**Birkbeck College University of London**
**Department of Computer Science and Information Systems**

# PROJECT PROPOSAL

# COMPARATIVE ANALYSIS OF TEXT SUPPORT METRICS

Author: **Simona Stavarache**
ID: **13174625**
Supervisor: **Stelios Sotiriadis**

MAY 2020

*A proposal submitted in partial fulfilment of the requirements for the MSc in Data Analytics (Advanced Computing Technologies)*

# ACADEMIC DECLARATION

*"This report is substantially the result of my own work except where explicitly indicated in the text. I give my permission for it to be submitted to the TURNITIN Plagiarism Detection Service. I have read and understood the sections on plagiarism in the Programme Handbook and the College website.*

*The report may be freely copied and distributed provided the source is explicitly acknowledged."*

## TABLE OF CONTENTS

1. Abstract ..... 4
2. Introduction ..... 5
    1 Problem & Motivation ..... 5
    2 Twitter ..... 5
    3 Natural Language Processing ..... 6
3. Literature Review ..... 6
    1 Use of NLP with Twitter ..... 6
    1 Sentiment Analysis ..... 7
    1 Use of similarity metrics ..... 8
4. Proposed Solution ..... 8
5. Aim and Objectives ..... 9
6. Project Plan ..... 10
7. References ..... 11

## 1. ABSTRACT

With the rise of social media platforms like Twitter, Facebook, Instagram, YouTube and many more, sharing ideas, beliefs and feelings between people has become easier, leading to the apparition of a propitious environment for behavioral research and text analysis. Moreover, this quick development of media industry has revolutionized the types and volume of information available to study, as well as the Natural Language techniques involved in analysis. According to previously developed research, the traditional NLP methods are no longer effective since posts tend to be smaller or restricted in size, which adds vagueness in the dataset. Additionally, grammatical errors, emoticons, language variations and unnecessary use of capital letters from social media posts can add to the ambiguity, thus adding some challenges.

In this research paper we will focus on Twitter data and NLP techniques to find the degree of support between two blocks of text. We will discuss and compare various similarity metrics and decide which one works best for our dataset, use sentiment analysis, clusterization and other machine learning techniques to find tweets that match semantically another text. The research will be conducted using tweets from various political and news channels from US and UK and a domain specific corpus will be composed using Streaming API from Twitter.

## KEYWORDS

NLP, similarity metrics, Twitter, sentiment analysis, keywords extraction, ML

# 2. INTRODUCTION

## 2.1. PROBLEM & MOTIVATION

Our main research question is the following: How can we define and measure the degree of approval/support between 2 digitized texts ?

The ideal characteristics of a medium in which we wish to study this problem would be: large collection of data, not overly complicated underlying message, access should be easy to obtain and a programmatic way of retrieving should exist. Therefore, Twitter seems like a valid study environment, especially because of the available retrieving API and the format of posts.

The better understanding of the nature of relationship between such texts, especially in the social media context can lead to improvement in multiple areas such as digital marketing, politic profiling or even detecting possible security threats. Nevertheless, our goal is not to provide an app for these specific uses, but a scientific tool to be used during their implementation.

## 2.2. TWITTER

Twitter was developed in 2006 in North America to be a microblogging and social networking service which gained worldwide recognition, managing to attract a pool of 100 million users by 2012 and having more than 340 million tweets per day[1]. Posts on Twitter are called tweets and they are constrained to 280 characters since 2017 to non-Asian languages, whereas before there were accepted only 140 characters. As a result of these limitations, tweets must be concise and the message might sometimes be vague, yet of a reasonably simple nature -- the problem of tweet interpretability stems rather from the format and syntax than the depth of the message.

With Twitter becoming extremely popular, at 330 million active twitter users and 500 million tweets posted per day in 2019[2], the interes in its data as a continuous stream has increased rapidly and has become widely used by analysts as dataset. Among other qualities, Twitter has the uniqueness of having real-life conversations in various subjects as news, sports, politics, fashion and many more. Twitter became not only a social media platform but it is also a place where politicians address to their people in a direct conversation.

