PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

SAAD DAHLEB BLIDA 01 UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE



MASTER’S INTELLIGENT SYSTEMS ENGINEERING

**NATURAL LANGUAGE PROCESSING**

REPORT

|  |  |
| --- | --- |
| Made by : |  |
|  | NAHLA YASMINE MIHOUBI  ABDELATIF MEKRI |
|  |  |

TWEET SIMILARITY ANALYSIS

WITH

TRANSFORMER EMBEDDINGS

Academic year : 2023-2024

Contents

[1. Introduction 3](#_Toc154398760)

[2. Objectives 3](#_Toc154398761)

[3. Project Structure 3](#_Toc154398762)

[4. Features 3](#_Toc154398763)

[5. Usage 4](#_Toc154398764)

[6. How It Works 4](#_Toc154398765)

[7. Usage Example 5](#_Toc154398766)

[8. Future Considerations 6](#_Toc154398767)

[9. Conclusion 6](#_Toc154398768)

# 1. Introduction

The objective of this project is to develop a robust model that evaluates the semantic similarity of two tweets and provides a similarity score, indicating the origine of the tweet beign the same user. Using techniques of natural language processing (NLP), particularly utilizing transformer architectures,

By employing a combination of data preparation techniques, data preprocessing methodologies, and a sophisticated model architecture utilizing transformer embeddings, we seek to create a reliable model capable of distinguishing between tweets authored by the same user and those originating from different users. This project delves into tweet analysis, with a focus on text representation, distance calculations, and evaluation metrics.

# 2. Objectives

- Model Development: Construct a model capable of analyzing the semantic similarity between pairs of tweets.

- Semantic Representation: Utilize transformer-based architectures for text representation, enabling the model to capture intricate semantic nuances present in tweet content.

- Similarity Score Generation: Develop a mechanism to generate a similarity score in this case Manhatan distance for each pair of tweets, indicating the likelihood that they originated from the same user.

- Data Preparation: Employ data sampling techniques, to ensure a balanced representation of tweet pairs from the same user and different users in the training dataset.

- Labeling Scheme: Assign appropriate similarity labels to tweet pairs, categorizing them as same-user pairs (label 1) or different-user pairs (label 0).

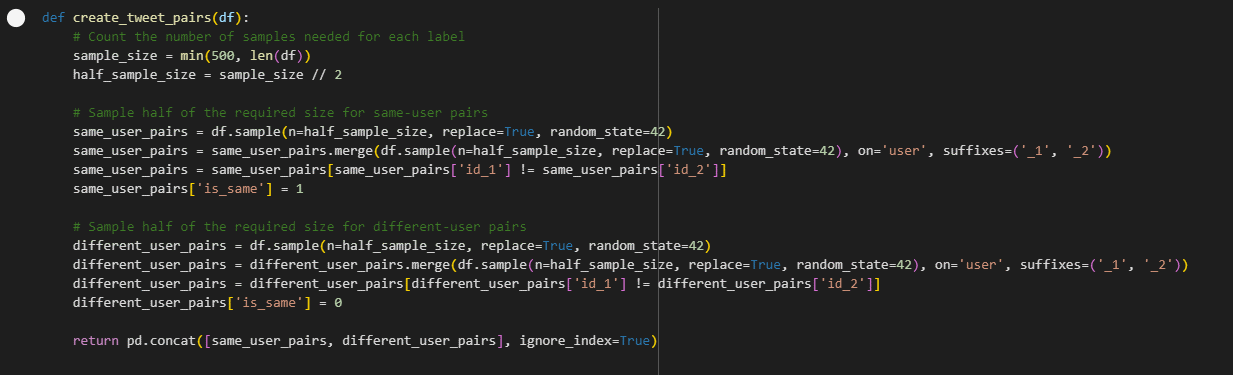
- Data Preprocessing: Implement text cleaning procedures, including lowercase conversion, punctuation removal, and tokenization, to prepare tweet content for model input.

- Evaluation Metrics: Assess model performance using standard evaluation metrics such as precision, recall, and F1 score, providing insights into the model's ability to accurately identify tweet similarity.

