

DESIGN AND DEVELOPMENT OF A WEB-BASED DATA VISUALIZATION
SOFTWARE FOR POLITICAL TENDENCY IDENTIFICATION OF TWITTER'S USERS
USING PYTHON DASH FRAMEWORK

ALEXANDROS BRITZOLAKIS

Bachelor of Applied Science (BA.Sc.) in Informatics Engineering,
Technological Educational Institute of Crete
(Now known as Hellenic Mediterranean University), 2016

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL
AND COMPUTER ENGINEERING

SCHOOL OF ENGINEERING

HELLENIC MEDITERRANEAN UNIVERSITY



2020

Approved by:

Major Professor
Dr. Nikolaos Papadakis

*“We must look beneath every stone, lest it
conceal some politician ready to sting us”*

*Aristophanes, 445-386 BC,
Ancient Greek comic playwright - Thesmophoriazusa*

Copyright

ALEXANDROS BRITZOLAKIS

2020

The content provided in this dissertation is registered under the copyrights of the Hellenic Mediterranean University – H.M.U as well as of its author and can be used only for research or educational purposes. Any reuse of the content material beyond of the academic context it is prohibited without the legal consent of the university and of its author.

Abstract

The rapid evolution of computers as well as the emerge of the internet brought a new era on the field of communication systems. Many individuals can instantly communicate with each other (through instant messaging or through a video conference). This technological breakthrough set the stage for the emerge of the first internet communities. As a result, social media platforms emerged, were a set of free services is provided such as the interactive communication, multimedia content uploading etc. This new type of communication shapes the way by which an opinion can be expressed. Twitter is the most popular micro-blogging platform since its users can post a text of 280 maximum characters for a variety of subjects such as famous brands - products, celebrities, prominent events including political ones. As a result, Twitter is a tool that politicians tend to use frequently as it is a source for obtaining voters. This master thesis presents a web-based application that will use Twitter's API in order to obtain the most recent Tweets of the top three Greek political leaders and thus identify their additional popularity. To achieve this a set sample of recent posted tweets will be obtained (e.g. 200 tweets) from their additional Twitter accounts. These tweets will be processed in order to extract structured and unstructured data and present them in a form of graph series through a web-page. The structure of this web-based application consists of a front-end part created with HTML/CSS and a back-end mechanism which is developed using Python and Dash framework for the visualization process as well as Tweepy module for the application's intercommunication with Twitter's servers in order to obtain the data. More specifically the extracted information will be the number of likes, re-tweets and characters per posted tweet as well as the number of followers where they have. Furthermore, sentiment analysis of the tweet's text is identified and visualized, using the Greek version of SpaCy module and labeled according to their corresponding expressed emotion. The extracted data will be used in order to display a set of charts that will present a comparison between these three political leaders. The research purpose of this dissertation is to present an engineering perspective on what data can mined from Twitter, how these data can be useful in order to estimate a political result as well as well as presenting the capabilities of Python Dash framework, Tweepy and SpaCy modules.

Table of Contents

List of tables	I
List of figures	II
Acknowledgments	V
Dedication	VI
Chapter 1: Introduction to Social Media analytics	- 1 -
1.2. The evolution of e-communication: How the Internet emerged	- 5 -
1.3. The transition from static (Web 1.0) to interactive content (Web 2.0)	- 12 -
1.4. The emerge of the Semantic Web (Web 3.0).....	- 15 -
1.5. Analyzing social networking and social media platforms	- 16 -
1.6. An overview on social media analytics	- 20 -
1.7. Summary	- 22 -
Chapter 2: Text mining methodologies over Twitter	- 23 -
2.1 Collecting and analyzing data from Twitter	- 23 -
2.2 Lexicon-Based Sentiment Analysis	- 27 -
2.3 Identifying the issues of sentiment lexicons	- 33 -
2.4 Sentiment Analysis with Machine Learning	- 34 -
2.5 Classification algorithms in Machine Learning	- 35 -
2.6 Summary	- 38 -
Chapter 3: AthPPA a tool for analyzing political popularity over Twitter	- 39 -
3.1 The purpose of this web application.....	- 39 -
3.2 Technical structure and methodology of AthPPA	- 39 -
3.3 AthPPA graphs results.....	- 44 -
3.4 Summary	- 54 -
References	- 56 -
Useful Links	- 58 -

List of tables

Chapter 1

Table 1.1: Most well-known chat based social networking platforms [23] - 18 -

Table 1.2: Most well-known chat based social networking platforms [23] - 19 -

Chapter 2

Table 2.1: Engagement features of Twitter [28] - 23 -

Table 2.2: Description of authentication keys used by the Twitter's API [28] - 24 -

Table 2.3: Describing the different levels of analysis [32] - 31 -

Table 2.4: Different types of sentiment analysis [43] - 33 -

Table 2.4: Summarization of different machine learning techniques [45] - 35 -

Chapter 3

Table 3.1: Python classes of AthPPA and their additional description - 40 -

Table 3.2: Sentiment values used for the labelling process by the sentiment analyzer used in AthPPA - 41 -

Table 3.3: Sentiment values used for the labelling process by the sentiment analyzer used in AthPPA - 45 -

Table 3.4: Identified negative hashtags per political party - 46 -

List of figures

Chapter 1

<i>Figure 1.1: (a) The Sumerian abacus system, (b) the Scytale (c) and the Antikythera mechanism</i>	- 1 -
<i>Figure 1.2: Emerged computational systems from Bronze age until the end of 19th century [1]</i>	- 1 -
<i>Figure 1.3: Emerged computing systems at the beginning of World War I and at the end of World War II [1]</i>	- 2 -
<i>Figure 1.4: Emerged computing systems from 1947 until 1973 (Post–World War II era) [1]</i>	- 2 -
<i>Figure 1.5: Emerged computing systems from 1974 until 1991 (End of the cold war) [1]</i>	- 2 -
<i>Figure 1.6: (a) the bell 101 modem, (b) FORTRAN documentation, (c) the IBM System/360, (d) C programming language logo, (e) first personal computer, (f) first laptop, (g) IBM 5100 one of the first IBM PCs</i>	- 3 -
<i>Figure 1.7: Technological advancements in computer science and engineering from 1992 until present [1]</i>	- 4 -
<i>Figure 1.8: (a) Logo of Boston Dynamics, (b) Apple newton, (c) first DVD player, (d) Universal Serial Bus, (e) Diamond Rio MP3 player (f) Napster logo, (g)Wikipedia logo, (h) Trek ThumbDrive one of the first USBs, (i) Google logo in 1998 [1]</i>	- 4 -
<i>Figure 1.9: Social media platforms and their additional logos organized by the date of their foundation</i>	- 4 -
<i>Figure 1.10: DARPA's former headquarters in the Virginia Square of Arlington, Virginia, USA and their official logo (source: Wikipedia)</i>	- 5 -
<i>Figure 1.11: An Interface Message Processor – IMP and its control panel</i>	- 6 -
<i>Figure 1.12: The expansion of ARPANET across the United States in 1970</i>	- 7 -
<i>Figure 1.13: Beginning from left Peter T. Kirstein, Vint Cerf and Bob Kahn pioneers of the TCP/IP</i>	- 8 -
<i>Figure 1.14: Hayes Smartmodem 300 baud modem designed in 1981 (source: wikipedia)</i>	- 9 -
<i>Figure 1.15: Logo of the MILNET domain</i>	- 9 -
<i>Figure 1.16: At left is the headquarters of NSF in Alexandria, Virginia and at right is the additional seal of the organization</i>	- 10 -
<i>Figure 1.17: Share of the population using the Internet from 1990 until 2017 per continental regions (source: World Bank and ourworldindata.org)</i>	- 11 -
<i>Figure 1.18: Share of the population using the Internet from 1990 until 2017 for the top 7 world superpowers (source: World Bank and ourworldindata.org)</i>	- 11 -
<i>Figure 1.19: At left is Darcy DiNucci, at right is a part from her article <i>Fragmented Future</i> published on Print Magazine in January 1999 [18]</i>	- 12 -
<i>Figure 1.20: Tim Berners Lee</i>	- 13 -
<i>Figure 1.21: Web 1.0 and Web 2.0 architectures</i>	- 14 -
<i>Figure 1.22: The “Vision” of Tim Berners-Lee for the Semantic Web [22]</i>	- 16 -
<i>Figure 1.23: First version of Twitter platform</i>	- 17 -
<i>Figure 1.24: First version of Blogger platform</i>	- 17 -
<i>Figure 1.25: First version of Facebook platform</i>	- 17 -
<i>Figure 1.26: First version of Youtube platform</i>	- 17 -
<i>Figure 1.27: Number of users using social media platforms from 2004 until present (source: Statista and The Next Web and ourworldindata.org)</i>	- 18 -
<i>Figure 1.28: Use of social media platforms by age group in the US in 2019 (Source: Pew Research Center 2019)</i>	- 20 -
<i>Figure 1.29: The three process steps of social media analysis</i>	- 21 -

Figure 1.30: Social Media Analytics procedure.....	21 -
--	------

Chapter 2

Figure 2.1: An abstract concept of how Twitter’s API data exchange procedure occurs [28]	24 -
Figure 2.2: Workflow of OAuth authentication model [30]	25 -
Figure 2.3: Google trends data related to the keyword sentiment analysis.....	28 -
Figure 2.4: Tasks of sentiment analysis as presented by Pozzi et al. [32].....	29 -
Figure 2.5: Distinction between subjectivity and polarity classifications as presented by Pozzi et al. [32]	30 -
Figure 2.6: Different levels of analysis as presented by Pozzi et al. [32]	31 -
Figure 2.7: Comparison between different machine learning algorithms [48].....	36 -
Figure 2.8: Machine learning algorithms and their respective classifiers [48]	37 -

Chapter 3

Figure 3.1: File Structure of AthPPA tool.....	39 -
Figure 3.2: How SpaCy module produces a Natural Language Processing linguistic object	41 -
Figure 3.3: Architecture of spaCy module	42 -
Figure 3.4: Results of the last Greek legislative election, showing the vote strength of the party winning a plurality in each electoral district.	45 -
Figure 3.5: Users likes and retweets per posted tweet mined from @kmitsotakis account (200 tweet sample)	46 -
Figure 3.6: Users likes and retweets per posted tweet mined from @PrimeministerGR account (200 tweet sample)	46 -
Figure 3.7: Users likes and retweets per posted tweet mined from @atsipras account (200 tweet sample)	47 -
Figure 3.8: Users likes and retweets per posted tweet mined from @FofiGennimata account (200 tweet sample)	47 -
Figure 3.9: Text length per posted tweet mined from @kmitsotakis account (200 tweet sample)	47 -
Figure 3.10: Text length per posted tweet mined from @atsipras account (200 tweet sample)	47 -
Figure 3.11: Text length per posted tweet mined from @FofiGennimata account (200 tweet sample).....	48 -
Figure 3.12: Users likes and retweets per posted tweet mined from @neademokratia account (200 tweet sample)	48 -
Figure 3.13: Users likes and retweets per posted tweet mined from @syrizagr account (200 tweet sample)	48 -
Figure 3.14: Users likes and retweets per posted tweet mined from @syrizagr account (200 tweet sample)	48 -
Figure 3.15: Mined tweets which include negative hashtag (#ΝΔ_Θελατε) for New Democracy party.....	49 -
Figure 3.16: Mined tweets which include negative hashtag (#ΝΔ_ξεφτιλες) for New Democracy party	49 -
Figure 3.17: Mined tweets which include negative hashtag (#ΝΔ_ρομπες) for New Democracy party	49 -
Figure 3.18: Mined tweets which include negative hashtag (#ΣΥΡΙΖΑ_ξεφτιλες) for SYRIZA party	49 -
Figure 3.19: Mined tweets which include negative hashtag (#συριζωα) for SYRIZA party.....	49 -
Figure 3.20: Mined tweets which include negative hashtag (#Συριζα_απατεωνες) for SYRIZA party	50 -
Figure 3.21: Number of registered subscribers per twitter account for the top three Greek political leaders (actual numbers).....	50 -
Figure 3.22: Number of registered subscribers per twitter account for the top three Greek political leaders (in percentage).....	50 -
Figure 3.23: Comparison of users likes per posted tweet for the top three Greek political leaders (600 tweet sample).....	51 -
Figure 3.24: Comparison of re-tweets per posted tweet for the top three Greek political leaders (600 tweet sample).....	51 -
Figure 3.25: Total comparison of Sentiment tweet for the top three Greek political leaders (600 tweet sample).....	51 -

Figure 3.26: Sentiment analysis per political leader (600 tweet sample).....- 52 -
Figure 3.27: Comparison of identified positive tweets (600 tweet sample).....- 53 -
Figure 3.28: Comparison of identified neutral tweets (600 tweet sample).....- 53 -
Figure 3.29: Comparison of identified negative tweets (600 tweet sample).....- 53 -
Figure 3.30: Comparison positive, negative and neutral tweets per political leader (600 tweet sample)- 54 -

Acknowledgments

With the completion of this dissertation I would like to thank Dr. Haridimos Kondilakis and Dr. Stelios Sfakianakis senior researchers at Computational BioMedicine Laboratory (C.BM.L.) of Foundation for Research & Technology – Hellas (FO.R.T.H.) for their guidance as well as to the entire team of the laboratory for giving me the chance to work with them and be a part of their team.

Dedication

I dedicate this work to my thesis supervisor Dr. Nikolaos Papadakis associate professor at the Hellenic Mediterranean University, for his constant guidance throughout my studies as well as to my family for their constant affection and support.

This Page Intentionally Left Blank

Chapter 1: Introduction to Social Media analytics

1.1. An overview on the evolution of computing systems

This sub-section presents a brief overview on the evolution of computing systems, their overall progress through the centuries and how modern computing systems emerged. To begin with this topic, we have to analyze how the computing systems operate, from where do they began, their current presence nowadays as well as their future potentials and implementations. Since the beginning of mankind, the curiosity to explore the unknown or to discover techniques in order to improve our life conditions was an utmost importance. The discovery of the wheel as well as the discovery of fire where some examples of how early mankind strived for progress. Although anthropological evolution is not the subject of this dissertation it is worthy to point out the need of mankind for progress through discovering new technologies that ultimately lead to better life conditions. Although the history of modern computers begins in mid 40s, there is an older recorded historical background of computing machines. The first recorded computing machine begins with the discovery of Abacus by Sumerians (2500 BCE) [1], the Scytale (700 BCE) [1] and Antikythera mechanism (150 BCE) [1][2][3] as well as Heron's Programmable Robot (10 AD-85 AD), discovered by the Ancient Greeks [4].

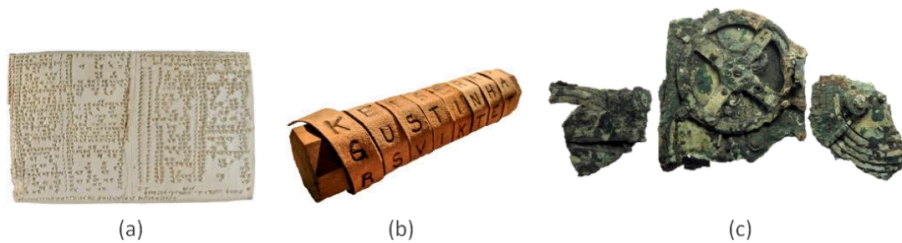


Figure 1.1: (a) The Sumerian abacus system, (b) the Scytale (c) and the Antikythera mechanism

Almost eight centuries later we have the deciphering cryptographic messages that was achieved by Abu Yusuf Ya'qub ibn Ishaq al-Sabbah Al-Kindi (801- 873) also known as the "Philosopher of the Arabs" who developed a systematic approach of breaking all existing encryption methods used in encrypted messages [5]. In 1470 the Cipher disk was merged by Leon Battista Alberti (1404–1472), where based on that a number of variations will be emerged over the years [6]. One example is Blaise de Vigenère, who took Alberti's concentric circles and turned them into a table that illustrated all of the possible substitutions for each letter. Later on, at 1613 the first Recorded Use of the Word Computer [7] emerges as well as Slide Rule 1621 [8] and the binary arithmetic in 1703 [1]. Figures 2 to 6 depict an overall timeline of the emerged computational system inventions during the course of centuries.

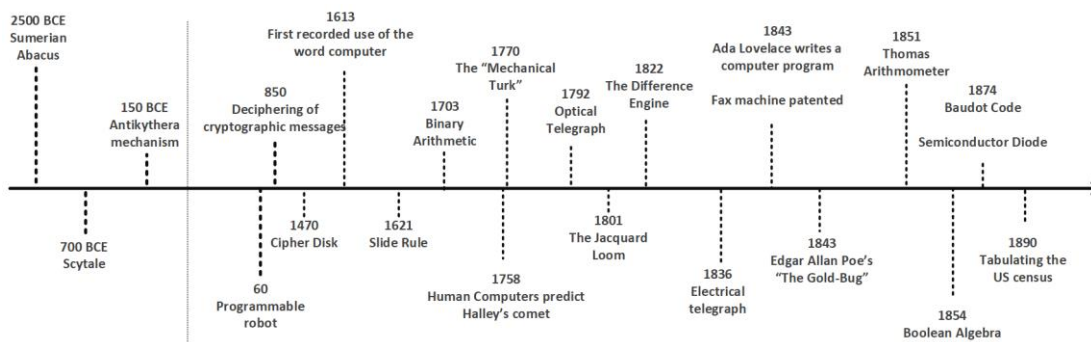


Figure 1.2: Emerged computational systems from Bronze age until the end of 19th century [1]

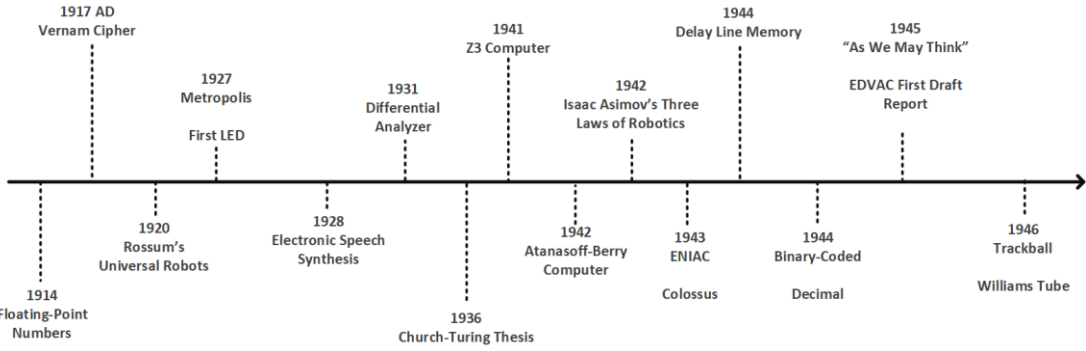


Figure 1.3: Emerged computing systems at the beginning of World War I and at the end of World War II [1]

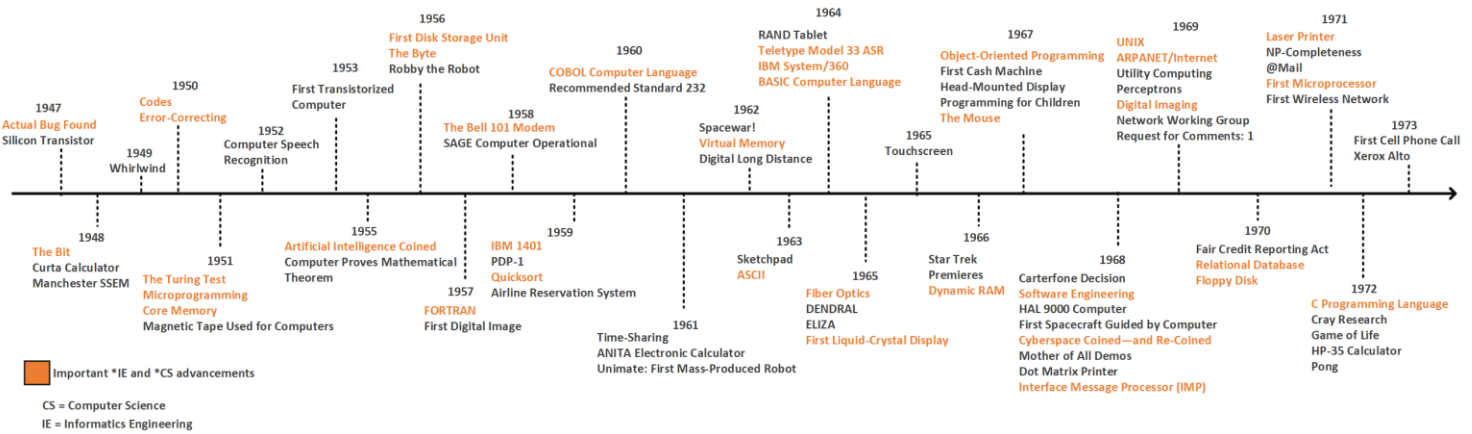


Figure 1.4: Emerged computing systems from 1947 until 1973 (Post-World War II era) [1]

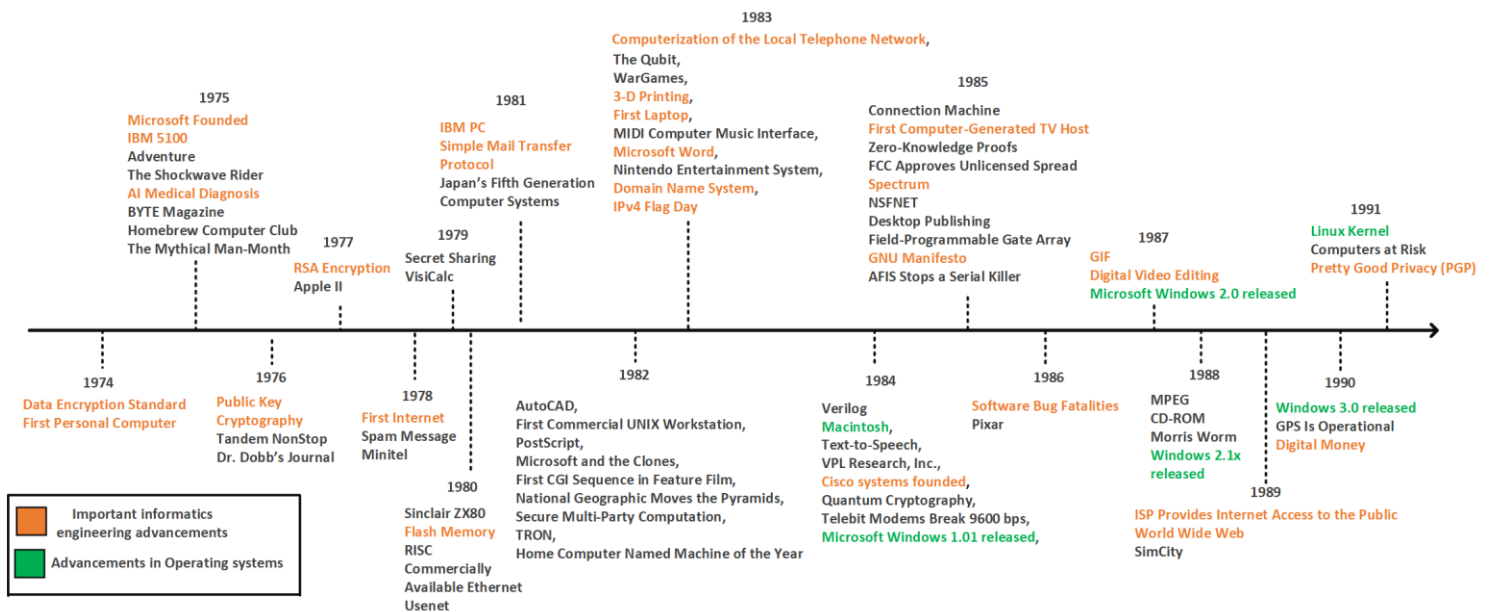


Figure 1.5: Emerged computing systems from 1974 until 1991 (End of the cold war) [1]

Garfinkel et al. [1] have recorded in their book the entire progress of computational and computing systems as well as the basic operation of each device. As we have observed from figures 3 until 6, the technological progress that has been made on the field of computing systems during the period of 1945 until 1991 (the beginning and at the end of cold war era) is outstanding with the most important ones being the discovery of bits in 1948, the Turing test and core memory in 1951, the coined of Artificial Intelligence (AI) in 1955, the byte in 1956, the FORTRAN programming language in 1957 the bell 101 modem in 1958, the COBOL computer language in 1960, the ASCII standard in 1963, the IBM System/360 and BASIC programming language in 1964, the fiber optics in 1965, the dynamic RAM in 1966, the UNIX systems and the ARPANET in 1969, the C programming language in 1972, the Data encryption standard and the first personal computer in 1974, the Public Key Cryptography in 1976, the RSA Encryption and Apple II in 1977, the first internet in 1978, the IBM PC and SMTP protocol in 1981, the Domain Name System, the Computerization of the Local Telephone Network, as well as the IPv4 Flag Day 1983, the first laptop, the Microsoft Word and 3D printing in 1983, the Text-to-Speech and Macintosh in 1984, the GNU Manifesto in 1985, the Software Bug Fatalities in 1986, the GIF and Digital Video Editing in 1987, the MPEG standard and CD-ROM in 1988, the Internet Access to the Public as provided by the ISP and World Wide Web in 1988, the operational GPS and Digital Money in 1990 and lastly the Linux Kernel and Pretty Good Privacy (PGP) in 1991. Figure 1.6 depicts some of those inventions which shaped the fields of computer science and computer engineering.

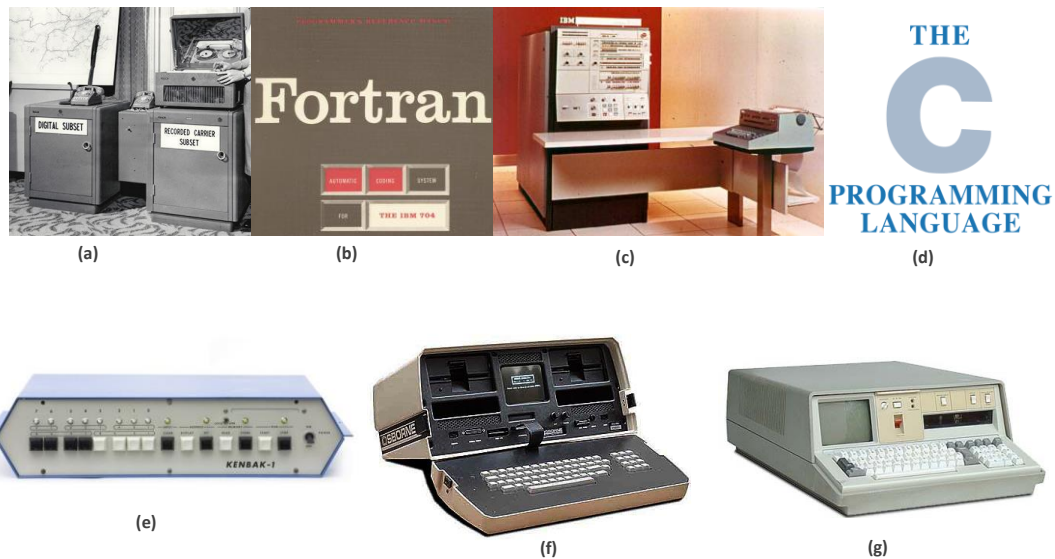


Figure 1.6: (a) the bell 101 modem, (b) FORTRAN documentation, (c) the IBM System/360, (d) C programming language logo, (e) first personal computer, (f) first laptop, (g) IBM 5100 one of the first IBM PCs

As we can assume from figures 4 and 5 the technological advancement on the field of computers was superb. The period of cold war was an important landmark for the emerge of the modern computers as well as an important step for the foundation of computer science and computer engineering fields. After the end of cold war there is also rapid advancement on both fields as the first operating systems emerged and the first social media platforms began their operation such as LinkedIn in 2002 and Facebook in 2004. After the creation of those social media platforms the emerge of YouTube occurs in 2005 and Twitter in 2006. Keep note on the last one as it is practically one of the most profound blogging platforms where in just a few years since its original release date it obtained millions of Internet users across the globe. A brief discussion is introduced in sub-sections 1.3, the emerge of social media platforms as well as in 1.6 introduction to Twitter analytics. Figure 1.6 presents the technological advancements on the field of Computer Science and Computer Engineering from 1991 until present.

