Data Glacier

Your Deep Learning Partner

# Twitter Hate Speech Detection using Transformers:
## Exploratory Data Analysis

**Group name:** VerbalVigilantes
**Name:** Zohra Bouchamaoui
**Email:** Zohra.Bouchamaoui@outlook.com

**Country:** United Kingdom
**Company:** Data Glacier

**Batch Code:** LISUM21
**Specialisation:** NLP

# Agenda

Problem Description

Data Description

Data Cleaning and Transformation

Exploratory Data Analysis (EDA)

Key Insights and Findings

Recommendations

Recommended Models

Data Glacier

Your Deep Learning Partner

# Problem Description

**Introduction:**

- Hate speech is defined as any form of derogatory or discriminatory communication targeting individuals or groups base on attributes.
- Objectives: Design a machine learning model to accurately detect hate speech.

**Approach:**

- Hate speech detection as a form of sentiment classification.
- Utilising a commonly used sentiment classification dataset for training.
- Twitter data is used for hate speech detection in this project.

# Data Description

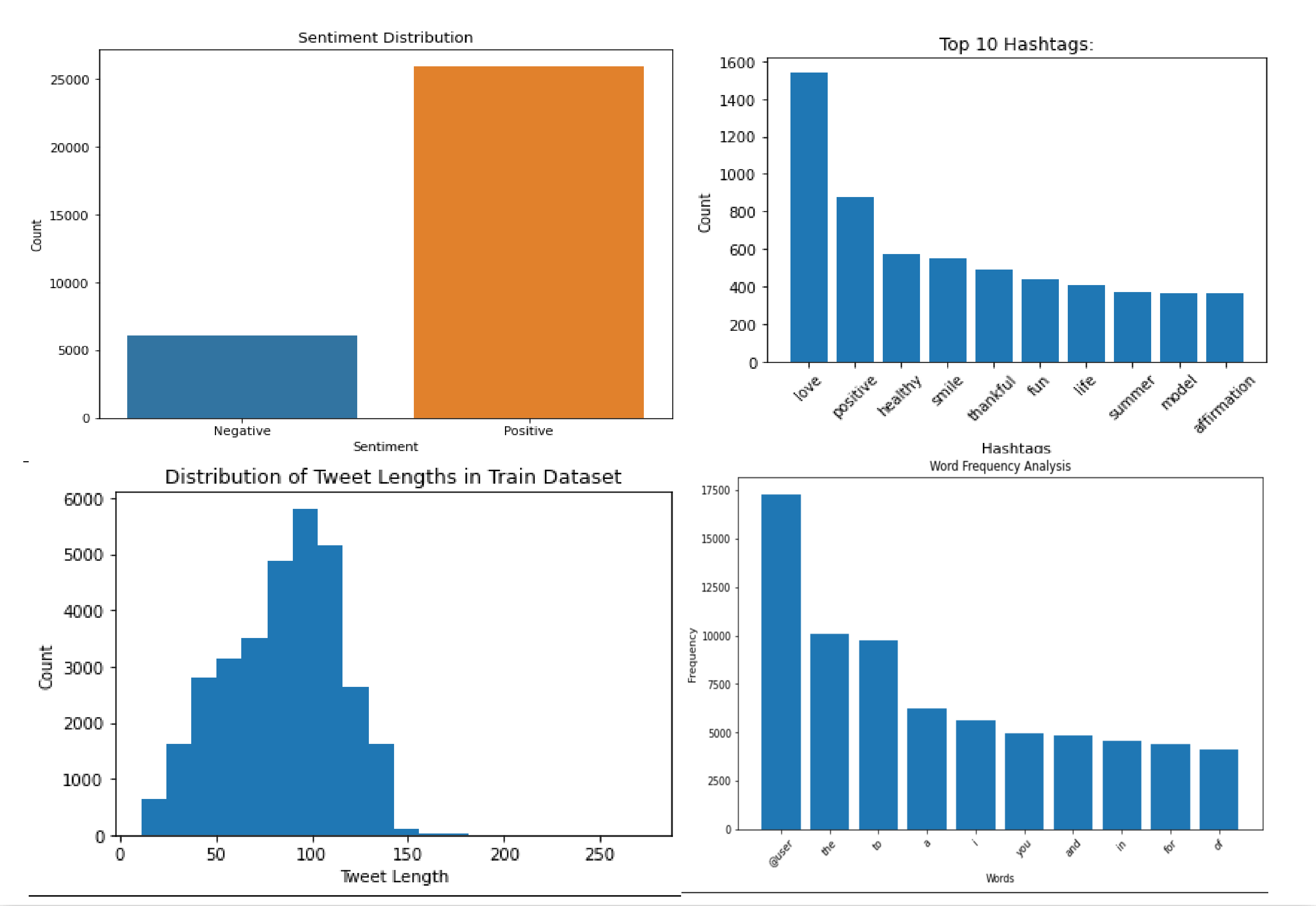
- Twitter dataset collected by Rahul Agarwal on Kaggle[1] , containing labeled tweets for hate speech detection.
- Datasets consists of two files train and test.
- Train dataset consists of tweets with labels indicating hate speech or non-hate speech (using 0 as non-hate speech and 1 as hate speech).
- Both train and test sets have unique identifiers (id) for each tweet and have a tweet column containing the text content of the tweets.

[1] https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv

# Data Cleaning and Transformation

- Data preprocessing techniques applied to improve data quality
- Highlights of the cleaning process:
  - Removal of URLs and special characters
  - Lowercasing the text
  - Handling missing values, if any

# Exploratory Data Analysis (EDA)



# Key Insights and Findings

- Dataset exhibits class imbalance with a larger number of non-hate speech tweets.
- Shorter tweets tend to have more negative sentiment.
- Common words include, *‘@user’* (freq. 17,291), *‘the’* (10,065), *‘to’* (9,768), *‘a’* (6,261), *‘i’* (5,655).
- Top hashtags include, *‘love’, ‘positive’, ‘healthy’, ‘smile’, ‘thankful’, ‘fun’, ‘life’, ‘summer’, ‘model’, ‘affirmation’*.
- Topic modelling showed 5 major topics within the tweets:

  - **Topic 1:** day, happy, new, love, father

  - **Topic 2:** love, life, time, bull, smile

  - **Topic 3:** like, amp, need, people, don

  - **Topic 4:** user, thankful, positive, amp, just

  - **Topic 5:** user, good, amp, gt, music

# Recommendations

- Enhancing hate speech detection algorithms to address class imbalance.
- Developing strategies to monitor and moderate specific hashtags associated with hate speech.
- Implementing user reporting mechanisms to identify and address hate speech instances.

# Recommended Models

**Model 1:** *Support Vector Machines (SVM) with TF-IDF vectorization*

- Preprocessing techniques: Stopword removal, TF-IDF vectorization
- Model evaluation metrics: Accuracy, precision, recall, F1-score

**Model 2:** *Recurrent Neural Network (RNN) with word embeddings*

- Preprocessing techniques: Tokenisation, word embeddings
- Model evaluation metrics: Accuracy, precision, recall, F1-score

**Model 3:** *XGBoost with bag-of-words (BoW) representation*

- Preprocessing techniques: Bag-of-words (BoW) vectorization
- Model evaluation metrics: Accuracy, precision, recall, F1-score

# Thank You

Data Glacier

Your Deep Learning Partner