INTERNATIONAL
HELLENIC
UNIVERSITY

Text Mining in Twitter with Spark and Scala

Twitter as Political Barometer in Greece

Adam Simitos

SID: 3306150011

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Mobile and Web Computing

DECEMBER 2016

THESSALONIKI – GREECE

Text Mining in Twitter with Spark and Scala

Twitter as Political Barometer in Greece

Adam Simitos

SID: 3306150011

Supervisor:	Prof. Apostolos Papadopoulos
Supervising Committee Members:	Prof. Christos Berberidis
	Prof. Apostolos Ampatzoglou
	Prof. Marios Gatzianas

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Mobile and Web Computing

DECEMBER 2016

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in “Mobile and Web Computing” at the International Hellenic University, Thessaloniki, Greece. Text Mining is a research area that tries to solve the document overabundance problem by using Data Mining, Machine Learning, Natural Language Processing, Information Retrieval, and Knowledge Management techniques. Text Mining’s main purpose is the automate documents categorization in classes.

People’s thoughts and opinions have always been studied and researched by the sciences of sociology and history. Social Media revolution has made opinion expression a very easy, simple and quick procedure. Thanks to Social Media an Internet user can propagate their opinion and read other users’ opinions as well. As a result, the Internet is “flooded” by a vast volume of data that is difficult to be managed. Social Media is one of the factors that contribute to the phenomenon called “Big Data” in computer science.

The object of this master thesis is the collection and manipulation of social media users’ opinions about political situation in Greece by using text mining methods. Specifically, the application developed crawls opinions for Greek parliament members from Twitter social medium and categorizes them in positive, neutral, and negative. Statistics produced are indicative for each member’s popularity.

Adam Simitos

23/12/2016

Contents

List of Pictures

List of Tables

1 Introduction

The perpetual increase of data volume produced by enterprises, organizations, multimedia, and social networks daily, has led computer science to new technologies development for easy and efficient data management. Knowledge extraction from complex, digital data ("Big Data") requires powerful processors and effective programming tools. Big data analysis uses high performance machines and intricate data mining algorithms in order to generate results timely. Spark (platform for distributed computer systems) and Scala (functional programming language) are two upcoming tools that were developed for Big Data analysis and management [1,2].

People's beliefs and thoughts have always been important information for decision making procedure. For example, a company wants to know customers' opinions about its products, in order to define the marketing strategy. Thanks to the Internet and social media networks, individuals have the opportunity to express themselves and share their opinions for various subjects and topics. Therefore, a huge volume of opinions that would be useful to be taken advantage of, is available in the Internet in the form of text data.

Twitter is one of the most popular social networks worldwide. It is a microblogging service, where users publish short messages called "Tweets". These tweets contain users' opinions and thoughts about different topics [3]. Twitter is now an evolving service for social and political exchanges. Most Twitter users post tweets expressing their opinions about politicians. The subject of this master thesis is the sentiment analysis of tweets that are aimed at Greek parliament members. Text mining techniques are applied in order to classify tweets into positive, neutral and negative according to their content. Sentiment classification has become a ubiquitous technology in the world of Twitter and it is used for many reasons.

The development of a system that extracts knowledge from a tweet is a quite difficult procedure, since the same sentiment can be expressed with a plenty of different words. Besides, ironic tweets make procedure more complicated, because words with positive meaning are used in negative context.

The structure of this dissertation is the following:

Chapter 2 introduces basic concepts of Big Data and modern technologies for Big Data management.

Chapter 3 outlines Twitter as one of the greatest social networks presenting basic functionalities and technologies used. It also describes its use for political purposes.

Chapter 4 contains mathematical modeling of text mining. It explains the mathematical formulas used for text mining analytically and the fundamental theory for document classification.

Chapter 5 describes the whole methodology followed for tweets collection, classification evaluation and mobile application development.

Chapter 6 presents all conclusions extracted from this research.

Chapter 7 mentions some topics for further research.

2 Big Data

In 2000 the telescope of the SDSS program (Sloan Digital Sky Survey), located in New Mexico City, produced more data than has ever been produced in the whole astronomy history in few weeks. After a decade of continuous work, the telescope scanned 14.555 square degrees, which are the 35% of the sky approximately, and produced 140 terabytes of data in image format. The new telescope that is located in Chile, produced the same volume of data in only 5 days [4]. Facebook users upload 350 million photos every day. All these photos are added to the already uploaded photos and as a result Facebook has now more than 240 billion photos [5].

People nowadays produce a huge quantity of digital information. Most of this information is produced by the Internet users and it is spread owing to social media instantly. The volume of data increases exponentially and data scientists are obliged to manage this data in a different way compared with the past. Therefore, the harness and the effective management of this growing information are issues of vital importance.

Furthermore, this information is characterized by an intense diversity, since it describes every aspect of society and every day routine. Data that is not restricted in volume and variety is called “Big Data”.

Examples of Big Data are:

- 340 million tweets are posted daily in Twitter, which are 4.000 tweets/ second
- More than 1 billion Facebook users interact each other
- 50 GB of video are uploaded in YouTube every minute
- 10.000 transactions with credit cards are made every second
- More than 5 billion subscribers make calls or video calls using their cell phones
- Data produced by enterprises will being doubled every 1,5 year
- The American multination company “Wal-Mart” makes more than 1 million transactions every hour. These transactions are stored in databases that contain 2.560 terabytes data; 167 times more data than this stored in the USA Congress library database. USA Congress library is considered as the biggest library in the world [6].

2.1 What is Big Data

The term “Big Data” is still an abstract concept, however it denotes enormous datasets generally. Datasets that cannot be captured, discerned, manipulated and processed by traditional computer technologies (software and hardware) within bearable time [7]. Big data management is one of the biggest challenges in computer science nowadays. Big data is extremely difficult to be analyzed and processed by the conventional RDBMS[1], since it is big in volume, is spread very quickly and don’t match with the architecture of the transitional RDBMS.

Big Data have the following three characteristics:

- Big volume of data: For example, the marketing department of a company is called to analyze terabytes of messages, posted by customers, in only one day in order to understand their reactions about new products.
- Big data spread velocity: Few seconds can be a long time span if there is time sensitive procedure such as information analysis for financial fraud. Millions of bank accounts and transactions must be scrutinized in real time for fraud detection and prevention.
- Big variety of data: Big data is not necessary to be in text format. Big data can be images, videos, audio, clickstreams (clicks that a user does during web navigation), data from sensors etc.

The three characteristics above are known as “the three V of Big Data” (volume, velocity, variety). Three Vs are fundamental elements of big data, since they define the nature of information and by extension the nature of information management.

Not only are computer scientists interested about big data, but politicians and entrepreneurs also. That happens because big data is an inexhaustible source of information, which hide a lot of knowledge [6].

2.2 Big Data Challenges

Mining and analyzing big data is an extremely difficult and complicated procedure. Since big data is unstructured data, because of its volume and diversity, the analysis is a time consuming procedure.

[1] Relational Database Management System

Moreover, most enterprises do not have the necessary infrastructure for big data analytics. Hewlett Packard made a world survey in 2013 and discovered that organizations don't have the suitable equipment for big data analysis and their staff is not trained for a concept like this. Another HP's survey in 2013 disclosed than one in three companies failed to perform big data analysis successfully.

Even though big data analysis is very possible to fail, 60% of organizations are willing to spend more than 10% of their budget for this. HP's CIO said that enterprises are benefited by big data analysis in real time. Decision taking is completed quickly making services and products better [6].

2.3 Managing Big Data

2.3.1 Spark

Big data cannot be processed by a single machine with a simple database software. This is why distributed computing technologies have been created in order to fulfill that need. Apache Spark is a unified, open source, cluster computing platform, which performs efficient big data analysis. It was developed at the University of California and released on May 2014. Spark is designed for clustering and parallel programming providing a single interface [8].

Spark constructs a data structure called RDD (Resilient Distributed Dataset). RDD is a read- only collection of items distributed over a cluster of machines that can be manipulated in parallel. If a RDD partition is destroyed, it can be rebuilt making Spark fault tolerant. All lost work is recovered and delivered immediately with no extra code (code reuse) or configuration.

Spark was developed in order to overcome the MapReduce[2] limitations [9,10,11]. It provides more types of computations and accomplishes better performance in complex applications than MapReduce [11]. Speed is an important factor in big data projects. RDDs make the application of iterative algorithms to large datasets better and more efficient. Training algorithms for data mining and analysis (Machine Learning) are iterative

[2] MapReduce is a programming pattern designed for processing big datasets in parallel way in one cluster. A program written in a *MapReduce* pattern performs filtering and sorting using the *Map* procedure and summary operation using the *Reduce* procedure. MapReduce organizes data transfer and communication among the several parts of the system guarantying redundancy and fault tolerance. However, MapReduce has been criticized regarding its spectrum of problems that can solve. It has been accused of low- level data manipulation and weakness in speed and performance [9,10,12,13,14,15,16].

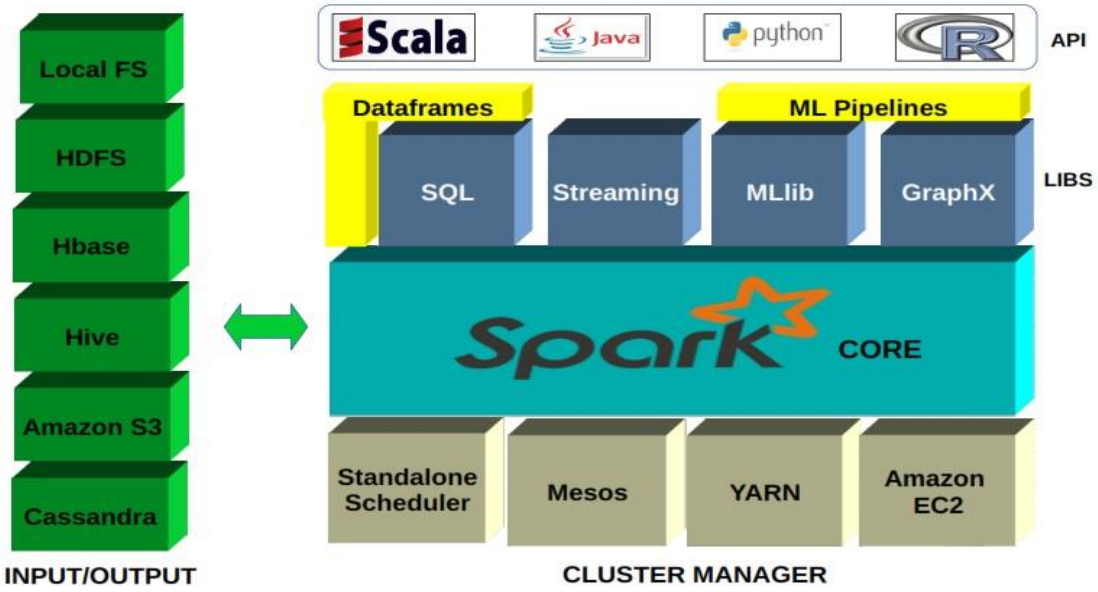
algorithms. These algorithms were the motivation for Apache Spark creation [17]. Furthermore, Spark achieves faster speed by executing computations into memory across the cluster [11].

In general, Spark is constructed to execute large datasets that previously need many distributed systems. Besides, Spark supports great object- oriented languages like Java, Python, Scala and R providing a simple API and many built- in libraries. It is, also, compatible with Hadoop[3] clusters and Hadoop data source [11].

Spark stack

Spark integrates multiple components. Spark's core schedules and distributes applications across multiple machines (clusters) and monitors them. Core is fast and provides high- level components suitable for a great variety of workloads (such as machine learning). The main idea of Spark's architecture is the tight integration. Low- level component can benefit high- level ones. For example, an improvement in Spark core affects SQL libraries positively.

What is more, costs can be reduced. An organization can run only one software system instead of several independent systems. These costs include all software lifecycle costs such as analysis, development, testing, maintenance etc. Evert time Spark creates a new component, every company or organization working Spark is able to integrate it [11,19].



Picture 1: Spark stack [20]

[3] Hadoop is an open source cluster computing platform, like Spark. It runs jobs in distributed and parallel systems using MapReduce process [18].

The most powerful advantage Spark offers is application building that combines many different processing models. For example, it is possible an application to collect and classify data real time. At the same time, the produced data can be queried through SQL and be accessed by Python or Scala also in real time. The IT department of the company has to support only one system [11].

Spark Core

As the names says, it is the core, the central unit of the system. Core is the base of the whole project. All the basic functionalities like memory management, distributed task scheduling and communication with other clusters (I/O functionalities) are implemented in here. The core also provides the API for RDD construction and manipulation.

Spark SQL

Spark SQL is the Spark's package (library) that supports relational databases and SQL queries. It also allows combination of SQL queries with Python, Java or Scala for RDDs management.

Spark Streaming

Spark streaming component enables live data streams processing. Live data streams are log files produced by web or application servers or messages sent by web service users. A suitable API is supported for data streams manipulation providing fault tolerance and scalability [11].

MLlib

One of the most important libraries offered is MLlib. MLlib contains all the basic Machine Learning functionalities and high quality algorithms for data import, classification, clustering, regression, collaborative filtering, association rules and model evaluation. Thanks to iterative computation, MLlib runs 100 times faster than MapReduce.

GraphX

GraphX is the library for graph mining and manipulation. A graph can be a Twitter's user network (followers). GraphX performs graph- parallel computations and constructs directed graphs very fast thanks to build- in graph algorithms. Data can be viewed as both graphs and collections.

Cluster Managers

Spark is famous for its scalability capabilities from several to thousands of computers. Spark achieves this by running over a big diversity of cluster managers like Apache Mesos and Hadoop. Spark's cluster manager is called *Standalone Scheduler* [11].

Spark Runtime Architecture

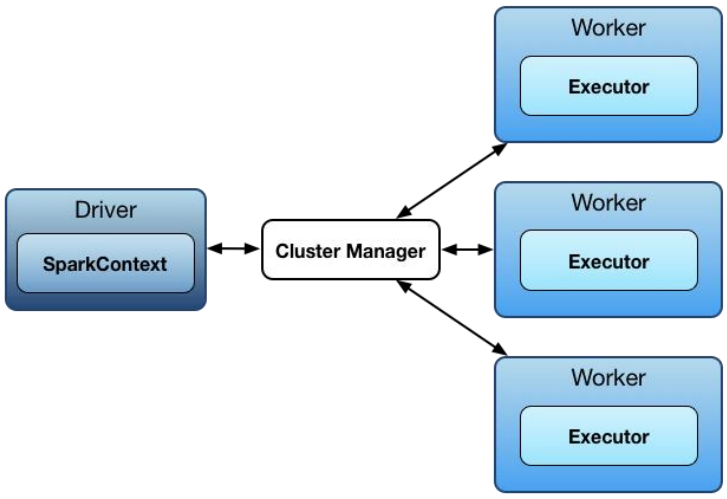
The main Spark's advantage is the computation scalability offered. More machines can be added in Spark system, which will run in cluster mode parallelly. This property makes Spark very flexible in applications execution. Since the Spark stack was described above, it is important the runtime architecture in distributed mode to be clarified.

As depicted in picture 2 the distributed mode of Spark is based on master/ slave architecture. The master is called *driver* and it is the central coordinator. Slaves are many distributed machines (workers) called *executors*. Driver and executors communicate each other through the *cluster manager*. The driver and each executor run their own java process (JVM) [11].

The Driver

Driver is the process that runs the main() method of each program. Driver grant access to Spark using a *SparkContext* object, which performs the connection with cluster. Driver has two responsibilities: a) Converts program into tasks and b) Schedules tasks on executors.

Spark driver splits a program into tasks. Tasks are execution units. Tasks are the smallest work units in Spark. A simple program may consist of thousands of individual tasks. The procedure is always the same: RDDs creation from input, new RDDs production and actions performed for data collection. Furthermore, Spark produces a Directed Acyclic Graph (DAG) of functions. DAG schedules tasks executions. Every DAG is a set of stages and each stage contains multiple tasks [11].



Picture 2: Distributed Spark application components [21]

Furthermore, driver schedules and assigns tasks on executors. Driver always has a complete view of all executors. Driver decides the executor to which tasks will be assigned. Tasks execution produces data that should be stored in a cache memory. Driver detects cached data and schedules new tasks that will access it [11].

Executors

Executors run individual tasks assigned by driver. Executors start working at the beginning of Spark application. Executors have 2 main roles: Firstly, they execute application tasks assigned by driver and send results back to it. Secondly, they have inner memory for RDDs caching.

Cluster Manager

Cluster manager mediates between Driver and executors. Cluster manager turns executors on and in some cases, does the same to driver. Besides, it helps Spark run with several external managers like YARN and Mesos [11].

2.3.2 Scala

Scala is a general- purpose programming language that supports both object- oriented and functional programming in order to provide high performance for big data projects. Scala runs on Java Virtual Machine (JVM) platform. Scala was first released in January 2004 by the Ecole Polytechnique Federal de Lausanne (EPFL), Switzerland. The name “Scala” stems from the words “SCAlable LAnguage" [23].

Picture 3: Scala logo [22]

Scala is based on the following concepts:

- Statically typed: Scala do not change the type of a variable during the runtime of a program (immutable variables).
- Object- oriented programming: Everything in Scala is object. All numeric types are objects. However, Scala sometimes uses primitive types in the runtime for performance reasons. Moreover, Scala implements traits. Traits are mixin[4] classes.
- Functional programming: Functional programming is a programming concept older than OOP, but it was sidelined. Functional programming has become “in-fashion” again, because it simplifies many problems like synchronization. Moreover, functions can be passed to other functions or be assigned to variables. Since everything in Scala is an object, functions are also objects.

[4] Mixin is a term of OOP. A class is called mixin class if it includes methods (functions) from other classes without having inheritance relationship with each other [24].

- JVM & .NET language: Scala supports both JVM and .NET. As JVM language can generate JVM byte code, whereas as .NET based language can produce Common Language Runtime (CLR). As a result, a Scala programmer can use libraries and interact with languages that are supported by these platforms.
- Succinct, elegant and flexible syntax: Scala is very concise language; a lot of functionality with the minimum commands. A function can be defined in another function and the name may include non- alphanumeric characters.
- Sophisticated type system: Scala has more generics[5] and typing constructs than java.
- Scalable architecture: Scala supports four mechanisms that foster the scalable development of systems: a) clear self-types, b) nested classes and functions, c) abstract type members and generics and d) traits.
- Scalable performance: Since Scala runs on JVM and CLR, it takes advantage from performance improvements provided by those platforms [23,25].

