



Τεχνολογικό
Πανεπιστήμιο
Κύπρου

Σχολή Επικοινωνίας και
Σπουδών Διαδικτύου

Bachelor's thesis:

**“A COMPUTATIONAL ANALYSIS OF THE ONLINE
CYPRIOT POLITICAL DISCOURSE – INTERPRETING BIG
DATA”**

Alexios Antoniou

Limassol 2020

CYPRUS UNIVERSITY OF TECHNOLOGY
FACULTY OF COMMUNICATION AND MEDIA STUDIES
DEPARTMENT OF COMMUNICATION AND INTERNET STUDIES

Bachelor's Thesis

A COMPUTATIONAL ANALYSIS OF THE ONLINE CYPRIOT
POLITICAL DISCOURSE – INTERPRETING BIG DATA

Alexios Antoniou

Supervisor

Dr. Constantinos Djouvas

Limassol, May 2020

Copyrights

Copyright© Alexios Antoniou, 2020

All rights reserved.

The approval of the thesis by the Department of of Communication and Internet Studies of Cyprus University of Technology does not imply necessarily the approval by the Department of the views of the writer.

I would like to thank my advisor Dr. Constantinos Djouvas for his guidance and the continuous support he provided me during the development of my thesis. I would also like to thank my parents, Antonis and Maria, for believing in me and for their endless and selfless assistance. Finally, I would like to thank my friends who were very supportive and they encouraged me throughout the execution of this thesis.

ABSTRACT

Due to the development of the Internet, the online information sharing is done with a high speed. Different tasks can be performed through it, such as information retrieving and socialising by using different social media. The amount of online information is enormous and constantly increasing in a rapid manner. Of high importance is information published that is related to politics. Politicians use different social media platforms such as Twitter to express their views and opinions on different issues. Moreover, most of the news media outlets use websites to communicate their articles to the masses. The aim of this study is to create a system that follows and analyses the news in the political sphere in Cyprus in real time through the utilization of big-data techniques and approaches. To be more precise, the system will collect different politic related documents uploaded in Twitter and/or in online newspapers and analyse them in real time. The analysis of the data will be achieved using machine learning techniques. Specifically, the main technique used is the so called Latent Dirichlet Allocation (LDA). It is an unsupervised technique for clustering big amount of information to create groups of similar documents. The system proposed to be developed will be online and accessible through the internet. The main features of the system will be to allow its users to monitor the Cyprus political sphere by creating different analyses in different time windows. Finally, the users will be able to observe the sentiment for each topic in each analysis which indicates the general vibe around the topic.

Keywords: Twitter, Online Media, Social Media, LDA, Topic Modelling, Sentiment Analysis

TABLE OF CONTENTS

ABSTRACT	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	ix
LIST OF ABBRIVATIONS.....	x
1 Introduction	1
2 Problem Description – Study Necessity	4
2.1 Problem Description.....	4
2.2 Study Necessity	4
3 Theoretical Framework	7
3.1 Political Communication.....	7
3.2 Data Collections - Crawling & Twitter API.....	8
3.3 Sentiment Analysis.....	10
3.4 Machine Learning Techniques	11
3.5 LDA.....	12
4 Literature Review	14
5 Methodology	16
5.1 Data Acquisition.....	16
5.1.1 Python Crawler.....	17
5.1.2 Twitter API.....	17
5.2 Research Tools	18
5.2.1 Web technologies	18
5.3 Procedure.....	19
6 Data Analysis	21
6.1 Texts processing.....	21
6.2 LDA Analysis.....	22
6.3 Sentiment of the Topics.....	23
6.4 Text Similarity to Topics.....	24
7 System Implementation and Development.....	25
7.1 Existing Results.....	25
7.2 Your Analysis.....	29
7.3 New Analysis	33
7.3.1 Naming, number of topics and size of the corpus	33
7.3.2 Twitter and Online newspapers data collection.....	35
8 Restrictions.....	37

8.1 Problems.....	37
8.1.1 Sentiment in Greek language.....	37
8.1.2 Sentiment analysis of each document.....	37
8.1.3 High perplexity due to the Greek Language.....	37
8.2 Future work	37
8.2.1 Not possible to create new analysis using existing data	38
8.2.2 Not possible to create model using historical data	38
8.2.3 Topics' sentiment value.....	38
9 Bibliography.....	39

LIST OF FIGURES

Figure 1: Platform's Architecture.....	16
Figure 2: Previous Results page	25
Figure 3: Legend of the previous results page.....	26
Figure 4: Previous Analysis words.....	27
Figure 5: Name the model	27
Figure 6: LDA analysis results graph.....	28
Figure 7: Topic's 5 representation	29
Figure 8: Topics, words and sentiment.....	30
Figure 9: Rename the topics	31
Figure 10: Similarity results	32
Figure 11: Similarity results on chart	32
Figure 12: New analysis interface	33
Figure 13: Corpus type	34
Figure 14: Source of data check boxes.....	34
Figure 15: Online newspapers categories.....	34
Figure 16: Twitter profiles and hashtags selection.....	35

LIST OF ABBRIVATIONS

1. **API:** Application Programming Interface
2. **DB:** Database
3. **GUI:** Graphical User Interface
4. **JS:** Javascript
5. **JSON:** Javascript Object Notation
6. **LDA:** Latent Dirichlet Allocation
7. **MA:** Machine Learning
8. **MEAN:** MongoDB, Express.js, AngularJS, Node.js
9. **MVC:** Model & View & Controller
10. **OMM:** Online Media Monitor
11. **OSM:** Online Social Media
12. **OSN:** Online Social Networks
13. **SMM:** Social Media Monitor

1 Introduction

Over the last years, the use of the internet constantly increases. Through it, people can perform various activities, such as, online shopping, seeking for information, watching movies, and socialising via the online social media (OSM). Except these, internet also allowed people to embed online news reading in their daily activities, mainly attributed to the development of the mobile and internet technologies. The 92% of the people who live in USA access news in different formats, including the online one. (Yadamsuren & Erdelez, 2010). By using the internet, people have greater variety of news from which they choose what to read in comparison to the traditional media (Tewksbury, 2003). The amount of available information regarding the different issues that occur every day can be found scattered across different resources, like online news media, and online social media. This work performs an automated computational analysis on online information regarding the political discourse on different political issues. In this work, we focus on two different types of online media for collecting data, that is online newspapers, and Twitter.

The history of the major online social networks dates back in 1997 when a platform with the domain name “SixDegrees.com” went online (Boyd & Ellison, 2007). Social network is defined as a web-based application which allows its users to complete various socializing tasks. Such tasks may be creating a public or semi-public profile, connecting with other people, or viewing other users’ profiles and interacting with them (Steinfeld, Ellison & Lampe, 2008).

Since the establishment of online social media, their users’ number is exponentially expanding. The number of adults who use social networks in America increased from 7% in 2005, to 65% in just 10 years and today is 70% (Dubras Kemp & Navarro, 2020). The usage of these media affects many different aspects of everyday life, concerning mainly one’s occupation, political stance, and health issues (Perrin, 2015). According to Kemp, (2019) the number of active online social media users around the globe in January 2019 was 3,484 billion people, which accounts for little less than 50% of the world’s population. The difference between the number of internet users and the number of online social media users is less than 1 billion, which consequently means that around 80% of the internet users are also online social media users.

Since the devise of online social media (OSM), people are utilizing them to express their opinions more freely and criticise political actions and decisions. To be more precise, the OSM were used to organize and trigger activities that require effort from many people in an attempt

to create the feeling that all the people together form a community with similar identity. The purpose of this creation is to attract the interest of the international community and seek for help (Eltantawy & Wiest, 2011). A great example of the usage of the online social media to criticise and provoke disturbances in the political scene was during the “Arab Spring”. In the Arabic countries where revolutions took place (Libya, Algeria, Morocco, Syria, etc.), the online social media had a crucial role in creating political debates, motivating and mobilizing people. Many times, the discussions made online triggered many people to act and demonstrate against their governments. The online social media were used as a mean to express democratic ideas (Howard et al. 2015). Additionally, people use them for other activities. Such activities include “like” of others’ activities, advertisement or any other kind of activity of social networks. Other people utilize online social networks (OSNs) to motivate their fellow citizens to vote or to encourage them to act against a matter they do not agree with (Rainie, Smith & Schlozman, 2012). Summarily, OSNs can be used by people as a mean to trigger other people to express themselves regarding different political issues.

The ability to publish anything in online social media, freely express themselves and get involved in online debates with other citizens, can result on a great impact on politics. There is no country in the world whose politicians do not use online social media to express themselves in order to reach the masses (Castells, 2004). In fact, online social media are constantly increasing their importance in shaping the political interaction between the politicians and the citizens (Stieglitz & Dang-Xuan, 2013).

Having the aforementioned abilities handed over to citizens, online social media can be perceived as a source of rich political related information since they are used by both politicians and the people to communicate, interact and discuss political issues.

Apart from social networks though, there is another rich source of information that can be used to gain knowledge about current political affairs. This source is the online newspapers. In the case under study, i.e., Cyprus, the majority of the daily printed newspapers also have online versions in which all the articles are published. Contrary to online social media, the content of newspapers is considered more reliable and valid, mainly because it is produced by professional journalists. However, even though they can be considered as a more reliable source for identifying issues in a timely manner, newspapers cannot be used for identifying the general public opinion about these issues. Thus, we combine data from the elitist, usually politically affiliated online news medias created by journalism and the user generated content published on online social media where citizens freely express their believes.

The purpose of this research is to create a real-time observatory that will monitor and analyse online content related to political issues utilizing different big-data approaches. The procedure will start with the collection of the data from online social media, Twitter in this case, as well as from articles published on different online newspapers through a simple and intuitive web interface. Since the subject of this research is related to politics, irrelevant articles, e.g., articles about sports or weather, will not be included in the dataset. The online media allocate the articles according to their subject. Each article is assigned to different category, for example a news article related to finance is distributed to the “economy” sub-category. For this research, the sub-categories that will be used are related to politics, Cypriot matters, economics, international affairs and Cyprus problem.

The data collection will be accomplished by developing different programs responsible for finding, retrieving and storing data into an online database. Having the retrieved data stored into a database, the next step will be to apply different techniques for the analysis of the dataset. The analysis includes topic modelling and sentiment analysis for identifying the stance of citizens against the identified issues. Finally, an intuitive Graphical User Interface (GUI) will be used for the representation of the analysed data.

2 Problem Description – Study Necessity

2.1 Problem Description

As mentioned above, online media are rapidly expanding and becoming more powerful throughout the years. As a result, the amount of online information is vast and in every second passing, the size of it is getting bigger. With such vast supply of information, it is impossible for someone to collect and analyse all the online data regarding different political events by using traditional approaches.

This information is located into many different places all over the web. It cannot be spotted in a single location. This may lead to the problem of wasting time when visiting websites while trying to collect the needed data. Furthermore, if the collection of this huge amount of information can somehow be achieved, it is really difficult and time consuming, if not impossible, to be analysed.

Another problem that cannot be toggled by traditional approaches is the speed in which the data are generated, updated and covered by different online sources. The ability given to people to post freely their opinions on online social media on one hand and the opportunity given to journalists to publish articles at any time on online newspapers on the other hand, has as a result the increase of the produced data in an enormous speed. Therefore, it is extremely difficult to keep up with updated information, unless this process is automated through the development of specialized software. Towards this end, this work aims to provide the ability to monitor the political sphere and politic topics discussed by Cypriots in real time.

