uc3m | Universidad Carlos III de Madrid

Master Degree in Big Data Analytics

Academic Year 2020-2021

Master Thesis

"Sentiment Analysis Classification in Covid-19 related tweets using Natural Language Processing with Deep Learning Techniques"

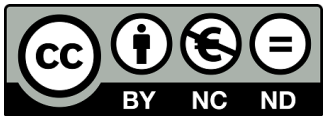
Jose Antonio Jijon Vorbeck

Advisor: Isabel Segura Bedmar

Madrid, April 2021

AVOID PLAGIARISM

The University uses the **Turnitin Feedback Studio** program within the Aula Global for the delivery of student work. This program compares the originality of the work delivered by each student with millions of electronic resources and detects those parts of the text that are copied and pasted. Plagiarizing in a TFM is considered a **Serious Misconduct**, and may result in permanent expulsion from the University.



SUMMARY

Due to the globalization of the COVID-19 pandemic, and the expansion of social media influence as the main source of information for many people, there has been a great amount of different reactions surrounding the topic. The WHO has announced in December 2020 that they are currently fighting an "infodemic" in the same way as they are fighting a pandemic'. An "infodemic" relates to the spread of information that is not controlled nor filtered, and can have a negative impact in the society. If not managed properly, an aggressive or negative tweet can be very harmful and misleading for the society. Therefore, the World Health Organization has called for action and asked the academic and scientific community to "commit to finding solutions and tools, ..., to manage the infodemic embedding the use of digital technologies and data science". The goal of this Thesis will be to develop and apply Natural Language Processing models using Deep Learning to classify a collection of tweets that refer to the Covid-19 pandemic. Several simpler and widely used models will be applied first and serve as a benchmark for methods based on Long Short Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT). Models and prediction scores will be presented. The results and conclusions of the different models and possible further development opportunities will also be discussed.

Keywords: Sentiment Analysis, Natural Language Processing, Machine Learning, Deep Learning, LSTM, BERT, Covid-19, Long Short Term Memory, Classification

DEDICATION

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Motivation

1.2. Objectives

1.3. Structure of the Document

2. BACKGROUD

2.1. Machine Learning

2.2. Neural Networks and Deep Learning

2.3. Natural Language Processing

3. STATE OF THE ART

3.1. Sentiment Analysis with NLP

3.2. Classical Machine Learning Approaches

3.3. Deep Learning Approaches

4. SENTIMENT ANALYSIS

4.1. Covid-19 Tweets Dataset

4.2. Data Preprocessing

4.3. Stemming and Lemmatization

4.4. Classical methods

4.5. Long Short Term Memory

4.6. BERT

4.7. Something else??

5. EVALUATION AND DISCUSSION

5.1. Multi Classification Performance Metrics

5.2. Methods evaluation

5.3. Methods discussion

5.4. Error Analysis

6. CONCLUSION AND FUTURE WORK

6.1. Conclusion

6.2. Future Work