[5] "*Generics extend Java's type system in order to allow a type or method to operate on objects of various types while providing compile- time type safety"* [26].

3 Twitter

The term “Social Media” refers to message and information exchange through several internet communities. Social Media offer a new type of digital socialization, through which internet users are shared text, images and videos. As we can understand, social media provide huge quantity of information that can be harnessed for social, political, commercial and financial reasons.

Picture 4: Twitter logo [27]

This can be understandable if we consider that millions of messages are published in Twitter every day. Twitter is a world social networking platform (micro- blogging service), where users can publish short 140- character texts, called “Tweets”. Tweets are restricted in 140 characters in order to look like the short messages of mobile phones. This restriction makes tweets easy to be read [28,29].

Users can write their feelings or opinions in a single tweet any time they want. Users, who want to watch another user’s tweets, should “follow” him. These users are called “followers” of the current user and constitute his/her network. Followers are notified every time the specific user makes a tweet. A user can answer to a specific tweet by clicking on “Reply” icon, publish it to his own network of followers by clicking on “Retweet” icon or like it by clicking on “Like” icon.

Twitter started its operation in July 2006. It has more than 300.000.000 users, 4.000 employees and it is considered as one of the 10 most successful social media websites worldwide [30]. Over 400.000.000 tweets are published every day globally [28].

Twitter was developed with “Ruby on Rails” framework and has its own API (Application Programming Interface). Jack Dorsey was the first person, who was inspired Twitter in 2005. Dorsey’s vision was the development of a web service, where his friends would be able to publish their habits, thoughts and actions. Twitter was established by a company called “Obvious” located in San Francisco, USA [31].

Nowadays, Twitter is a prominent communication medium thanks to its speed and ease of use [28]. Since Twitter has millions of users around the world, all opinions are spread

immediately formulating ways of thinking. Twitter has played an important and fundamental role in social, political, art and sports events like Arab Spring (2010- 2012), Turkish coup attempt (Summer 2016), Olympic Games 2012 and 2016, elections, concerts etc.

It is said that Twitter is a powerful tool for historians of the future, since each fact is described by many perspectives. In past, the only sources a historian had, were newspapers and written documents, which were not absolutely objective informing. This is why social media are ideal for researches and surveys [32,33].

3.1 Twitter Analytics

Twitter offers a statistic tool called "Analytics" to all users. Analytics tool produces useful statistics about user's activity in Twitter. Users can be informed about the success of their tweets and how these tweets affect other users. Besides, they can explore their followers' interests, needs and geographic locations.

Twitter Analytics measures the success of a users' tweets and guide them how to make them more successful. A dashboard is offered which depicts how user's network is affected by published tweets using charts. For example, users can evaluate the success of their tweets by viewing the number or replies, retweets, likes, clicks or impressions tweets received.

Users are also able to view the number of mentions (how many times users mentioned them), profile visits and new followers they have per month or totally. This is very practical, since it indicates if people react to a single tweet and how they react. Additionally, Twitter Analytics provide details about how the tweet content is shared through all Twitter users. All these metrics are available for downloading for further processing (in excel, statistica or other software programs) [34].

3.2 Crawling Twitter Data

As mentioned above, Twitter produces millions of tweets every day. Twitter provides some of these tweets to Data Scientists, researchers and developers through its API for free. Twitter's API is divided in two categories: a) REST API and b) Streaming API. The REST API uses the *pull* method for data crawling. If a user wants to collect tweets, he must request for them. The Streaming API is suitable for real time data collection. Once a user requests for data, Streaming API offers tweet perpetually. Requests to Twitter API

can contain keywords, hashtags[6] or mentions (Twitter usernames starting with @). Twitter responses are in JSON (JavaScript Object Notation) format, which is a popular format in web.

Of course, Twitter responses only to authenticated requests. *Open Authentication* is a mechanism developed by Twitter, which sends user the necessary credentials in order to validate the requests. Also, Twitter sets a maximum the number of requests within the time window. This is called *rate limit*. If a user reaches this rate limit, Twitter stops sending responses and user has to wait for 15 minutes till the next request [28].

3.2.1 Open Authentication

Twitter has adopted the *Open Authentication* (OAuth) mechanism in order to provide authorized access to its API. OAuth is an open standard for user authentication. Since passwords are considered powerless and easy to be theft, OAuth affords a three- way handshake making communication and authentication safer. Furthermore, the user's confidence in the application is improved, as user's credentials are not notified to third- party services or applications.

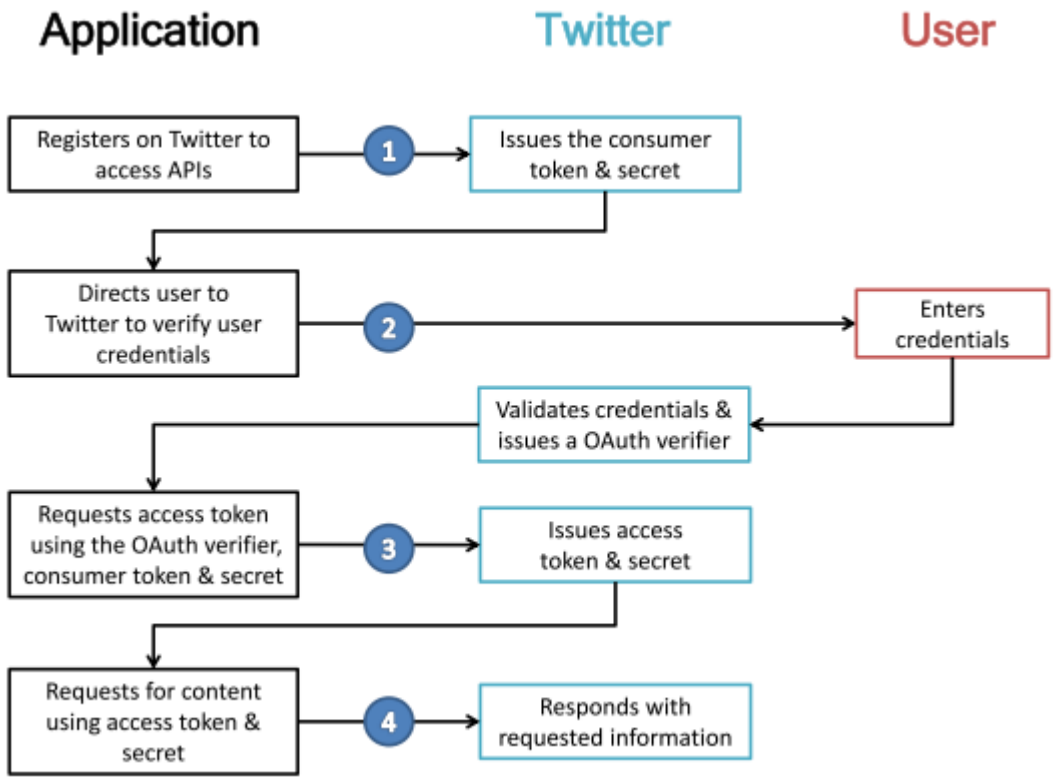
As it is reasonable, API requests are authenticated by OAuth. Figure 5 depicts the OAuth mechanism for Twitter API access.

1. OAuth provides application with keys and tokens. Specifically, application asks for access to Twitter API. The application is "equipped" with "Consumer Key" (API Key) and "Consumer Secret" (API Secret). Twitter authenticates application checking the Consumer Key and Consumer Secret and creates a link between them (API and application).
2. After establishing the link its time for user to be authenticated. User enters his Twitter account credentials in order to validate the application and authenticate himself. Twitter validates credentials and produces OAuth verifier.
3. User provides the OAuth verifier to the application. Application requests Twitter for the "Access Token" and "Access Token Secret" using the OAuth verifier. Access Token and Access Token Secret are unique for every user.

[6] The word "Hashtag" stems from the union of the words "Hash" (aka the symbol #) and the word "Tag" (aka label). An example of Hashtag is #Greece or #Brexit. Twitter first introduced hashtags in 2007 and since then they are used daily in order to categorize internet discussions and posts. A hashtag indicates the topic of a post/ tweet. Usually, hashtags refer to events, celebrities, TV programs, places or describe situations like love, funny etc. [35]. A hashtag is also a hyperlink. If a user clicks on it, all posts/ tweets containing the current hashtag will be returned to user's device.

4. After validating Access Token and Access Token Secret, application authenticates user on Twitter. Then, it makes requests to Twitter on behalf of the user.

All the aforementioned keys and tokens can be changed if the user asks for it [28].



Picture 5: OAuth workflow [28]

3.2.2 Collecting search results

Search queries to Twitter API can contain hashtags, simple keywords or phrases, usernames and userIDs.

Collecting tweets using REST API

Twitter disposes the *GET search/tweets* API in order to deliver the desired tweets. GET search/tweet API returns a compilation of relevant tweets matching a specific query. As Twitter informs us in developers documentation, it does not perform exhaustive search. Only the most recent tweets of the past 10 days are returned. The rate limit of the REST API in this case is up to 450 requests in the time window or up to 180 requests for a single user [28,36].

Collecting tweets using Streaming API

Streaming API searches tweets based on hashtags, keywords, usernames perpetually. Once a tweet is published, it is delivered to the interested user immediately. Streaming

API supports two methods: *GET statuses/sample* and *POST statuses/ filter*. The GET method chooses only a small random sample of tweets (statuses), whereas the POST method chooses all tweets matching the search criteria [36].

Steaming API sets limit on the parameters that can be applied in a query. For example, maximum 400 keywords and 5.000 usernames are acceptable. As happens in REST API, Steaming API does not return all tweets ever published. The rate limit in this case is the 1% of the total tweets posted on Twitter [28].

3.3 Tweets Sentiment Analysis

In most cases we don't care about what people say, but how they say it. Sentiment analysis is the procedure that calculates a sentiment (emotional) score for a given text. This score can be a positive or negative. Sentiment analysis helps analysts understand how people react in a given topic.

Sentiment analysis is performed on a per- tweet basis. Words containing in a tweet are compared and contrasted with words of other tweets, which have been previously labeled as "negative" and "positive". If the words in a current tweet are appeared more in positively labeled tweets, then this tweet is characterized as "positive". In this way, the algorithm concludes if a given tweet is positive or negative.

A lexicon is necessary for the algorithm. Lexicon is a dictionary of words, where every word is assigned a positive and a negative score. The choice of the suitable source that will build the lexicon demands thoughtfulness, since words can have different sentimental meanings in different contexts. For example, the word "bomb" may have positive meaning in a film review ("That film was bomb") and negative in a journal article ("Bomb was exploded…") [28,37].

3.4 Twitter and Politics

3.4.1 Twitter for political communication

Twitter is an immediate, fast and simple way for political communication and commending. It, also, promotes dialogue. The short nature of texts (only 140 characters as mentioned above) make users (politicians and voters) laconic. They don't chatter and as a result tweets are full of meaning. Furthermore, Twitter gives the opportunity to users to

choose the moment they like to transmit tweet, unlike the traditional media like TV or newspapers [38].

Twitter allow voters connect, communicate and interact with politicians in a unique way that was never possible before. It is able to influence voters and elections results and this is why all modern politicians maintain a Twitter account, so that they can express themselves and approach potential voters. Voters react and answer making themselves members of e- democracy society. It is very common a single tweet to replace an official press release of a political party [39].

Twitter was first used for political communication during the USA presidential elections in 2008. Barack Obama tried to talk to youth using all social media (Facebook, MySpace, Twitter) spreading the message "Yes we can". This new medium of communication made Obama beloved to voters, journalists and celebrities. John McCain, Obama's political opponent, didn't achieve Obama's rhetoric in social media. On November 5th of 2008, after winning the elections, Obama wrote in his Twitter account: *"We just made history. All of this happened because you gave your time, talent and passion. All of this happened because of you. Thanks"* [30]. By 2012 Barack Obama has had more than 100 employees administrating his Twitter account.

Twitter is used by politicians in order to affect people, who will affect other people later and goes on. Politics through Twitter is based on the word of mouth. Twitter is a miniature of society, which can define political agendas and priorities. It applies big lobbies practices. Someone can think: "How can Twitter be so influential when its users are fewer than the whole electorate?". Users legislate and decide what it is important and what is not [40].

The power of Twitter in politic communication has made the famous social networking service establish the "Twitter Government and Elections team", which has published the absolute *"Twitter Government and Elections Handbook*[7]*"*. This handbook gives advice and instructions to current or future politicians about how to set their Twitter account up (for example how to choose the perfect username), how to navigate to their profiles correctly, how to share photos or videos, how to use hashtags correctly, how to make live-tweeting and how to use Twitter Analytics in order to achieve a successful political communication.

[7] Available on: https://g.twimg.com/elections/files/2014/09/16/TwitterGovElectionsHandbook.pdf

3.4.2 Twitter users as voters

A survey made by N. Diakopoulos and D. Shamma in 2008 analyzed Tweets that commended the first debate between Barack Obama and John McCain. Results showed that Barack Obama had more positive tweets than McCain. A second survey was held for the second debate in September 2008. Analysis presented that the big volume of Tweets was posted after finishing debate, but most Tweets were posted when debate discussion topics were the financial crisis and the foreign and security policy [40]. This finding showed the hot topics that voters are interested in.

Tweets were also examined by I. Sonnefeld during the Angela Merkel and Frank Walter debate in German federal election in 2009. The survey presented that 3.507 users posted tweets commending on the debate. 30 of them wrote more than 38 tweets, while the majority of users wrote only one. The 100 most retweeted tweets were ironic by 26%, informative by 19% and annotative by 17% [41, 42].

Users do not comment politician in debate only, but in political talk shows as well. N. Anstead and B. O'Loughlin analyzed tweets published during a popular political talk show on British TV, 2009. Users had a greater desire to express their political opinions during the TV show broadcast. Some users commented the political dialogue offering additional opinions, whereas some others expressed different points of views questioning politician's statements [41, 42].

3.4.3 Twitter in Greek political reality

In 70s and 80s political speeches were held in balconies. In 90s political candidates presented themselves in TV shows. Nowadays, all politicians "tweet". Since 2009 all Greek political parties have perceived the power of Twitter and have decided to be active on this. Twitter has obliged politicians to introduce their political profiles to voters in a fast and intelligent way. Twitter constitutes journalistic tool and its value is precious for political communication.

Indisputably, 2015 was the year of Twitter in Greece. Two parliamentary elections, many eurogroups, referendum, third memorandum, intraparty elections of the opposing party were hot topics of many conversations in Twitter. When the prime mister of Greece announced the referendum in summer 2015, Twitter was overloaded by thousands of tweets. From the first moment, the hashtag *#Greferendum* became the first hashtag used

worldwide showing referendum's popularity. Proponents of "YES" or "NO" advertised their opinions through Twitter writing tweets in order to persuade the opponents.

The first days after the referendum announcement the majority of tweets supported "NO" [38]. Before capital controls imposed 25.976 tweets were published, where 71% of them supported "NO". The first day after capital controls imposed the proportion remained the same (70%- 30%). But the second day "NO" fell dramatically in 38% of total tweets making "YES" reach the 62%. The situation has been changed again few days before voting, where "NO" in Twitter dominated over "YES". As we can understand Twitter is a perfect medium for opinions gathering working as a political barometer and poll [38]. Yanis Varoufakis' tweet *"Minister No more!"* the day after referendum was one of the most typical tweets, which prove that Twitter does not follow the news, but it produces news [30].

Political "battles" among politicians, among politicians and users or among users are common phenomenon in Greece. Besides, intense clash and fanaticism are noticed. Users fight and swear each other when they disagree and the conversation level is dropped dramatically.

Twitter is the new reality in political life of everyone. Followers are now the new voters and favorites and retweets replace the old- fashioned exit polls. Dialogues among politicians or among politicians and users, caustic comments, humorous or ironic tweets differentiate Twitter from any other social media platform. Twitter in Greece is not as popular as in USA, but it is able to influence people and cause reactions [43].

4 Text Mining

Text mining is a newly established field of data mining, which tries to extract useful information from a text of Natural Language. The term *"Text Mining"* describes a system that analyzes hefty amount of natural language text and detects lexical and linguistic norms for knowledge extraction [44].

Semantic web has given the opportunity of text databases development [45]. Text databases are databases that contain the complete text of a document (books, articles, journals etc.) [46]. Data stored in text databases are semi-structured, which means that they are neither completely structured, nor unstructured [45].

A text consists of two basic units: *document* and *term*. A *document* can be a typical document book, journal, scientific paper, magazines, newspapers, web articles, e-mail messages, web pages [47]. A document can be characterized by structured fields like title, author, date etc, but the rest of it (the whole text) is unstructured [45]. A *term* is a word or phrase in the document [47].

How can we harness unstructured data like text? A lot of scientific research has been done in the field of **Information Retrieval**. As [freedictionary] informs us: "*Information Retrieval (or Text Retrieval) is the technique of storing, searching, recovering and interpreting information from large amounts of stored data*" [48]. Since there is an abundance of texts, information retrieval is used in many applications; Web search engines are the most important and frequently used applications [45].

Users, who want to find documents that are close to their desires in terms of semantic content, should use an Information Retrieval system. *Natural Language Processing (NLP),* computer's ability to "understand" text of natural language, has always been a big challenge in the field of Artificial Intelligence. A computer with NLP capabilities transforms the ASCII code of a document into well- defined conceptual model. *Polysemy* and *Synonmy* are two of the basic problems of NLP in text comprehension. Polysemy is the multiple meanings of a single word and Synonmy stands for the various different ways that the same thing can be described. Therefore, NLP techniques, which model and extract the conceptual substance of a text, are not the base of most IR systems used today [47].

4.1 Text Retrieval Methods

Text retrieval methods are separated in two categories: The *document selection problem* and *the document ranking problem*.

In *document selection method* user provides a query of Boolean expression (such as "milk" and "cocoa") and the system returns documents that indulge this expression. Because of the difficulty in defining user's information with a Boolean query exactly, this method is effective only when the user knows a lot about the corpus and can express the query in the right way [45].

Document ranking method ranks all documents of the collection in the order of relevance. This method is suitable for simple, ordinary users who do not know a lot about the corpus and they have no knowledge about Boolean queries. All modern information retrieval systems (e.g. search engines) use document ranking method and return a ranked list of documents in reply to user's query. Document ranking methods use mathematical formulas from algebra, trigonometry, logic, probability and statistics. The purpose is a score to be assigned to each document depending on how well this document matches the query. The document is ranked according to the frequency of words in the document and the whole corpus [45].