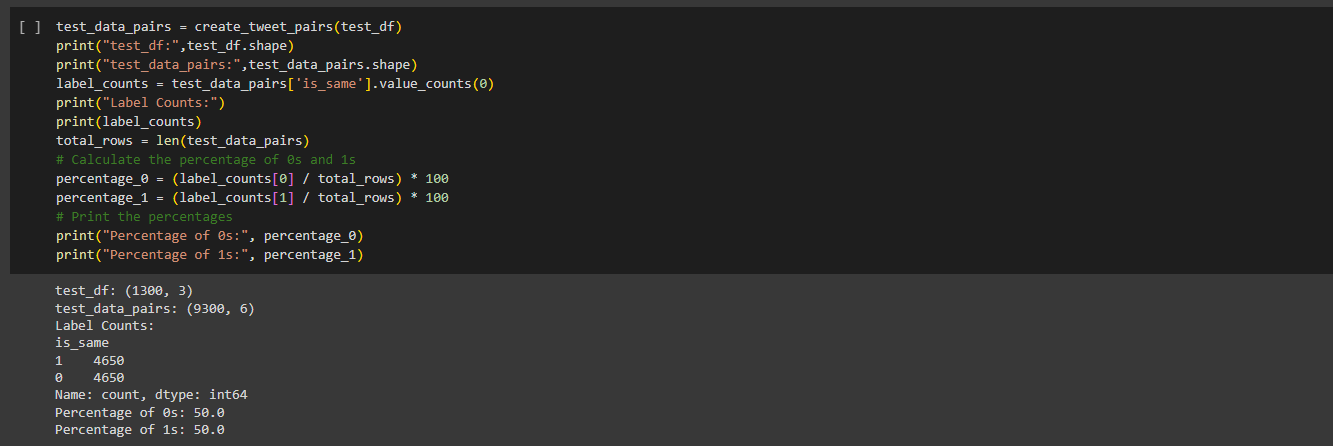
# 3. Implementation

## 3.1 The Data

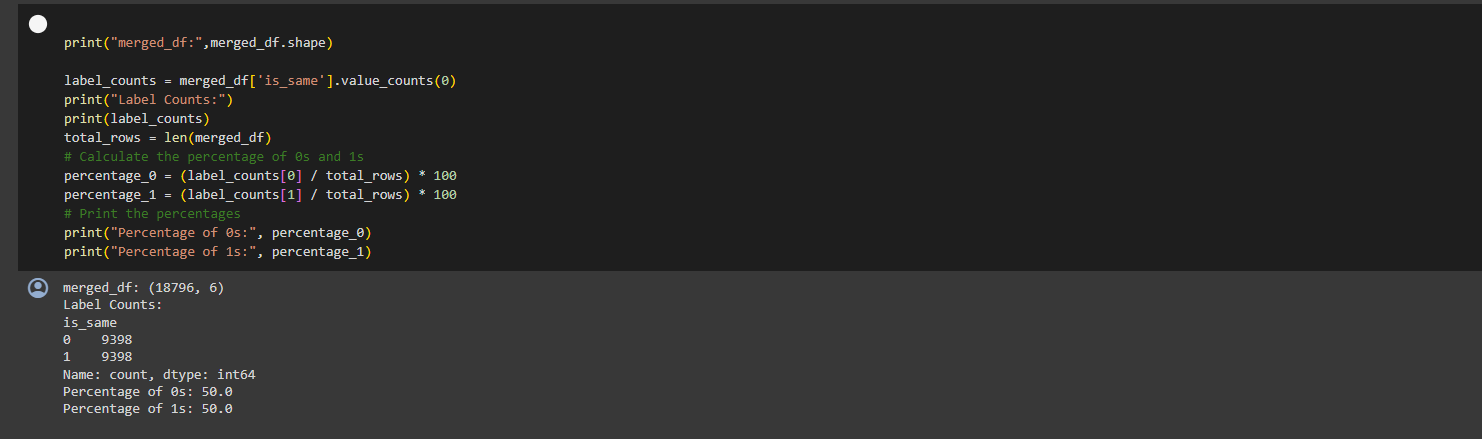
3.1.1 Creation of tweet pairs



3.1.2 Calculation equal label ‘is\_same’ values :



3.1.3 Showing the results of the merged data

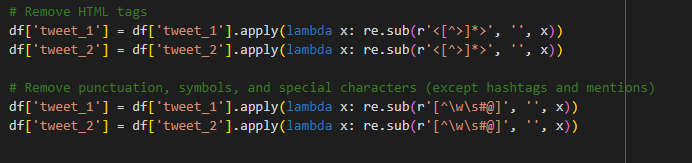


## 3.2 Data Preprocessing

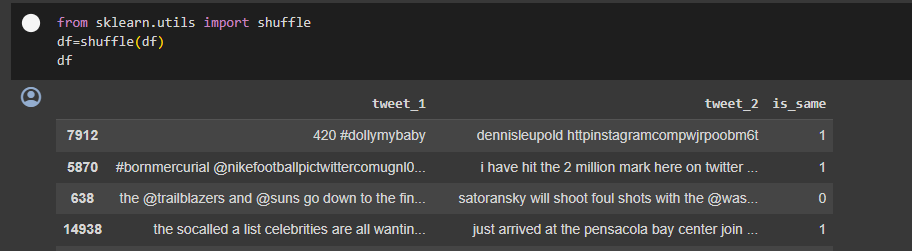
3.2.1 converting all text to lowercase



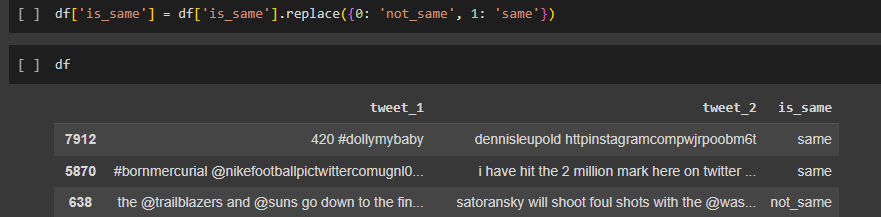
3.2.2 stripping the data