Having this huge pool of unaltered data, researchers went beyond and applied natural language processing and machine learning techniques to have a deeper understanding of the humanity’s way of thinking and demographic trends.

### 2.3. NATURAL LANGUAGE PROCESSING

Gathering information about the public's opinion, political views, preferences and many more from social media has captured the interest of the scientific community as new challenges emerged with the unstructured data, the limited number of characters, internet slang, emoticons, abbreviations, misspellings and typos that leads to a large vocabulary. However, a big interest comes from the business world, as social media analytics offers a remarkable advantage in market forecasts.[4] To tackle this type of analysis the following terms are introduced: natural language processing (NLP) and sentiment analysis.

NLP combines Linguistics and Computer Science as a tool that evolved in an unprecedented way in the last twenty years.[5] As humans, we understand each other by making semantic and syntactic analogies, extracting keywords to synthetize. In order to imitate this human logic ability, NLP based methods use *sentence segmentation* to break the text into more sentences, *word tokenization* as preprocessing operation used to divide into words, then performing syntactic order to identify each word in a sentence, technique also known as *Part-of-Speech tagging or POS*. The previously described steps are strictly required for the construction of categories of words with different syntactic forms but the same meaning -- text *lemmatization*. Furthermore, *stopping words* can be identified and removed as they do not add any additional insights to the analysis process. Overall, some other steps might be present in various approaches, but generally speaking after the operations described above the result is fed to a machine learning model, predicting relationships between words.

However, since our end goal is to be able to determine acceptance/support relationships between a written paragraph and tweets we must deploy sentiment analysis techniques in order to have a better understanding of the texts as a whole.

## 3. LITERATURE REVIEW

### 3.1. USE OF NLP WITH TWITTER

Given the popularity ascendence of Twitter, various researchers grasped the opportunity of having an unlimited source of data to develop their theories and analytics over large populations. Even though Twitter offers a free API to extract data, it is limited to 2400 tweets per day. This lead to the development of platforms such as Twitter Vigilance (by DISIT Lab University of Florence) presented and experimented with by Daniele Cenni et. all in [6], which offers researchers *"a cross-domain, multi-user tool for collecting and analyzing*

*Twitter data"*. The possibility of retrieving only a limited number of daily tweets, further broken down into smaller parts during the day which adds in difficulty when following a big event and the fact that retweets are also considered tweets are concerns raised in [6].

Another challenge that Twitter brought is related to tweets' limited size. Due to Twitter's restrictions related to the length of a tweet it was thought by previous researchers that traditional NLP techniques cannot face the ambiguity that results from the limitations. Additionally, *"A tweet might not always be grammatically correct"*[7] which increases the vagueness of the message, thus normal POS tagging might not be sufficient for twitter analysis. A proposed solution by Weerasooriya[7] consists in adding two additional rule-based parsers to remove noise and include any missed keywords by the Stanford CoreNLP tagger. As result, the distinguishing accuracy increased from 50% to 83% during the Turing test -- examines a machine's ability to be *"linguistically indistinguishable from humans"*[8] -- which shows remarkable improvement.

Another study done by Di Giovanni et. all [9] on political inclinations of Twitter users revealed that *"people belonging to the same political party write in similar ways"*. Interestingly to say is that after testing and comparing different vectorization methods (count vectorizer -- CV --, hashing vectorizer -- HV --, term frequency vectorizer -- TF -- and term frequency - inverse document frequency vectorizer -- TF-IDF --) with different classification methods (multinomial logistic regression, K-neighbors classifier, decision tree, random forest, support vector classifier and multilayer perceptron classifier) the results showed increased accuracy (0.89) when using only nouns, vectorized with TF-IDF with multinomial logistic regression or multiplayer perceptron classifier. These results show that politicians from different parties use preponderantly different nouns to express themselves.