2.2 Study Necessity

Online social media can be considered as a place where political interaction between the citizens and the politicians is taking place (Kalsnes, Larsson & Enli, 2017). It is widely accepted that politicians should serve citizens' needs. A system that can analyse data and present the attitudes of the citizens towards a subject can transform this interaction into a feedback mechanism for politicians. By receiving feedback, politicians will be able to understand better the community's satisfaction or dissatisfaction about different political issues. If the citizens are unhappy with the attention given by the government to a specific topic, a system that will represent the public opinion about the issue can be used as a pressure lever and force politicians to act.

Politicians at the same time, can use online social media during election campaigns. As an example, before being elected as the president of the United States, Donald Trump used online social media in order to inform people about different activities of his campaign. By doing this, he managed to keep his profile high in popularity among the Americans' public opinion and he was able to communicate directly with the voters (Francia, 2018). Given his election, one can easily conclude that such system can be beneficial during the political campaigns for the candidates too. By having the online social media analytics data, the nominees will know where to focus and how to approach different topics and communicate them to the voters.

Such systems can be beneficial for the citizens as well – which is actually our main consideration. As mentioned above, these systems help politicians to gain information about the beliefs and opinions of the citizens regarding an issue and then use it in their favour, like Cambridge Analytica which was believed to have used Facebook users data to impact the voting behaviour of the US citizens (Peruzzi, Zollo, Quattrocioni & Scala, 2018). There is a limited number of online tools that process online politically related data in real-time and presenting aggregated and simple to interpret and understand results to the users. By developing this platform, citizens will have access to analysed data regarding the Cyprus political landscape. This will be highly beneficial for them for different reasons such as taking the opportunity to be more active. By having more information about a current affair gathered in one place, people can evaluate it and compare it to their own views. By knowing what the opinions of the political actors are, it will be easier and clearer for the citizens to agree or disagree with them and act for or against them. Moreover, since the data will be collected through published online documents, the system will be able to represent the context of the decisions and actions made by the politicians. People will be able to see clearly how the political scene is changing and which topics are discussed in timeliness.

The OSM can be used by their users as a mean to express their opinions and thoughts about political issues. In fact, in USA around 39% of adults posted material regarding the political activity (Rainie, Smith & Schlozman, 2012). Consequently, online social media cannot be ignored in any study investigating the public opinion of a community on political issues.

Nevertheless, their volume is huge and thus special systems relying on Big Data approaches must be implemented and utilized for the wider coverage of the politics – related information. At the same time, the newspapers, at least the ones selected for this research, publish articles created by journalists who are working for well-known organisations. They are considered responsible for the articles they publish. For this reason, they should aim to provide the news

to the most objective level. Furthermore, the work done by them usually is checked before being published and as a result inappropriate or written mistake content is removed and corrected. For this reason, the online media could be considered as a more reliable and valid source of data.

The system that will be developed will merge these two pillars of information in an effort to collect as much and diverse content as possible related to different political issues. The extracted content will then be analysed in order to extract meaningful information regarding different political issues. Interpreting the enormous data will no longer matter since the aggregated results to be produced will reflect the public opinion towards a topic in a more intuitive way. Last but not least, an interesting fact that makes this study important is that there are not many relevant studies in Cyprus. To the best of our knowledge, there is only one research paper that utilized similar approaches but in an offline mode (Triga, Mendez & Djouvas, 2019). The current study can be considered as the only existing Online Media Monitoring (OMM) system that is based on Cypriot data. Having all the data analysed and presented through intuitive representation can also help political scientists who want to study Cypriots' attitude towards politics. It can also contribute to other fields of study such as sociology and political science. The system will be based on the posts being uploaded by the members of the community, expressing their views and opinions about the political issues and current affairs. The material collected could be considered as a primary source of data for further studies of the aforementioned fields. All data collected and analysed as well as data derived will be saved online and will be easily accessible.

3 Theoretical Framework

This study touches upon different theories. In this chapter the theories that will be used for this study are presented and explained in detail. Due to the approaches adopted and to the techniques applied, the use of computer theories and tools were deemed necessary. At the same time, an explanation about political communication will be given since this is the topic which this research is based on.

3.1 Political Communication

One can find different definitions of “political communication” in the literature. McNair, (2018) separated the definition in three parts. The first one points out to the communication produced by the political actors in an effort to achieve their objectives. The second part includes the communication towards politicians by non-politicians such as the citizens who vote. The third part is about the communication describing the activities of the politicians.

Nowadays, research about political communication has advanced to a level where it can be characterized as an interdisciplinary field combining sciences such as politics, journalism rhetoric and others (Kaid, 2004). Political actors do not use political communication techniques only during the election period but also in many other occasions. It can be used to investigate the communication produced by political parties or to examine the strategies they followed. Political communication can be considered as a professional procedure. Experts with different backgrounds and research interests cooperate aiming to help politicians achieve a better communication with voters. To be more precise, according to Holtz-Bacha (2007), the professionalisation of the political communication is defined as the procedure of merging the political system and the use of the online social media. In other words, the professionalisation of the political communication can be characterized as a way that allows the examination of the issues that do not belong to the current affairs (Negrine, Holtz-Bacha, Papathanasopoulos & Mancini, 2007).

Moreover, some people believe that microblogging networks, such as the Twitter, can increase the interaction between the politicians and the citizens. Twitter is used by the people as well as by the politicians to spread their ideas and opinions (Stieglitz & Dang-Xuan, 2013). This is a new tool that allows the political actors to share their activities faster and in first-hand. In the past this was not possible since neither the internet was invented nor the online media.

Currently, for different reasons mainly associated to the economic crisis, some people shift toward more radical and extremist ideologies which results in the emergence of new political parties (Kriesi, 2014). Utilizing OSM, the supporters of these parties are actively participating and sometimes may even contribute to different parties' political communication activities. Online social media can be used by right-wing populists resulting into a nativist spiral communication according to a research made by Heiss & Jörg (2019). Moreover, extreme ideas can be spread to people by news articles in newspapers according to a study made by Blassing, Engesser, Ernst & Esser (2019) titled "Hitting a Nerve: Populist News Articles Lead to More Frequent and More Populist Reader Comments". As it can be easily understood, new types of political communication techniques and strategies arise by the usage of the online media.

Systems like the one proposed, can predict the development of such extreme parties by analyzing the daily agenda. When the representation of extremist opinions starts to be more intensive to the daily agenda, the system which uses real-time data will include these opinions and the results of the analysis will represent the increase of the far right or left parties. A few years back, in 2016, Germans neo-Nazi people used the Twitter to celebrate Hitler's 127th birthdate. His fans tweeted text including hashtags like "#Hitler", "#AdolfHitler" and "#HappyBirthdayAdolf" showing that the internet and the OSM can help the spread of fascism (Fuchs, 2017). Similar events could be predicted via the usage of the proposed system. Clearly, all the above suggest that an online platform monitoring and analysing politically related data on real time can be of high importance for understanding and getting insights in contemporary political communication techniques.

3.2 Data Collections - Crawling & Twitter API

With the evolution of the web from web 1.0 to web 2.0 the amount of the online information became huge. The number of the webpages is continuously increasing and they make more data available to public. With such an amount of web pages, the task of collecting information from a specific website could be a difficult and a time consuming process.

Specific software called "crawler" can be created to scrape the webpage automatically. According to Manning, Raghavan and Schütze, "Web crawling is the process by which we gather pages from the Web, in order to index them and support a search engine" (2009 page 443). In this study, the crawlers will be used for collection of online newspapers articles. Web crawlers are the basic way to collect online data by crossing the web and keeping the information the creator of the crawler needs.

The process of crawling consists of three steps. The first step deals with the pages the crawler will visit. A url or a list of urls should be provided to the crawler in order to visit them. The second one has to do with the downloading of any visited page. In other words, the crawler, after visiting a web page, collects all, or part, of its content. The last step is to store the content of the downloaded page (Udapure, Kale & Dharmik, 2014). One example of the crawler's usage is the retrieval of the information displayed on the main page of the online version of the Reuters. To achieve that, the url of the webpage("www.reuters.com") is given to the software. Then the software visits and downloads the page and finally stores it (usually in a database).

Crawlers can also be used to retrieve data from twitter but usually a different and a lot easier and faster approach is preferred. In fact, Twitter allows third-party developers to use a service it offers called API (Makice, 2009). In the case of tweets collection, the Twitter Application Programming Interface (API) will be used. API is a method used by applications that store data to allow people to access and retrieve some of them (Makice, 2009).

Through Twitter API third-party developers can retrieve information about a single tweet or numerous tweets. According to the official twitter site ("www.twitter.com", 2020) everyone can have access to the API but first they must create a twitter application. The API by default is set to allow the developers to retrieve data that were published by public profiles. The API gives the ability to developers to have access to different type of data like users' accounts, tweets uploaded, ads, etc.

Having created a twitter application, developers can use different commands that the API offers to collect different type of information. These commands are also known as endpoints. Each endpoint has different parameters that should be set in order to specify what kind of data the API will return. An example of an endpoint that returns tweets that include the words "politic" and "news" is <https://api.twitter.com/1.1/search/tweets.json?q=%20politic%20news>. The response will be in JavaScript Object Notation (JSON) format. JSON is defined as a "text format used to exchange data between platforms" (Basset, 2015 page 1). The format of the JSON object consists of pairs of words defined as key-values pairs. For each value that needed to be exchanged between platforms, a key is assigned. An example of JSON is "JSON_obj = {country: "Cyprus", age: 23, city: "Nicosia"}". This JSON object is consisted by 3 keys-values pairs. The first word of each pair (country, age, city) is the key that describes the value passed to the platforms. The JSONs can be very complex with many key-value pairs and nested objects (i.e., the value part can be a JSON object itself).

The JSONs returned by the Twitter API have many of key-value pairs regarding the content of the tweet, the user who uploaded the tweet, the timestamp when the tweet was uploaded, how many times the tweet was retweeted and many more. JSON also gives the ability to the developer to select only a specific value of the collection of key-value pairs by writing the name of the object then dot (‘.’) and then the name of the key. For the selection of the value that describes the country the command should look like “JSON_obj.country”. The method of selecting specific values from the JSONs returned by the Twitter API is used many times in the system.

The crawling and the Twitter API are two different methods that are used to collect online data. The difference is that in the Twitter API, the information is given by the organization (Twitter) in a well-structured format and there is no need for a software to visit the page and collect it manually. On the contrary, the crawling technique is used when the organization (online versions of newspapers) does not provide the data they publish through an API; structure to the collected data should also be created manually.

3.3 Sentiment Analysis

Nowadays, with the advances of online services, people can use different online services for expressing their views and ideas on several issues. In general, the content they publish expresses an opinion and it can be classified as supportive, discouraging or neutral towards a topic or a person. This classification can be achieved via different techniques. One of these techniques is the Sentiment Analysis or Opinion Mining. Sentiment analysis is a computational method which classifies the sentiment of a text towards individuals and affair (Liu, 2010). The main aim of the sentiment analysis is to identify whether a text expresses a positive, negative or neutral approach towards a matter. To attain this task, sentiment analysis takes into consideration the emotional expressions, the weight of the emotions and the correlation of the text to the issue (Nasukawa & Yi, 2003).

There are different ways to calculate the sentiment of a text. One way is by using a method called “Rule-based” sentiment analysis models. According to this approach, each text is labelled as positive, negative or neutral. The facts that contribute to labelling the text is the existence of emoticons and other specific words. The emoticons are separated into three categories according to the vibe they send, positive, negative or neutral. An example of positive emoticon could be a happy face (:)), an example of negative emoticon could be a sad face (:() and a neutral emoticon could be a face that express no vibe. Apart from the emoticons though,

in this method, lexicons that include words related to sentiment are used. An example of such lexicon is the SentiWordNet (Chikersal, Poria & Cambria, 2015). The benefit of the “rule-based” models is that they are more accurate than the automatic ones. On the other hand though, a setback of this approach is that the model creation is a timely and hard process (Borromeo & Toyama, 2015).