1.2. The evolution of e-communication: How the Internet emerged

The beginning of 60s proved to be an important chronological landmark in terms of new electrical appliances and communication. During that period the Cold War was in full scale where Government officials of the United States of America as well as of the North Atlantic Treaty Organization – NATO, worried of a possible nuclear war conflict with the Soviet Union and with the countries of the Warsaw Pact. At that time, communications were vulnerable and vital information's were not protected [9]. At that time the communication networks were based on a switched-based architecture where it demanded rigid routing structures prone to single point of failure. In 1960s Paul Baran (or Pavel Barasov) a Polish migrant who was an M.Sc. graduate of Electrical Engineering, worked on the RAND organization as part of a research for the United States military. This research involved the structure of a survivable communication network in case of a nuclear event as well as its secure communication establishment. During his research he came to the realization that normal telephone network was very vulnerable [9]. From 1961 until 1964 Baran wrote a variety of research papers describing a mesh network with many nodes and with no central control where small messages would be sent by any available route. Although an important research implementation his work had little notice on that time. After Baran's work, Donald Watts Davies continued a similar approach on a project that he was in charge as a technical manager. Donald Watts Davies received physics and mathematics degrees from the Imperial College of London. That gave him the chance to work on the early computers of Alan Turing at the National Physical Laboratory where he was part of a team that completed the ACE computer. In 1963 he was a technical manager of the advanced computer techniques project where he was interested in the communication establishment across computing machines. In 1965 Davies realized that the messages were as he called them as “bursty”, consisted by short burst of information which were followed by relatively long periods when nothing happened while either the computer or the user digested it. This was an example of a dedicated connection.

This concept was not sending the entire message but instead it partitioning it into fixed-size small chunks which contained an index information regarding with its address and its position in the entire message. These small chunks were passed into any route from node to node of the entire network. The receiving machine would gather and reassemble all these chunks advising their index information in order to synthesize the entire message. Davies recorded his thoughts and in March of 1964 where he gave a presentation into interested parties [10]. From the attendants one government official from the U.S. Ministry of Defense drew his attention on Paul Baran's work [11]. Beyond the differences of both implementations a combination of them could form a communication network. The combination of both implementations made a new communication one by which differ from the simple telephone system, where here we have the segmentation of the information and its assemblance when all chunks reach to the sender. This relation means that this new communication network had the ability to handle more traffic information than a telephone network would have had. Chunks at each node had to be temporally stored before being sent on. The small size of the chunks meant that the amount of storage needed was small and known. Also, the delays to the messages through the system were minimized. Another advantage was that there was no central control and hence any method could be used to send on the chunks [2]. The Defense Advanced Research Projects Agency - DARPA was the initial step for the foundation of the Internet. When Soviet Union orbit the first Satellite in space (namely as Sputnik) in 1957, the United States founded DARPA in order to compete the space race as the latter had fallen behind.



Figure 1.10: DARPA's former headquarters in the Virginia Square of Arlington, Virginia, USA and their official logo (source: Wikipedia)

By 1966 the Advanced Research Project Agency had several other research centers throughout the U.S. and thus they decided to connect all the computing machines of all those branches through a network. For this purpose, ARPA recruited a person with a strong background in computer communications and this one was Larry Roberts. In October of 1967 he presented a paper on an ACM conference with title “*Multiple Computer Networks and Intercomputer Communication*” where it was an outline of what he was planning to design. At the same meeting, Roger Scantlebury, who was running the NPL plan to implement Donald Davies packet switching network, presented a paper of their detailed work. It was a revelation to Roberts, particularly as Scantlebury also referred to Baran’s work.

After the meeting, Wesley Clark a government official of the US army came up to him and said what he needed was to use a small computer alongside the main machines to handle all the communications. Roberts adopted this idea and called them Interface Message Processors - IMPs. The three involved groups in this packet-switching network idea (the ARPA, NPL and RAND) held a meeting in order to discuss this in detail. All the basic blocks were now in set in order to construct the ARPANET.

By 1968 Roberts issued a specification to contractors to quote to build the IMP units where it was sent to 140 companies. From them only 12 responded [9][11]. The eventual winner was a small company called Bolt, Beranek and Newman (nowadays known as BBN). By the August of 1969 the first IMP was ready and was installed at the University of California - UCLA at Los Angeles. The University of California was the perfect place to start this project as the Head of the University was Leonard Kleinrock a pioneer of computer networks who contributed a lot on the early theoretical work around computer networks. In September of the same year, the UCLA connected their host computer to the IMP and the first node of the ARPANET was ready. Soon the next IMP arrived at Stanford Research Institute - SRI and further ones at the University of California at Santa Barbara and at the University of Utah. They were rapidly connected to their host computers and to leased telephone lines and thus the ARPANET was born [9].

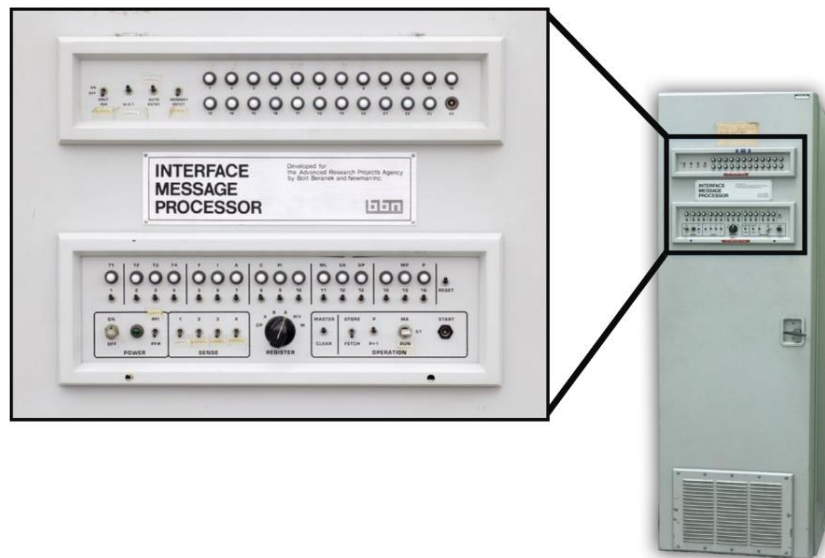


Figure 1.11: An Interface Message Processor – IMP and its control panel

The IMPs were consisted by commercial minicomputers and their functionalities were utilized by a software and a set of specific protocols. In every communication method there has to be a set of rules in order to achieve its success. For example, when there are many individuals and one of them is speaking, the rest of them have to listen him/her and wait for their turn, otherwise there will be a chaos. In ARPANET the idea is the whole network is completely depending on the protocols. On the ARPANET there is no need for exchanges or central controllers whereas this thing happens with the telephone systems. The several IMPs receive a chunk of data (namely as packets) and when they receive it, they read its instructions. These had to conform to a strict protocol or it would become confused [10].

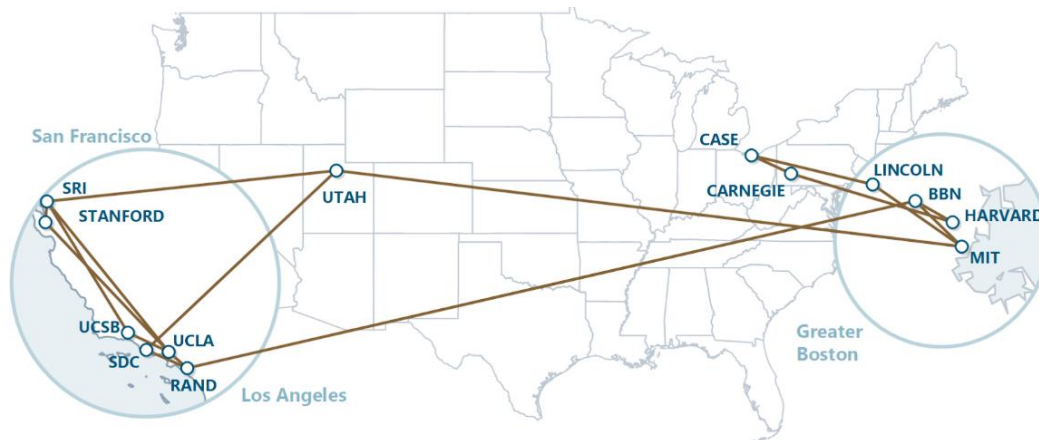


Figure 1.12: The expansion of ARPANET across the United States in 1970

In the following years, ARPA decided that the concept of the ARPANET had to be released into the public. In October of 1972 ARPA setup an exhibition at the International Conference on Computer Communications in a hotel located in Washington by inviting any relevant person that they could have think of [11]. In the center of the room, there was one minicomputer which was connected with 40 terminals. Visitors were able to sit in one of those terminals and run the demonstrated programs on computers connected to the ARPANET network located on the other side of the USA. The result was a complete success as the participants had to be ejected after midnight each day or they would have stayed the whole week. This exhibition set a strong boost on the ARPANET and in the next year 23 computers were connected.

After the release of the ARPANET several other inventions were created such as the emerge of the first email application by Ray Tomlinson who was a BBN employee. He also contributed to the development of TENEX operating systems and to the developed a file transfer program namely as CPYNET in order to transfer files over the ARPANET [9][13] and implementations of Telnet and on self-replicating programs such as the Creeper and Reaper. Tomlinson had a task from the BBN company to change a time-sharing computer message sender program namely as SNDMSG in order to run on TENEX operating system. He added code he took from CPYNET to SNDMSG so messages could be sent to users on other computers and thus forming the very first email [13].

Important note

ARPANET was an important technological milestone one by which government officials, scientists and companies worked together for a common purpose. It is considered as a success story in terms of project management due to the reason that brought high-profile researchers working on top institutions. It also encouraged the ethos of collaboration with the most profound example, the design of the central protocols that the ARPANET network used which was a task appointed to the Network Working Group which was a team of University students. The second crucial aspect of the ARPANET was that it provided an interesting case of study to the point by which technologies are socially shaped. Specifically, the social design aspect of the ARPANET was done by its users where they were also its designers as well. The early purpose of the ARPANET was to support the ability of resource sharing between computers, although this evolved to another greater purpose which was the communication between different computer users across the USA (in the places where the network was established as shown by the figure 1.12). With term communication we refer to the exchange of text between different ARPANET users and to the ability of file and software sharing among them. Users saw ARPANET as a

communication tool rather than a tool for communication across devices and thus it was rapidly embraced by the society. In addition to that, proceeding on the technical aspects of the ARPANET, the network itself has restrictions as well as flaws that users wanted to improve them and ARPA did not encouraged this as it was a network for military and defense purposes. These restrictions will lead to the design of new technologies such as the OSI and TCP/IP models.

Meanwhile the British had setup their own network as we mentioned above as NPL Network or NPL Data Communications Network where it was a local are computer network. This local network was operated by a team of the National Physical Laboratory in London who constructed the idea of packet switching. By 1971 their network was fully operational having 60 active connections in their NPL network. Larry Roberts who was one of the creators of the ARPANET thought that it would be useful to connect NPL network with the ARPANET [9]. This idea seemed to be also supported by the NPL team, although Donald Davies who was a scientist working on the NPL network and pioneer of the packet switching method was unable to convince the British government for funding this project. United Kingdom at that time was in a process of joining the European Economic Community and thus necessary economic measures had to be taken leaving no gap for the government to spend money [9][14]. However, they managed to find a way, ARPA had a satellite connection with the Norwegian Seismic Array - NORSTAR and a cable link was setup from it to the UK. Donald Davies managed to obtain a small funding from the UK government and thus ARPA agreed to loan an interface processor which was one of the terminal varieties known as a TIP [9]. Later on, Peter Thomas Kirstein a British computer scientist managed to setup a connection with a computer located in Rutherford High Energy Laboratory of the UK to the ARPANET network. This was unique achievement as it was the very first computer that it was connected to the ARPANET network which resided outside from the US. He was also known to be the pioneer of TCP/IP model alongside Vint Cerf and Bob Kahn which they were scientific members of the ARPANET network [9].

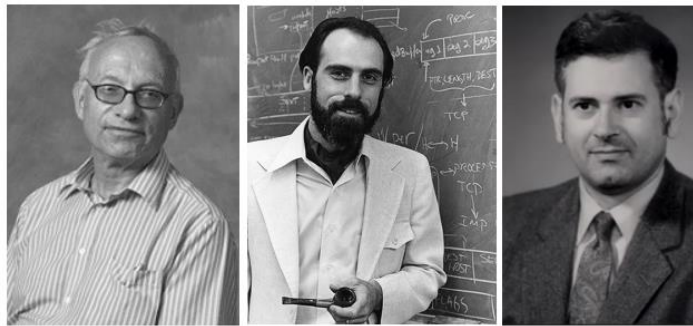


Figure 1.13: Beginning from left Peter T. Kirstein, Vint Cerf and Bob Kahn pioneers of the TCP/IP

In 1973 a project began namely as internetworking where the main task was to design a new computer interconnection system that would be able to incorporate a variety of different networks. The main issue of the ARPANET is that it was a unitary network and thus a transition into something new that would have the ability to incorporating different networks was necessary. One was to allow networks wishing to join the new ‘internetwork’ to retain their existing protocols and simply construct ‘gateway’ computers that would translate those into a common set of conventions [9]. The other was to require that all candidate networks adopted a new set of protocols, which would become the lingua franca of the new overarching network. In the end, the second option was adopted, and a suite of interlocking protocols centered on two new ones – TCP and IP – evolved. In this way TCP/IP became the cornerstone of the new “network of networks”. The great advantage of this approach was the possibility of organic growth, as long as a given network supported TCP/IP it was free to join the

Internet, furthermore the system was not owned or controlled by anybody (unlike the ARPANET), there were no gatekeepers to control admission to it [9][12].

The ARPANET continued to operate alongside TCP/IP, as the new network took shape. But since it was no longer an experiment project, ARPA started looking for a new owner to host it. AT&T, the federally owned telecommunications monopoly, was approached first, but the enterprise turned it down. In 1975, the government passed it to the network's organizational obligation to the Defense Communications Agency - DCA, which delivered communications facilities to the U.S. military. This had negative impact as the network itself started to lose its research interest as the military officials wanted a secure communication network rather than a network where every individual on the academia can contribute on it. In the midst of this, the Internet initiative was gaining interest and the network family TCP/IP was being finalized [9]. The main target was to extend adoption of the new protocols to the point where network effects came into operation. This turned out to be more difficult than anticipated: many nodes were reluctant to devote the necessary resources to configure their operations around the new protocols. It was at this point that DCA management of the ARPANET was to prove decisive. In March 1981, the Pentagon announced that all ARPANET hosts would be required to adopt TCP/IP by January 1983. It is also worth to mention that at the same year the first smart modem was introduced namely as the Hayes smart modem. This device could support a rate of 300bits per second and it was directly connected into a small microcontroller that inspected the data stream for certain character strings represented by additional commands. This feature of data and command exchange was sent through a serial port and these commands were known as Hayes command set [9]. The instructions that this set of commands used was for hanging up the phone, dialing numbers, and answering calls, among others. Furthermore, the smart modem could be connected to any computer with an RS-232 port, which every microcomputer had back in those days. Figure 1.12 depicts the first smart modem.



Figure 1.14: Hayes Smartmodem 300 baud modem designed in 1981 (source: wikipedia)

Resuming to the TCP/IP where not all sites were able to meet the deadline, but by the middle of 1983, every ARPANET host was running TCP/IP. A few months before that, DCA concern about the security of the network and that led to the decision of splitting it into civilian and military domains. From October 1982, one domain (the ARPANET) would continue as a research enterprise; the other one (the labelled MILNET) would henceforth be entirely devoted to military communications. The switchover was implemented in January 1st 1983 where civilian and military sides of the network were created [9][13].



Figure 1.15: Logo of the MILNET domain

With the creation of MILNET domain, the ARPANET regained its former status as a researched-based network used by universities and research institutions, which was the initial purpose of this technology, although the US military monitored the network as it was the main funder. The second important step was the deploy of TCP/IP technology into the computer industry. For this task, ARPA funded some operators in order to design TCP implementations for a variety of operating systems (e.g. Unix10). This procedure occurred in order to help

computer manufacturers to implement TCP/IP software in their machines with overall funding cost to 20 million dollars. By the 1990 the TCP/IP model was available in most of computers at least on the US market industry. Although there was a main issue of that of host addressing as users in order to visit another host, they had to remember its additional IP address which is a long number e.g. 132.120.78.19. for computers this is probably fine but humans it is a bit difficult to memorize it or remembering them especially if the host we want to visit are many [9]. For this reason, in 1983 Paul Mockapetris and Jonathan Bruce Postel of the University of California, Irvine wrote the first implementation of Domain Name System - DNS where it converted the IP addresses into domain names where people can easily memorize (e.g. google.com). Although DNS has one main disadvantage which is the need for central look-up in order to achieve a successful conversion of IP address into a domain name. As we've previously mentioned the access to the ARPANET was allowed to a certain personal of research institutions which had a contract with the ARPA organization. Although, after the ARPANET was released to the public, it seemed difficult for a simple user to learn it or to navigate/communicate through it. As a result, in the early 1980s the National Science Foundation - NSF was founded by the United States in order to fund the creation of a Computer Science Network [9][13].



Figure 1.16: At left is the headquarters of NSF in Alexandria, Virginia and at right is the additional seal of the organization

It's worth to mention that by the 1980s the Computer Science and Engineering fields were gradually accepted as disciplines by many of Universities [11]. Although the ARPANET was only used by researchers which had some kind of relation with the ARPA, the Computer Science Network was encouraging computer scientists of any institution to pay an annual feed in order to use the ARPANET (beyond that its commercial use was prohibited by the National Science Foundation Network). As a result, the ARPANET network expanded from 2000 hosts 1985 to 185.000 in October of 1989 and 1.776.000 in July of 1993. The National Science Foundation Network or NSFNET was a program funded by the United States from 1985 until 1995 in order to encourage the advancement of research and networking throughout its federal states. By the mid 1984 the NSFNET began to fund the establishment of supercomputing centers throughout the USA, where a community of a national network as well as researchers was required. Furthermore, by the mid-1980s the NSF, NASA and DOE started to design the first Wide Area Networks based on the TCP/IP standard with NASA to develop its NASA Science Network, the NSF to develop the CSNET and the DOE to develop the Energy Sciences Network or ESNNet. The CSNET designed by NSF in 1981 was connected with the ARPANET using TCP/IP over X.25, but it also supported non sophisticated network connections which used dial-up or mail exchange. At 1986 the NSFNET was created by the NSF with a 56 kbit/s. Furthermore, the NSF funded the establishment of supercomputing centers throughout the USA. The NSFNET Beyond its 56 kbit/s it became quickly overloaded as it provided support for regional research and education networks throughout the US as well as the connection of college campus with the regional networks. Thus, by the 1988 the NSFNET was upgraded to 1.5 Mbit/s and 45 Mbit/s in 1991 [9]. United Kingdom also build up its own network namely as JANET in order to connect its universities with the NSFNET network [12]. With the NSFNET to be a fast-growing network, the almost 20 years old then ARPANET was deprecated and thus it was replaced by the NSFNET. In 1991 it was clear that the NSFNET had dominated on the communication across computers and thus by the year 1993 it was allowed to be used as fully commercial product. Thus, the first commercial Internet Service Providers - ISPs were created. In order for a user to be connected to the network a monthly registration fee to an ISP company was

necessary. In 1995 just two years after the creation of the first ISPs, the NSF withdraw its support from the NSFNET as it was functional and thus the Internet start to form. In just 8 years the NSFNET managed to grow from 2000 users in 1985 to more than two million in 1993 [9][12][13].

Back in those days, the main drawback of e-mails was that it supported only text message transferring. This meant that pictures, drawings or any other non-text information was unable to be sent. Nathaniel Borenstein and Ned Freed managed to solve this issue as they provided a method where the attachment of images and documents was possible through an e-mail. This was the beginning of Multipurpose Internet Mail Extensions - MIME which allowed the ability of attaching documents and pictures on emails. The next problem which the newborn internet had was a way to view the webpages. Early version of web browsers such as the Mosaic browser developed by National Center for Supercomputing applications was introduced and within a few months from its release it gained a million of users. The same team behind the development of Mosaic browser decided to launch the release of a new improved browser which that was the Netscape Navigator. Microsoft followed this and they also launched the release of the Internet Explorer. At that time, it was difficult for users to know what website existed on the web. By the 1998 the Google was launched its release by providing a page where a user could use in order to search websites through the internet [9][12][13]. That's when the surface web, deep web and dark web took shape. Surface web is what it is considered as a legitimate website by google, any website which is accessible through the internet it is considered in this category. Websites that are not accessible through the Google are considered as websites which belong to the deep web or dark web.

Figure 1.17 and 1.18 presents the Internet growth rate per continental regions for Europe, Asia, Africa, North America and the pacific continents as well as for the to the top seven global superpowers. As counted population it is considered a person which has used (or continues using) the Internet over the last 3 months via a computer, a mobile phone, a personal digital assistant, game machines, digital TVs etc.

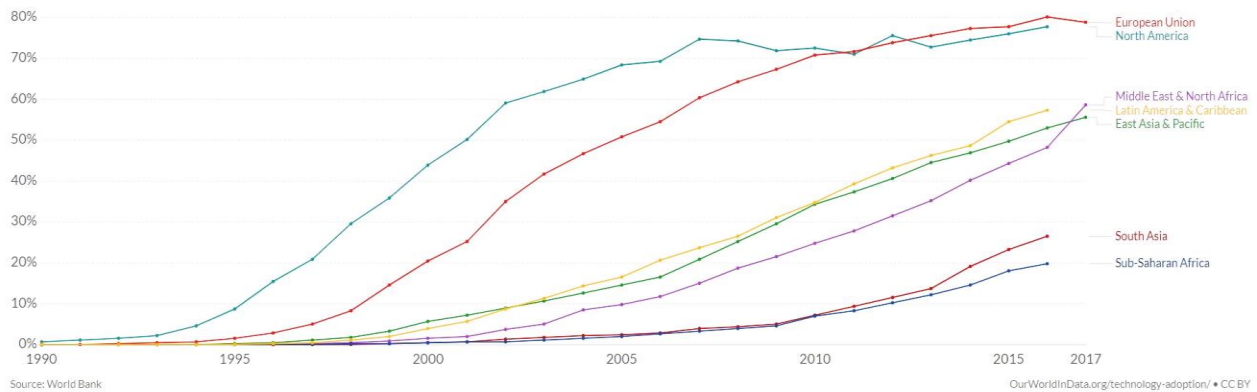


Figure 1.17: Share of the population using the Internet from 1990 until 2017 per continental regions (source: World Bank and ourworldindata.org)

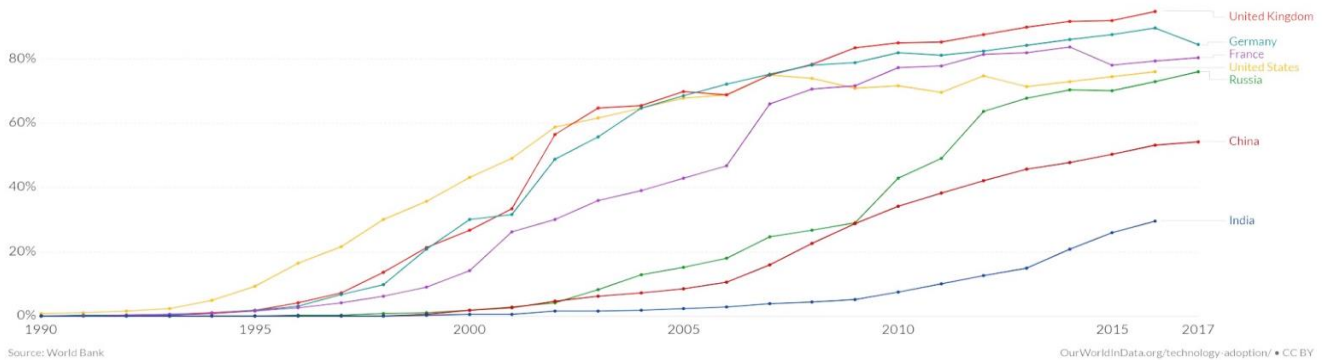


Figure 1.18: Share of the population using the Internet from 1990 until 2017 for the top 7 world superpowers (source: World Bank and ourworldindata.org)

1.3. The transition from static (Web 1.0) to interactive content (Web 2.0)

As the computer industry and Internet was evolving and thus graphical user interface - GUI become available to many applications. From 1985 (when the NSFNET became available) until the mid-1995 (with the emerge of the first Internet), more and more simple users tend to use this service as it provided a graphical interface by which a user could use in order to visit an address of an another pc, instead of a console where the user has to know its additional commands set and insert them through a terminal. Hjorth et al. [15] mention that the main factor for the growth of the Internet was the development of World Wide Web in 1992 which provided a graphical method instead of a textual one, for the navigation of the user to the service. This feature made the Internet accessible to a larger community of users who weren't familiar with console commands. Another crucial factor was of course the development of the Internet in 1993 and the emerge of the first ISP companies as it was analyzed in section 1.2.