## 3.2. SENTIMENT ANALYSIS

In his book, *"Sentiment Analysis: Mining Opinions, Sentiments, and Emotions"*, Liu [10] has not only covered the basics of sentiment analysis -- sentiment classification, opinion search, emotion identification --, but also describes various sentiment related applications such as opinion detection, debate analysis and others. He reasons about its multiple applications in social media, as more and more people seek advice on social platforms, they become a huge pool of data that has the power to give true insights in users lifes, thoughts and opinions. The potential of sentiment analysis provided strong motivations for corporations to develop research projects and practical applications.

The key idea of sentiment analysis in Twitter is to determine the polarity of text and classify tweets as positive, neutral or negative. The main issue that rises is in choosing the right classifier. Generally, techniques as Naive Bayes, SVM or Logistic Regression are used, but we will follow research done in the ensemble world. Monisha Kanakaraj in [11] undertakes a comparative research study regarding the performance of ensemble methods over the classical ones. The experiments conducted showed an increase in performance of the ensemble ones by 5%. Compared as a whole, ensemble algorithms outperformed Naive

Bayes, SVM, Baseline and MaxEnt methods while out of the ensemble pool Extremely Randomized Trees have performed better than Random Forest, Decision Trees and AdaBoost. Another research done by Ankit et. all in [12] on ensemble classifiers on Twitter, defined an ensemble classifier by joining four base methods: Naive Bayes, Random Forests, SVM and Logistic regression. Each algorithm calculates and predicts the sentiment score, than to each tweet a probability is assigned. After, each classifier receives a weight in the ensemble technique based on each accuracy and the polarity is predicted for each tweet. The experiments showed that the proposed ensemble algorithm outperformed the classical ones.

### 3.3. USE OF SIMILARITY METRICS

Besides the sentiment analysis part, our paper focuses on the different ways of measuring similarity between texts. Different approaches on the most common ones have been researched on various topics such as Chunyu Xia et. all analysis in [13] about similarities in law documents where word2vec technique -- text space to vector space -- is described and compared to other metrics such as Jaccard similarity or N-grams. Jaccard similarity is defined by the division of the size of the union over the size of the intersection divided by the size of union and is used to determine the diversity level of samples while the N-grams technique is used to predict the following item in a sequence using Markov model. The research showed better results for word2vec. In 2013, professor Mikolov said in [14] that the distant words are less related to the current word, the more distant the words are the less weight will receive. Chia-Yang Chang et. all [15] has described a weighted word2vec technique as it is considered that “the word distance in the context bear certain semantic sense which can be exploited to better train the network model”. The results when comparing weighted to classical word2vec technique were favourable to the weighted one as it scored more correct answers than the original. When comparing Cosine similarity -- cosine angle between two vectors --, Jaccard similarity and Dice similarity -- the division between twice the size of the intersection over the sum of the sets sizes -- , Sazianti Mohd Saad at [16] has concluded that the last two surpass the Cosine similarity.

## 4. PROPOSED SOLUTION

In this section we will discuss about our solution and the steps we need to take in order to achieve the desired result.

Our proposed solution will focus on Twitter data, mostly gathered from different news pages or individuals on the politics and economics theme using the StreamingAPI offered by Twitter.

Studies showed that usually when the word order changes in sentence, the meaning also changes. Because we follow not only similarity at syntax level, but also semantically our solution is to represent documents as vectors of features and compare the distance between

characteristics. We will compare a list of similarity text metrics and investigate their performance: *Jaccard Similarity, K-means, Cosine Similarity, Word2Vec with Smooth Inverse Frequency with Cosine Similarity, LSI with Cosine Similarity, LDA with Jensen Shannon distance, Word Mover Distance, Variational Autoencoder, Universal Sentence Encoder, Siamese Manhattan LSTM, Knowledge Based* measures. We are going to try these metrics with different sentence embeddings such as *Bag-of-Words (BOW), Term Frequency - Inverse Document Frequency (TF-IDF), Continuous BOW (CBOW), SkipGram, Word2Vec, Glove, fastText, Poincarre, Node2Vec*. Our goal is to compare how the similarities behave by checking their accuracy, precision and recall.