Another, different, way to distinguish the sentiment of a text is called “Automatic sentiment analysis”. According to this approach an algorithm requires two datasets. The first dataset is used to create and train a model by using machine-learning techniques. The second dataset is the data from which the sentiment should be extracted. The procedure starts by the creation of the training dataset. After that, each word of the second dataset is assigned a score based on its evaluation with the training set. In the final stage, a probabilistic model is used to calculate and characterise the text as positive, negative or neutral. In contrast to the “rule-base” models, the automatic ones require less effort and time to build and use. On the negative side though, the lack of the usage of lexicons to calculate the sentiment could result in less efficient and of lower accuracy results (Borromeo & Toyama, 2015).

In this research the sentiment analysis model that will be used is “rule-based”. The model is provided by a python library called Textblob. Among other options the library is able to calculate the sentiment of a sentence by giving results of polarity and subjectivity (Loria, 2018). According to the library’s documentation the value of polarity, which shows how positive or negative a text is, is between -1.0 and 1.0, while the subjectivity ranges between 1.0 and 0. To calculate the sentiment level it uses a built-in lexicon with sentiment values (from -1.0 to 1.0) assigned to each word. The selection of this way to calculate the sentiment was made due to its simplicity.

3.4 Machine Learning Techniques

In addition to sentiment analysis, another analysis that will be applied on the collected data is data classification and data clustering according to their topic, e.g., economy, foreign policy, etc. This can be achieved using different methods. The technique that will be used for this research is called Topic Modelling, a technique that belongs to the field of Machine Learning (ML). Machine learning is a sub-field of the computer science’s field of Artificial Intelligence that studies how the software can evolve and be more efficient autonomously (Hosch, 2019). There are two types of machine learning. The first one is called supervised machine learning and the second one unsupervised machine learning.

Supervised machine learning is defined as the process of the development of algorithms that can create archetypes and hypothesis by handling user provided data in order to calculate new unknown to the algorithm instances (Singh, Thakur & Sharma, 2016). Contrariwise to the supervised machine learning is the unsupervised machine learning where algorithms are trying to find hidden patterns in a dataset without using any trained model (Lloyd, Mohseni & Rebentrost, 2013). Each of two types of the machine learning, supervised and unsupervised, uses different techniques. Specimens of supervised machine learning techniques are the Linear/Polygonal Regressions and the Classification. On the other hand, examples of unsupervised machine learning techniques are the Clustering and the Dimensionality Reduction.

Topic modelling can be characterised as a tool that uses a piece of information, like text, or even a collection of texts, like corpus, and tries to find patterns between the words mentioned in the text and the semantic meaning (Graham, Weingart & Milligan, 2012). An application of Topic Modelling is the Latent Dirichlet Allocation (LDA). LDA is a “probabilistic model for collections of discrete data such as text corpora” (Blei, Ng & Jordan, 2003, page 993). In fact, LDA uses the vocabulary in each document in order to shape the latent topics of them (Newman, Asuncion, Smyth & Welling, 2008). The main reason why an unsupervised model, LDA in our case, was selected for this research, is because the main feature of the system is to automatically create clusters of articles (and/or Tweets) automatically each time a new analysis will be made.

3.5 LDA

As it has been already mentioned, in this research, the machine learning technique that will be used is an unsupervised approach that clusters a corpus of documents. To achieve that, an LDA model provided by the Gensim python library will be used. In general, LDA can be characterised as a model that uses probabilistic theories about a collection of documents (Blei, Ng & Jordan, 2003). The main idea of this model “is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words” (Blei, Ng & Jordan, 2003, page 996).

The LDA model starts the process by taking as an input a text corpus. Then the model breaks down all the words of the texts to different tokens. Each different token is used to shape a new dictionary that will be used later by the model. Having created the dictionary with each unique word of the corpus, the model continues the process by assigning ids to every word in the

corpus. The ids assigned to the words are referred to the id number that was given to each word in the dictionary. At this point it is essential to say that all words should have the same format to avoid creating duplicate values in the dictionary, e.g., all lower case, all without punctuation, etc. If a word is written in a different way, the model will consider that as two different words, even without any relation among them.

The corpus given to the LDA, can also include n -grams. The n -grams are defined as a part of a string that is created by using n number of words (Cavnar & Trenkle 1994). In the case where the n equals to two, the n -gram is named bigram and when the n -gram is created by using three words is called trigram. The n -grams are used in various procedures including the field of information retrieval. The reason that n -grams will be used in this research is to separate issues that are described with more than one word. For example, the bigram “Cyprus dispute” describes the topic of the Turkish invasion in Cyprus. At the same time, the two words used in isolation have a completely different meaning vastly different from the previous one.. Thus, using bigrams, will treat these two words together preserving the original meaning. In this study, we support both bigrams and trigrams.

After pre-processing the corpus for creating bigrams and trigrams and for removing some stop-words (i.e., words that have no internal meaning, e.g., “and”, “or”, etc.), the model clusters the data into different groups. The number of groups, called topics in LDA, that will be created must be defined by the user. Different methods can be used to identify the ideal number of clusters that are based on the coherency of the generated clusters. At the current status of the platform, such method is not supported and the number of clusters should be declared by the user without any support.

4 Literature Review

Online social media play a significant role in the political discourse process everywhere around the world (Conover et al., 2011). More specifically, among other social networks, Twitter is used by politicians to communicate with voters. To clarify, Twitter has been used in many different occasions such as the pre-election campaigns and other type of campaigns (Larsson & Moe, 2014). The introduction of online social media in political context first occurred during Obama's presidential run in 2008. Many analysts have characterised his win as a result of Twitter's usage (Budak, 2011). However, using the Twitter cannot guarantee the majority of the votes for the candidate. This can be justified by the results of the 2011 United States elections, which showed no correlation between the candidates who used the Twitter to promote themselves and the ones who did not (Lassen & Brown, 2011).

In the most recent USA elections (2016), both main candidates used the online social media to come closer to voters despite the fact that they used it differently. According to Enli (2017), Hilary Clinton's campaign was considered to be closer to election campaign theories. While Trump's was inept, at the same time it was more authentic. Everybody knows the outcome of those elections. A research that was published few years ago shows that Twitter evolves the communication process (Parmelee, 2014). With a modern different approach to the news represented, Twitter can boost or lessen the promotion of an issue to the public sphere.

According to Go, Bhayani and Huang (2009), the tweets shared in the Twitter platform can be classified according to their sentiment. What is needed for the classification though, is a term that according to the tweet's content, it will be characterized either as positive, negative, or neutral. This kind of sentiment analysis can be beneficial for politicians in many different occasions. Some of these cases are when politicians want to find out how satisfied or dissatisfied the citizens are regarding policies taken by their colleagues or to envision the outcome of an upcoming election (Yaqub, Chun & Alturi, 2017). The collection of tweets that are available for analysis help the researchers to shape an image of the public attitude towards politicians and their actions. Without the Twitter platform that gathers all the data in one place, politics related studies would be a lot more difficult since the opinions would not be published and easily accessed (Clemence, Doise & Lorenzi-Ciodi, 2013).

Furthermore, politicians using Twitter's data to achieve their goals in their campaigns is common nowadays. The example of Obama's presidential run, as well as other politicians' campaigns, attracted the attention of many researchers to study the platform's importance

during election periods (Yaqub, Chun & Alturi, 2017). In a study conducted during the 2009 congress elections the researchers found out that in addition to the actual text, which was included in the tweets, hyperlinks were also found. The huge majority of the hyperlinks linked the tweets to announcements made by the candidate, were longer than the limit of 140 characters that Twitter sets. However, there were some rare cases where the link sent the user to another person's data. The rest of the tweets, i.e., the tweets that did not include any links, were referring to facts that happened in the past or about decisions taken in the past referring to politics. (Golbeck, Grimes & Rofers, 2010). Having examined the way politicians use Twitter, what can be said is that politicians are very active on twitter, utilizing it as a mean to promote themselves and reach the masses. In some cases, by informing the citizens with short but meaningful texts, while in some other occasions via connecting their Twitter profiles with external sources with more detailed information.

Apart from online social media, online newspapers publish articles related to politics. The role of these articles is to inform the citizens about the current affairs and in more general to present a country's political background. The online editions of newspapers help politicians to achieve their goals, either by the high speed of the news articles publishing or by the details provided about issues (Singer, 2003). In a study made by Triga, Mendez & Djouvas (2019), they used the information presented by the mainstream media as research's data. In fact, their dataset was based on a collection of media articles related to the 2018 presidential elections published in the online versions of Cypriot news corporations. So, it can be said that the online media, by publishing political related articles, can represent the political background of a community.

As a conclusion, the brief review of previous studies on Twitter usage by politicians, points to the conclusion that Twitter is a medium that is mainly used by politicians to expose themselves in public. At the same time, the articles that are published in newspapers can be useful to shape a country's political framework. By having this type of information, politicians' activities and political related articles published online and being accessible by everyone, can help experts to use it as data for their studies. Taking all the aforementioned studies into account and having a huge amount of data related to a field, like politics, it gives the opportunity to scientists to explore and shape the public opinion of the citizens regarding a topic. In that way, this study will use data that have been produced by Cypriot politicians as well as the politics-related articles published in Cypriot newspapers in an effort to analyse and create a visual representation of the public opinion about certain issues political issues that are currently discussed by Cypriots.

5 Methodology

This thesis proposes the creation of a novel online platform, called an “Online Media Monitoring” (OMM) platform that will collect, analyse and represent graphically the political related data collected from online news media and from Twitter.

For the development of the OMM platform, a well-defined pipeline of modules is proposed; that is the data acquisition module, the data analysis module, and the online representation of the results module.

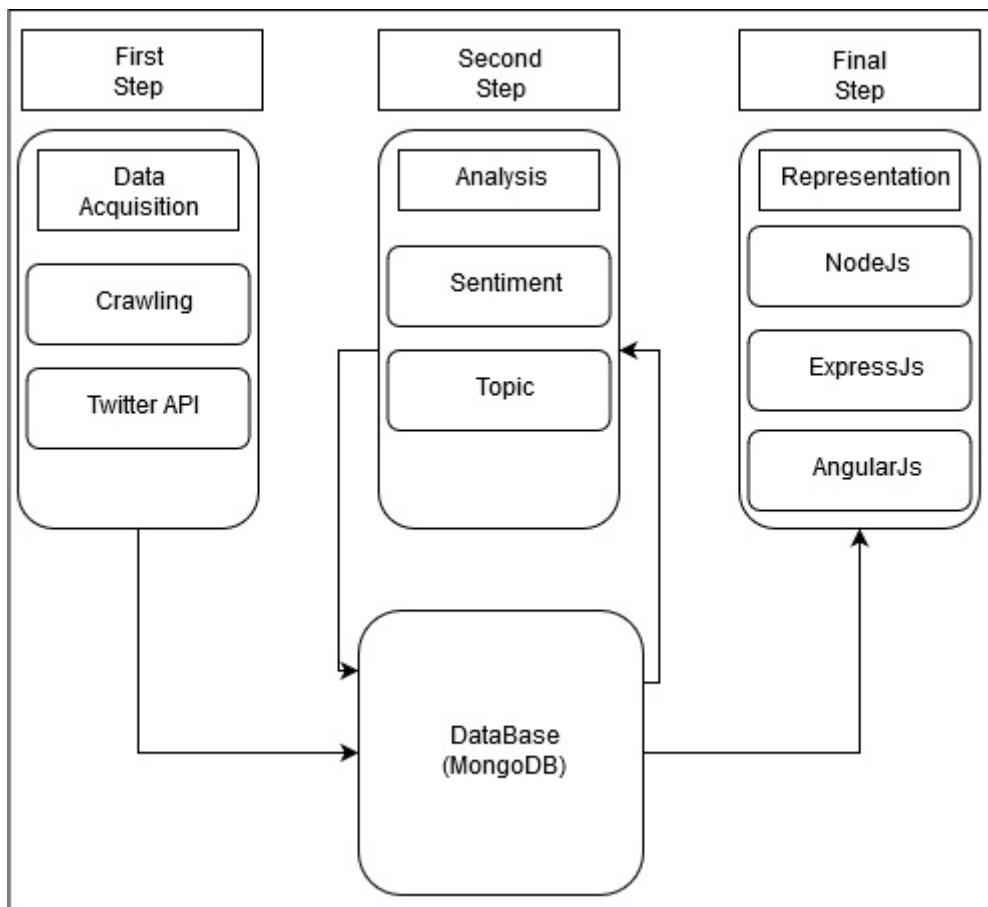


Figure 1: Platform’s Architecture

The development of the three modules will follow a logical order to avoid errors that may cause problems and malfunctions to the platform.

