Unbalanced Datasets in Text Classification: Impact and Solutions

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—**

***Index Terms*—sentiment Analysis, Roots, GloVe, word vector representations , Machine learning**

I. Introduction

One of the known issues in natural language processing is assigning text to classes dubbed ”Text Classification.” It’s one of the most active research areas, especially with the advent of social media platforms, where a massive quantity of texts in many domains are generated in a matter of seconds. We concentrated on sentiment analysis, which is a part of Text Classification, in this paper.

The goal of sentiment analysis is to forecast and categorize feelings expressed in comments or tweets into three categories: positive, negative, and neutral. In the literature, many strategies have been used to address this issue. Deep learning and related algorithms are becoming increasingly popular. The act of preprocessing is crucial to the creation of models. To feed machine learning algorithms, words and texts must be vectorized.

An appropriate dataset is required for creating performant deep learning models. In sentiment analysis, there is a major exception: the neutral class dominates the comments and tweets, resulting in datasets that are practically unbalanced. Different approaches were employed in recent studies to overcome this drawback.

Approaches such as oversampling, undersampling, and synthetic balance have been used. Everyone has advantages and disadvantages. We compare these strategies in text categorization, particularly in sentiment analysis, where we apply a prominent preprocessing methodology and the deep learning BERT algorithm, in our study. The F1-score and recall measures are used to evaluate our system.

The remainder of the paper is structured as follows: The second section of the paper goes over the existing literature. Section 3 discusses the strategies for balancing datasets and the deep learning algorithm used to extract polarity from tweets. We look about the proposed approach and methodology for extracting polarity from tweets in Section 4. Finally, we offer the results of the system’s evaluation on an unbalanced dataset, as well as some concluding observations and perspectives.

II. Related Work

A. Results and Discussion:

III. Conclusion