**NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS**

**SCHOOL OF SCIENCE**
**DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS**

**GRADUATE PROGRAM**
**"COMPUTER SCIENCE"**

**MSc Thesis**

# Sentiment Analysis on Twitter Data and Social Trends: The Case of Greek General Elections

**Georgios I. Trachanas**

**Supervisor:** **Christina Alexandris,** Professor

**ATHENS**

**APRIL 2023**

**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**
**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ**
**“ΠΛΗΡΟΦΟΡΙΚΗ”**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

# Ανάλυση συναισθήματος σε δεδομένα του Twitter και κοινωνικές τάσεις: Η περίπτωση των ελληνικών εκλογών

**Γεώργιος Ι. Τραχανάς**

**Επιβλέπουσα:** **Χριστίνα Αλεξανδρή,** Καθηγήτρια

**ΑΘΗΝΑ**

**ΑΠΡΙΛΙΟΣ 2023**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

# Ανάλυση συναισθήματος σε δεδομένα του Twitter και κοινωνικές τάσεις: Η περίπτωση των ελληνικών εκλογών

**Γεώργιος Ι. Τραχανάς**

**Α.Μ.:** 7115112100025

**ΕΠΙΒΛΕΠΟΥΣΑ:** **Χριστίνα Αλεξανδρή,** Καθηγήτρια

**ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ**: **Μανόλης Κουμπαράκης,** Καθηγητής
**Ιωάννης Παναγάκης,** Αναπληρωτής Καθηγητής

Απρίλιος 2023

**MSc Thesis**

# Sentiment Analysis on Twitter Data and Social Trends: The Case of the Greek General Elections

**Georgios I. Trachanas**

**S.N.:** 7115112100025

**SUPERVISOR:** **Christina Alexandris,** Professor

**EXAMINATION COMMITTEE:**
**Manolis Koubarakis,** Professor
**Ioannis Panagakis,** Associate Professor

April 2023

## ΠΕΡΙΛΗΨΗ

Η ανάλυση συναισθήματος και εξόρυξη γνώμης (Sentiment Analysis-Opinion Mining) είναι η διαδικασία χρήσης επεξεργασίας φυσικής γλώσσας και διαφόρων τεχνικών (μηχανική μάθηση, λεξικά) για τον εντοπισμό και την εξαγωγή υποκειμενικών πληροφοριών από δεδομένα κειμένου. Χρησιμοποιείται συνήθως για τον προσδιορισμό του συνολικού συναισθήματος ενός κειμένου, όπως αν είναι θετικό, αρνητικό ή ουδέτερο.

Σκοπός της παρούσας Διπλωματικής Εργασίας είναι η ανάλυση του συναισθήματος σε δεδομένα του Twitter. Πιο συγκεκριμένα, εφαρμόστηκε μια προσέγγιση βασισμένη σε λεξικό για την ανάλυση του συναισθήματος σε κείμενο tweet που σχετίζεται με τις Βουλευτικές Εκλογές του 2019 στην Ελλάδα. Τα tweets είναι στην ελληνική γλώσσα και ταξινομούνται ως θετικά, αρνητικά και ουδέτερα με βάση το συνολικό συναίσθημα που εκφράζουν. Μέσω της ανάλυσης συναισθήματος στα σύνολα δεδομένων με τη χρήση της γλώσσας προγραμματισμού Python, εξάγουμε συμπεράσματα σχετικά με τις κοινωνικές τάσεις που αναπτύσσονται στο προεκλογικό twitter σε σχέση με τα έξι (6) πολιτικά κόμματα που εξέλεξαν βουλευτές σε αυτές τις εκλογές. Τα αποτελέσματα παρουσιάζονται με σαφείς οπτικοποιήσεις με τη χρήση του εργαλείου Tableau για πληρέστερη κατανόηση. Εκτός από την περιγραφή της υλοποίησης, παρουσιάζονται οι κυριότεροι περιορισμοί και οι προκλήσεις και δυσκολίες που προέκυψαν στην προσπάθεια επεξεργασίας της ελληνικής γλώσσας. Τέλος, επιχειρείται η  να επισήμανση ορισμένων πτυχών της ανάλυσης συναισθήματος και εξόρυξης γνώμης που χρήζουν βελτίωσης, τόσο στη προτεινόμενη εφαρμογή που παρουσιάζεται εδώ όσο και σε άλλες υπάρχουσες.



**ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ**: Επεξεργασία Φυσικής Γλώσσας, Ανάλυση Συναισθήματος
**ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ**: Εξόρυξη δεδομένων, Προσέγγιση με βάση λεξικό, Κατηγοριοποίηση κειμένου, Twitter, Γενικές εκλογές

## ABSTRACT

Sentiment analysis and Opinion Mining involve the process of using natural language processing and various techniques (machine learning, lexicons) to identify and extract subjective information from text data. Sentiment analysis and Opinion Mining are commonly used to determine the emotional tone of a piece of text, such as whether it is positive, negative, or neutral.

The purpose of the present Thesis is to analyze sentiment in Twitter data. More specifically, a lexicon-based approach has been implemented to analyze sentiment in tweet texts related to the 2019 general elections in Greece. The tweets are in the Greek language and are classified as positive, negative, and neutral based on the overall sentiment they express. Sentiment analysis implemented on the datasets using the Python programming language allows insights and conclusions about the social trends that develop in pre-election twitter in relation to the six (6) political parties that elected Members of Parliament (MPs) in the 2019 elections. The results are presented with visualizations using the Tableau tool targeting to a clear and more complete understanding.

In addition to the description of the implementation, the main challenges, limitations, and difficulties encountered in trying to process the Greek language are presented, along with aspects of the implementation that can be improved, as well as other existing issues in Sentiment analysis and Opinion Mining.



**SUBJECT AREA**: Natural Language Processing, Sentiment Analysis

**KEYWORDS**: Data Mining, Lexicon-based Approach, Text Classification, Twitter, General Elections

*Η διπλωματική εργασία αφιερώνεται στην οικογένεια μου, στους φίλους μου και στη νησιώτικη παραδοσιακή μουσική που με στήριξαν με το τρόπο τους στην υλοποίησή της.*

## ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ ιδιαίτερα τη καθηγήτρια μου, κ. Αλεξανδρή, για τη πλούσια συνεισφορά της και τις πολύτιμες προτάσεις της για την υλοποίηση της παρούσας εργασίας.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1. INTRODUCTION

Sentiment Analysis and Opinion Mining are rapidly evolving fields gaining significant attention in both academia and industry. The explosive growth of social media platforms, online reviews, and customer feedback allows an extremely large amount of data that can be analyzed to understand people's opinions and sentiments towards products, services, and events.

Sentiment analysis involves using natural language processing (NLP) techniques to extract subjective information from text data, such as emotions, attitudes, opinions, and judgments. Sentiment Analysis and Opinion Mining have become an essential tool for businesses to monitor their brand reputation, improve customer experience, and gain insights into consumer behavior. However, Sentiment Analysis and Opinion Mining have recently spread into other fields beyond the business sector, such as Journalism and Politics, Socio-Cultural Studies and Psychology, especially for governments, organizations and academia.

The present Thesis aims to explore a lexicon-based approach to sentiment analysis, and its application to several political elections. Specifically, we have collected tweets in Greek, which relate to an important political event, the Greek general elections of 2019. Six (6) different datasets have been formed which correspond to the top six (6) political parties in terms of votes in the 2019 elections. For the tweets of each dataset, after being preprocessed, we collect the set of positively and negatively charged terms. Based on these totals, we estimate the sentiment of each tweet and, thus, infer the overall sentiment received by each of the six (6) parties.

The sentiments we compute are three: positive, negative and neutral. In addition, an effort is made to manage special issues such as the negation cases so that we can correctly estimate sentiment content. Furthermore, in the overall sentiment estimation for each party, other factors such as the total number of retweets, “likes”, the quota of positive/negative/neutral tweets and the ratio of positive/negative tweets are taken into account.

These datasets are examined without distinguishing between the users who post them and by distinguishing between vocal minority and silent majority users.

Finally, we present our estimates and conclusions using visualizations. For a clear and comprehensive presentation of data and results, the use of the Tableau tool is included.

The main target of the present Thesis is to achieve a robust lexicon-based implementation approach in sentiment analysis on datasets related to the Greek political scene.

# 2. SENTIMENT ANALYSIS

## 2.1 General methodology of Sentiment Analysis in Natural Language Processing

### 2.1.1 Definition

Analysts frequently comment on the rapid growth of technology leading to a huge increase in internet usage, with daily use of various online applications such as e-commerce, social media, video and movie platforms, and blogs. These applications generate extremely large amounts of data (both structured and unstructured), which may also be characterized as “vast” amounts of data. This “vast” data, known as "Big Data," can be analyzed through various techniques and tools to provide insights in multiple scientific and commercial fields. Social media platforms have particularly contributed to the progress in the analysis and processing of Big Data, allowing millions of users to share their opinions on various topics, from major issues of national and/or international interest to everyday life problems and so-called “niche” issues. This extremely large amount of data produced by social media, is valuable for organizations, governments, and companies to understand public opinions, preferences, and reputations. As a result, the research community and academia have been working on Sentiment Analysis since the beginning of the 21st century.

**Sentiment analysis (SA)** may be defined as the computational research “field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes.” [1]

Sentiment Analysis (SA) is also known as Opinion Mining. The definitions in the literature may be considered equivalent, however some argue that they are two slightly separate subjects. Opinion Mining extracts and analyzes people’s opinion about an entity while Sentiment Analysis identifies the sentiment expressed in a text then analyzes it. Therefore, the target of SA is to find opinions, identify the sentiments they express, and then classify their polarity [2] [3].

For the scope and requirements of present Thesis, we consider both definitions to be equivalent, however, the term Sentiment Analysis (SA) is observed to be more widely used. We note that Sentiment Analysis (SA) is a broad field of research and applications and it is also considered an “umbrella” term, including the research and development of objects and applications with slightly different characteristics and requirements.

The analysis of the data to extract underlying user’s opinion and sentiment is considered to be a challenging task. For a machine, the Sentiment Analysis problem can be formulated as a 5-entity interaction: an owner (a user) adopts a View on a Topic and expresses it through Sentiment at a specific Time. To a machine, “opinion is a “quintuple”, an object made up of 5 different things:

$$(O_j, f_{jk}, SO_{ijkl}, h_i, t_l),$$

where $O_j$= the object on which the opinion is on, $f_{jk}$ = a feature of $O_j$, $SO_{ijkl}$ = the sentiment value of the opinion, $h_i$ = Opinion holder, $t_l$ = the time at which the opinion is given” [1].

### 2.1.2 General Methodology

The general methodology in the Sentiment Analysis task includes two main stages, namely Data Selection-Extraction and Data Analysis, the latter including the Data Visualization process. The steps of various sentiment analysis tasks are presented at Figure 1.

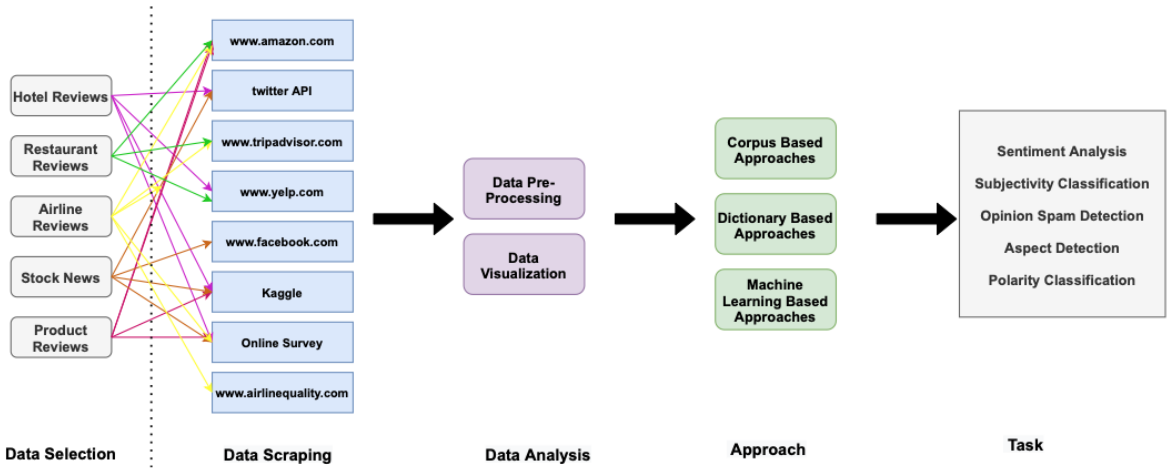



**Figure 1 Steps of various Sentiment Analysis tasks**

#### 2.1.2.1 Data Selection-Extraction

In the Data Selection-Extraction processing stage of sentiment analysis, data is selected and extracted from the existing sources. There is a wide variety of sources on the Internet, which determines the format and content of the data used in Sentiment Analysis applications. A user can collect data, which comes from text, images, audio or even video. Popular websites offer either free or paid various APIs (Application Programming Interface) for data extraction. There are many popular APIs developed either by major social media platforms (such as Facebook, Twitter, Instagram, Sina-Wiebo) or by the research community (Table 1). Through these APIs, a user can extract large datasets from posts and other information made public by millions of users and exploit them in their application. [4]

**Table 1: Social media platforms and their extracting data tools**

| API | Social Media Platform |
|---|---|
| Twitter API | Twitter |
| Twitter Streaming API | Twitter |
| Twitter4J | Twitter |
| Facebook Graph API | Facebook |
| Tancent | Sina-Wiebo |

### 2.1.2.2 Data Analysis

#### 2.1.2.2.1 Text Preprocessing

The Data Analysis stage of sentiment analysis involves the text-preprocessing task. In the text-preprocessing task, text data that is collected from various sources often requires preparation before conducting a full analysis. Data preprocessing is a technique used in data mining to transform raw data into a usable and efficient format for a machine to improve its performance.

Some popular preprocessing steps are based on [5]:

- **Lower casing:** Convert the input text into the same casing format so that ‘text’, ‘Text’ and ‘TEXT’ are treated in the same way.
- **Remove punctuation:** Remove all the punctuation symbols and characters, like !"#$%&\'()*+,-./:;<=>?@[\\]^_{|}~`, so that the strings “thesis” and “thesis!!” are treated in the same way.
- **Remove stop words:** Stop words are commonly occurring words in a language like 'the', 'a' and so on. They can be removed from the text most of the times, as they don't provide valuable information for analysis. In some cases, like Part of Speech tagging, we should not remove them as they provide valuable information about the POS.
- **Stemming:** Stemming is the act of reducing words that have been changed through inflection or derivation to their root or base form. For instance, the words "walks" and "walking" would be reduced to "walk" through stemming. However, in another example with the words "console" and "consoling," the stemmer would result in "consol," which is not a proper English word.
- **Lemmatization:** Lemmatization is like stemming in reducing inflected words to their word stem but differs in the way that it makes sure the root word (also called “lemma”) belongs to the language
- **Remove URLS, links:** URLs are useless in analysis, and therefore should be removed from data.

The preprocessing steps are executed in a sequential manner, where the outcome of one step serves as the input for the subsequent step. To clarify, here are a few preprocessing examples for the sentences below at Table 2:

- Twitter is a microblogging, social networking service owned by American company Twitter, Inc., on which users post and interact with messages known as "tweets".
- #NASA #SPACE On Dec. 8 at 11am ET (1600 UTC), Administrator @SenBillNelson and agency leaders will discuss @NASAClimate research and our role as a global leader in understanding how the planet is changing. How to watch: https://go.nasa.gov/3P92hHi

**Table 2 Preprocessing techniques in example sentences**

| Preprocessing Technique | Sentence | Text after preprocessing |
|---|---|---|
| **Lower casing** | (1) | twitter is a microblogging, social networking service |

| | | |
|---|---|---|
| | | owned by american company twitter, inc.,<br>on which users post and interact with messages known as "tweets". |
| **Remove punctuation** | (1) | twitter is a microblogging social networking service owned by american company twitter inc<br>on which users post and interact with messages known as tweets |
| **Remove Stop Words** | (1) | twitter microblogging social networking service owned american company twitter inc users post interact messages known tweets |
| **Stemming** | (1) | twitter microblog social network servic own american compani twitter inc user post interact messag known tweet |
| **Lemmatization** | (1) | twitter microblog social network servic own american compani twitter inc user post interact messag known tweet |
| **Remove URLs and links** | (2) | #NASA #SPACE On Dec. 8 at 11am ET (1600 UTC), Administrator @SenBillNelson and agency leaders will discuss @NASAClimate research and our role as a global leader in understanding how the planet is changing. How to watch: |
| **Remove mentions** | (2) | #NASA #SPACE On Dec. 8 at 11am ET (1600 UTC), Administrator and agency leaders will discuss research and our role as a global leader in understanding how the planet is changing. How to watch: |
| **Remove hashtags** | (2) | On Dec. 8 at 11am ET (1600 UTC), Administrator and agency leaders will discuss research and our role as a global leader in understanding how the |

| | | |
|---|---|---|
| | | planet is changing. How to watch: |

#### 2.1.2.2.2 Data visualization

The Data Analysis stage of sentiment analysis also involves the data visualization task. The graphical representation of information and data is defined as “Data visualization” [6]. Data visualization tools make it simple to obtain information and comprehend trends and patterns in data using what are known as "visual elements," such as charts, graphs, and maps. Data presentation by academics and professionals to non-expert receivers and a larger audience is thought to benefit from the effectiveness and clarity of data visualization tools.

The use of tasks pertaining to the structure and content of text data, such as average word length analysis, sentence length analysis, and word frequency analysis, allows for a more thorough comprehension of the text in question [6].

### 2.1.2.3 Approach selection

For the implementation of Sentiment Analysis, three main approaches are mainly chosen. These approaches are the Lexicon Based Approach, the Machine Learning Approach, and the Hybrid Machine Learning Approach [7]. Each of these approaches is divided into several sub-solutions. A diagram with all the approaches and each subcategories is shown in Figure 2.

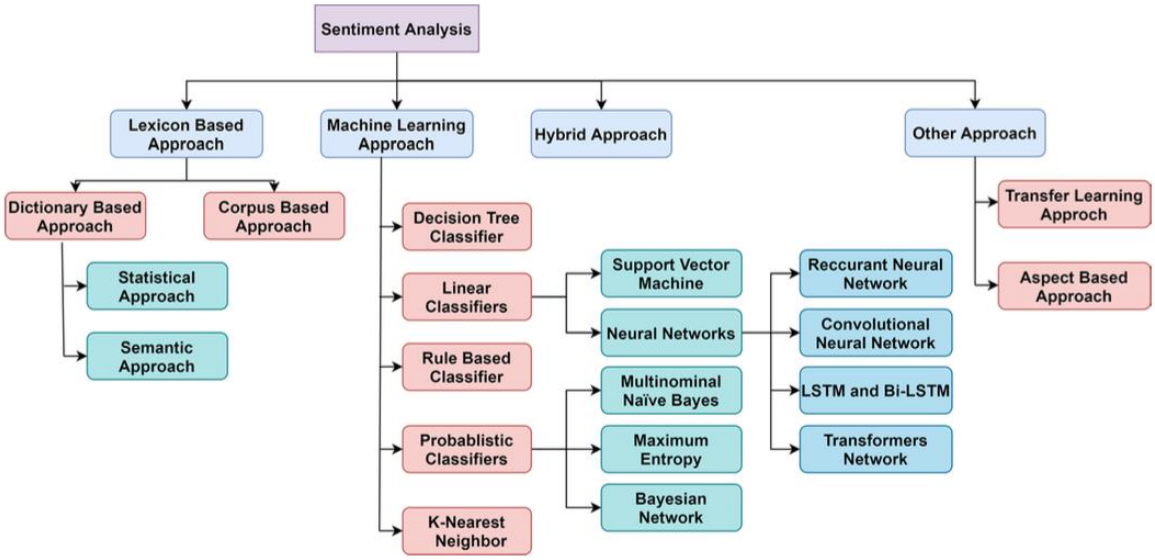



**Figure 2 Sentiment Analysis and its various approaches**

#### 2.1.2.3.1 Lexicon Based Approach

The Lexicon-Based Approach involves the use of lexica. In the case of Sentiment Analysis, lexicons are defined as a “collection of tokens where each token is assigned

with a predefined score which indicates the neutral, positive, and negative nature of the text" [7].

Specifically, in the Lexicon-Based Approach, text analysis and processing are based on the initial task of dividing the text into tokens. The second step involves the calculation of a positive, negative, or neutral score for each separate token [7].

In other words, tokens in Sentiment Analysis are linked to a positive, negative, and neutral content, namely, "polarity" [8]. Specifically, each token has a polarity score.

The most used polarity scores are the value range [-1, +1] [8], ranging from "strong positive", "positive", "neutral" to "negative" and "strong negative". Specifically, the most positive sentiments linked to a token approach the +1 (highest) score (i.e., "strong positive") while the most negative sentiments linked to a token approach the -1 (lowest) score (i.e., "strong negative"). Neutrality is indicated with a value of "0" (zero) for a token's score.

In the final stage, the overall polarity of the text is calculated based on the highest value of individual scores. This is an unsupervised technique, since pretrained data are not considered necessary. On the other hand, a necessary requirement of crucial importance is the use of a "sentiment lexicon" comprising positive and negative terms. Sentiment lexica are used with satisfactory results for Sentiment Analysis at both sentence level and document level. [7]

However, the Lexicon-Based Approach may often result to errors, due to linguistic parameters of a natural language, especially semantics and socio-cultural factors and associations. For example, a word may have a positive sentiment in a particular context and a negative one in another context. These cases and situations are considered a risk to the accuracy of models employed in the Lexicon-Based Approach. A characteristic example is the word "short" [7] and in the context-sentences "The queue at the cinema is short" and "Sam is too short for this role" [7], where "short" can be linked to a positive and a negative score respective, if it considered that short queues are preferred whereas a short height is usually undesirable for males. These types of problems can be managed or avoided by developing a domain-specific sentiment lexicon or by adapting an existing vocabulary [7].

#### 2.1.2.3.2 Machine Learning approach

Another main approach for Sentiment Analysis is the Machine Learning Approach, based on machine learning techniques and algorithms for sentiment analysis. In particular, the machine learning technique "uses syntactic and/or linguistic factors for detecting and classifying sentiment type (sentiment classification) with a classification model associating the characteristics of the text or text segment (record) with a label linked to a respective sentiment type (class label) [7]. The classification model is, subsequently, used to predict a class label for a given instance of an unknown (sentiment) class. In the Machine Learning Approach, it is considered that the sentiment analysis algorithms can be trained to read beyond simple definitions to understand information about context, sarcasm, and misapplication of words" [7].

Characteristic examples of widely used techniques used in machine learning approaches are the "Naïve Bayes" technique, the "Support Vector Machine (SVM)" technique, the "Logistic Regression" technique, the "Maximum Entropy" technique and the "AdaBoost (Adaptive Boosting)" technique.

The **Naive Bayes** technique for machine learning approaches involves the **Naive Bayes** family of probabilistic machine learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. The various versions of the algorithm differ mainly by the assumptions they make about the distribution of the input data [9].

Naive Bayes classifiers typically use bag-of-words features, an approach commonly used in text classification, resulting to their popularity as a popular statistical technique in Natural Language Processing (NLP) (for example, to mark the topic of a news article [9]) and in specialized applications such as e-mail filtering.

Naive Bayes classifiers are highly scalable, requiring several parameters linear in the number of variables (features/predictors) in a learning problem. For Naive Bayes classifiers, maximum-likelihood training can be executed by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

Naive Bayes classifiers can also be trained on and make predictions for multiple classes. In this case, these classifiers are named “Multi-class Naive Bayes” classifiers. [9]

The **Support Vector Machine (SVM)** technique for machine learning approaches concerns a type of supervised learning algorithm that can be employed for classification or regression tasks [10].

The SVM algorithms (SVMs) are considered to be particularly well suited for problems with high-dimensional feature spaces, such as text classification and image classification [10].

Regarding processing data and performing data classification, the main idea behind SVM algorithms is finding the best boundary (or "hyperplane") that separates the data into different classes. [10] This boundary is chosen in such a way that it maximizes the so-called “margin”, which is the distance between the boundary and the closest data points from each class. These data points are also known as “support vectors” [10]. It is considered that the data points that are closest to the (class) boundary are the most difficult to classify and have the most impact on the position of the boundary [10].

Apart from processing problems with high-dimensional feature spaces (i.e. text classification, image classification), SVMs can also be used to solve non-linear classification problems by applying the so-called "kernel trick", which maps the input data into a higher-dimensional space where a linear boundary can be found [10]. It is noted that the most common kernels are linear, polynomial and radial basis function (RBF). The SVM technique can also be used for regression problems, in which case, it finds the boundary that best fits the data by minimizing the margin of error [10].

**Logistic Regression** is a supervised machine learning technique used for classification problems. It is a type of generalized linear model that uses a logistic function to model a binary dependent variable [11].

The main goal of logistic regression is to find the best set of coefficients (weights) that minimize the difference between the predicted probabilities and the true class labels [11]. In the case of Sentiment Analysis, sentiment type class labels. Finding the best set of coefficients (weights) is typically performed by using an optimization algorithm such as gradient descent.

The logistic function, also called the sigmoid function, maps any input value to an output value between 0 and 1, which can be interpreted as the probability of the input data

belonging to a certain class. The logistic regression model learns a set of coefficients (also called weights) that can be used to make predictions on new input data. Given an input vector x and a set of coefficients w, the model produces a predicted probability "y_hat" of the input data belonging to the positive class. This probability can then be thresholded to produce a binary prediction (0 or 1) [11].

**Maximum entropy** is a technique in machine learning that is used for modeling probability distributions, based on the principle of maximum entropy. The principle of maximum entropy states that among all possible probability distributions that are consistent with a given set of constraints, the one with the highest entropy is the most unbiased and least informative [12].

The maximum entropy machine learning technique is often used for natural language processing tasks (NLP), such as part-of-speech tagging and text classification, where the goal is to estimate the probability of a certain output given an input [12].

The maximum entropy model is trained using a set of labeled training data. The labeled training data is used to estimate the parameters of the probability distribution that best fit the (input) data [12].

**AdaBoost (Adaptive Boosting)** is a machine learning technique that is used to improve the performance of so-called "weak learners" (i.e. models that perform only slightly better than random guessing) by combining them into a single "strong learner" [13].

It is considered that the basic idea behind AdaBoost is to repeatedly train weak learners on different subsets of the data. The subsets of the data are chosen in such a way that the examples that were misclassified by previous weak learners have a higher probability of being included. In particular, the algorithm starts by training a weak learner on the entire training data set, and then it assigns a weight to each training example based on how well it was classified by the weak learner [13]. The examples that were misclassified are given a higher weight, and the examples that were correctly classified are given a lower weight. The algorithm then trains a second weak learner on the same data set, but this time, the examples are chosen with the new weights. The second weak learner is combined with the first weak learner by giving them different weights, and the combination is used to classify the examples. The process is then repeated for several rounds, each round the examples are reweighted according to how well they were classified by the previous weak learners. The final classifier is a combination of all the weak learners [13].

#### 2.1.2.3.3 Hybrid approach

Finally, it is noted that for Sentiment Analysis, a Hybrid Approach can be employed. The Hybrid Approach combines machine learning and lexicon-based approaches. In particular, the term "Hybrid" refers to the combination of machine learning and lexicon-based techniques for Sentiment Analysis. The combination of machine learning and lexicon-based techniques results to the remarkable popularity of the Hybrid Approach in Sentiment Analysis, with sentiment lexicons playing a significant role in most systems and including both statistical and knowledge-based methods for polarity recognition [7].

### 2.2 Levels of Sentiment Analysis

Sentiment Analysis is generally considered to be performed in three (3) levels of detail: the sentence level, the document level and the aspect and entity level [1]. Sentiment

analysis at the sentence level and at the document level concern the detection of the overall sentiment in a sentence or document respectively. Sentiment analysis at the aspect and entity level concern the detection of the overall sentiment in a sentence or document considering a particular aspect or characteristic [1].

**Sentence level:** The overall sentiment of a statement is evaluated at this level. This level of analysis is closely related to subjectivity classification, which divides sentences that represent knowledge (called objective sentences) from words that express subjective ideas and opinions (called subjective sentences) [7]. For example, the line "iPhone 10 is an outstanding cell phone" communicates a favorable opinion about a single thing, iPhone 10. In another case, the line "iPhone 10 is superior than Xiaomi Redmi 10" reflects one negative and one positive attitude. The sentence level is not as precise in these circumstances as desired [7].

**Document level:** The overall sentiment of a document is evaluated at this level. The document type varies. It could be a brief news item, a tweet, an article, or a product or film review. Given a product review, for example, the system assesses if the review represents an overall good or negative assessment of the product. At this level, we assume that the document reflects a feeling or an opinion regarding a certain entity. Otherwise, the model's precision is inadequate. Because a document consists of one or more sentences, the document level is a generalization of the sentence level. Each document's average sentiment is determined by the average sentiment of each sentence [7].