In order to present our work, we will create a webpage that allows a user to write a paragraph which will generate a tweet search. The returned tweets will have to match, in meaning, to an extent the users text. Other functionalities might include an editor for the user to write his text and drag & drop the tweet that better matches his content, the ability to share the created post on other platforms such as Facebook, Twitter, Instagram etc., but these are not the main concern of this paper.

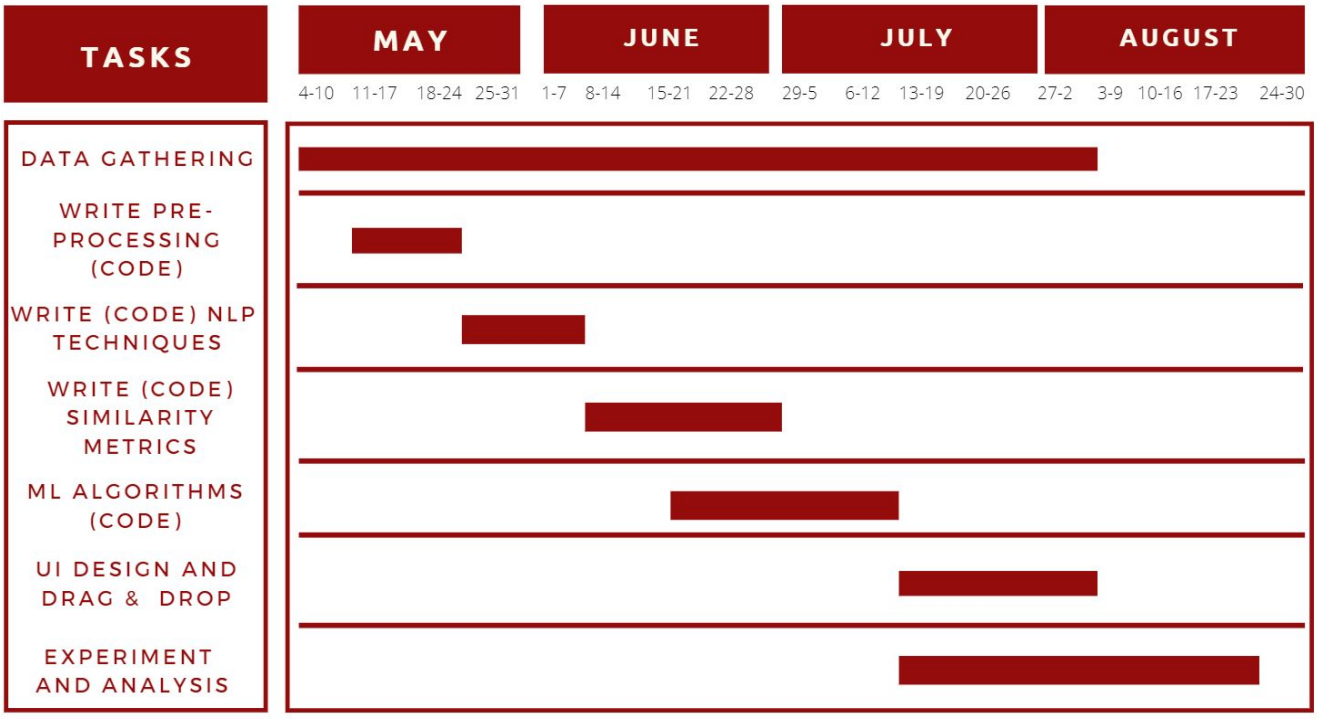
## 5. AIM AND OBJECTIVES

The aim of this research is to find a similarity measure that best fits our model in order to sustain semantically a group of text. To achieve this we will follow the list of objectives presented below.