5.1 Data Acquisition

Currently, data acquisition module consists of two tools, one for collecting tweets and one for collecting online new media articles creating two distinct datasets. The first dataset includes tweets that are posted by the official twitter accounts of the six parliamentary political parties in Cyprus. The second part of the sample consists of the political related articles published in

online Cypriot newspapers. The newspapers used to collect articles are the online versions of the two of the biggest news corporations in Cyprus; that is Sigmalive and PolitisNews. In future releases, additional online newspapers will be used since the architecture supports the easy addition of new crawlers.

5.1.1 Python Crawler

As already mentioned in the chapter 3, a crawler is a software that its purpose is to browse the web and download webpages that match specific criteria. Developers can use crawlers provided by different libraries. One such library in the python programming language, that provides the crawling technique and it is used in this research is named “Scrapy” (A Fast and Powerful Scraping and Web Crawling Framework, 2020).

The information needed for this research is scattered in many different webpages all over the web. A crawler selects the information that is enclosed by specific HTML tags in the webpage’s source code. Each webpage has different structure with different tag names so a designated crawler for each online media that needs to be crawled is needed.

In this research the crawling technique starts by selecting the links of news articles that exist in the newspaper’s webpage. Then the XPath (XPath is the path that one should follow starting from the root element of an XML file for reaching particular item(s)) of the news articles will be used to extract the link of each article. For example, starting from the generic page of the economy, e.g., “<https://politis.com.cy/category/oikonomia/>” for PolitisNews, all the links to the economy articles will be collected. Having collected the links, they will be stored in the mongo database as a JSON object along with some extra information indicating the time when the links were collected. Then, a second crawler retrieves the saved links from the database and starts the process of downloading the corresponding content of each article. After downloading the content of all articles, again by using XPath, the crawler retrieves the main text of each article and stores it back in a different collection in the database again as JSON object with the same time-related information. This process continues until downloading all the webpages.

5.1.2 Twitter API

Twitter allows third-party companies and/or developers to have access to posts published by its users. It provides companies and developers access to the data by numerous Application Programming Interfaces (API) outlets according to their needs. Twitter provides a wide publicly accessed Twitter data through APIs. In this study, the Twitter API will be used to

acquire posts on Twitter made by the official accounts of the big six Cypriot political parties (Akel1926, Diko1976, edek1969, DISY1976, cygreens, SymmaxiaPoliton).

In this study, the twitter API was used through a python library called “Tweepy”. The library uses the Twitter API and retrieves data from the platform according to the developer’s needs (Roesslein, 2009). The returned values are again in JSON format. The JSON object contains information regarding the published time of a tweet, the text of the tweet, the creator etc. An example of some of the returned values by the Twitter API is:

```
{  
  'created_at': 'Tue Apr 28 15:01:43 +0000 2020',  
  'id': 1255150205508349954,  
  'full_text': 'Παραγωγική διαβούλευση το πρωί, μέσω τηλεδιάσκεψης, της  
  Επ.Οικονομικών και αμέσως μετά της Επ. Νομικών .Αριθμός νομοσχεδίων και  
  πρωτοβουλιών πήραν τον δρόμο τους. Αύριο Επ. Εμπορίου και την Πέμπτη, 17.00,  
  Ολομέλεια του σώματος. Κοινή η προσπάθεια για αντιμετώπιση και έξοδο από την  
  κρίση. https://t.co/aJ6JQJ9LEC'  
  'name': 'Δημήτρης Συλλούρης',  
  'screen_name': 'Syllouris'  
}
```

The example shows only a limited number of the returned values of the JSON sent by the Twitter API. The complete JSON is not shown due to its huge size. In our case, we are only interested on the “full_text” field.

5.2 Research Tools

The tools required for this research are technologies related to the web, data storage, web scrapping and other tools that will be used to process and analyse the data collected.

5.2.1 Web technologies

The first step is to create a new website by using the “MEAN stack” technology. MEAN stands for “MongoDB”, “Express.js”, “Angular.js” and “Node.js”.

AngularJs is a javascript framework built and maintained by google. Its purpose it to create front-end webpages. It works under the Model-View-Controller (MVC) approach. Model files

are responsible to manage the data between the webpage and the place where they are stored, usually in a database. The View files are responsible for the data rendering and the appearance of the webpage while the controller files are the middleware between the models and views. This technology was selected because is one of the state of the art technique for web development. Moreover, it provides the developer with huge number of third-party libraries that make the representation of the information friendly to the user.

The Node.js is an environment that enables the execution of JavaScript code on servers. The Node.js package coordinates the tasks performed on the front-end (the web browser) and the back-end (the web server and the database). The express.js is a framework for the Node.js and it handles the requests send by the user through its browser

Finally, the last part of the MEAN stack is the MongoDB. MongoDB is not a relational database that will be used for storing the data collected. MongoDB is a non SQL database. The data are stored as “documents” in “collections”, instead of “records” in “tables” that are used in traditional SQL databases. The reason for this selection was due to the database’s document-oriented approach that fully matches the format of the collected data, as well as the high performance.

5.3 Procedure

The procedure starts with a twofold process for data acquisition. The first step is the information retrieval from the news media sources and the Twitter. To accomplish that, the development of the crawlers is needed. The crawlers will scrap specific online news media and collect online news articles. Having retrieved the information from the twitter, the retrieval of information published in online newspapers follows. To achieve that, another crawler will be developed which will use user-defined queries for collecting the related articles as explained in the section 5.1.1. All the information collected will be stored in the mongoDB as JSON objects without any initial processing. The aforementioned functionality will be handled by the user through a simple and intuitive web interface.

The next step of the pipeline is the text processing and the analysis of the processed data. The analysis is based on an unsupervised machine learning technique for clustering data into groups. An LDA model will be used to group the corpus in topics. Each group formed can be characterised by its most important words that describe the subject of the topic. The most representative 15 words of each group will also be stored back in the database to indicate the actual topic of each group.

In the final step of the proposed approach, the aggregated results regarding different analyses will be presented graphically on the platform's website. There will be a page for the representation of the pre-built analyses and another one where the users' analyses will be shown. Each analysis will depict the results of the analyzed and classified data on graphs.

6 Data Analysis

6.1 Texts processing

The first step in data analysis is to collect the data which in this case are texts drawn from Twitter and/or the online Cypriot newspapers and are stored in the database. Since the data are taken from online sources in real time, each time a new analysis takes place the used data will differ. Moreover, the user who wants to run an analysis can choose among a variety of sources for collecting data. Furthermore, when it comes to the newspapers' articles, the user can specify the categories (e.g. Politics, economy, etc.) of the selected articles through the web interface (Figure 15). This can easily be done since the articles are uploaded to different categories and the crawlers can choose from which pages to collect links to the articles to be scrapped. Additionally, the user has also the option to select twitter users' accounts and/or hashtags to collect the texts of different tweets. In other words, the data used for every analysis can be vastly different.

Before storing the data in the database, it is of great importance to remove some words, numbers and/or symbols, that may appear in the texts and are irrelevant to the actual meaning of the content. Such words could be HTML tags, e.g. <div>, video and image urls. At this point is essential to mention that the hashtags appearing in any type of text, either in Twitter posts or in newspapers articles, are preserved. This is because people (twitter users, journalism, etc.) use hashtags as an effort to summarize/categorize and emphasize on the main meaning and topic of the tweet or the article published.

Furthermore, before storing the texts in the database another pre-processing step is applied. That is letter capitalisation in order for all data to be in the same case, capital in our case. As said in the chapter 3.5, all words must have the same format, either capital or lower case. So, for this study all the words were capitalised. Having the same word with a different format, the algorithm used will consider it as a different word and as a result the word will not have the same weight and importance. Consequently, with the same words written in a different way can cause error and malfunctions to the LDA model and make it less efficient. The same words should be written in the same way without any differences because of a letter being capital or not.

When the texts processing is completed, the processed texts will be stored in the database and will be ready to be used for further analysis. In addition to the processed text, some extra information for administration purposes is added to the database, e.g., the user who wants to

create an analysis. Each piece of document that was processed is finally stored in the database along with the source text, the date and the system's user id, his/her email in our case.

6.2 LDA Analysis

Having stored the texts in database everything is set for the creation of a new LDA analysis. The first step is to retrieve the data from the database and begin the formation of a new model. Then, we apply some further processing that is specific to the type of analysis that is performed, i.e., LDA.

In our case, since we are dealing with Greek language, some further processing is required to deal with accents used in Greek words. Since accents are not yet removed from the data, the procedure to create a new analysis starts by removing the accents from each word. Each Greek vowel letter with an accent is replaced with the same vowel letter without the accent symbol.

Next, the corpus is cleaned from some words that have no contribution to the meaning and consequently to the topic discussed in an article or tweet. Having a meaningless word appearing so many times in a corpus, the machine will interpret it wrongly as an important word. For example, the word "and" does not have any meaning and it should be removed from the corpus. If it is not removed the model after locating so many times it will consider it as a word of great importance. These words are usually called as "stopwords". According to Wilbur & Sirotkin (1992), stopwords are considered the words that have the same possibility to occur in a document related to a query to documents that are irrelevant to the same query. Examples of such words in English are the words "and", "or", "to", "as" and many more.

Successfully completing the removal of the accent from the words and the Greek stopwords from the texts the corpus is ready to be transformed in a form that is acceptable by the machine and be used to create a new analysis. To each word in the corpus a unique number is assigned to by the model (this is an internal process made by the Gensim library), so that each corpus is treated by the program as a list of numbers. There is also the option for the user, using the web interface, to make bigrams or even trigrams in the model. As explained earlier, bigram is a pair of two words and trigram is a sequence of three words. If the user chooses to use bigrams or trigrams, the program will shape them through different methods and then the newly created words (bigrams or trigrams) will also be assigned a unique number as an id (the concatenation of two or three words will be treated as a new word). The final step before the creation, is to declare the number of topics, n , that the model will cluster the corpus (again, this is a parameter specified by the user through the web interface – see Figure 5).

There are many different libraries that implement the LDA algorithm. However, in this research the LDA model used, is created, and provided by the library called Gensim. The choice of this library was made because of its simplicity and easy to use way and the reputation of the Gensim library in the research community. LDA will process the corpus and finally will split it to the n clusters. Having finished the clustering phase, the next step is the labelling of each cluster with a meaningful label, e.g., the cluster consisting of the words related to the economy should be labelled “economy” rather than the generic label, e.g., Topic 1, assigned by Gensim. Gensim, in order to assist the user, presents the most prevalent and representative words for each newly clustered topic created.

Having completed all the above steps, the created analysis is ready for observation and use. The user who created the analysis will be able to inspect and name the topics according to the top 15 words of each topic (Figure 4). The top words of each shaped topic are stored back in the database.