**Aspect and entity level:** At this stage, a sentence or document is assessed based on a specific aspect or characteristic. The main concept is that an opinion expresses a sentiment (positive, negative, neutral) towards a particular target. Without identifying the target, an opinion is not very helpful. Recognizing the significance of opinion targets is crucial in gaining a better understanding of sentiment analysis. For instance, the opinion "our new car is quite nice, but the price was high" pertains to two targets: the car, for which a positive opinion is expressed, and the car's price, for which a negative opinion is expressed. In many cases, a target entity is described globally and in terms of its individual aspects. Aspect-level analysis is more challenging than sentence and document-level analysis as it attempts to determine specific likes and dislikes rather than a general feeling about an entity. This level comprises several sub-problems that are currently tackled in research [7].

## 2.3 Sentiment Analysis Applications

### 2.3.1 Business Analysis

Business analysis is a professional field that identifies business needs and solves business challenges. Solutions frequently contain a component of software development, but they may also include process improvements on products and services delivered, organizational change, or strategic planning and policy formulation. Large and small businesses, as well as international, public, and private institutions and organizations, can develop new marketing strategies or improve existing ones. It is feasible to measure consumer input on a product or service and, therefore, enhance it by applying the Sentiment Analysis capabilities [7].

### 2.3.2 Products Reviews

The increased acceptance of e-commerce has opened new opportunities for enterprises. Businesses can acquire useful insights into how to improve their offerings and make better purchase decisions by studying client comments on products and services using Sentiment Analysis. This can be done at the phrase or aspect level, and can focus on certain keywords or product features, such as food, service, or cleanliness, to discover and analyze only the most relevant information [7].

### 2.3.3 Market Research and Competitor Analysis

Sentiment analysis allows a corporation or institution to research the situation in a market or area of the economy in which it wishes to invest. As a result, a company may anticipate all of its competitors, their market position, and the overall level of service they deliver. By this strategy, it might build a focused marketing campaign with a goal to gaining a larger portion of the market "pie". Sentiment analysis may collect data from various platforms such as Twitter, Facebook, and blogs, give meaningful and usable results, and solve challenges in business analytics [7].

### 2.3.4 Review Analysis

Sentiment Analysis applications are widely used in the entertainment business. Various internet portals (e.g., Imdb) and applications allow the public to voice their opinion (and rate) on movies, TV series, and short films. This possibility allows the audience to select the highest quality information, while excellent and great works gain greater visibility and recognition. Sentiment Analysis at the sentence level has proven useful in this area [7].

### 2.3.5 Customers Reviews

Sentiment Analysis has numerous advantages for the tourism and restaurant industries. On the one hand, it allows a consumer to make the best possible decision based on their wants and preferences, and on the other, it provides input to the owners on areas that need to be improved. Aspect-based sentiment analysis on hotels and restaurants will assist in identifying the aspect with the most positive and negative evaluations, on which owners can work to improve [7].

### 2.3.6 Stock Market

Sentiment analysis has promising applications in the stock market, as the progress of the market generates corresponding trends that can lead to either profits or losses for investors. Correctly assessing these trends can result in accurate predictions of stock prices and increased profits, while incorrect assessments can lead to losses. To make the most appropriate forecast, all available news on the stock market can be analyzed with sentiment analysis, using data from sources such as Twitter, news articles, and blogs. By performing sentence-level sentiment analysis on these texts, the overall polarity of news related to a particular company can be determined. If the overall sentiment is negative, a price drop is predicted, while a positive sentiment indicates a price increase. Recently, sentiment analysis frameworks have been used to predict stock prices in the cryptocurrency market [7].

### 2.3.7 Social Media Monitoring

Social media is a never-ending supply of information from millions of users. Clients can voice their opinions about a person or organization, a product or a service at any time through these applications. Thus, using sentiment analysis, one may estimate the overall view that dominates social media regarding an entity in real time. If the image is not satisfactory, one has the option of improving it through various measures so that it improves in the eyes of the public [7].

## 2.4 Challenges of Sentiment Analysis

A basic characteristic of Sentiment Analysis is that it applicable to wide variety and diversity of data, each of which presents challenges. Therefore, Sentiment Analysis had to deal with a considerable set of challenges and effectively address several key difficulties to improve accuracy. Typical challenges and difficulties include, from a linguistic aspect, sarcasm, informal style of writing grammatical and syntactic errors while, from the practical aspect of implementation, computational cost and availability of data pose additional problems. Special issues such as the adaptations and heterogeneity of a natural language are also included in challenges to be dealt with by Sentiment Analysis applications.

- **Sarcasm**

In regard to **Sarcasm**, it should be noted that, here, Sentiment Analysis has to detect and process a complex linguistic phenomenon, often not perceived by non-native and even native speakers of a natural language. Sarcasm is language-specific and often individual-specific.

There are many definitions of Sarcasm. According to one definition [14]: " Sarcasm refers to the use of words that mean the opposite of what you really want to say, especially to insult someone, or to show irritation, or just to be funny. The intention of the speaker is to insult or mock someone [14]."

Furthermore, it should be noted that Sarcasm and Irony are often indistinguishable phenomena, even for humans, in everyday communication. Therefore, the detection of irony and sarcasm is an even more complex and difficult task for a machine and a challenge for NLP. Specifically, beyond the words and sentences that make up a text (e.g., a dialogue or a tweet), additional (world) knowledge concerns the attitude of the persons involved and the context in which they are exchanging views [15].

- **Informal style of writing**

If it is considered that Sentiment Analysis is mostly performed on raw text data, mostly written in an informal writing style**,** the **Informal style of writing** constitutes a typical challenge in processing and evaluating texts [7]. For example, reviews of a movie, a tweet or Facebook post, or a comment on a product are not usually fully compatible to typical syntactic and linguistic rules. Users may use acronyms, idioms, abbreviations, emojis and emoticons, abusive and insulting expressions. These features are linked to their distinct complexity factors, for example, acronyms are not universally the same, so specificity is needed for each country and for each region. Similar issues apply to linguistic phenomena such as idioms. The characteristic features of the informal writing style in text content pose additional obstacles to the effectiveness of Sentiment Analysis tools. [7].

- **Grammatical and syntactic errors**

The informal writing style confronted in most cases of raw text data is linked to a high probability of syntax and spelling errors. In other words, the less formal a text is (such as

raw text data), the greater the possibility of such errors. Problems regarding grammatical and syntactic errors can be resolved with spelling and syntax checking. However, it should be considered that spelling and syntax checking is not always an easy task nor are there tools available for all languages [7].

- **Computational cost**

In regard to the practical aspect of implementation, **Computational cost** may become a challenging issue for Sentiment Analysis. In particular, achieving the best possible level of accuracy entails the increase of training data size and, therefore, complicating the model. This results to the exponential increase of the computational cost of the model for training [7]. GPU (Graphics Processing Unit) may be required to train a model with a huge corpus. Models like SVM (Support Vector Machine), NB (Naïve Bayes) are not computationally costly, but neural networks and attention models have shown that they are computationally costly [7].

- **Availability of data**

The **Availability of data** is another challenge for Sentiment Analysis. The online availability of hundreds of different datasets does not guarantee that for every Sentiment Analysis application there is a complete and properly structured dataset [7]. Furthermore, it should be considered that not every available dataset can be used in all cases of Sentiment Analysis. Specifically, it is observed that there are not enough labeled datasets for all languages and manual labelling is a time-consuming process [7]. Additionally, it is also observed that not all sentiment classification models fit all application cases [7]. For example, a model trained on a stock market dataset cannot be used for movie reviews [7]. However, it is noted that, for the convenience of users, several popular sites provide several easy-to-use APIs (Application Programming Interfaces) for data mining and creating appropriate datasets.

As observed with most Natural Language Processing (NLP) applications, the more often a language is used, the more likely it is that datasets (including natural language resources – corpora and lexica) and tools are available for processing. Unlike highly resourced languages such as English, less-resourced languages are, therefore, most likely to pose challenges in their processing, including Sentiment Analysis. These challenges also apply to (Modern) Greek, a language considered to be in the middle of the spectrum between highly resourced and low-resourced languages. Therefore, some challenges are expected to be confronted in the present dataset, which is mostly in the (Modern) Greek language.

- **Adaptations of language**

An additional challenge in Sentiment Analysis is the heterogeneity and adaptations of a natural language, often confronted in raw text data. A language, apart from its common and uniform features, shows various heterogeneities from region to region and even within the same country. Such differences have to do with local idioms, language tradition and literacy level. For example, Greek is spoken in Greece and Cyprus, but there are differences in vocabulary and pronunciation. A Sentiment Analysis model even in one language should take these differences into account [7].

## 2.5 Overview of Present Approach

The present research attempts to implement a lexicon-based sentiment analysis method for a set of datasets. These datasets are about the Greek general elections in 2019, specifically the top six parties in terms of electoral share. These are written in Greek

language. Our article is an initial attempt to answer the research question: "Is there a link between the Twitter sentiment and the social trends exhibited in the elections? And, if so, how prevalent is this? ".

In general, the actions we have taken are as follows:

1. Using adequate preprocessing, eliminates "noise" from datasets
2. It employs a lexicon-based technique in which we gather positively and negatively charged terms for each tweet. In this manner, we compute the polarity and overall emotion of each tweet.
3. We update the dataset for each political party with the associated polarity and sentiment values.

We use additional data to analyze each tweet, such as the number of “likes” and retweets, and each dataset, such as the ratio of positive and negative tweets, in addition to sentiment analysis.

In conclusion, we find a link between the general mood expressed on Twitter regarding the election problem and societal trends and opinions about the same subject.

# 3. TWITTER AS A SOCIAL MEDIA PLATFORM

## 3.1 The Twitter Platform

The Twitter platform is one of the most popular social media platforms used worldwide by millions of users. Twitter is a popular social media platform for following the news, the area of social life people are interested in, for commenting and debating about social affairs. Microblogging via the Twitter platform facilitates the quick communication with audiences quickly, with a micro blog allowing quick, conversational connections with other people instead of writing pages of text [16], [17].

The trend of microblogging emerged at the same time with the emergence of social media. For business, microblogging allows quick engagement with customers through various formats such as audio, video, images, and text and also helps keep customers informed about longer content on websites [16], [17]. Popular platforms like Facebook, Twitter, Pinterest, Instagram, Tumblr, and Reddit, not only offer this functionality, but also generate a large amount of data [16], [17].

Regarding the definition of the Twitter platform, Twitter may be defined as:

“ a “microblogging and social network” service owned by American company Twitter Inc., on which users post and interact with messages known as "tweets". Registered users can post, like, and retweet tweets, while unregistered users only have a limited ability to read public tweets. An example tweet is shown in Figure 3. Users interact with Twitter through browser or mobile frontend software, or programmatically via its APIs” [18].

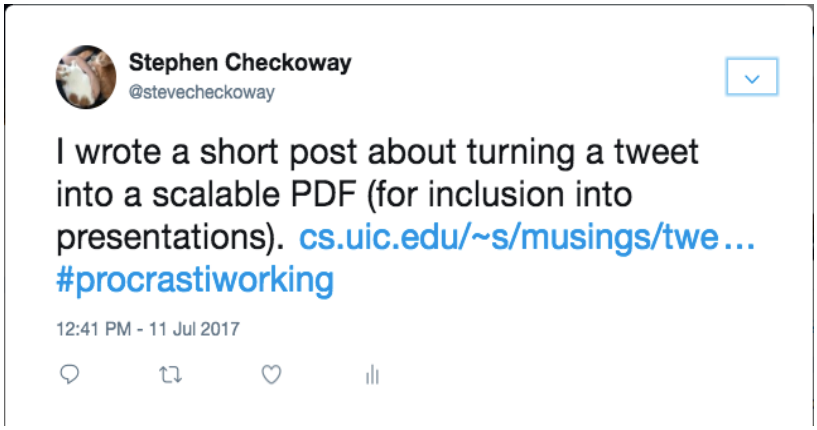



**Figure 3 Tweet's format**

Twitter messages, often known as "tweets," have a character limit of 140 characters. This format was created to be compatible with phone SMS text messaging standards. Most accounts also limit audio and video tweets to 140 seconds.

- Twitter users can use one or more hashtags to designate the topic of a tweet. A hashtag is a term that begins with the symbol '#'; for example, if a tweet includes the hashtag #Tesla, it is about the Tesla automobile firm. The quantity of tweets with the same hashtag influences Twitter's so-called “hot” trends, and users can search messages using their hashtags.

- Twitter users can address other users by using the "mention" tag followed by the "@" symbol. The 140-character limit limits the number of persons who can be tagged in a tweet. For example, if a tweet includes the word '@DonaldTrump,' it is directed at Donald Trump, and the user will be notified.
- Twitter also features a "retweet" option that allows users to republish another user's tweet on their timeline. This distributes the tweet to their followers, who may or may not follow the original user. Retweeting aids in the preservation of information by allowing users to republish anything that they think interesting or relevant that would otherwise be lost.
- Twitter users can approve of a tweet by "liking" it. The amount of “likes” and retweets that a tweet receives show its popularity.

## 3.2 Social media usage in Greece

Social media refers to online platforms and applications that enable users to create and share content or participate in social networking. Examples include Facebook, Twitter, Instagram, LinkedIn, and TikTok. These platforms allow users to connect with others, share information, and engage in a variety of online activities.

Over time, social media has gained a lot of popularity and use by millions of users. Of course, the use of each platform is not the same across the population. Differences in usage are related to country, level of Internet adoption, gender, age, social class and other social and cultural factors.

Before evaluating the data on Greek Twitter, we need to describe the overall use of social media in Greece. The more active users a particular social media platform gathers, the more popular it becomes and, therefore, the more it can gather social trends. For the use of social media in Greece we have relied on the report [19], which is an excellent presentation of the digital reality for various countries.

- There are 8.5 million Internet users in Greece (Figure 4-January 2022). This means that internet adoption is at 82.2 % of the total population. In just one year (2021-2022), during the COVID-19 pandemic, it increased by +3.5 % (about

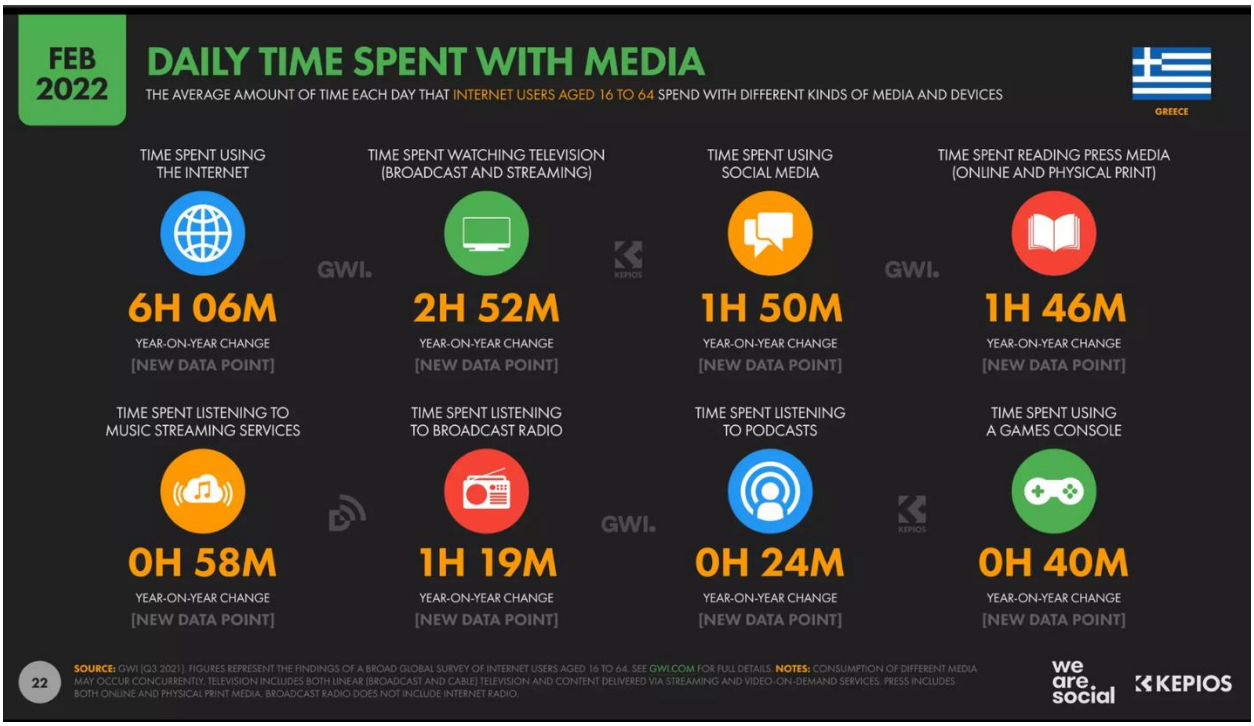



**Figure 4 Daily time spent in social media in Greece (February 2022)**

285000 new users).

- Within this population, there are 7.4 million social media users (Figure 5). This number corresponds to approximately 71.5 % of the total population of Greece.
- The average time spent on social media every day is about 1 hour and 50 minutes (Figure 4). Most platforms allow users over 13 years old. Thus, the percentage of

"eligible" users reaches 80.4 % of the population over 13 years old. In this number, men are 51 % and women 49 %.

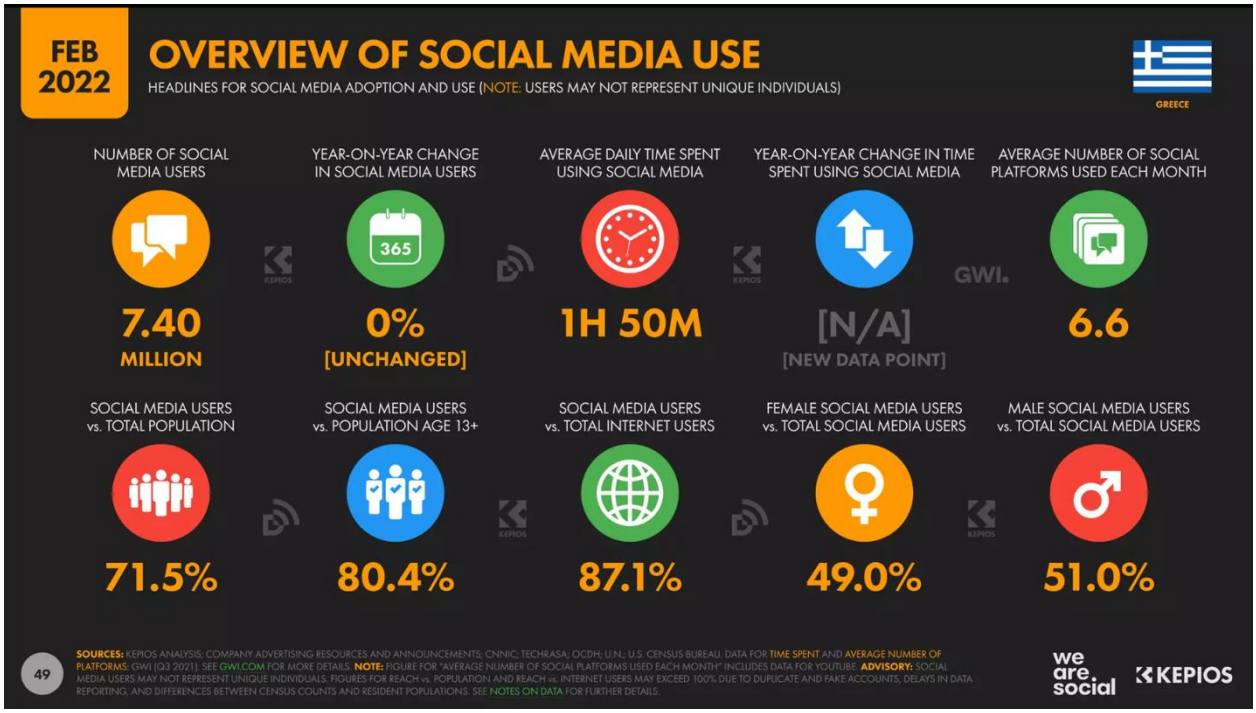



**Figure 5 Overview of social media use in Greece (February 2022)**

- The use of social media has not always been so massive and frequent (Figure 6). We see a steady increase from 2014 to the present, with a significant increase during the COVID-19 pandemic. The figures in the year 2021-2022 remain

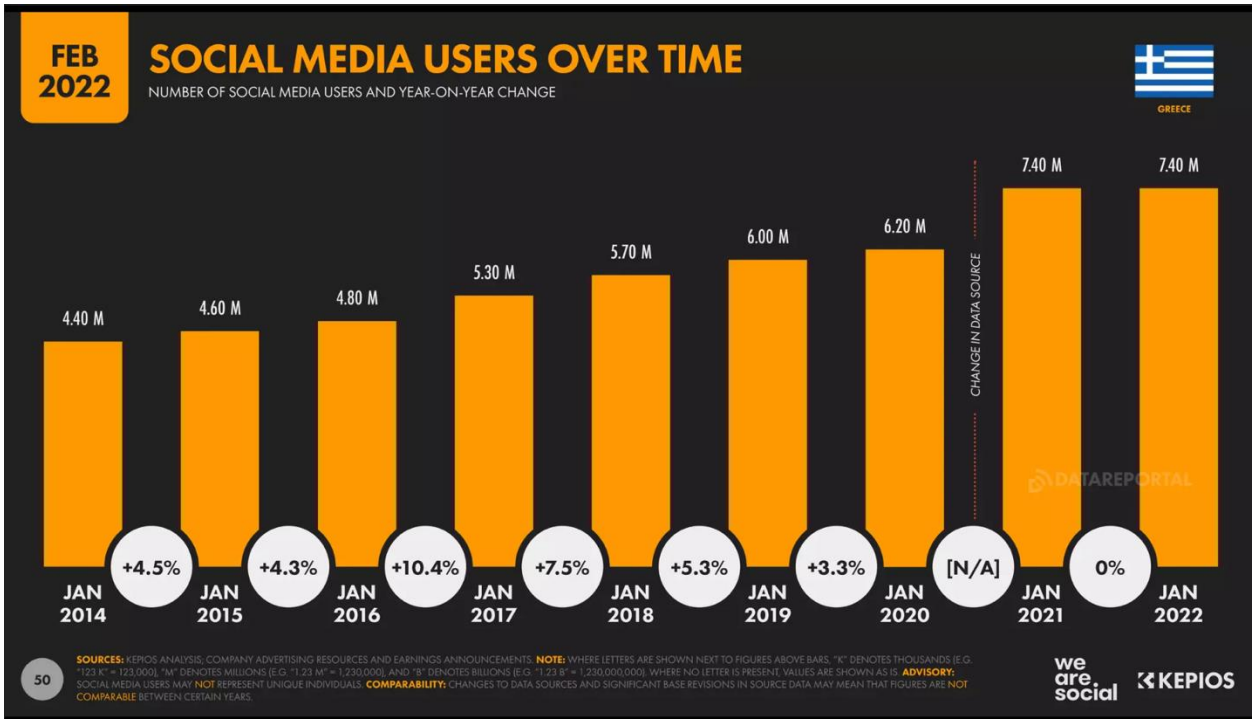



**Figure 6 Social media users over time in Greece (January 2014-January 2022)**

unchanged. In total from January 2014 to January 2022 we have 3 million new social media users in Greece (an increase of 68.18 %).

- Among the main reasons for using social media in Greece, 57.9 % are for informational purposes and 26.2 % are for sharing opinions (Figure 7).

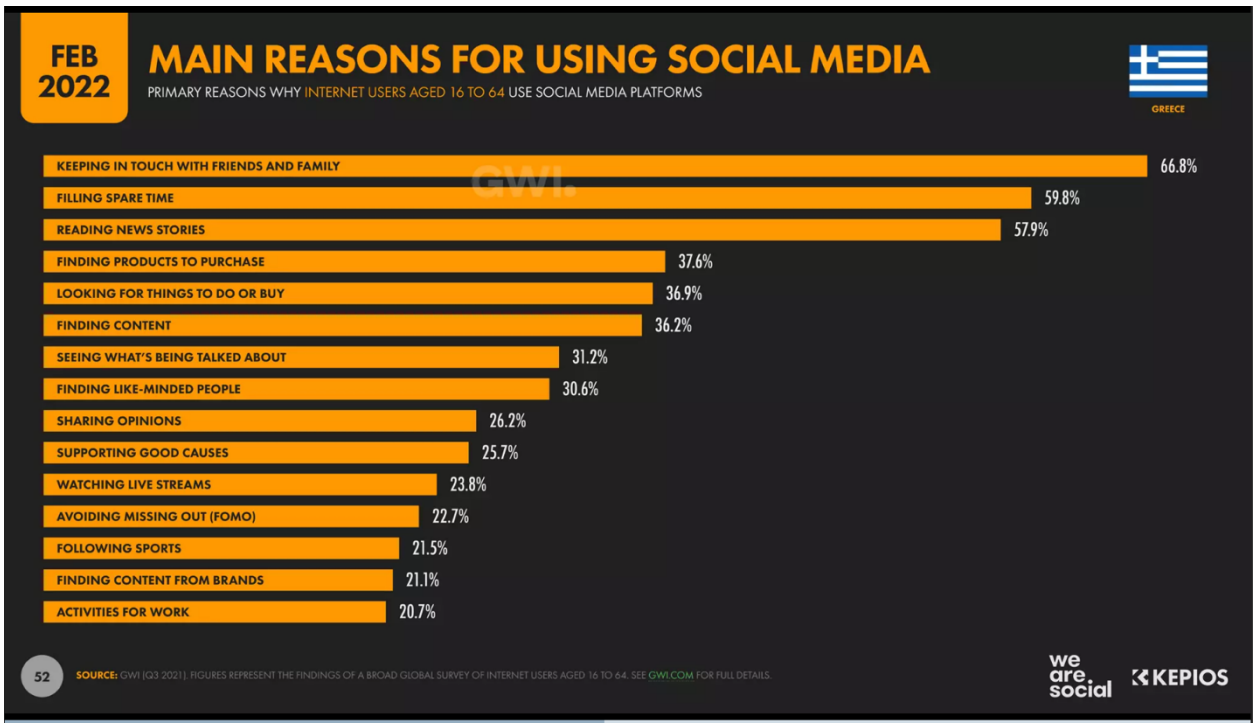



**Figure 7 Main reason for using social media in Greece (February 2022)**

- For the use of popular social media, in the 16 to 64 age group, Twitter comes in 8th place, accounting for 31.2% of online users in this age group (Figure 8).

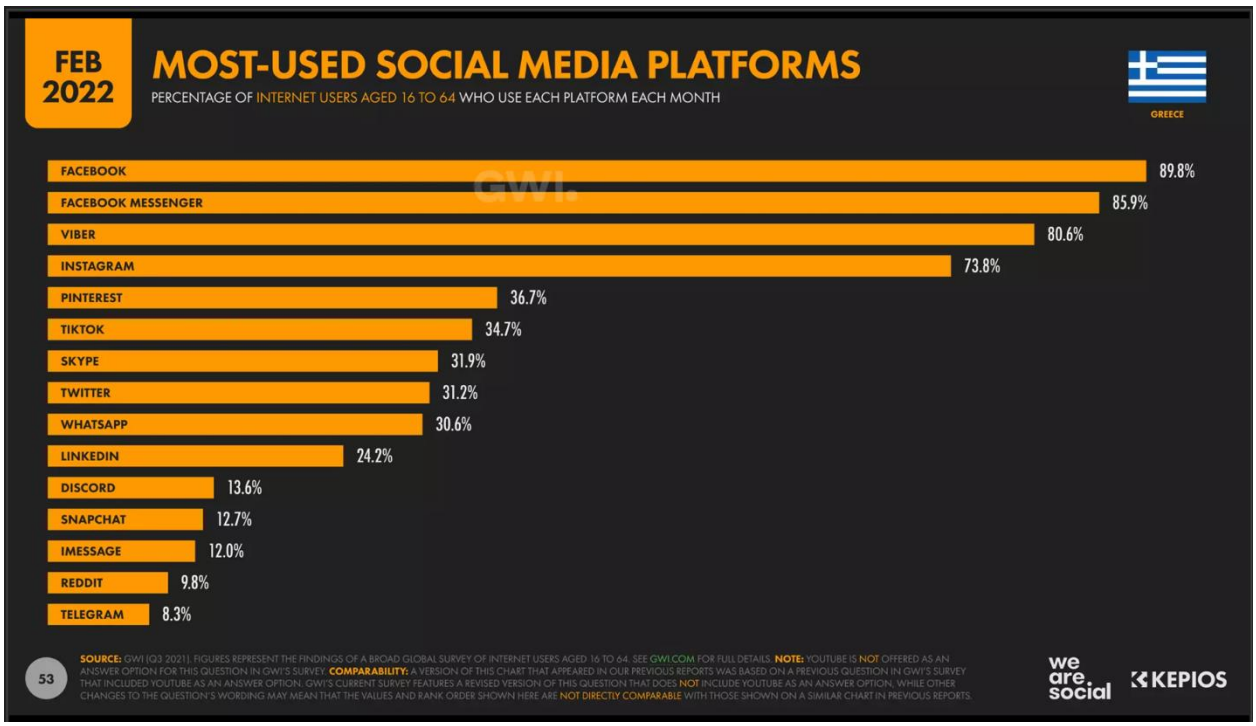



**Figure 8 Most-used social media platforms**

But beyond the percentages, we also need to know the absolute numbers for these platforms. Specifically, the absolute numbers of the platforms are presented in Table 3:

**Table 3 Social media platforms and their usage statistics**

| Platform | Total Users (millions)[1] | Percentage of population (%) | Percentage of "eligible[2]" population (%) | Female (%) | Male (%) |
|---|---|---|---|---|---|
| **Facebook** | 5.15 | 49.8 | 55.9 | 48.1 | 51.9 |
| **Instagram** | 4.35 | 42.1 | 47.3 | 51.2 | 48.8 |

[1] No social media platform publishes accurate data about its users. This data is derived from the advertising resources of each platform.