The most popular document ranking method is the *"Vector Space Model"*. The Vector Space Model is based on the following idea: Documents are represented as vectors in high- dimensional space, where every term (word) is a dimension[8].

The first step before performing information retrieval is a procedure called *Tokenization*. Tokenization is the procedure of splitting the text in tokens (words), which are candidates to be indexed. After that the text retrieval system removes words with no semantic information. This is achieved using a *stopword list* containing useless words (a, an, the, of, and, or, by, to etc). Stopwords removal reduce index size requirements and speed- up query processing.

Subsequently, *lemmatization* and *word stemming* take place. Lemmatization reduce inflected forms of a word so that they are treated as a single term (am, was, were, being ➔ be). A text mining system ought to detect words with common root such as *drink, drinks, drunk.* Stemming reduces tokens to the "root" form [45,49].

[8] Vector Space Model is described in details in the next section (4.2).

Let's study now the most fundamental technique for information retrieval "TF-IDF". TF-IDF is the acronym produced by the "Term Frequency- Inverse Document Frequency". TF-IDF considers a corpus o *d* documents and a set of *t* terms. TF-IDF calculates scores that point out the relative importance of terms in the documents. TF is the term frequency in a document and IDF is the term frequency in the whole corpus. The mathematical product of TF and IDF (TF-IDF= TF * IDF) indicates the importance of a term in the current document [45,50].

4.2 Finding Similar Documents

Once we have queried and found the documents of our interest, we have to pinpoint similar documents. Vector space model uses a similarity measure to compute the similarity between document and query. This similarity measure is C*osine Similarity*. Whereas TF-IDF gives us the ability to restrict the corpus, cosine similarity is one of the most famous methods of finding similar documents. Values produced by similarity measure rank documents [45,50].

Each document can be modeled as a vector *v* in the *t* dimensional space, where *t* is the number of terms. This is why this technique is called Vector Space Model. As we know from Discrete Mathematics: "Vector *u* is a list of numbers $\alpha_1, \alpha_2, \dots \alpha_n$". A vector like this is symbolized as:

$$u = (\alpha_1, \alpha_2, \dots \alpha_n)$$

Numbers α_i are called components or entries of *u*. Two vectors *u* and *w* are equal if they have the same number of components and the corresponding components equal. Two basic operations of vectors are the *dot product* (or *inner product*) and the *length* (or *norm*). If we have two vectors $u = (\alpha_1, \alpha_2, \dots \alpha_n)$ and $w= (b_1, b_2, \dots b_n)$, then their dot product is [51]:

$$uw= \alpha_1b_1+\alpha_2b_2+\dots+\alpha_nb_n$$

The length of a vector u is: $\|u\| = \sqrt{uu} = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$

A vector expresses a *direction* from an origin point to another point and a *magnitude*, which is the distance from the origin point. A vector can be depicted with a line segment between the origin point and another point in an n- dimensional space by plotting a line between the origin point with the other point [50].

As mentioned above, each document is represented as a vector in n- dimensional space. Similar documents will have similar term frequencies and, hence, similar vectors. Vectors very close to each other indicate similar documents. Scientists have proposed many metrics in order document similarity to be measured. Cosine similarity is one of the most important and common techniques for documents comparing. Cosine similarity is defined as:

$$sim(u,w) = \frac{u\,w}{|u||w|}$$

and calculates the cosine of the angle between any two vectors.

Let's consider the following example so as TF-IDF and cosine similarity methods to become understandable. We have the following corpus of 4 texts:

Table 1: Example texts

Text1	Breakthrough drug for depression.
Text2	New depression drug.
Text3	New approach for treatment of depression.
Text4	New hopes for depression patients.

First, the TF-IDF algorithm computes the frequency of each term (TF) for the each text and after that the Normalized Term Frequency (NTF), which is the TF divided by the words number of the text. Tables 2, 3, 4 and describe the calculation of TF and NTF for each text of the corpus.

Table 2: TF and NTF for Text1

Text1	Breakthrough	drug	for	depression
TF	1	1	1	1
NTF	1/4	1/4	1/4	1/4

Table 3: TF and NTF for Text2

Text2	New	depression	drug
TF	1	1	1
NTF	1/3	1/3	1/3

Table 4: TF and NTF for Text3

Text3	New	approach	for	treatment	of	depression
TF	1	1	1	1	1	1
NTF	1/6	1/6	1/6	1/6	1/6	1/6

Table 5: TF and NTF for Text4

Text4	New	hopes	for	depression	patients
TF	1	1	1	1	1
NTF	1/5	1/5	1/5	1/5	1/5

Now, it its time IDF value for each term to be calculated. The IDF is given by the formula:

$$IDF=\log_{10}\frac{N}{dft}$$

,where N is the number of documents in corpus (in our case N=4) and dft is the frequency of a term in the whole corpus. So, we have

Table 6: IDF calculation for each term in the whole corpus

	Breakthrough	drug	for	depression	New	approach	treatment	of	hopes	patients
NTF_{Text1}	1/4	1/4	1/4	1/4	0	0	0	0	0	0
NTF_{Text2}	0	1/3	0	1/3	1/3	0	0	0	0	0
NTF_{Text3}	0	0	1/6	1/6	1/6	1/6	1/6	1/6	0	0
NTF_{Text4}	0	0	1/5	1/5	1/5	0	0	0	1/5	1/5
IDF	0,60	0,30	0,12	0	0,12	0,60	0,60	0,60	0,60	0,60

Now we have to find the product of TF on IDF. We multiply the NTF of each term with the corresponding IDF. So,

Table 7: TF-IDF calculation for each term in whole corpus

	Breakthrough	drug	for	depression	New	approach	treatment	of	hopes	patients
Text1	0,15	0,075	0,03	0	0	0	0	0	0	0
Text2	0	0,1	0	0	0,04	0	0	0	0	0
Text3	0	0	0,02	0	0,02	0,1	0,1	0,1	0	0
Text4	0	0	0,024	0	0,024	0	0	0	0,12	0,12

In a similar way, we compute the TF, NTF and IDF for the query "New approach".

Table 8: NTF, IDF and TF-IDF for query

	New	approach
TF	1	1
NTF	1/2	1/2
IDF	0,12	0,60
TF-IDF	0,06	0,3

The only thing remaining is the cosine similarity computation between the query and each text.

DotProduct(query,Text1)= $(TF - IDF_{New_{query}} * TF - IDF_{New_{Text1}})$ +

$(TF - IDF_{approach_{query}} * TF - IDF_{approach_{Text1}})$ = 0,06*0 + 0,3*0= 0

||query|| = $\sqrt{0{,}06^2 + 0{,}3^2} = \sqrt{0{,}0936} = 0{,}305$

||Text1|| = $\sqrt{0^2 + 0^2} = 0$

CosineSimilarity1 = $\frac{DotProduct(query,Text1)}{||query|| \, ||Text1||} = \frac{0}{0{,}305 * 0} = 0$

Similarly,

CosineSimilarity2 = $\frac{DotProduct(query,Text2)}{||query|| \, ||Text2||} = \frac{0{,}0024}{0{,}0122} = 0{,}196$

CosineSimilarity3 = $\frac{DotProduct(query,Text3)}{||query|| \, ||Text3||} = \frac{0{,}0312}{0{,}0308} \approx 1$

CosineSimilarity4 = $\frac{DotProduct(query,Text4)}{||query|| \, ||Text4||} = \frac{0{,}00144}{0{,}00732} = 0{,}196$

Cosine similarity between query and Text3 is close to 1, so Text3 is similar to the query [52].

4.3 Document Classification Analysis

Since there is a huge number of on-line documents and web pages, document classification has been made a very important task of text mining. It is imperative text classification techniques to be developed in order to separate documents into classes. This procedure will facilitate document analysis and retrieval. Document classification is used for labels assigning to documents (characterization of documents) and identification of document writing styles (finding the authors of anonymous documents) as well.

The general procedure of text classification is the following way: First, a set of pre-classified documents is divided in two subsets; a training set and a testing set. The training set is used by the classification algorithm (classifier) in order to construct a classification scheme. Testing set is used for the evaluation of classification scheme accuracy. If the accuracy of the model (scheme) is satisfying, it can be used for classification of other web documents. Among the many classifiers, five of them have satisfying results in text classification. These are: nearest- neighbor classification, feature selection methods, Bayesian classification, support vector machines and association- based classification.

As we mentioned above, two documents are similar in the vector space model if their vectors are close to each other. K- nearest neighbor classifier assigns documents that have adjacent vectors (thus big angle cosine) to the same class. A test document can be considered as query to the IR system. K documents of the training set that are most similar to the query, will be retrieved. K is a mutable number. The test document class label can be defined by the class label distribution of its k nearest neighbors. K- nearest neighbor algorithm has high requirements in memory (in order to store training data) and time.

Vector space model may compute big scores for rare terms ignoring the class distribution characteristics. This method can result to wrong classification. Let's look an example of TF-IDF computation where there is problematic classification. Suppose we have 200 training documents and two classes C1 and C2 each containing 100 training documents. We also have two terms t1 and t2 in these two classes. 5 documents in each class contain the term t1 (5% of the corpus). 20 documents of C1 class contain only the term t2 (10% of the corpus). TF- IDF will assign higher value (score) to term t1, because it is rarer, whereas term t2 has stronger distinctive power in this case.

For cases like the one above, feature (or attribute) selection can be used. Feature selection decreases dataset size by eliminating irrelevant or useless attributes (or dimensions). Feature selection aims to minimize the set of attributes so that the resulting probability distribution of classes is as close as possible to the initial distribution acquired using all attributes. With this technique, terms uncorrelated with the class labels are removed from the training documents. By reducing the set of terms, efficiency and accuracy are improved.

One more classification algorithm is the Bayesian classification. As we know, document classification is the calculation of statistical distribution of documents in classes. Bayesian classifier calculates the document distribution P(d|c) of documents d to class c

and, hence, trains the model. Subsequently, it checks the class that is most likely to produce the test document. Bayesian classifier can manage high-dimensional datasets and this is why it provides effective document classification.

Support vector machines also offer effective classification since they can support high dimensional datasets. They assign a number to each class and create a mapping function which matches terms with classes.

Finally, association- based classification uses associated and frequent text patterns and keywords in order to classify documents. Association- based classification assigns high values only to non-frequent terms, because these have discriminative power. At first, information retrieval and association analysis techniques are used in order keywords and terms to be extracted. Secondly, WordNet[9] or keyword classification systems are applied, because they provide concept hierarchies of keywords and terms. Associated term rules are discovered (with the use of association rules), because they lead to better class distinction. As a result, association rules are computed for each document class. Such classification rules have strong discriminative power and this is why they are used to classify new documents [45].

4.4 Text retrieval evaluation methods

Since we have performed the text retrieval, we have to estimate the accuracy of the algorithm. Evaluation metrics like Precision, Recall and F-score have been developed for this reason. We define as {Relevant} all documents that are relevant to the query and as {Retrieved} all the documents that are retrieved by the system. Based on the Venn diagrams {Relevant}∩{Retrieved} are all documents that are both relevant and retrieved.

- Precision: Indicates exactness. The percentage of retrieved documents that are actually relevant to the query. The formula is:

$$\text{precision} = \frac{\{Relevant\} \cap \{Retrieved\}}{\{Retrieved\}}$$

- Recall: Indicates completeness. The percentage of relevant to query documents that were retrieved. The formula is:

[9] *"WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet superficially resembles a thesaurus; in that it groups words together based on their meanings."* [53]

$$\text{recall} = \frac{\{Relevant\} \cap \{Retrieved\}}{\{Relevant\}}$$

- F-score: Is the harmonic mean of precision and recall and tries to compromise them. The formula is:

$$\text{F-score} = \frac{recall \times precision}{(recall + precision)/2}$$

Although these three measures are fundamental for the evaluation of the text classification algorithm, they are not helpful for comparing two ranked lists of documents, because they are not affected by the internal ranking of the documents in the retrieved set. If we want to evaluate the quality of ranked list of documents, we ought to calculate the average of precisions at all ranks, where a new relevant document may be returned [45].

4.5 Latent Semantic Indexing

Opponents of TF-IDF method claim that TF-IDF cannot "understand" synonymous words. User may include words in their queries, which are different from the terms indexing a document. For example, a query containing the terms "Data mining", will not return documents about knowledge discovery. One solution to that problem is the creation of a database that will link semantically and conceptually related words (terms) together [47].

Another alternative solution to the problem above is the Latent Semantic Indexing (LSI). LSI (also called Latent Semantic Analysis in text analysis) analyzes the relationships of the terms in a set of documents and constructs a concept context relevant to the documents and terms. Its aim is to detect similarities among documents. As the name insinuates, LSI extracts hidden semantic structure from documents rather than using just terms frequencies [47,54]. The LSI's main philosophy is to keep the most representative features without affecting the final result [45].

LSI supposes that similar words are appeared in similar pieces of text. An initial matrix M×N is constructed, where columns represent text paragraphs and rows represent unique words. The cells of the matrix contain the number of each unique word in the current paragraph (column).

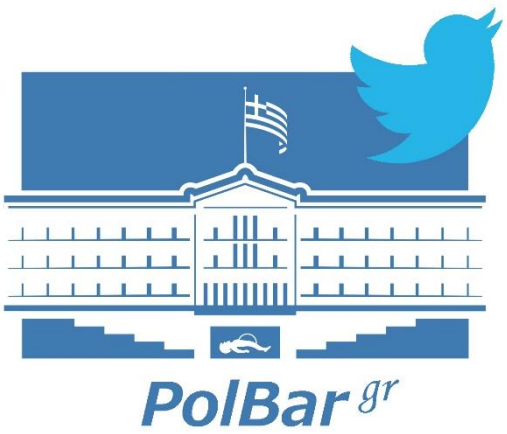
LSI reduces the dimensionality of the matrix (document) using the Singular Value Decomposition (SVD) technique [45]. Singular Value Decomposition is applied in order the number of rows to be reduced and, as a result, a much smaller matrix KxN to be

constructed, where K<<M. Typical value of M could be 50.000, while value of K can be 100 without information deprivation.

Subsequently, the cosine similarity of two vectors (rows of matrix) is computed and the words are compared with each other. Cosine value close to 1 indicates very similar words, whereas value close to 0 indicates different words. LSI is able to compare two documents (or a document and a query) and find similarities between them. Based on these similarities, LSI can be used, also, for text classification [54].

5 The PolBar Application

"PolBar" is an Android application that implements text mining techniques. It collects, analyzes, and categorizes tweets written by simple Twitter users for Greek parliament members. Tweets are categorized in positive, neutral, and negative and useful statistics about the popularity of each deputy (percentages of positive, neutral and negative tweets) are exported. PolBar is the acronym of the phrase "Political Barometer", since this application estimates political trends in Greece. The PolBar logo is shown above (picture 6) and it is a combination of the Greek parliament logo and the Twitter logo.



Picture 6: PolBar logo

PolBar is an innovative application, since an application like this does not exist in Greece or in any other country. Conventional polls conducted by newspapers, magazines or TV channels are not always reliable, since they can benefit specific politicians or parties serving political interests and misleading electorate. PolBar's results are determined by Twitter users' opinions exclusively maintaining credibility.

PolBar is implemented in 4 steps:

1) Tweets collection,
2) Tweets storage
3) Tweets analysis
4) Results presentation

5.1 Tweets Collection

5.1.1 Communicating with Twitter API

As mentioned in chapter 3 Twitter offers tweets for further analysis through Twitter API. In order access to Twitter API to be acquired, a Twitter account is demanded. Using this

Twitter account, we visit the "Developers" service of Twitter (https://dev.Twitter.com/) so as access to Twitter API to be granted. By clicking on "My Apps" choice, Twitter Developer provides us all the necessary credentials[10] needed for user authentication and tweets provision.

Credentials provided above are inserted into java code. An application written in java communicates with Twitter API, crawls tweets based on a specific keyword (hashtag, username, word etc.) and stores them in SQL database. Necessary component for this java application is the "Twitter4j" library.

Twitter4j11 is a free, external java library for communication with Twitter API. It supports both REST and Streaming API. PolBar uses REST API, since there is no online server available. Twitter4j integrates java programs with Twitter API making tweets collection a very simple procedure. It is developed in 100% pure java code and works on any java virtual machine version 5 or later.

Picture 7: Twitter4j logo [55]

Furthermore, Twitter4j offers a great variety of extra features. It is compatible with Android platform and Google App engine. It also supports threads giving the opportunity of concurrent method call. Besides, it transforms Twitter API JSON responses into simple strings or numbers for easy data processing and manipulation. Moreover, it also renders tools for easy proxy server configuration for server- side applications. "Bugs Report Service" gives developers the capability of reporting possible bugs in order Twitter4j to being improved [56].

Twitter4j contains many implemented methods that return useful information about users and tweets they publish. For example, Twitter4j retrieves tweet's geolocation or country or even street address that it came from, date and time published, number of favorites or retweets, and the user's username who wrote the current tweet. Regarding users, Twitter4j returns data about account creation date, number of tweets or followers or friends or favorites they have, profile image or background image, profile URL, profile color, status, time zone and many more.

[10] Consumer Key (API Key), Consumer Secret (API Secret), Access Token, Access Token Secret
[11] http://Twitter4j.org/en/index.html

5.1.2 Choosing the suitable search keyword

Next step was finding the suitable keyword for tweets collection. Hashtags would be a good solution, but unfortunately Twitter users have created many different hashtags for the same person (e.g. #firstnamelastname, #lastnamefirstname, #firstname#lastname, #lastname (both in Greek and greeklish[12] characters)). So, searching deputies only by hashtags is not a practical technique, since a single deputy may have multiple hashtags. The use of just the last name of the deputy is not indicated, because there are many individuals with the same last name. So, the only solution was the use of deputies' Twitter username, since Twitter usernames define each person uniquely.

At first, all parliamentarians were searched one- by- one through the official Greek parliament website[13.] Afterwards, all deputies' Twitter accounts were searched in Twitter. The Twitter usernames used by PolBar are quoted in footnote 14 below[14] (Total number: 187). One more thing that should be mentioned here is that PolBar uses deputies whose

---

[12] Greeklish: Greek words written with Latin alphabet characters.