The early web had non search engines such as Google, Yahoo or DuckDuckGo and that made the navigation through the system difficult for a user. Although, before Google there were some forms of early search engines such as Lycos in 1994 and Altavista in 1995 which provide basic search throughout web [16]. Hjorth et al. [15] suggest that the first search engine with the most extensive index was Yahoo which had an immense collection of categorized websites, but as the web grew the management of their index became very difficult. User-created content was crucial since the creation of the Internet. Some examples of user-created content can be personal home pages which they were hosted on their own computers or on free hosting services like GeoCities. On the early days of the Internet, websites were created with HTML which was released in 1993 and thus technical experience was essential. As the number of Internet users was increasing so did the commercial companies increase their presence on the service. Yahoo was gradually starting to deteriorate as the number of web pages was increasing on the web and that gave space to other search engines with more optimized index list to step in, were in 1998 Google emerged and gradually took the leading role as a web search engine. The American dot-com bubble, a historic period of excessive speculation which occurred from 1994 until 2000 was a period of immense growth of the Internet usage. The Nasdaq Composite stock market index, which many internet companies where included on it, reached its peak on 10th of March 2000 before it crashes. This bubble was known as the dot-com crash which was lasted from 11th March 2000 until 9th of October 2002 [15]. Dot-com crash affected many companies specifically the online shopping ones such as pets.com or boo.com as well as communication companies such as worldcom, NorthPoint Communications and Global Crossing which eventually shut down. It also effected large companies such as Cisco whose stock declined by 86% as well as Qualcomm. Following the events of dot-com crash many internet companies and services began emerging which followed a set of common principles with the top ones giving its focus in user-generated content as well as the ability of websites to exchange data between each other and with other applications. Darcy DiNucci an expert web designer wrote an article in January 1999 namely as "Fragmented Future" which specifically stated [17][18].



"The relationship of Web 1.0 to the Web of tomorrow is roughly the equivalent of Pong to the Matrix. Today's Web is essentially a prototype - a proof of concept. This concept of interactive content universally accessible through a standard interface has proved so successful that a new industry is set on transforming, capitalizing on its powerful possibilities. The web as we know it now, which loads into a browser window in essentially static screenfuls, is only an embryo of the Web to come. The first glimmerings of Web 2.0 are beginning to appear, and we are just starting to see how that embryo might develop... The web will be understood not as screenfuls of text and graphics but as a transport mechanism, the ether though which interactivity happens."

*Darcy DiNucci ,
Design and new media: Fragmented future,
Print Magazine, 1999*

Figure 1.19: At left is Darcy DiNucci, at right is a part from her article *Fragmented Future* published on *Print Magazine* in January 1999 [18]

This statement of DiNucci marked the beginning of the of the Web 2.0 where a few years later it re-appeared as a term in the first Web 2.0 conference in 2004 organized by Tim O'Reilly. By 2003 user-content sites start to emerge such as WordPress (blogging platform) and MySpace (Social Networking) in 2003 as well as Flickr (photo sharing) in 2004 [16]. Websites using Web 2.0 provided storage for the hosted generated user content, allowed users to create their accounts on the systems and encouraged the sharing of information between them. Web 1.0 services provided online data storage but few things on the technical support. With Web 2.0 the uploading of information was simplified as well as additional features like tagging in order to make content easier to discover. In more simple terms Web 2.0 applications provide users the ability of interaction and collaboration in order to create virtual communities. For example, Flickr provided a relatively easy way for users to upload their photographs and thus make them available throughout the Internet. The entire process of content uploading was simplified since users could use an interactive web-based interface in order to upload it [16].

Important note

With the emerge of World Wide Web by Tim Berners Lee in 1991 the Internet obtained a hyper-text feature where it made possible the establishment of online communities on the web. Weblogs, list-servers and e-mail services contributed to the emerge of the first online communities. Until the late 1999, online communities were exclusively consisted by generic services by which a user could join or to create a social group but the Internet service itself did not provided the automatic connection to others. With the Web 2.0 which emerged at the early 2000, online services started to evolve into something more interactive

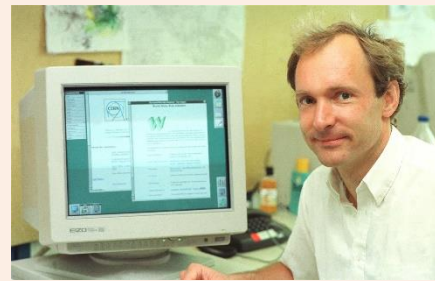


Figure 1.20: Tim Berners Lee

and not just providing channels for network communication. The impact that the Web 2.0 had was so immense that it was considered as an initial global Internet infrastructure just like electricity or irrigation systems in our daily lives [19]. As we've observed in section 1.1 the technological advancement on the field of communications over the past two centuries shaped our social daily practices. Some examples of these technological advancements are the evolution of the telephone communication which has been developed a lot or even the emerge chatting platforms and the ability of sending short messages from one computer onto another, using social media platforms or chatting applications over the Internet [19][20].

In just few years, Web 2.0 was established as a functional infrastructure over the Internet and thus users started to use more often the online communication environments for their daily social activities. As we've mentioned previously, before Web 2.0 was implemented, the operation of websites can be described as conduits for social activity. With the emerge of the new Web 2.0 platforms, websites started to be transformed into applied services, making the use of Internet easier for its users [19].

Web 2.0 applications provide users with the ability of interaction and collaboration in a variety of ways and thus forming virtual communities. Wikipedia is an example of interactive collaboration as users have the ability to edit the content of a particular information and create new section blogs. The word Blog derives from the words web log and it can be described as a place where users add information about a subject or a problem (e.g. Labor Day in Wikipedia which is a section referring to what this event is) or thoughts for discussion (e.g. chat forums). So, Blogs are interactive in the way that users who read a topic that is written in a blog section can leave comments, upvote or downvote a comment and reply to additional comments posted by other users. This an example of a virtual community where users can interact each other and discuss the latest trends etc. A user can post a comment without technical knowledge and contribute to the content of a blog or of a webpage. This

successful attempt of interactive communication led to 150 billion public blogs by the year 2011 [17]. Figure 1.19 depicts the overall architectures of web 1.0 and 2.0.

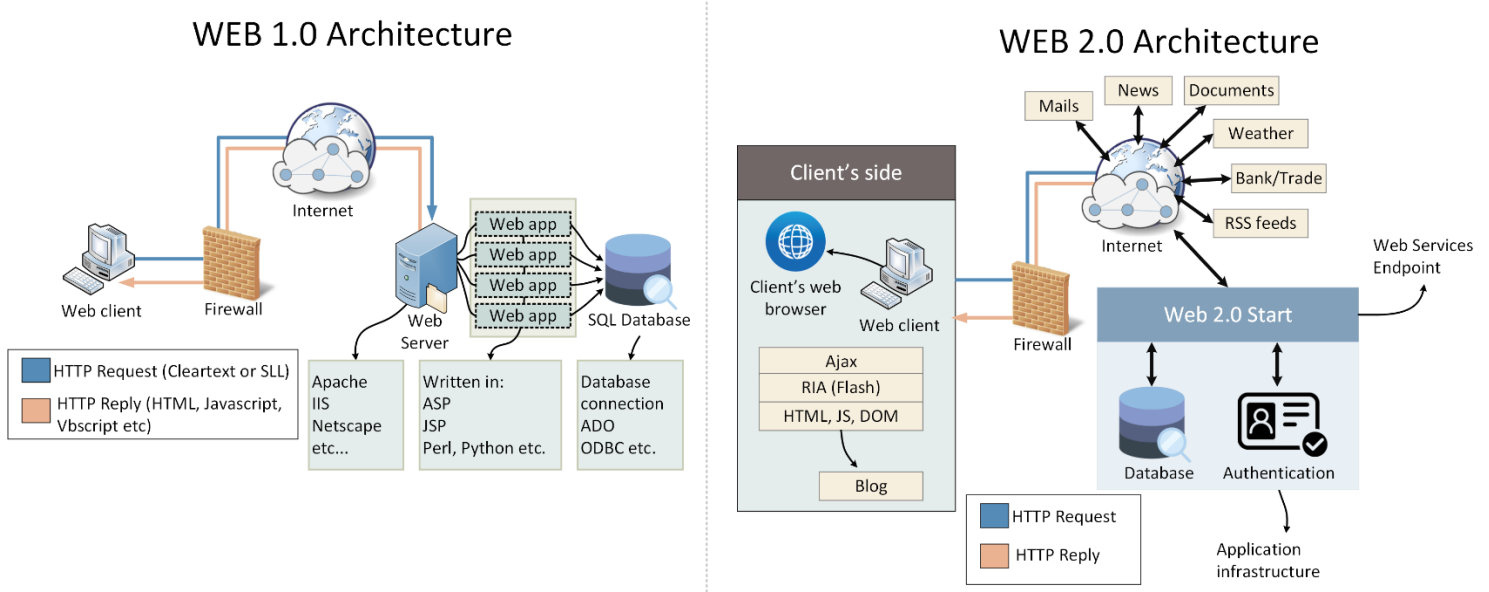


Figure 1.21: Web 1.0 and Web 2.0 architectures

With the term wiki website, we refer to a website with the ability of user interaction meaning that users are able to interact and do socializing activities with other users by adding, removing or editing the content of the webpage. Wikipedia is one example of a wiki website as it is an online encyclopedia which most of its sections have been created by Internet users. The policies of Wikipedia are based on 5 aspects, where the first is that it is an encyclopedia, the second that each information added to the site has a neutral point of view, the third is its content is free, fourth is its users respect other users and its community and lastly the absence of strict rules. Wikipedia is a great example of a Web 2.0 website where since its creation, users have an increasing number of collaborative creations until nowadays [17]. Another important feature of Web 2.0 is the ability for users to create their own tags in order to bookmark their information such as photos. Many social media platforms provide photo tagging services such as Facebook, Flickr, Instagram and so on, where this feature contributes for the creation of online communities. Lastly the Web 2.0 provides the ability of mash-ups creations which means the ability of websites or applications of allowing its users to combine data from multiple websites (e.g. a webpage for product searching such as ebay.com) [17]. In the next section 1.4 we will analyze the social networking and media platforms and present additional statistics. The list below presents the features of Web 2.0 as well as its capabilities.

1. *Software Scalability*: The ability of gathering behavioral data through cookies and clickstream analytics. Clickstream analytics is path or order pages were users select when navigating through a webpage. It is notable to mention that when clickstreams are applied to the Web, there are categories off websites where users visit on a regular basis and it is used for retargeting purposes. Furthermore, users can be co-developers as the software begins harvesting the collective intelligence of developers and users [21].
2. *Leveraging the Long Tail*: This phrase was fist coined by Chris Anderson in 2004 and later became popular as one of O'Reilly's Web 2.0 patterns where he described it as "*the collective power of the small sites that make up the bulk of the webs content*". Originally, this term didn't have the web as its main aim cause but in recent years it has been used in order to describe the strategies that the internet companies use in order to get advantage of the online market. A traditional retail store cannot have a large quantity of products while virtual online shopping can have thousands of products plus the required labor cost for the distribution of a product is also a major issue. Furthermore, online shopping

can provide recommendations based on the user's needs or preferences. Some examples are that of the Amazon, eBay and iTunes [21].

3. *New business models*: eServices which are based on an economic collaborative business model such as Airbnb and Uber challenging the survival of the traditional business models such as hotels and regional taxi services [21].
4. *Device agnostic software development*: Programmers have the ability to utilize mobile software development toolkits such as the Android SDK which eases the development of programs throughout mobile devices [21].
5. *Web Analytics*: As long as a platform measures Web 1.0 and 2.0 activities so does Web Analytics mature [21].
6. *Search engines*: These engines select specific keywords (provided by the user) and display webpages regarding to the given string. Such examples are Google, Yahoo etc [21].
7. *Semantic Web*: An extension of the Internet which makes the Web intelligent and intuitive by separating the presentation content from its meaning. The semantic web simplifies the way by which a user can customize/personalize its communications [21].
8. *Social media* and networking platforms: Web communication platforms where users can communicate with each other in a textual form or upload multimedia content such as video or audio and share it with other users or through the Internet. Additional examples are Facebook, YouTube, Twitter, Instagram etc [21].
9. *Geo-location*: Any mobile device broadcasts a signal to the Internet using the Global Positioning System (GPS), Near Field Communication (NFC) and georeferencing. With GPS the user's geographical navigation is fairly simplified as it eliminates the need for physical map advising which is a very time consuming and inaccurate procedure [21].

1.4. The emerge of the Semantic Web (Web 3.0)

The Web 3.0 is the next generation of the World Wide Web where it marks the beginning era of a connected Web operating system with most software elements (e.g. application programs, operating systems) and data processing reside on the Internet. Web 3.0 is more efficient in terms of speed, intelligence as well as very reliable with connecting data concepts, applications and users. The scope of the Semantic Web (also known as Web of data) is to create a relationship of a great portion of data with datasets so they can be accessible through the Web in a machine-readable format (such as Resource Description Framework - RDF) in order for applications to query it. RDF can be described as a general-purpose language for information representation with an "on demand approach" [21]. Due to this change the World Wide Web faced the following changes:

1. *Democratization*: Every individual can create a content or either a consumer. Most corporations are constantly trying to transform in order to align with the new capabilities developed through Web 1.0 and Web 2.0 with the automation and extensibility of Web 3.0. One example can be bloggers who can create their blog, post public content on them as well as collaborate with other content creators in order to make their posts viral [21].
2. *Smart applications*: These applications customize/personalize the web experience of user with insights based on geo-location services, Big Data and predictive analytics. The invention of smart applications had as a result the creation of smart devices which set the foundations for the Internet of Things [10].
3. *Collaborative Economy*: Leverage of the Internet and of Social Networks as business models in order to satisfy needs such as Uber and Airbnb [21].
4. *Smart Fabric*: The location of a shared or processed information whether if it is in cloud computing, iBeacons, Hadoop, data lakes or intelligent devices [21].

The rapid advancement of the internet gave space to new platforms such as search engines, websites, social media, portals and so on. Combining data together from all these different platforms makes a massive source of

valuable information. Nowadays internet users have the ability of searching information through search engines and find the additional results. Wouldn't be more exciting if we could ask an electronic device about something and to be able to process it and answer it back to us?

That's what semantic web trying to achieve by filling by providing a knowledge database to applications in conjunction with Natural Language Processing or machine learning techniques and thus be capable of answering a question. We will be focusing on Natural Language Processing on the next Chapter, for now let's understand the basic idea behind Semantic Web. As a term appeared during 1999 by the Tim Berners-Lee creator of World Wide Web where he specifically stated.



I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize.

*Tim Berners-Lee,
Weaving the Web. HarperSanFrancisco,
March 1999*

Figure 1.22: The "Vision" of Tim Berners-Lee for the Semantic Web [22]

We can describe the Semantic Web as a complex concept based on a simple idea of connecting entities (URLs, pages and content) on the web using relations by the underlying implementation is difficult at scale due to the sheer volume of entities presented on the internet. To create those links between pages and content based on relations markup languages have to be utilized such as Resource Description Framework - RDF and Web Ontology Language - OWL. RDF and OWL languages give the ability to content creators of adding meaning to their documents where machines could process for reasoning or inference purposes and thus allow automating many tasks on the web.

1.5. Analyzing social networking and social media platforms

As we've mentioned from section 1.3 social media platforms appeared when the first Web 2.0 platforms started to emerge due to the reason that a web 2.0 application promoted interoperability, sharing and multiple-way communication which are fundamental aspects in social media. The very first social media platform is Blogger developed by Pyra Labs in 1999 and bought by Google in 2003 where it provides a blog-publishing service to multiple user blogs with a time stamp record. Many other social media platforms existed before the emerge of Facebook in 2004 with the form of MySpace and other platforms [23]. Although, when Facebook was launched in 2004 the use of social media platforms was skyrocketed. Since then more than every five people around the globe have started to use some sort of a social media platform on a regular basis. YouTube for example receives more than a half million visits every month and half million users join Twitter on a daily basis. To understand better the impact of social media more than 10% of the world uses Facebook alone, not to mention for other social media platforms [10]. The popularity of social media can be traced from ages from 18 until 30 years old and this is due to the adaptability of younger ages with the technology and computers itself [21]. Figures 1.21 until 1.24 depicts the first versions of Blogger, LinkedIn, Facebook, YouTube and Twitter platforms and Figure 1.24 depicts the number of users per social media platform.

What your friends are doing. (over the last 24 hours)

- 
Jack Just received a txt message from a person in Brazil wanting to twtr. (7 minutes ago) x
- 
Dom It would be impossible to surf Linda Mar with the short board, but it won't stop teh Stewie! (13 minutes ago) x
- 
Crystal my krissy behind it's fine all of the time. (23 minutes ago) x
- 
donnie I'm hopin it comes true! (31 minutes ago) x
- 
Garett Gulf shrimp w artichoke broth and hearts of palm mmmh :) (about 1 hour ago) x
- 
biz Just had a good workout! (about 1 hour ago) x
- 
Crystal eyes feeling dizzy. hoping my intro back into super-aerobics is ok after my poison oak forced break (about 1 hour ago) x
- 
lisa saw two crying women on two different subways this evening. (about 1 hour ago) x

What are you doing?

0/1

twtr from your phone!

twtr is easy to use from your phone. Just txt your updates to the number **40404**.

TXT NOTIFICATIONS: **OFF**

291 txt messages this month through twtr on your **Cingular** phone. **Need to increase your txt plan?**

Your friends. **Add more!**

↓ WAITING FOR THESE PEOPLE TO ACCEPT

- 
cancel
- 
cancel

Star someone to get a txt message every time they update. Those in **bold** (40) have starred you.

Figure 1.23: First version of Twitter platform

"I'm very impressed. I've been planning to set up my own server using Frontier, but I think Blogger might do everything I want."
- Michael Zajac

WHAT'S THIS?

pyrAlert! is our (mostly) industry-focused blog, featuring a sporadic mix of news, links, and commentary about various technologies, companies, products and ideas we're interested in.

pyrAlert! is actually a subset of an internal blog we have been using since the first days of Pyra. One day, we realized a lot of the people who are interested in what we're doing would be interested in a lot of the stuff we posted to that blog. So we decided to make it part of our site. (And to do it easily, since we don't host this site in-house, we wrote the code that eventually became Blogger.)

pyrAlert! is also available as a My Netscape channel, a My.Userland channel, and in raw RSS and ScriptingNews formats (in case you'd like to syndicate us).

PYRALEAT!

<< archive index | current >>

Friday, August 27, 1999
SETools, from SiteExperts, is a bookmarklet-implemented web designer utility for IES that gives you a quick control for: resizing the current window to standard resolutions, switching any page into full-screen mode, testing different colors schemes, and hiding images. Pretty nice.
-ev, 3:59:03 PM

Thursday, August 26, 1999
LinkCount.com tells you how many links there are to your site (according to AltaVista and InfoSeek) and lets you graphically compare that number with your "competitors" (or whoever).
-ev, 9:47:56 PM

Is your weblog automated? [Answer here.](#)
-ev, 9:07:29 PM

[Searchbutton.com](#) allows you to add full-text searching to your site without installing or administering any software. Here's the interesting part: It can look like your visitors haven't left your site at all.
-ev, 8:09:28 PM

There's a very interesting thread on Outliners.com about various outlining/flowcharting tools and their merits as web site architecture design tools.
-ev, 6:08:23 PM

Interesting thought regarding hosted, web-based applications: If a user doesn't like a new version, unlike with desktop-installed apps, he or she doesn't have the option of not upgrading. He or she either gets used to the changes or stops using it. An interesting corollary to the many ways web-based apps make life easier for developers.
-ev, 6:03:33 PM

Wednesday, August 25, 1999
New Blogger blog: Lake Effect
-ev, 6:34:14 PM

Dink Kuford: "Been there. Done that." It's a total, fatal attitude in



[Home](#) | [Pyra](#) | [Blogger](#) | [Pyra](#) | [Company](#) | [pyrAlert!](#)

Figure 1.24: First version of Blogger platform



[login](#) | [register](#) | [about](#)

[Welcome to Thefacebook]

Email:

Password:

[Welcome to Thefacebook]

Thefacebook is an online directory that connects people through social networks at colleges.

We have opened up Thefacebook for popular consumption at **Harvard University**.

You can use Thefacebook to:

- Search for people at your school
- Find out who are in your classes
- Look up your friends' friends
- See a visualization of your social network

To get started, click below to register. If you have already registered, you can log in.

about contact faq terms privacy
a Mark Zuckerberg production
Thefacebook © 2004

Figure 1.25: First version of Facebook platform



Upload, tag and share your videos worldwide!

[Sign Up](#) | [Log In](#) | [Help](#)

[Home](#)

[Videos](#)

[Channels](#)

[Friends](#)

[Upload](#)

[My Videos](#)

[My Favorites](#)

[My Messages](#)

[My Subscriptions](#)

[My Playlists](#)

[My Profile](#)

Watch

Instantly find and watch 1000's of fast streaming videos.

Upload

Quickly upload and tag videos in almost any video format.

Share

Easily share your videos with family, friends, or co-workers.

Recently Viewed (1 - 4 of 5) [More Recently Viewed...](#)



In Love
1 second ago



soccer violence comes...
2 seconds ago



The Rock meets the nWo
2 seconds ago



My Level 20 video
4 seconds ago

Today's Featured Videos... [See More Videos](#)



tokyo
tokyo at night
Tags // [tokyo](#)
Channels // [Travel & Places](#)
Added: 1 day ago by [videmul](#)
Runtime: 00:30 | Views: 4671 | Comments: 17
★ ★ ★ ☆ (9 ratings)



REMEMBERING CHICAGO
february 2005. a visit to IIT in chicago to see rem koolhaas' new addition to the student center. it required buying tickets to ride the el-train just so we could pass through the tunnel and back.
Tags // [chicago](#) [architecture](#) [koolhaas](#) [iit](#)
Channels // [Arts & Animation](#) [Odd & Outrageous](#) [Videoblogging](#)
Added: 1 day ago by [kammannohia](#)
Runtime: 01:04 | Views: 645 | Comments: 1
★ ★ ★ ☆ (6 ratings)

Sign up for your free account!

Nano a Day Giveaway Extended!

We're giving away a 4GB iPod Nano every day through the end of the year! Increase your chances of winning by:

- [Signing Up](#)
- [Inviting Your Friends](#)
- [Uploading Videos](#)

[View full contest details](#)

See all nano winners!

Explore New Features

Find the videos you want.

Channels

Subscribe to videos from your favorite users.

Subscriptions

Subscribe to videos from your favorite users.

Holiday Video Contest

Win an Xbox 360!



Figure 1.26: First version of YouTube platform

- 17 -

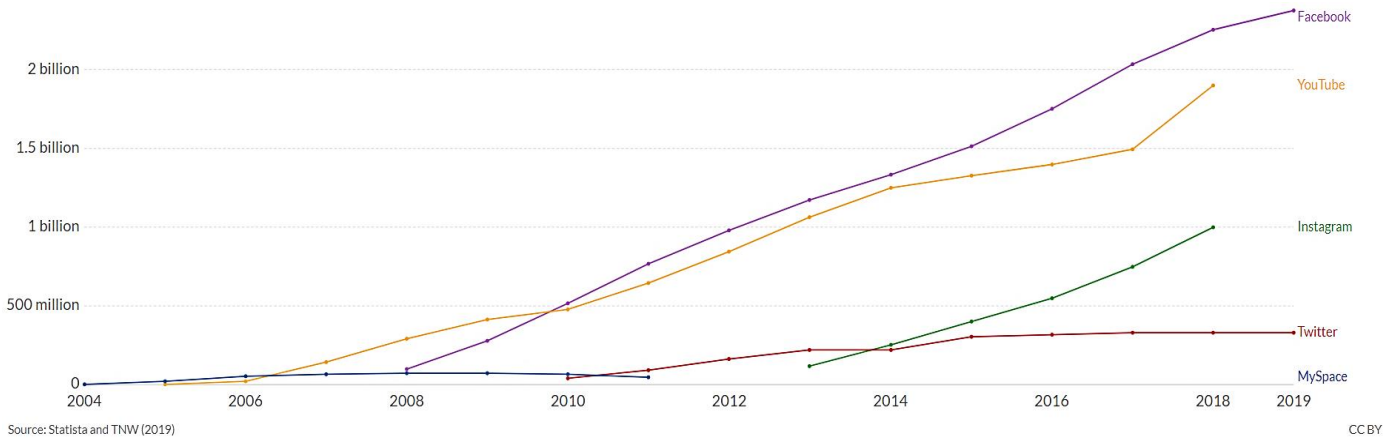


Figure 1.27: Number of users using social media platforms from 2004 until present (source: Statista and The Next Web and ourworldindata.org)

The reason that social media became popular is because of the ease of sharing an information. Humans are social beings by nature have the tendency to share their ideas or achievements to other individuals. Social media provide an easy way for someone to share an information with a fast way as a post of a user can be seen instantly by his friends and to react on his posts as well. Furthermore, this easy and fast way of sharing things through social media has made them a tool for news. For example, newspapers, televisions etc. have a centralized authority by providing rules or certain news that they will share to the public. Instead social media have the ability of sharing the information quickly and thus news can reach fast to the people [23]. This is also a negative aspect as sometimes news is not filtered or getting validated and thus can be false information (or fake news). The table 1.1 depicts a list with the socializing groups where can be formed through the most prominent social media as well as a small description about them.

Social networking platforms	Structure	Description
Facebook	Friends, Groups, Pages	A chat-based platform where users can create their friend list by adding other users. Chatting can be publicly through a post/comment or privately through message.
Twitter	Followers	A micro-blogging platform where its users can follow other users and can be informed about the latest news using hashtag trends or post a tweet which can be public or private (depending on user settings). It is a platform which is used frequently by politicians.
LinkedIn	Connections	A job seeking platform where its users can apply for jobs add other users. It is used mostly by unemployed individuals and recruiters. A user can add a post which is shared publicly or send a message to one or many of his/her connections.
Hi5	Friends	A chat-based platform where its users can create friend lists. It is similar with Facebook service.

Table 1.1: Most well-known chat based social networking platforms [23]

From the above table 1.1 we've analyzed some networking platforms, although these platforms differ a lot with each other. Social media platforms such as Facebook, are consisted by social networks built in a way where they can reflect the offline reality of human relationships. Although, there are users who do not use social media

properly by adding unknown to them persons or by providing false identity (fake accounts) which does not reflect the reality of offline human relationships [23]. It's important to mention the impact of Twitter as it is one of the most popular micro-blogging platforms and an outstanding source of data as its API has a lot of capabilities, allowing a developer to obtain plenty and useful information [21]. This research thesis focuses on Twitter data and in Section 1.8 we will discuss why Twitter data is being used.

Important note

The spreading and growth use of social networking platforms occurs due to their networking features that they provide which reflects their effectiveness and influence over the global population and on the market industry [12]. The process of gathering information for a specific person or a social target group has been simplified with social media as users tend to express and share their opinion through those platforms. So, if we want to identify the success for the spreading of social media platforms or let's say if we want to build a social networking platform ourselves the key for making it popular is by having or by integrating some sort of social networking features or functionalities [12].

There are also the media platforms which emphasize and influence uses to create and upload multimedia content as well as share it with others. With the term multimedia content, we refer to videos, images and audio. Over the recent years the growth of computing power has been enormous thus any computing mobile device from smartphones to tables, has the ability of creating multimedia content making the contribution to those platforms easier. Table 1.2 depicts the most prominent media platforms over the Internet.

Social media platforms	Structure	Description
YouTube	Video	A video sharing platform where users upload their videos by creating an account. Each uploaded video has a comment section below so that users can express their opinion.
Flickr	Photos	It is an image and video hosting service where users can share multimedia content with each other.
Instagram	Mobile Photos	A photo sharing platform for mobile devices. It has also private chatting features and users can create stories of their photos.
Vimeo	Edited "High-End" Video	A video sharing platform just like YouTube.
Pinterest	Variety of contents (Photo blogging mostly)	A photo-sharing platform where users upload their photos and share them publicly.