6.3 Sentiment of the Topics

To achieve the sentiment analysis of each topic, the procedure applied had some extra steps since the texts are in Greek – to the best of our knowledge there is no reliable sentiment analysis tool that works for Greek. First, the top words of each topic are put together and shape sentences with the 15 most representative words of each topic. It was impossible to use all the words of each topic since in order to avoid the deadlock of sentiment analysis of Greek text, we had to translate the text into English. However, there is no tool which allows an unlimited number of translations. The library used for the translations restricts the size of the text that will be translated. So, the translation from Greek to English was necessary to find the sentiment of each text. The technique of using only the top 15 representative words of each topic to calculate the sentiment of each topic is not the best one. Clearly, a better approach to calculate the sentiment of each topic would be to use the average sentiment of all of its documents. However, this is not a feasible approach in our case. Nevertheless, it can be easily implemented if someone is willing to pay the premium for unlimited translations.

Afterwards, the translated text is used and the python library “Textblob” calculates the sentiment of the 15 top words and it shows a value between -1 to +1. The lowest sentiment value (-1) represents the most negative level and on the other side the highest value (+1) represents the most positive level of the text. In addition to the sentiment value of the text, there is also another value the library calculates. That is the subjectivity value. The rate of the

subjectivity value ranges from 0 to 1. The closer to 0 the value is, the more objective the predicted sentiment is. On the other hand, the higher subjectivity value, the less objective the calculated sentiment is. These values are also stored in the database along with the top words.

6.4 Text Similarity to Topics

The final analysis that takes place in this study is the text similarity to the newly formed topic clusters. The users can type a text in the web interface and by using the stored analyses, the system can identify and return the topic similarity to the given input by a value between 0 and 1. The topics allocation of the new text is achieved by using a build-in method of the LDA model. The similarity test process returns results values from 0.00000001 to 1 to indicate how close to each topic the text is. The higher the value, the closer a text is to a specific topic.

According to similarity results, a radar graph is created which displays the distance between the text and each topic. The visual representation aims to help the user to have a better understanding of the relation of the text to each topic (Figure 11).

7 System Implementation and Development

The platform developed is an online web application accessible through a browser from everywhere in the world. Even though it is publicly available, the users must create an account in order to be granted access to all features available.

The first page a user sees is the landing page with a descriptive text about the service. On the top left corner there is a button and when clicked, a drop-down menu appears allowing the user to navigate through the website. If the user is not logged in, only limited functions of the website are available. To be more precise, without being logged in, the visitor can only read the text in the landing page and the “About Us” page. Even though the “Previous results” are accessible, there is no analysis displayed unless the visitor logs in or creates an account.

If the visitor logs in, then the existing results, i.e., existing results from previous analyses stored in the database, are shown and more options on the top left navigation menu appear. The new pages allow the user to access his/her stored analyses and to create new analyses.

7.1 Existing Results

The platform offers a set of existing results created previously by the administrators of the system so that everyone can immediately use. The image below (Figure 2) shows how the previous results page looks like.

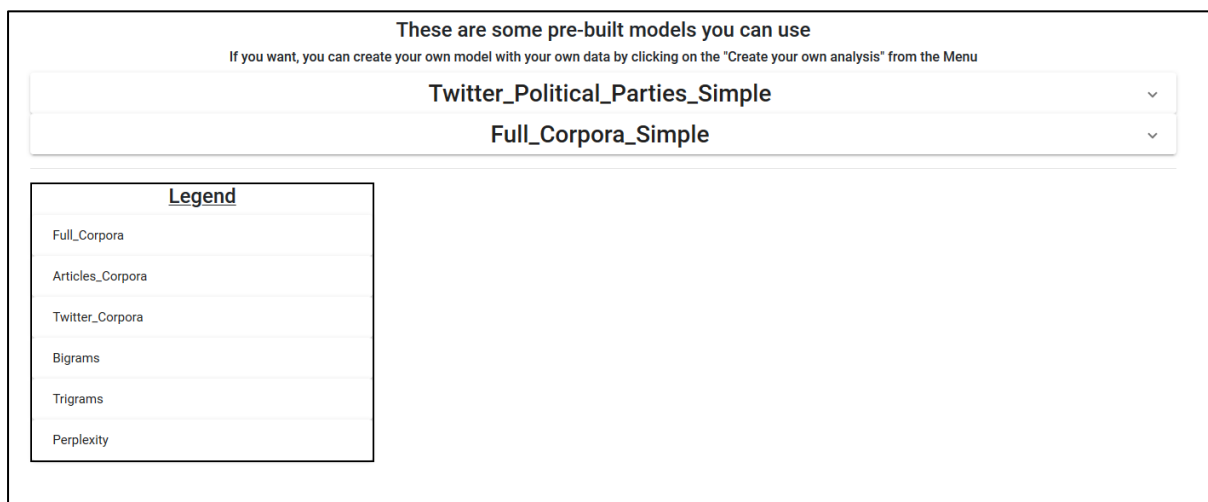


Figure 2: Previous Results page

At this page, all the existing analyses are displayed and the users can choose which one to use. As it has already been mentioned, each analysis differs from the others. This is because each new analysis is done with more recent data, since crawlers and the Twitter API collect the latest

available data. Here I should mention that currently new analysis using data that is already stored into the database is not supported. In future extensions and improvements of the platform, this feature will be available.

The legend that appears on the left of the screen consists of the names of the existing analyses. By clicking on a button, a drop-down menu opens (Figure 3) and a more detailed explanation is given to the user.

<u>Legend</u>
Full_Corpora
Full Corpora indicates the models that were created by using all the articles and tweets that were collected
Articles_Corpora
Twitter_Corpora
Bigrams
Trigrams
Perplexity

Figure 3: Legend of the previous results page

By clicking on an analysis box, the box expands displaying two tabs. The first tab provides the top words of each topic, whereas the second one is a visual representation of the clusters created. Top words are defined as the words that contributed the most in the clustering process.

Twitter_Political_Parties_Simple

Words Graph

Topic 1
 CYPRUS,ΠΡΟΕΔΡΟΥ,ΠΡΟΕΔΡΟΣ,ΟΙΚΟΝΟΜΙΑ,ΑΠΟΦΑΣΗ,ΚΥΠΡΙΑΚΟ,ΚΡΙΣΗΣ,ΣΤΕΦΑΝΟΥ,ΥΓΕΙΑΣ,ΠΑΝΔΗΜΙΑ,ΕΚΠΟΜΠΗ,ΕΠΙΣΤΟΛΗ,ΣΥΝΑΝΤΗΣΗ,COVID_19,ΑΠΡΙΛΙΟΥ

Topic 2
 ΑΚΕΛ,ΚΥΠΡΟΣ,ΠΟΛΙΤΩΝ,CYGREENS,ΤΥΠΟΥ,ΕΚΠΡΟΣΩΠΟΥ,ΑΝΑΣΤΟΛΗ,ΚΙΝΗΜΑΤΟΣ,ΕΠΑΝΑΠΑΤΡΙΣΜΟ,ΕΡΓΑΖΟΜΕΝΟΥΣ,ΟΙΚΟΛΟΓΩΝ,ΖΗΤΟΥΜΕ,ΔΗΣΥ,ΚΥΒΕΡΝΗΣΗΣ,ΣΥΝΕΡΓΑΣΙΑ

Topic 3
 CYPRUS,COVID_19,ΚΥΠΡΟΣ,CYGREENS,ΚΟΡΩΝΟΪΟΥ,COVID2019,ΕΕ,ΠΑΣΧΑ,ΜΕΤΡΩΝ,CORONAVIRUS,EDEK,COVID19,ΠΟΛΙΤΙΚΗ,ΔΟΣΕΩΝ,ΜΕΤΡΑ

Topic 4
 CYPRUS,ΤΡΑΠΕΖΕΣ,COVID,CYGREENS,ΑΝΤΙΜΕΤΩΠΙΣΗ,ΚΥΠΡΟ,ΚΡΑΤΟΣ,ΚΟΡΩΝΟΪΟΥ,ΕΠΙΧΕΙΡΗΣΕΙΣ,ΚΥΒΕΡΝΗΣΗ,ΚΑΛΟ,COVID_19,COVID19,ΝΟΜΟΣΧΕΔΙΟ,ΠΡΟΕΔΡΟΣ

Topic 5
 CYPRUS,ΘΕΣΕΙΣ,ΔΕΛΤΙΟ,ΜΕΝΟΥΜΕΣΠΙΤΙ,COVID,STAYSAFE,ΕΙΔΗΣΕΩΝ,COVID_19,ΕΝΟΙΚΙΩΝ,ΠΟΛΙΤΙΚΗΣ,ΡΕΠΟΡΤΑΖ,ΕΡΓΑΣΙΑΣ,DISY,ΚΟΜΜΑΤΩΝ,ΑΡΧΗΓΩΝ

Topic 6
 CYPRUS,ΜΕΤΡΑ,ΣΤΗΡΙΞΗ,ΠΡΟΤΑΣΕΙΣ,ΠΕΡΙΣΣΟΤΕΡΑ,ΟΙΚΟΝΟΜΙΑ,ΑΚΕΛ,ΚΥΒΕΡΝΗΣΗ,ΠΑΝΔΗΜΙΑΣ,ΚΥΠΡΙΑΝΟΥ,ΟΙΚΟΝΟΜΙΚΩΝ,ΥΠΟΥΡΓΟ,ΟΙΚΟΝΟΜΙΑΣ,ΑΝΤΡΟΥ,ΕΠΙΧΕΙΡΗΣΕΩΝ

Perplexity is: **-8.370**

This analysis was created on: **2020-04-23**

[Name the model and use](#)

Full_Corpora_Simple

Figure 4: Previous Analysis words

The first tab (Figure 4) consists of the topics with generic titles (from Topic 1 to Topic N, where N is the number of topics). Users can rename the analysis and use a more descriptive name by clicking on the “Name the model and use” button (Figure 5), e.g., Signalive_May_2020. After providing a descriptive name to the analysis, users can also assign names to Topics. These are usually names describing a category according to the articles assigned to the specific topic, e.g., economy, foreign policy, sports, etc.

[Name the model and use](#)

Name the model

Give a name for our model *

[Use](#)

Figure 5: Name the model

In the text-area “Give a name for your model” the users can write the name they like to give to the analysis. By clicking on the “Use” button the name changes. Having renamed the analysis, the website redirects the user automatically to the “Your Analysis” menu where he/she can see their newly renamed analysis.

The second tab, called “Graph”, contains the visual representation created during the analysis. This graph (Figure 6) is generated by the “Gensim” python Library, the library used for the analyses.