[2] Most platforms only accept users over the age of 13.

| | | | | | |
|---|---|---|---|---|---|
| **Messenger** | 4.5 | 43.5 | 48.9 | 48.9 | 51.1 |
| **TikTok** | 2.35 | - | 27.1 | 52.9 | 47.1 |
| **LinkedIn** | 1.9 | 18.4 | 21.9[3] | 44.4 | 55.6 |
| **Pinterest** | 1.69 | 16.3 | 18.4[3] | 80.3 | 14.6 |
| **Snapchat** | 0.85 | 8.3 | 9.3 | 64.7 | 35.1 |
| **Twitter** | 0.706 | 6.8 | 7.7 | - | - |

### 3.3 Sentiment Analysis on Twitter data

Twitter is a micro-blogging platform where users post "tweets" that are broadcast to their followers or sent to another user. Along with its huge and rapid growth, Twitter has been a fertile field for developing and testing sentiment analysis applications and approaches, as a tweet often expresses a user's feeling or opinion about an entity. Any user can comment on any topic related to social reality with a message of up to 140 characters. For this reason, TSA (Twitter Sentiment Analysis) is sentiment analysis at the sentence level [20].

Researchers have developed various approaches and tools to monitor Twitter and evaluate different events and phenomena, in order to capture valuable information about a wide range of topics. This includes things like reviews of movies or products, discussions about new laws or political movements, changes in public opinion about elections, movements in the market, or warnings about adverse medical events.

TSA (Twitter Sentiment Analysis) depends closely on the nature and topic addressed by the aggregated tweets. The research community has developed several techniques for implementing SA on datasets from twitter and continues in this orientation with the aim of improving them.

Some features of Twitter and other microblogging platforms that affect sentiment analysis are [20]:

- The writing style of a tweet is informal. There is use of acronyms, idioms, slang, abbreviations, emoticons and emojis.
- The informal style of writing is conducive to spelling and syntax errors.
- A tweet may include irony, sarcasm or other sentiment that is too complex for a machine to recognize.
- The maximum length of a tweet is 140 characters, which is why users typically refer to the topic in a wide variety of short and irregular forms.

[3] Eligible users on this platform are 18 years of age or older.

- The inability to appreciate the context and circumstances in which a tweet is written is a difficult problem for Sentiment Analysis.

Twitter Sentiment Analysis is a challenging task due to the unique features of the language used in tweets, which can negatively impact the accuracy of developed techniques. Despite these difficulties, research in this field has been ongoing for several years now, as the insights gained through TSA (Twitter Sentiment Analysis) are valuable across a wide range of business and social areas. Dozens of research papers have been published in recent years, highlighting the importance of this field. The two main motivations for TSA research are:

- To apply TSA to gain insights into various business or social issues, predict key indicators, or monitor Twitter for emerging information or events.
- To innovate and develop improved techniques for TSA, recognizing the value of information derived through accurate TSA.

For the implementation of twitter sentiment analysis there are 4 approaches-strategies [20]:

1. Supervised Machine Learning-based methods
2. Ensemble methods
3. Lexicon-based methods
4. Hybrid methods

### 3.3.1 Supervised Machine Learning-based

Beyond the individual approaches that anyone can take in the application they are implementing, a supervised machine learning approach has the following methodology:

In a supervised learning approach, a tagged dataset is used to train a model. The labeled dataset will consist of text data (such as tweets, customer reviews or survey responses) that have been previously labeled with a sentiment (positive, negative, neutral). Once the model is trained, it can then be applied to new, unlabeled text data to predict the sentiment of the text. There are several algorithms that can be used for sentiment analysis, such as Naive Bayes, Logistic Regression, Support Vector Machines (SVMs) and Recurrent Neural Networks (RNNs), such as LSTMs. These algorithms are trained on the labeled dataset and then used to classify new text data into one of the predefined sentiment categories.

There is an extensive literature on the use of supervised machine learning techniques in sentiment analysis. Some notable research contributions are the following:

- In [21] the importance of n-grams and emoticons was exploited in the implementation of a machine learning application. The researchers concluded that the implementation with SVM is better than that with Naive Bayes classifier. Also, among their experimental efforts, they concluded that the best performance was SVM with unigram feature extraction. This effort achieved 81 % precision and 74 % recall.

- In [22] labeled datasets were constructed by utilizing the twitter API and emoticon symbols. The classification performed was multi-class, distinguishing tweets into

positive, negative, and neutral. Static and linguistic analysis was performed on the data with attention to the frequency distribution of terms. Classifiers were trained using various techniques such as Multinomial Naive Bayes, SVM, and CRF. Among the various experimental efforts, the researchers concluded that Multinomial Naive Bayes approach using Part of Speech tags and n-grams performed optimally.

- The researchers in [23] created a machine learning application for sentiment analysis based on AdaBoost classifier. This classifier was trained with a varied combination of different parameters and features. Such parameters are lexicon, unigrams, bigrams, POS tags, microblogging features. Among the various experimental efforts, they concluded that the AdaBoost approach combined with n-grams, lexicon and microblogging features performed optimally.

- The researchers in [24] implemented a semantics merging approach with POS tags and unigrams. The main classifier they used in their approach is Naive Bayes. The utilization of semantics proved to be more effective than the unigrams-POS tags combination.

- Malhar and Ram in [25] developed a model based on SVM classifier. They concluded that the combination of SVM and principal component analysis (PCA) using n-grams (unigrams, bigrams) can reach an accuracy of 92%.

### 3.3.2 Ensemble methods

An ensemble method in sentiment analysis is a technique that combines the predictions of multiple individual models to make a more accurate overall prediction. This can be done by averaging the predictions of the individual models, by weighting the predictions based on the performance of the individual models, or by training a new model to make a final prediction based on the outputs of the individual models. Ensemble methods can help to improve the performance of sentiment analysis by reducing the variance and bias of the predictions, and by leveraging the strengths of multiple models to make more accurate predictions. The following are some notable research contributions:

- Xia et al [26] implemented a sentiment analysis application by combining 3 classifiers (Naive Bayes, Maximum Entropy, SVM). At the same time, 2 specific feature types, POS tags and word-relations, were exploited. We need to focus on their 2 main conclusions: 1) Ensemble classifiers achieve better results than single classifiers and 2) that combinations of different feature types give remarkable results.

- Lin and Kolcz [27] followed an approach based on the logistic regression classifier. The training set varied from one to 100 million of examples with ensembles of 3 to 41 classifiers. We need to focus on their main conclusions: 1) ensemble approaches show better results than single classifiers and 2) ensemble approaches have an additional computational cost, which is proportional to the number of classifiers they use. Among their various experimental efforts, the highest level of accuracy was achieved when the classifiers were set to 21 and the number of sample instances used were 100 million, resulting in a classification accuracy rate of 0.81.

- Da Silva et al [28] implemented an ensemble method, which was based on 4 different classifiers: SVM, random forest, logistic regression, Multinomial Naive Bayes. They experimented with different scenarios using bag of words (BOW) and feature hashing. Their implementation showed significant improvements in accuracy on specific datasets.

- Hagen, Matthias et al [29] implemented a model combination of 4 specific classifiers. They named their model "Webis". The main idea for their implementation was to use each classifier on a different feature set. Based on this, they get a confidence score for the predictions of the labels in the final.

- Chalothorn and Ellman [30] in turn also implemented an ensemble method with the combination of BoW and lexicon features. They achieved their best performance (F-score) with the combination of SVM, SentiStrengh and stacking methods.

### 3.3.3 Lexicon-based methods

A lexicon-based method in sentiment analysis is a technique that uses a predefined set of words, known as a lexicon or a sentiment dictionary, to classify the sentiment of a given text. The lexicon contains a list of words and their associated sentiment scores, which can be positive, negative, or neutral. To classify the sentiment of a text using a lexicon-based method, the text is first tokenized into words, and then the sentiment scores of each word are looked up in the lexicon. The overall sentiment of the text is then calculated by aggregating the sentiment scores of the individual words. This can be done by summing the sentiment scores, by taking the average sentiment score, or by counting the number of positive, negative, and neutral words in the text. Dictionary-based approaches do not require labelled data. They belong to the category of unsupervised techniques. Some notable research contributions are mentioned here:

- Paltoglou and Thelwall [31] used a dictionary-based method to measure the level of emotional intensity in order to make predictions. This method was suitable for identifying subjective texts expressing opinions and for classifying the polarity of emotion to determine whether the text was positive or negative. Their proposed dictionary-based approach achieved F1 scores of 76.2, 80.6 and 86.5 for the Digg, MySpace and Twitter datasets and outperformed all other supervised classifiers.

- Masud et al. [32] used a vocabulary-based system to classify sentiment in tweets, categorizing them as positive, negative or neutral. Their system was able to identify and score the slang used in tweets. The results of their experiments showed that their proposed framework outperformed previous methods, achieving 92% accuracy in dual classification and 87% in multi-category clustering. However, the system still needs to improve the accuracy in negative cases and further investigate the neutral cases.

- Asghar et al. [33] proposed a method to improve dictionary-based sentiment classification by incorporating a rule-based classifier to address data sparsity and increase accuracy. The approach involves using multiple classifiers, such as those using emoticons or negative modifier, and domain-specific or SWN-based classifiers, in turn to classify tweets based on sentiment polarity. The technique

was found to achieve F1 scores of 0.8, 0.795 and 0.855 for datasets with drug, car and hotel reviews respectively.

### 3.3.4 Hybrid methods

A hybrid-based method in sentiment analysis is a technique that combines multiple methods to classify the sentiment of a given text. This can be done by combining the predictions of different models, such as lexicon-based, rule-based and machine learning models, or by using a combination of different features, such as words, n-grams, and syntactic information.

Hybrid-based methods can be useful for sentiment analysis because they can leverage the strengths of different methods and features to make more accurate predictions. For example, a lexicon-based method can be used to identify the sentiment of individual words, while a machine learning model can be used to capture the sentiment of the text as a whole. Additionally, hybrid-based methods can also help to mitigate the limitations of individual methods. For instance, a lexicon-based method can be improved by incorporating additional features such as syntactic information, or a machine learning model can be improved by incorporating additional knowledge from a rule-based method. It's also worth noting that hybrid-based methods are more complex than single method-based techniques, they may require more computational resources, and they may be more difficult to interpret and explain. Some important research achievements are listed below:

- Balage, Filho and Pardo in [34] implemented a hybrid application for sentiment analysis in tweet text. Their system utilizes a combination of 3 classification methods: machine learning (SVM algorithm), rule-based and dictionary-based (SentiStrength dictionary). The results obtained from the experiments showed that the hybrid system outperformed the individual classifiers, achieving an F- measure of 0.56 compared to 0.14, 0.448 and 0.49 obtained by rule-based, dictionary and SVM-based classifiers respectively.

- Ghiassi et al [35] proposed a hybrid system, which operated with a combination of n-gram features and a dynamic artificial neural network (DAN2). They created a reduced dictionary from Twitter text and used SVM and DAN2 for the classification task. The results collected showed that the DAN2 learning method performed slightly better than the SVM classifier, even when the same Twitter-specific dictionary was incorporated.

- Khan et al [36] introduced a framework for Twitter opinion mining (TOM). This framework is a combination of SentiWordNet analysis, emoticon analysis and an enhanced polarity classifier. Based on all experimental approaches, the proposed framework achieved an average performance of 83.3%.

- Asghar et al [37] developed a hybrid application for emotion analysis in twitter that works with a combination of 4 classifiers: a slang classifier (SC), an emoticon classifier (EC), a general purpose sentiment classifier (GPSC) and an enhanced domain-specific classifier (IDSC). The application mines the sentiment of tweets after detecting the use of slang and emoticons. From the experimental analyses

conducted, it was found that computing the sentiment polarity of slang expressions and emoticons improves the accuracy of the model

## 3.4 Twitter in Politics

The use of Twitter as a source of data for making predictions or possible estimates is not limited to the commercial world but could also be applied to the field of politics. The use of twitter in the field of politics concerns both those in authority such as governments, international organizations and institutions (EU, IMF), political parties and those not in authority such as workers and ordinary people. It is also another way of interaction. Some important aspects of this issue are presented below:

- The importance of Twitter in elections and in politics in general was assessed in serious terms for the first time by Barack Obama and the Democratic Party in the US. On 29 April 2007, the then Democratic Party candidate, Barack Obama, launched his election campaign by publishing the following tweet: "Thinking we're only one signature away from ending the war in Iraq. Learn more at http://www.barackobama.com". Obama's first tweet, shown in Figure 9, was published about a year after the first tweet by Jack Dorsey. The use of twitter in this election battle contributed to Obama's electoral victory in this election. This move was to play a catalytic role in the utilization of twitter as a political instrument. The same logic was used again by Obama in the 2012 election. Since then, the road to the widespread use of Twitter by political figures has been opened wide. Nowadays, there is no political party that does not use social media to spread its agenda.


**Figure 9 Barack Obama's first tweet**

- Beyond the use of Twitter and other social media in election campaigns, the collection and processing of tweets from millions of users can probe the mood of public opinion. It is well known that on social media platforms, many users express their anger or dissatisfaction with a political figure or a political development. Assessing the mood of public opinion about a person or a bill or a political event is quite valuable for both governments and political parties. Social media is an informal parliament, with thousands of users "voting" electronically via their tweets on whatever is happening. So useful is this appreciation of public opinion that politicians use either paid users or bots to improve their online image with fake retweets and “likes”.
- The use of Twitter and other social media encourage popular discontent. The ability of every user to post anything that interests them has opened up new horizons in what we call "online activism". Various movements, through the rapid

exchange of information, have evolved into real uprisings, acting as a catalyst for political developments in one country and globally. Typical examples are the movements in the countries of Mediterranean Africa or the movement of "indignant citizens" in Spain and Greece, which evolved from simple protests into mass anti-government movements. The ability of every user to play the role of a simple journalist and the widespread use of Twitter as a means of expressing political opinion and, above all, criticism, are important weapons in the hands of many in the struggle for a better life.

- Therefore, social media platforms, including Twitter, are dangerously exploited many times, either to manipulate public opinion or to spread fake news and conspiracy theories. Politicians and big business collaborate in political advertising in order to mislead public opinion. The scandal of the large consulting firm, Cambridge Analytica, which illegally collected the personal data of 87 million users and in this way helped the 2016 election campaign of Ted Cruz and Donald Trump, is widely known. At the same time, the widespread use of Twitter to spread fake news has taken on huge proportions. The fake news phenomenon was gigantic during the COVID-19 pandemic. The major platforms have developed mechanisms to curb this phenomenon by disabling accounts (even Donald Trump's) and cutting off content.

## 3.5 Twitter and General Elections

The number of people who are social media users has attracted the interest of a number of organizations such as governments, political parties, multinational companies, etc. The huge amount of data generated by users on platforms such as Twitter is widely exploited for political and social use. This mode of use has added an additional role to twitter, namely, the role of a political tool.

Twitter is a digital environment with millions of users who are also active citizens and voters. Within this environment, in all countries, political figures, parties and organizations organize their political campaigns and contact their potential voters. Voters express their political opinions and preferences, which provide valuable feedback for political organizations. By using techniques such as Sentiment Analysis, one can determine the approval or disapproval of a political person or political party by estimating the overall like or dislike towards them. Thus, analyzing social media data on political issues can reveal useful socio-political trends from public opinion.

So, if we can analyze what these people wrote on Twitter and learn their political tendencies, can we predict the election results? More importantly, is there a correlation between Twitter posts and political tendencies? If there is such a correlation, to what extent is it valid? These are a series of questions that researchers have been trying to answer since 2010. Dozens of studies have been carried out in different countries to investigate the link between political trends in society and those on Twitter.

This effort has multiple advantages for political analysis of the period. First of all, the sample of participants is much larger than in conventional polls. Millions of users publicly share their personal opinions and their utilization is more complete in shape than the small statistical samples of surveys and polls. In addition, it is much lower in cost and much faster. Collecting and analyzing millions of tweets from Twitter is a process that can be handled in a very short time after setting up the first model and with just a little bit of computer work.

The research activity in this field is of interest to both the research community and the general public alike. Several applications have been developed, with datasets of different

origins (from different countries), which use a variety of approaches, with the most popular and typical ones presented in Table 4, as well as in research papers [38].

**Table 4 Twitter's dataset sentiment analysis and various approaches**

| Reference | Dataset | Approach | Conclusions |
|---|---|---|---|
| [39] | They include tweets about the 6 political parties that had MPs in the German parliament and their leaders | Emotion analysis was performed with LIWC text analysis software | It has been found that the correlation between the number of tweets mentioning the parties and their corresponding voting share in the election results is consistent. |
| [40] | Tweets about the Irish general election of 2011 with the hashtag “#GE11” | In their research, they employed a combination of volume-based analysis and sentiment analysis, with the assumption that a party's voting share is correlated with the volume of related content on social media. The tweets were evaluated utilizing a supervised learning sentiment analysis method developed by them. | Correct prediction of the number of votes with mean absolute error of 1.61% with proportioning the total number of tweets related to a party to the sum of all the tweets |
| [41] | Collected 64,395 tweets a week before the election to estimate the 2011 Dutch election results. | Manual sentiment analysis | Prediction of the voting rate with absolute difference of 17.4 %. |
| [42] | They collected 7,541,470 tweets over 3 months for the prediction of the 2012 USA Presidential Election result. | They didn't take into account the number of tweets, but the user accounts that posted them. They used the AFINN library to implement sentiment analysis. They made the assumption that each tweet with positive sentiment | They arrived at a candidate's total support by combining the results of the multiplication. Through this analysis, they were able to accurately predict that Obama would win the 2012 election prior to the actual event. |

| | | | |
|---|---|---|---|
| | | represents a positive vote for that candidate | |
| [43] | For estimating the results of the 2013 election in Pakistan, they took into account the number of tweeters. They identified 24 individuals who tweeted about the election and amassed a total of 9000 tweets during the election period. | They manually specified 40 words for or against the parties. According to these keywords in a tweet, they used RapidMiner grouping models to predict which party the tweet supports and which party it is against. | Failed to predict the correct result of the general elections. |
| [44] | They collected tweets during the 10 days preceding the election, using 27 manually selected keywords such as party hashtags, party names, leader names, and general election hashtags related to the 2012 Albertan general election in Canada. As a result, they collected 181,972 tweets from 28,087 accounts. | To develop their model, they took into account a number of characteristics of tweets such as the number of “likes”, retweets and interaction with political parties and their candidates. They used the same model for Pakistan's 2013 general elections | They concluded that their model can predict the political preferences and trends of Twitter users. |
| [45] | They collected tweets for Donald Trump and Hillary Clinton, two candidates who participated in the 2016 USA Election, containing the names of the parties and their candidates. Their dataset consisted of more than 60000 tweets. | Used the VADER algorithm to perform sentence level sentiment analysis. After that, they used Multinomial Naïve Bayes and SVM for their machine learning models. They distinguished the tweet in positive and negative. | The method of estimating the election outcome by dividing the total number of positive tweets about a candidate by the total number of tweets related to that candidate was found to be inaccurate. |

| | | | |
|---|---|---|---|
| [46] | They gathered English tweets that included keywords such as republican, democrat, Hillary Clinton, Donald Trump and location information that pertained to parties or candidates participating in the 2016 US presidential election. | In order to perform sentiment analysis on tweets, they used Sentiment140 tweet corpus which includes 1,600,000 datasets (800,000 for training positive and negative emotions) and 497 test data (181 positives, 177 negatives, and 139 neutral) along with an abbreviations dictionary. They applied sentiment analysis using 3 different sentiment classifiers: Binarized Multinomial Naïve Bayes Classifier, SentiWordNet, and AFINN. | Based on their analysis, they classified tweets about a party with positive emotion as a positive vote for that party, and tweets with negative sentiment as a positive vote for the opposing party. This approach likely resulted in the incorrect assumption that an individual would cast positive votes for multiple parties based on different tweets. As a result, their study incorrectly predicted that Hillary Clinton (Democrat) would win the election with 253 electoral votes, while Donald Trump (Republican) would receive only 219 electoral votes. |
| [47] | They gathered tweets that pertained to the 2017 French elections, which were posted prior to the election and included the names of the candidates. | They performed conventional Sentiment Analysis. They estimated the election result taking into account not only the positive or negative but also the number of neutral tweets about the parties | They calculated the daily election results based on the formulas they established using the tweets they collected. On the final day, they predicted the election results with only a 2% margin of error. They stated that their vote rate estimates, which included neutral tweets, were more accurate than methods that only considered positive and negative tweets |

# 4. PROJECT DESCRIPTION

## 4.1 General description and research question

The present Thesis is an attempt to implement a Sentiment Analysis technique for tweet datasets in the Greek language. The case under consideration is datasets related to the 2019 general parliamentary elections in Greece. The elections took place on Sunday 7 July 2019. Of the political parties that participated, six (6) of them elected members of Parliament (MPs).

The Greek political parties are “New Democracy” (Liberal-Conservative, Center-Right Wing), “SYRIZA” (Left Wing), “KINAL-PASOK” (Social-Democratic Party), KKE (the Communist Party), “Elliniki Lysi” (Right Wing – Far Right/ Right Wing Populist) and Mera25 (Left Wing -“European Realistic Disobedience Front”).

The respective percentages and seats of each a political party are presented in Table 5.

**Table 5 Results of Greek general elections at 07/07/2019**

| Political Party | Percentage (%) | Seats won (300 seats) |
|---|---|---|
| **Νέα Δημοκρατία (ΝΔ) [ND-New Democracy]** | 39.85 | 158 |
| **Συνασπισμός Ριζοσπαστικής Αριστεράς (ΣΥΡΙΖΑ) [SYRIZA]** | 31.53 | 86 |
| **Κίνημα Αλλαγής (ΚΙΝΑΛ-ΠΑΣΟΚ) [KINAL-PASOK]** | 8.1 | 22 |
| **Κομμουνιστικό Κόμμα Ελλάδος (ΚΚΕ) [KKE]** | 5.3 | 15 |
| **ΕΛΛΗΝΙΚΗ ΛΥΣΗ [Elliniki Lysi]** | 3.7 | 10 |
| **ΜέΡΑ25 [Mera25]** | 3.44 | 9 |

In the present research, we describe a lexicon-based technique for sentiment analysis on 5 datasets consisting of tweets. Two approaches have been followed:

1. Sentiment analysis is implemented on each dataset without distinguishing between users posting content.

2. Sentiment analysis is implemented on each dataset after having made the distinction between active and inactive users. Active users (“vocal minority”) are

those who produce a large amount of content in a short period of time and inactive users (“silent majority”) are those who produce little or no content.

Each dataset refers to one of the political parties that managed to elect Members of Parliament (MPs) in the 2019 elections. Using this implementation, plausible answers regarding the following research question are targeted:

*"Is there a correlation between*
*(a) the overall sentiment of a set of election-related tweets and*
*(b) the social trends that characterized the election?*
*And, if so, (c) to what extent does this correlation exist?"*

The key motivation for the present Thesis is both the absence of a variety of Sentiment Analysis applications for the Greek language and the absence of corresponding approaches that are politics-oriented and specific to the Greek political scene. In particular:

1. The absence of a variety of Sentiment Analysis applications for the Greek language. There are dozens of papers published which address the issue of SA for various languages, such as English, Chinese, German and others. However, SA tools based on Greek are few and do not cover the whole range of SA applications and topics.

2. The absence of corresponding approaches that are politics-oriented and specific to the Greek political scene. There is a considerable number of scientific publications and applications for the prediction and estimation of social trends and election results based on Twitter for several countries, as described in chapter 3. Of course, the size of the corresponding literature for the case of Greek elections (any election) is minimal to zero.

Regarding the implementation of the Sentiment Analysis application, the programming language used to write the code for the application is Python. Python [48] is a high-level, general-purpose programming language that is widely used in a variety of fields, including web development, data science, scientific computing, and artificial intelligence. It is known for its simplicity, readability, and flexibility, as well as its extensive standard library and large community of users and developers. One of the key features of Python is its support for powerful and easy-to-use libraries and frameworks for tasks such as data manipulation, machine learning, and web development. This makes it a popular choice for rapid prototyping and development of complex applications.

For the graphs presented, the Tableau tool has been used. Tableau [49] is a data visualization and business intelligence software. It allows users to connect to various data sources, including excel spreadsheets, SQL databases, and web services, and then create interactive visualizations, dashboards, and reports. It provides a drag-and-drop interface that makes it easy to create charts, graphs, and maps, and allows users to analyze and explore data in real-time. Tableau also offers a wide range of options to customize the look and feel of the visualizations and provides tools to share and publish the data insights to the web or embedded them in other applications. It is widely used by business and data analysts, data scientists, and other professionals who need to make sense of large data sets and communicate their findings to others.

## 4.2 Datasets

The data used for the implementation comes from Twitter. For the formation of the datasets, the tool "snscrape" has been used. Snscrape [50] is a scraper for social networking services. Using it, a user can gather data from various platforms based on the desired search criteria such as hashtags, usernames, date range, term in tweet text. The Snscrape tool is compatible with several platforms such as Facebook, Twitter, Instagram, Mastodon, Reddit, Telegram, VKontakte and Weibo. As a first step, data collection and processing involve the identification of tweets for each political party. Of all the political parties that participated in the 2019 elections, I chose to collect data only for the 6 that managed to elect members of parliament.

The search performed in the present analysis and implementation concerned both specific hashtags as well as specific terms. Specifically, for each political party under study:

1. Tweets were searched based on specific hashtags. If a tweet contains even one hashtag from those searched, then it is stored in the dataset.

2. Tweets were searched based on specific terms. If a tweet includes even one term in its text from those searched, then it is stored in the dataset.

Tweets for each political party from 24/06/2019 to 07/07/2019 (13 days in total), i.e., up to two weeks before the start of the election, are searched for. For each hashtag and term, there is a limit of 10000 tweets, so as not to create huge datasets. It is not impossible to have the same tweets in 2 or more datasets, as there is a possibility that they share the same hashtags or terms. The snscrape tool does not return retweets, i.e., tweets that have been republished by other users.

Based on this process, for each political party, two (2) datasets are created which are then merged into one. The columns and the content of each column are presented in Table 6.

**Table 6 Columns of the dataset and its content**

| Column | Content |
|---|---|
| **ID** | The unique ID of each tweet |
| **Username** | The username of the account published the tweet |
| **Hashtags** | The hashtags included in a tweet |
| **Datetime** | The timestamp of the tweet |
| **“Likes”** | The number of “likes” that a tweet receives |
| **Retweets** | The number of retweets that a tweet receives |
| **Location** | The location of the account posted the tweet |

For each individual political party, a specific search has been performed. We note that the names of Party Leaders and their variations (signalized here as “[PL]”), as well as the Party names and their variations (signalized here as “[PN]”) are commonly linked to the Hashtags concerning the respective parties. Mottos including a specific Party name are (signalized here as “[Motto]”) also linked to the respective Hashtags, as well as particular programs, organizations, and parliament members-politicians (signalized here as “[Prt]”).