Table 1.2: Most well-known chat based social networking platforms [23]

The above-mentioned platforms as described in table 1.2 support a variety of different media types but they are known preliminary for one type. YouTube is considered as a video sharing platform but its users can also communicate through video comments section or through live commentary systems. Some media platforms

feature a specific form of one type of media, one common example is the blogs where they have features for large or small textual information. For example, Blogger or WordPress services can support large textual information whereas Twitter supports small textual information. Generally, Twitter can support a text with maximum limit of 280 characters. Figure 1.26 depicts the percentage of age groups who use social networking or media platforms over the Internet on the United States.

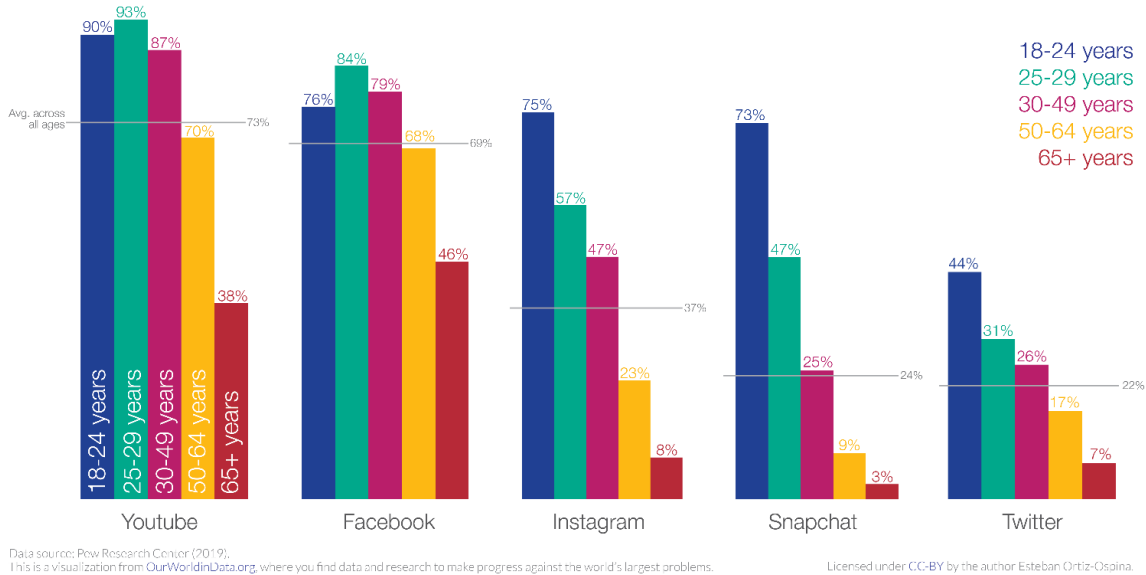


Figure 1.28: Use of social media platforms by age group in the US in 2019
 (Source: Pew Research Center 2019)

Figure 1.26 reveals the usage of the most prominent social media and networking platforms for a variety of different age groups where it reveals that younger ages tend to use them frequently rather than the older ones. This is reasonable since the elderly do not have experience in using computers and the Internet itself. It is worth to mention that some platforms appear to have more users from younger populations with examples such as Snapchat with 75% of young population (under the age of 25) using it as well as Instagram with additional 75%. Although these platforms are barely new to the Internet and it is not sure if the numbers of usage will be the same or if it will be increased over the years [24].

1.6. An overview on social media analytics

With the term social media analytics, we refer to the procedures of obtaining and processing data from social networks such as Facebook, Twitter and Instagram. Social media analysis is currently used by marketers in order to track comments or reviews made by Internet users about companies or products and extract if these opinions are positive or negative. Social media analytics can be defined as the art of science in order to extract useful hidden insights from a variety of social media data [24]. With the term "social media data" we refer to any kind of information obtained from social media. Social data have many forms and can be separated into two categories the structured and unstructured data [25]. To better understand the booth terms of structured and unstructured data we will use the Twitter's API data example which is fairly simple. With the term structured data, we refer to numeric or quantifiable information for the user of a particular social network, for example user's activity such as the number of likes per posted tweet or additional retweets, the location (georeferenced) where the user resides, the number of user's subscribers and so on. With the term unstructured data, we refer to information which has not any form of numeric value, for example the comments or tweets which a user uploads are in a textual form and have not any structural value as well as posted videos or images. These values can be rich in information but not as direct to analyze as structured data. To analyze unstructured data a processing

mechanism is necessary in order to get some form of numeric value. This mechanism can use artificial intelligence algorithms such as Naïve Bayes, SVM or Natural Processing Techniques meaning a lexicon as well as a corpus of words which the program will advise.

When analyzing data from social media there are three necessary steps that have to be followed. The first step is the process of data identification, the second step is the process of data analysis and the third step is the process of information interpretation. Data analysts define the question that needs to be answered in every point of the process in order to maximize the obtained value. These questions can help us in determining the proper data sources to evaluate which can affect the type of analysis that can be performed [26]. Figure 1.27 depicts a diagram regarding with the entire process of social media analysis.

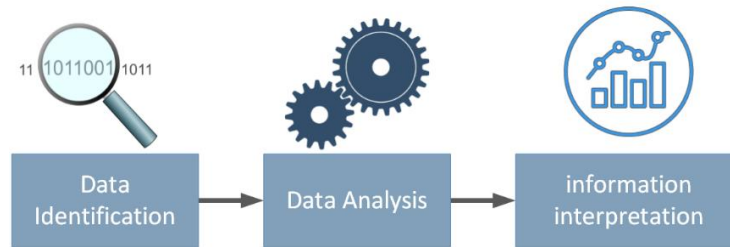


Figure 1.29: The three process steps of social media analysis

The next step is the process of data analysis where it is the point that filtered data are taken as input and get transformed into data values for the analysts. There can be a variety of analysis types through social media such as data which include procedures like sentiment analysis (whether if the text expresses positive, neutral or negative opinion) as well as geographic, demographics etc. To begin with the step of data analysis we need to define the problem, its solution and having the available data in order to have a meaningful result. Once we have enough sufficient data for analysis then we need to design a data model. The construction of a data model is the procedure where we organize the data elements and set a standard regarding of how separate data elements relate with each other. This step is crucial if a computer program is executed over the data and thus the computer has to know a set of words are important if there are any related words within an exploration topic. When analyzing data, it is important to have some available tools in order to obtain a reliable result. Let's consider an example of a word cloud for example "IT architect" and construct a word cloud, there is no doubt that the largest word of that word cloud would be the word architect. Some tools may be efficient for determining sentiment while other tools may be more efficient into analyzing a text into a grammatical form and thus enable us to get a better grasp of the meaning by using various words of phrases [15]. The last stage is information interpretation where it is the process of visualizing the information using the values that we have obtained from the data analysis procedure [27]. In data visualization there are three crucial aspects that we have to consider:

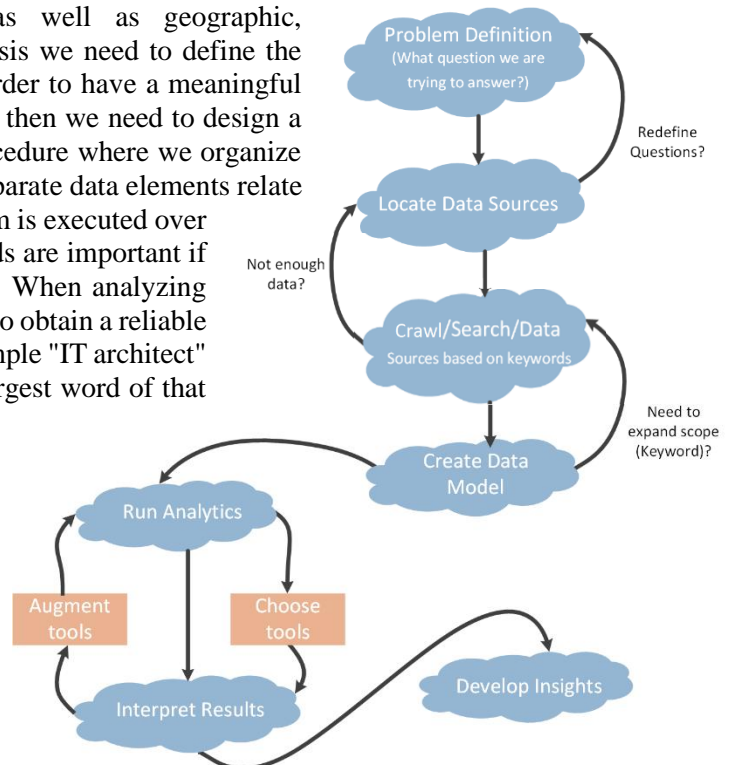


Figure 1.30: Social Media Analytics procedure

1. *Understanding the audience:* Before we proceed in any visualization within a given a set of data it is important to understand its target group and the people that interests them. In more simple terms we have to ask ourselves the question "Who is the audience" and how capable is to know the used technologies [27].

2. *Setting a clear framework:* A data analyst need to make sure that the visualization process is syntactically and semantically correct. For example, when using an icon, the element should bear resemblance to the thing it represents, with size, color, and position all communicating meaning to the viewer [27].
3. *Provide a description:* Any analytical information is difficult to be understood or to comprehend therefore a data analyst has to ensure that there is sufficient information for the viewer in order to understand them. Describing a story to the viewer it is crucial since it helps him/her of understanding the visualizing concept of data [27].

1.7. Summary

From this chapter we gained basic knowledge of how the Internet emerged, how the first network of computers was created the emerge of Web 1.0 and Web 2.0 technologies as well as the emerge of the Semantic Web (3.0). We've also analyzed how social media emerged, their overall evolution as well as its importance on semantic web. In addition to that we've also discussed the steps of analyzing information obtained by social media. On the next Chapter 2 we will discuss the ways of extracting data over social media and especially Twitter, we will also discuss briefly how Twitter's API works, what data we can obtain from this networking platform as well as the raw information and how to format it into a processable information for data visualization.

Chapter 2: Text mining methodologies over Twitter

2.1 Collecting and analyzing data from Twitter

Twitter is a micro-blogging platform and social networking service where its users can submit their opinion publicly in a form of a limited text or video known as "tweets". Its registered users can interact to the tweets of other users with likes and retweets. Beyond likes and re-tweets there are also other forms of engagement which Twitter's users can use such as replies, mentions and favorites. This form of interaction makes Twitter an interesting platform for online marketers as this form of engagement can "tell" if a user would likely buy their commercial products due the reason that it incorporates one-to-one conversations as well as promotion to their circles or influence. Table 2.1 describes the engagements that can be used by Twitter's users [28].






Engagement Icon	Engagement type	Description
	Reply	Users can reply to posted tweets by other users and say their opinion about the context of the tweet. It is a feature which is used frequently. Online marketers rely a lot to this feature as from the replies we can extract the opinion of a user for a commercial product.
	Retweet	Users can share a posted Tweet by re-sharing it to their circle of followers in their accounts. Companies rely a lot in this feature as retweets enable the free promotion of their commercial product as it reaches to other users.
	Mentions	Users can mention the account name of another user. In politics, voters who use twitter tend to post a tweet and mention the politician that relates with it. For example, “@RealDonaldTrump Nice campaign in back in 2016” or “@RealDonaldTrump Is the wall ready?”.
	Likes	Users can use this feature when they agree or like the context of the tweet. It is a useful as from it we can detect if the users like a politician or not. The same thing goes for commercial products as well.
	Hashtags	It is a feature that can be used by users that marks the tweet as a public one. Users can use a hashtag regarding to the context that the tweet contains. For example, “Donald Trump has successfully built the wall in order to protect our country #Trump2016 #OurPresident” or “Donald Trump lied about the wall #TrumpDictator #NotMyPresident”. Hashtags may indicate negative or positive opinion for a political leader and it is used by data analysts who analyze the political landscape of a country.

Table 2.1: Engagement features of Twitter [28]

To start with analyzing and collecting data from Twitter we need to be a registered user. Once we do that then we need to create a Twitter API, in general terms Twitter provides the ability of creating an API through their platform in order to obtain access to data posted by the users of the platform. The creation of an API is

useful as without it, developers cannot access a social media web service neither obtain data from them. Twitter has simplified the creation of their API where developers can visit to the additional developer's page (developer.twitter.com) and send a request for its creation. Once the request is approved the developer gets notified with an email by Twitter with the link of their requested Twitter API [28]. Figure 2.1 depicts the structure of data exchange procedure in Twitter's API and table 2.2 depicts a description about the four Twitter's API "keys" set.

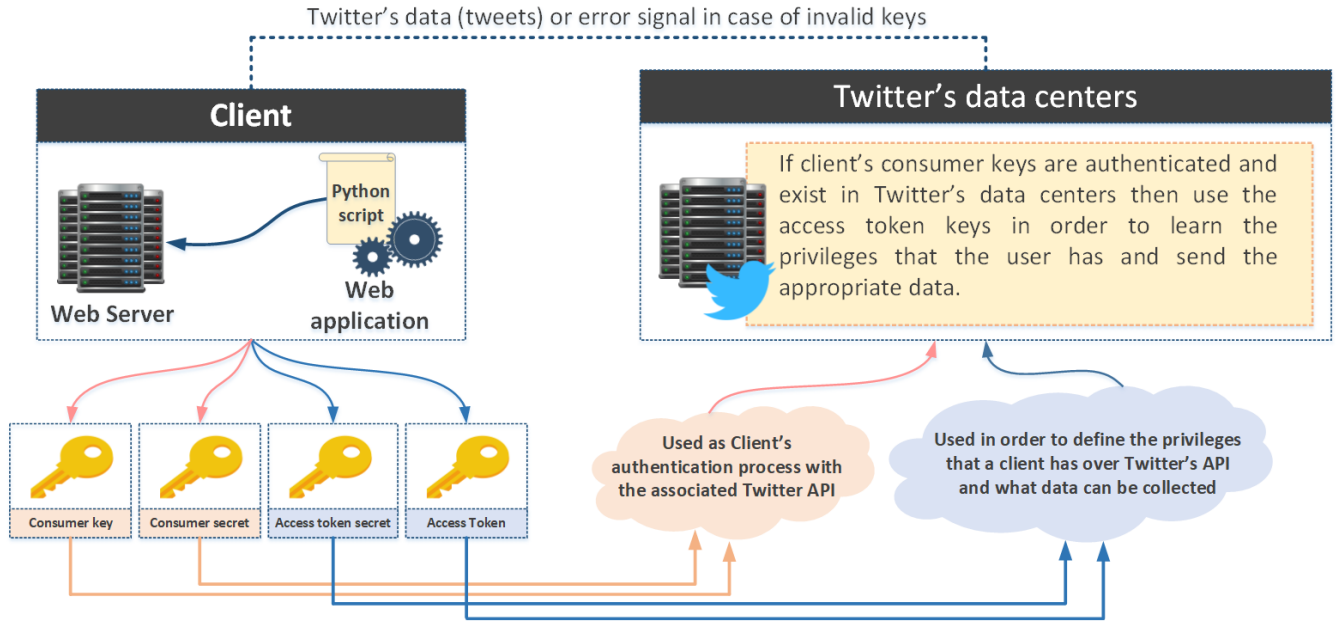


Figure 2.1: An abstract concept of how Twitter's API data exchange procedure occurs [28]

Key Type	Description
Consumer Key	It is the API key which associates with the Twitter micro-blogging platform and identifies uniquely the client. With the term client we refer to a user which uses a script (Python or Java) or a web service in order to access Twitter's resources.
Consumer Secret	It is the unique ID of the client which is used for the authentication process of the authorized Twitter's server.
Access Token	It is issued once the client is successfully authenticated (using the consumer key and secret) with the Twitter's server. Typically, it defines the set of privileges that the client has and the amount of data that can be collected from Twitter's data center.
Access Token Secret	It is the ID of the issued access token.

Table 2.2: Description of authentication keys used by the Twitter's API [28]

Twitter has adopted an open standard for their API authentication process namely as Open Authentication (OAuth) where it provides access to protected information. Due to the reason that passwords are vulnerable and can be obtained by third non-trusted individuals (commonly known as hackers), OAuth is safer alternative approach to the traditional authentication methods which require the use of three-way handshake process. Twitter's API uses OAuth authentication process for its API requests where these are generated by an application which uses it (e.g. an application which obtains a sample of tweets and visualizes the additional likes and retweets which they have) [30]. Bellow we analyze the detailed steps that are carried out for API call from an application which leverages Twitter's using OAuth authentication process:

1. Applications which leverage Twitter's data are known as consumers where all of them need to be registered with Twitter. The consumer application makes use of the consumer and secret keys in order to be authenticated with Twitter.
2. The consumer application uses the consumer and secret keys in order to create a distinct Twitter link by which the user is directly forwarded for authentication. Afterwards the user authorizes the application by verifying his/her identity on Twitter. Once the identity of the user is verified by Twitter an OAuth verifier (or PIN) is issued [30].
3. The OAuth verifier is provided by the user towards the application. Once the application has it then sends a request in order to gain the corresponding access token and access secret keys.
4. With access token and secret obtained the application can authenticate the user on Twitter by issuing API call on his/her behalf [30].

It is worth to mention that the “access token” and “access secret” keys for a user do not change therefore the entire authentication process occurs once. In figure 2.2 it is depicted an illustration example of the above authentication steps of OAuth authentication model.

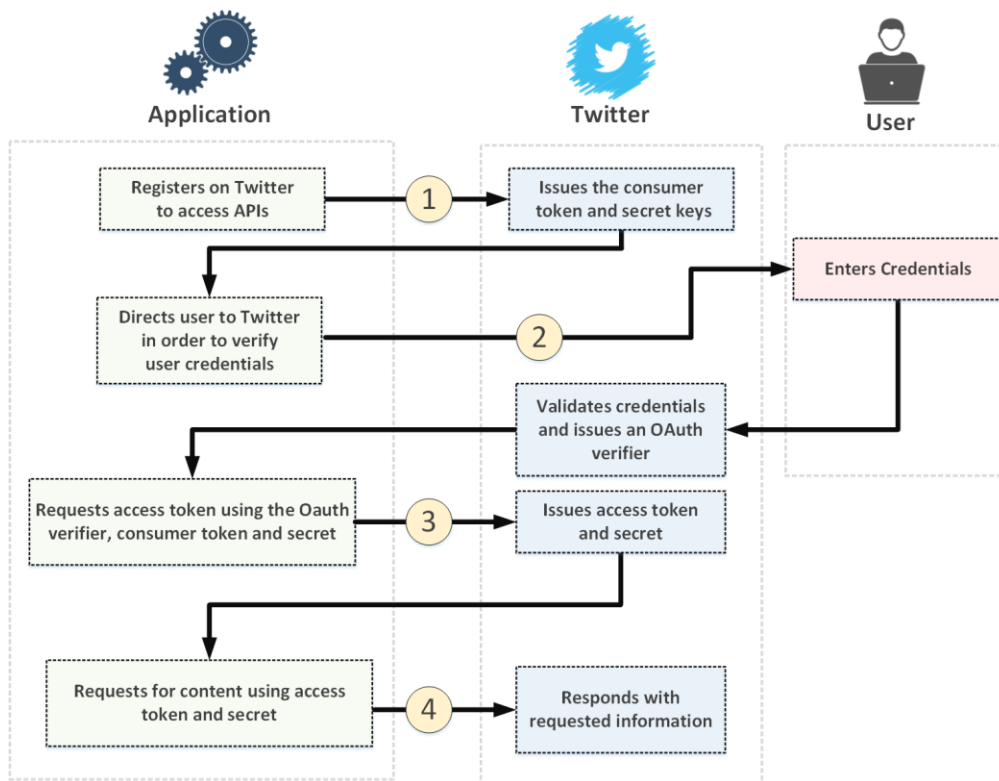


Figure 2.2: Workflow of OAuth authentication model [30]

Since that we've covered how authentication process occurs through Twitter's API, we can discuss how we can obtain data through Twitter. Before we start to analyze data from Twitter, we have to understand the structure of a tweet. Generally, tweets are short messages that consist by a maximum of 280 characters. Since Twitter is a micro-blogging and news information service (quick and short messages), its users tend to use acronyms, making spelling mistakes as well as using emoticons and/or any other special characters in their entire text. On the previous chapter 1, we've mentioned about structured and unstructured data. The process of obtaining information from a tweet provided by Twitter's API is structured data while the process of extracting information within from the text of a tweet is unstructured information. The process of obtaining unstructured data from tweet we will discuss it in detail on the next section 2.2 for the moment let's focus on structured data and what information we can retrieve using Twitter's API. By default, a script uses Twitter's API that a developer has created on the official developer page of Twitter. Using the four-key set with the use of a module it is possible to retrieve structured data from a tweet. Typically, two programming languages are well known for information extraction through Twitter the Python and Java programming languages. These two programming languages have additional modules that allow a programmer to leverage the capabilities of Twitter's API. Java uses Twitter4J while Python uses Tweepy module. For the purposes of our research Python is used as it provides more capabilities for data analysis as well as its simplicity and due to the fact that it is a well-known programming language on the field of data science. Let's suppose that we have a Python script and we want to obtain structured data from a tweet. To do this we need to create an API through Twitter's website on developer section and install Tweepy module. Tweepy provides us the ability of searching data through Twitter. To better understand this let's see the example bellow.

```
#comment 1: Import Tweepy module
import tweepy

#comment 2: Import csv library
import csv
import pandas as pd

#comment 3: variables that contain the keys from the created API
consumer_key = 'the given key code from the created API'
consumer_secret = 'the given key code from the created API'
access_token = 'the given key code from the created API'
access_token_secret = 'the given key code from the created API'

#Comment 4: OAuth authentication process
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth,wait_on_rate_limit=True)

#Comment 5: Open/Create a file to append data
csvFile = open('ua.csv', 'a')
#Comment 6: Use csv Writer
csvWriter = csv.writer(csvFile)

#Comment 7: For loop take 100 recent tweet sample and store them on a csv file
for tweet in tweepy.Cursor(api.search,q="#unitedAIRLINES",count=100, lang="en").items():
    print (tweet.created_at, tweet.text)
    csvWriter.writerow([tweet.created_at, tweet.text.encode('utf-8')])
```

The above Python script uses Tweepy module and uses the four set keys from the Twitter’s API created by a developer. The basic functionality of the script is to collect a sample of 100 recent tweets that contain the hashtag #unitedAIRLINES. From this sample store on csv file the date that the tweet was posted as well as its corresponding text. The structured data that we can obtain from a tweet are the following.

1. text: the text of the tweet
2. created_date: the date that the tweet was posted
3. favourite_count: the number of likes that a tweet has
4. retweet_count: the number of re-tweets that a tweet has
5. place, geolocation: location of the user that posted the tweet
6. user: the full user profile entities: the list of entities like URL's, @mentions and #hashtags

From the above list we can realize that there are interesting data that we can obtain from a tweet. We can check the popularity of an account whether from the number of likes that have in each tweet to the number of his/her subscribers and so on. Although these are just a small portion of what structured data, we can obtain from Twitter using Tweepy module. There is a detailed list provided in their official documentation [31]. The most crucial one is that we can obtain the text of the tweet. This is very useful as it might contain useful information for a subject or a topic. For example, extracting how positive or negative is the text of a tweet over a generic subject. There are many ways where we can achieve this and we will analyze it on the next section 2.2 as it is part of sentiment analysis procedure.

2.2 Lexicon-Based Sentiment Analysis

With the term sentiment analysis (or opinion mining) we refer to the automatic extraction process of subjective information from a text (unstructured data) using natural language processing techniques [32] which we will analyze them on the next section 2.3. The extracted information might be the opinions as well as the sentiment of a user regarding a subject through text analysis. Sentiment analysis has been an active research area on the field of natural language processing since the early 2000s [33]. Pozzi et al. [32] mention that there has been unsurprising confusion among researchers of the field by debating whether if the field should be called sentiment analysis or opinion mining and the difference between those two terms. Merriam-Webster's Collegiate Dictionary defines the word sentiment as a thought, attitude or judgment motivated by the feelings of a crowd while the word opinion is defined as the judgement, view or appraisal that someone has over a subject. The latter word is more egocentric as it defines the thoughts of one individual while the other one defines the thoughts of a group of individuals or a crowd. Let’s compare to phrases the phrase is “*The current political situation concerns me*” and the phrase “*I think that politics is not going well*”, the first phrase expresses a sentiment while the second one expresses an opinion. For the first phrase the most logical response of a person would be “*I have mutual thoughts*” while for the other phrase would be something like “*I do/don’t agree with you*”. Although the two phrases have similar meaning the first phrase tends likely to sentiment expression while the second phrase tends to an opinion about something. Ironically, the first phrase indicates a negative opinion about politics which is the same with what the second phrase implies. There can be also phrases which imply positive opinion for example “*My belief is that he will win the next parliamentary elections*” implies a positive opinion. Liu et al. [33] suggest that an opinion can be defined as a quintuple as shown by the equation 2.1.

$$\left(e_i, a_{ij}, s_{ijkl}, h_k, t_l \right), \tag{2.1}$$

Where e_i is the name of an entity, a_{ij} is an aspect of e_i , s_{ijkl} is the sentiment on aspect a_{ij} of the entity e_i , h_k implied the opinion holder, and t_l is the time the opinion is expressed by h_k . The sentiment s_{ijkl} can be positive, neutral or negative or it can be expressed differently strength/intensity levels such as the 1-5 stars system which many review websites use such as Amazon.

Let's suppose an individual with the name Alex, where the previous day he decided to buy the new iPhone. The day after he bought it, he was testing it for the entire day and in the afternoon let's say at 20:00 o'clock on 12/3/2020 decided to submit a review regarding to that product. He commented the following text, “*The new iPhone is good enough, although battery life and its security need to be checked*”. Let's highlight the following words "iPhone", "battery life", "security" and index them as 1, 2 and 3. Alex has indexed as 4 and the time when he wrote the sentence is indexed as 5. Then Alex is the opinion holder h_4 and t_5 (20:00 at 12/3/2020) is the time when the opinion was expressed by h_4 (Alex). The term "iPhone", $s_{1245} = \text{neg}$ is the sentiment on aspect a_{12} ("battery life") of the entity e_1 ("iPhone") and $s_{1345} = \text{neg}$ is the sentiment aspect a_{13} ("security issues") of entity e_1 ("iPhone"). When an opinion is on the entity itself as a whole the special aspect "GENERAL" is used to denote it. As we mentioned above with the term sentiment analysis, we refer to the definition of automatic tools capable of extracting subjective information in order to produce structured data or a knowledge about something. From the above quintuple-based definition example we can realize that it provides a framework for transforming unstructured data into structured ones and thus extract sentiment from a given text. Onwards this extracted information forms a rich-full portion of qualitative and quantitative set of information where trend analyses can be performed utilizing database systems and online analytical processing tools. Due to the importance of the sentiment analysis to the corporations and to the society itself, it has reached from computer science/engineering to management and social sciences. In just few years sentiment analysis has been embraced from the industrial sector with colossal companies such as Microsoft, Google and IBM developing additional tools. Due to its strong potentials as a new emerging field, sentiment analysis has an active presence both on academia and industrial sectors. These factors set the field of sentiment analysis as a growing trend, figure 2.3 depicts a graph created in google trends where it depicts the trend keyword sentiment analysis as it was searched by Internet users for the last decade.