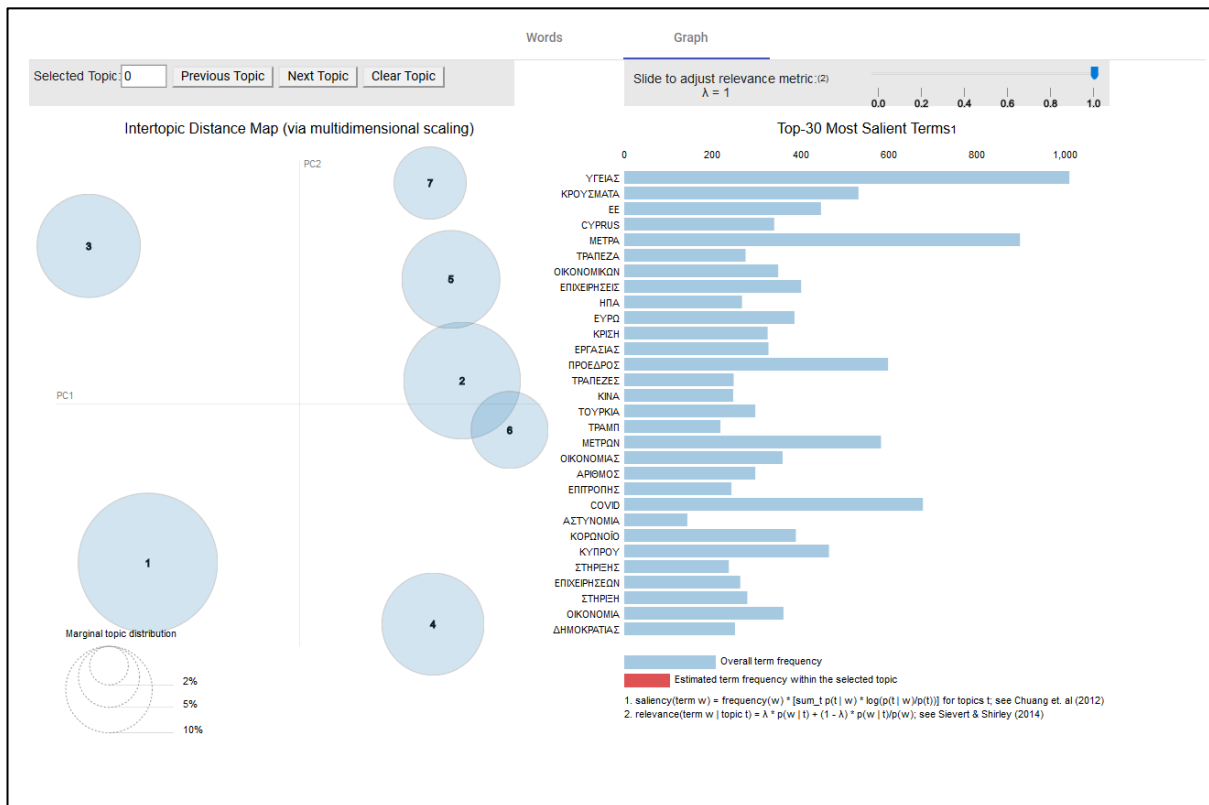


Figure 6: LDA analysis results graph

Each circle represents a topic and the list of words on the right show the overall terms’ frequency in the corpus. As an example, in the case of the word “ΥΓΕΙΑΣ” it appears nearly 1000 times in articles of the corpus. The words presented on the right side are the 30 most salient words that were assigned by the model. Saliency is calculated by the model during the analysis by using a specific formula regarding the word’s probability to be generated by a latent topic. On the right side one can also check how relevant each word is to the selected topic. To enable this feature, the user must click on a topic-circle and then slide the top right bar. The higher the value of the λ metric, the more relevant the words are to the topic, where lower λ

values denote exactly the opposite. If the user does not click on any topic, the words displayed are the ones with the highest value as per the analysis.

By clicking on a topic, represented as a circle – cluster (Figure 7), the graph changes and provides information regarding the topic’s top words and an estimation about the frequency of each word appearing in the topic. There is also a comparison between the topic’s most important words (with red line) and the most important words of the entire corpus (light blue colour).

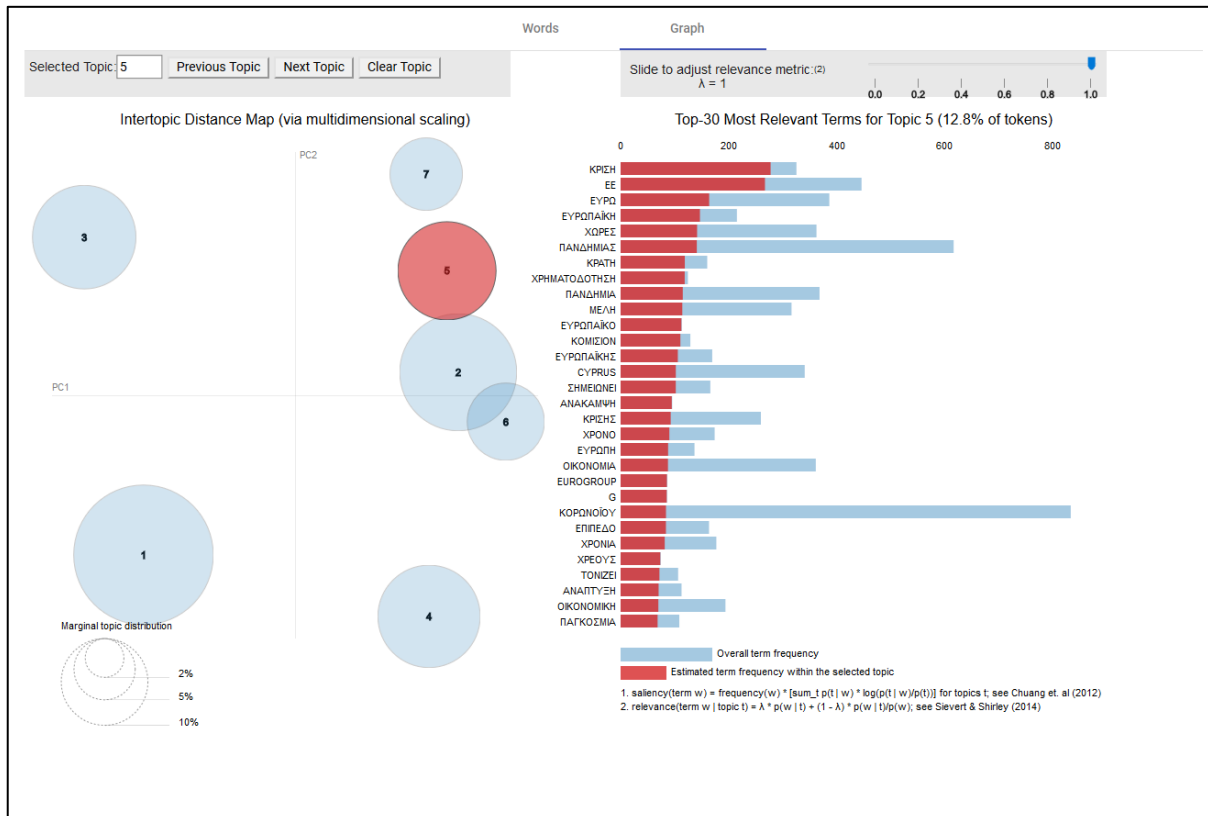


Figure 7: Topic's 5 representation

This graph is the first graph presented to users before proceeding to other types of analysis. In the following steps, there will be more graphical representations of the generated results regarding the text similarity among topics. These extra features are currently available only under the “Previous Analysis” menu.

7.2 Your Analysis

The “Your Analysis” menu is the place where users’ analyses are listed. This menu contains the analyses users created, among other new features like the similarity test and a chart related to the similarity results. Each one of these features are displayed in a different tab. The first

two tabs are the same as the “Previous Results” menu (explained above in section 7.1); however, the other two tabs are available only in this menu. In the first tab, users can observe the words and the topics’ titles as well as the sentiment of the top 15 words of each topic (Figure 8). Each analysis has the name they chose either when they created a new model or when they used one from the pre-built list.

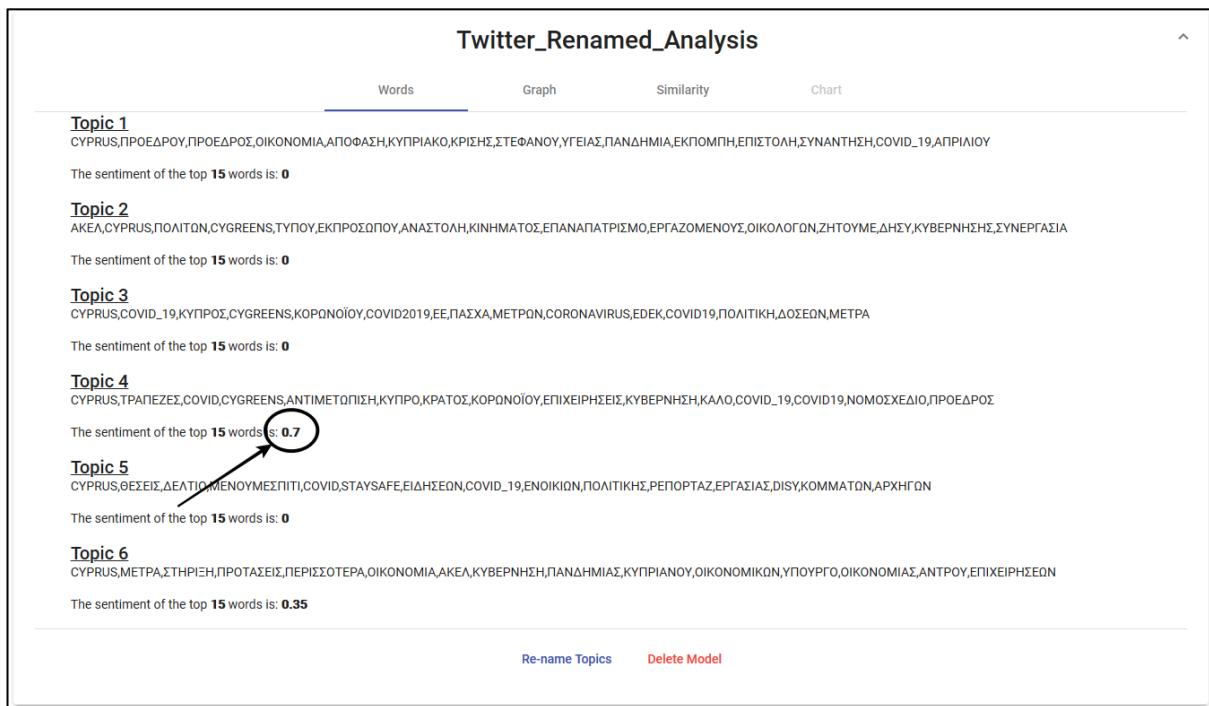


Figure 8: Topics, words and sentiment

The sentiment of the top 15 words quite often has neutral (0) or low value. The explanation for this is that usually the top 15 words of each topic are nouns which express no sentiment. To address this, we plan to expand the number of words used until finding enough words that express sentiment.

At the same tab, there is a functionality that allows the user to rename the topics. The procedure to delete an analysis is quite simple and easy for the users. More specifically, when they click on the red “Delete Analysis” button, a warning pops-up asking the user if he/she is sure about deleting the analysis. If the user confirms that he/she wants to delete the model, the analysis will be deleted. At his point it is important to note that if anyone deletes an analysis it is gone forever, it is impossible to restore it. It is possible, of course, to create a new one but with different data. So, it is highly recommended to not delete any analysis even if the user does not need it any more.

The other button refers to the renaming of the topics. By clicking on it, a list of n-inputs (n=the number of the topics) appears and the user can provide a new name for each topic as shown in the figure below (Figure 9). Having provided the new names, the user should click on the “Re-name” button and the names will change. At the same time, the page refreshes and the analyses are shown with the newly assigned topic names.



Re-name Topics Delete Model

Re-name the topics

Topic 1 *

Topic 2 *

Topic 3 *

Topic 4 *

Topic 5 *

Topic 6 *

Re-name

Figure 9: Rename the topics

In the third tab named “Similarity” (Figure 8), there is a text area where the user can write a text and the system will check how close it is to each topic. The result values (Figure 10) vary from 0.00000001 to 1. A lower value indicates less similarity to a topic and on the contrary, a higher value indicates a higher similarity to the topic. The similarity level of a given text is calculated by the generated model. The LDA model used to create the analysis provides the ability to the users to compare how close to each topic a word is.

As it can be easily understood, the sum of all values should be 1. In some cases, the text is distributed to some of the topics. Figure 10 shows the similarity results of the word “πρόεδρος”. As it is depicted, this word has over 60% probability of belonging to the “First” topic and considerably lower probability of belonging to any other topic.