**Table 7 Hashtags and terms used to create the examined datasets about the six greek political parties**

| Political Party | Hashtags | Terms in tweet text |
|---|---|---|
| **ΝΔ [ND-New Democracy]** | 'Mitsotakis' [PL] 'kyriakosmitsotakis' [PL], 'Kyriakos' [PL], 'Kiriakos' [PL], 'NeaDemokratia' [PN], 'neadimokratia' [PN], 'ND' [PN], 'KyriakosMitsotakis' [PL], 'KMitsotakis' [PL], 'NewDemocracy' [PN], 'NEA_ΔΗΜΟΚΡΑΤΙΑ' [PN], 'Μητσοτάκης [PL]', 'Κυριάκος_Μητσοτάκης' [PL], 'ΝΔ' [PN] | 'Μητσοτάκης', 'Μητσοτάκη', 'Μητσοτακη', 'ΝΔ', 'ΝέαΔημοκρατία', 'Κυριάκος Μητσοτάκης', 'Με τον Κυριάκο' |
| **ΣΥΡΙΖΑ [SYRIZA]** | 'Tsipras' [PL], 'AlexisTsipras' [PL], 'atsipras'[PL], 'ΤωραΣΥΡΙΖΑ ' [Motto], 'SYRIZA' [PN], 'ΣΥΡΙΖΑ' [PN], 'ΤώραΑποφασίζουμεΓιαΤηνΖωήΜας' [Motto],, 'μόνο_συριζα' [PN], 'alexistsipras' [PL], 'τώραΣΥΡΙΖΑ' [PN], 'primeMinisterGr'[PL], 'ΠροοδευτικηΣυμμαχία' [PN], 'μετονΑλεξη' [PL] | 'Τσίπρας', 'Τσίπρα', 'Αλέξης Τσίπρας', 'ΣΥΡΙΖΑ', 'Προοδευτική Συμμαχία' |
| **ΚΙΝΑΛ-ΠΑΣΟΚ [KINAL-PASOK]** | "PASOK" [PN], "ΠΑΣΟΚ"[PN], 'kinimallagis'[PN], "ΚΙΝΑΛ"[PN], "ΚίνημαΑλλαγής [PN]", 'ΦωφηΓεννηματα' [PL], | ΠΑΣΟΚ', 'ΚΙΝΑΛ', 'Κίνημα Αλλαγής', 'Γεννηματά', 'Φώφη', 'Φώφη Γεννηματά' |

| | | |
|---|---|---|
| | 'Φώφη'[PL], 'ΚΙΝΗΜΑ_ΑΛΛΑΓΗΣ', 'gennhmata'[PL], 'Γεννηματά [PL]', 'ΚΙΝΑΛ_ΠΑΣΟΚ'[PN], 'Kinima_allagis'[PN], 'Κίνημα_Αλλαγής'[PN], 'ΚΙΝΑL'[PN], 'ΠΑΣΟΚ_ΚΙΝΗΜΑ_ΑΛΛΑΓΗΣ'[PN], 'ΠΑΣΟΚ_Κιναλ'[PN] | |
| **ΚΚΕ [ΚΚΕ]** | "ΙΣΧΥΡΟ_ΚΚΕ" [Motto], "ΚΚΕ"[PN], "ΚΚ"[PN] "Ριζοσπάστης" [Prt], "Ριζοσπάστη"[Prt],, "ΚΝΕ" [Prt],, "Μενει_ΚΚΕ" [Motto],, "ΕΥΓΕ_ΚΚΕ"[Motto],, "ΚΚΕ_ΤΩΡΑ"[Motto],, "ΤΩΡΑ_ΚΚΕ"[Motto],, "ΚΚΕ_ισχυρό"[Motto],, "ΚΚΕ_με_το_νου"[Motto],, "epomeni_mera_ΚΚΕ" [Motto], "Κουτσούμπας"[PL], "τώραΚΚΕ"[Motto],, "Ψηφίζεις_ΚΚΕ"[Motto],, "ΚΚΕ_Επειδή"[Motto],, "ΚΚΕ_η_μονη_κερδισμενη_ψηφος"[Motto],, "ΠΑΜΕ [Prt], "ΔημήτρηςΚουτσούμπας" [PL] "kane_ti_diafora" [Motto],, "mono_ΚΚΕ"[Motto],, "ΜΟΝΟ_ΚΚΕ" [Motto], | "ΙΣΧΥΡΟ_ΚΚΕ", "ΚΚΕ", "ΚΚ" "Ριζοσπάστης", "Ριζοσπάστη", "ΚΝΕ", "Μενει_ΚΚΕ", "ΕΥΓΕ_ΚΚΕ", "ΚΚΕ_ΤΩΡΑ", "ΤΩΡΑ_ΚΚΕ", "ΚΚΕ_ισχυρό", "ΚΚΕ_με_το_νου", "epomeni_mera_ΚΚΕ" "Κουτσούμπας", "τώραΚΚΕ", "Ψηφίζεις_ΚΚΕ", "ΚΚΕ_Επειδή", "ΚΚΕ_η_μονη_κερδισμενη_ψηφος", "ΔημήτρηςΚουτσούμπας" "kane_ti_diafora", "mono_ΚΚΕ", "ΜΟΝΟ_ΚΚΕ |
| **ΕΛΛΗΝΙΚΗ ΛΥΣΗ [Elliniki Lysi]** | 'Βελόπουλος' [PL], 'ΕλληνικήΛύση'[PN] 'ελληνική_λύση', [PN] 'Ελληνική_Λύση', [PN] 'velopky' [PL] | 'Βελόπουλος', 'Ελληνική Λύση', ‘ελληνική_λύση', 'Ελληνική_Λύση', 'velopky' |
| **ΜέΡΑ25 [Mera25]** | 'diem25', [PN] 'MEPA25', [PN] 'Βαρουφάκης' [PL], 'Γιάνης' [PL], 'yanisvaroufakis'[PL], '7Τομές' [Prt], 'YanisVaroufakis' [PL], | 'diem25', 'MEPA25', 'Βαρουφάκης', 'Γιάνης', 'yanisvaroufakis', '7Τομές', 'YanisVaroufakis', |

| | | |
|---|---|---|
| | 'Diem25', [PN]<br>'ΜέΡα25', [PN]<br>'Varoufakis' [PL],<br>'ΠρόγραμμαΜέΡΑ25' [Prt],,<br>'Ψηφίζω_Μέρα25' [Motto],,<br>'MeRA25', [PN]<br>'ξημερώνει_ΜέΡΑ25'<br>[Motto],<br>'κλέωνγρηγοριάδης [Prt],',<br>'ΨηφιζουμεΜεΡΑ25<br>[Motto],',<br>'DiEM25', [PN]<br>'ΚρίτωνΑρσένης'[Prt],<br>'ψηφίζουμε_ΜέΡΑ25'<br>[Motto] | 'Diem25',<br>'ΜέΡα25',<br>'Varoufakis',<br>'ΠρόγραμμαΜέΡΑ25',<br>'Ψηφίζω_Μέρα25',<br>'MeRA25',<br>'ξημερώνει_ΜέΡΑ25',<br>'κλέωνγρηγοριάδης',<br>'ΨηφιζουμεΜεΡΑ25',<br>'DiEM25',<br>'ΚρίτωνΑρσένης',<br>'ψηφίζουμε_ΜέΡΑ25' |

For each political party, the most frequently used terms and the most common hashtags have been selected and are presented in Table 7. Each tweet can include more than one of the above hashtags. This factor led to the existence of many duplicate tweets, which were subsequently deleted from the respective datasets. Also, the large number of hashtags does not imply many tweets. It is worth noting that, in addition to the name of the political party (in all its Internet variants), a search has been performed based on the name of its political leader and/or other prominent figures.

### 4.3 Exploratory Data Analysis

Datasets differ in the number of tweets they have stored. This is a common phenomenon, which is also a limitation for the application. Each tweet has a unique field, its ID. Before processing, duplicates based on the ID have been removed. The removal of the duplicates results to the respective number of tweets for each political party, as presented in Table 8.

Using Python's WordCloud library [51], a wordcloud was formed, with a dataset for the 100 most frequent terms in each dataset. Figures 10-15 show the respective wordclouds.

**Table 8 Number of tweets and the corresponding political party**

| Political party | Number of tweets |
|---|---|
| **ΝΔ [ND-New Democracy]** | 31023 |
| **ΣΥΡΙΖΑ [SYRIZA]** | 33062 |
| **ΚΙΝΑΛ-ΠΑΣΟΚ [KINAL-PASOK]** | 11481 |
| **ΚΚΕ [KKE]** | 6644 |
| **ΕΛΛΗΝΙΚΗ ΛΥΣΗ [Elliniki Lysi]** | 2761 |
| **ΜέΡΑ25 [Mera25]** | 3879 |


**Figure 10 Wordcloud for ND political party**

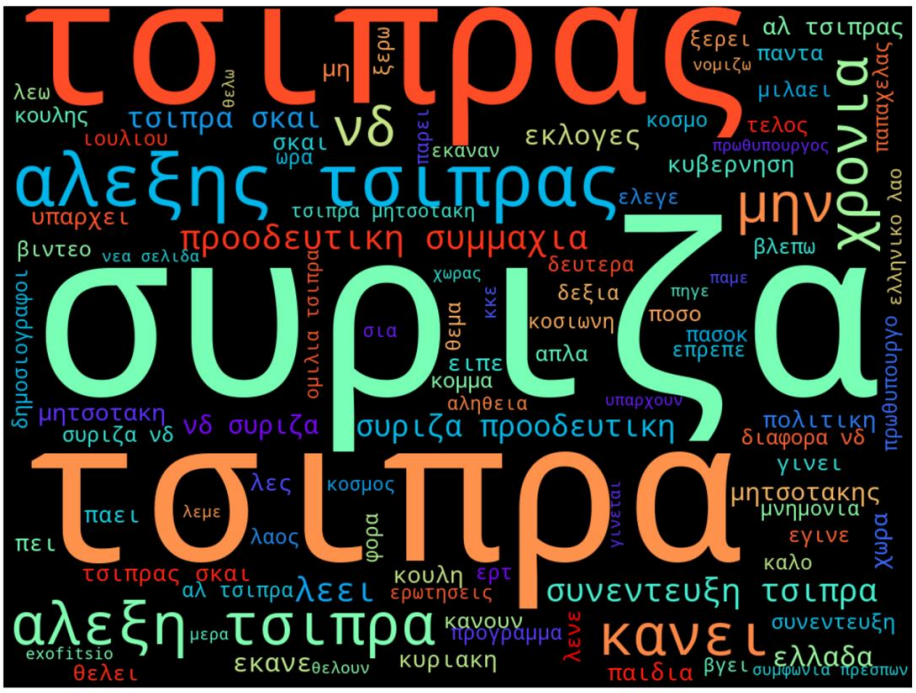



**Figure 11 Wordcloud for SYRIZA political party**

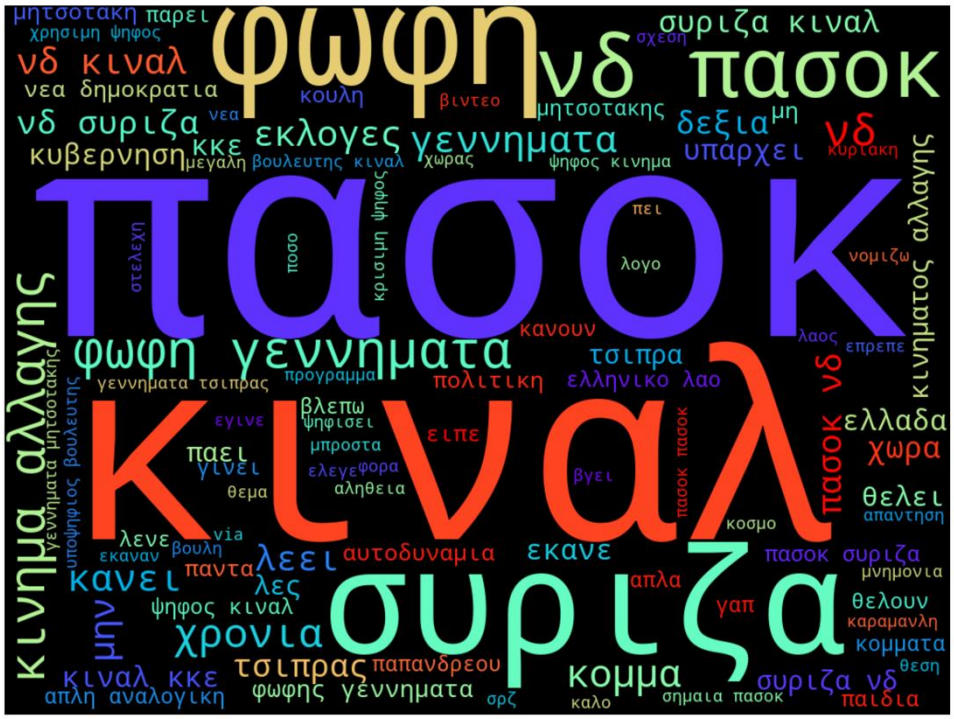



**Figure 12 Wordcloud for KINAL-PASOK political party**

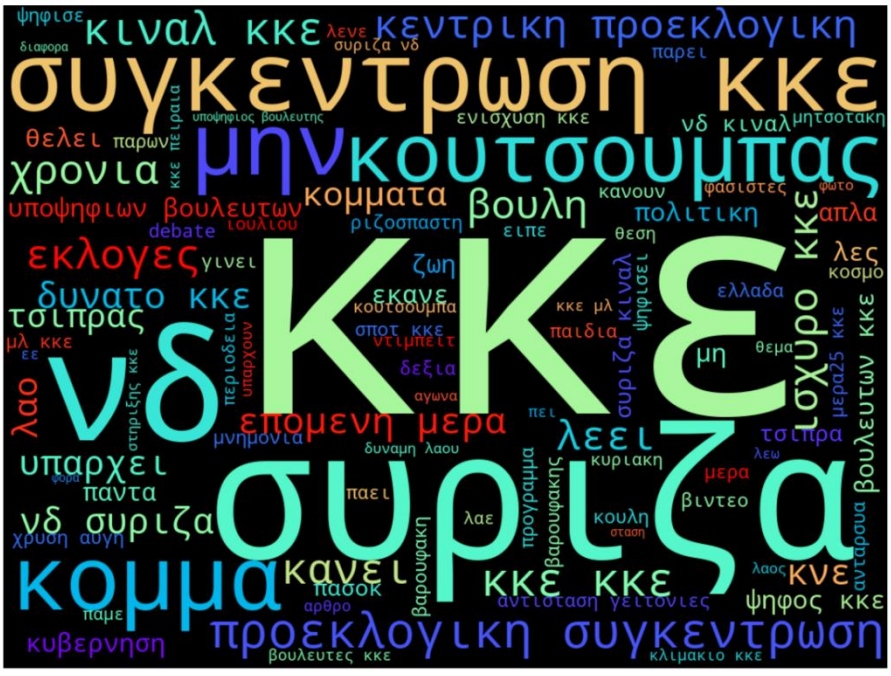



**Figure 13 Wordcloud for KKE political party**


**Figure 14 Wordcloud for Mera25 political party**

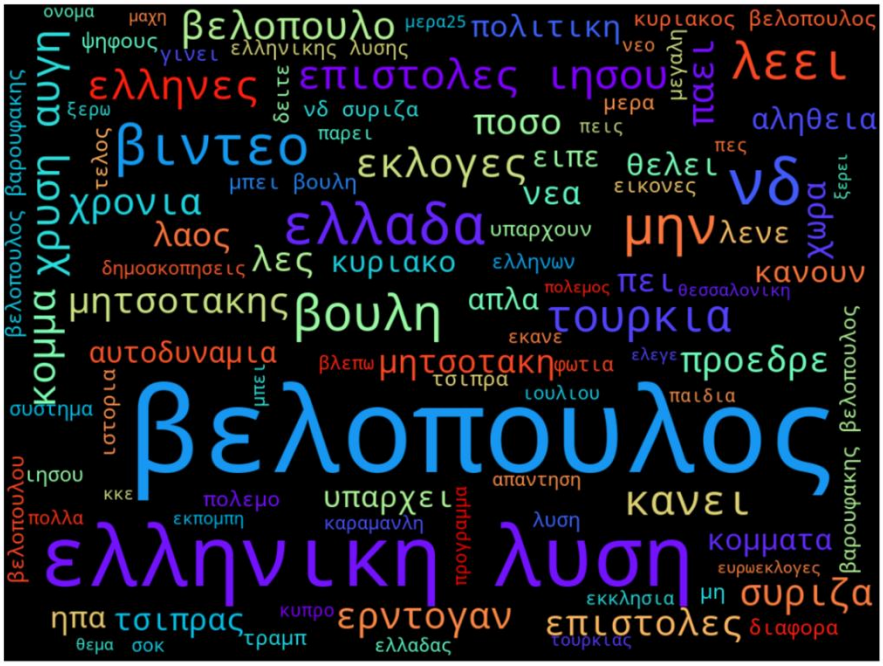



**Figure 15 Wordcloud for Elliniki Lysi political party**

Beyond the most commonly occurring words and terms in each dataset, particular features and other elements emerging from the analysis of the datasets are observed.

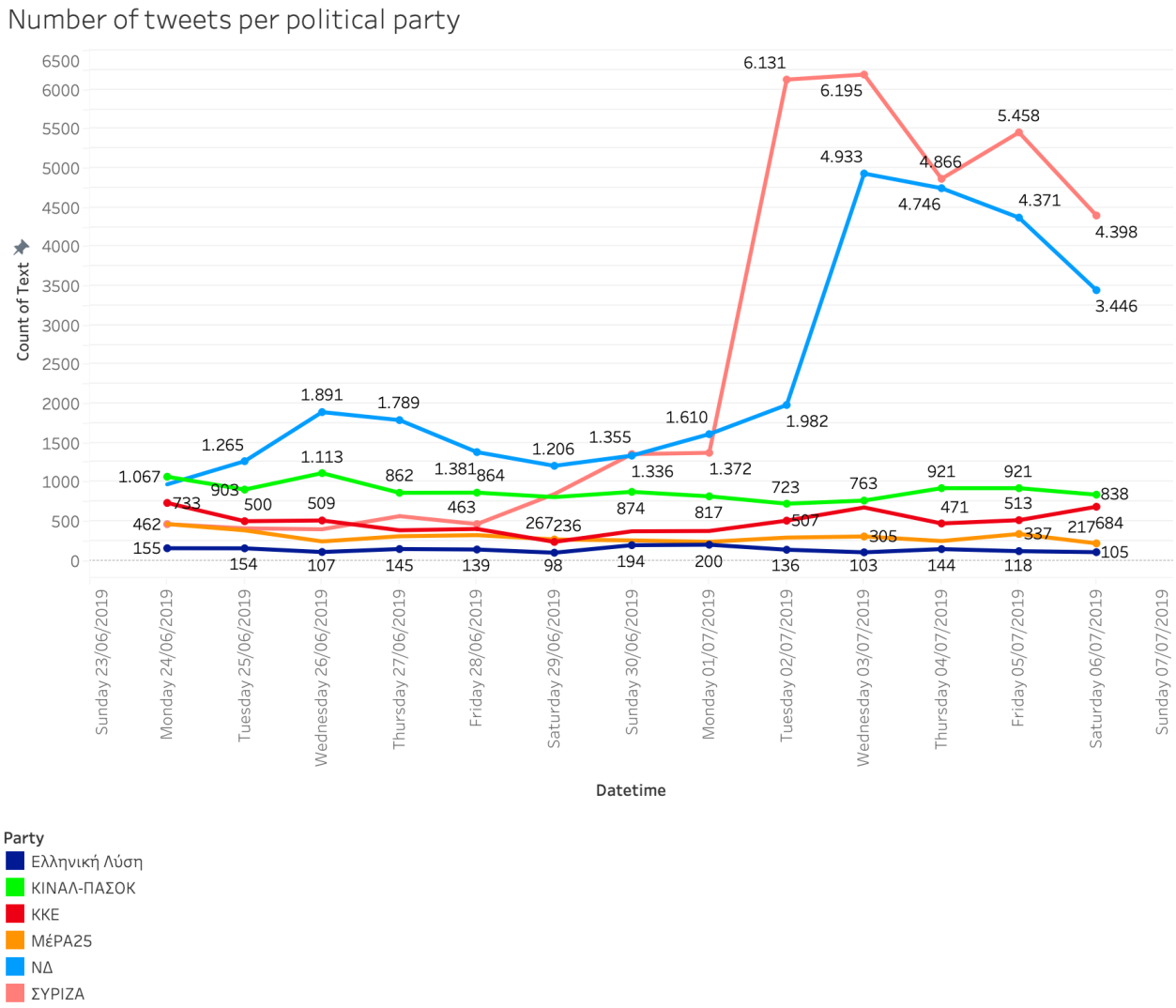



**Figure 16 Number of tweets per political party in general approach**

As shown on the graph below (Figure 16), the New Democracy (ND) party starts from quite low in the number of daily tweets, but from the beginning of the second week of the election period it manages to reach a maximum of 4.933 tweets per day, closing at 3.446 tweets daily one day before the elections.

The SYRIZA party (SYRIZA) starts from even lower numbers than the New Democracy (ND) party, but in the 2nd week of the election it manages to make a spectacular shift, approaching 6131 tweets (its maximum) and closing at the top with 4398 tweets the day before the elections.

The other political parties (KKE, KINAL-PASOK, Elliniki Lysi, Mera25) are limited to low numbers, failing to exceed 1000 tweets a day, with a few exceptions of KINAL-PASOK, which has the lead in this set of parties. The last place is occupied by Elliniki Lysi's party with a maximum number of daily tweets with a maximum of 200 tweets.

Unfortunately, no conclusions can be reached regarding the location of published tweets, as few users have set a clear and precise location. If similar information were available, an estimate of the sentiment of tweets for all prefectures of the Greek territory would be possible. This information would be quite useful in case we wanted to estimate the social trends of the elections by prefecture and region, beyond the total territory of the country.

## 4.4 Data Preprocessing

The preprocessing process involves a series of techniques, with which we process the datasets we have created in such a way that they are suitable for use in the task we are implementing. In other words, through various steps we are able to reduce the noise in the data we have as input. The steps and techniques that a user implements in the preprocessing stage are related to the requirements of the application. A technique that may reduce noise in one task, in another task with different requirements may be detrimental to the effectiveness of the model. For example, removing emoticons or emojis is a useful technique for the case of text classification. However, in the case of Sentiment Analysis, emojis and emoticons contribute to the sentiment estimation of a document.

The critical field in concerning the present application is the 'Text' column in the datasets collected. Using the libraries of the python programming language, the following tasks have been performed:

1. duplicate rows have been eliminated (those rows containing the same text)
2. all null values have been removed (all rows that have the value null in the “Text” column)
3. all text less than 5 characters long have been converted to “nonestring”
4. all hashtags have been removed from the text (#hashtag)
5. all mentions in twitter accounts have been removed
6. all URLs and links have been removed, i.e., strings starting with www, http, https
7. all emails have been removed, i.e., strings of the form (the strings of the form abc@xyz.com)
8. punctuation symbols have been removed
9. all text characters have been converted to lower case
10. the newline character ('\n') has been removed
11. the multiple space characters have been replaced by a single blank character
12. the accents have been removed from the Greek letters

An important and often used preprocessing technique is the removal of stop words. However, it was decided not to remove the stop words of the Greek language, because the original text would become even more difficult for syntactic and morphological analysis. Correct verification of syntactic and morphological rules is an auxiliary factor in the natural language processing task implemented in the present research.

## 4.5 Datasets Sentiment Analysis

The implementation used is a lexicon-based approach, comprising two basic steps, namely, creating the emotional lexicons and scanning the texts for each dataset. The lexicon-based approach concerns the detection and processing of 3 typical and characteristic categories of linguistic data expressing sentiment and overall attitude towards a political party (in Greece), in particular:

1. Mottos (typical of specific political parties in Greece)
2. Negative statements with the use of negation
3. Characteristic expressions with irony

In the present stage, the lexicon-based approach targets to identify and process only the most typical and characteristic categories of linguistic data for the purposes of signalizing

the positive, negative or neutral sentiment and overall attitude towards a political party. It is considered that the present approach may also function as an initial processing tool for the further detection, evaluation and processing of more complex cases of linguistic information expressing sentiment and overall attitude. More complex cases usually concern connotative features and implied information and often require expert knowledge for their detection and evaluation.

The overall process concerning the implementation is depicted in Figure 17. The text of each tweet, after it has been preprocessed and noise removed, is the input for the process. Then, the existence of negation in the text is checked. If a negation exists, a method is taken to perform lexicon analysis for the negation. After processing, the total polarity and sentiment of the text are calculated as a component of the emotionally charged terms present in the text. At the end, the results for each dataset are gathered.

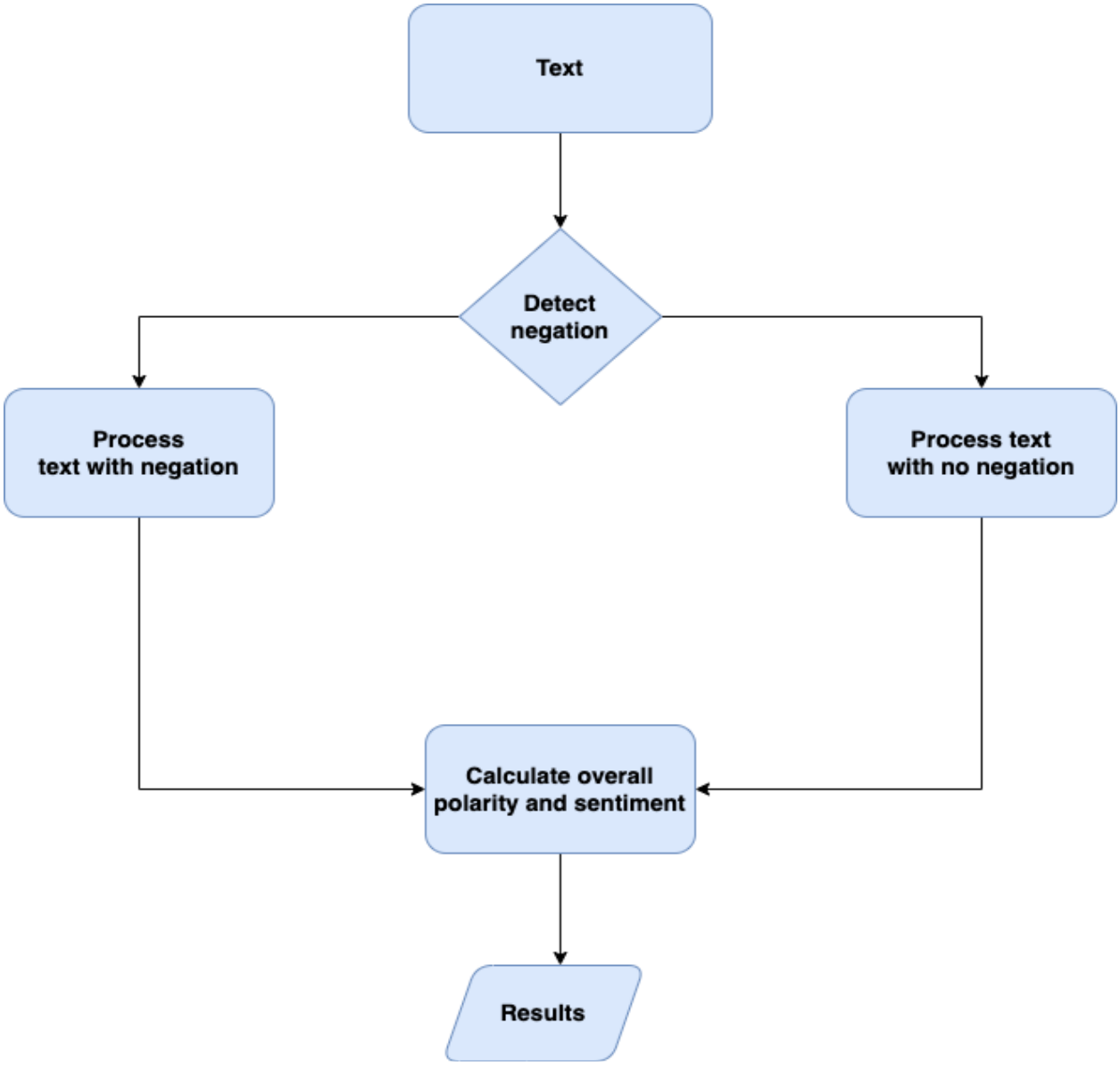



**Figure 17 Main implementation's flow chart**

### 4.5.1 Creating the emotional lexicons

Specifically, two lexicons have been created in Greek, one with positively charged terms and the other with negatively charged terms. As a basis for the two (2) dictionaries, the ones available in [52] are used. At this reference, dictionaries of positive and negative terms are available for 81 languages. The dictionaries used in the application were then enriched with terms from the dictionary in [53]. This is sentiment dictionary, which has been exploited in another sentiment analysis application in R programming language [53] for the Greek language.

In parallel, by reading and manually estimating the texts of the tweets, we collected other useful terms for extending and enriching our dictionaries. We also added words that are politics oriented and have specific emotional content for someone who is familiar with the Greek political scene. For example, the words ‘κούλης’, ‘ζαίος’, 'γερμανοτσολιάς', 'νενέκος' are politically charged words and have been considered to assess the sentiment of each tweet. In other applications that are not political in nature, they would be of no use to anyone.