[13] www.parliament.gr

[14] @atsipras, @AdonisGeorgiadi, @kmitsotakis, @nikospappas16, @NikosKotzias, @K_Hatzidakis, @PanosKammenos, @olgagerovasili, @nasosa8anasiou, @MakisVoridis, @BKegeroglou, @YDragasakis, @TheanoFotiou, @fortsakis, @stamatis_dim, @nkerameus, @PappasXA, @theocharop, @Paparigaleka, @nikosfilis1, @chrisvernard, @olgakef, @Dora_Bakoyannis, @nkaklamanis, @KSkandalidis, @ElenaKountoura, @Mar_Georgiadis, @kouroumplis, @pskourl, @giorgosdimaras1, @cspirtzis, @TsironisGianni, @dbbda1ef08c242b, @annetakavadia, @giorgoskyritsis, @v_meimarakis, @NikosDendias, @AnnaKaramanli, @Sofia_voultepsi, @AnnaAsimakopoul, @GerGiakoumatos, @kpapakosta, @GermenisGiorgos, @ipanagiotaros, @EleniZaroulia, @grpsarianos, @amirasgiorgos, @PapachristTh, @ProedrosEK, @PanagoulisSt, @SalmasMarios, @k_karagounis, @Barbarousis, @dkonstantop, @Yannis_Maniatis, @KostasVlasis, @Odysseas_, @VasilisTsirkas, @gstylios, @tsakalotos, @PanosSkourolia1, @gpantzas, @GeorgiaMartinou, @Vlachos_G, @IliasKasidiaris, @evichrist, @dimitris176, @sia_anag, @katsaniotis, @papatheodorou_t, @IasonFotilas, @EKarakostas8, @kyriazidisdim, @xarakefalidou, @nectarsant, @dgakis, @kamateros_hlias, @manoskonsolas, @KremastinosD, @kaisasgeorgios, @NatasaGkara, @pavpol2222, @g_stathakis, @amixelis, @EVenizelos, @JSarakiotis, @cstaikouras, @k0stopanagiotou, @igglezikaterina, @GVagionas, @Vkikilias, @FofiGennimata, @htheoharis, @a_loverdos, @GKoumoutsakos, @Mavrogiorgos, @VroutsisGiannis, @DimGeoRizos, @ADimoschakis, @geakriotis, @SimosKedikoglou, @MBalaouras, @Tzavaraskon, @koutsoukosilia, @FKarasarlidou, @a_vesiropoulos, @l_avgenakis, @SpyrosDanellis, @marios_katsis, @Amanatidis_Gian, @tasoskourakis, @MarkosBolaris, @tr_alexandros, @gioulekaskostas, @KalafatisSt, @Elena_Rapti, @AntonisGregos, @KostasZouraris, @Saridis_Ioannis, @mardas55, @TheoKaraoglou, @savanastasiadis, @Arvanitidis_Geo, @kat_markou, @meropitzoufi, @npanagioto, @MarAntoniou, @FotiniVaki, @IAIVATIDIS, @Parastatidis_th, @XA_KILKIS, @johntheofilakto, @mtheleriti, @g_psychogios, @chrisdimas1980, @syrmal2000, @adavakis, @lgrigorakos, @annavayena, @pantamaximos, @kellas_xristos, @kbargio, @g_katsiantonis, @plakiotakis, @XarAthan, @A_Meikopoulos, @chboukoros1, @Panagiotisiliop, @gkatr, @pkonstantineas7, @samaras_antonis, @dimkoukoutsis, @SGiannakidis, @elen_stamataki, @Ggennia, @katsafados, @dimitrisvitsas, @ninemesis, @EviKarakosta", @tzakri, @johnsaxinidis, @Kastoris_Aster, @Koukodimos, @kostasbarkas, @andreasxanthos, @gkefalogiannis, @ArampatziFotini, @MTZELEPIS, @MariaKolliaTsar, @amegalomystakas, @sakis_papad, @p_dritseli, @XrSimorelis, @KostasSkrekas, @akaranastasis, @dimvett, @ant_balomenakis, @nmitarakis, @MVarvitsiotis, @ZoeKonstant, @rachelmakri,@koukoulopoulos, @tpangalos, @e_stylianidis, @thanosplevris, @NikNikolopoulos, @TeoPelegrinis

Twitter account is active. 5.200 tweets on average are collected for these 187 parliament members weekly.

5.1.3 Organizing keywords

As we already know from chapter 3, Twitter API provide us the last 100 tweets or the tweets of the last 10 days (keeping the maximum tweets number equals 100). Based on this criterion deputies were divided into two categories: a) deputies of high interest, who receive more than 100 tweets every day or couple of days, and b) deputies of low interest, who receive maximum 100 tweets in 10 days.

Tweets of first category deputies are collected every two days, while tweets of second category deputies are collected every 10 days (sampling in both cases). All parliament members were checked one- by- one about the number of tweets they receive daily and weekly. Only 4 parliamentarians belong to the first category of high interest (the first 4 mentioned in the footnote 14 above (p.31)).

5.1.4 Crawling and preprocessing tweets

The java program that communicates with Twitter API, crawls tweets, which are written only in Greek language so as tweets in greeklish to be excluded. In addition, tweets coming from reportorial (journalistic) accounts (online newspapers, magazines etc) are also excluded, since they don't contain any useful information (opinion or criticism) just news about the politician. Furthermore, retweets are blocked because they are repeated tweets, while unique tweets are desirable.

Besides, java application makes a basic preprocessing in tweets before being stored in database. Firstly, all single quotes (΄) are deleted from tweet, because they create exceptions in the INSERT command of SQL. "Enters" are, also, taken away for tweet simplification and easy insertion in database table. Moreover, possible links (sentences starting with "http" or "https" or "www") are also removed, since they are useless information that cannot be harnessed.

Emoticons could help the classification procedure, since a single emoticon can characterize the "mood" of a tweet immediately. Unfortunately, emoticons have different codification from that used in database *(utf8mb4_unicode_520_ci)* and exceptions are caused. The only solution is the replacement of emoticons with a white space.

5.2 Tweets Storage

All tweets are stored in MySQL database table. There are 3 tables. The "Data" table which stores the tweet, a unique ID for this tweet (auto- increment integer number), the person (politician) for whom the tweet was written, the status (label) of tweet, the day, month, year that the tweet was published and the date that the tweet was inserted in our database. The "ID" column is necessary in order every inserted tweet to be defined unlikely. Columns "Person" and "Tweet" are varchar type, whereas "Status", "Day", "Month" and "Year" are integer type. "InsertedIn" is date type. A structure of "data" table is show below:

Table 9: "Data" table structure

ID	Person	Tweet	Status	Day	Month	Year	InsertedIn

When a new tweet is added in the table, the status is set NULL by default. Afterwards, Scala classification application retrieves all tweets with NULL status, classifies them and updates this characteristic with an integer number. "Status" can take 3 different numbers: 0 for neutral, 1 for negative and 2 for positive tweets. Negative numbers (such as -1) cannot be used, because classification algorithms accept only non- negative values.

Next table is the "Month Statistics" table. This table stores for every deputy the sum of positive, neutral, and negative tweets and the percentage of them for every calendar month as well. It also stores the month and the year of each statistic (record) and the sum of all tweets (Sum column) that a politician has received for the current month and year in order month statistics for every deputy to be exported.

Table 10: "Month Statistics" table structure

ID	Person	Positive	Neutral	Negative	Perc_Positive	Perc_Neutral	Perc_Negative	Sum	Month	Year

The third table is the "Total Statistics" table. This table has the same columns as the "Month Statistics" table apart from the ID, the month and year. The "ID" column is not necessary here, because all data is deleted from the table before new data being inserted,

so there is no reason for numbering. “Total Statistics” table summarizes statistics for each deputy from the beginning of time (September 2016) giving a general overview.

Table 11: “Total Statistics” table structure

Person	Positive	Neutral	Negative	Perc_Positive	Perc_Neutral	Perc_Negative	Sum

Separate java programs were developed in order month and total statistics to be calculated and inserted into the suitable table of the database. Instances of tables are quoted in appendixes A, B and C. In future improvements of the app, one more table will be added that will store the Twitter usernames of deputies, their real first name and last name, the political party, and the constituency they belong to. This new table will be connected to other tables in “Person” (username) column, which is common for all.

5.3 Tweets Analysis and Classification

After collecting and storing tweets in an organized way, data analysis is needed. Scala is the language that is used for tweets analysis and classification. As mentioned above, all Scala does is retrieving tweets with NULL status (which means that they haven’t been classified yet), applying a classification algorithm on them (the one with the best accuracy) and updating the “Status” characteristic. After tweet analysis, classification algorithm produces an integer number (0,1 or 2), which describes the class of the tweet, and runs UPDATE query to the table. Scala offers implemented classification algorithms for both binary (two classes) and multilabel (many classes) classification.

5.3.1 Creating the training dataset

Necessary precondition for classifiers evaluation and selection of the best is the creation of the training set. As mentioned in chapter 4 training set is a set of classified data that trains the classifier (algorithm) in order the latter to “know” how to classify data (This method is called “supervised learning” in data mining). The first 2.000 tweets were classified manually[15]. During the manual classification of these 2.000 tweets, “problematic” tweets were discovered. These are tweets which are difficult to be classified even by human for many reasons. Table 12 gives few examples of these “problematic” tweets.

[15] 197 positive, 718 neutral, 1085 negative.

Table 12: "Problematic" tweets

Tweet	Comment
"@politicianA I like you the best. @politicianB is useless."	This tweet is both positive for @politicianA and negative for @politicianB. Furthermore, this tweet will be collected twice; once when crawling tweets for @politicianA and once when crawling tweets for @politicianB.
"@UserA I agree with you. @politicianA"	Twitter user agrees with @UserA's opinion for @politicianA, but we don't know if this opinion is positive or negative.
"@politicianA Look at this" (Tweet contains a photo)	Photo is not collected.
"@politicianA @politicianB @politicianC All government does is tax imposing" (where @politicianA is governmental deputy and B, C are not)	This tweet is negative for @politicianA and neutral for B and C. It will also be crawled three times (one for each politician mentioned).
"@userA You are mean. @politicianA"	This is a negative comment for user, not for politician.
"@politicianA Congratulations for the great job!"	We don't know if this is an ironic (negative) tweet or not. Possibly ironic (based on statistics).
"@politicianA Why doesn't it?"	If user derides, it is an ironic, negative tweet. If user asks a question based on politician's statement, it is neutral.

Some other problems found in tweets analysis is the absence of punctuation marks and intonation, misspellings and joined or cropped words (so as users to save characters), which reduce classifier accuracy. Joined words are read by classifier as one word. A misspelling or a cropped word is not recognized as the same with the grammatically correct one.

Moreover, timeliness changes every month (or sooner) and new discussion topics are appeared. Thus, tweets contain new words, which are not included in training set. Therefore, training dataset should be updated frequently, otherwise, the classifier will "meet" new words without knowing how to classify them.

Stopwords

Stopwords are very frequent words, that don't not carry semantic or important information. Stopwords are removed from the training and test dataset and from unlabeled (and stored) tweets speeding up the classification process and improving classifier accuracy[16]. Stopwords used in PolBar are quoted in footnote[17] 17.

A good practice would be the removal of all 2-letters words, which are articles, particles, conjunctions etc. However, that practice would remove the greek, rude word "re" (ρε) (there is no translation in english) and the word "xa" (χα) (vocal display of laugh), which are used for mockery in all negative tweets. Words "Re" (ρε) and "xa" (χα) indicate a 100% negative tweet and therefore it must be included in training set and in unlabeled tweets. So, all the other 2-letters words were included in stopword list. This stopword list is not applied only to the 2.000 labeled tweets, but to all unlabeled tweets during the analysis procedure also.

5.3.2 Classifiers evaluation

Since 3 categories (positive, neutral, negative) are used, classifiers that support multiple classes (multilabel classification), should be examined. As the official Spark Apache website [57] informs us, multiclass classification is offered by "Logistic Regression", "Naïve Bayes", "Decision Trees" and "Random Forest" classifiers. In all evaluations, the 2.000 labeled tweets were randomly split into 60% (1.200 tweets) training set and 40% (800 tweets) testing set. This division (60%- 40%) is applied only in evaluation process. During the unlabeled tweets analysis, all 2.000 labeled tweets are used as training set.

The measurement used for classifiers evaluation is "F- measure" as a single measurement of the system, since it combines Precision and Recall metrics. From now onwards "F- measure" will be called (abusively) "Accuracy" for simplicity reasons.

Logistic Regression

Logistic regression is an approach to prediction. It is a statistical technique designed for provision of a categorical, dependent variable using quantitative and qualitative independent variables. Independent variables determine the outcome. This provision is

[16] Further analysis about accuracy improvement in section 5.5.

[17] απ, από, απο, ΑΠΟ, τη, την, ΤΗ, ΤΗΝ, για, ΓΙΑ, τν, τς, να, ΝΑ, αν, ΑΝ, ας, ΑΣ, θα, ΘΑ, και, ΚΑΙ, που, ΠΟΥ, το, Το ,ΤΟ, τον, ΤΟΝ, στο, στον, ΣΤΟ, ΣΤΟΝ, οι, ΟΙ, στη, ΣΤΗ, στην, ΣΤΗΝ, τα, ΤΑ, του, ΤΟΥ, εγώ, εγω, ΕΓΩ, εσύ, εσύ, Εσύ, ΕΣΥ, τι, Τι, ΤΙ, πως, ΠΩΣ, οτι, ότι, ΟΤΙ, σε, τις, Τις, ΤΙΣ, της, Της, ΤΗΣ, με, μου, Μου, σου, Σου , Μας, μας, σας, Σας, τους, Τους

achieved by estimating probabilities with the use of a logistic function. Logistic regression is used in cases, where we want to predict the presence or absence of a characteristic or event. This type of classification is applied in binary problems mainly, but it, also can be extended to multiclass classification problems creating the *multinomial logistic regression*. In this case, we want to model and predict the probabilities of K classes without their sum exceeding the [0,1] space [45,58].

For K possible classes (outcomes), one of the classes can be defined as "pivot". The rest K-1 classes can be regressed apart from the others against the pivot class. Spark's machine learning library chooses always the first (0) class as "pivot". For multiclass classification problems, the classifier will construct a multinomial logistic regression model. In this model, K-1 logistic regression models regressed against the 0 class will exist. Therefore, K-1 models will be constructed and run. The class (outcome) with the largest probability will be the predicted class [59]. In our experiment, number of classes are equal to 3 and the resulted accuracy was 57.35%.

$$\mathbf{Accuracy_{LogisticRegression} = 57.35\ \%}$$

Naïve Bayes

Naïve Bayes is a statistical classifier that performs probabilistic prediction. It is based on Bayes' theorem according to which the current probability is related with the initial one. Naïve Bayes classifier supports multiclass classification and predicts class memberships probabilities with strong independence assumptions among features. Each feature is a term whose value is the frequency of the term [60, 61].

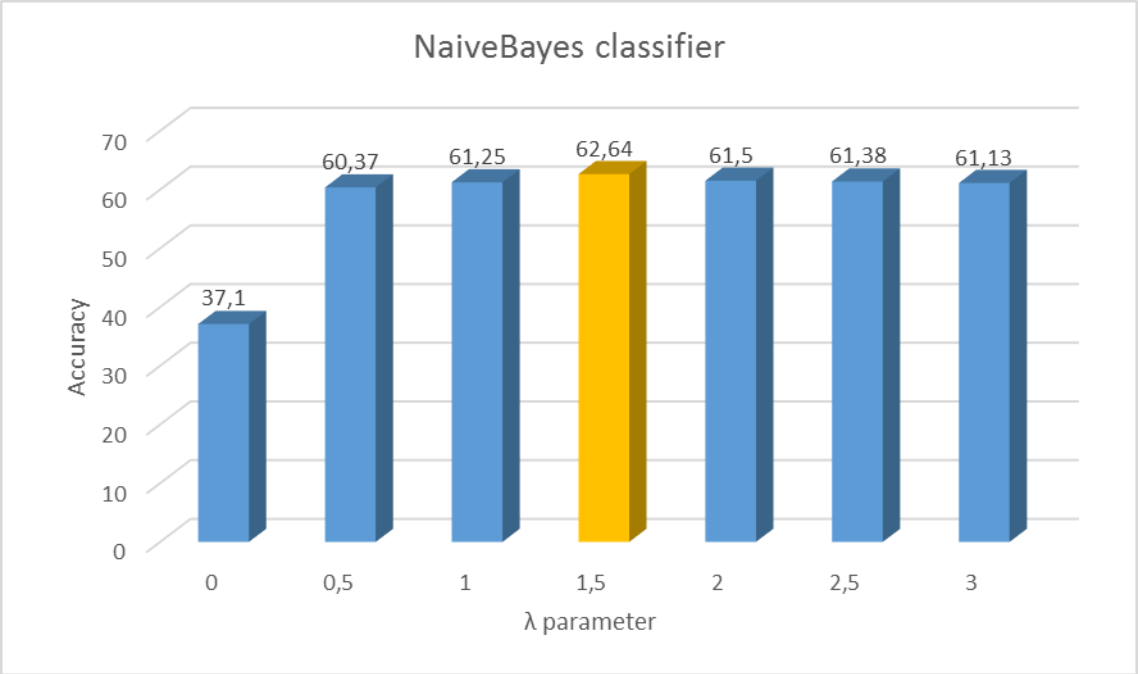
Table 13: Naïve Bayes accuracy

λ	Accuracy
0	37.10
0.5	60.37
1	61.25
1.5	**62.64**
2	61.50
2.5	61.38
3	61.13

Naïve Bayes classifier represents each text X as an n-dimensional vector, where n is the number of words (attributes). For K number of classes Naïve Bayes classifies the text X to the class, which has the highest posterior probability conditioned on X [54].

A Scala developer experiments with two parameters: modelType and lamda (λ). Model type can take two values: *multinomial* and *Bernouli*. Bernoulli Naïve Bayes supports only binary classification, so all experiments were held with multinomial version of the algorithm. Lamda (λ) parameter adjusts the "Additive smoothing". Additive smoothing is a technique used for categorical data smoothing [62]. Accuracy of the model was examined for 7 different values of the λ parameter. The results of the experiments are show in the

table 13 numerically above and in picture 8 graphically below. The best accuracy (62.64%) was achieved for λ=1.5.



Picture 8: Naïve Bayes classifier evaluation

Decision Tree

A decision tree is a supportive tool in decision making, which uses a tree graphical depiction including all posible decisions, all influence factors and all posible results. Every possible "decision point" is represented by a node. Nodes in decision trees test a particular attribute. Usually, the test performed in a node is actually a comparison of an attribute value with a constant. Sometimes, two attributes are compared with each other. Every possible choice that can be taken in a node (decision point) is depicted with an arrow to a child- node [63].

One parameter that was taken into account is the *maxBins*. This variable defines the maximum number of bins[18]. Incresing the number of bins, split decisions turn to be more fine- grained and decision rules more optimal. However, the computation (processing) time is augmented too.

