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 : |  |
|  | Benamara Ichrak  Messar Aya |
|  |  |

TWEET SIMILARITY ANALYSIS

WITH

TRANSFORMER EMBEDDINGS

year : 2023-2024

# 1. Introduction

# Developing a model that assesses the semantic similarity between two tweets and provides a similarity score between 0 and 1, indicating the probability that they were authored by the same user. Leveraging transformer-based architectures for text representation and distance computation, our model aims to capture nuanced semantic relationships, enabling accurate determination of tweet authorship.

# 2. Objectives

To build a model that analyzes the semantic similarity of two tweets and

provides a similarity score (between 0 and 1) indicating the likelihood that they came

from the same user. This model will utilize transformers for text representation and

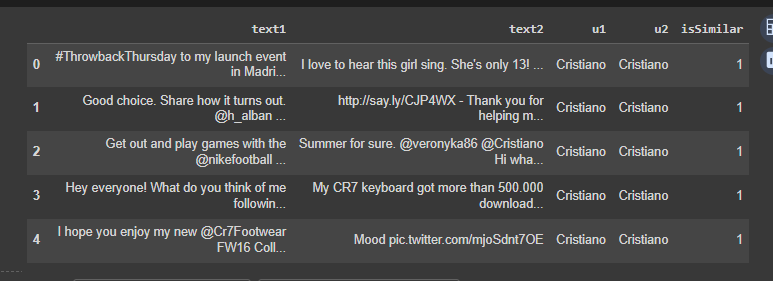
distance calculations.

# 3. Implementation

## 3.1 The Data

3.1.1 Creation of tweet pairs :

This code defines a function pairs(dataset) to generate pairs of tweets from a given dataset. It iterates over unique authors, randomly selects tweet pairs from the same author and different authors while ensuring each tweet is used only once. The resulting pairs are stored in a DataFrame with columns for the two tweets, their respective authors, and a label indicating whether they are similar (1) or not (0)

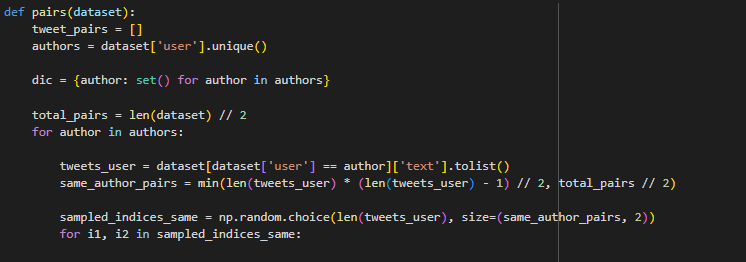


## 3.2 :

## This code defines a function pairs(dataset) to create pairs of tweets from a given dataset. It iterates over unique authors, randomly selects pairs of tweets from the same author and different authors, ensuring each tweet is used only once. The resulting pairs are stored in a DataFrame with columns for the two tweets, their respective authors, and a label indicating their similarity.

## 

**3.3.3 Tokenization :**



## 3.4 Results :

## 