Finally, various terms have been removed from the already prepared dictionaries on the same grounds. For example, the term 'economy' refers to the sharing of resources at the individual level, while at the state level it refers to the economic activity of countries, organizations, and populations. The dictionaries of my implementation have the following format: (Lemma, Lemma without Greek accents, Term, Term without Greek accents) each lemma and term is followed by itself with the Greek accents. Lemma is a morphological element of Term, which indicates its dictionary form. In Table 9, an example of lemmatization process is presented:

**Table 9 Lemmatization process**

| Term | Lemma |
|---|---|
| Καταδιώκουν | Καταδιώκω |
| Καταγγείλει | Καταγγέλω |
| Κακόβουλα | Κακόβουλος |
| Θεαματική | Θεαματικός |
| Εργάστηκε | Εργάζομαι |
| Περιβάλλοντος | περιβάλλον |

### 4.5.2 Scanning the texts for each dataset

For each row of the dataset, we read the text of the tweet after it has been processed in the preprocessing stage. For the text of each tweet, we create a dictionary (a key/value data structure in python) with which we keep the number of positive and negative terms in the tweet. The text of the tweet is sliced and the tokens that make it up form a list. For each token, we obtain its lemma using the trusted tool of [54]. The NLP toolkit in [54] is based on [55] and is the current most complete implementation for the case of the Greek language. The existing Python libraries for Greek language processing produce several errors in the lemmatization process, which affects the accuracy of our model. Lemmatization is a crucial process, because only by this way is there an equal treatment of words of the same family. For example, the words "χειρότερος" (“worse”) and “χείριστος” (“most worst”) belong to the same word family as the word "bad". So, we do

not treat them separately, but we use their lemma, which in both cases is the word "bad". For each lemma, we check whether it is a member of a sentiment lexicon or not. If a lemma is included in the sentiment lexicon, then the sentiment counter is incremented. The respective process is done with the lexicon of negative terms. Tokens, which do not belong to a sentiment lexicon, are not counted. Below, we present pseudocode of the tweet text processing for tweet_text in tweets:

```
list_of_tokens = split(tweet_text)
sentiment_dictionary = {pos: 0, neg: 0}
for token in list_of_tokens:
    lemma = token.lemma
    if lemma in positive_terms:
        sentiment_dictionary[pos] += 1
    if lemma in negative_terms:
        sentiment_dictionary[neg] += 1
```

At the same time, an effort has been made to handle negation in the text of tweets. The existence of negation affects the emotional content of a text, reversing the polarity of emotionally charged terms. The table below (Table 10) presents illustrative examples of the importance of the correct handling of negation in the case of Sentiment Analysis.

**Table 10 Negation handling examples**

| Text | Sentiment |
|---|---|
| Ο Κυριάκος Μητσοτάκης είναι χαρισματικός<br>Kyriakos Mitsotakis (ND) is charismatic | Positive |
| Ο Κυριάκος Μητσοτάκης είναι μη χαρισματικός<br>Kyriakos Mitsotakis (ND) is not charismatic | Negative |
| Ο υπουργός της κυβέρνησης του ΣΥΡΙΖΑ είναι απατεώνας.<br>The minister of the SYRIZA administration is a con-artist | Negative |
| Ο υπουργός της κυβέρνησης του ΣΥΡΙΖΑ δεν είναι απατεώνας.<br>The minister of the SYRIZA administration is not a con-artist | Positive |
| Η ηγεσία του υπουργείου Οικονομικών χαρακτηρίζεται από αξιοπιστία.<br>The leadership of the Ministry of Finance is characterized by trustworthiness and reliability | Positive |
| Η ηγεσία του υπουργείου Οικονομικών χαρακτηρίζεται από έλλειψη αξιοπιστίας.<br>The leadership of the Ministry of Finance is characterized by lack of trustworthiness and reliability | Negative |

For English and other languages there are Python libraries that detect the existence of negation. For Greek there is no respective tool. For this reason, for each tweet, performed a lexical and syntactic analysis was performed, utilizing Part-of-Speech tags (POS tags), to cover the most frequent cases of negation. For finding the most frequent instances of negation in the tweet text of the datasets, an n-gram analysis of the text of each dataset was performed.

The following characteristic cases of negation were considered:

1.[δε, δεν (not) ] + [VERB] + να (to-finite verb) + [VERB]

Example: δε μπορώ να ωφελήσω

(not able(I) to benefit/help) = “I cannot be of (any) help”

2.[δε, δεν(not)] + [AUXILIARY VERB] + να(to-finite verb) + [VERB]

Example: δε πρέπει να λυπάσαι

(not must(you) to be-sad) = “You must not be sad”

3.[δε, δεν(not)] + [VERB or AUXILIARY VERB] + [ADJECTIVE or ADVERB]
Example: δεν κάνω καλά, δε προσφέρω βοήθεια
(not do(I), not offer(I) help)
= “I am not doing the right thing, I am not offering help”

4.[δε, δεν(not)] + [VERB] + [NOUN]
Example: δε προσφέρω βοήθεια
(not offer(I) help) = “I am not offering help”

5.[δε, δεν(not)] + θα (will-finite verb) + [VERB]
Example: δε θα δυσαρεστήσω
(not will offend/sadden(I) ) = “I will not offend/sadden”

6.[δε, δεν] + [VERB]
Example: δε δυσφορώ
(not feel-annoyed/displeased(I) ) = “I am not feeling annoyed/displeased”

7.μη + [ADJECTIVE]
Example: μη χαρισματικός
(non charismatic) = “non-charismatic”

8.όχι + [ADJECTIVE ή ADVERB]
Example: όχι ωφέλιμος
(no beneficial / no helpful) = “not beneficial/helpful”

9.[NOUN like έλλειψη, αδυναμία] + [NOUN]
Example: έλλειψη ζωντάνιας, αδυναμία νίκης
(lack-of vivacity, lack-of victory) = “Lack of vivacity, lack of victory”

### 4.5.3 Calculation of Sentiment Polarity

Finally, after parsing all the words of the tweet and assessing the existence of negation, the overall sentiment of the tweet is calculated, based on the formula:

$$\boldsymbol{Overall\ polarity} = \frac{\#\boldsymbol{positive\ terms\ found\ in\ the\ text} - \#\boldsymbol{negative\ terms\ found\ in\ the\ text}}{\#\boldsymbol{terms\ that\ make\ up\ the\ text}}$$

- #positive terms found in the text = sum of all positive terms found in the text

- #negative terms found in the text = sum of all negative terms found in the text
- #terms that make up the text = number of tweet's terms before preprocessing

Based on this, three (3) different values are obtained for the results:

- if overall polarity > 0, then the sentiment of the tweet is positive.
- If overall polarity < 0, then the sentiment of the tweet is negative.
- If overall polarity = 0, then the sentiment of the tweet is neutral.

This calculation treats each term as equal in its contribution to the estimation of the sentiment of the text, regardless of its intensity. In other words, for example, although the adjective "worst" in Greek expresses a negative emotion of greater intensity than the adjective "bad" in Greek, it is valued in the same way in the calculation of the overall emotion.

In an extension of the present application, a "weight" factor indicating the intensity of the corresponding emotion could be added. After the sentiment calculation has been implemented for all tweets in a dataset, the total number of positive, negative and neutral tweets is calculated for each political party candidate and output result in a different dataset.

## 4.6 The "one tweet one vote" approach

With the "one tweet one vote" approach, we make the following assumption: each tweet is the product of the publication of a different-separate account. For example, we assume that if we have 30000 tweets, then 30000 different accounts participate in this network. This approach is simpler to implement, as it does not distinguish users based on their activity, but considers them equal. It does not fully correspond to the real life situation and "departs" from reality for reasons that will be explained in a later chapter (see also Section 4.7), but we felt it necessary to incorporate it in the current chapter, as it is an attempt to study the big picture of datasets. Apart from positive/negative/neutral tweets, in our implementation we also use a number of other data for sentiment analysis such as the

number of “likes” and retweets of a tweet, the ratio of positive/negative tweets overall for a political party. Figures 18-21 describe the values for the respective political parties.

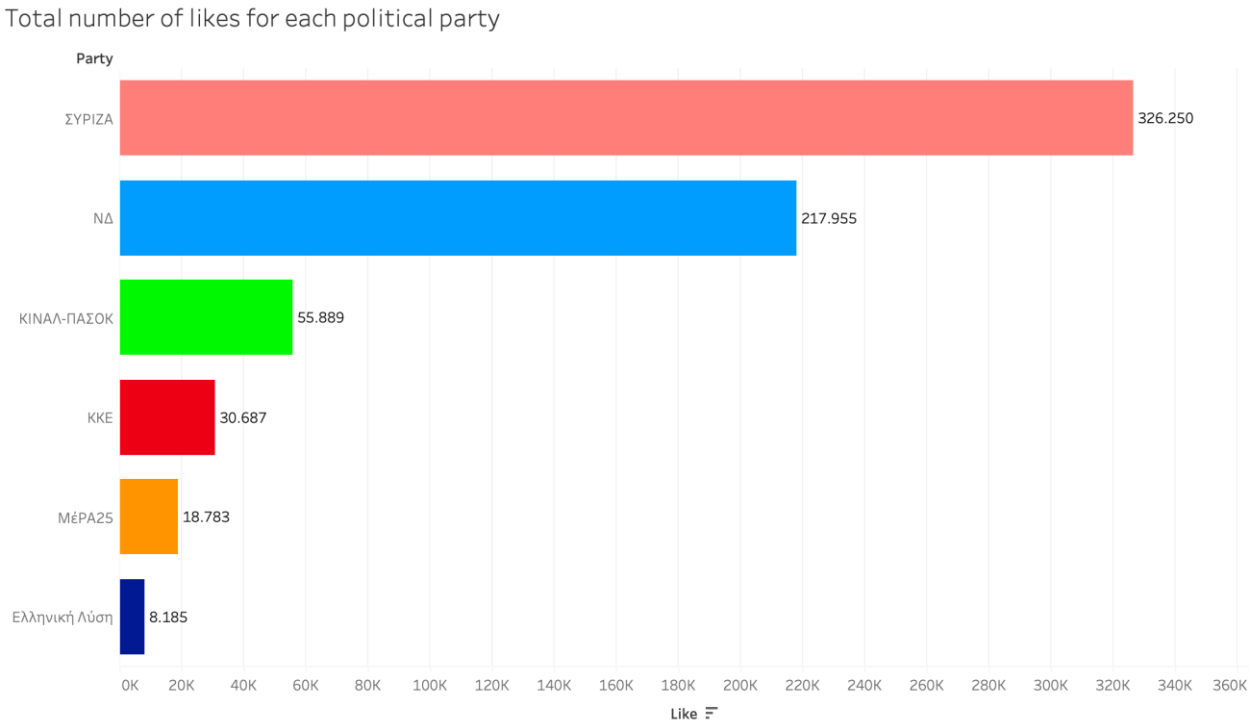



**Figure 18 Total number of “likes” for each political party ["one tweet, one vote"]**

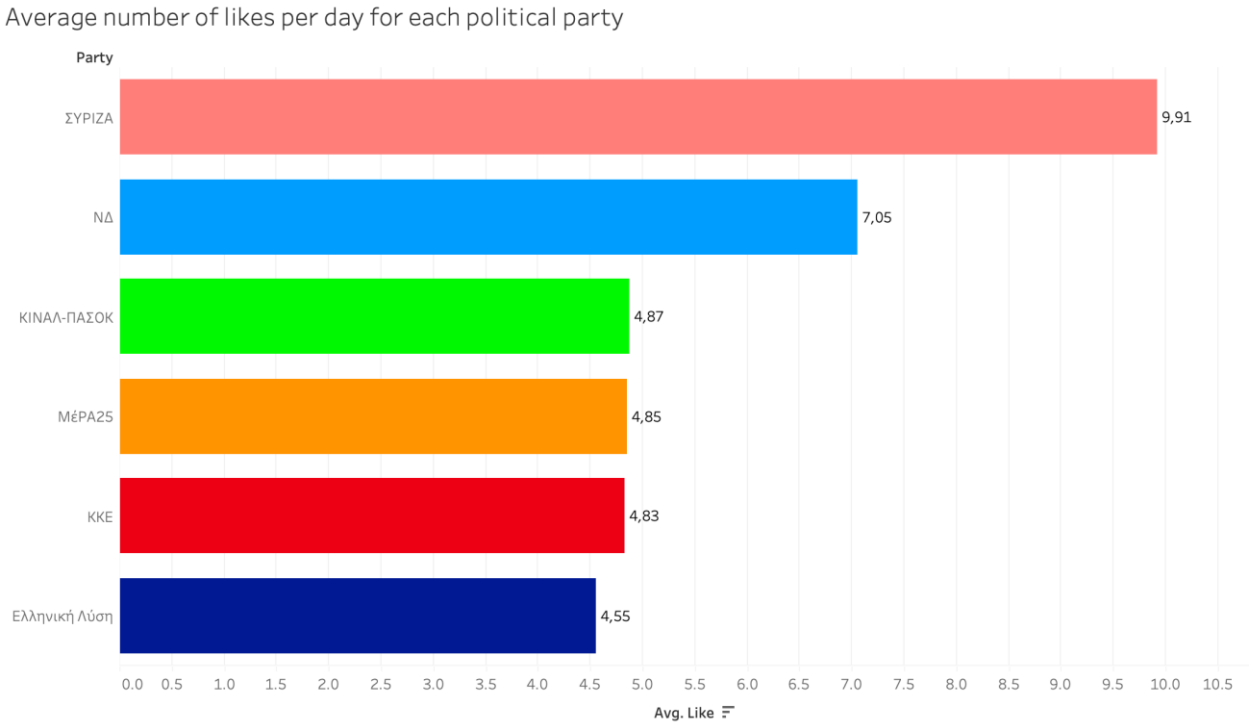



**Figure 19 Average number of “likes” per day for each political party ["one tweet, one vote"]**

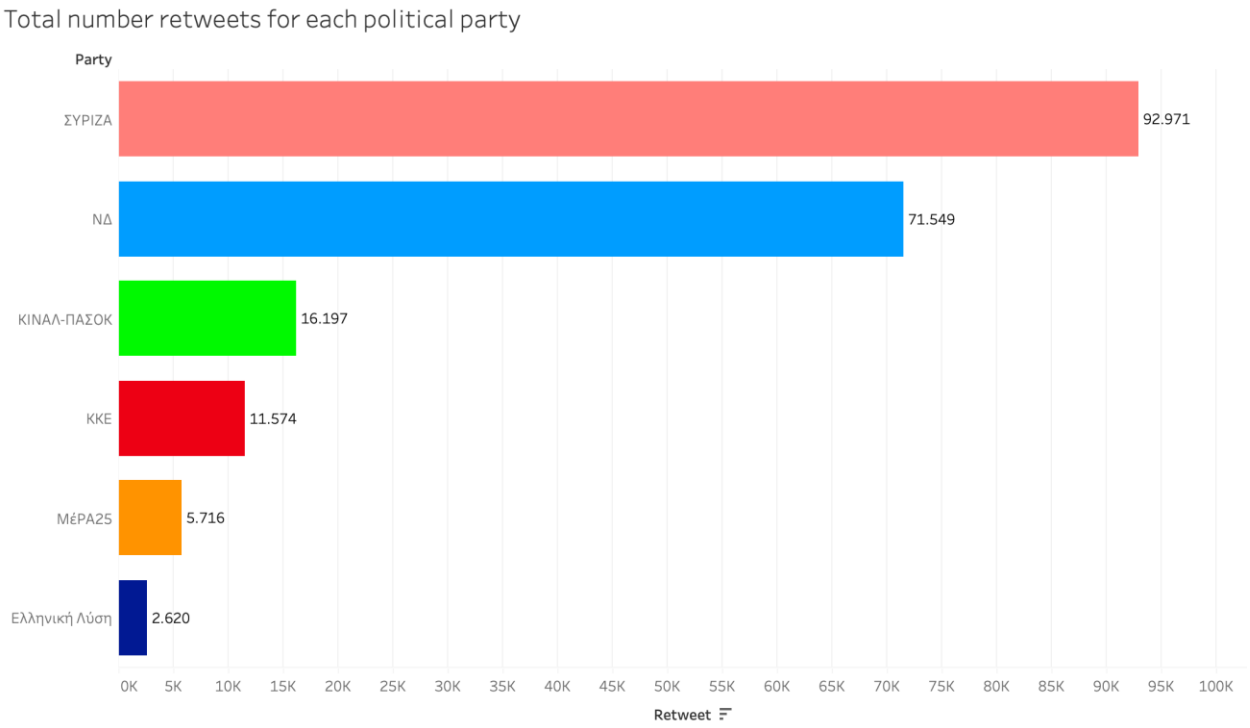



**Figure 20 Total number of retweets for each political party ["one tweet, one vote"]**

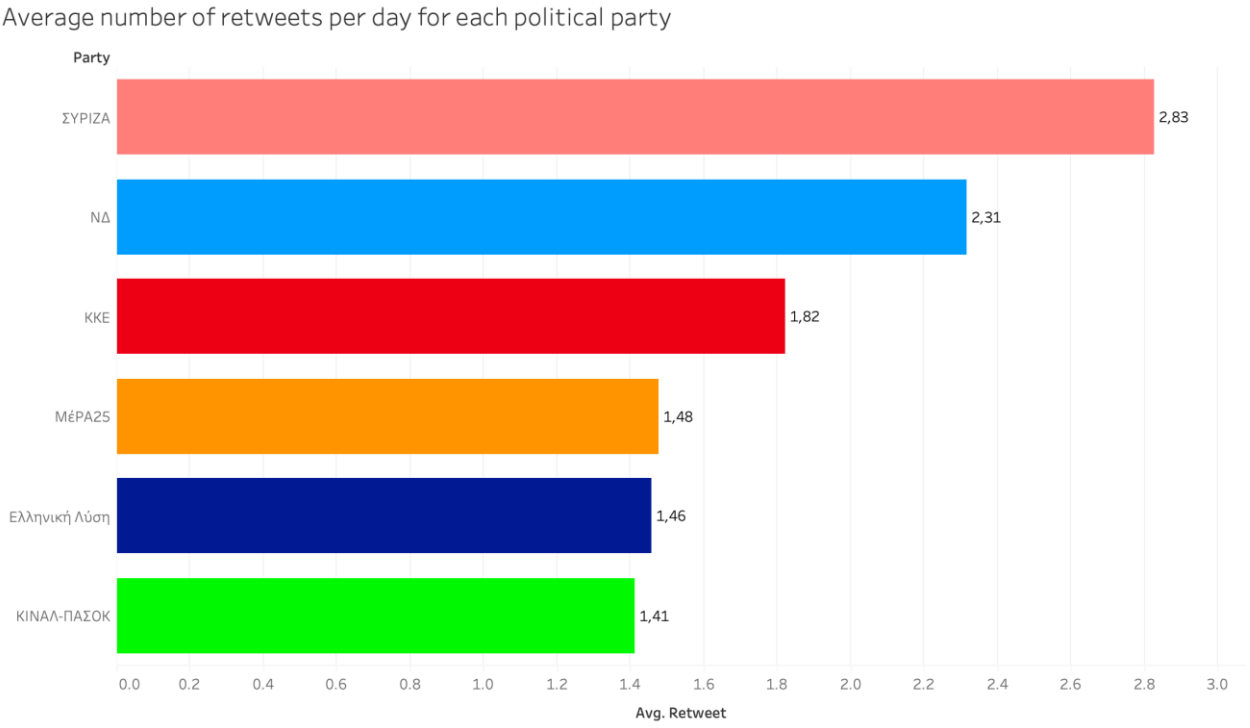



**Figure 21 Average number of retweets per day for each political party ["one tweet, one vote"]**

In the "one tweet, one vote" approach, the case of “likes” and retweets does not fully confirm the outcome of the election. In both cases (total and average), SYRIZA is first in the ranking and with a difference compared to the others, which can be explained by the fact that it was in government before the election result. We are able to conclude that

"likes" and retweets (in absolute number and percentage) are a measure of popularity for those who post the tweets, without specifying its content (whether it is negative, positive

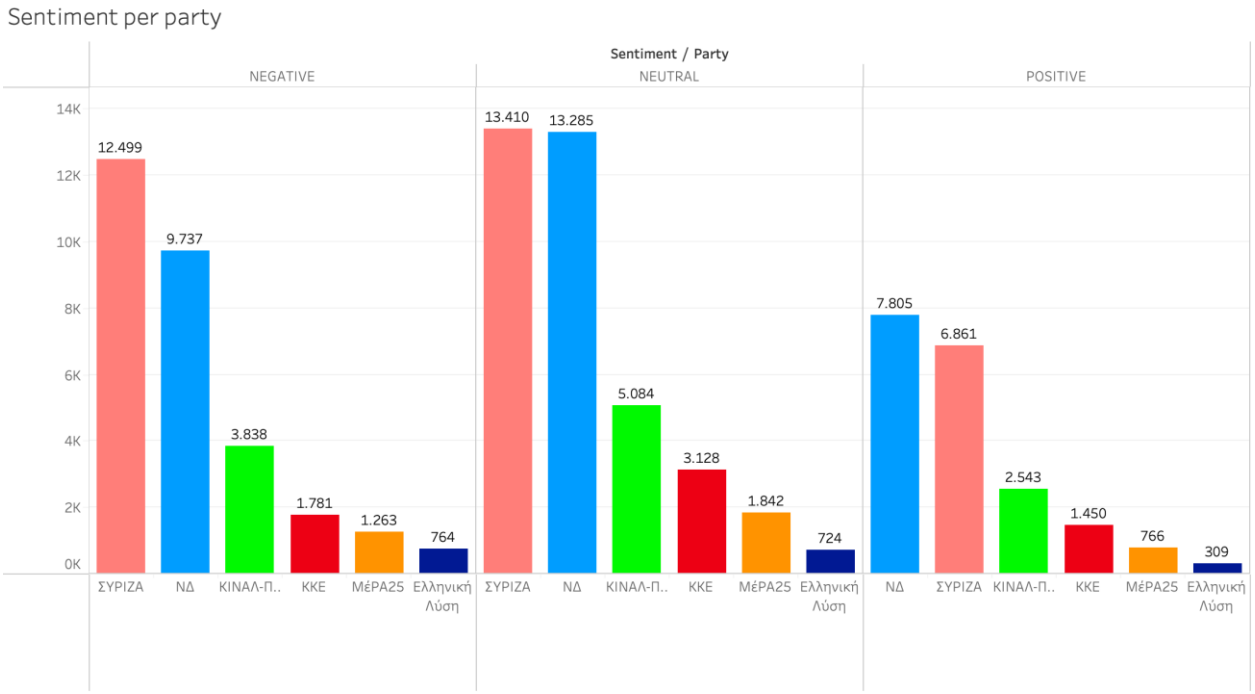



**Figure 22 Sentiment per party ["one tweet, one vote"]**

or neutral). In other words, the more retweets and "likes" the tweets that have a reference to an entity or organization accumulate, the more popular it is within the network. Thus, they can be a safe criterion for detecting and assessing social trends. However, we cannot draw safe conclusions about election results based on these factors.

The graph in Figure 22 shows a complete picture of the overall sentiment of tweets regarding political parties. For this, we need to dwell on some observations-conclusions:

- SYRIZA is the political party that gathers the most tweets of negative sentiment, almost twice as many as positive tweets. This fact is indicative of the social trends expressed in the 2019 general elections. SYRIZA was defeated in the election and came in 2nd place.
- ND has the most positive tweets. This data is indicative of the social trends expressed in the 2019 elections, as ND emerged victorious in 1st place.
- Considering that each positive tweet is considered a positive vote for the respective political party, the sub-graph of positive tweets confirms the ranking in the 2019 election results for 4 first political parties.