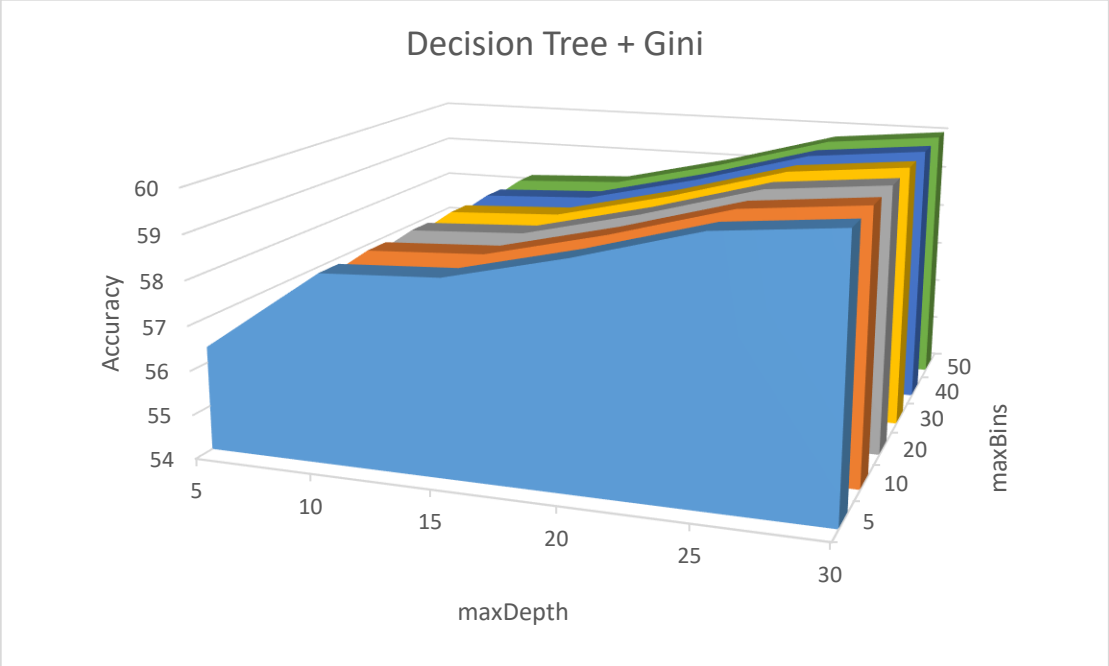
[18] Bins is a term of statistics science. The number of bins represent the width of a column (or class) in a histogram. The number of bins (column/ class width) sets the number of values (objects) that will "fall" in each class [64].

Besides, the algorithm contains the parameter “Impurity”. Impurity calulates the labels’ homogeneity at each node. Impurity can take two values: *Gini* or *Entropy*. Both Gini and Entropy use almost the same formula with the difference that entropy uses the logarithm of the label’s frequency at each node, whereas Gini does not.

The last parameter examined is the *maxDepth*. *maxDepth* defines the depth of the tree. The deeper a tree is, the more accurate it is. Deepers trees are very expressive, but their construction and trainning cost in time and memory. The maximum number that can be assigned to maxDepth parameter is 30 [65]. Evaluation results of Decision Tree algorithm with Gini impurity are presented below (numerically and graphically).

Table 14: Decision Tree (Gini) evaluation results

Bins / maxDepth	5	10	20	30	40	50
5	56.35	56.35	56.35	56.35	56.35	56.35
10	58.23	58.23	58.23	58.23	58.23	58.23
15	58.36	58.36	58.36	58.36	58.36	58.36
20	58.99	58.99	58.99	58.99	58.99	58.99
25	59.74	59.74	59.74	59.74	59.74	59.74
30	**60**	**60**	**60**	**60**	**60**	**60**



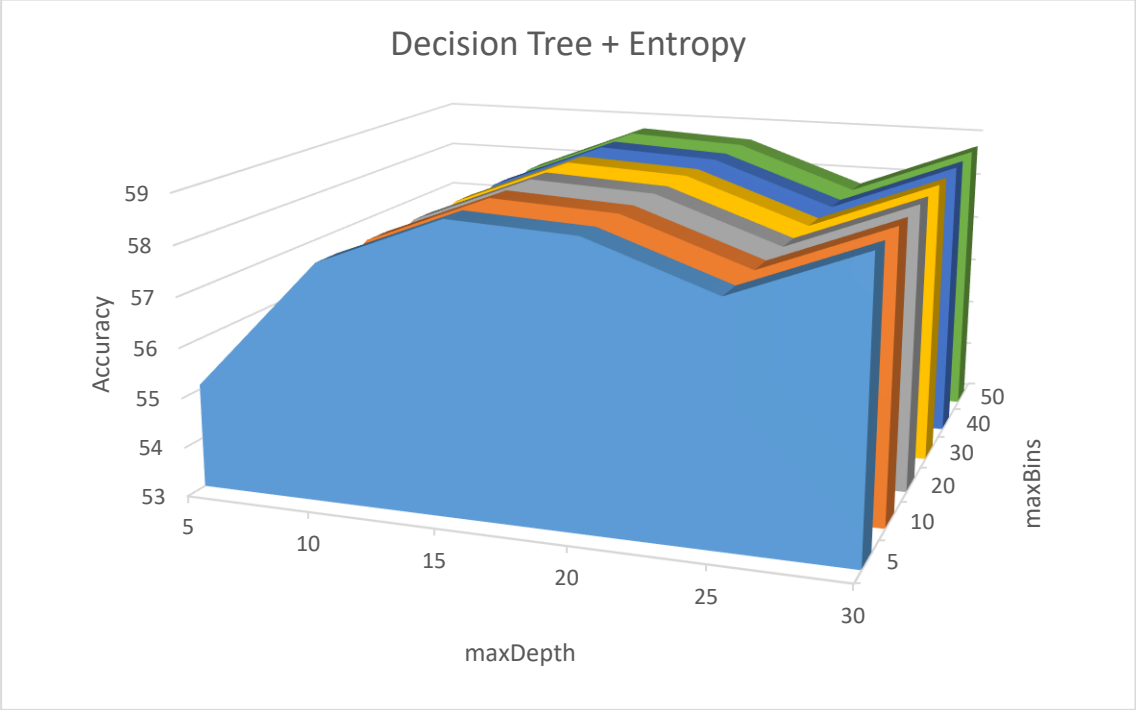
Picture 9: Decision Tree (Gini) evaluation results

As we can notice, the number of bins does not affect the accuracy of the algorithm, since the number of classes, we use, is quite small. The best accuracy (60%) was achieved with the maxDepth equals 30 regardless the Bins number.

Evaluation results of Decision Tree algorithm with Entropy impurity are presented below.

Table 15: Decision Tree (Entropy) evaluation results

Bins / maxDepth	5	10	20	30	40	50
5	55.09	55.09	55.09	55.09	55.09	55.09
0	57.73	57.73	57.73	57.73	57.73	57.73
10	**58.74**	**58.74**	**58.74**	**58.74**	**58.74**	**58.74**
20	58.61	58.61	58.61	58.61	58.61	58.61
25	57.73	57.73	57.73	57.73	57.73	57.73
30	**58.74**	**58.74**	**58.74**	**58.74**	**58.74**	**58.74**



Picture 10: Decision Tree (Entropy) evaluation results

As we notice again, the number of bins does not affect the accuracy of the algorithm. Besides, the best accuracy (58.74%) was achieved for two different values of max Depth (10, 30) regardless the bins number. Gini impurity have achived better accuracy than Entropy.

Random Forest

Random Forest is a group of decision trees. Random forest algorithm constructs several decision trees (forest) and trains them separately. The classifier is called "Random", because the training process of each decision tree is performed randomly. As a result, all trees differ with each other. Each tree produces a result suggesting a class. Subsequently, the algorithm combines the results of each tree and chooses the class that was suggested by the most trees.

Random Forest is suitable for large datasets, since it can manage thousands of inputs variables and evaluate which of them (the variables) are meaningful. Moreover, it estimates possible missing data and preserves accuracy when big proportion of data is missing. Random forest can be, also, applied to unlabeled data, leading to unsupervised learning (clustering) [45, 58, 66].

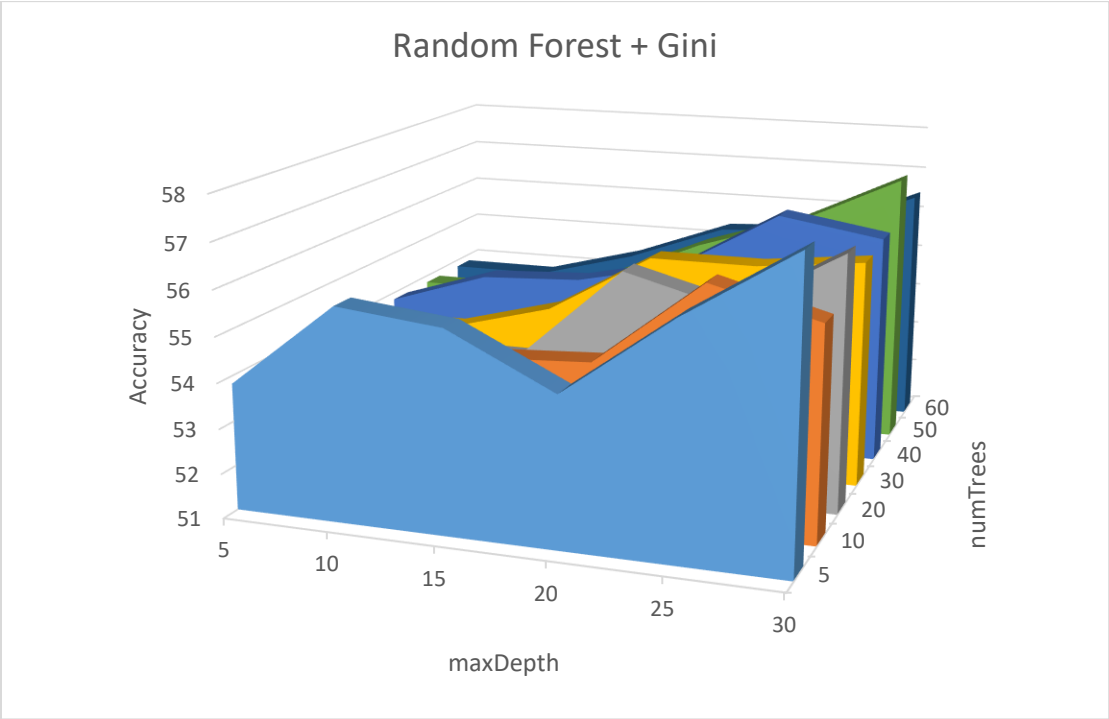
Random Forest classifier is considered as one of the most popular and successful algorithm for machine learning. Thanks to the multiple Decision Trees, the risk of overfitting is reduced. Since Random Forest uses decision trees, it can support categorical features.

We experiment with two parameters of the algorithm: *numTrees* and *maxDepth*. The parameter *numTrees* represents the number of trees in the forest. Increasing the number of trees in the forest, more accurate predictions are achieved but, also, the run time of the algorithm increases linearly. *maxDepth* is the depth of each decision tree, as described above (p.39). There is, also, the impurity parameter, which takes two values: *Gini* and *Entropy* (also described in p.38,39).

Evaluation results of Random Forest algorithm with Gini impurity are presented below.

Table 16: Random Forest (Gini) evaluation results

numTrees / maxDepth	5	10	20	30	40	50	60
5	53.96	53.83	53.83	53.83	53.83	53.83	53.83
10	53.96	54.21	53.83	54.71	53.83	53.83	53.83
15	55.59	53.83	54.84	54.33	54.71	54.71	54.46
20	55.84	56.22	56.47	55.97	55.22	55.84	54.84
25	55.59	57.10	56.98	56.22	56.10	57.23	56.35
30	**57.86**	**57.35**	**57.10**	**57.61**	**57.73**	**57.35**	**57.23**



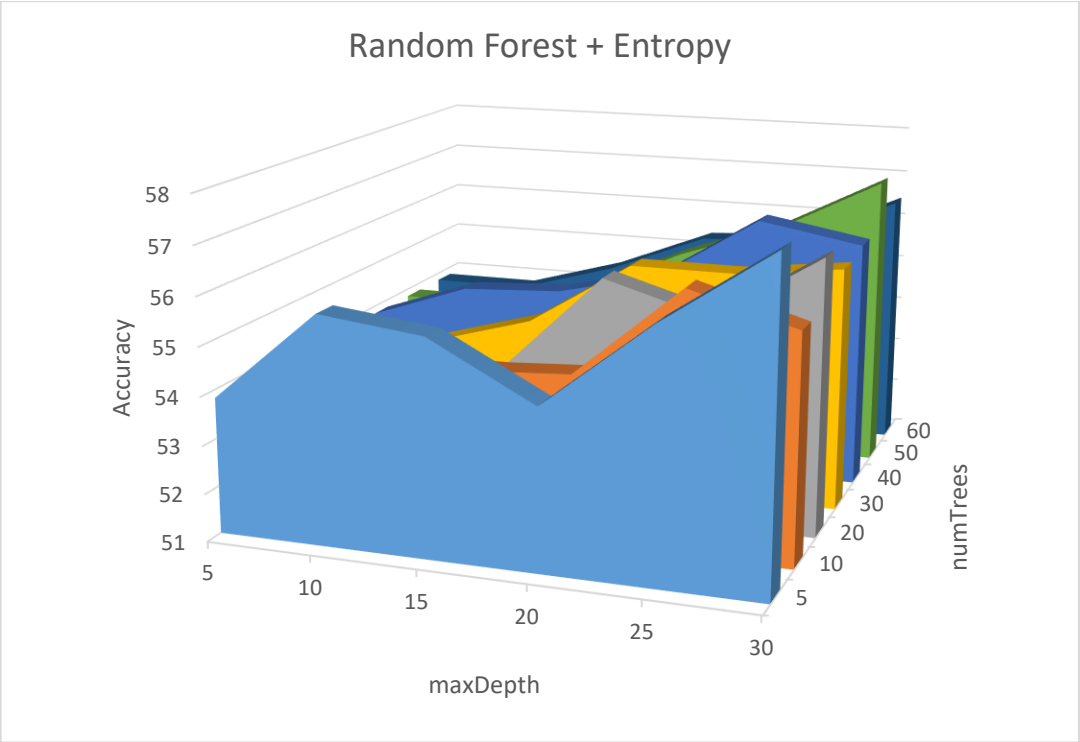
Picture 11: Random Forest (Gini) evaluation results

As we can notice, accuracy is improved by increasing the number of maxDepth. Augmentation of trees number (forest) does not offer better accuracy. In some cases, a big number of trees caused accuracy reduction (keeping the maxDepth number constant). The best accuracy (57.86%) was succeeded with the minimum numTrees (5) and the maximum maxDepth number (30).

Evaluation results of Random Forest algorithm with Entropy impurity are presented below.

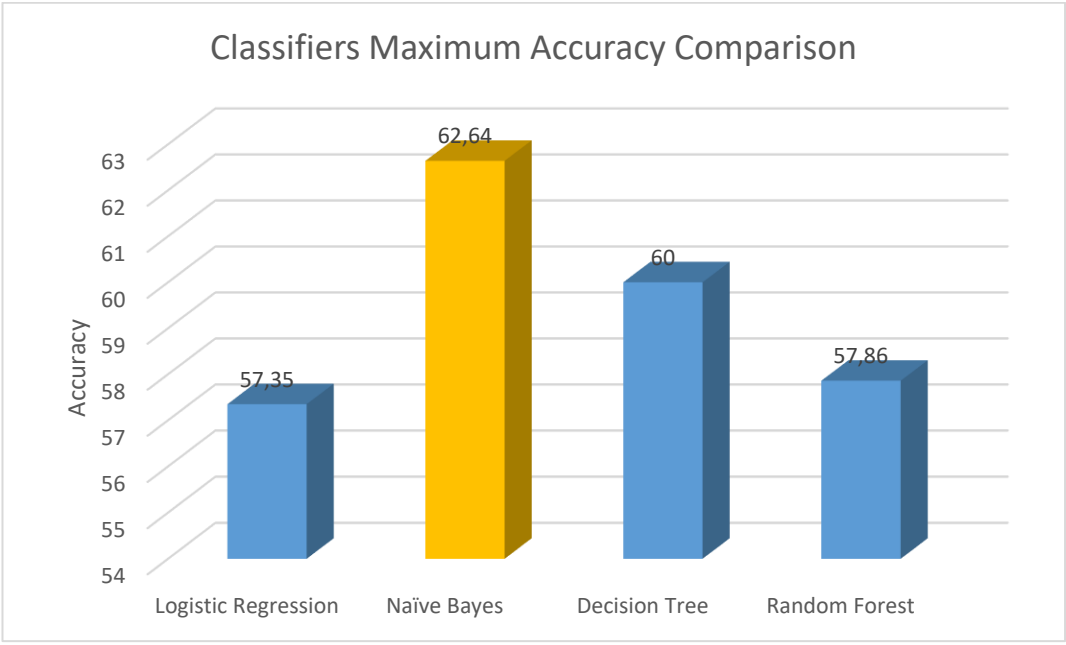
Table 17: Random Forest (Entropy) evaluation results

numTrees / maxDepth	5	10	20	30	40	50	60
5	53.83	53.83	53.83	53.83	53.83	53.83	53.83
10	55.72	53.83	53.83	53.83	54.59	53.83	53.83
15	55.47	54.33	53.83	54.46	54.71	54.33	54.46
20	54.33	54.46	55.97	55.84	55.22	55.34	55.34
25	56.10	56.35	55.34	55.84	**56.60**	56.10	55.47
30	**57.61**	**55.72**	**56.72**	**56.10**	56.22	**57.23**	**56.47**



Picture 12: Random Forest (Entropy) evaluation results

As happened in Gini impurity case, accuracy is improved by increasing the number of maxDepth, not by the number of trees. The best accuracy (57.61) was succeeded again with the minimum numTrees (5) and the maximum maxDepth number (30). All in all, among the 4 classifiers described in this section, Naïve Bayes presented the best accuracy (62.64%) and that is used for the classification of the rest of tweets.