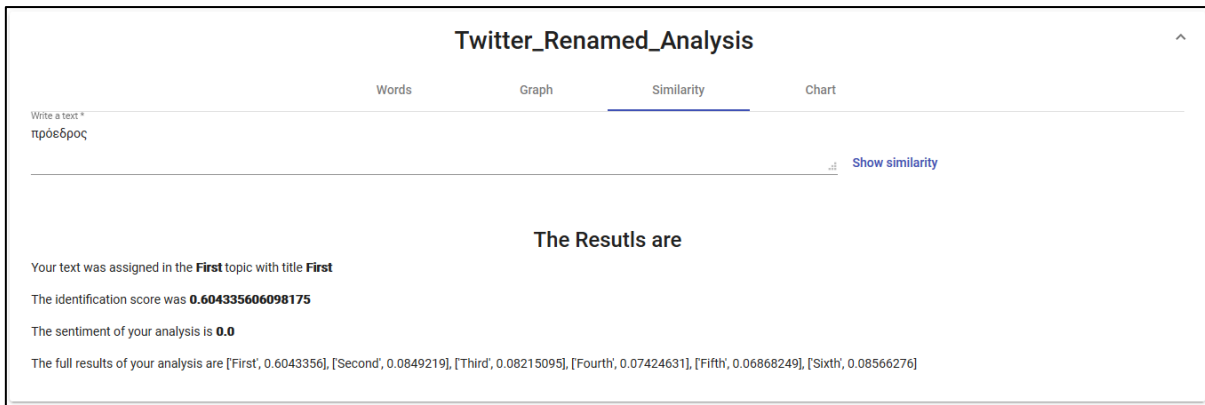


Figure 10: Similarity results

Apart from the similarity results, the sentiment of the text is shown to the user. That sentiment is calculated according to the words written in the text area.

Having finished the similarity test, the fourth tab, “Chart”, is activated and the user can check the results of the test on a radar chart (Figure 11) rather than in a text mode shown under the “Similarity” tab. Every time the user tries the similarity of a different text, the chart changes and it adapts according to the new data.

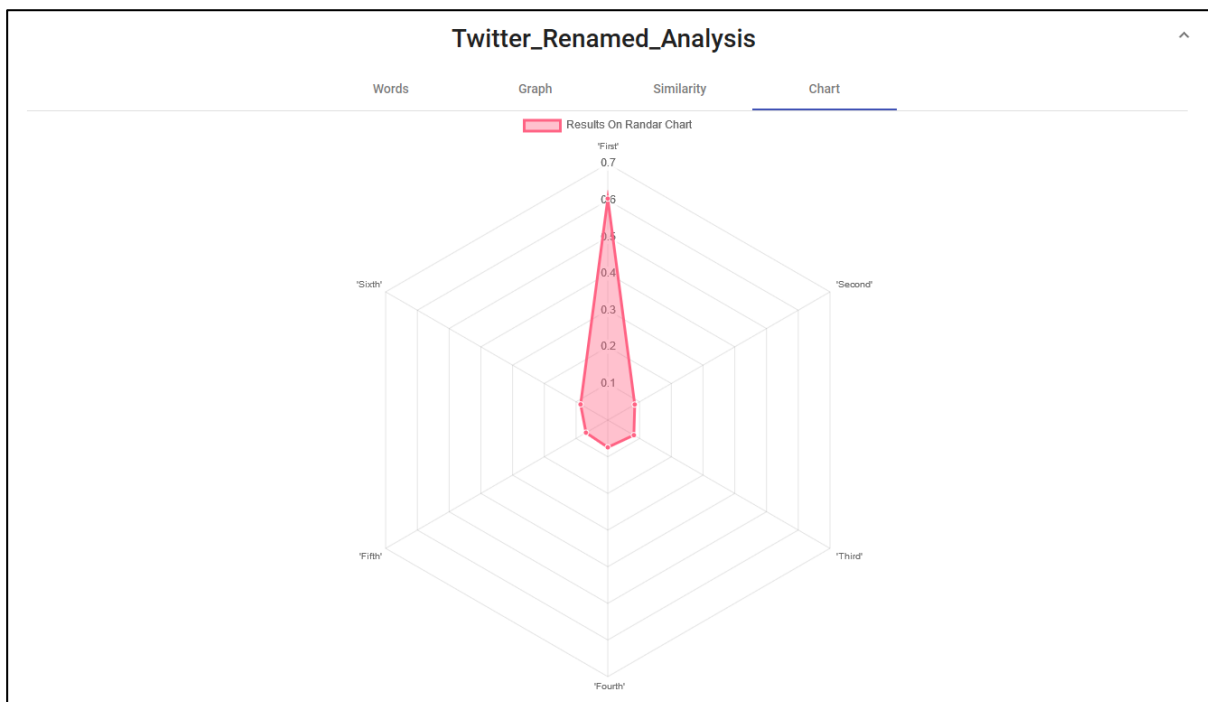
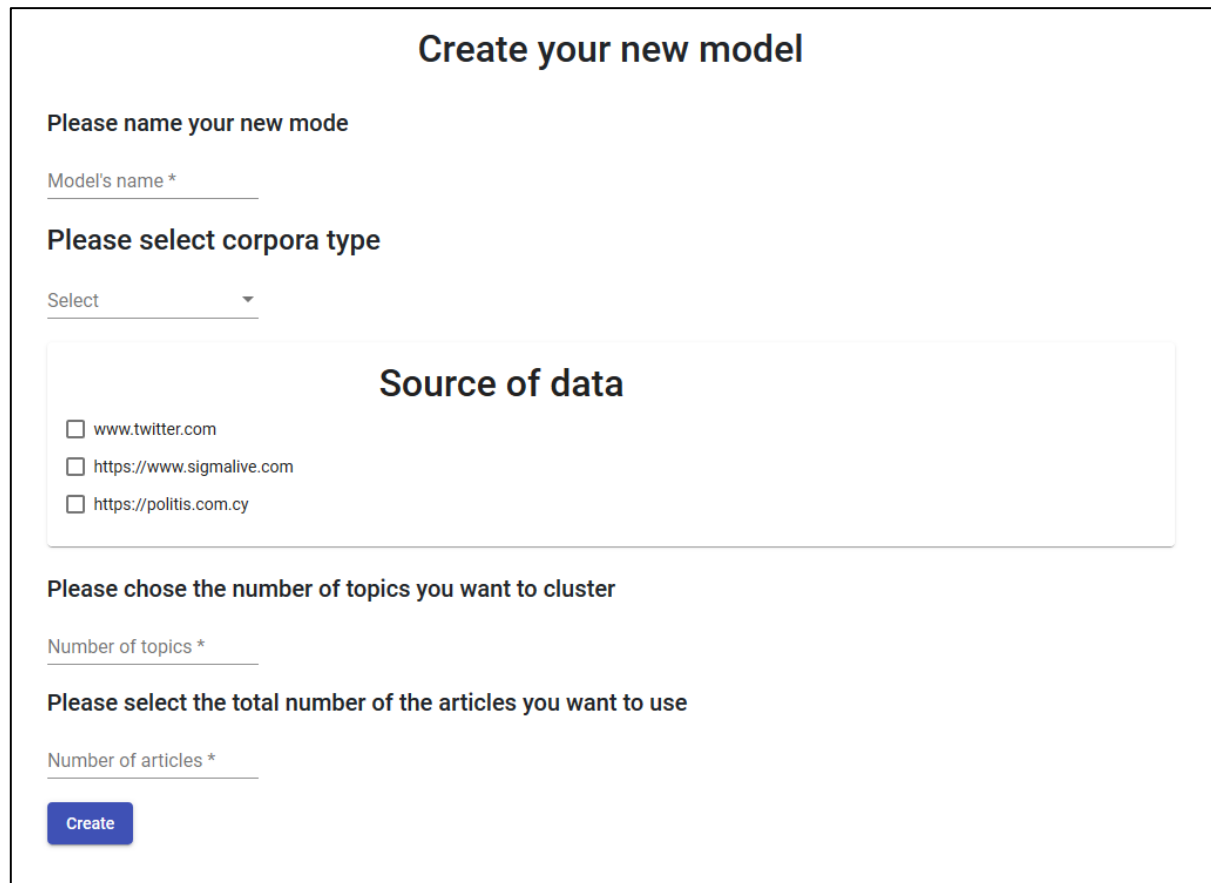


Figure 11: Similarity results on chart

7.3 New Analysis

The last feature provided to the user is the option to create a new analysis based on his/her preferences. The image below (Figure 12) shows the information that users need to set in order to create a new analysis.



Create your new model

Please name your new mode

Model's name *

Please select corpora type

Select

Source of data

www.twitter.com

https://www.sigmalive.com

https://politis.com.cy

Please chose the number of topics you want to cluster

Number of topics *

Please select the total number of the articles you want to use

Number of articles *

Create

Figure 12: New analysis interface

7.3.1 Naming, number of topics and size of the corpus

The creation of a new analysis starts by giving it a name. The name can be anything the user wants; however, it should contain no whitespaces. Whitespaces in the title may cause misbehaviour to the system. To avoid this situation, the spaces in the title are automatically replaced with underscore (“_”) when the new analysis begins.

The second option that the user must set is the type of the corpus. If users want to use bigrams or trigrams, then they must select the corresponding option from the drop-down menu (Figure 13). If nothing is selected, the corpus will be processed without using either bigrams or trigram.

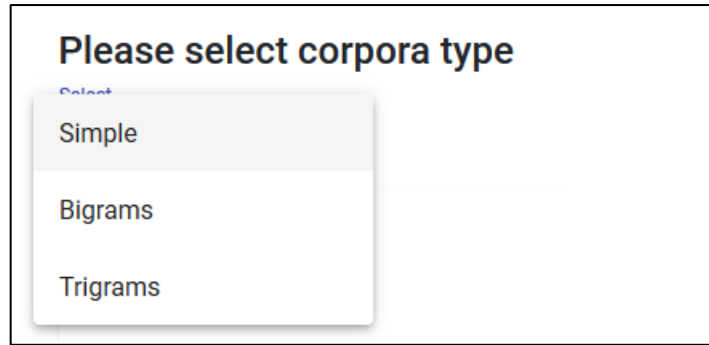


Figure 13: Corpus type



Figure 14: Source of data check boxes



Figure 15: Online newspapers categories

The size of the corpus depends on the data source. Currently, two different types are supported. The first source is the Twitter and the second one is online newspapers. The users must choose the source they want to collect the data from by selecting the corresponding checkboxes (Figure 14). By clicking on the Twitter check box, the system shows two options the users should choose from: a) to search twitter using profiles and b) to search twitter using hashtags. If Twitter is not selected, then the online newspapers must be used. By clicking on the name of the organisation, represented as home URLs (Figure 15), a few more options will appear. Then,

the user must choose the categories of the articles that will be collected. Moreover, it is necessary to write a number which indicates how many articles will be collected.

Having selected the source and number of the articles to be collected, the last step is to declare the number of topics the LDA will create.

7.3.2 Twitter and Online newspapers data collection

Users can be very specific about which part of an online document is to be collected and thus analysed. If they select Twitter as a data source, they can select if the data will be posts made by specific users or the texts that include specific hashtags. Of course, in both cases there should be a limit on the number of tweets to be collected. The limit is set by the users. The image below (Figure 16) shows an example of selecting the last 100 tweets sent by the top six political parties in Cyprus as well as the first 200 tweets that include the hashtag #covid19.