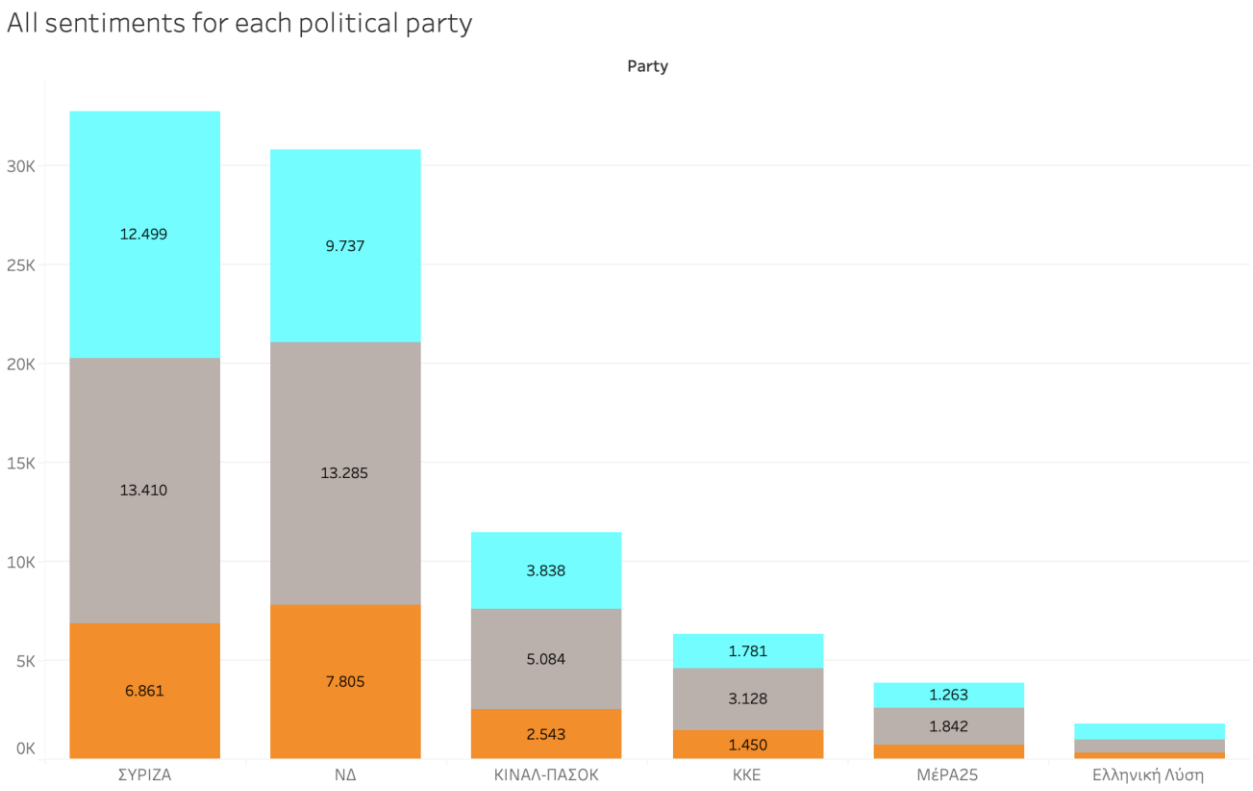



**Figure 23 Total sentiments per day ["one tweet, one vote"]**

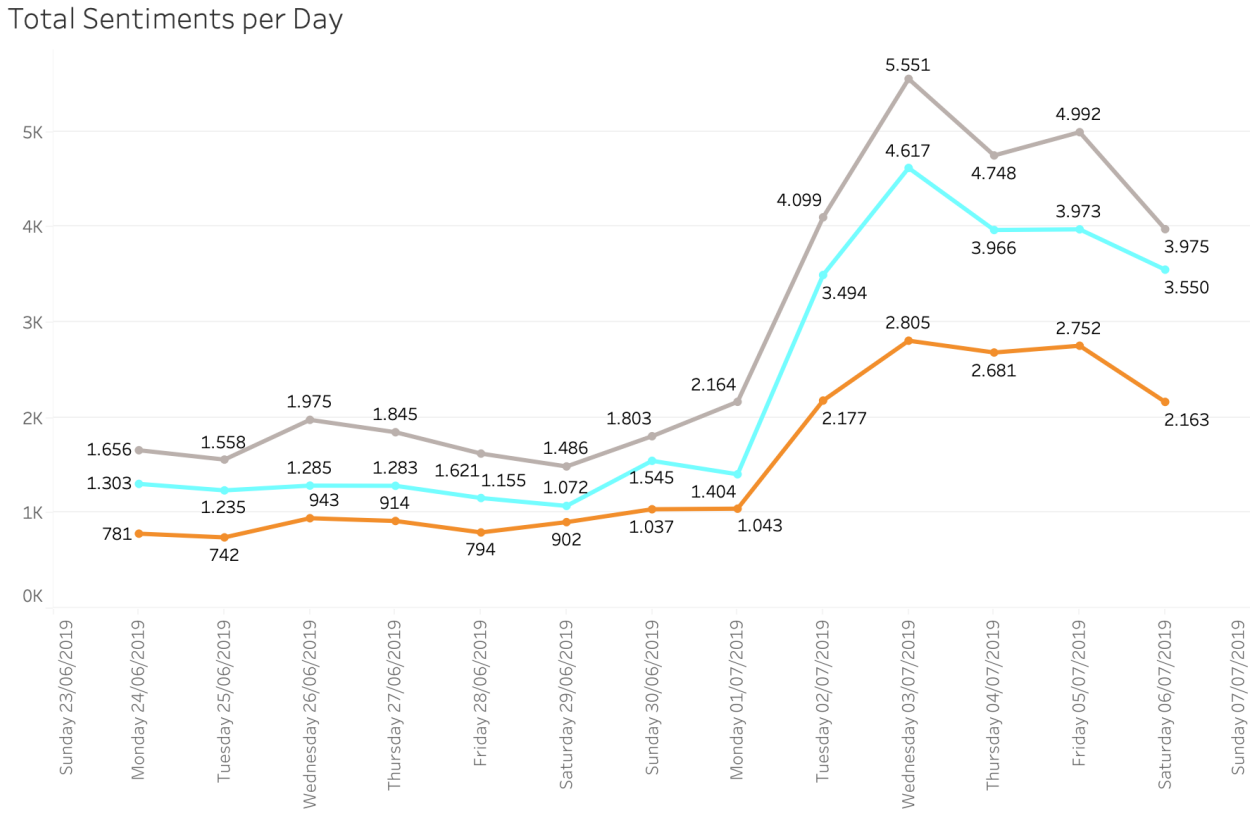



**Figure 24 All sentiments for each political party ["one tweet, one vote"]**

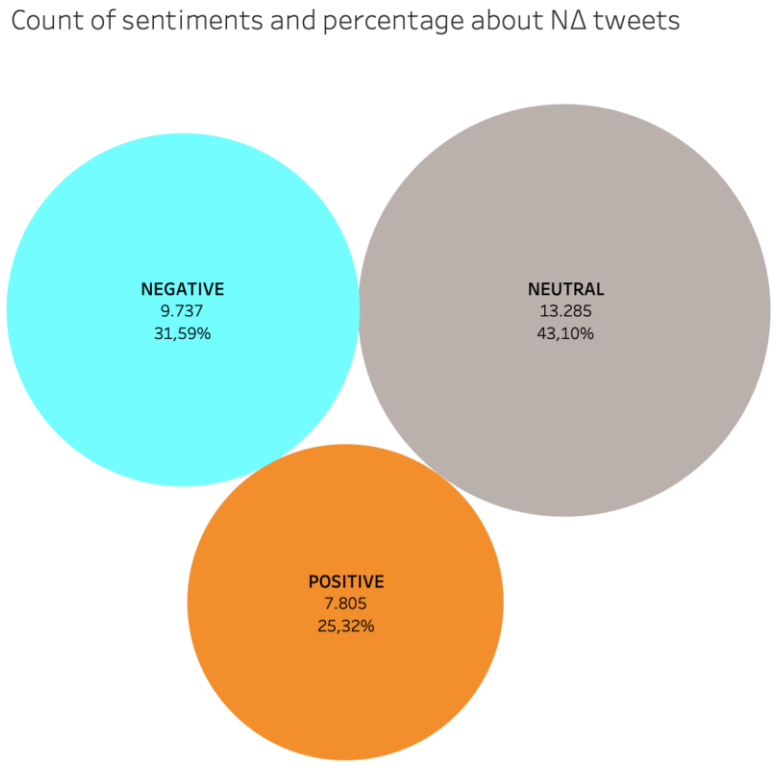



**Figure 25 Count of sentiments and percentage about ND tweets ["one tweet, one vote"]**

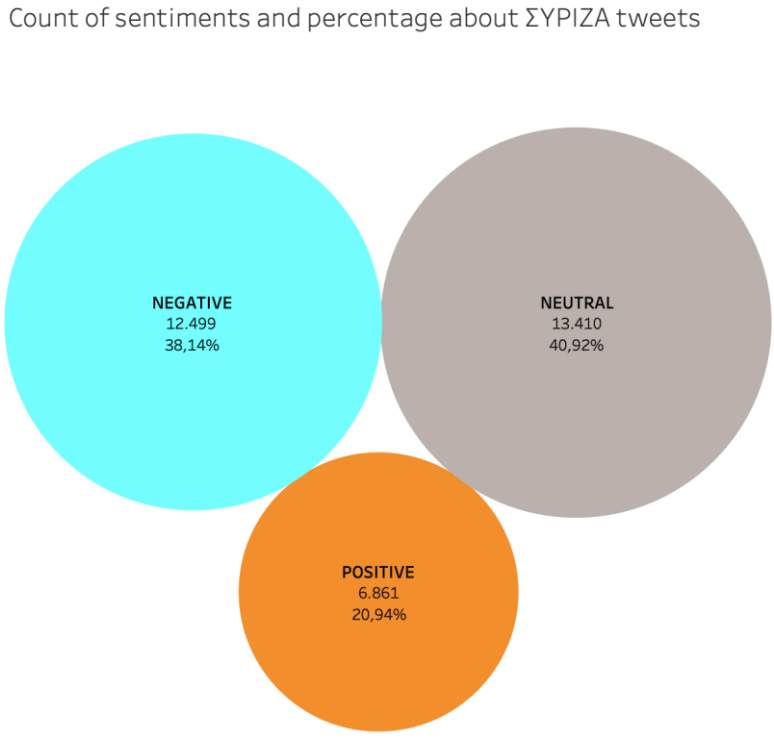



**Figure 26 Count of sentiments and percentage about SYRIZA tweets ["one tweet, one vote"]**

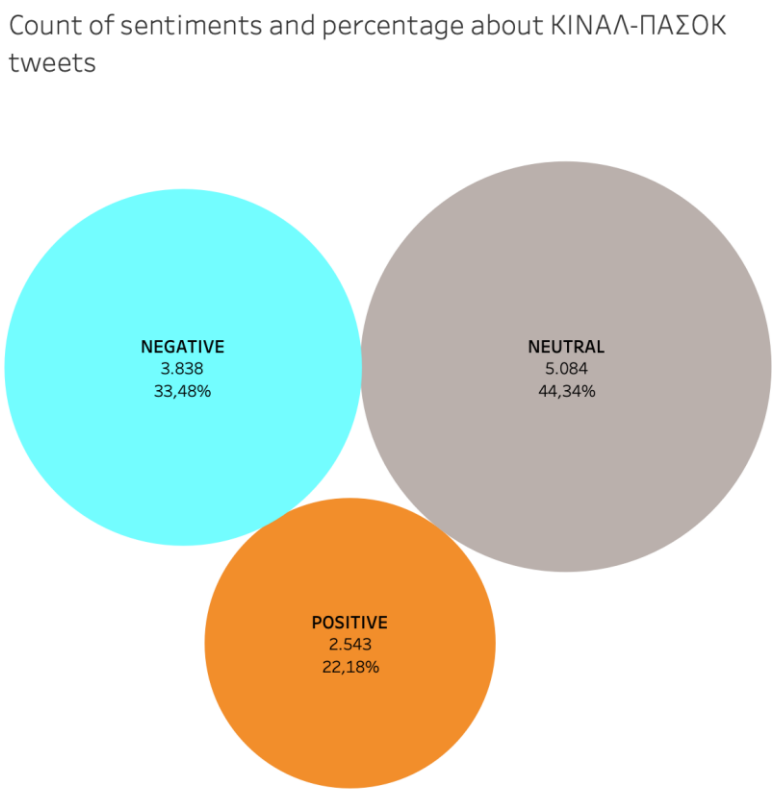



**Figure 27 Count of sentiments and percentage about KINAL-PASOK tweets ["one tweet, one vote"]**

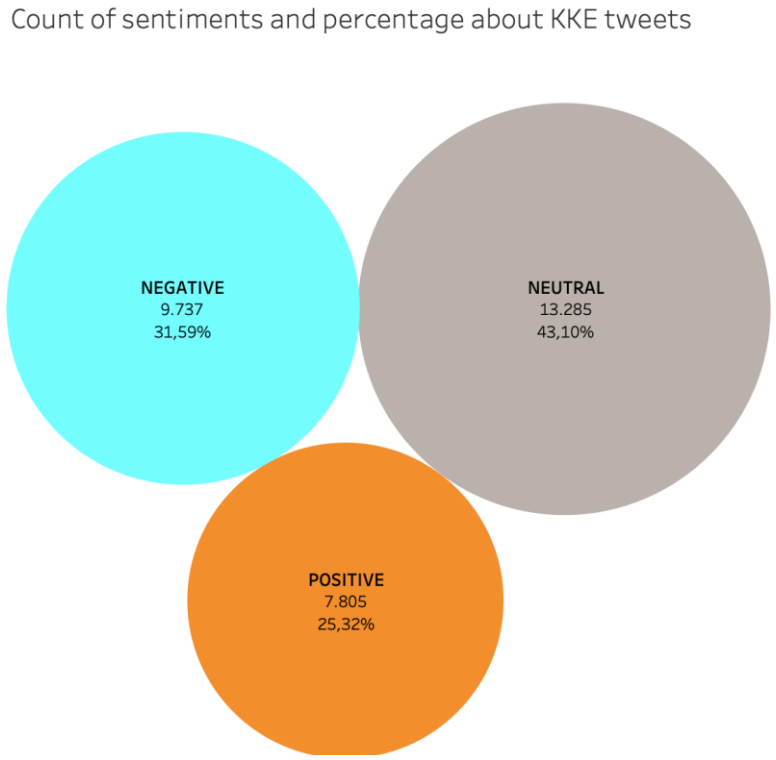



**Figure 28 Count of sentiments and percentage about KKE tweets ["one tweet, one vote"]**

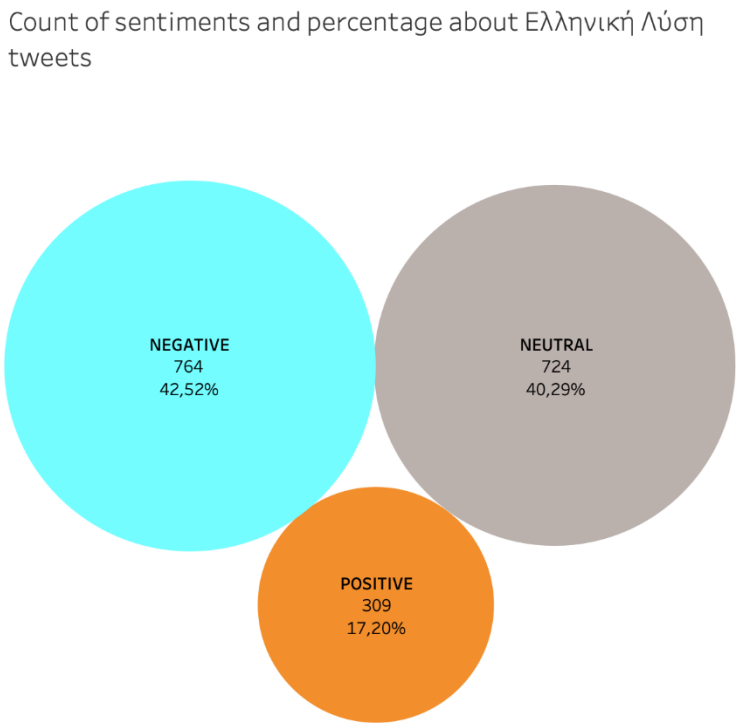



**Figure 29 Count of sentiments and percentage about Elliniki Lysi tweets ["one tweet, one vote"]**

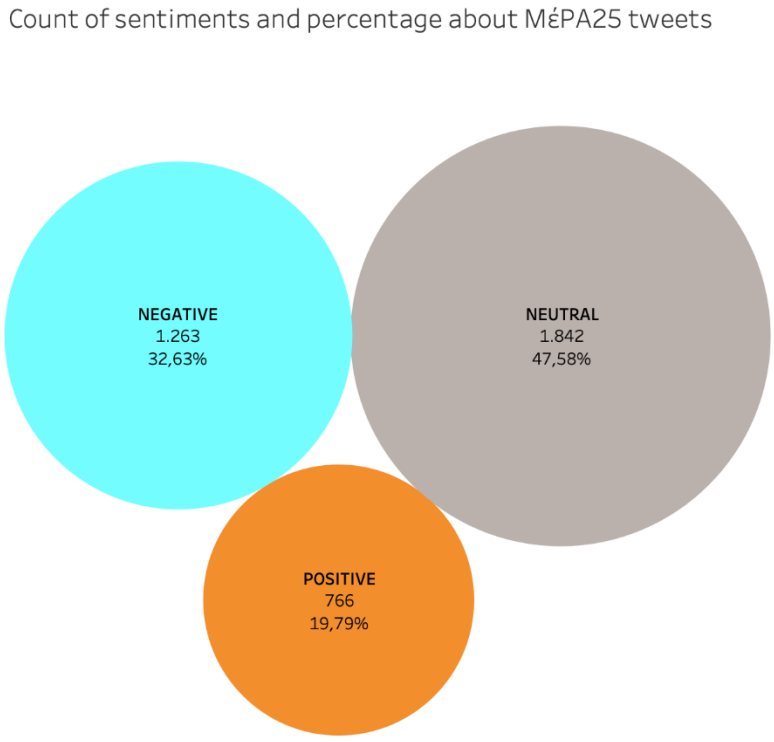



**Figure 30 Count of sentiments and percentage about Mera25 tweets ["one tweet, one vote"]**

**Table 11 Percentages of different sentiments on tweets per political party ["one tweet, one vote"]**

| Political Party | Positive (%) | Negative (%) | Neutral (%) |
|---|---|---|---|
| **ND** | 25,23 | 31,59 | 43,10 |
| **SYRIZA** | 20,94 | 38,14 | 40,92 |
| **KINAL-PASOK** | 22,18 | 33,48 | 44,34 |
| **KKE** | 31,59 | 25,32 | 43,10 |
| **Elliniki Lysi** | 17,2 | 42,52 | 40,29 |
| **Mera25** | 19,79 | 32,63 | 47,58 |

All political parties, regardless of the number of tweets they collect, have a relatively similar quota of positive, negative and neutral tweets (Figure 23). Neutral tweets are first in number and percentage, followed by negative and then positive tweets. Also, their growth throughout the election period, as shown in Figure 24, follows a similar rate. From this data we can draw the following conclusions:

- It is true that a large volume of tweets come from news sites that produce informative content, which is most often of neutral sentiment. Based on this fact, the predominance of neutral tweets for all political parties is justified.
- It is observed, based on the predominance of negative tweets over positive ones, that the use of twitter is more for negative criticism and denunciation of an organization or entity than to endorse and/or support its position or stance. This phenomenon is more pronounced in the debate on twitter about political issues, which can be observed from a close reading of the datasets. The negative tweets are dominated by the mood of criticism and trolling towards the respective political organization or entity.
- The ND and KKE parties have the highest percentage of positive tweets. This fact confirms the result of the elections, as on the one hand ND came out victorious, increasing its electoral percentage, while on the other hand KKE had the smallest percentage losses compared to the other parties.
- The Elliniki Lysi and SYRIZA parties have the highest percentage of negative tweets. In general, on twitter the Elliniki Lysi party does not have a high visibility and appeal, having an audience from social categories that do not use this social media. SYRIZA has a high percentage of negative tweets, which is confirmed by its defeat in the elections and its loss of votes.

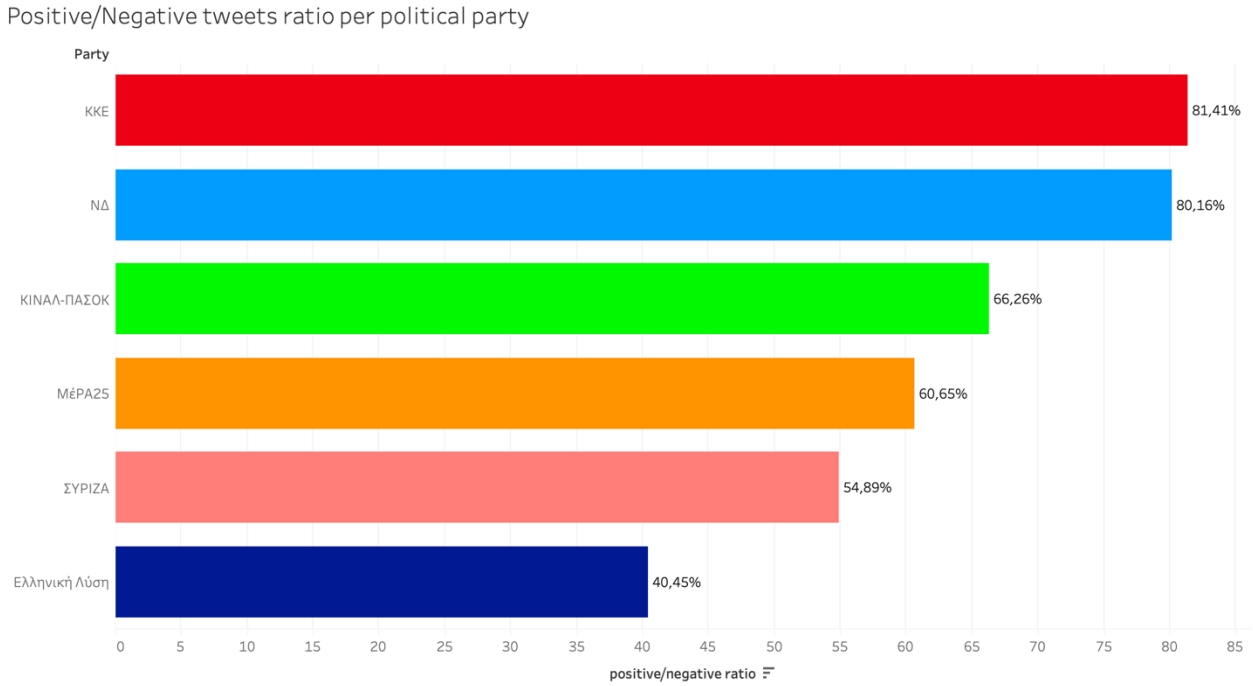



**Figure 31 Positive/Negative tweets ratio per political party ["one tweet, one vote"]**

Another metric that could be used to compare the results is the ratio of positive/negative tweets for each political party, shown in Figure 31. This metric enables us to compare results between datasets that differ in size. We reach the following conclusions-observations:

- The KKE and ND parties have a fairly good positive/negative ratio. For about every 8 positives we have 10 negatives. Clearly, this result once again confirms the social trends expressed in the elections, with ND emerging victorious and the KKE having the smallest losses compared to other parties.
- The SYRIZA party has a fairly low percentage, with only 1 positive tweet for every 2 negative ones. This confirms the result, as this party suffered a defeat and fell to 2nd place.

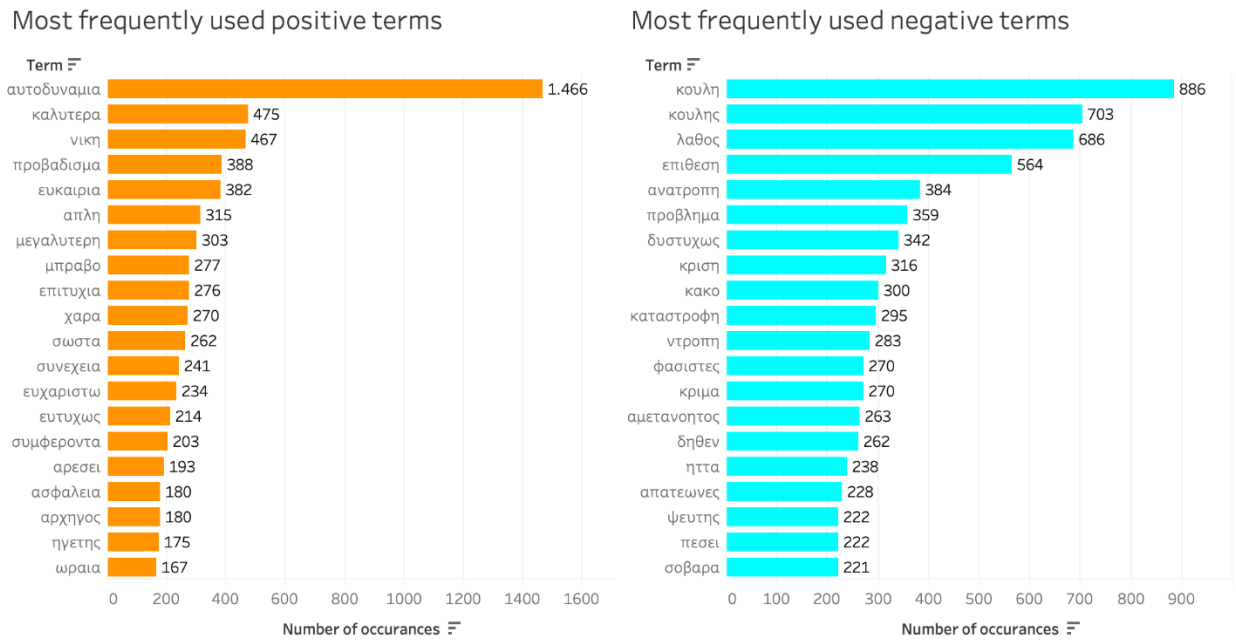



**Figure 32 Most frequently positive/negative terms in project's datasets**

Figure 32 shows the 20 most frequently used positive and negative terms, as derived from the application of our sentiment analysis to the datasets of political parties. We observe that the term "αυτοδυναμία" (absolute parliamentary majority) is far superior in frequency (1466 times) to the positively sentiment-charged terms. This term is a politically charged term, which could have no value in sentiment analysis applications of other content. We also observe something similar in the case of negatively charged terms. In the first two (2) positions, we find the terms "κουλη" and "κουλης", which are negative diminutives of the first name of the then leader of the opposition, Kyriakos Mitsotakis.

## 4.7 Vocal minority vs Silent Minority

Based on the paper [56], social media users do not have equal participation in network platforms. There is a variety of purposes for using social media, hence Twitter, resulting in sets of users with different online behavior. The user group sets produce content at different rates, structures, and characteristics. We distinguish two main groups, based mainly on the rate of content production, the vocal minority, which produces many tweets and is composed of a few users, and the silent majority, which instead produces little or no content, but involves many users.

In terms of the structure of tweets, the vocal minority produces tweets more with links, mentions, hashtags, with the purpose of spreading opinions widely, unlike the silent minority, which has no such purpose. The behavior of the vocal minority is more similar to the behavior of media, political, official or unofficial supporters of political parties. Any attempt to analyze for datasets with such characteristics should consider the different characteristics and different user groups that are objectively shaped by their Twitter usage. For this reason, in this paper, we have performed sentiment analysis in two phases:

1) phase 1 approaches users as equals (i.e. that they produce equal content)
2) and 2) phase 2 approaches users based on the number of tweets they have posted during the election period and separates them into vocal minority and silent majority

For phase 2, a threshold equal to 26 tweets in total has been subjectively selected. This means that, if a user has posted more than 26 tweets in the 13 days of the election period, he/she is included in the vocal minority. Otherwise, he/she is included in the silent majority.

Below, in Figure 33, we observe the distribution of the number of tweets with respect to the number of users. In both phase's 2 approaches, apart from the positive/negative/neutral tweets, in our implementation we use a number of other data for sentiment analysis such as the number of "likes" and retweets of a tweet, the ratio of positive/negative tweets overall for a political party.

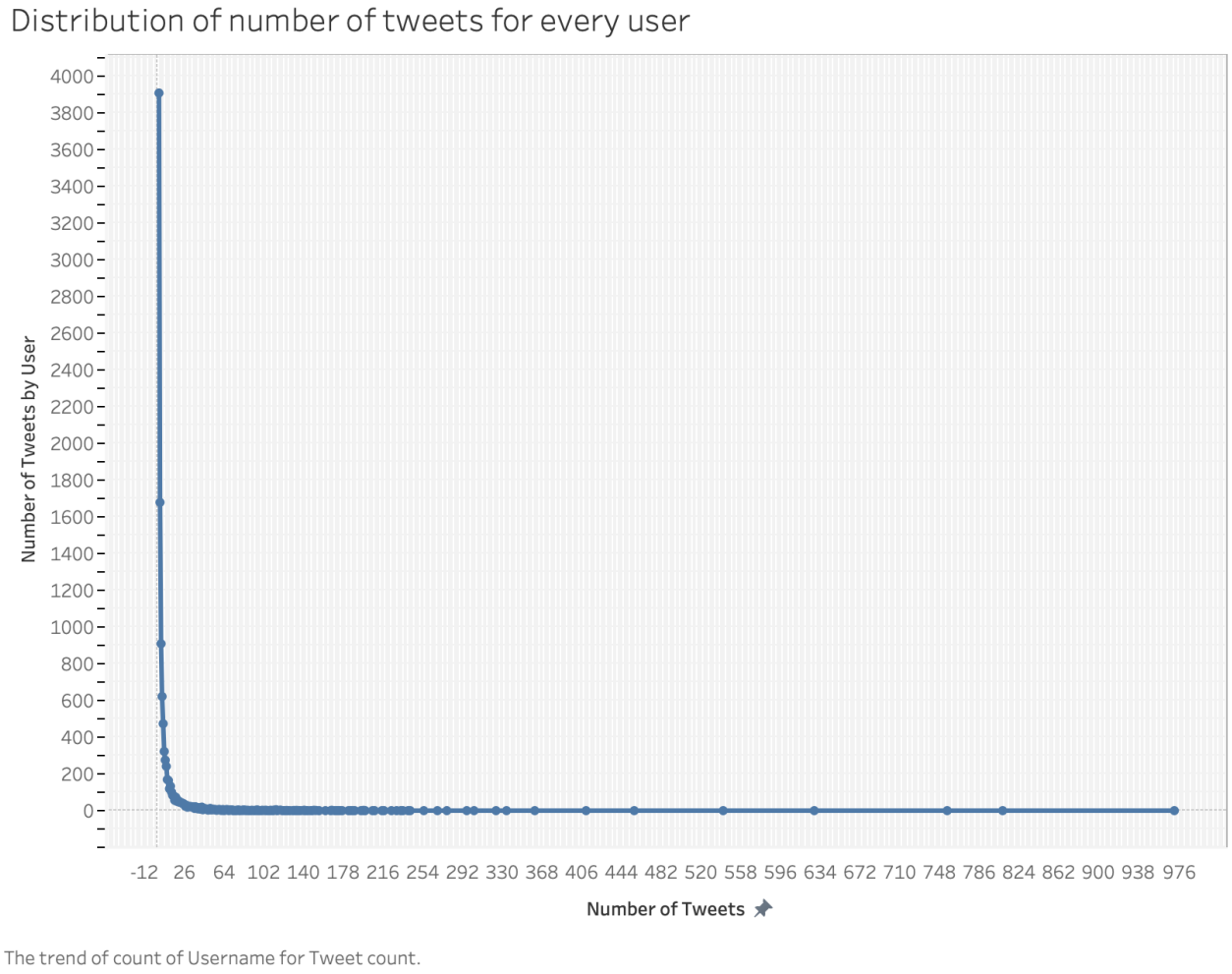



**Figure 33 Distribution of number of tweets for each user**

### 4.7.1 Vocal Minority

In the "vocal minority" approach, we have similar results for political parties to the "one tweet, one vote" approach in both absolute numbers and percentage terms. Once again, SYRIZA manages to come out 1st party in "likes" and retweets, reiterating that these data do not express negative or positive sentiment, but are an indicator of popularity.

Figures 34 - 37 display total and average number of "likes" and retweets for each political party. And in the case of the vocal minority, the SYRIZA party is first in terms of "likes" and retweets, either in absolute number or percentage.

