, and determines the action based on the predictions.
- DEMO: ->>>>>>>>>>>>>>>>>>

Problems with Current Solution

Current Problems

- The training data used doesn’t really represent the status of messaging in today’s life
- Must have large data for both classes to avoid overfitting
- Model might not always make accurate predictions



Conclusion

References

- https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
- https://www.kaggle.com/datasets/team-ai/spam-text-message-classification
- https://github.com/DeshDSingh/SMS-SPAM-Detection/blob/master/sms_spam.csv
- https://huggingface.co/datasets/sms_spam
- https://docs.pyrogram.org/
- https://www.tensorflow.org/tutorials/keras/save_and_load
- https://www.tensorflow.org/api_docs/python/tf
- https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2
- https://www.geeksforgeeks.org/how-to-handle-imbalanced-classes-in-machine-learning/

Thank You

All questions and inquires are welcomed