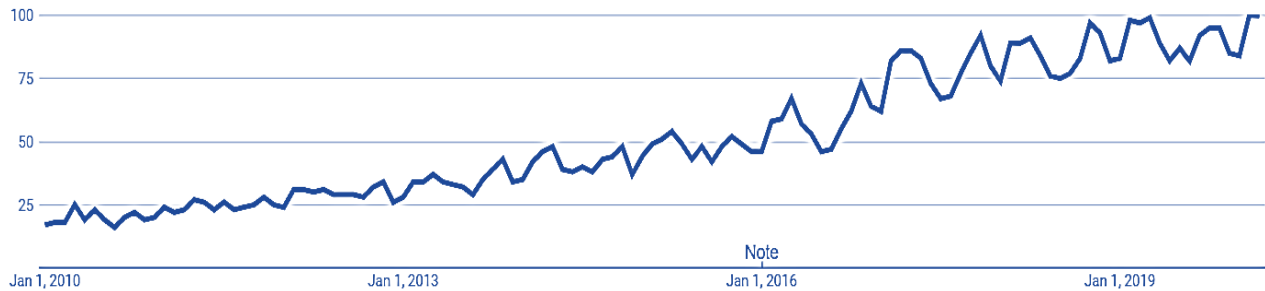


Figure 2.3: Google trends data related to the keyword sentiment analysis

As the growth of social networks is constantly rises, so does the presence of sentiment analysis, thus forming the field of social media analytics. Due to the immense diffusion and their prominent role in our nowadays society, it reflects one of the most profound novelties in recent years and captures the interest of researchers, companies, governments and journalists. This immense interconnection between active users sets the basis for a dialogue that is capable of inspiring and involving individuals in a greater agora; connecting citizens to shared interests and promoting different types in collective action. With that said social media platforms set the foundations for a digital revolution which triggers the speeding of ideas and thoughts around the globe in just a matter of seconds. Data from social media which indicate an opinion if are gathered and analyzed properly allow us the ability not only of understanding and predicting social events. The current progress on technological advancements allow us the efficient storage and retrieval of big-size data. Currently the main focus is on how we can extract information from these data and thus form knowledge from raw information. Social networks reflect a brand-new emerging field (in terms of big data) where natural language expressions of users can be easily reported through short textual information (posts, tweets etc.) and thus creating a unique content. If we successfully and effectively analyze this unique content, actionable knowledge is formed that can be utilized for decision making processes. Pozzi et al. [32] mention that immense amount of constant

textual information created by the users of social networks have to be processed in real time in order make informed decisions, calls for two main types of radical progress:

1. A change in study strategy across the transition from data-constrained towards data-enabled paradigms
2. The integration of psychology, sociology, natural language processing and machine learning in order to form a multi-disciplinary field.

Corporations and businesses tend to show their interest towards data from social media as its users tell their opinion about their products and helps them in order to observe if they like them or not. The same thing applies for government institutions as from social media we can extract the opinion users regarding to social or political events. Pozzi et al [32] mention that sentiment analysis sub-task aims towards positive, negative or neutral sentiments extraction from texts. These extracted sentiments are also called polarities, with this said it is worthy to mention that sentiment analysis must not to be confused with polarity classification. Depending on the data subject which an application analyzes the term “sentiment analysis” can also be met as opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis and review mining). Figure 2.4 presents a taxonomy regarding to the most popular sentiment analysis tasks as mentioned by Pozzi et al. [32].

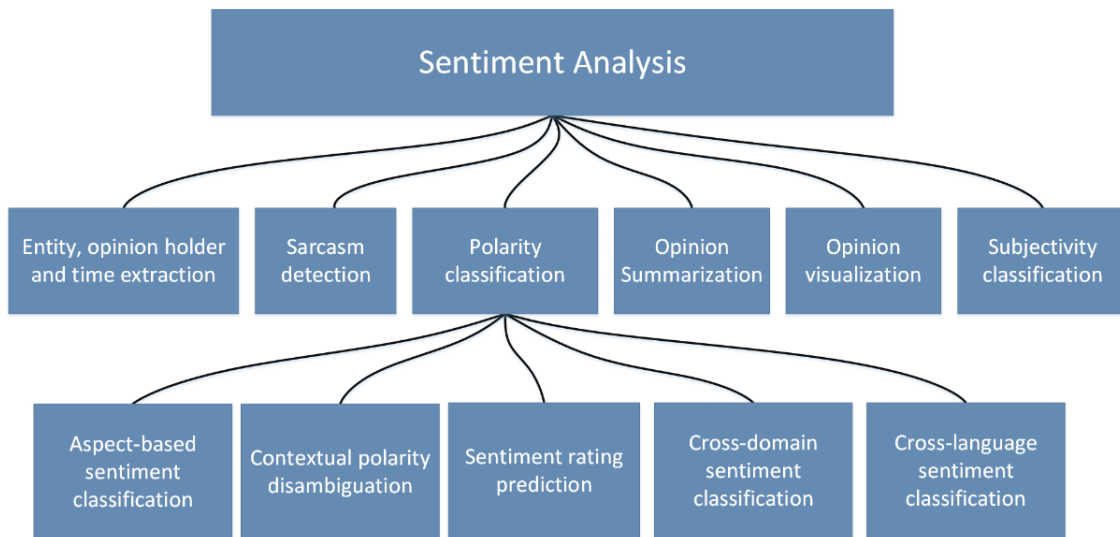


Figure 2.4: Tasks of sentiment analysis as presented by Pozzi et al. [32]

Important note

With the immense growth of social media (the emerge of 1 to 5 stars voting reviews, discussion forums, blogs, micro-blogs, Twitter as well as commenting and posting services) on the Internet, corporations and businesses showed their interest for obtaining and analyze data from those platforms in order to perform a decision. For example, if an individual wants to buy a commercial product let's say a mobile phone there is likely the chance of leaving a review regarding to this device, sharing its thoughts with other users and thus affect them. With this system a corporation doesn't have to conduct surveys, asking for the opinion of consumers since every opinion is publicly available. Thus, automated sentiment analysis systems are crucial in order to extract information from the data contained in social media. Liu et al. [33] mention that in recent years, there is an increasing rate of users who tend to express opinionated posts in social media and

that helped the reshape of businesses as well as in the political aspect since public sentiment and

emotions can be extracted from social media. One example of the effect that social media have over political landscape is the Arab spring which was a movement created through social media. Due to the demand of automated sentiment analysis systems more and more applications which make use of sentiment analysis mechanisms have flourished from consumer products and healthcare to social and political events. Colossal companies such as Microsoft, Google, Hewlett-Packard have created additional sentiment analysis mechanisms or applications for a variety of needs. Beyond from real-life applications there have also been many research-oriented applications where they have been published in additional research papers. One profound example of them is by Liu et al. [34] sentiment analysis model which was proposed to predict sales performance using blogs. Another interesting study is by O'Connor et al. [35] which they present Twitter text sentiment analysis and its linking with the public opinion polls. There is also a plethora of other research-based applications which use Twitter's data for analysis which include prediction of political election results [37], movie ratings prediction using data from Twitter [37].

When we want to analyze a text and thus extract sentiment analysis, we need to make a distinction between subjective and objective sentences. If we have a sentence that has been classified as objective then there are no other required fundamental tasks. In addition to that, if a sentence has been classified as subjective then we have to calculate its polarity meaning if it implies positive, negative or neutral sentiment. Figure 2.5 depicts a visualized example of this distinction based on Pozzi et al. [32] example.

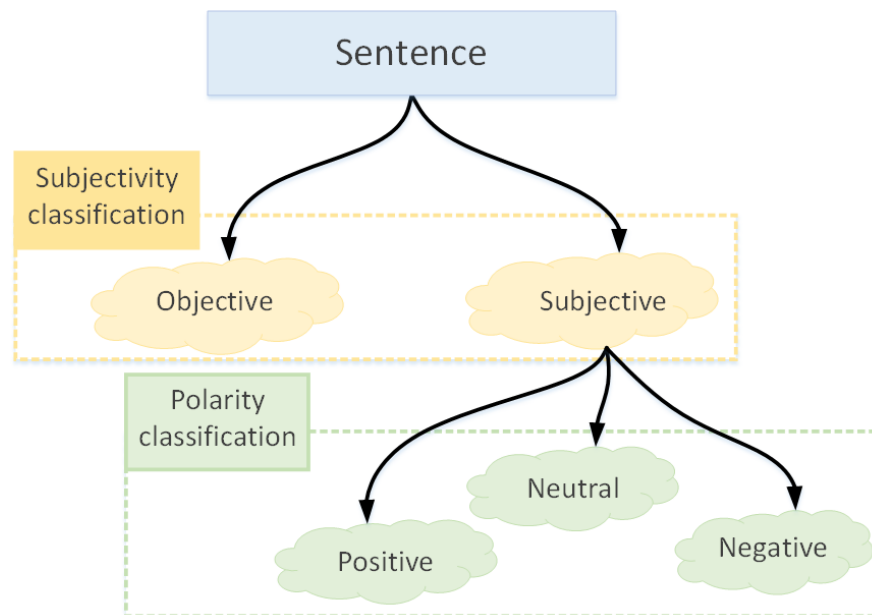


Figure 2.5: Distinction between subjectivity and polarity classifications as presented by Pozzi et al. [32]

As we previously mentioned subjectivity classification is the process that distinguishes sentences that have objective information (objective sentences) with sentences which express subjective views and opinions (subjective sentences). To understand this let's give an example with two sentences the first one is "Nokia 7 Plus

is a smartphone” and the second is “*Nokia 7 Plus is awesome*”. The first sentence indicates a fact or something which exists while the second sentence indicates something positive, in the case of our second sentence indicates something positive for a particular smartphone. The first sentence can be classified as an objective one while the second sentence can be classified as a subjective one.

As it is mentioned above, the purpose of sentiment analysis is “the definition of automatic tools which can be capable of extracting subjective information from texts in natural language”. When we want to apply sentiment analysis our first task is to define what textual information will be analyzed. Sentiment analysis which is applied in social networks can be separated into three layers as figure 2.6 depicts.

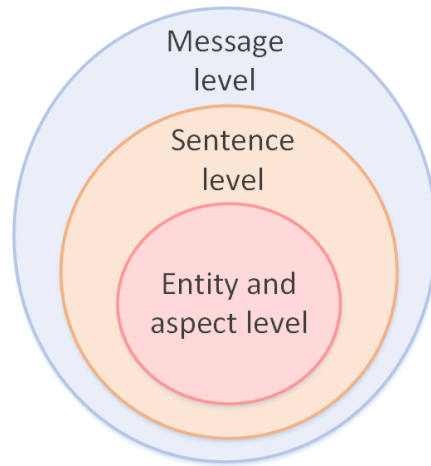


Figure 2.6: Different levels of analysis as presented by Pozzi et al. [32]

<i>Level</i>	<i>Description</i>
<i>Message or document</i>	The purpose at this level is the polarity classification of the entire message that indicates an opinion. For example, let's suppose that we have a tweet text from an individual, the system will determine if the entire text of the tweet expresses a positive or a negative opinion about something e.g. a politician, an event, a commercial product etc. Therefore, the whole message indicates one opinion over a single entity. Liu et al. [32] suggest that this level is not applicable for documents which evaluate or compare multiple entities.
<i>Sentence</i>	At this level it is determined the polarity of each sentence that a text message contains. Therefore, each sentence depicts a message which is classified accordingly on the entire text of the document.
<i>Entity and aspect</i>	A fine-grained analysis is performed comparing to the previous levels. At this level the opinion is consisted by a sentiment and a target (of option). For example, the sentence “Samsung phones are good but do not have efficient battery” indicates to opinions one that Samsung phones are good which is positive and a negative which “do not have efficient battery”.

Table 2.3: Describing the different levels of analysis [32]

As we previously mentioned, an opinion can be classified as positive, neutral or negative, although an opinion might indicate different meaning from what it seems to have. There are mainly two category groups the

explicit and the implicit ones. The explicit means a phrase or a statement that declares subjectivity and they can be separated into two categories the Regular and Comparative opinions. As its name implies, the explicit regular or “standard” opinion can be separated into two different types the direct opinion which refers directly into a single entity and the indirect opinion which refers indirectly onto a single entity based on its impact on other entities. For example, the phrase “After I switched to iPhone, I had problems with battery performance” describes the effect of battery inefficiency over a product that the user switched with. The explicit comparative opinion expresses a relation of similarities/differences between two or more entities and/or a preference of the opinion holder is based on some shared aspects of the entities [38]. For example, the sentences “Linux is better OS than Windows” and “Linux is the best OS” is an example of two comparative opinions. The second category which is the implicit opinion defines an expressed opinion over a desirable or undesirable fact. For example, the phrases “Better Call Saul new season series starts tomorrow at 20:00 on Netflix I can’t wait until it starts” and “Breaking bad series was more violent than Better call Saul.”. The first sentence implies that there is a good expectation for the Better Call Saul series but it is not explicated with words. In the case of the second sentence it is not easy to uncover its sentiment indications even for humans. Some people might find violence as a good characteristic for the entire plot of a series while others might not. It is obvious that from the two given types of opinions the implicit ones are easier in order to detect the underlying sentiment of an entire sentence.

To be able of analyzing user expressions through social media it is important to analyze the semantics of the language. Therefore, the content of the textual expression is evident to achieve successful sentiment analysis and thus discover the underlying sentiment of the text. If we analyze ourselves a sentence, we might lead to an assumption that it is positive or a negative one based on our own perspective whereas from the perspective of sentiment analysis this might be the completely opposite. Let's see an example the sentences "I just watched the most terrifying horror movie. It was really spooky! GEETTT PANNNIIICCCC" initially would be interpreted as negative. However, taking into consideration the sense in which these views are articulated (i.e. community of horror film fans) as well as other lexical views characteristic of the vocabulary of the social network, we can draw a (actual) optimistic assessment. Lexica, corpora and ontology need to be correctly developed and then used to grasp the meaning of natural language in social networks online.

Important note

The assumption basis of using sentiment analysis based on social networks is that the provided texts by the users are independent and identically distributed. Although several efforts have been made in order to handle complex characteristics of the language in social networks as there is still an open issue when it comes of dealing with content of the user generated texts. A first tentative solution is to deal with the true existence of social network information relevant to the principle of homophily [39] (the propensity of people to identify and connect with similar individuals, as in the quote "Feather birds flock together" [40]). In that sense, associations with "friendship" could be used to deduce whether related users may have similar views. Even so, a sentiment analysis system would take into consideration that the presumption regarding friendship connections may not match the reality adequately, where two related individuals may have opposing views on the same subject. Based on this observation, to properly reflect user and post relations, many other pieces of contextual knowledge may be derived from the social network itself. Interactions focused on or expressing an affection could be more insightful than a casual friendship.

Dealing with figures of speech is also an important aspect in order to perform accurate sentiment analysis. For the term figure of speech, we refer to an artful etymology of the normal style of expression or writing. According to Aristotle's tradition, figures of speech can be separated into two parts the schemes and the tropes. Schemes and tropes have the role of doing a sort of transference; schemes are defined by a transition in time, whereas tropes are defined by a transference of context. For starters, irony and sarcasm, which are grouped under

the tropes section, are the most controversial figures of speech in natural language processing. Although irony is sometimes used to illustrate events that deviate from the intended, such as twists of nature, sarcasm is often used to express subtle critique with a particular entity/person as its goal [32]. To get a better understanding of this let's see some examples of ironic and sarcastic sentences.

Sarcasm: (Note: Nick dislikes Kate's set of squirrel dolls)

- *Nick: What a huge collection of squirrel dolls. Remind me how old are you yet?*

- *Kate: Ha! Who cares?*

Irony (Note: Nick and Kate watch a movie although it wasn't as good as they expected)

- *Nick: Thank God it has finished! What an investment of our times.*

- *Kate: Couldn't agree more on that.*

On the irony example there was no sarcasm due to the reason that Nick doesn't had intentions to hurt Kate with his sayings. Nick was ironic to express his disappointment for the movie and for the time that he spends watching it. In the example of sarcasm Nick used sarcasm to show to Kate that he disliked the set of squirrel dolls that she has [32].

2.3 Identifying the issues of sentiment lexicons

On the previous section we referred some basic aspects of sentiment analysis, the types of opinions and the issues that exists for the creation of automated sentiment analysis mechanisms [42]. Although there are other crucial aspects that automated sentiment analysis mechanisms use in order to identify the polarity of each word. Perhaps one of the most usual forms of research on the field of sentiment analysis is the creation of resource dictionaries which include terms, expressed opinions, words with polarity (marked as positive, negative or neutral) emotion expression [43]. The last method is identical for political sentiment analysis due to the reason that an automated sentiment analysis mechanism will advise a resource dictionary in order to identify if the words contained on the text are positive, negative or neutral based on the indication [33]. Each word contained on a lexicon has a rating which indicates how positive or negative is. The negative type can be indicated from -2 (very negative) to 2 (very positive). Based on that there are a variety of different types of sentiment analysis, table 2.4 depicts all those types.

Type of sentiment analysis	Description
<i>Standard</i>	This type is used regularly as it classifies the emotional tone of the overall expression as positive (1), negative (0), or neutral (-1).
<i>Fine-grained</i>	This type uses a scale from very positive (-3) to very negative (3) and thus it provides a more fine-grained polarity of the overall sentiment.
<i>Emotion detection</i>	This type identifies specific emotions such as happiness, anger, frustration, sadness etc.
<i>Aspect-based</i>	This type of sentiment analysis classifies the text according to its related topic. It identifies if the content of a text and guesses its expressed topic. For example, the sentence "I like Olympic Games and specifically Pentathlon" it is related to the sports topic.
<i>Intent detection</i>	This type attempts to identify the actions of the person who wrote a text. It is best to be used for real-time sentiment analysis.

Table 2.4: Different types of sentiment analysis [43]

Liu et al. [33] mention that the most profound indicator of sentiments are the sentiment words. These are mainly words which are regularly used in order to identify positive or negative sentiments. Words such as

“good”, “wonderful”, “amazing” can be categorized as positive words while words such as “bad”, “poor”, “terrible” are categorized as negative ones. Beyond words themselves there are also phrases and idioms for example. Sentiment words and phrases are a key for sentiment analysis. Thus, with the term sentiment lexicon we can refer to a document which contains a set of sentiment words which indicate mostly a positive or a negative result.

2.4 Sentiment Analysis with Machine Learning

The term Machine Learning reflects the capability of a system to acquire and integrate knowledge by using an automatic way. To classify the emotion underlying from a text several techniques can be applied. There are three machine learning categories the supervised, unsupervised and semi-supervised. Supervised learning is the process of learning a function that maps an input to an output based on the example of input-output pairs [44]. A machine learning system which uses supervised techniques when it begins the learning process it will require to have an input namely as training data in order for the classifier to recognize and obtain knowledge for more representative differences between texts belonging to different categories [42]. Basically, in supervised machine learning we have an input variable X then this input is transferred to an algorithm in order to obtain knowledge (learn the mapping function from the input) and thus produce an output variable Y. This can be expressed by the following 2.2 equation.

$$y = f(x) \tag{2.2}$$

Supervised learning has two types of problems the regression and classification. The classification problem is when an output variable is a category e.g. “green” and “yellow” or “decision” and “no decision” while the regression problem is when the output has a real value e.g. “euros” or “weight”. Several specific categories of issues developed on top of classification and regression require estimation of the recommendation and time series, respectively. Some profound examples of supervised machine learning algorithms are the Linear Regression, the Random Forest, Support Vector Machines – SVMs, Naïve Bayes – NB as well as Maximum Entropy [45]. Unsupervised learning systems are driven only by arriving inputs (data) in order to uncover potential secret structures among them with main purpose to classify them into groups that reflect a certain similarity. In more simple terms, in unsupervised learning we have input data x but non any produced output. The concept idea in unsupervised learning is to get the data and model their underlying structure in order to learn as much as possible information’s about them [42]. As its name implies (“unsupervised”) these systems do not provide any guarantee for correct answers and there are none training models. Algorithms decide themselves to present the interesting structure within the data. An unsupervised system is in the position to know the existence of numbers and properties of groups and the need for a training set for extraction of feature vectors is not a prerequisite. In addition to that, unsupervised systems use pre-built emotion dictionaries for various and thus the various terms contained in a text are labeled by the overall polarity arises [42]. There are two types of unsupervised learning problems [45]. The first type is namely as clustering occurs when we want to discover the inherent groupings in the data such as grouping customers by purchasing behavior and the second type namely as association rule occurs when we want to discover rules that describe a large amount of our data, for example when people are buying x will likely have the tendency to buy y item as well. Some profound examples of unsupervised algorithms are k-means used in clustering problems and Apriori algorithms for association rule learning problems. In semi-supervised learning the system tries to learn with limited text data which are already labeled with their corresponding emotion content. Spatiotis et al. [42] mention that this method is time consuming as it needs the creation of a dictionary but it considered as the most secure [45]. A variety of machine learning problems exist on this area as it is time consuming to label data as it requires domain experts as well as time consuming or expensive. Generally semi-supervised leaning algorithms are characterized by a large amount of data X and a small portion Y labeled data. Table 2.5 summarizes the above-mentioned types of machine learning algorithm categories [45].

Types of machine learning	Description
<i>Supervised</i>	Every data is labeled and the algorithm tries to learn in order to predict an output from the given data
<i>Unsupervised</i>	Every data is unlabeled and the algorithm tries to learn to inherent structure from the input data
<i>Semi-Supervised</i>	A small portion of data is labeled while most of them are unlabeled and a combination of supervised and unsupervised techniques can be used

Table 2.4: Summarization of different machine learning techniques [45]

2.5 Classification algorithms in Machine Learning

In this section we will analyze some basic machine learning algorithms. For starters, with the term “classification” we refer to a system/application which uses machine learning methods and it is capable of learning from the input data and process them in order to categorize a new observation. This set of data can be bi-class (such as finding the gender of an individual, spamming or non-spamming emails) or it could also be multi-class. Voice identification, hand writing, biometric authentication, document categorization are some examples of issue categorizations. Bellow we will discuss all the machine learning categorization algorithms.

1. *Naïve Bayes – NB (Generative Learning Model)*: It is a probabilistic classifier were its main function is to count the combinations of frequencies and values within a dataset under consideration and calculation of a set of probabilities [46]. Bayes equation is the foundation of this algorithm. Using this equation, the algorithm has the ability to assume whether all the characteristics are entirely autonomous within a value series of a class variable. Specifically, Naive Bayes classifier can predict the existence of a specific feature within a class which has no relation to the existence of any other features [46]. Although these features are strictly depended on each other characteristics, all these characteristics assist independently to the probability. Naive Bayes prototype can easily be designed and can be a very important tool for immense large datasets. Furthermore, it has the ability of outperforming even higher sophisticated classification methods [47].
2. *Bayes Net - BN*: It is a neural network-based system were its main purpose is to analyze and reflect the uncertainty models. A Bayesian network can learn the casual relationships and use them for implementing incremental learning. In order to perform such a classification authors, suggest that feedback modules have to be set with the indication [46]. Afterwards output modules can be consulted and examined with the use of Bayesian network inference standard [47].
3. *Discriminative Multinomial Naive Bayes - DMNB*: It is a well-known classifier for document classification [46] where it has been evaluated to yield satisfactory performance [47]. This classifier accepts a document as input and it considers it as an aggregation of words. For every class c , $P(w|c)$ the training data is utilized in order to calculate a probability in order to observe the word w towards any given class. This can be achieved by determining the relative occurrence frequency of each word as it uses a set of training documents of that specific class. In addition, Deshwali et al. [46] indicate that a prior probability is required for this classifier, $p(c)$ which is straightforward prediction. Let’s imagine a word w which appears nwd times in a document d [46][47]. Under evaluation the probability of the class C can be expressed as follows:

$$P(c|d) = \frac{P(c) \prod_{w \in d} P(w|c)^{nwd}}{P(d)} \quad (2.2)$$

4. *Sequential Minimal Optimization - SMO*: It as a training procedure method of Support Vector Machines - SVM classification algorithm where it is comprised by a set of configurations specifically developed in order to boost the reasoning performance of vast datasets [46]. This algorithm has been created in order to ensure the coverage result even in harsh conditions. It functions by splitting an issue into a small portion set of sub-problem. This can be solved with the use of analytic method [47].
5. *HyperPipes*: It as a technique which generates a “hyperpipe” for each dataset's class. Typically, these classes are collected information build around in a specific template object [46]. Furthermore, authors suggest that HyperPipes work tremendously fast and appear to be very efficient [47].
6. *Random Forest*: It as an algorithm for classification process which creates a set of many trees. This can be achieved by classifying a new element belonging to an input vector [46]. After that we can set the vector against the forest on each of the trees. Every tree generates a classification. In few words, this specific class is voted by the tree and the classification which has the most votes is picked by the random forest method across all of the trees. Furthermore, this method has a very good efficiency on large datasets [47].

At this point, it is useful to present a comparison of the above-mentioned classification algorithms. For the purposes of the topic of this section we will analyze studies which presents a comparison between different supervised machine learning algorithms used for sentiment analysis. These machine learning algorithms are Naïve Bayes, Support Vector Machine, Maximum Entropy and Random Forest. Desai et al. [48] mention that the above-mentioned supervised machine learning algorithms provide average accuracy in majority of domains as well as with different types of data. Also, these algorithms provide average speed of classification process regardless the size of the input data while handling the outliers [48]. Based on this extended survey, Desai et al. [48] identified several parameters like the comprehension of complexity, empirical training acceleration, small proportion of assertions efficiency and classification type in order to came to the conclusion which algorithms are accurate or efficient in their overall performance. Figure 2.8 presents a comparison between those machine learning algorithms.