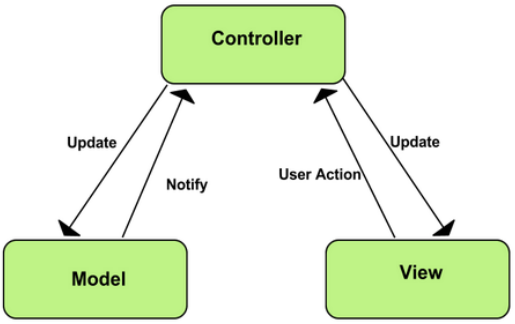
Picture 13: Classifiers maximum accuracy comparison

5.4 Results Presentation

Since statistics are ready in database, it is time for them to be presented through a mobile application. PolBar is a native android application developed for use in mobile devices (tablets and smartphones), which run Android operating system. Native application development was chosen, because it offers reliability, stability, and high performance.

PolBar was implemented in Android Studio 2.2.2[19] (programming language is Java) with Minimum SDK (Software Development Kit[20]) Android 4.2 (Jelly Bean) with API 17[21]. The Minimum Required SDK is the earliest version of Android operating system, which an application supports. This SDK was chosen, because, according to android studio, 87.4% of android devices run Jelly Bean version globally. So, PolBar will be run by the 87.4% of total android devices in the world [68].

PolBar was built using MVC (Model, View, Controller) model, because this model is advisable for application with graphical interface. *Model* stores application data (MySQL database), *View* visualizes data for presentation to user (Statistics Activities (classes) in java code with the corresponding xml files) and *Controller* synchronizes View with Model and checks data flow between them [69].



Picture 14: MVC model [69]

PolBar navigation is very clear and understandable. When user starts the application, Splash activity is appeared. Splash activity is the startup screen showing the PolBar logo for 3 seconds. After passing 3 seconds, the main menu is presented on user's screen. Main menu contains 2 choices: "Total Statistics" and "Month Statistics". Data are retrieved from "Total Statistics" and "Month Statistics" tables and are presented on device screen visually. All results are presented in a ListView layout[22]. In general, the app functionality is very simple. Screenshots of the application are introduced below.

---

[19] JRE: 1.8.0_76-release-b03 amd64, JVM: OpenJDK 64-Bit Server VM by JetBrains s.r.o

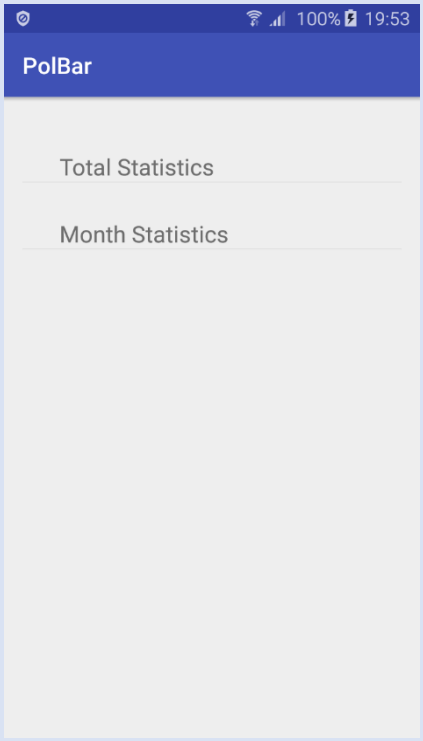
[20] *"The SDK includes tools, sample code and relevant documents for creating Android apps."*[67]

[21] *"API level is an integer value that uniquely identifies the framework API revision offered by a version of the Android platform"* [68]

[22] ListView layout is described in p.46.

Picture 15: Initial Screen (Splash Activity)

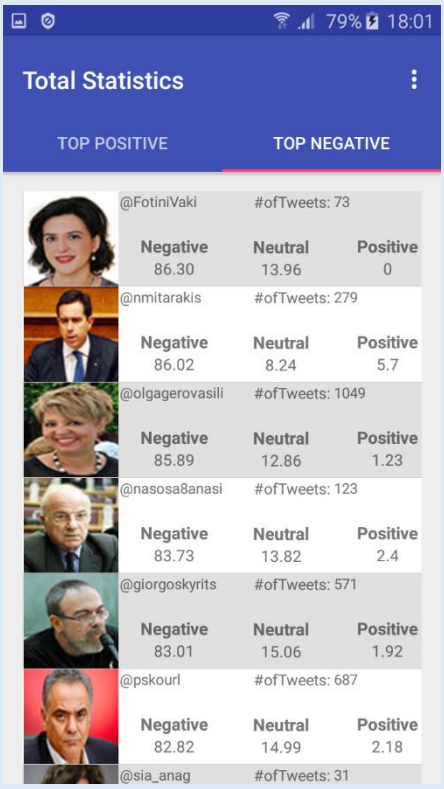
This is the initial screen (Splash Activity) appeared for 3 seconds when the app starts showing the PolBar logo.



Picture 16: Main Menu

This is the main menu where the user chooses what he wants to see.

Picture 17: Total Statistics (Top Positive tab)



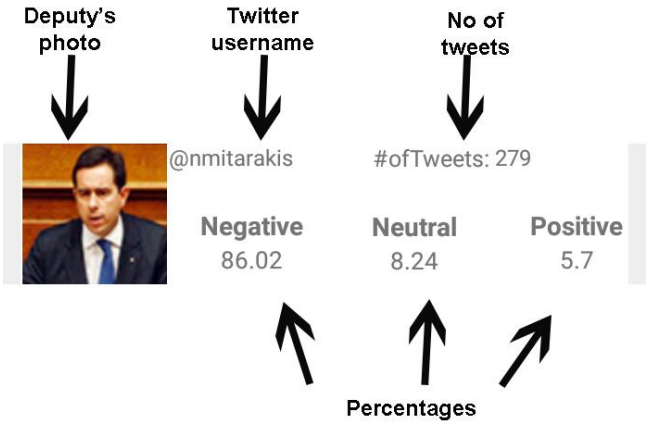
Picture 18: Total Statistics (Top Negative tab)

Every statistics screen (Activity) has the main title on the top so that the user knows what type of statistics are displayed. Every statistics screen contains two tabs: “Top Positive” tab and “Top Negative” tab.

In “Top Positive” tab all deputies are sorted in descending order starting with the person that has the greatest positive percentage. Similarly, in “Top Negative” tab all deputies are sorted in descending order starting with the person that has the greatest negative percentage.

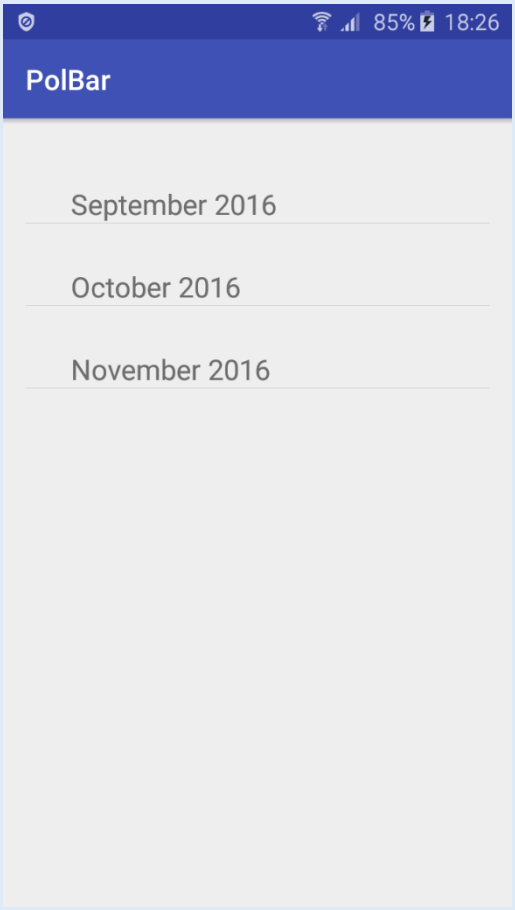
Both tabs display the results in ListView layout. Listview is an Android view that shows a list of scrollable items. Each item contains a parliamentarian’s picture (30x30 px), the Twitter username, the number of tweets that have been written for him/her (for the current statistics screen) and the percentages of positive, neutral and negative tweets.

The background colors of ListView items are white and light grey alternately in order to be distinguished.



Picture 19: List item details

List item description

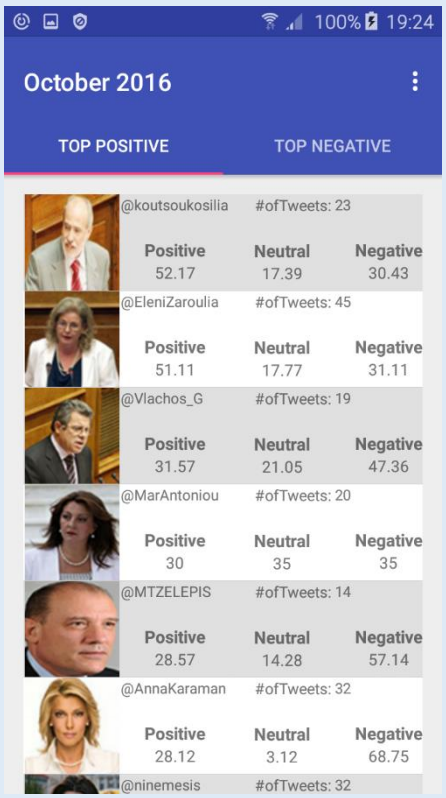


Picture 20: Month menu

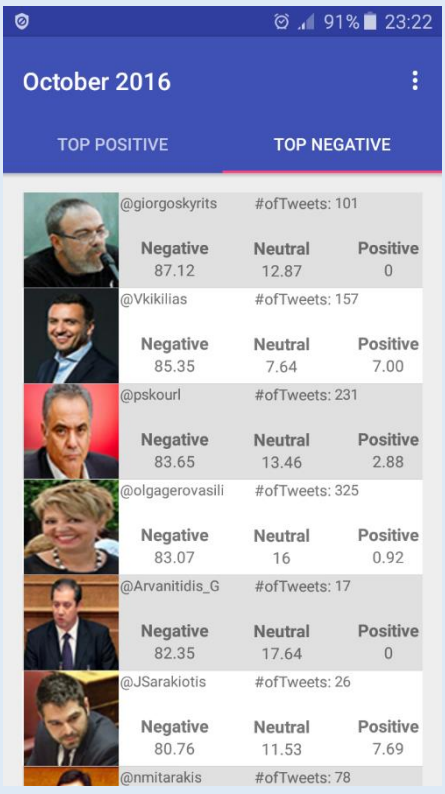
This is the month menu where user chooses the desired month.

Picture 21: September 2016 statistics (Top Positive tab)

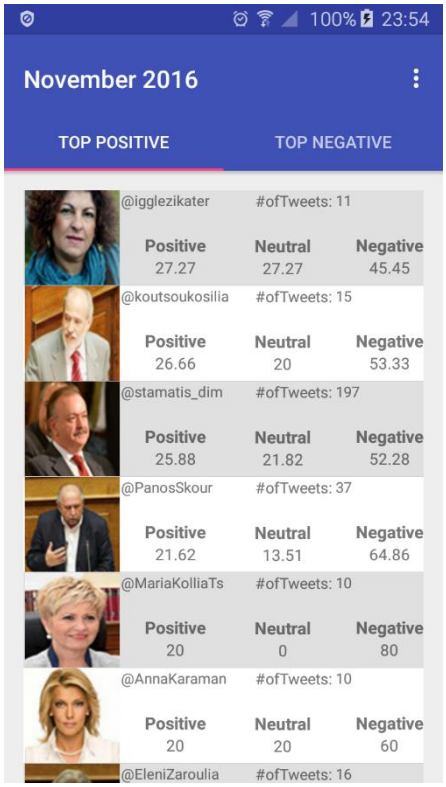
Picture 22: September 2016 statistics (Top Negative tab)



Picture 23: October 2016 statistics (Top Positive tab)



Picture 24: October 2016 statistics (Top Negative tab)



Picture 25: November 2016 statistics (Top Positive tab)

Picture 26: November 2016 statistics (Top Negative tab)

5.5 Extra Experiments

This section presents two experiments. These experiments were conducted in order the accuracy variation of classifiers to be examined when 2 factors in training set change: Type of dataset (dataset with no stopwords vs dataset with stopwords), and Size of dataset (dataset of 1.500 records vs dataset of 2.000 records). The number of classes in datasets are not changed (still 3 classes).

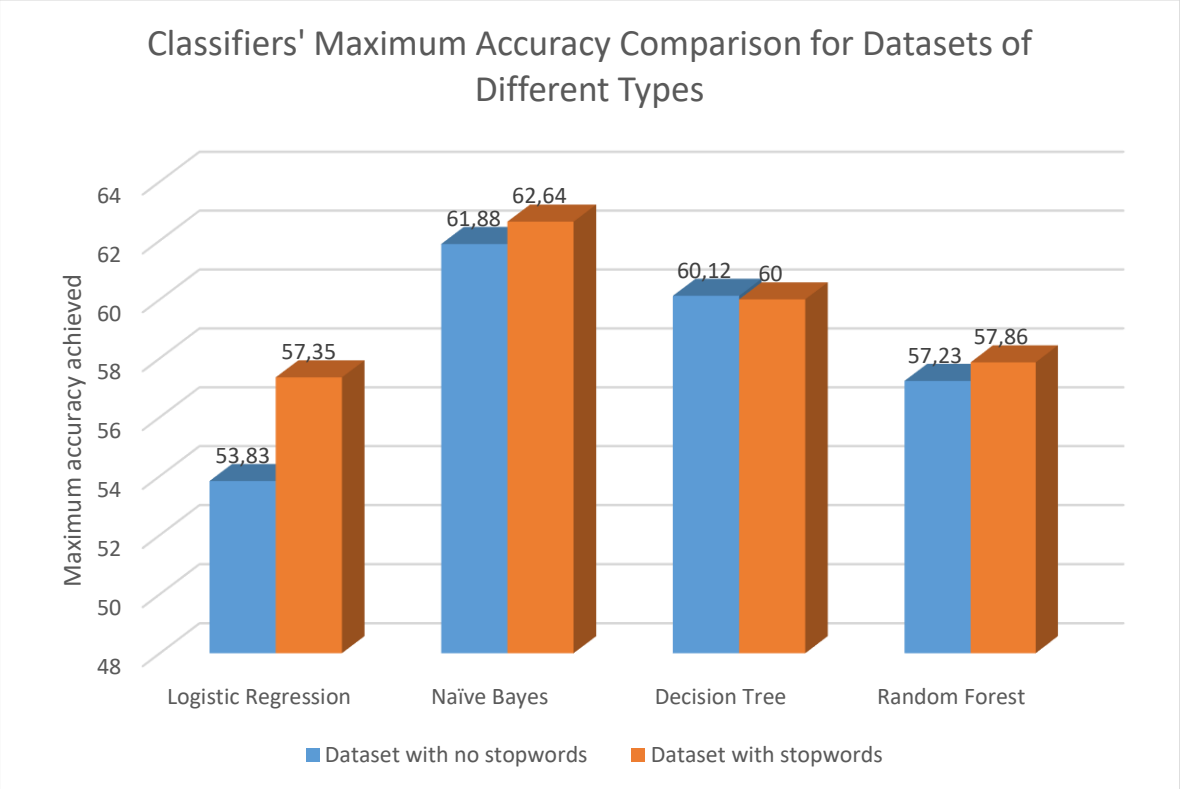
All charts quoted below, present the maximum accuracy achieved by each classifier testing several values for all possible parameters (λ parameter for Naïve Bayes, numBins and maxDepth for Decision Tree and numTrees and maxDepth for Random Forest). We don't care which parameters achieve the maximum accuracy, but what the maximum accuracy value is for each classifier and for each dataset. All experiments results introduced in next sections are come with tables and graphic charts for easy understanding. Firstly, the factor of dataset type is examined and secondly the factor of dataset size.

5.5.1 Experiment with different datasets types

This experiment uses one dataset of 2.000 records without stopwords and one with the same number of records with stopwords. The results are presented in table 18 and picture 27. As we can notice all algorithms (apart from Decision Tree) performed better accuracy in datasets where stopwords applied than in datasets where stopwords were not applied. Only "Decision Tree" presented higher accuracy in the first case (dataset with no stopwords) than in the second one, but the difference is not remarkable (0.12%). On the contrary, all other classifiers performed noticeably better accuracy in second case than in first one. This experiment helps us understand the utility of stopwords in text mining.

Table 18: Classifiers' maximum accuracy for two different types of datasets

Classifier	No stopwords	Stopwords
Logistic Regression	53.83	57.35
Naïve Bayes	61.88	62.64
Decision Tree	60.12	60
Random Forest	57.23	57.86



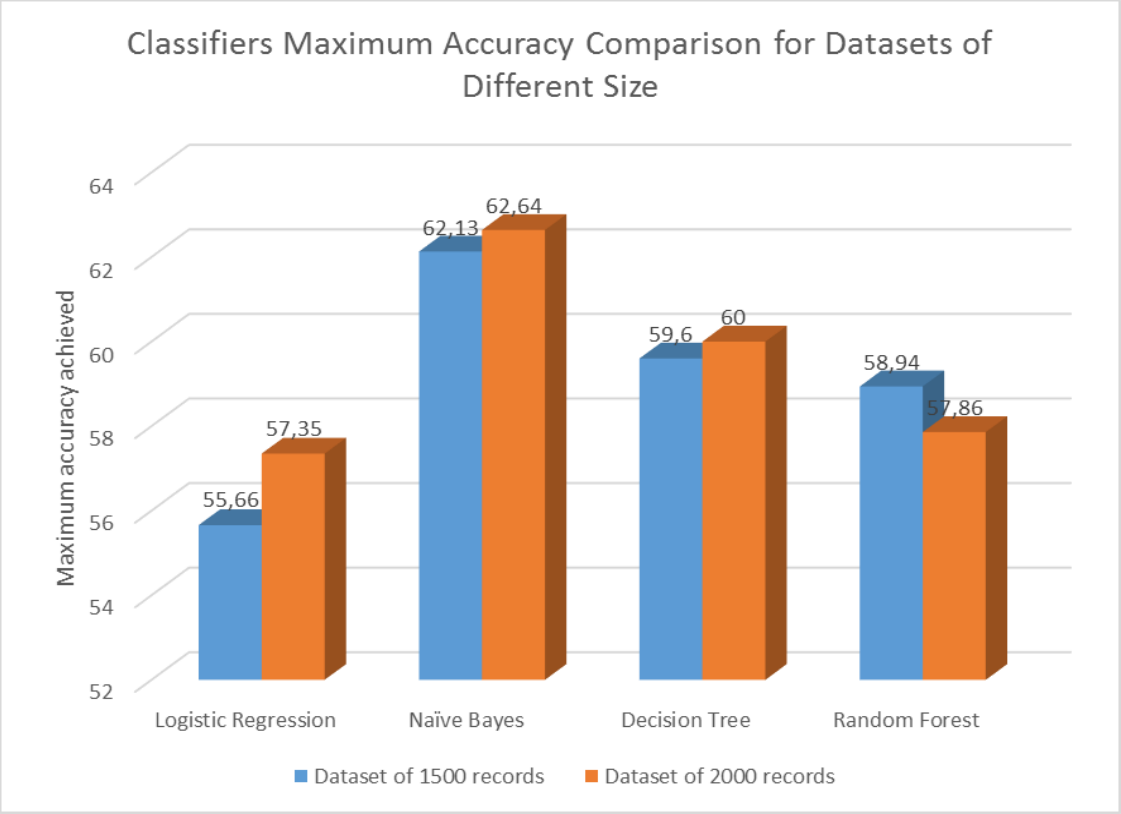
Picture 27: Classifiers' maximum accuracy comparison for datasets of different types

5.5.2 Experiment with different datasets size

This experiment uses two datasets: one with 1.500 records (159 positive, 490 neutral and 851 negative) and one with 2.000 records (197 positive, 718 neutral, 1085 negative). Stopwords were applied in both datasets. Table 19 and picture 28 present the results. 3 of 4 classifiers performed higher accuracy with the large dataset than the short one. "Random Forest" was the classifier that had better accuracy with the small dataset (1.08% better). As expected, the larger a dataset is, the better accuracy is achieved. So, if we want a much more accurate classifier, we need the largest possible dataset.

