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"

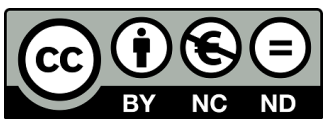
Jose Antonio Jijon Vorbeck

Advisor: Isabel Segura Bedmar

Madrid, June 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.



[Include this code in case you want your Master Thesis published in Open Access University Repository]

SUMMARY

Due to the globalisation 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 Organisation 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

- Impact of Covid in Social Networks
- Why is it important to make Sentiment Analysis
- How is SA related to tweets (how to use it)
- Link to BigData Analytics and NLP and DL

1.2. Objectives

- Brief presentation of the Data set
- History about the data set, who, where, why
- Set Main objectives for the research

1.3. Structure

- Describe the structure of the Report

2. BACKGROUND

2.1. Natural Language Processing

- introduction to NLP and main aspects
- Latest developments and usages

2.2. Deep Learning

- What is Deep learning
- Neural Networks
- How do they work, main aspects and short intro

3. STATE OF THE ART

3.1. Previous Work on the Data Set

- Talk about the Kaggle competition
- What other methods and techniques have been used
- Most popular methods and results

3.2. Classical Machine Learning Approaches

- Methods in ML used
- Best methods for large data sets

3.3. Deep Learning Approaches

- DL methods used
- methodologies and results

3.4. Personal Contribution

- Personal contribution
- What have I done differently
- Comparison between different models

4. METHODS

4.1. Machine Learning Models

- Support Vector Machine
- Logistic Regression
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Random Forest

4.2. Deep Learning Models

- Multi layer Perceptron
- Long Short Term Memory
- Bidirectional Encoder Representations for Transformers

5. EVALUATION

5.1. Data Set

- Visual Representation of the Data set
- Examples of the different classes
- Description of the classes
- Prepossessing implementation

5.2. Metrics

- Multi-classification problems
- What metrics are used
- Short description of every metric (formulae)

5.3. Results and Discussion

- Results of Classical Approaches
- Results of the different Deep Learning approaches
- Comparison between the different results
- Advantages and disadvantages from each kind

6. CONCLUSION AND FUTURE WORK

- Repetition of the main problem
- Have we helped solving the issue ??
- What is the conclusion from the work
- Best methods and why
- What future work could be made to further increase the research