Machine Learning Model



Prepared By:

Ibrahim Munther

The National ICT Up skilling program –Data Science section

Al-Hussain Technical University

Prepared for:

Mr. Mohammad AbuAlhoul

Mr. Ibrahim AbuAlhoal

Mrs. Rawan Alkurd

February 16, 2021

# Abstract

This project addresses the problem of sentiment analysis on Twitter. The goal of this project was to predict sentiment for the given Twitter post using Python. Sentiment analysis can predict many different emotions attached to the text, but in this report, only 2 major were considered: positive and negative. The training dataset was small (just over 45000 examples) and the data within it was highly skewed, which greatly impacted on the difficulty of building a good classifier. After creating a lot of custom features, utilizing bag-of-words representations and applying the logistic regression algorithm, the classification accuracy at the level of 75% was achieved.

Table of Contents

[Abstract 2](#_Toc6176378)

[1 Introduction 5](#_Toc6176379)

[2 Background 6](#_Toc6176380)

[3 Methodology 7](#_Toc6176381)

[4 Results and discussion 8](#_Toc6176382)

[5 Conclusions and recommendations 8](#_Toc6176383)

[6 References 9](#_Toc6176384)

# Introduction

So, from this project, I receive an opportunity to work in sentiment analysis field. Also, it will definitely be beneficial for my startup. Because while dealing with the reviews of customers, we want to interpret what user tends to portray so that we can give him best recommended results.

Apart from this, Sentiment analysis has been an interesting field of study. This is still an evolving subject, it has functions that are too complicated to understand by the machines such as sarcasm, negative emotions, hyperbole etc.

Because I am part of the industry, I know the potential in sentiment analysis. It adds a lot of value to the industry. Sentiment analysis bases its results on factors that are so inherently humane, it is bound to become one the major drivers of many business decisions in future.

**Statement of the Problem**

The company collected this dataset to provide Arabic sentiment corpus for the research the company doing to investigate deep learning approaches for Arabic sentiment analysis.

**Significance of the project**

The biggest benefit to this Data science application is to interpret what user tends to portray so that we can give him best recommended results.

# Background

This dataset we collected in April 2019 by an company. It contains 58K Arabic tweets annotated in positive and negative labels, The company collected this dataset to provide Arabic sentiment corpus for the research the comapny doing to investigate deep learning approaches for Arabic sentiment analysis. The dataset is balanced and collected using positive and negative emojis lexicon.

The ability to understand the public sentiment in social media is increasingly considered as an important tool for market understanding and customer segmentation.

# Methodology

Our Arabic Tweets Dataset divide the Tweets into two categories Positive or negative

the very first thing you should do is to identify which behavour the tweet belong to.

So, based on the text contents it tells us from the tweets text if it’s positive or negative, if the emojies and the text have positive vibes then its classified as positive, otherwise its negative.

And based on that the company will divide them as positive or negative as we said earlier based on the text and its emotions embedded with it

# Results and discussion

In my opinion, the results show that by offering tweets it be classified in negative or positive and it will tell us how is the people emotions in those tweets based on the text it’s self and it’s emotions and emojies.

# 5 Conclusions and recommendations

The increase of microblogging sites like Twitter offers an unparalleled opening to form and employ approaches & technologies that search and mine for sentiments. The work presented in this paper specifies an approach for sentiment analysis on Arabic Twitter data. To unseal the sentiment, we extracted the relevant data from the tweets, added the features.

The overall tweet sentiment was then calculated using a model that presented in this report. This work is exploratory in nature and the prototype evaluated is a preliminary prototype.

The models showed that prediction of text sentiment is a non-trivial task for machine learning. A lot of preprocessing is needed just to be able to run an algorithm. The main problem for sentiment analysis is to craft the machine representation of the text. Alot of additional features were created basing on common sense (length of the words, number of the characters, number of sentences etc). I think that a slight improvement in classification accuracy for the given training dataset could be developed, but since it included highly skewed data (small number of negative cases), the difference will be probably in the order of a few percents. The thing that could possibly enhance classification outcomes will be to add a lot of additional examples (increase training dataset), because given 45275 examples clearly do not include all sequence of words used, further - a lot of emotion-expressing information certainly is missing.

# References

My portfolio: <https://github.com/ibrahimmun96>

# 