Table 19: Classifiers' maximum accuracy for two datasets of different size

Classifier	Dataset of 1.500 rec	Dataset of 2.000 rec
Logistic Regression	55.66	57.35
Naïve Bayes	62.13	62.64
Decision Tree	59.6	60
Random Forest	58.94	57.86



Picture 28: Classifiers' maximum accuracy comparison for different dataset sizes

6 Conclusions

Twitter is a global social network where anyone can express themselves. This project's aim was the collection, analysis and classification of tweets written by users for Greek politicians and the export of statistics for each politician's popularity using big data and text mining technologies. Statistics produced by tweets' sentiment analysis are presented by an Android mobile application visually. Several different technologies (Java, MySQL database, Scala and Android studio) were combined for the final result. Since this application uses Big Data management technologies, it is able to analyze more than thousands of tweets.

Tweets were classified in three categories (positive, neutral and negative) and four different classifiers were examined. All classification algorithms were described briefly. The classifier with the best accuracy was selected for the classification of all tweets. The maximum accuracy achieved (62.64%) is not impressive. It could be better if the training set and stopwords set were larger. Two extra experiments done, proved that accuracy can be improved if training set increases and stopwords are applied in it.

The majority of tweets are negative tweets. Twitter users post tweets in order to express their indignation and dissatisfaction to Greek politicians. Tweets collected for the 3 months (September, October, November) of autumn 2016 (56.021 tweets) were 61.43% negative, 31.93% neutral and only 6.63% positive. Twitter is used as a protest medium for political situation in Greece.

In general, tweets analysis and classification is a very difficult procedure, since every tweet is very different from all others. Criticism through Twitter differs from a simple review significantly. A review summarizes organized thoughts about a topic, whereas tweets are more casual, customized, targeted, character limited and lack of thought.

Every individual has its own way of writing that is unique. Accuracy decreases, because users tend to join or crop words so as to save characters. Tweets contain joined or cropped words that are impossible to be analyzed by the classifier correctly. This problem is getting worse by users who misspell words or do not use intonation.

"Problematic" tweets also affect accuracy negatively, since they are not clear where they are addressed. Most "problematic" tweets contain more than one Twitter usernames confusing both human and machine. One solution would be filtering tweets with more

than one usernames, so as only tweets with one username to be collected and stored. This method will reduce the number of collected tweets, but tweets will be unambiguous to the politician judged and, as a result, accuracy will be improved.

Furthermore, timeliness changes rapidly. New topics are arisen and discussed and new words are used that may not be included in training set. Therefore, the training set should be updated frequently (maybe every month or sooner). Otherwise, the classifier will "meet" new words without knowing how to classify them.

Moreover, it is not fair parliamentarians with thousand tweets to be compared with parliamentarians of tens tweets. A comparison like this is not representative, since the same percentage can produced by different number of tweets. Besides, this application need updates after elections, since new politicians are elected.

The primary goal of this study was to show how political reality of a country is reflected in social media, Twitter specifically. Electorate's feelings were analyzed satisfactorily. Analysis can be improved with further study. Politics is being held in Twitter and new findings will be discovered in near future.

7 Future Prospects

More actions can be done for accuracy improvement. As mentioned in previous chapter, “problematic” tweets can be faced with a filter that keeps tweets with only one username. The problem of joined and cropped words can be confronted with stemming and lemmatization procedures, which will make tweet analysis better. Emoticons harnessing is also very useful. A database with codification compatible with emoticons’ codification is necessary.

Besides, improvements can be done in mobile application. List items can provide more information for each depicted politician like the political party or the constituency they belong to. Of course, this demands the existence of an extra table in the database.

Statistics can be visualized with pie, line or bar charts. By clicking on a specific politician, aggregated month statistics can be returned to user in a chart format.

Finally, one extra option can be “My List” option. A simple user may be interested for some deputies only, not all. A list of favorite politicians is very useful. Users will choose the parliament members whose statistics they want to be displayed on screen.

Bibliography

1) Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, Samee Ullah Khan, *"The rise of "big data" on cloud computing: Review and open research issues"*, Information Systems 47 (2015) 98-115, Elsevier August 2014
2) Domenico Talia, *"Clouds for Scalable Big Data Analytics"*, IEEE Computer Society, 2013
3) Alec Go, Richa Bhayani, Lei Huang, *"Twitter Sentiment Classification using Distant Supervision"*,
4) Cukier K., *"Data, data everywhere",* (www.economist.com/node/15557443), 25/02/2010
5) Smith C, *"By the numbers: 17 Amazing Facebook User and Demographic Statistics"*, (http://expandedramblings.com/index.php/by-thenumbers-17-amazing-facebook-stats/), 25/08/2013
6) Iason (Jason) Katsamenis, *"Εξόρυξη Δεδομένων και Καταγραφή του Περιεχομένου του Διαδικτύου και των Κοινωνικών Δικτύων" (Data Mining and Internet and Social Networks Content Recording)*, Bachelor thesis, School of Applied Mathematics and Physical Sciences, National Technical University, September 2014
7) Min Chen, Shiwen Mao, Yunhao Liu, *"Big Data: A Survey"*, Springer Science + Business Media New York, Mobile Netw Appl (2014) 19: 171-209, 2014
8) Apache Spark official web site (www.spark.apach.org)
9) Jeffrey Dean, Sanjay Ghemawat, *"MapReduce: Simplified Data processing on Large Cluster"*, OSDI, 2004
10) Derrick Harris, *"Apache Mahout, Hadoop's original machine learning project is moving on from MapReduce"*, (https://gigaom.com/2014/03/27/apache-mahout-ha-doops-original-machine-learning-project-is-moving-on-from-mapreduce/), March 2014
11) Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, *"Learning Spark, Lightning- Fast data analysis"*, O' Reilly editions, 2015

12) Stephen Shankland, *"Google spotlights data center inner workings"*, (https://www.cnet.com/news/google-spotlights-data-center-inner-workings/), May 2008
13) Ralf Lammel, *"Google's MapReduce programming model- Revisited"*, ScienceDirect, Science of Computer Programming 70 (2008) 1-30, Elsevier 2007
14) Peng, D., Dabek, F. *"Large-scale Incremental Processing Using Distributed Transactions and Notifications."*, In OSDI (Vol. 10, pp. 1-15), October 2010
15) Typical Programmer, *"Relational Database Experts Jump the MapReduce Shark"*, (http://typicalprogrammer.com/mapreduce/)
16) Reza Bosagh, Gunnar Carlsson, *"Dimension Independent Matrix Square using MapReduce (DIMSUM)"*, October 2014
17) Derrick Harris, *"4 reasons why Spark could jolt Hadoop into hyperdrive"*, (https://gigaom.com/2014/06/28/4-reasons-why-spark-could-jolt-hadoop-into-hyperdrive/), June 2014
18) Welcome to Apache Hadoop (http://hadoop.apache.org/)
19) Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills, *"Advanced Analytics with Spark- Patterns for learning from data at scale"*, O'Reilly editions, 2015
20) Oracle & Beyond (http://oraclenbeyond.blogspot.gr/)
21) GitBook.com, (https://jaceklaskowski.gitbooks.io/mastering-apache-spark/content/spark-architecture.html)
22) Scala logo- Unixstickers.com (http://www.unixstickers.com/stickers/coding_stickers/scala-language-logo-shaped-sticker)
23) Jason Swartz, *"Learning Scala- Practical functional programming for the JVM"*, O' Reilly editions, 2015
24) Chuck Esterbrook, *"Using Mix-ins with Python"*, (http://www.linuxjournal.com/article/4540), Linux Journal, April 2001
25) Dean Wampler, Alex Payne, *"Programming Scala"*, O' Reilly editions, 2009
26) Generics in Java- Java programming language (http://docs.oracle.com/javase/1.5.0/docs/guide/language/index.html)
27) Twitter logo (http://dwan.tk/twitter-logo/)
28) Shamanth Kumar, Fred Morstatter, Huan Liu, *"Twitter Data Analytics"*, Springer, August 2013

29) Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, Alok Choudhary, *"Twitter Trending Topic Classification"*, 11th IEEE International Conference on Data Mining Workshops, 2011

30) Panagiota Kavatha, *"Twitter, το απόλυτο πολιτικό trend" (Twitter- The absolute political trend)*, (*http://news247.gr/eidiseis/politiki/twitter-to-apolyto-politiko-trend.3965305.html*), March 2016

31) Aikaterini Tsagalidou, *"Semantic definition of views and subjective classification of posts in social networks"*, Case study Twitter. Aristotle University of Thessaloniki, 2011

32) Ahlqvist, T., Asta, B., Halonen, M., Heinonen, S., *"Social media road maps exploring the futures triggered by social media"*, 2008

33) Stross, R. *"When History Is Compiled 140 Characters at a Time."*, The New York Times, (http://www.nytimes.com/2010/05/02/business/02digi.html?scp=1&sq=twitter%20+%20history&st=cse&_r=1&), May 2010

34) Twitter Analytics (https://analytics.twitter.com)

35) Theodoros Konsoulas, *"Τι είναι το hashtag και πως το χρησιμοποιώ;" (What is hashtag and how do I use it?)*, (http://www.socialmedialife.gr/110564/ti-einai-to-hashtag-kai-pos-to-xrisimopoio/), October 2014

36) Twitter Developers (https://dev.twitter.com/)

37) Wei Gao, Fabrizio Sebastiani, *"Tweet Sentiment: From Classification to Quantification"*, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015

38) Alexandra Kassimi, *"Το δημοψήφισμα διεξάγεται ήδη στο Twitter" (Referendum has already been held in Twitter)*, (http://www.kathimerini.gr/821995/article/epikairothta/politikh/to-dhmoyhfisma-die3agetai-hdh-sto-twitter), July 2015

39) Jisue Lee, Hohyon Ryu, Lorri Mon, Sung Jae Park, *"Citizens' Use of Twitter in Political Information Sharing in South Korea."*, iConference 2013 Proceedings (pp. 351-365). doi:10.9776/13210, 2013

40) Lifo Team, *"Παρακμιακός πολιτικός λόγος ή πως το Twitter διαμορφώνει τη νέα πολιτική ζωή" (Decadent political discourse or how Twitter form the new political life)*, (http://www.lifo.gr/articles/digital-media_articles/109419), August 2016

41) Andreas Jungherr, *"Twitter in Politics: A Comprehensive Literature Review"*, February 2014

42) Ferran Pla, Lluis Hurtado, *"Political Tendency Identification in Twitter using Sentiment Analysis Techniques"*, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 183–192, Dublin, Ireland, August 23-29 2014

43) Anders Olof Larsson, Hallvard Moe, *"Studying political microblogging. Twitter users in the 2010 Swedish election campaign"*, New Media Society, 2010

44) Marti Hearst, *"What is Text Mining?"*, (http://people.ischool.berkeley.edu/~hearst/text-mining.html), October 2003

45) Jiawei Han, Micheline Kamber, Jian Pei, *"Data Mining- Concepts and Techniques (3rd edition)"*, Morgan Kaufmann editions, 2012

46) Techopedia- Full text database, (https://www.techopedia.com/definition/7348/full-text-database)

47) David Hand, Heikki Mannila, Padhraic Smyth, *"Principles of Data Mining"*, MIT press, 2001

48) Freedictionary- Information Retrieval, (http://www.thefreedictionary.com/information+retrieval)

49) Charu C. Aggarwal, ChengXiang Zhai, *"Mining Text Data"*, Springer Science + Business Media, 2012

50) Matthew A. Russell, *"Mining the Social Web- Data mining, Facebook, Twitter, Linkedin, Google+, Github, and more (2nd edition)"*, O' Reilly editions, 2014

51) Seymour Lipschutz, Marc Lars Lipson, *"Schaum's Outline of Theory and Problems of Linear Algebra (3rd edition)"*, McGraw- Hill Inc, 2001

52) TFxIDF Repository, (http://kak.tx0.org/IR/TFxIDF)

53) WordNet- A lexical database for English, (https://wordnet.princeton.edu/)

54) Sushmita Mitra, Tinku Acharya, *"Data Mining- Multimedia, Soft Computing and Bioinformatics"*, Wiley editions, 2003

55) Twitter4j logo, (https://wiki.smu.edu.sg/is480/IS480_Team_wiki%3A2014T1_Chie_Project_Documentation_Technologies)

56) Twitter4j official site, (http://twitter4j.org/en/index.html)

57) Apache Spark- Classification and Regression, (https://spark.apache.org/docs/latest/mllib-classification-regression.html)

58) Trevor Hastie, Robert Tibshirani, Jerome Friedman, *"The Elements of Statistical Learing- Data Mining, Inference and Prediction (2nd edition)"*, Springer editions, August 2008

59) Apache Spark- Linear Methods, (https://spark.apache.org/docs/latest/mllib-linear-methods.html)
60) Apache Spark- Naïve Bayes, (https://spark.apache.org/docs/latest/mllib-naive-bayes.html)
61) Jason D.M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger, *"Tackling the Poor Assumptions of Naïve Bayes Text Classifiers"*, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003
62) C.D. Manning, P. Raghavan and M. Schütze. "*Introduction to Information Retrieval*", Cambridge University Press, p. 260, 2008
63) Ian H. Witten & Eibe Frank, *"Data Mining- Practical Machine Learning Tools and Techniques (2nd edition)"*, Morgan Kaufmann editions, 2005
64) Statistics How To- What is a Bin in Statistics, (http://www.statisticshowto.com/what-is-a-bin-in-statistics/)
65) Apache Spark- Decision Trees, (https://spark.apache.org/docs/latest/mllib-decision-tree.html)
66) Aparche Spark- Ensembles, (https://spark.apache.org/docs/latest/mllib-ensembles.html)
67) Techopedia- Software Development Kit, (https://www.techopedia.com/definition/3878/software-development-kit-sdk)
68) Android developers- What is API level?, (https://developer.android.com/guide/topics/manifest/uses-sdk-element.html#ApiLevels)
69) What are the benefits of MVC? , (http://blog.iandavis.com/2008/12/what-are-the-benefits-of-mvc/), 2008

Appendix A

Instance of Data Table

ID	Person	Tweet	Status	Day	Month	Year	InsertedIn
46573	@YDragasakis	.@VasilisSkouris Το μεγάλο “όχι” του @YDragasakis σε πρόωρες εκλογές και… οικουμενική ://t.co/FeDFd9p0s3 … #Ysterografa @atsipras	0	6	11	2016	11/10/16
46574	@YDragasakis	@YDragasakis @Real_gr δεν ξέρετε τι σας γίνεται. Και μας τακαιπωρειτε. Ντρέπομαι για σας	1	5	11	2016	11/10/16
46575	@YDragasakis	@YDragasakis @Real_gr Αυτά να προσέξετε αγαπητέ ...://t.co/o1JD2CY5hf	1	5	11	2016	11/10/16
46576	@YDragasakis	.@YDragasakis @Real_gr ΓΚΕΟΥΡΟ, ΠΡΑΝΙΚ ΧΗΛΙΝΓΚ, ΚΑΤΕΒΑΣΜΕΝΑ ΑΚΚΑΟΥ ΚΑΙ ΤΖΕΡΟΝΥΜΟ ΡΟΥΛΣ.	0	5	11	2016	11/10/16
46577	@YDragasakis	@YDragasakis όταν λες "χειροπιαστά" ποιό ακριβώς ...στοιχείο έπιασες; Άσε μας κουκλίτσε μου!	1	3	11	2016	11/10/16
46578	@YDragasakis	@YDragasakis Τι "κάθεστε και τους ρωτάτε" βγάλτε τα τανκς. Άλλωστε μόνοι σας είπατε ότι έγινε πραξικόπημα κ το 4ο Ράιχ εισέβαλε στην Ελλάδα.	2	2	11	2016	11/10/16
46579	@TheanoFotiou	@TheanoFotiou καλησπέρα θα ήθελα να ρωτησω το οικογενειακό επίδομα τέταρτη δόση ποτέ θα δοθεί για το 2016	1	10	11	2016	11/10/16
46580	@TheanoFotiou	@TheanoFotiou δεν εξοργίστηκε αμάν η προπαγάνδα... της έκοψε απλά το αυγολέμονο..	1	10	11	2016	11/10/16
46581	@TheanoFotiou	@TheanoFotiou απατεωνες ειστε;	1	9	11	2016	11/10/16
46582	@TheanoFotiou	@TheanoFotiou Κυρά Θεανώ μη μιλάτε πλέον για ΑΞΙΕΣ.Τις γελοιοποιήσατε όλες	0	6	11	2016	11/10/16
46583	@TheanoFotiou	@TheanoFotiou Ετσι. Γαργάρα.	0	6	11	2016	11/10/16
46584	@TheanoFotiou	@TheanoFotiou μπορείς σε δύο παραγράφους να απαριθμήσεις αυτές τις "αξίες" & τα τυπικά προσόντα της νέας υπουργού για τη συγκεκριμένη θέση;	1	6	11	2016	11/10/16
46585	@TheanoFotiou	Λολ ρε @TheanoFotiou!Αξία το λέμε τώρα; ://t.co/TOWn2ELhmd	1	6	11	2016	11/10/16
46586	@TheanoFotiou	@TheanoFotiou @NikSfikas Ξέρει κι αυτή να πετάει μαλακίες για γεμιστά?!	1	5	11	2016	11/10/16
46587	@TheanoFotiou	@PeriplanomenosP @TheanoFotiou Αυτό μάλλον το εξέλαβε ως ευχή για αυτό δεν το διαγράφει	1	5	11	2016	11/10/16
46588	@TheanoFotiou	@TheanoFotiou Η μόνη σας αξία είναι τα γεμιστά. Για τις άλλες είστε τόσο περηφανες που τις διαγράφετε για να μη φανεί η κωλοτουμπα.	1	5	11	2016	11/10/16
46589	@TheanoFotiou	@TheanoFotiou δε γαμιεσαι ρε βλημα κι εσυ	1	5	11	2016	11/10/16
46590	@TheanoFotiou	@TheanoFotiou πες τα θεανω μου.πες μας και για τη μνημονιακη στροφη της	1	5	11	2016	11/10/16