1. Data gathering: we will create a python script that uses StreamingAPI to collect data from various news pages and individuals in english language.
2. Preprocessing data (validation, duplicate removal, conversion of letters to upper or lower, converting numbers in words, removing diacritics or accent marks)
3. Apply NLP techniques (sentence segmentation, word tokenization, Part-Of-Speech tagging, lemmatization, stop words, dependency parsing, finding noun phrases, named entity recognition, coreference resolution)
4. Try ML algorithms for predictions and sentiment analysis
5. Build model
6. UI design: editor window to build a post/tweet/article starting from the input paragraph and supported by the “references” (drag & drop tweet from list into your paragraph)

## 6. PROJECT PLAN

In this section we will present a Gantt chart containing the objectives discussed in section 5 over the next period of time until September.



## 7. REFERENCES

[1] https://en.wikipedia.org/wiki/Twitter
[2] https://www.omnicoreagency.com/twitter-statistics/
[3] Silvio Amir, Miguel Almeida, Bruno Martins, Joao Filgueiras and Mario J. Silva, "TUGAS: Exploiting Unlabelled Data for Twitter Sentiment Analysis". Available online at: https://www.aclweb.org/anthology/S14-2120.pdf
[4] Erik Cambria, Bjorn Schuller, Yunqing Xia, Catherine Havasi, "New Avenues in Opinion Mining and Sentiment Analysis", IEEE Intelligent Systems, Volume 28, 2013. Available online at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6468032
[5] I. Zitouni, Natural Language Processing of Semitic Languages, Springer, 2016.
[6] Daniele Cenni, Paolo Nesi, Gianni Pantaleo, Imad Zaza, "Twitter vigilance: A multi-user platform for cross-domain Twitter data analytics, NLP and sentiment analysis", 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation. Available online at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8397589
[7] Tharindu Weerasooriya, Nandula Perera and S.R. Liyanage, "A method to extract essential keywords from a tweet using NLP tools", 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), 2017 IEEE. Available online at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7829895
[8] K. Lacurts, "Criticisms of the Turing Test and Why You Should Ignore (Most of) Them", Official Blog of MIT's Course: Philosophy and Theoretical Computer Science, 2011. Available online at: people.csail.mit.edu/katrina/papers/6893.pdf
[9] Marco Di Giovanni, Marco Brambilla, Stefano Ceri, Florian Daniel, Giorgia Ramponi, "Content-based Classification of Political Inclinations of Twitter Users", 2018 IEEE International Conference on Big Data (Big Data). Available online at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8622040
[10] Bing Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions", Cambridge University Press, 2015. Available online at: www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
[11] Monisha Kanakaraj, Ram Mohana Reddy Guddeti, "Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques", Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing. Available online at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7050801

[12] Ankit and Nabizath Saleenam "An Ensemble Classification System for Twitter Sentiment Analysis", Procedia Computer Science, Volume 132, 2018, Pages 937-946. Available online at: www.sciencedirect.com/science/article/pii/S187705091830841X

[13] Chunyu Xia, Tieke He, Wenlong Li, Zemin Qin, Zhipeng Zou, "Similarity Analysis of Law Documents Based on Word2vec, 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C). Available online at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8859429

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", 2013

[15] Chia-Yang Chang, Shie-Jue Lee, Chih-Chin Lai, "Weighted word2vec based on the distance of words", 2017 International Conference on Machine Learning and Cybernetics (ICMLC). Available online at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8108974

[16] Sazianti Mohd Saad, Siti Sakira Kamarudin, "Comparative analysis of similarity measures for sentence level semantic measurement of text", 2013 IEEE International Conference on Control System, Computing and Engineering. Available online at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6719938