3.2.3 reshuffling the merged file after cleaning

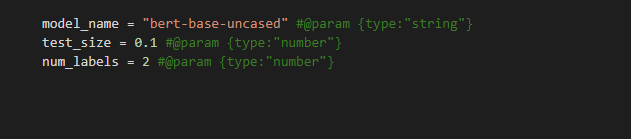


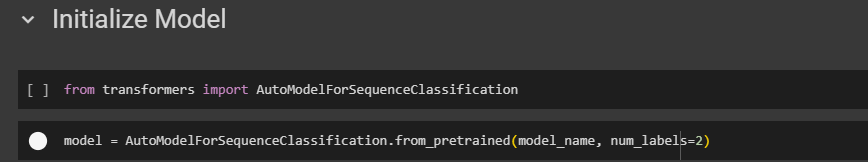
3.2.4 Attributing lables instead of boolean values for ‘is\_same’



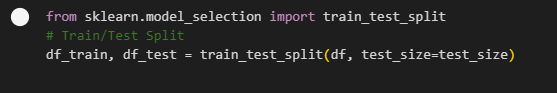
## 3.3 Configuration

3.3.1 Initialisation

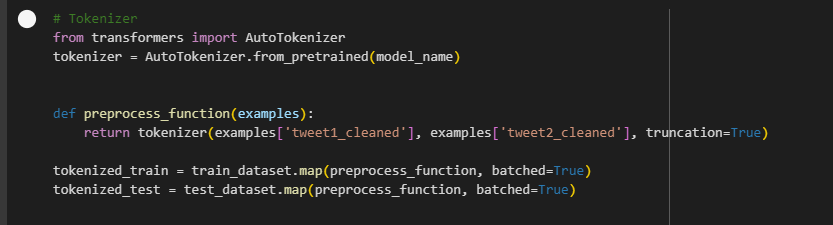




3.3.2 Data splitting



3.3.3 Tokenization



## 3.4 Results :

The following results reprsent the training results of the model with 3 epoch , and a min sample size of at least 500 