		αριστερας που τοσο πιστα ακολουθεις.φερτε κ το 4ο να γουσταρουμε					
46591	@TheanoFotiou	@TheanoFotiou αυτή τη δουλειά θα κάνουμε όλο το βράδυ;	1	5	11	2016	11/10/16
46592	@TheanoFotiou	@TheanoFotiou Γιατί τα διαγράφεις; Δεν αντέχει την αλήθεια η αριστερή φασιστική οργάνωση ΣΥΡΙΖΑ;	1	5	11	2016	11/10/16
46593	@TheanoFotiou	@TheanoFotiou η τελευταία πράξη του δράματος ΠΡΕΠΕΙ να παιχτεί σε δικαστήριο	0	5	11	2016	11/10/16
46594	@TheanoFotiou	@TheanoFotiou Θα μπορούσα μέχρι το ξημέρωμα να αναρτώ τις "αξίες" της κατοχικής κυβέρνησης του 4ου αριστερού Ράιχ	1	5	11	2016	11/10/16
46595	@TheanoFotiou	@TheanoFotiou Η αριστερή “δικαιοσύνη” που καταργεί το “άδικο” επίδομα των χαμηλοσυνταξιούχων ακόμη μια "ανεκτίμητη… ://t.co/wRass459WO	1	5	11	2016	11/10/16
46596	@TheanoFotiou	@TheanoFotiou Αστους Θεανω μου τους φιλελεδες να λενε. Αλήθεια στα γεμιστα βαζεις κιμα ή τα κανεις ορφανα;	0	5	11	2016	11/10/16
46597	@TheanoFotiou	@TheanoFotiou Είστε τόσο περηφανες για τις αξίες σας που σβηνετε τα γραφόμενα σας για να μη φανεί η κωλοτουμπα.	0	5	11	2016	11/10/16
46598	@TheanoFotiou	@TheanoFotiou Τις αξίες της αριστεράς θα τις συζητήσουμε στο Ειδικό Δικαστήριο	2	5	11	2016	11/10/16
46599	@TheanoFotiou	@TheanoFotiou Άλλη μια αριστερή "αξία" η υποδούλωση της Ελλάδας για 99 χρόνια... ://t.co/OXtcA34idF	1	5	11	2016	11/10/16
46600	@TheanoFotiou	@TheanoFotiou Η μόνη αξία σας είναι τα γεμιστά.	1	5	11	2016	11/10/16
46601	@TheanoFotiou	@TheanoFotiou Το ξεπούλημα της δημόσιας περιουσίας κ η ολοκλήρωση του έργου της δεξιάς οι βασικές αριστερές “αξίες”… ://t.co/ZfUuhktLKF	2	5	11	2016	11/10/16
46602	@TheanoFotiou	@TheanoFotiou το άσχημο είναι ότι έφυγε ο κατρουγκαλος	1	5	11	2016	11/10/16
46603	@TheanoFotiou	@TheanoFotiou Την ξέρουμε από την σχολή!Κάνει καλό κρεβάτι!Μπράβο της που έγινε Υπουργός!Έχω γαμήσει μία υπουργό!!Και είναι και εύκολη!!	0	5	11	2016	11/10/16
46604	@TheanoFotiou	@TheanoFotiou Για τα φημολογουμενα γκομενικα της λέει;	1	5	11	2016	11/10/16
46605	@TheanoFotiou	@TheanoFotiou από αξίες αριστεράς γκώσαμε ! Κυρίως τις ακροδεξιές αξίες π πρεσβεύει ο Καμμένος. όσο για τη νεα υπουργό... γεμιστά φτιάχνει ?	1	5	11	2016	11/10/16
46606	@TheanoFotiou	@TheanoFotiou Και την αριστερά μάθαμε καλά Και την νεα...80ετών Αχτσιογλου μάθαμε πολύ νωρίς ://t.co/Dajng30Eon	1	5	11	2016	11/10/16
46607	@TheanoFotiou	.@TheanoFotiou ΕΠΕΙΓΟΝΤΩΣ ΠΡΑΝΙΚ ΧΗΛΙΝΓΚ ΣΤΟ ΚΡΑΝΙΟ, ΝΑ ΦΥΓΕΙ ΤΟ ΦΟΡΤΩΜΑ.	2	5	11	2016	11/10/16
46608	@TheanoFotiou	Συγνώμη ρε παιδιά, συγκρίνεται η @TheanoFotiou με αυτή την κρεμανταλού την Αχτσιογλου;; τι ξέρει από μαγειρική αυτή… ://t.co/qaiST7Sb1V	1	5	11	2016	11/10/16

Appendix B

Instance of Month Statistics Table

ID	Person	Positive	Neutral	Negative	Perc_Positive	Perc_Neutral	Perc_Negative	Sum	Month	Year
1	@NikosKotzias	13	147	85	5,30612	60	34,6939	245	9	2016
2	@K_Hatzidakis	29	91	217	8,60534	27,003	64,3917	337	9	2016
3	@PanosKammenos	64	77	279	15,2381	18,3333	66,4286	420	9	2016
4	@olgagerovasili	8	45	350	1,98511	11,1663	86,8486	403	9	2016
5	@nasosa8anasiou	0	0	6	0	0	100	6	9	2016
6	@MakisVoridis	68	79	221	18,4783	21,4674	60,0543	368	9	2016
7	@BKegeroglou	40	100	276	9,61539	24,0385	66,3462	416	9	2016
8	@YDragasakis	39	80	216	11,6418	23,8806	64,4776	335	9	2016
9	@TheanoFotiou	16	61	154	6,92641	26,4069	66,6667	231	9	2016
10	@fortsakis	1	21	7	3,44828	72,4138	24,1379	29	9	2016
11	@stamatis_dim	51	71	215	15,1335	21,0682	63,7982	337	9	2016
12	@nkerameus	3	28	31	4,83871	45,1613	50	62	9	2016
13	@PappasXA	39	30	135	19,1176	14,7059	66,1765	204	9	2016
14	@theocharop	0	0	0	0	0	0	0	9	2016
15	@Paparigaleka	0	0	0	0	0	0	0	9	2016
16	@nikosfilis1	8	97	279	2,08333	25,2604	72,6562	384	9	2016
17	@chrisvernard	0	1	5	0	16,6667	83,3333	6	9	2016
18	@olgakef	4	78	73	2,58065	50,3226	47,0968	155	9	2016
19	@Dora_Bakoyannis	15	180	189	3,90625	46,875	49,2188	384	9	2016
20	@nkaklamanis	4	32	70	3,77358	30,1887	66,0377	106	9	2016
21	@KSkandalidis	2	2	15	10,5263	10,5263	78,9474	19	9	2016
22	@ElenaKountoura	5	72	64	3,5461	51,0638	45,3901	141	9	2016
23	@Mar_Georgiadis	0	3	1	0	75	25	4	9	2016
24	@kouroumplis	4	26	84	3,50877	22,807	73,6842	114	9	2016
25	@pskourl	6	49	176	2,5974	21,2121	76,1905	231	9	2016
26	@giorgosdimaras1	4	0	4	50	0	50	8	9	2016
27	@cspirtzis	0	0	0	0	0	0	0	9	2016
28	@TsironisGianni	0	23	10	0	69,697	30,303	33	9	2016
29	@dbbda1ef08c242b	0	0	2	0	0	100	2	9	2016
30	@annetakavadia	0	19	28	0	40,4255	59,5745	47	9	2016
31	@giorgoskyritsis	7	58	302	1,90736	15,8038	82,2888	367	9	2016
32	@v_meimarakis	1	18	7	3,84615	69,2308	26,9231	26	9	2016
33	@NikosDendias	20	81	295	5,05051	20,4545	74,4949	396	9	2016
34	@AnnaKaramanli	2	8	21	6,45161	25,8065	67,7419	31	9	2016
35	@Sofia_voultepsi	3	2	7	25	16,6667	58,3333	12	9	2016
36	@AnnaAsimakopoul	11	115	118	4,5082	47,1311	48,3607	244	9	2016

37	@GerGiakoumatos	0	5	15	0	25	75	20	9	2016
38	@kpapakosta	0	8	21	0	27,5862	72,4138	29	9	2016
39	@GermenisGiorgos	0	5	2	0	71,4286	28,5714	7	9	2016
40	@ipanagiotaros	4	46	50	4	46	50	100	9	2016
41	@EleniZaroulia	25	13	24	40,3226	20,9677	38,7097	62	9	2016
42	@grpsarianos	0	16	31	0	34,0426	65,9574	47	9	2016
43	@amirasgiorgos	2	13	15	6,66667	43,3333	50	30	9	2016
44	@PapachristTh	2	0	2	50	0	50	4	9	2016
45	@ProedrosEK	13	64	136	6,10329	30,0469	63,8498	213	9	2016
46	@PanagoulisSt	0	0	0	0	0	0	0	9	2016
47	@SalmasMarios	0	0	0	0	0	0	0	9	2016
48	@k_karagounis	0	4	6	0	40	60	10	9	2016
49	@Barbarousis	5	7	22	14,7059	20,5882	64,7059	34	9	2016
50	@dkonstantop	1	14	5	5	70	25	20	9	2016
51	@Yannis_Maniatis	6	22	55	7,22892	26,506	66,2651	83	9	2016
52	@KostasVlasis	0	0	1	0	0	100	1	9	2016
53	@Odysseas_	22	181	217	5,2381	43,0952	51,6667	420	9	2016
54	@VasilisTsirkas	0	1	3	0	25	75	4	9	2016
55	@gstylios	0	0	0	0	0	0	0	9	2016
56	@tsakalotos	3	62	82	2,04082	42,1769	55,7823	147	9	2016
57	@PanosSkourolia1	8	19	55	9,7561	23,1707	67,0732	82	9	2016
58	@gpantzas	0	1	6	0	14,2857	85,7143	7	9	2016
59	@GeorgiaMartinou	0	0	0	0	0	0	0	9	2016
60	@Vlachos_G	0	0	0	0	0	0	0	9	2016
61	@IliasKasidiaris	17	172	70	6,56371	66,4093	27,027	259	9	2016
62	@evichrist	6	25	63	6,38298	26,5957	67,0213	94	9	2016
63	@dimitris176	0	0	3	0	0	100	3	9	2016
64	@sia_anag	1	3	17	4,7619	14,2857	80,9524	21	9	2016
65	@katsaniotis	0	0	4	0	0	100	4	9	2016
66	@papatheodorou_t	1	25	4	3,33333	83,3333	13,3333	30	9	2016
67	@IasonFotilas	7	45	51	6,79612	43,6893	49,5146	103	9	2016
68	@EKarakostas8	0	0	0	0	0	0	0	9	2016
69	@kyriazidisdim	2	0	0	100	0	0	2	9	2016
70	@xarakefalidou	2	2	2	33,3333	33,3333	33,3333	6	9	2016
71	@nectarsant	0	5	17	0	22,7273	77,2727	22	9	2016
72	@dgakis	0	0	0	0	0	0	0	9	2016
73	@kamateros_hlias	0	0	0	0	0	0	0	9	2016
74	@manoskonsolas	1	9	11	4,7619	42,8571	52,381	21	9	2016
75	@KremastinosD	0	0	0	0	0	0	0	9	2016
76	@kaisasgeorgios	0	1	0	0	100	0	1	9	2016
77	@NatasaGkara	0	0	0	0	0	0	0	9	2016
78	@pavpol2222	7	30	196	3,00429	12,8755	84,1202	233	9	2016
79	@g_stathakis	0	53	45	0	54,0816	45,9184	98	9	2016
80	@amixelis	0	0	0	0	0	0	0	9	2016

Appendix C

Instance of Total Statistics Table

Person	Positive	Neutral	Negative	Perc_Positive	Perc_Neutral	Perc_Negative	Sum
@NikosKotzias	78	305	342	10,7586	42,069	47,1724	725
@K_Hatzidakis	92	288	628	9,12698	28,5714	62,3016	1008
@PanosKammenos	147	202	905	11,7225	16,1085	72,1691	1254
@olgagerovasili	13	135	901	1,23928	12,8694	85,8913	1049
@nasosa8anasiou	3	17	103	2,43902	13,8211	83,7398	123
@MakisVoridis	186	245	799	15,122	19,9187	64,9594	1230
@BKegeroglou	86	239	592	9,37841	26,0632	64,5583	917
@YDragasakis	81	202	551	9,71223	24,2206	66,0671	834
@TheanoFotiou	29	98	266	7,37913	24,9364	67,6845	393
@fortsakis	4	86	29	3,36134	72,2689	24,3697	119
@stamatis_dim	179	217	590	18,1542	22,0081	59,8377	986
@nkerameus	9	114	112	3,82979	48,5106	47,6596	235
@PappasXA	87	78	287	19,2478	17,2566	63,4956	452
@theocharop	0	3	6	0	33,3333	66,6667	9
@Paparigaleka	0	0	0	0	0	0	0
@nikosfilis1	17	213	563	2,14376	26,86	70,9962	793
@chrisvernard	2	10	35	4,25532	21,2766	74,4681	47
@olgakef	23	255	232	4,5098	50	45,4902	510
@Dora_Bakoyannis	82	496	573	7,12424	43,093	49,7828	1151
@nkaklamanis	25	157	370	4,52899	28,442	67,029	552
@KSkandalidis	8	29	84	6,61157	23,9669	69,4215	121
@ElenaKountoura	18	130	181	5,47112	39,5137	55,0152	329
@Mar_Georgiadis	3	6	11	15	30	55	20
@kouroumplis	8	66	166	3,33333	27,5	69,1667	240
@pskourl	15	103	569	2,18341	14,9927	82,8239	687
@giorgosdimaras1	4	2	8	28,5714	14,2857	57,1429	14
@cspirtzis	0	0	0	0	0	0	0
@TsironisGianni	2	40	30	2,77778	55,5556	41,6667	72
@dbbda1ef08c242b	0	3	6	0	33,3333	66,6667	9
@annetakavadia	7	81	186	2,55474	29,562	67,8832	274
@giorgoskyritsis	11	86	474	1,92644	15,0613	83,0123	571
@v_meimarakis	4	119	43	2,40964	71,6867	25,9036	166
@NikosDendias	43	284	700	4,18695	27,6534	68,1597	1027
@AnnaKaramanli	13	11	52	17,1053	14,4737	68,4211	76
@Sofia_voultepsi	4	8	9	19,0476	38,0952	42,8571	21
@AnnaAsimakopoul	40	271	324	6,29921	42,6772	51,0236	635
@GerGiakoumatos	1	22	35	1,72414	37,931	60,3448	58

@kpapakosta	8	121	139	2,98507	45,1493	51,8657	268
@GermenisGiorgos	2	14	22	5,26316	36,8421	57,8947	38
@ipanagiotaros	17	126	139	6,02837	44,6809	49,2908	282
@EleniZaroulia	51	30	51	38,6364	22,7273	38,6364	132
@grpsarianos	22	80	150	8,73016	31,746	59,5238	252
@amirasgiorgos	10	47	131	5,31915	25	69,6809	188
@PapachristTh	4	0	6	40	0	60	10
@ProedrosEK	33	119	292	7,43243	26,8018	65,7658	444
@PanagoulisSt	0	0	0	0	0	0	0
@SalmasMarios	0	0	1	0	0	100	1
@k_karagounis	0	8	26	0	23,5294	76,4706	34
@Barbarousis	9	28	68	8,57143	26,6667	64,7619	105
@dkonstantop	2	29	6	5,40541	78,3784	16,2162	37
@Yannis_Maniatis	17	73	152	7,02479	30,1653	62,8099	242
@KostasVlasis	2	2	1	40	40	20	5
@Odysseas_	58	388	514	6,04167	40,4167	53,5417	960
@VasilisTsirkas	0	2	8	0	20	80	10
@gstylios	1	2	1	25	50	25	4
@tsakalotos	12	141	204	3,36134	39,4958	57,1429	357
@PanosSkourolia1	18	25	86	13,9535	19,3798	66,6667	129
@gpantzas	2	2	7	18,1818	18,1818	63,6364	11
@GeorgiaMartinou	0	0	0	0	0	0	0
@Vlachos_G	6	7	10	26,087	30,4348	43,4783	23
@IliasKasidiaris	41	304	161	8,10277	60,0791	31,8182	506
@evichrist	16	82	173	5,90406	30,2583	63,8376	271
@dimitris176	0	0	4	0	0	100	4
@sia_anag	1	5	25	3,22581	16,129	80,6452	31
@katsaniotis	0	4	13	0	23,5294	76,4706	17
@papatheodorou_t	6	49	14	8,69565	71,0145	20,2899	69
@IasonFotilas	11	96	114	4,97738	43,4389	51,5837	221
@EKarakostas8	0	0	0	0	0	0	0
@kyriazidisdim	2	2	0	50	50	0	4
@xarakefalidou	2	15	17	5,88235	44,1176	50	34
@nectarsant	0	10	30	0	25	75	40
@dgakis	0	0	0	0	0	0	0
@kamateros_hlias	0	0	0	0	0	0	0
@manoskonsolas	4	18	22	9,09091	40,9091	50	44
@KremastinosD	1	2	2	20	40	40	5
@kaisasgeorgios	0	1	0	0	100	0	1
@NatasaGkara	0	0	0	0	0	0	0
@pavpol2222	15	87	385	3,08008	17,8645	79,0554	487
@g_stathakis	3	148	125	1,08696	53,6232	45,2899	276
@amixelis	0	0	0	0	0	0	0
@EVenizelos	93	487	619	7,75646	40,6172	51,6264	1199