The screenshot shows a web form titled "Source of data". At the top, there are two checked checkboxes: "www.twitter.com" and "Twitter Users" (with "Twitter Hashtags" also checked). Below these, a list of usernames is displayed: "AKEL1926, DIKO1976, edek1969, DISY1976, cygreens, SymmaxiaPoliton". There are instructions: "Please separate the users names with a comma (,)" and "Please make sure that you write the usernames correctly. Otherwise the analysis will fail". A field for "Tweets per account *" contains the number "100". Below this is a large horizontal bar. Further down, the hashtag "#covid19" is entered. Another instruction says "Please separate the hashtags with a comma (,)", and a field for "Tweets per hashtag *" contains the number "200".

Figure 16: Twitter profiles and hashtags selection

On the other hand, the user can also specify which categories of news articles the system will collect. The current version of the developed system allows its users to collect data from only two online newspapers, “https://www.sigmalive.com” and “https://politis.com.cy”. In future releases crawlers of more Cypriot newspapers will be added. As shown in Figure 15, the user

can choose the category of the news articles from the two online newspapers. The difference in the number of categories that appear is due to the fact that the two sources have different structure with different categories. The two online newspapers group the articles they publish in a different way and so there is a difference in the number of the available categories for the user to select.

After the selection of the categories, the system calculates how many articles should be collected from each category. The reason behind this calculation is to avoid any kind of bias if more articles from one source were collected. The algorithm tries to separate the data in a way that each source has the same or very similar number of articles.

8 Restrictions

Although the system designed is working and can display results related to Topic Modelling and Sentiment Analysis, there are some limitations that restrict the performance of the system performance as described below. Future improvements are encouraged and suggested in order to enhance the system, and provide its users with more abilities.

8.1 Problems

8.1.1 Sentiment in Greek language

The model that was used to calculate the sentiment in both cases, for the top 15 words and the text that the user wrote, does not work for the Greek content. To resolve this problem, all text used for sentiment is translated into English. Moreover, the library that was used to translate the text into English has restrictions on the number of words to be translated per day. In other words, the system cannot calculate the sentiment of all words in each topic. Also, the user cannot calculate the sentiment of a text for an unlimited amount of times.

8.1.2 Sentiment analysis of each document

The way the system was developed and implemented the sentiment of each separate document is not calculated. The sentiment and subjectivity levels of the documents are not stored in the database. Consequently, it is not possible to calculate the overall sentiment level of each cluster created by using the average sentiment value of the assigned documents to each topic.

8.1.3 High perplexity due to the Greek Language

The system was developed using the Python programming language. In Python there is an enormous number of libraries that the developers can use to program what they need.. However, not all libraries are available in every spoken language. The available tools for the Greek language do not work well. For example, the available libraries for word stemming, the transformation of the words into their root form, is creating words that do not exist in the Greek language

8.2 Future work

Even though there are many improvements that can be made to improve the performance and the features of the system, at the current stage three suggestions are made for future work.

8.2.1 Not possible to create new analysis using existing data

The way the system currently works involves collection of data in real time and then processing and analyses. It is not yet possible for users to apply analysis on stored data. Although the proposed architectures support this feature, it is not yet implemented.

8.2.2 Not possible to create model using historical data

Once again, the data used for each analysis are collected in real time from online sources. When the system collects new data, by default, it collects the last published articles or posts by the sources. In the future, a feature is proposed to let users specify between which periods of time or between which dates the system should collect data.

8.2.3 Topics' sentiment value

Currently the sentiment analysis is not the best possible one. In future work, an update in the online system is proposed in order to store into the database the required information regarding the sentiments of each document used in the LDA. The main aim for this update is to enable the calculation of the average sentiment of all the documents assigned to each topic and therefore to find out the average sentiment of each topic.

9 Bibliography

- A Fast and Powerful Scraping and Web Crawling Framework. (n.d.). Retrieved May 9, 2020 from <https://scrapy.org/>
- About Twitter's APIs. (n.d.). Retrieved April 28, 2020 from <https://help.twitter.com/en/rules-and-policies/twitter-api>
- Basset, L. (2015). *Introduction to JavaScript Object Notation*. Beijing: O'Reilly
- Blassing, S., Engesser, S., Ernst, N., & Esser, F. (2019). Hitting a nerve: Populist news articles lead to more frequent and more populist reader comments. *Political Communication*, 1–23.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Borromeo, R. M., & Toyama, M. (2015). Automatic vs. crowdsourced sentiment analysis. *ACM International Conference Proceeding Series*, (CONFCODENUMBER), 90–95. <https://doi.org/10.1145/2790755.2790761>
- Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230. <https://doi.org/10.1111/j.1083-6101.2007.00393.x>
- Budak, A. (2010). *Facebook, Twiter and Barack Obama: New Media and the 2008 Presidential Elections*. Georgetown University.
- Castells, M. (2004). The Information Society Reader. In *the Information Society Reader* (First, pp. 138–149). London: Routledge.
- Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94 3rd Annual Symposium on Document Analysis and Information Retrieval.
- Chikersal, P., Poria, S., & Cambria, E. (2015). *SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning*. (SemEval), 647–651. <https://doi.org/10.18653/v1/s15-2108>
- Clemence, A., Doise, W., & Lorenzi-Cioldi, F. (2013). *The quantitative analysis of social representations*. New York: Routledge.

- Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *Fifth International AAAI Conference on Weblogs and Social Media*, 8(2), 125–136. <https://doi.org/10.1023/A:1018556721104>
- Dubras, R., Kemp, S., & Navarro, P. (2020, February 11). Digital 2020: The United States – what you need to know. Retrieved April 27, 2020, from <https://wearesocial.com/us/blog/2020/02/digital-2020:-the-united-states---what-you-need-to-know>
- Enli, G. (2017). Twitter as arena for the authentic outsider: exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election. *European Journal of Communication*, 32(1), 50–61. <https://doi.org/10.1177/0267323116682802>
- Francia, P. L. (2018). Free Media and Twitter in the 2016 Presidential Election: The Unconventional Campaign of Donald Trump. *Social Science Computer Review*, 36(4), 440–455. <https://doi.org/10.1177/0894439317730302>
- Fuchs, C. (2017). Fascism 2.0: Twitter Users’ Social Media Memories of Hitler on his 127th Birthday. *Fascism*, 6(2), 228–263. <https://doi.org/10.1163/22116257-00602004>
- Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *Processing*, 1–6.
- Golbeck, J., Grimes, J. M., & Rogers, A. (2010). *Twitter Use by the U.S. Congress*. 61(8), 1612–1621. <https://doi.org/10.1002/asi>
- Graham, S., Weingart, S., & Milligan, I. (2012). Getting Started with Topic Modelling and Editor’s Note. *The Editorial Board of the Programming Historian*.
- Heiss, R., & Matthes, J. (2019). Stuck in a Nativist Spiral: Content, Selection, and Effects of Right-Wing Populists’ Communication on Facebook. *Political Communication*, 1–26.
- Holtz-Bacha, C. (2007). Professionalisation of Politics in Germany. In *the Professionalisation of Political Communication*. Bristol: Intellect.
- Hosch, W. L. (2019, October 8). Machine learning. Retrieved April 28, 2020, from <https://www.britannica.com/technology/machine-learning>
- Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., & Mazaid, M. (2015). Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? *SSRN Electronic Journal*, 1–30. <https://doi.org/10.2139/ssrn.2595096>

- Kaid, L. L. (2004). *Handbook of Political Communication Research*. New Jersey: Lawrence Erlbaum Associates.
- Kalsnes, B., Larsson, A., & Enli, G. (2017). The social media logic of political interaction: Exploring citizens' and politicians' relationship on Facebook and Twitter. *First Monday*, 22(2). <https://doi.org/10.5210/fm.v22i2.6348>
- Kemp, S. (2019). Digital 2019: Global Internet Use Accelerates. Retrieved November 23, 2019, from <https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates>
- Kriesi, H. (2014). The Political Consequences of the Economic Crisis in Europe: electoral punishment and popular protest. *Mass Politics in Tough Times*, 297–333.
- Larsson, A. O., & Moe, H. (2014). Twitter in Politics and Elections. In *Twitter and Society*. New York: Lang Publishing.
- Lassen, D. S., & Brown, A. R. (2011). Twitter: The electoral connection? *Social Science Computer Review*, 29(4), 419–436. <https://doi.org/10.1177/0894439310382749>
- Liu, B. (2010). *Sentiment Analysis: A Multi-Faceted Problem*. (1).
- Lloyd, S., Mohseni, M., & Rebstrost, P. (2013). *Quantum algorithms for supervised and unsupervised machine learning*. 1–11. Retrieved from <http://arxiv.org/abs/1307.0411>
- Loria, S. (2018). textblob Documentation. *Release 0.16.0*. Retrieved from <https://textblob.readthedocs.io/en/dev/index.html>
- Makice, K. (2009). *Twitter API: up and running*. Farnham: O'Reilly.
- Manning, C., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval*. Cambridge: Cambridge University Press.
- McNair, B. (2018). *An Introduction to Political Communication*. New York: Routledge.
- Nasukawa, T., & Yi, J. (2003). Sentiment analysis: capturing favorability using natural language processing. *Proceedings of the 2Nd International Conference on Knowledge Capture*, 70–77. <https://doi.org/10.1145/945645.945658>
- Negrine, R., Holtz-Bacha, C., Papathanasopoulos, S., & Mancini, P. (2007). *The Professionalisation of Political Communication*. Bristol: Intellect.

- Newman, D., Smyth, P., Welling, M., & Asuncion, A. U. (2008). Distributed inference for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 1081–1088.
- Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, 12(2–3), 140–157. <https://doi.org/10.1080/19312458.2018.1455817>
- Parmelee, J. H. (2014). *Political journalists and Twitter: Influences on norms and practices*. 2753. <https://doi.org/10.1386/jmpr.14.4.291>
- Perrin, A. (2015). *Social Media Usage: 2005-2015*. (October), 2005–2015. Retrieved from www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/.
- Peruzzi, A., Zollo, F., Quattrocchi, W., & Scala, A. (2018). How News May Affect Markets' Complex Structure: The Case of Cambridge Analytica. *Entropy*, 20(10), 1–12. <https://doi.org/10.3390/e20100765>
- Rainie, L., Smith, A., Schlozman, K. L., Brady, H., & Verba, S. (2012). Social Media and Political Engagement. *Pew Internet {&} American Life Project*, 1–13.
- Roesslein, J. (2009). tweepy Documentation. *Online*] <http://tweepy.readthedocs.io/en/v3.5>.
- Singer, J. B. (2003). Campaign contributions: Online newspaper coverage of election 2000. *Journalism & Mass Communication Quarterly*, 80, 39–56.
- Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 1310–1315. New Delhi.
- Small, T. A. (2011). What the hashtag?: A content analysis of Canadian politics on Twitter. *Information Communication and Society*, 14(6), 872–895. <https://doi.org/10.1080/1369118X.2011.554572>
- Steinfeld, C., Ellison, N. B., & Lampe, C. (2008). Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 29(6), 434–445. <https://doi.org/10.1016/j.appdev.2008.07.002>

- Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*, 3(4), 1277–1291. <https://doi.org/10.1007/s13278-012-0079-3>
- Triga, V., Mendez, F., & Djouvas, C. (2019). Post-crisis Political Normalisation? The 2018 Presidential Elections in the Republic of Cyprus. *South European Society and Politics*, 24(1), 103–127. <https://doi.org/10.1080/13608746.2018.1511347>
- Udapure, T. V., Kale, R. D., & Dharmik, R. C. (2014). Study of Web Crawler and its Different Types. *IOSR Journal of Computer Engineering*, 16(1), 01–05. <https://doi.org/10.9790/0661-16160105>
- Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, 18(1), 45–55.
- Yaqub, U., Chun, S. A., Atluri, V., & Vaidya, J. (2017). Analysis of political discourse on twitter in the context of the 2016 US presidential elections. *Government Information Quarterly*, 34(4), 613–626. <https://doi.org/10.1016/j.giq.2017.11.001>