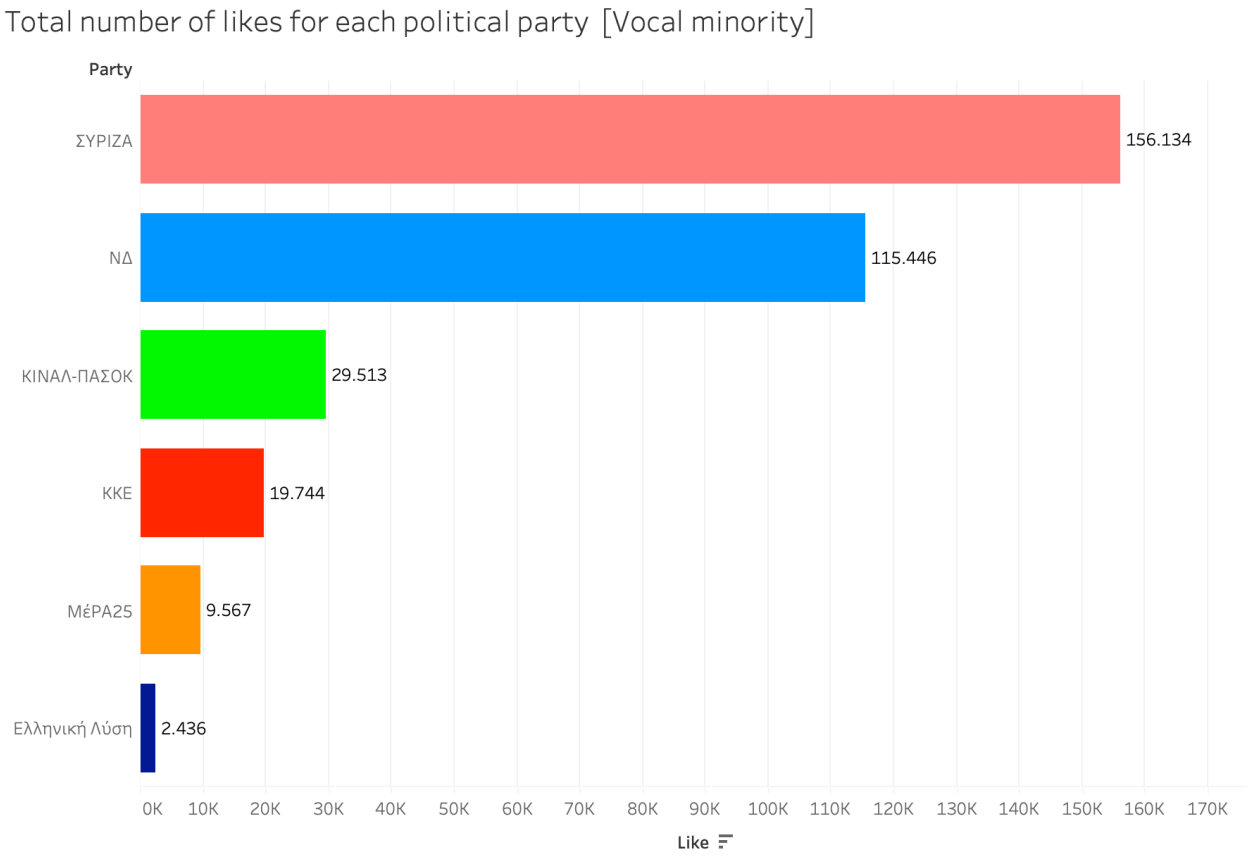



**Figure 34 Total number of "likes" per political party [Vocal minority]**

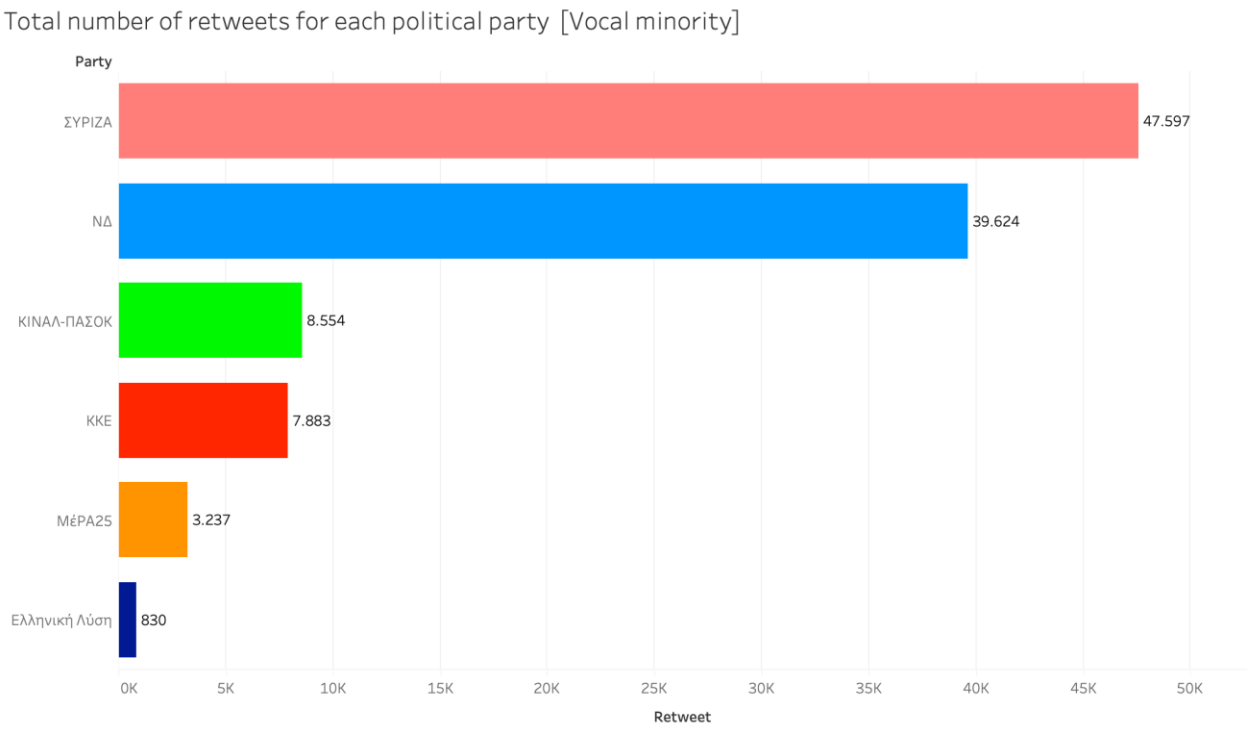



**Figure 35 Total number of retweets for each political party [Vocal minority]**

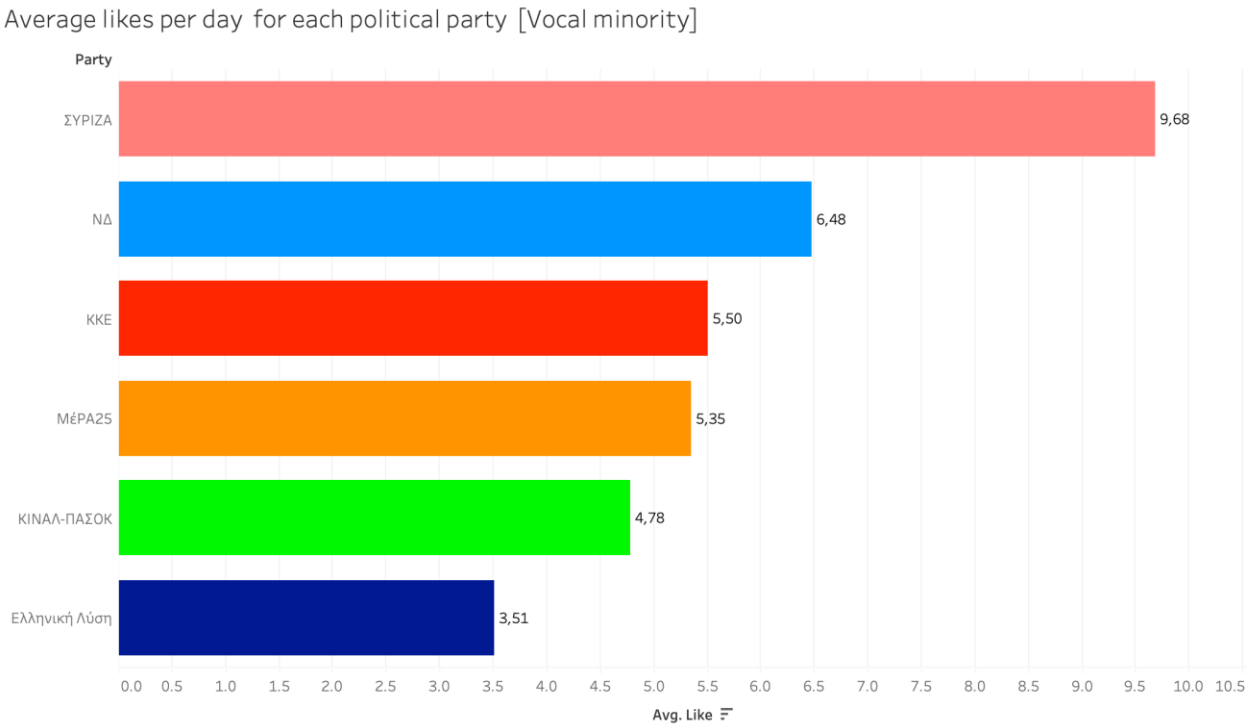



**Figure 36 Average "likes" per day for each political party [Vocal minority]**

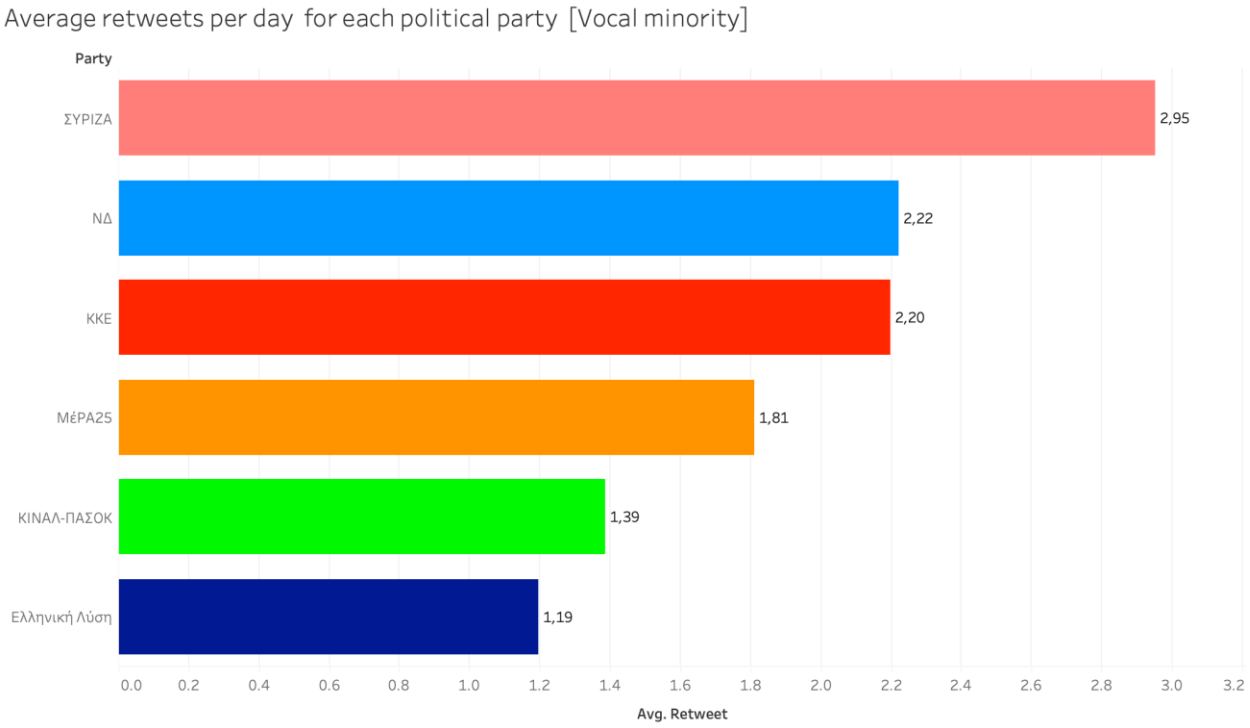



**Figure 37 Average retweets per day for each political party [Vocal minority]**

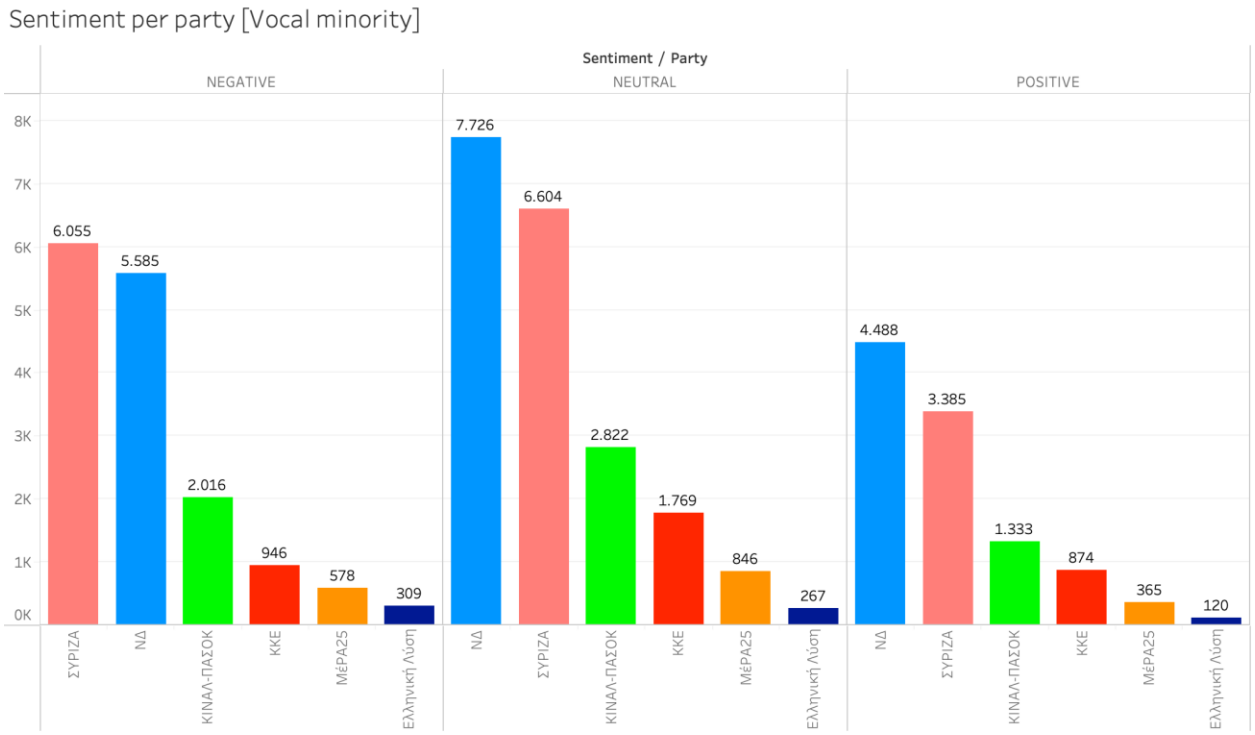


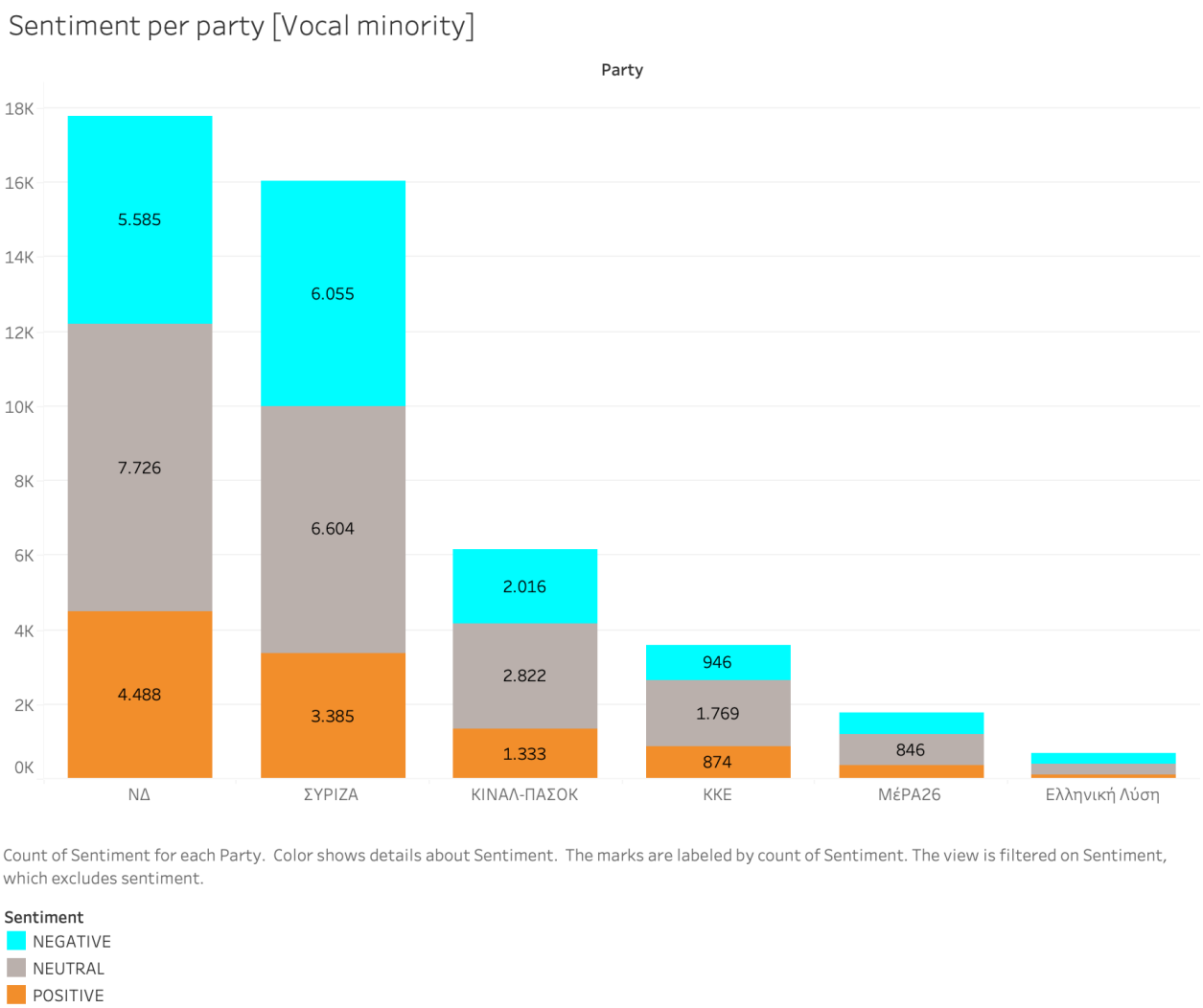

**Figure 38 Sentiment per party [Vocal minority]**

The graph in figure 38 shows a complete picture of the overall sentiment of tweets regarding political parties for the vocal minority approach. For this, we need to dwell on some observations-conclusions:

- The SYRIZA party is in 1st place in the tweets of negative sentiment. However, in this case, it manages to significantly reduce the difference from ND. Specifically, in the "one tweet, one vote" case, the difference between SYRIZA and ND is 2,762 tweets, while in the vocal minority, the difference is reduced to only 470 tweets. Once again, the defeat of SYRIZA in the elections is confirmed.
- The ND party has the most positive tweets, widening the gap with SYRIZA. In the case of the "one tweet, one vote" approach, the difference between ND and SYRIZA is 944 tweets, while in the vocal minority it is 1,103 positive tweets. In this case too, however, the ND's lead in the elections is verified.
- And in the case of the vocal minority, assuming that each positive tweet is considered a positive vote for the respective political party, the sub-signature of positive tweets confirms the ranking in the 2019 election results for 4 first political parties.

All political parties, regardless of the number of tweets they collect, demonstrate a relatively similar quota of positive, negative and neutral tweets and in the case of the "vocal minority" (Figures 41 - 46) . Neutral tweets are first in number and percentage, followed by negative and finally positive tweets in this case as well. Also, the growth of all tweets throughout the election period, as shown in Figure 40, follows a similar rate. From this data we can draw the following conclusions:

- The ND party has a lead also in this case in positive tweets. Next comes the KKE. Once again, the social trend expressed in the elections of 2019 for these political parties is confirmed.




- The Elliniki Lysi and SYRIZA parties are in the top positions in terms of negative tweets, with the first party collecting 44.4% of negative tweets, followed by SYRIZA. This confirms the social trends expressed in the 2019 elections.

And in the case of the "vocal minority", the ranking based on the positive/negative ratio (Figure 47) metric is the same for all political parties using the "one tweet, one vote" approach. We need to make some observations-estimates about the result:

- The KKE party improved its result with an increase to 92.39%, in contrast to the 81.41% of the general approach ("one tweet, one vote"). This result approximates a 1 to 1 ratio of positive/negative tweets.
- The ND party shows a similar percentage with 80.36 %, slightly higher than the general approach.
- The SYRIZA party also shows a low percentage here, with just 1 positive tweet for 2 negative tweets, as in the general approach.