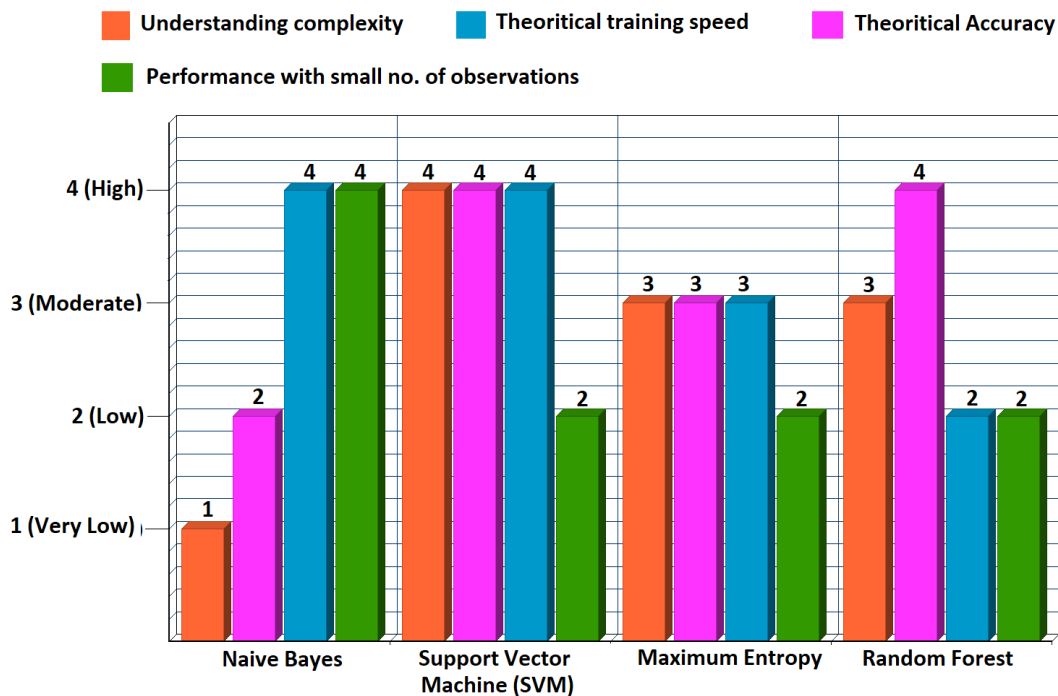


Figure 2.7: Comparison between different machine learning algorithms [48]

The above-mentioned supervised machine learning algorithms make use of a classifier model as a main mechanism in order to predict a result with a given set of data (e.g. a set of tweets). Figure 2.7 depicts the above-mentioned machine learning algorithms and type of classifier that make use.

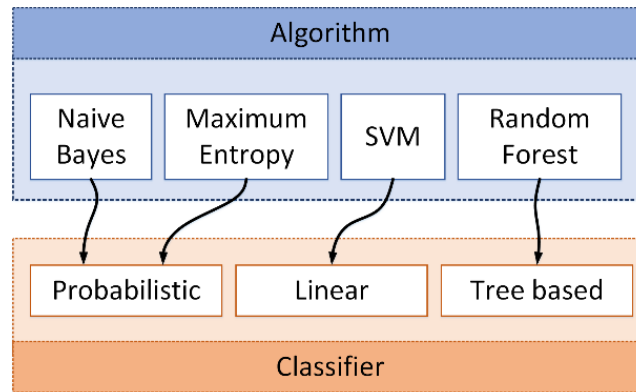


Figure 2.8: Machine learning algorithms and their respective classifiers [48]

To understand this better let's suppose a set of tweets that have been gathered and sentiment analysis is performed on them in order to predict a result. Beginning with the first phase which is the process and formation of the given dataset of tweets into a specific processable structure for the machine learning algorithm. The second phase is where all the major features are mined from the formatted text using a selection method of features. The third phase is where the part of data is labeled manually with tags such as negative/positive Tweets in order to strengthen the training set. The final phase is where the extracted features are provided as input for the built classifier in order to organize the remaining data i.e. evaluation set. There are three achievable ways in order to gather Tweets as a set of data for analyzing political sentiment. These are data repositories like UCI, Friendster, Kdnuggets, and SNAP. Twitter's API offers two types, the search API and the stream API [49]. When we want to collect data on Twitter, we use search API on the basis of hashtags and when we want to obtain real-time data from Twitter, we use steam API. Social media CRM tools are additionally classified into superior tools like Radian6^[1], Sysomos^[2], Simplify360^[3], Lithium^[4] as well as non-superior tools like Keyhole^[4], Topsy^[5], Tagboard^[6] and SocialMention^[7].

When data are gathered and assembled their form is unprocessed and may contain non useful characters. For this reason, authors suggest that it is necessary to implement a classifier in order to perform proper formation on Tweets. With this way we can remove any non-useful characters. Desai et al. [48] refers to this method as cleaning pre-process of the raw data. This procedure includes uniform casing, removing hashtags or any other Twitter characters like (@. RT), emoticons, URLs, stop words, decompression of slang words (like g8 to Group of Eight) and compression of elongated words (like happyyyy to happy).

When it comes to feature extraction methods Desai et al. [48] suggest that there are several ways where we can extract distinctive elements like adjectives, verbs and nouns. These aspects are categorized as negative/positive in order to identify the polarity of the entire phrase. Specifically, there are three methods that Desai et al. [48] use as methods of feature extraction. The first is the frequency of terms and their presence where it is used in order to imply individual as well as distinct words followed with the number of their appearance in a text. The second method is the negative phrases/words where they can change the context of the extracted opinion. Therefore, it is notable to consider the use of negative words/phrases.

Footnote

^[1] <https://socialstudio.radian6.com/>, ^[2] <https://sysomos.com/>, ^[3] <https://simplify360.com/>,
^[4] <https://khoros.com/platform>, ^[5] <https://keyhole.co/>, ^[6] <http://topsy.thisisthebrigade.com/>,
^[6] <https://tagboard.com/>, ^[7] <https://www.social-searcher.com/social-mention/>

The third method is the parts of speech like finding nouns, adjectives, verbs etc. due to the reason that they are essential counters of opinions. For sentiment classification techniques authors suggest that two of them exist for finding the sentiment of a text namely as knowledge or lexicon-based and machine learning. The main scope of Lexicon-based techniques is to derive the opinion-based lexicons from the text and find the polarity from these lexicons. Authors mention that lexicons are a set of known and pre-compiled sentiment terms. While in machine learning techniques the main scope is the creation of the algorithm, which optimizes system efficiency, through training data. Machine Learning techniques provide a solution for sentiment classification problems in two sequential phases. The first phase is where the development and training of the model occurs using a set of labeled data. The second phase is where the classification of the non-labeled data occurs, based on the trained model. Machine learning techniques can be classified to administered and non-administered methods. According to authors administered or supervised machine learning algorithms are commonly used and provide effectiveness in a variety of domains or with various data types. Furthermore, these algorithms are very efficient and fast when it comes to classification processing.

From the above information we can make the assumption that Naïve Bayes algorithms are simple as well as easy to comprehend and create them in contrast with Support Vector Machine and Maximum Entropy. Although, due to its simple Bayesian probability assumption they fail in high accuracy. In addition, Maximum Entropy algorithms have better accuracy however the ability of feature automated learning is not supported. Random forest uses a method namely as "decision tree" which provides high precision with instant learning of features. Implementing accuracy in the above-mentioned algorithms is depended on a plethora of issues such as the chosen domain, the source and amount of data as well as the pre-processing method applied on data.

2.6 Summary

In this chapter we covered the structure of the obtained data from Twitter as well as how we can access and use its API in order to obtain vital unstructured information. Furthermore, we analyzed what is sentiment analysis, as well as its various approaches such as lexicon-based and machine learning techniques. Finally, we analyze the issues of lexicons and what they are as well as presenting and comparing a variety of supervised machine learning algorithms. In the next chapter we will describe the structure of the designed application namely as Athena Political Popularity Analysis – AthPPA tool, how it performs sentiment analysis, what data we decided to visualize and for what reason.

Chapter 3: AthPPA a tool for analyzing political popularity over Twitter

3.1 The purpose of this web application

AthPPA (which stands for Athena Political Popularity Analysis) is a tool for identifying the popularity among different political leaders as well as of the latest trends that occur in social media and especially in Twitter. The scope of AthPPA is the design of a web application capable of presenting Twitter statistics for the top three well known political leaders in Greece. With that said we will choose structured and not structured data as both of them indicate the sentiment of the user. For the purposes of our analysis three Twitter accounts have been identified which are @PrimeMinisterGR, @kmitsotakis, @atsipras and @FofiGennimata as well as the respective official political party accounts in Twitter which these politicians represent. The tweet sample is by default 200 and we analyze the number of likes, re-tweets and characters per posted tweet as well as the total number of subscribers from the above-mentioned Twitter accounts. We have also identified negative hashtags for these political parties which are currently trends on Twitter by its users. We compare those structured data and present them into graphs using Python Dash framework. The technologies used for this application will be described on section 3.2. Furthermore, we have used SpaCy module for identifying Greek language and the open source sentiment analyzer of NLP Buddy. Sentiment analyzer of NLP buddy reaches to 90% accuracy for Greek language and does text subjectivity analysis. On the next section we will analyze the methodology that was followed for the creation of this application as well as its overall structure presenting its file structure, and source code. There is also a Github repository with the source code as well as a website version of this web application. Both links are provided on the Links section at the bottom of this thesis

3.2 Technical structure and methodology of AthPPA

For the creation of this application Python 3.8 was used in accordance with Tweepy a Python module which allows the application to communicate with Twitter and fetch data from the platform. Furthermore, Python Dash is used which is a Python web framework with many capabilities and features. Figure 3.1 depicts the file structure of AthPPA web application.

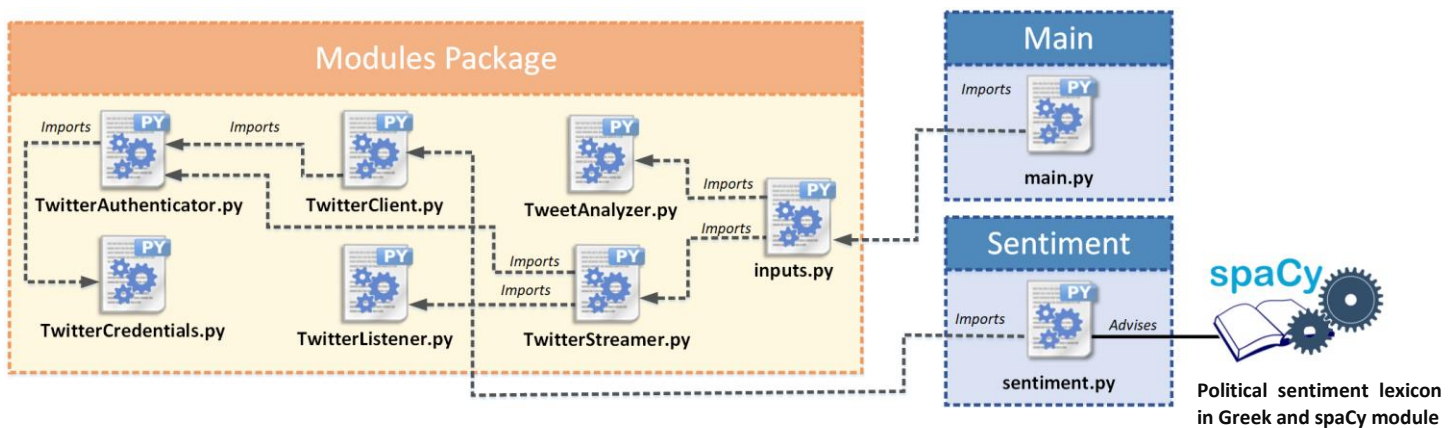


Figure 3.1: File Structure of AthPPA tool

The following table 3.1 describes each Python class on the Modules package and its functionalities.

Python Class	Description
TwitterCredentials.py	It contains the four-key set of the created Twitter's API
TwitterAuthenticator.py	It authenticates the application using the four key set and to do this it imports the twitter credentials.
TwitterClient.py	It defines what the application will "ask" from Twitter's Servers and it imports TwitterAuthenticator in order to authenticate the application against the platform
TwitterListener.py	It is a basic interface which "listens" towards the communication socket. If data are streamed properly, if data error occurred during the streaming or if TCP handshake error occurred and thus the overall connection failed.
TweetAnalyzer.py	It adds all structured data into a data frame in order to be used by the input class
TwitterStreamer.py	It does basic streaming of tweets using TwitterAuthenticator and TwitterListener.
Inputs.py	It contains the basic data which graphs use in main class of the program
Sentiment.py	This class searches for tweets of a twitter account where it calculates and prints the overall sentiment per posted tweet. To achieve this, it uses a Greek sentiment lexicon designed especially for political sentiment analysis as well as spacy-nightly module and its additional el_core_news_md corpus.
Main.py	This class contains accepts the data from input class and contains the basic structure of the webpage. Typically, it is the place where the graphs are called and visualized.

Table 3.1: Python classes of AthPPA and their additional description

The Python classes TwitterAuthenticator.py, TwitterClient.py, TwitterCredentials.py, TwitterListener.py and TwitterStreamer.py have the responsibility of guarantying the data exchange between AthPPA web application and Twitter's Servers. The Python classes Inputs.py and TweetAnalyzer.py have the responsibility of gathering the tweets and process them in order to extract useful structured information. The structured data taken from Twitter are the total number of likes, retweets and text length per posted tweet as well as the total number of subscribers per account for a set sample of 200 tweets. Furthermore, negative hashtag counter is also a functionality which these classes perform. We have identified three negative hashtags for New Democracy and SYRIZA parties as well as one for the KINAL party. The Sentiment.py Python class it gathers a set sample of 200 tweets from the Twitter accounts of the most well-known Greek political party leaders and from these ones it extracts only the textual information of each tweet. Speaking of textual information, we refer to the entire text of each tweet which means that this tweet might contain special characters such as emojis and other unnecessary elements such as exclamation marks. For this reason, AthPPA implements a text format parsing of each obtained tweet using regex where it removes all these unwanted characters. The reason for doing this is that emojis and special characters such exclamation marks and punctuations do not indicate a value for sentiment analysis and SpaCy module will simply do not process the entire text of the obtained tweet properly in order to extract its overall sentiment. When the tweet is processed and formatted properly it is stored on a data frame in

order to be analyzed by SpaCy module and the implemented sentiment analyzer. Note that the implementation for sentiment analyzer was inspired by an additional open source project namely as NLPBuddy which uses Greek SpaCy module. A few changes have been made through the source code of this sentiment analyzer in order to analyze the set of Tweets stored on the data array. A loop checks all the elements of the data frame and each sentence is processed by the sentiment analyzer in order to extract a proper value. Table 3.2 depicts the sentiment values that the sentiment analyzer assigns on tweets based on their mood expression.

Sentiment Value	Emotion Score Type
3	Happiness
2	Surprise
1	Sadness
0	Neutral
-1	Fear
-2	Disgust
-3	Anger

Table 3.2: Sentiment values used for the labelling process by the sentiment analyzer used in AthPPA

The sentiment analyzer analyzes the sentence and assigns on it the proper emotion (as presented on the table 3.2) based on the indicative expression of the sentence. Neutral value has been added to the sentiment analyzer and as neutral it counts every sentence which has not any indicative expressed opinion. The following example presents a sentence in Greek language with the produced output from the sentiment analyzer.

Sentence:

Έχω μείνει έκπληκτος! Πώς γίνεται αυτό; Η έκπληξη είναι τόσο μεγάλη! Α, τώρα εξηγούνται όλα.

Produced Output:

Subjectivity: 16.6%

Main emotion: Surprise.

Emotion score: 33.3%

To achieve this, it advises a lexicon which contains a set of labelled words. SpaCy is used for processing the text using a variety of NLP techniques. Figure 3.2 presents the structure by which SpaCy module works where a text is given as input and it produces a doc object. This doc object can be used in order to apply several Natural Language Processing techniques that SpaCy provides.

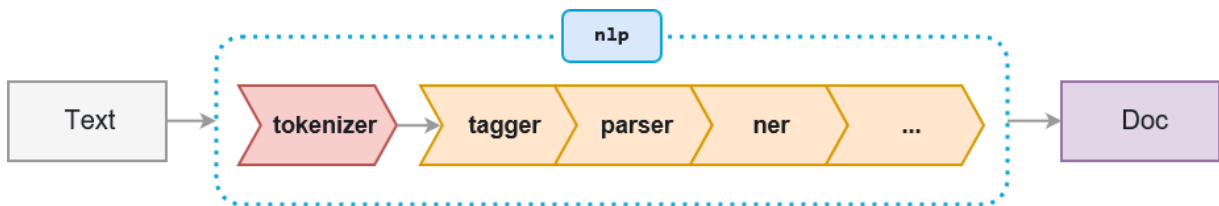


Figure 3.2: How SpaCy module produces a Natural Language Processing linguistic object

Specifically, when a text is given to spaCy it tokenizes the text in order to produce a Natural Language Processing linguistic object namely as doc. Afterwards, the doc can be processed with several steps which are called as pipeline processing. This pipeline processing includes a tagger, parser and an entity recognizer. Each pipeline component returns the processed result from the doc which can also be forwarded onto the next components.

Important note

SpaCy is comprised by two main data structures which are the Doc and the Vocab. The Doc object is constructed by a tokenizer. It can be modified by a pipeline and owns tokens and annotations. The vocab object owns lookup tables for information common to all documents. With this way we have centralized strings, lexical attributes and word vectors. This eliminates the need for storing multiple copies of this data which saves memory and ensures that there is a single source of truth [52]. Text annotations have been designed in order to allow a single source of truth where the Doc object owns the data and Span and Token are views that point into it. The doc object is designed by the tokenizer and can be modified with the use of the pipeline components. The Language object defines the coordination of these components where a raw text is given as input through the pipeline and the produced output is an annotated document. It also sets the procedures of training and serialization [53].

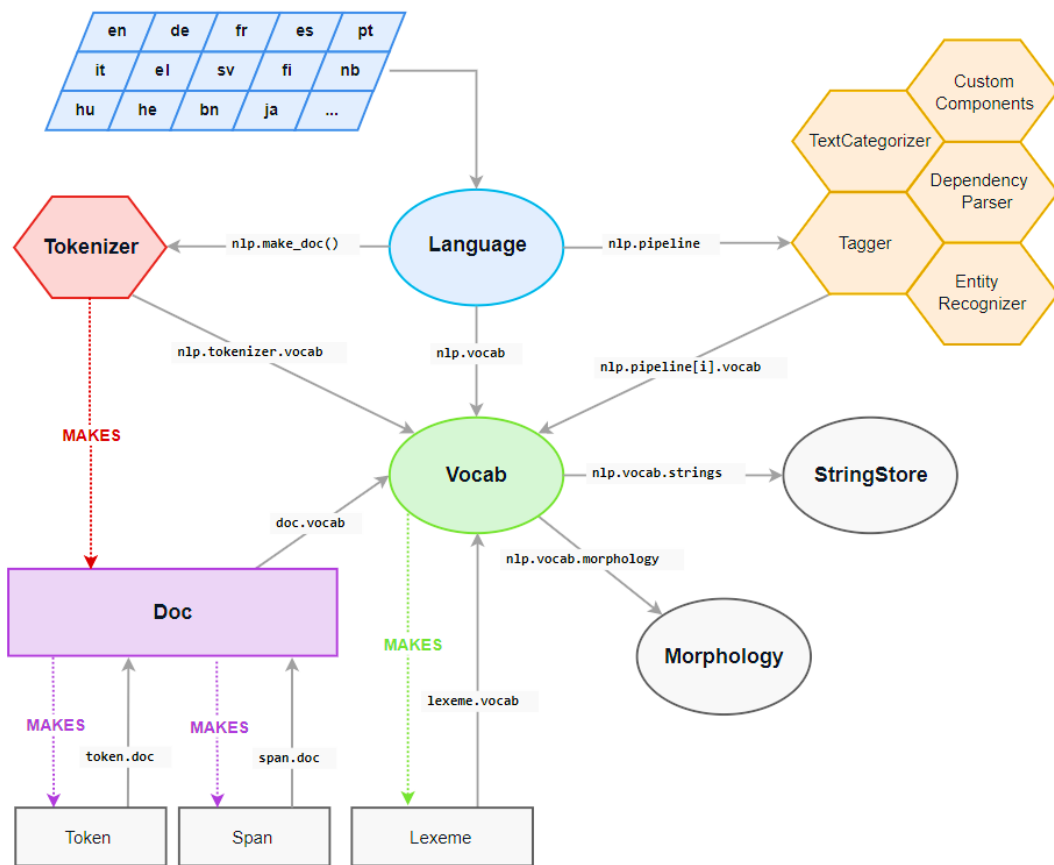


Figure 3.3: Architecture of spaCy module

The following code examples present the basic Natural Language Processing capabilities of spaCy module. Beginning with the Linguistic annotations which provide insights into a text and present its grammatical structure. This process includes the identification of parts of speech and related words within the text.

Example

```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")
for token in doc:
    print(token.text, token.pos_, token.dep_)
```

Result

```
Apple PROPN nsubj
is AUX aux
looking VERB ROOT
at ADP prep
buying VERB pcomp
U.K. PROPN compound
startup NOUN dobj
for ADP prep
$ SYM quantmod
1 NUM compound
billion NUM pobj
```

Code example 3.1: Identifying linguistic annotations with spaCy module

The first step when a text is being processed by spaCy is to tokenize it. With the term tokenization we refer to the process word segmentation from the sentence. In more simple terms words are identified from the text and are separated independently.

Example

```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")
for token in doc:
    print(token.text, token.pos_, token.dep_)
```

Result

```
Apple
is
```

```
looking
at
buying
U.K.
startup
for
$
1
billion
```

Code example 3.2: Tokenizing a text into words with spaCy module

The text processing begins from left to the right and on every substring, it performs two checks. The first check is if the substring matches to the exception rule of the tokenizer. Let's imagine that we have the following sentence and we want to provide as input to the Doc.

Sentence: "I don't want to move on the U.K."

Highlight two words "don't" and "U.K.". On the first highlight word there is not any whitespace although it is not one word and can be separated into two words the "do" and "n't". On the second highlight word there is no need to be separated into two words and should remain one single word. These words are expressed as tokens. The second check is the removal of prefixes and suffixes for e.g. commas, hyphens, quotes or periods. There are other functionalities that spaCy provides such as named entity detection but for the sake of this thesis we will skip them.

In AthPPA's sentiment script identifier, regex expressions have been deployed in order to remove punctuations, hashtags, emoticons or any other non-processing character from the obtained tweets. Once the text is processed the analyzer advises a Greek lexicon designed specifically for political sentiment analysis in order to locate words within the text [54]. The overall sentiment is extracted from the average of sentiment words. For example, if there is a sentence with 5 words where the 3 of them indicate Anger then the overall sentiment of the sentence will be anger by a 70%. Emotion lexicons and word matching are a well-known form of sources for creating knowledge when it comes for sentiment analysis over social networking. While their efficiency has been analyzed and proven on the field of sentiment analysis, and a plethora of them has been created for the English language there is not availability for the Greek language. In addition to that, Spatiotis et al. [42] suggest that the most common way of research approach for sentiment analysis is by implementing resource dictionaries which include a set of terms as well as expressed opinions which indicate positive or negative sentiment commonly known as opinion words. Tsakalidis et al. [52] in their study are trying to resolve this issue by implementing a vast collection of such tools, spanning from manually annotated lexicon to semi-supervised word matching vectors and annotated data sets for various activities. It is worth to mention that all of them can be accessed publicly. Tsakalidis et al. [54] mention that experiments have been performed using a variety of algorithms and parameters on their implemented resources to present promising results over standard baselines. Specifically, Tsakalidis et al. [54] achieved a 24.9% relative improvement in F-score on the cross-domain sentiment analysis task when training the same algorithms with their resources, compared to training them on more traditional feature sources, such as n-grams. Furthermore, Tsakalidis et al. [54] mention that their resources were built with the primary focus on the cross-domain sentiment analysis task but they also show promising results in related tasks, such as emotion analysis and sarcasm detection.

3.3 AthPPA graphs results

On the following chapters we will present the graphs and discuss their results. The last legislative elections in Greece was held on 7th of July 2019 where the New Democracy Centre-right conservative party of Kyriakos Mitsotakis won with 158 from the overall 300 seats of the Greek parliament leading to an outright majority. Figure 3.4 depicts the election results from the last Greek legislative elections

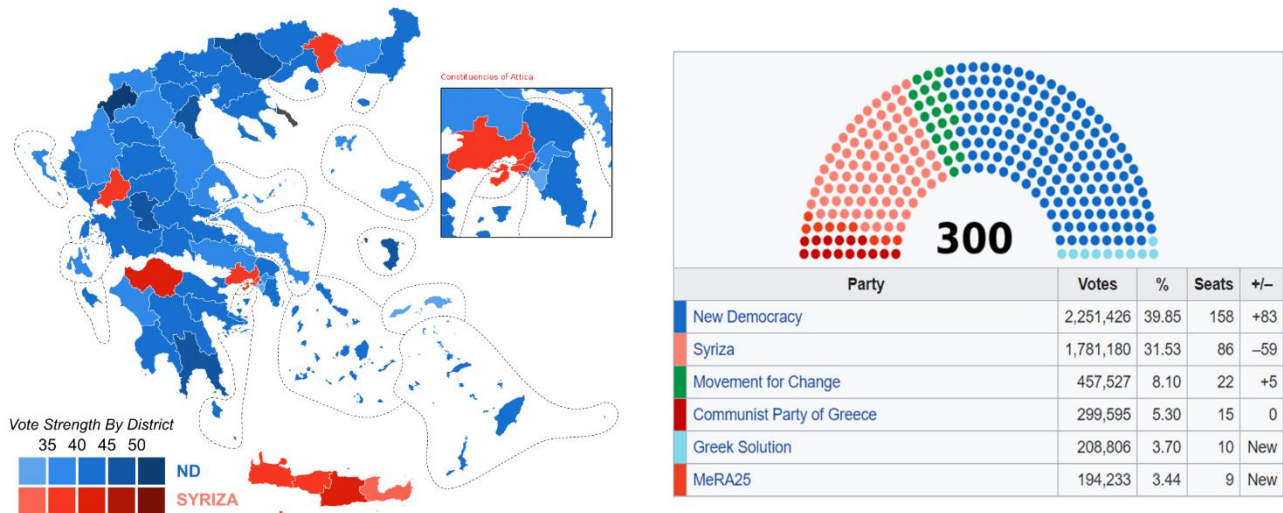


Figure 3.4: Results of the last Greek legislative election, showing the vote strength of the party winning a plurality in each electoral district.

Figure 3.4 presents the Greek parliament seat ratio among the different political parties. SYRIZA (Coalition of the Radical Left) won 86 seats making it the official opposition party in Greece, Movement for Change (or KINAL) which is a Centre-democratic socialist party won 22 seats and afterwards follow the rest political parties which have a small share of seats in the Greek parliament (34 seats in total). For the purposes of this research the three most prominent political parties and their representing leaders have been included. These ones are the current Greek Prime Minister Kyriakos Mitsotakis which is leading the New Democracy party, afterwards is Alexis Tsipras which is leading the Coalition for the Radical Left making it the official opposition party and finally Fofi Gennimata which is leading the Movement of Change party. Table 3.3 depicts the identified Twitter accounts which the data are obtained.

Twitter Account	Type	Person/Entity	Representation
@PrimeministerGR	Politician	Kyriakos Mitsotakis	ND (Majority)
@kmitsotakis	Politician	Kyriakos Mitsotakis	ND (Majority)
@neademokratia	Political Party	New Democracy	ND (Majority)
@atsipras	Politician	Alexis Tsipras	SYRIZA (2 nd Opposition)
@syriza_gr	Political Party	SYRIZA	SYRIZA (2 nd Opposition)
@FofiGennimata	Politician	Fofi Gennimata	KINAL (3 rd Opposition)
@kinimallagis	Political Party	KINAL	KINAL (3 rd Opposition)

Person: Entity:

Table 3.3: Sentiment values used for the labelling process by the sentiment analyzer used in AthPPA

From those accounts we obtain for each one of them a dynamic sample of 200 posted tweets using Tweepy Python library. The data have been visualized are the number of likes, re-tweets and text character count for each tweet of the obtained sample of 200 tweets. We have also identified negative hashtags for the two prominent political parties that Twitter’s users tend to use in their posted tweets. Table 3.4 depicts those identified hashtags.

Identified hashtag	Hashatag relation	Tweet sample	Data visualized
#ΝΔ_θελατε	Negative for ND	100	Date posted frequency
#ΝΔ_ξεφτίλες	Negative for ND	100	Date posted frequency
#ΝΔ_ρομπες	Negative for ND	100	Date posted frequency
#ΣΥΡΙΖΑ_ξεφτίλες	Negative for SYRIZA	100	Date posted frequency
#συριζωα	Negative for SYRIZA	100	Date posted frequency
#Συριζα_απατεώνες	Negative for SYRIZA	100	Date posted frequency

Table 3.4: Identified negative hashtags per political party

At this stage we can present the graphs and discuss the visualized results. Beginning from the number of likes and re-tweets for each political leader. The following figures depict the amount of likes and retweets that each tweet has from the obtained sample of 200 tweets taken from the accounts mentioned in table 3.3

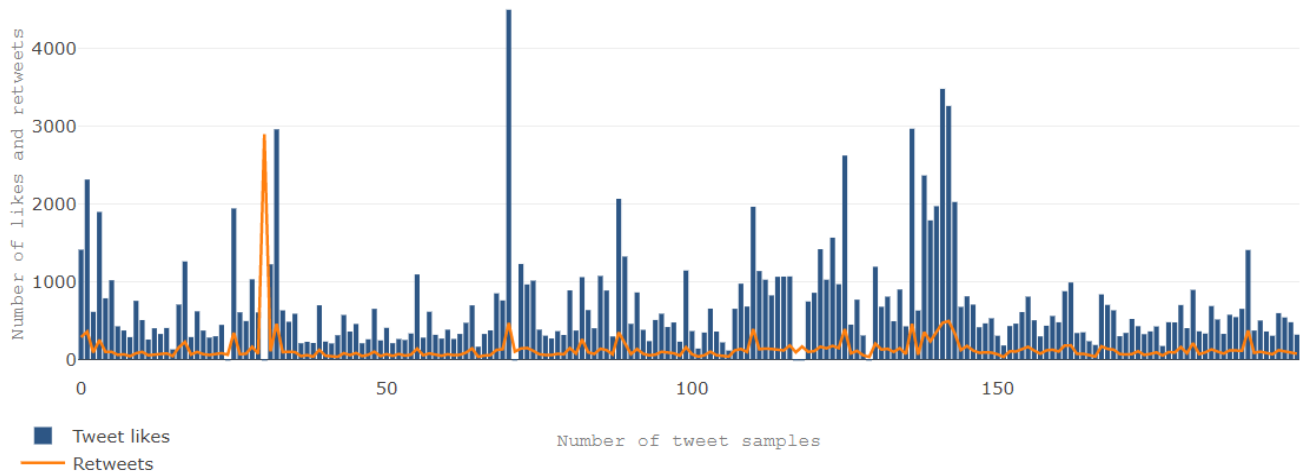


Figure 3.5: Users likes and retweets per posted tweet mined from @kmitsotakis account (200 tweet sample)

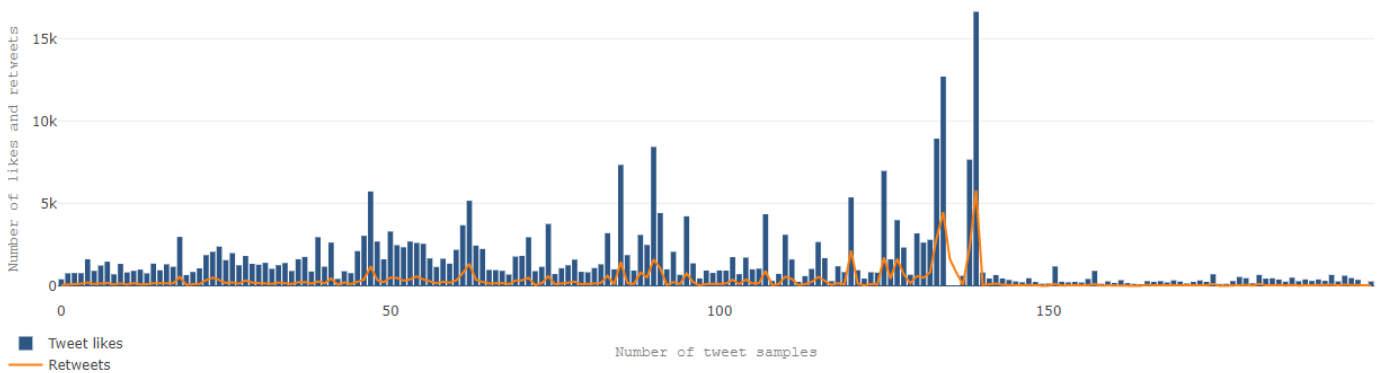


Figure 3.6: Users likes and retweets per posted tweet mined from @PrimeministerGR account (200 tweet sample)

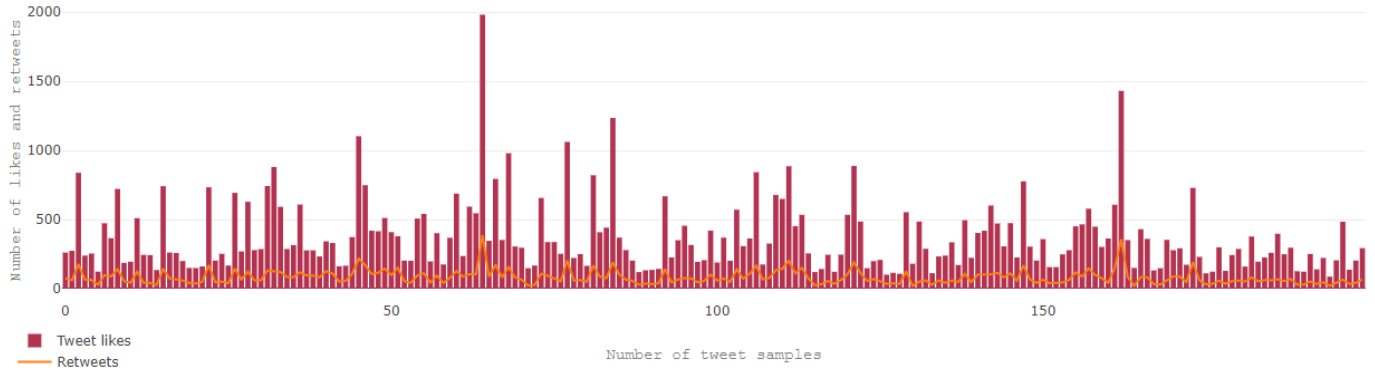


Figure 3.7: Users likes and retweets per posted tweet mined from @atsipras account (200 tweet sample)

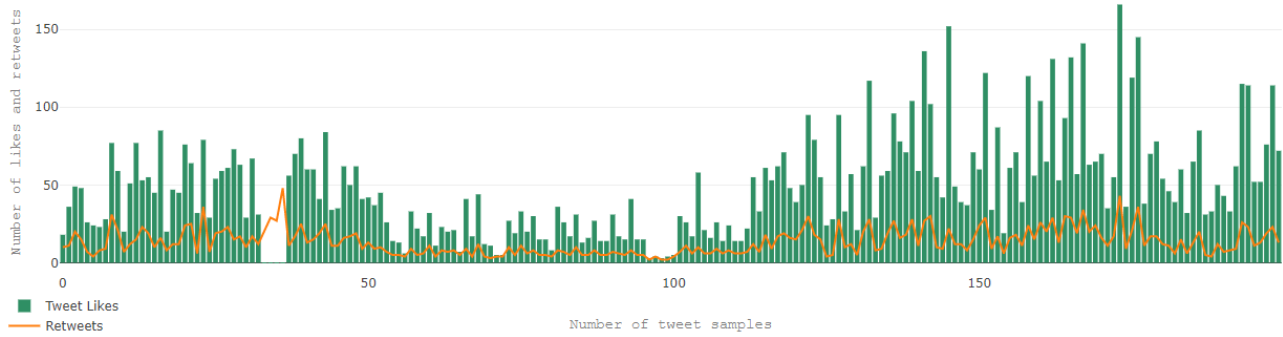


Figure 3.8: Users likes and retweets per posted tweet mined from @FofiGennimata account (200 tweet sample)

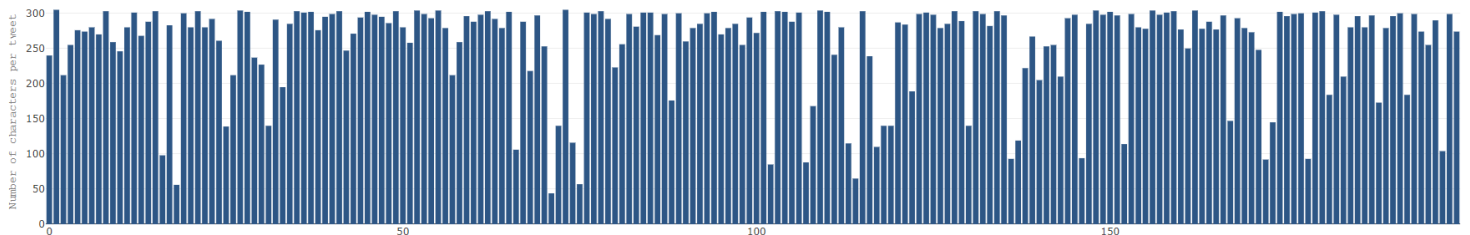


Figure 3.9: Text length per posted tweet mined from @kmitsotakis account (200 tweet sample)

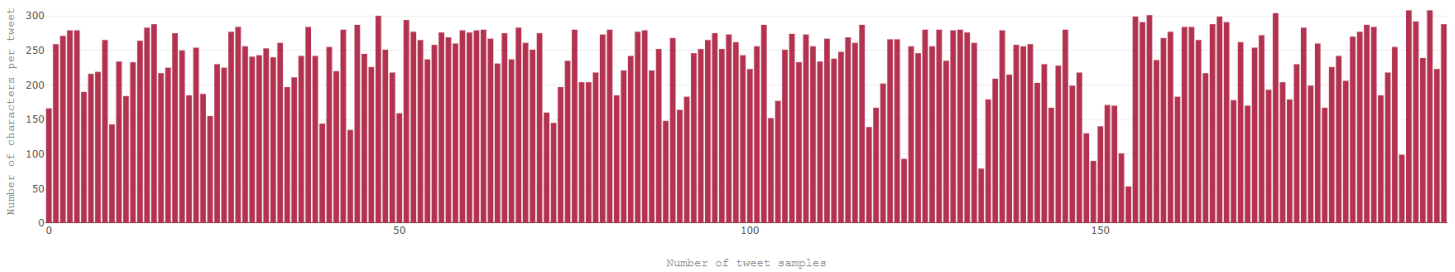


Figure 3.10: Text length per posted tweet mined from @atsipras account (200 tweet sample)

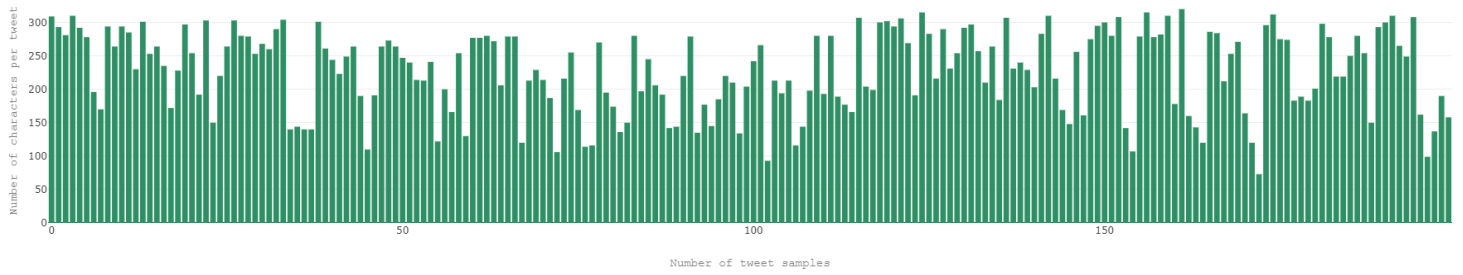


Figure 3.11: Text length per posted tweet mined from @FofiGennimata account (200 tweet sample)

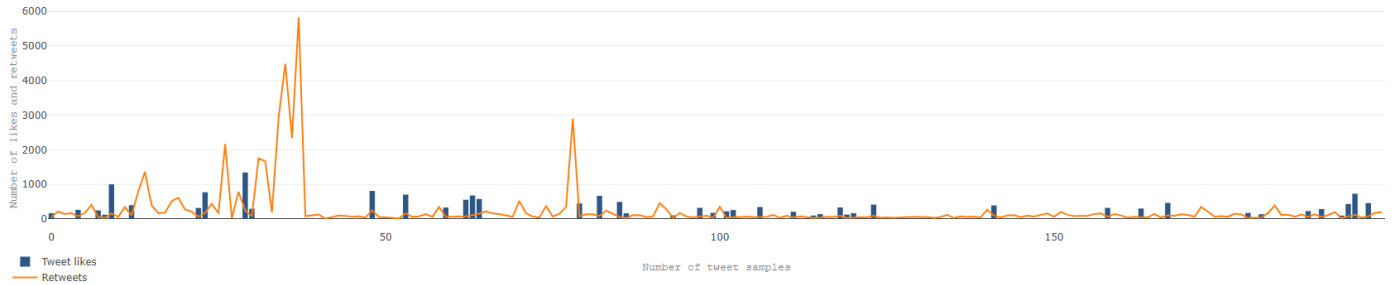


Figure 3.12: Users likes and retweets per posted tweet mined from @neademokratia account (200 tweet sample)

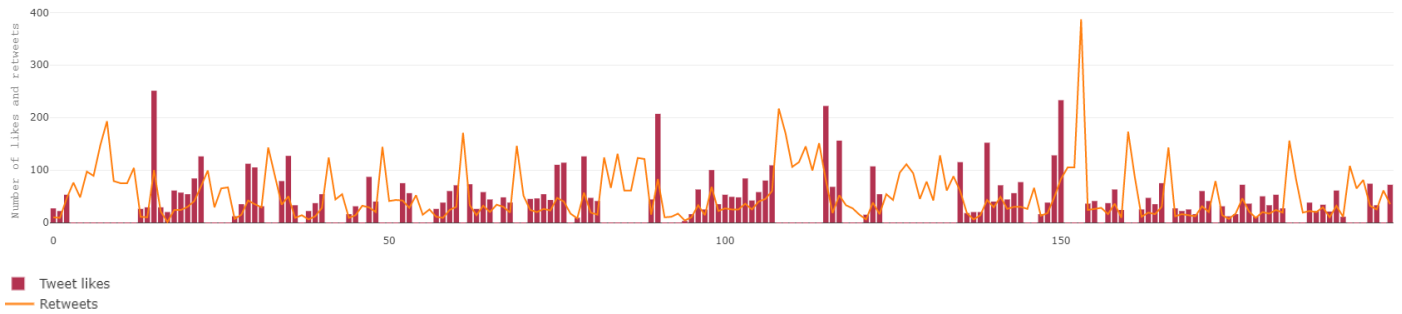


Figure 3.13: Users likes and retweets per posted tweet mined from @syriza_gr account (200 tweet sample)

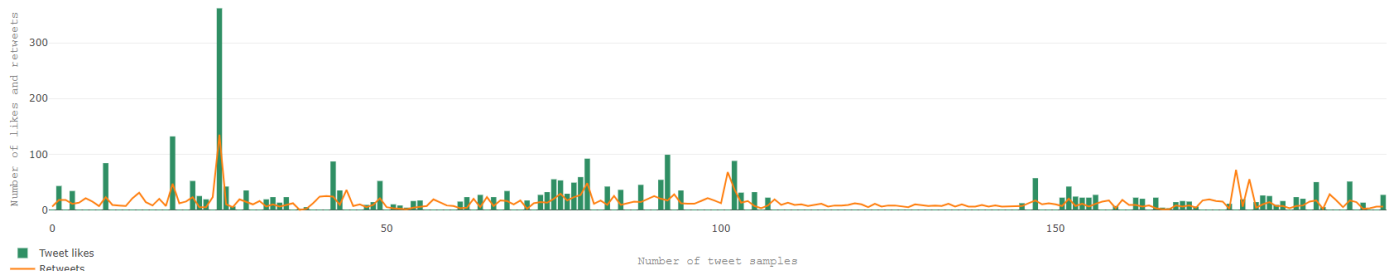


Figure 3.14: Users likes and retweets per posted tweet mined from @kinimallagis account (200 tweet sample)

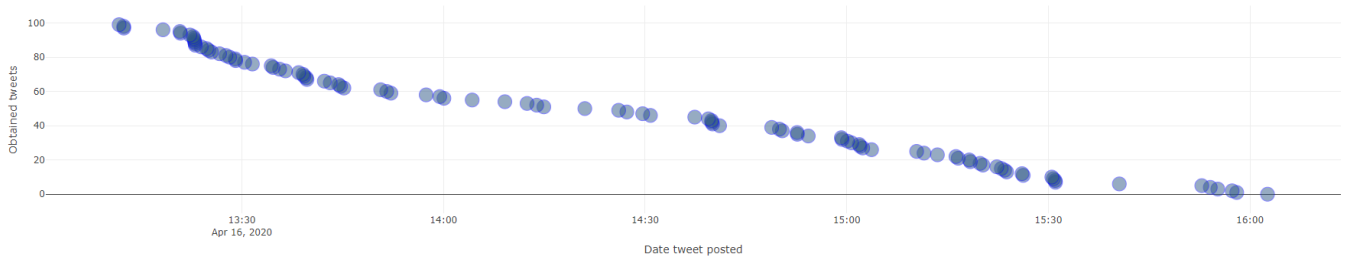


Figure 3.15: Mined tweets which include negative hashtag (#ΝΔ_Θελατε) for New Democracy party

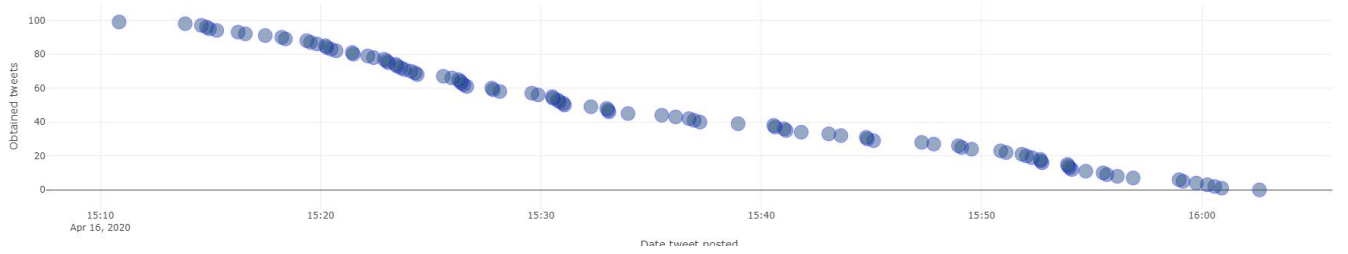


Figure 3.16: Mined tweets which include negative hashtag (#ΝΔ_ξεφτίλες) for New Democracy party

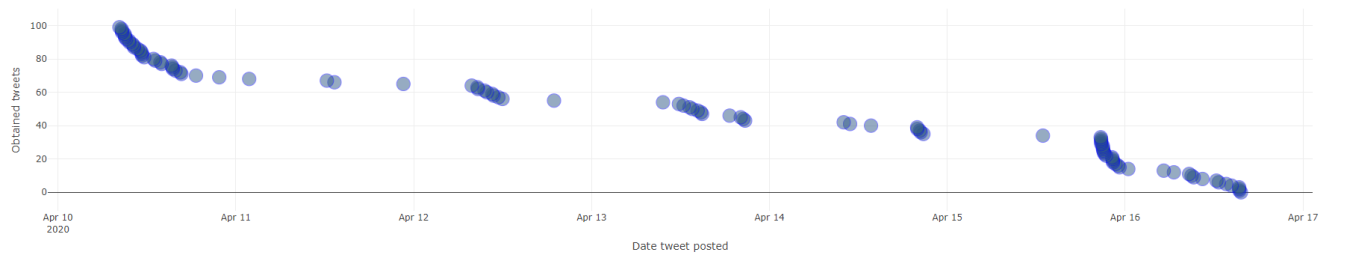


Figure 3.17: Mined tweets which include negative hashtag (#ΝΔ_ρομπες) for New Democracy party

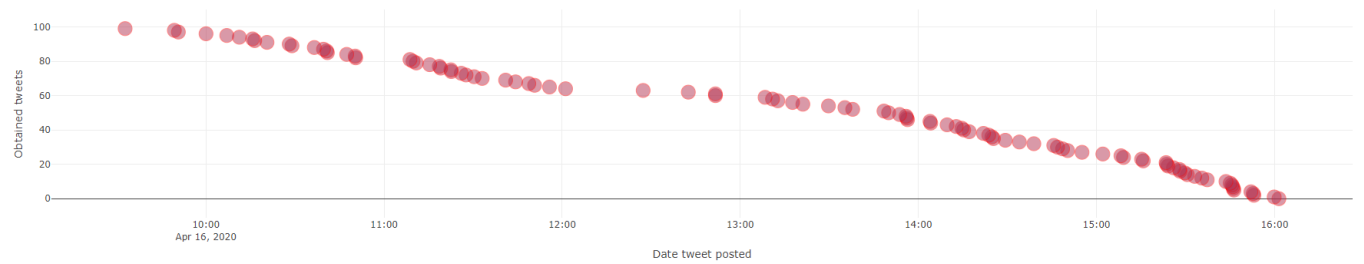


Figure 3.18: Mined tweets which include negative hashtag (#ΣΥΡΙΖΑ_ξεφτίλες) for SYRIZA party

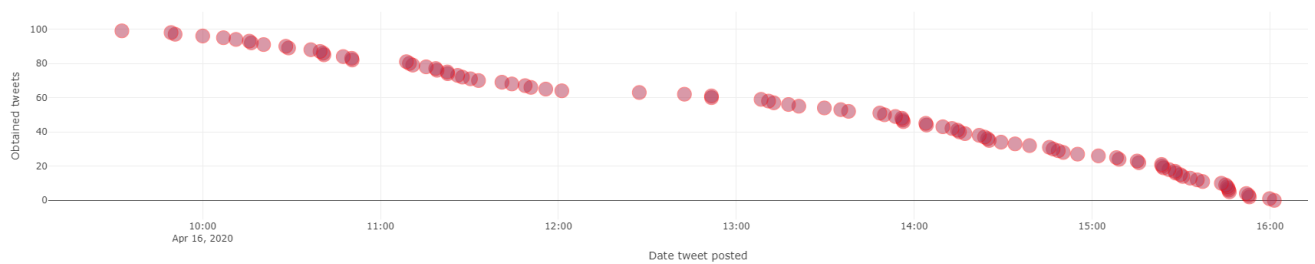


Figure 3.19: Mined tweets which include negative hashtag (#συριζωα) for SYRIZA party

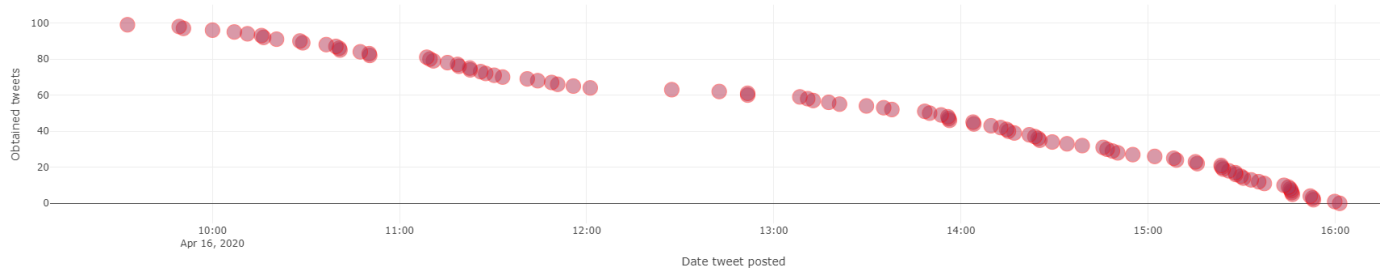


Figure 3.20: Mined tweets which include negative hashtag (#Συριζα_απατεωνες) for SYRIZA party

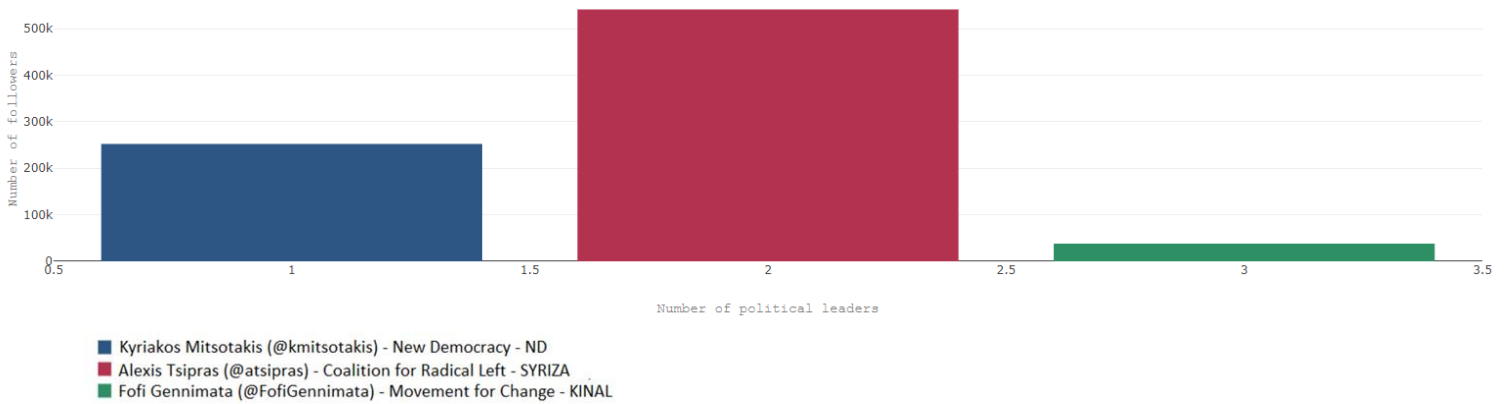


Figure 3.21: Number of registered subscribers per twitter account for the top three Greek political leaders (actual numbers)

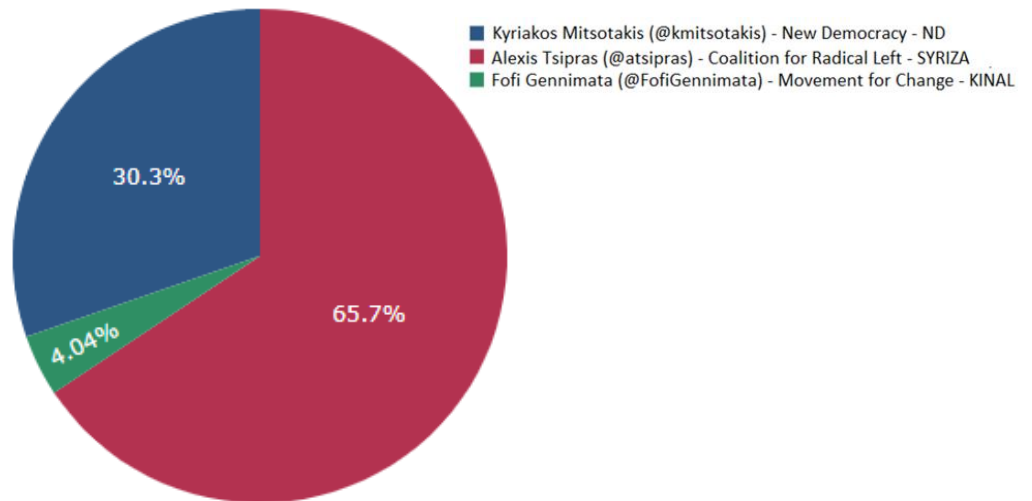


Figure 3.22: Number of registered subscribers per twitter account for the top three Greek political leaders (in percentage)

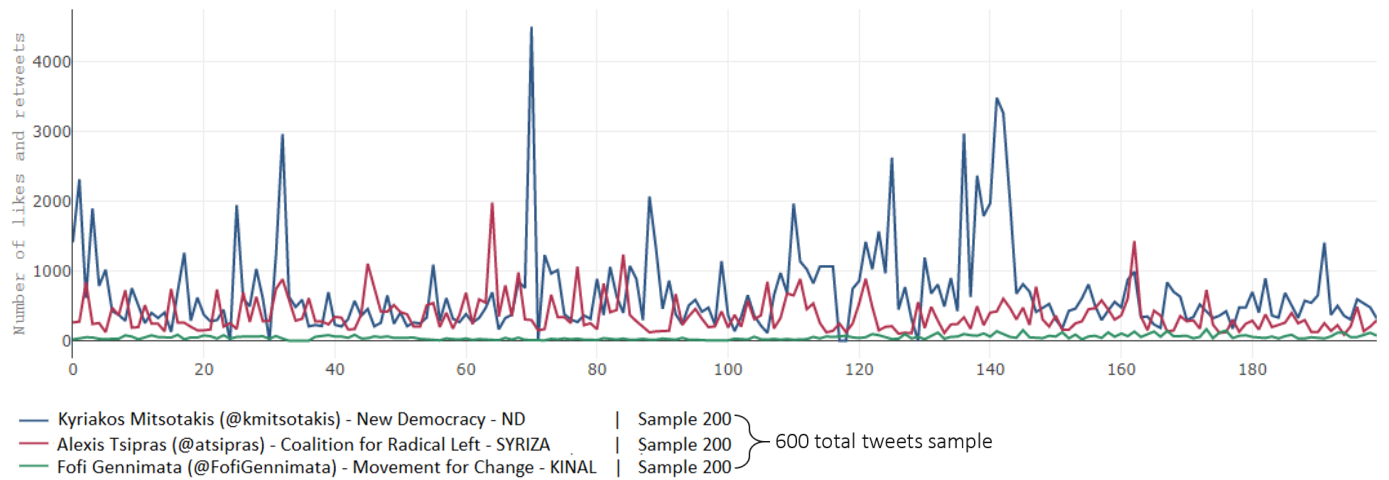


Figure 3.23: Comparison of users likes per posted tweet for the top three Greek political leaders (600 tweet sample)

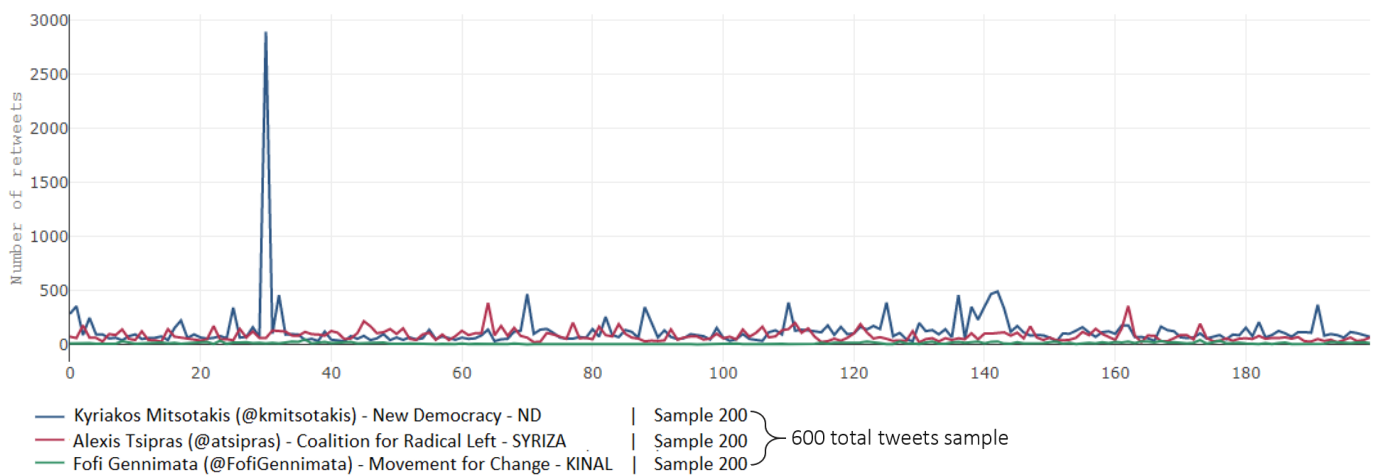


Figure 3.24: Comparison of re-tweets per posted tweet for the top three Greek political leaders (600 tweet sample)

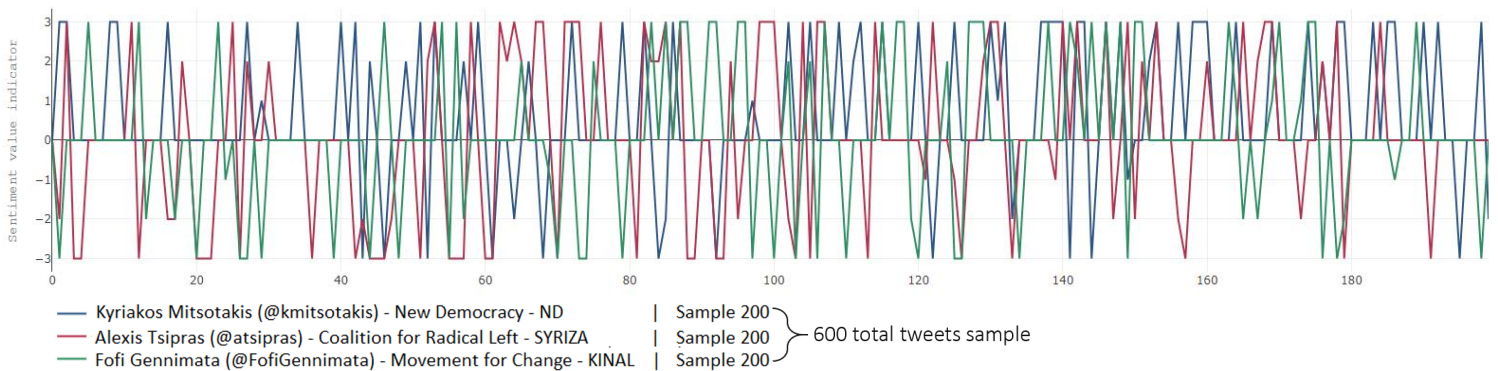


Figure 3.25: Total comparison of Sentiment tweet for the top three Greek political leaders (600 tweet sample)



Figure 3.26: Sentiment analysis per political leader (600 tweet sample)

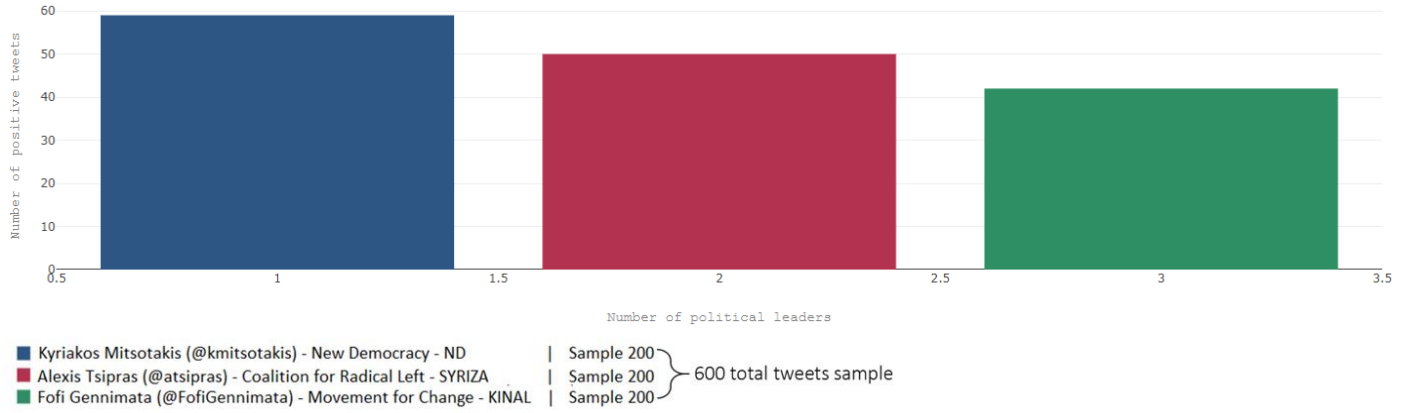


Figure 3.27: Comparison of identified positive tweets (600 tweet sample)

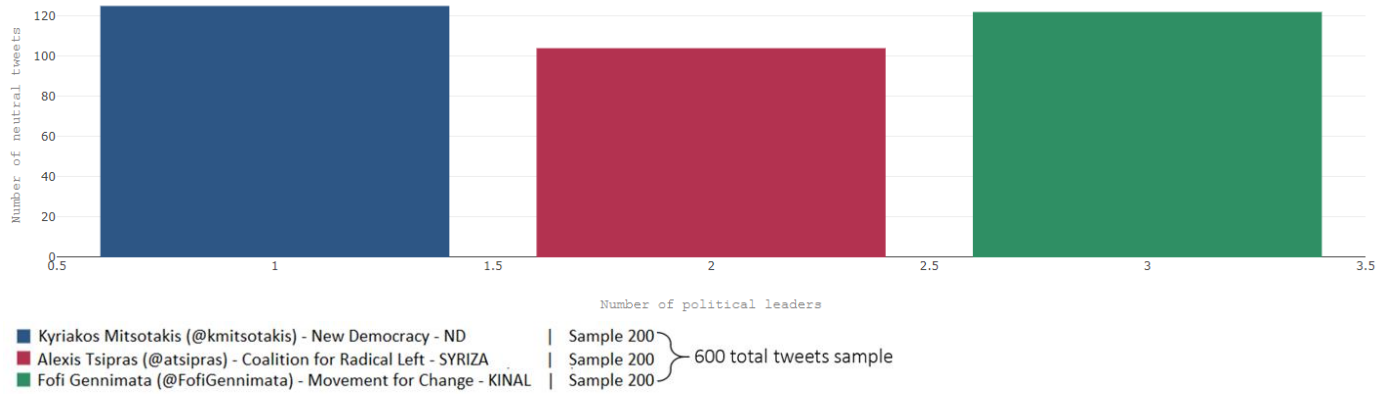


Figure 3.28: Comparison of identified neutral tweets (600 tweet sample)

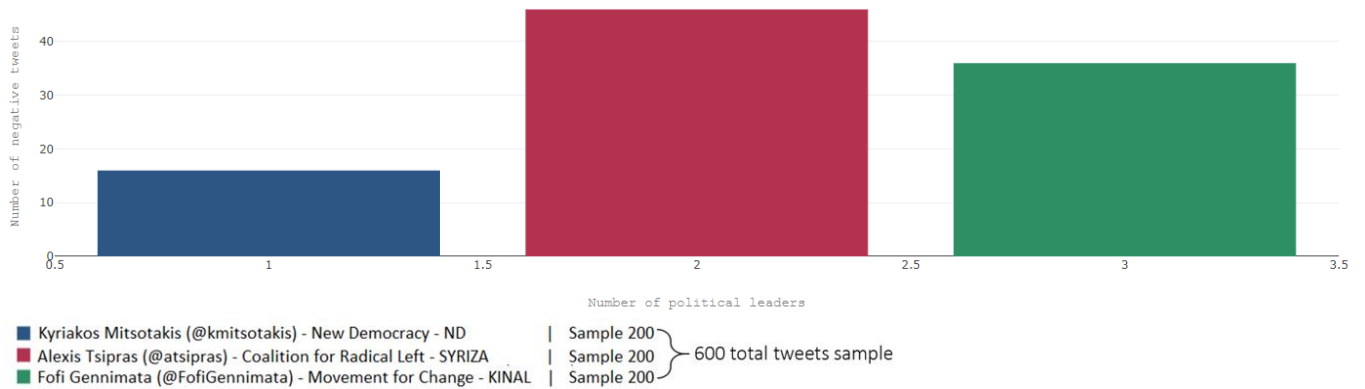


Figure 3.29: Comparison of identified negative tweets (600 tweet sample)

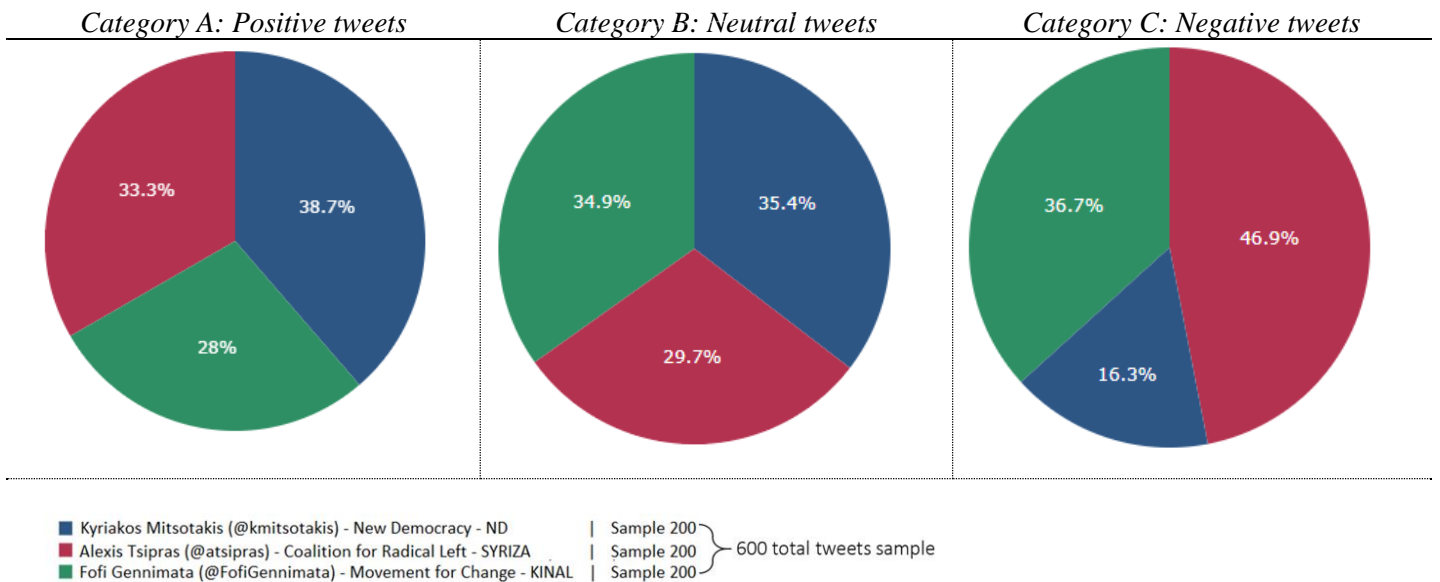


Figure 3.30: Comparison positive, negative and neutral tweets per political leader (600 tweet sample)

3.4 Summary

We presented a web-based data visualization application for political tendency identification of Twitter’s users. Twitter is a useful tool to extract the sentiment of users and to predict a political result. We have identified crucial structured data such as the number of likes, re-tweets, text length, number of subscribers per account as well as the frequency of negative hashtags that users include in their posted tweets. The number of likes and retweets allows us to observe how popular is a political leader and how many active users a politician has on Twitter. This can help us of identifying the fan base group of a politician which is a strong indicator of political popularity. The same occurs for political parties that have their official accounts in Twitter where likes and re-tweets per posted tweet are an indicator of a fan base pool, voters or supporters. In addition to that, political parties that have official accounts in Twitter are the first source of advertising a political leader which represents it. For example, on the Greek political standards if a political leader let’s say leader A posts something on his/her account it is likely that this post will be re-tweeted by the official Twitter account of the political party that leader A represents. Also, the number of subscribers that political leaders have on their accounts is another crucial indicator of political popularity as it indicates the interest of users to stay tuned to the posts and news that a political leader tweeted. On the other hand, some of these users might be fake accounts but that would certainly be a small number, although AthPPA presents unstructured data as well such as sentiment analysis. This includes also the negative hashtags as users might use them as trend rather than a true source of opinion indicator. For this reason, we are focusing mostly on how frequently a tweet with negative hashtag about a political party is posted on the Twitter rather than the emotion that these tweets imply. Finalizing with the structured data that we've obtained from Twitter, the text length of a posted tweet is an indicator to what extend a politician uses Twitter for example small texts are an indicator for expressing an announcement or something very crucial that wants to reach onto many people and that will ensure that many will read it while long texts indicate an effort to express an opinion or to affect the thoughts of a certain group of people e.g. invocation of emotion etc. although they have high risk of being unnoticed by users as they might be bored of reading it. That’s why the character limitation that Twitter provides is a crucial restriction in order to make sure that the message of the tweet will reach to everyone (or at least everyone will read it). As we've mentioned previously, structured data are a fairly good source of obtaining the popularity of a politician or a political party but there are also fake accounts that might lead us onto false results. That’s why we are also analyzing the text of the tweet

that a politician posts on Twitter in order to identify its expressed emotion or its impact that this tweet might have on users. SpaCy is an efficient Natural Language Processing tool that helps us to deploy Natural Language Processing techniques on a text and it is also an efficient commercial tool with strong community with a good documentation. Also, the lexicon made by Tsakalidis et al. [54] is a good option on or case as we want to extract the emotion that a text implies rather than simply identifying how negative or positive that text is based on a simple scale of positive or negative words. Emotion based lexicons are efficient when it comes on political sentiment analysis as the emotions are the main indicators that affect the opinion of a voter (e.g. anger about a tweet that a politician tweeted etc.). Furthermore, Python as programming language is an outstanding tool when it comes on data visualization as Dash framework provides many capabilities and features for creating graphs and charts on a web-based environment.

To conclude this dissertation presented a web-based data visualization tool for political sentiment analysis, similar applications have been created with machine learning or Natural Language processing techniques although not many of them are web-based tools that people can observe and usually exclude structured data. The research question of this thesis is to depict how our daily activities through social media have impact over political landscape especially in countries with political crisis like Greece. Furthermore, we can conclude that social media analytics can be proven useful for analyzing political landscape of countries.

References

- [1] Garfinkel, Simson L., and Rachel H. Grunspan. *The Computer Book: From the Abacus to Artificial Intelligence, 250 Milestones in the History of Computer Science*. Sterling Swift Pub Co, 2018.
- [2] Ellison, Carl. "Cryptography Timeline." (2001) [<http://world.std.com/~cme/html/timeline.html>].
- [3] Kelly, Thomas. "The myth of the skytale." *Cryptologia* 22.3 (1998): 244-260.
- [4] Sharkey, N., and A. Sharkey. "Electro-mechanical robots before the computer." *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 223.1 (2009): 235-241.
- [5] Al-Kadit, Ibrahim A. "Origins of cryptology: The Arab contributions." *Cryptologia* 16.2 (1992): 97-126.
- [6] Kahn, David. *The Codebreakers: The comprehensive history of secret communication from ancient times to the internet*. Simon and Schuster, 1996.
- [7] Gray, Jonathan. "'Let us Calculate!': Leibniz, Lull, and the Computational Imagination." (2016).
- [8] Kulstad, Mark, and Laurence Carlin. "Leibniz's Philosophy of Mind." *The Stanford Encyclopedia of Philosophy*. Last modified November 11, 20.
- [9] Williams, John B. *The electronics revolution: inventing the future*. Springer, 2017.
- [10] Campbell-Kelly, Martin. "Data communications at the national physical laboratory (1965-1975)." *Annals of the History of Computing* 9.3/4 (1987): 221-247.
- [11] Pelkey, James. "A History of Computer Communications 1968–1988." (2014).
- [12] Katie Hafner, and Matthew Lyon. *Where wizards stay up late: The origins of the Internet*. Simon and Schuster, 1998.
- [13] David Roessner, et al. "The Role of NSF Support of Engineering in Enabling Technological Innovation." *Final Report to the National Science Foundation*. SRI International, Washington, DC (1997).
- [14] Davies, Donald W. "An historical study of the beginnings of packet switching." *The Computer Journal* 44.3 (2001): 152-162.
- [15] Hjorth, Larissa, and Sam Hinton. *Understanding social media*. SAGE Publications Limited, 2019.
- [16] Wall, Aaron. "History of search engines: From 1945 to Google today." *Search engine history* (2015): 2006-2017.
- [17] Anthony JG Hey and Gyuri Pápay. *The computing universe: a journey through a revolution*. Cambridge University Press, 2014.
- [18] Darcy DiNucci, "Fragmented future." (1999): 32-33.
- [19] José Van Dijck. *The culture of connectivity: A critical history of social media*. Oxford University Press, 2013.
- [20] José Van Dijck, Thomas Poell, and Martijn De Waal. *The platform society: Public values in a connective world*. Oxford University Press, 2018.
- [21] Sponder, Marshall, and Gohar F. Khan. *Digital analytics for marketing*. Routledge, 2017.
- [22] Berners-Lee, Tim, and Mark Fischetti. "Weaving the Web. HarperSanFrancisco. chapter 12." (1999).
- [23] Ravi Gupta and Hugh Brooks. *Using social media for global security*. John Wiley & Sons, 2013.
- [24] *The Rise of Social Media*, Esteban Ortiz-Ospina - <https://ourworldindata.org/rise-of-social-media>
- [25] Chatterjee, Siddhartha, and Michal Krystianczuk. *Python Social Media Analytics*. Packt Publishing Ltd, 2017.
- [26] Matthew Ganis and Avinash Kohirkar. *Social media analytics: Techniques and insights for extracting business value out of social media*. IBM Press, 2015.
- [27] Stikeleather, Jim. "The Three Elements of Successful Data Visualizations." *Harvard Business Review* 19 (2013).
- [28] *Measured, Simply*. "The complete guide to Twitter Analytics: How to analyze the metrics that matter." (2014).
- [29] Shah, Chirag. *A Hands-On Introduction to Data Science*. Cambridge University Press, 2020.
- [30] Kumar, Shamanth, Fred Morstatter, and Huan Liu. *Twitter data analytics*. New York: Springer, 2014.
- [31] Roesslein, Joshua. "tweepy Documentation." [Online] <http://tweepy.readthedocs.io/en/v3.5> (2009).
- [32] Pozzi, Federico Alberto, et al. *Sentiment analysis in social networks*. Morgan Kaufmann, 2016.
- [33] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5.1 (2012): 1-167.

- [34] Liu, Yang, et al. "ARSA: a sentiment-aware model for predicting sales performance using blogs." *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007.
- [35] O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." *Fourth international AAAI conference on weblogs and social media*. 2010.
- [36] Tumasjan, Andranik, et al. "Predicting elections with twitter: What 140 characters reveal about political sentiment." *Fourth international AAAI conference on weblogs and social media*. 2010.
- [37] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." *international Journal of Engineering and Technology* 7.6 (2016): 1-7.
- [38] Jindal, Nitin, and Bing Liu. "Identifying comparative sentences in text documents." *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2006.
- [39] Lazarsfeld, Paul F., and Robert K. Merton. "Friendship as a social process: A substantive and methodological analysis." *Freedom and control in modern society* 18.1 (1954): 18-66.
- [40] Ferguson, Niall. "The False Prophecy of Hyperconnection: How to Survive the Networked Age." *Foreign Aff.* 96 (2017): 68.
- [41] Edward, P. J., and J. C. Robert. "Classical rhetoric for the modern student." (1971): 40.
- [42] Spatiotis, Nikolaos, et al. "Sentiment Analysis for the Greek Language." *Proceedings of the 20th Pan-Hellenic Conference on Informatics*. 2016.
- [43] Qazi, Atika, et al. "A systematic literature review on opinion types and sentiment analysis techniques." *Internet Research* (2017).
- [44] Stuart, Russell, and Norvig Peter. "Artificial intelligence: a modern approach." (2003).
- [45] Müller, Andreas C., and Sarah Guido. *Introduction to machine learning with Python: a guide for data scientists*. "O'Reilly Media, Inc.", 2016.
- [46] Deshwal, Ajay, and Sudhir Kumar Sharma. "Twitter sentiment analysis using various classification algorithms." *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2016.
- [47] Witten, Ian H., Eibe Frank, and Mark A. Hall. "Practical machine learning tools and techniques." *Morgan Kaufmann* (2005): 578.
- [48] Desai, Mitali, and Mayuri A. Mehta. "Techniques for sentiment analysis of Twitter data: A comprehensive survey." *2016 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2016.
- [49] O'Reilly, Tim, and Sarah Milstein. *The twitter book*. "O'Reilly Media, Inc.", 2011.
- [50] Gokulakrishnan, Balakrishnan, et al. "Opinion mining and sentiment analysis on a twitter data stream." *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*. IEEE, 2012.
- [51] "Eellak/Nlpbuddy". Github, 2020, <https://github.com/eellak/nlpbuddy>.
- [52] SpaCy 101: Everything You Need to Know · SpaCy Usage Documentation. "SpaCy 101: Everything You Need to Know, spacy.io/usage/spacy-101
- [53] Architecture · SpaCy API Documentation. "Architecture, spacy.io/api.
- [54] Tsakalidis, Adam, et al. "Building and evaluating resources for sentiment analysis in the Greek language." *Language resources and evaluation* 52.4 (2018): 1021-1044.
- [55] Ribeiro, Filipe N., et al. "Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods." *EPJ Data Science* 5.1 (2016): 1-29.

Useful Links

- [1] Website for AthPPA: <https://athppa.cs.hmu.gr/>
- [2] Github link for AthPPA: <https://github.com/CodeBrakes/AthPPA>
- [3] Greek SpaCy: <https://github.com/eellak/gsoc2018-spacy>
- [4] Greek sentiment lexicon: <https://github.com/MKLab-ITI/greek-sentiment-lexicon>