**Figure 39 All sentiments for each political party [Vocal minority]**

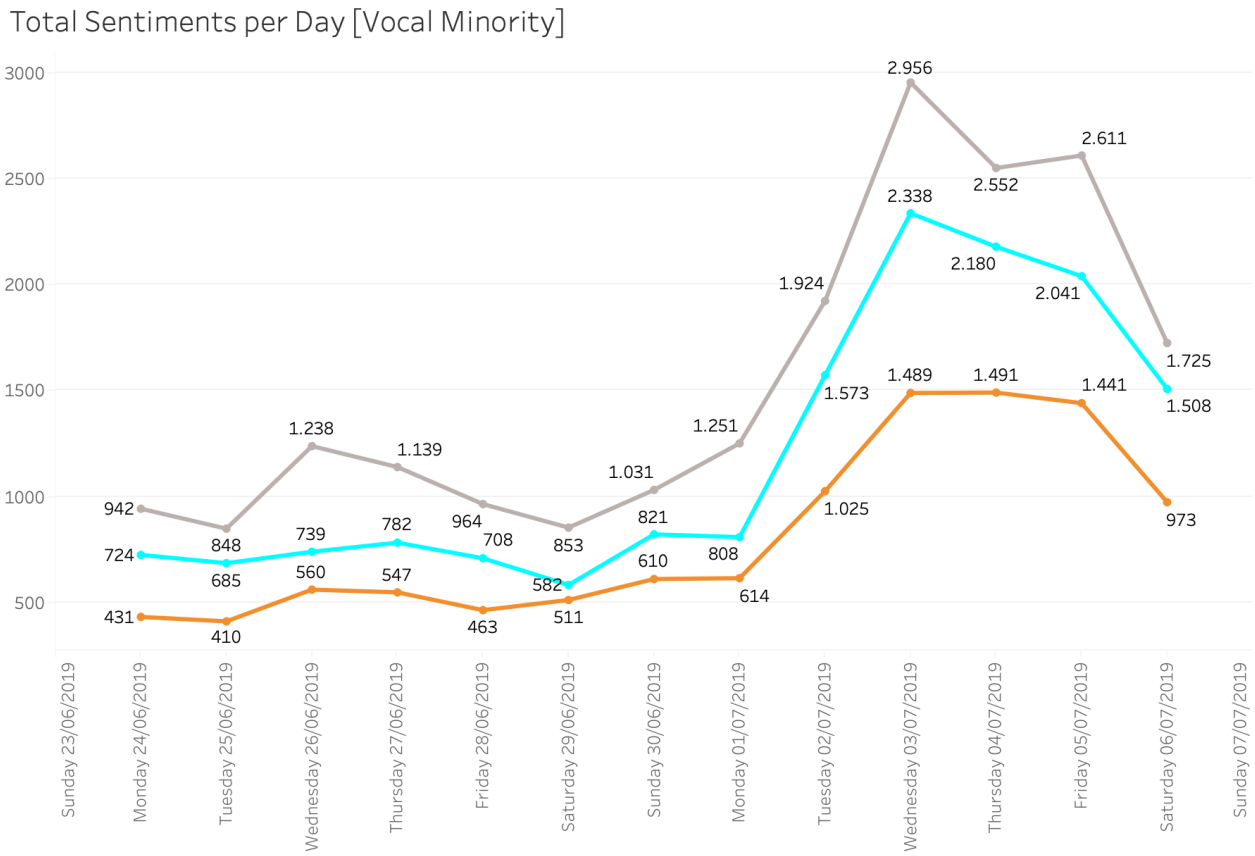



**Figure 40 Total sentiments per day [Vocal minority]**

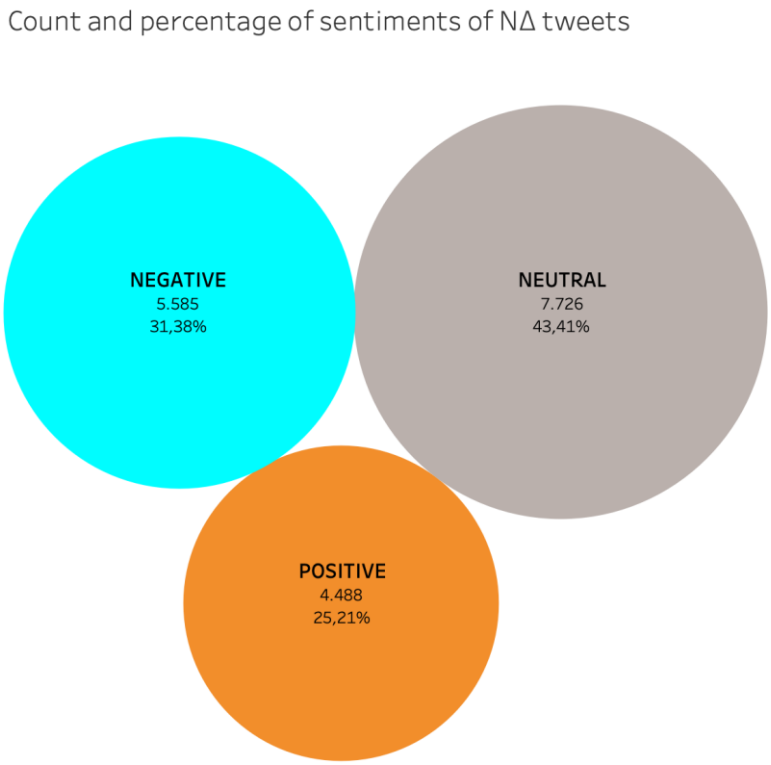



**Figure 41 Count and percentage of sentiments of ND tweets [Vocal minority]**

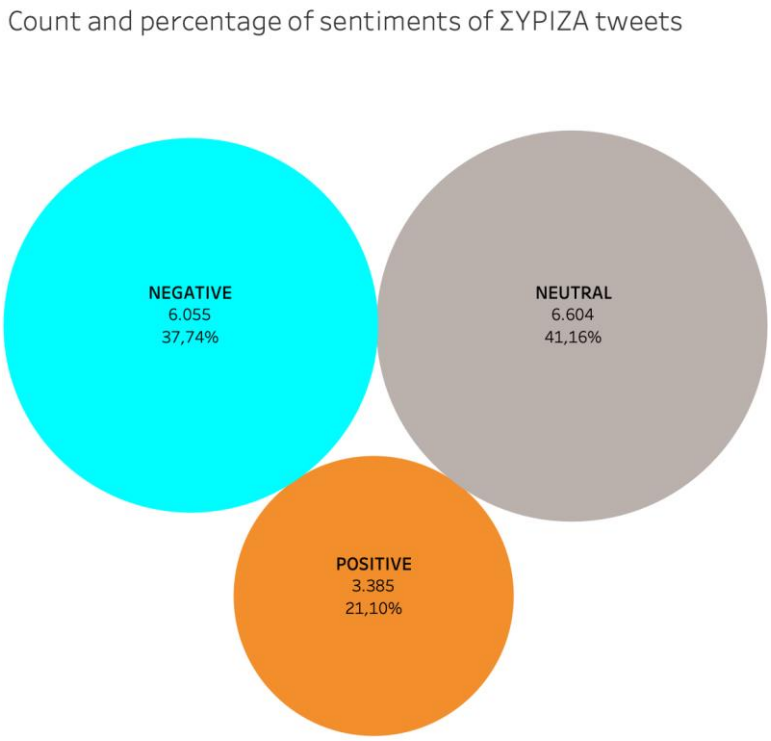



**Figure 42 Count and percentage of sentiments of SYRIZA tweets [Vocal minority]**

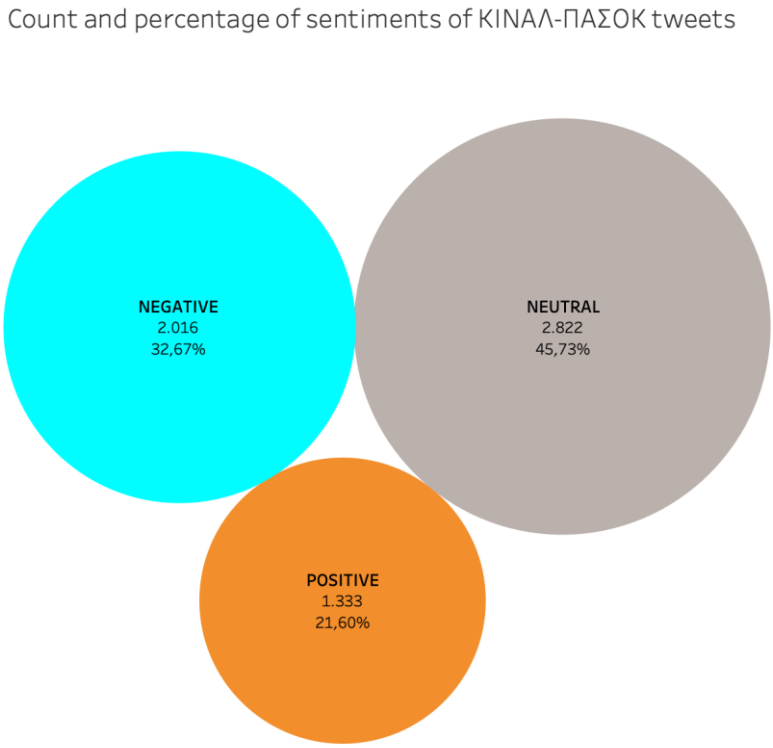



**Figure 43 Count and percentage of sentiments of KINAL-PASOK tweets [Vocal minority]**

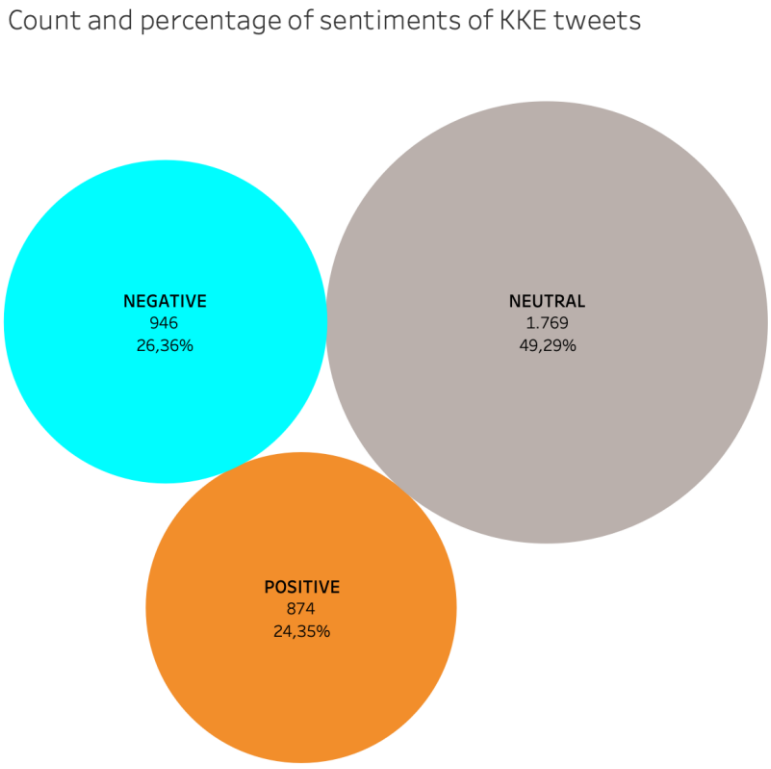



**Figure 44 Count and percentage of sentiments of KKE tweets [Vocal minority]**

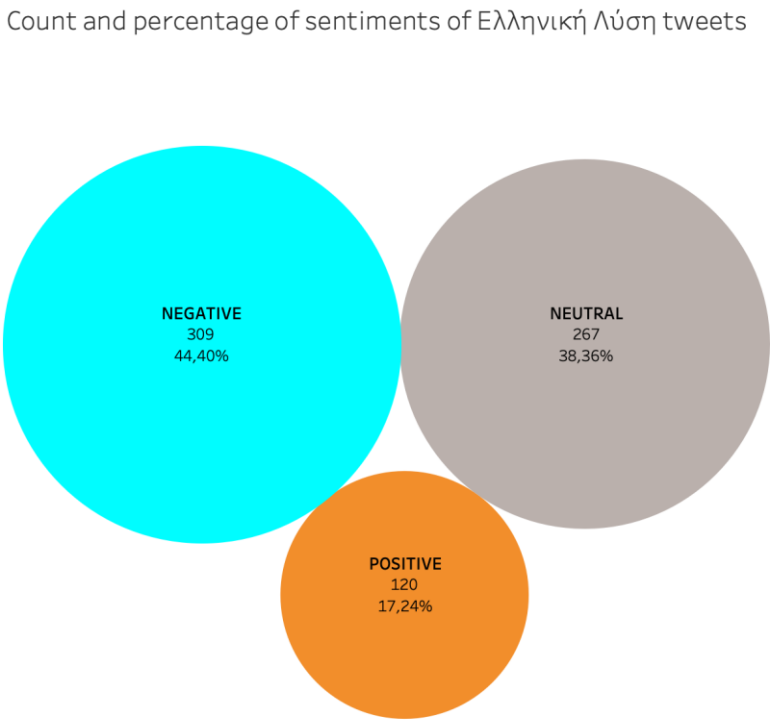



**Figure 45 Count and percentage of sentiments of Elliniki Lysi tweets [Vocal minority]**

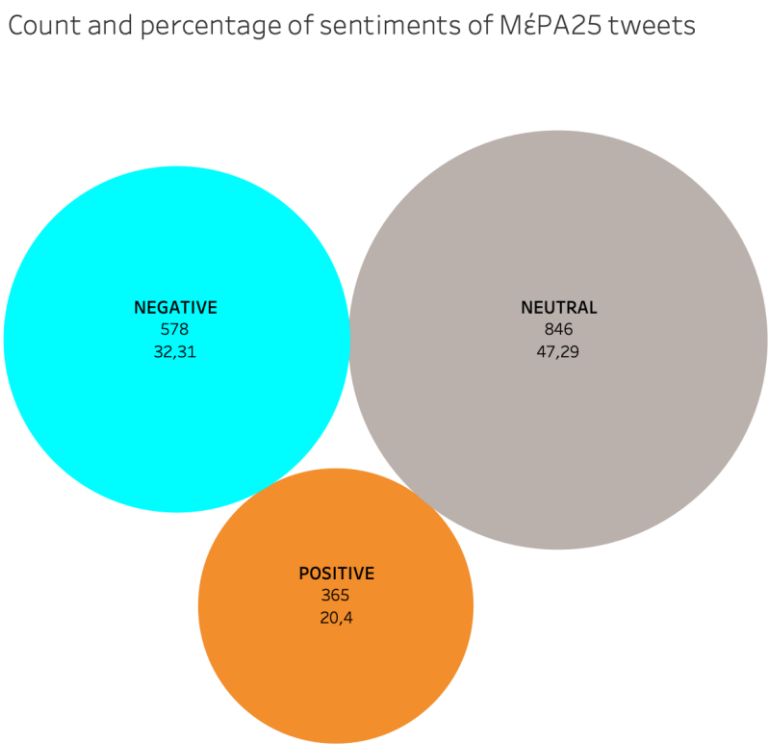



**Figure 46 Count and percentage of sentiments of Mera25 tweets [Vocal minority]**

**Table 12 Percentages of different sentiments on tweets per political party [Vocal minority]**

| Political Party | Positive | Negative | Neutral |
|---|---|---|---|
| **ΝΔ** | 25,21 | 31,38 | 43,41 |
| **ΣΥΡΙΖΑ** | 21,10 | 37,74 | 41,16 |
| **ΚΙΝΑΛ-ΠΑΣΟΚ** | 21,16 | 32,67 | 45,73 |
| **ΚΚΕ** | 24,35 | 26,36 | 49,29 |
| **Ελληνική Λύση** | 17,24 | 44,4 | 38,36 |
| **ΜέΡΑ25** | 20,4 | 32,31 | 47,29 |

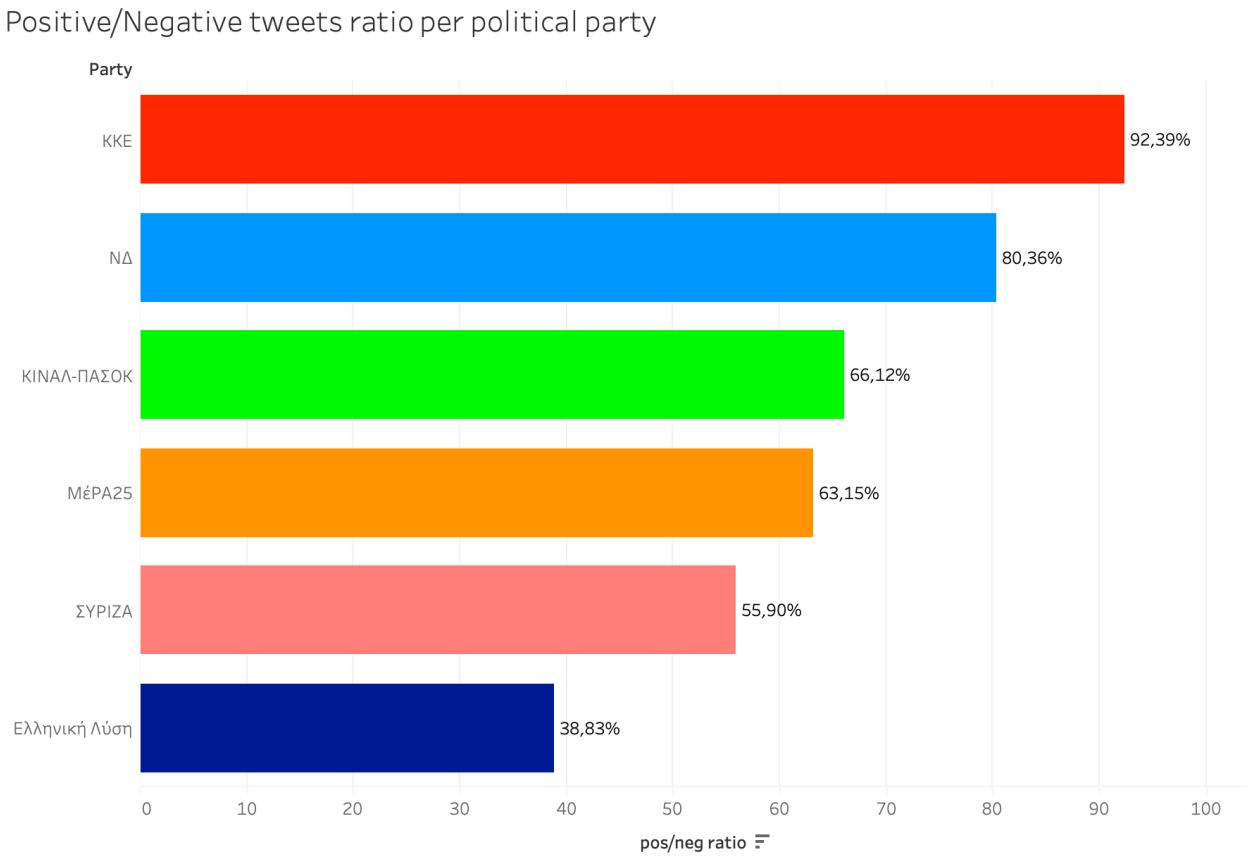



**Figure 47 Positive/Negative tweets ratio per political party [Vocal minority]**

### 4.7.2 Silent majority

In the case of the "silent majority", for the data of “likes” and retweets, we have the same ranking as the other two approaches (general, vocal minority). Once again, SYRIZA manages to come out 1st party in “likes” and retweets, reiterating that these data do not express negative or positive sentiment, but are an indicator of popularity.

Figures 48 – 51 display total and average number of "likes" and retweets for each political party. The SYRIZA party is first in terms of "likes" and retweets, either in absolute number or percentage in the case of “silent majority”.

The graph in Figure 52 shows a complete picture of the overall sentiment of tweets regarding political parties for the silent majority approach. For this, we need to dwell on some observations-conclusions:

- The SYRIZA party is in 1st place in the tweets of negative sentiment. In this case, however, the difference increases by far compared to the case of the "vocal minority". In particular, it reaches 2,292 negative tweets, while in the case of the "vocal minority", the difference between SYRIZA and ND was 470 negative tweets.
- The SYRIZA party manages to take the lead in positive tweets, reversing what we have observed in the other 2 approach cases. It manages to get a difference of 159 positive tweets.

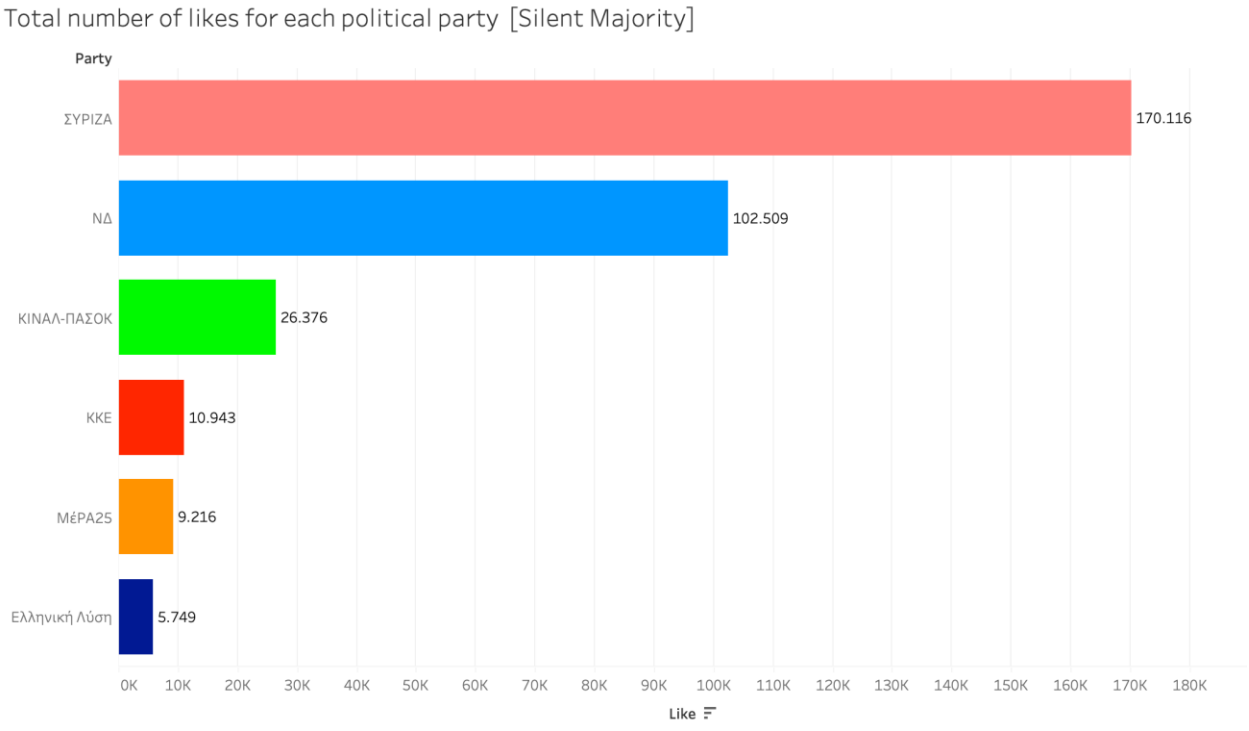



**Figure 48 Total number of “likes” for each political party [Silent majority]**

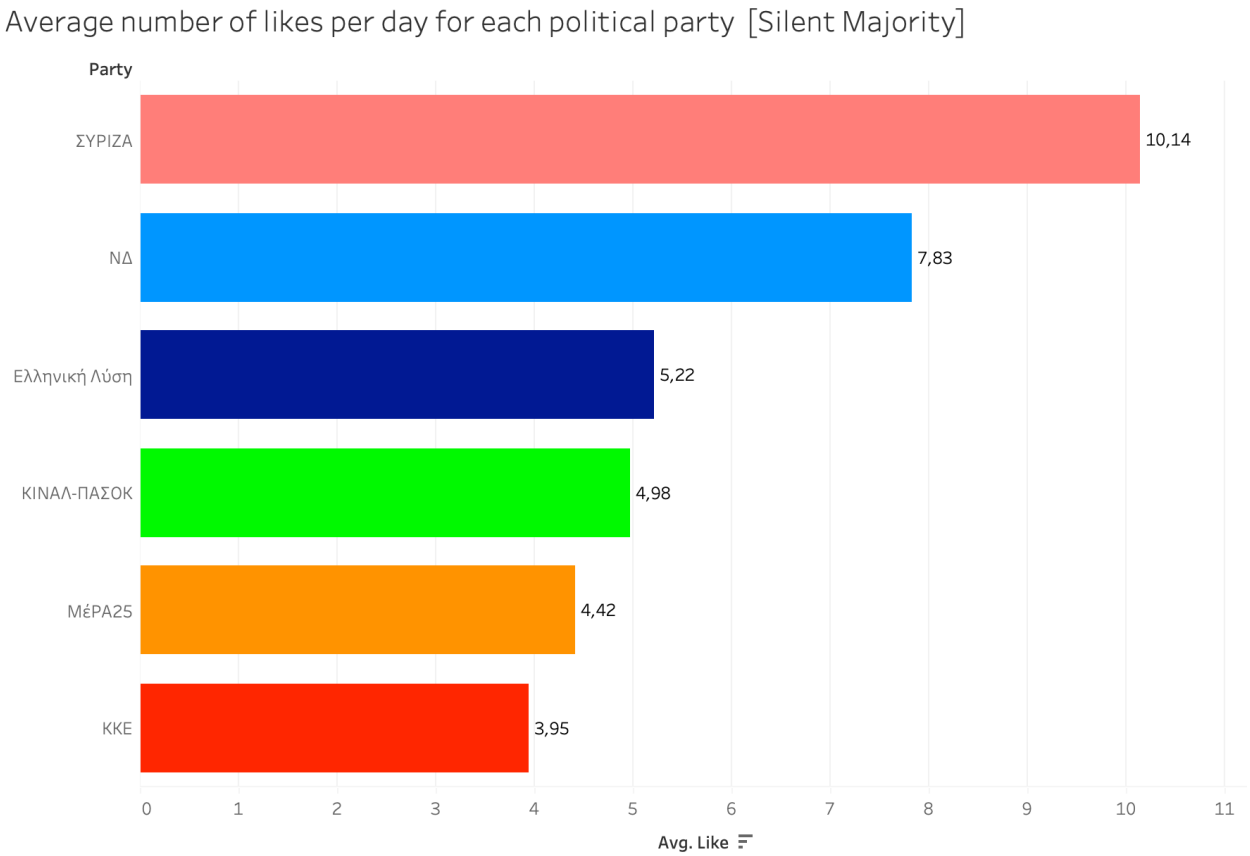



**Figure 49 Average number of "likes" per day for each political party [Silent majority]**

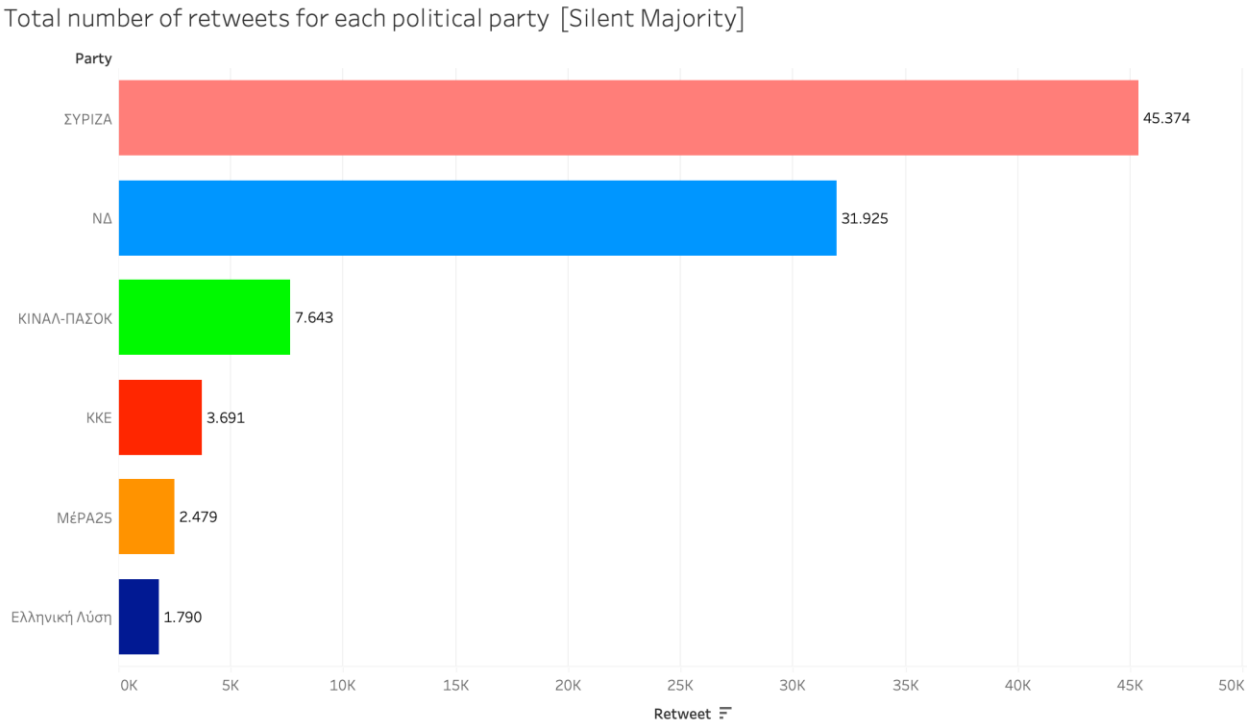



**Figure 50 Total number of retweets for each political party [Silent majority]**

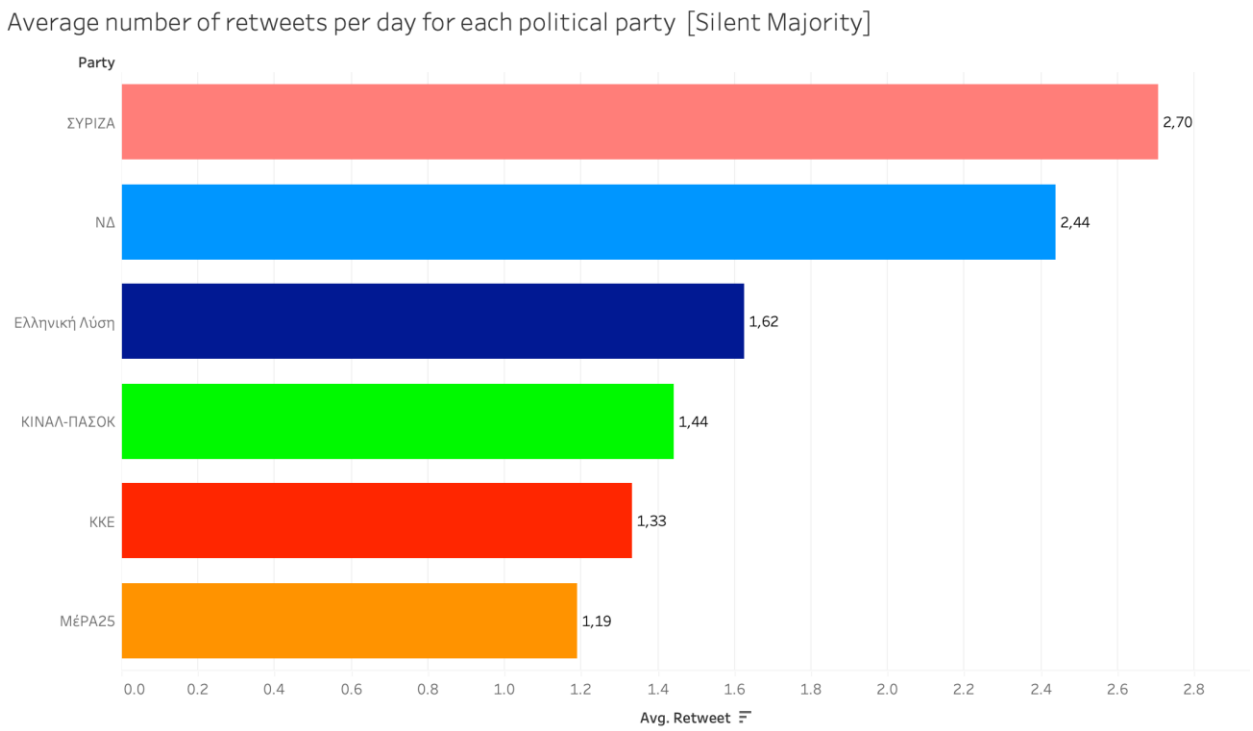



**Figure 51 Average number of retweets per day for each political party [Silent majority]**

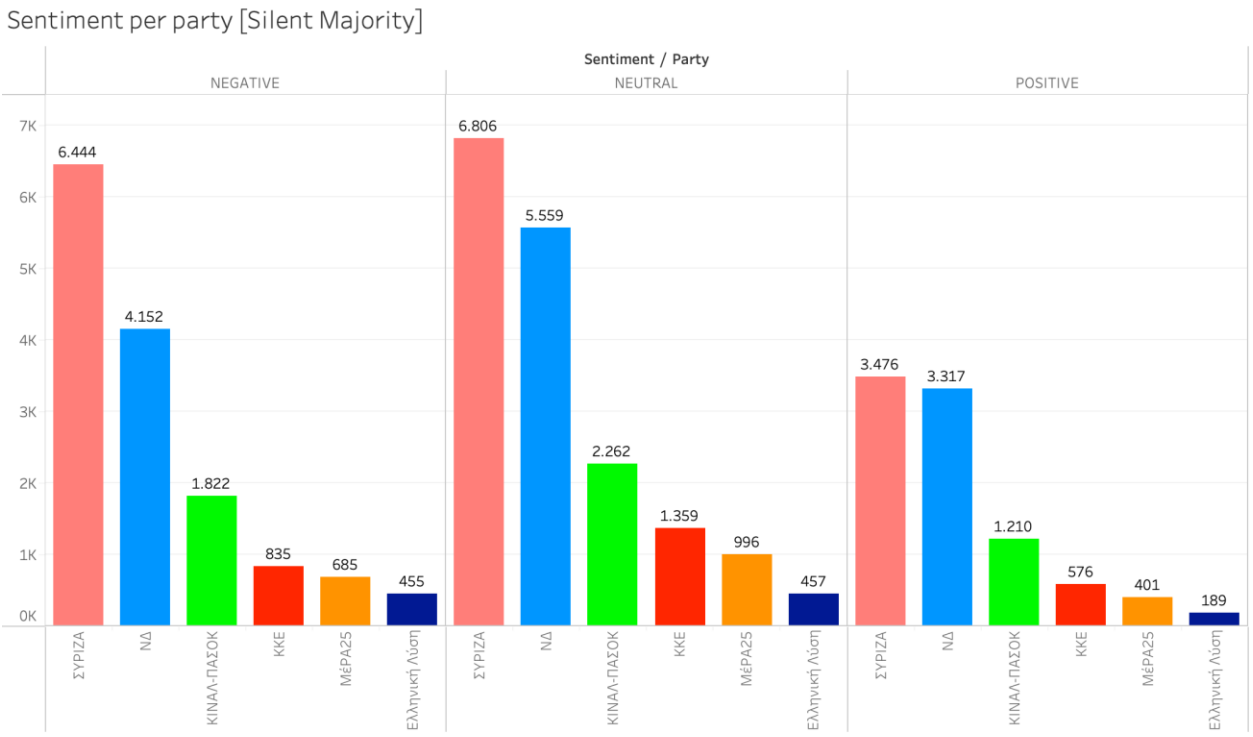



**Figure 52 Sentiment per party [Silent majority]**

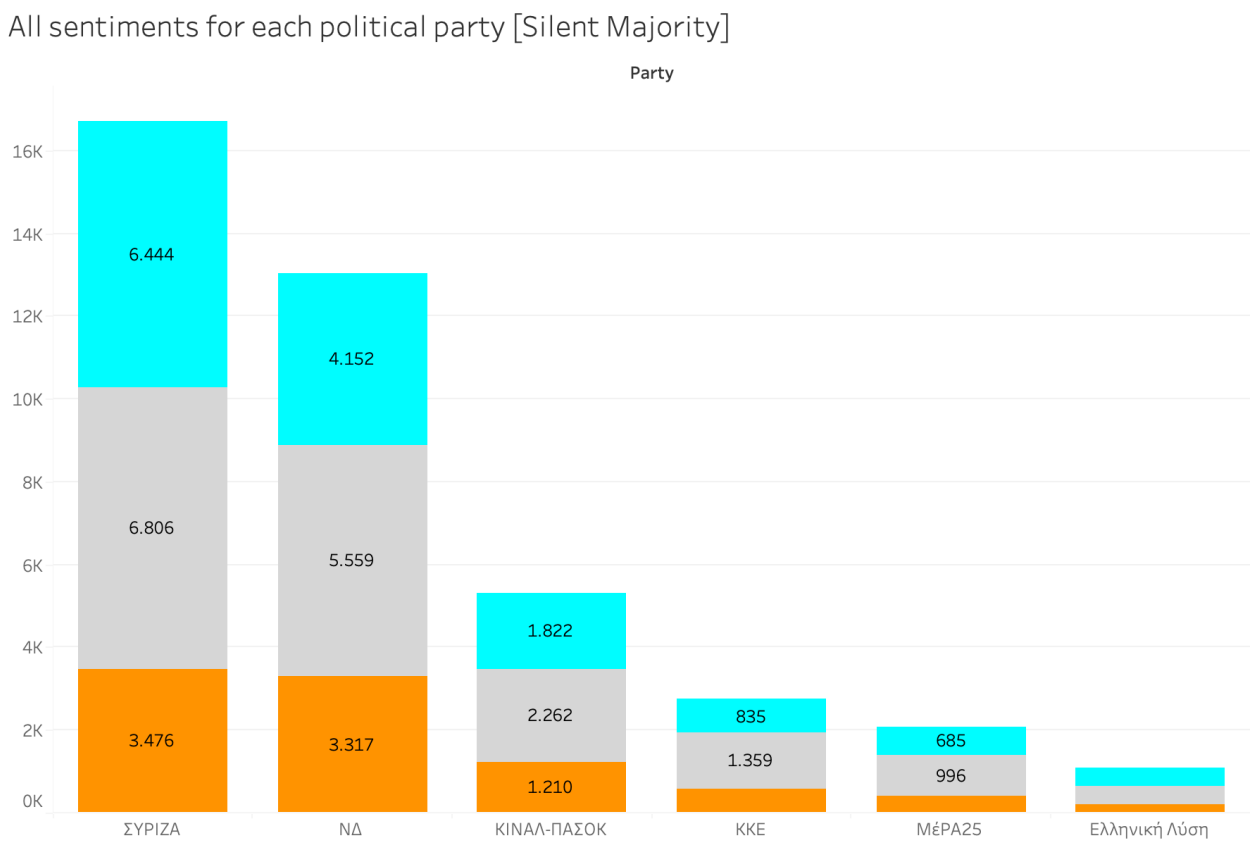



**Figure 53 All sentiments for each political party [Silent majority]**

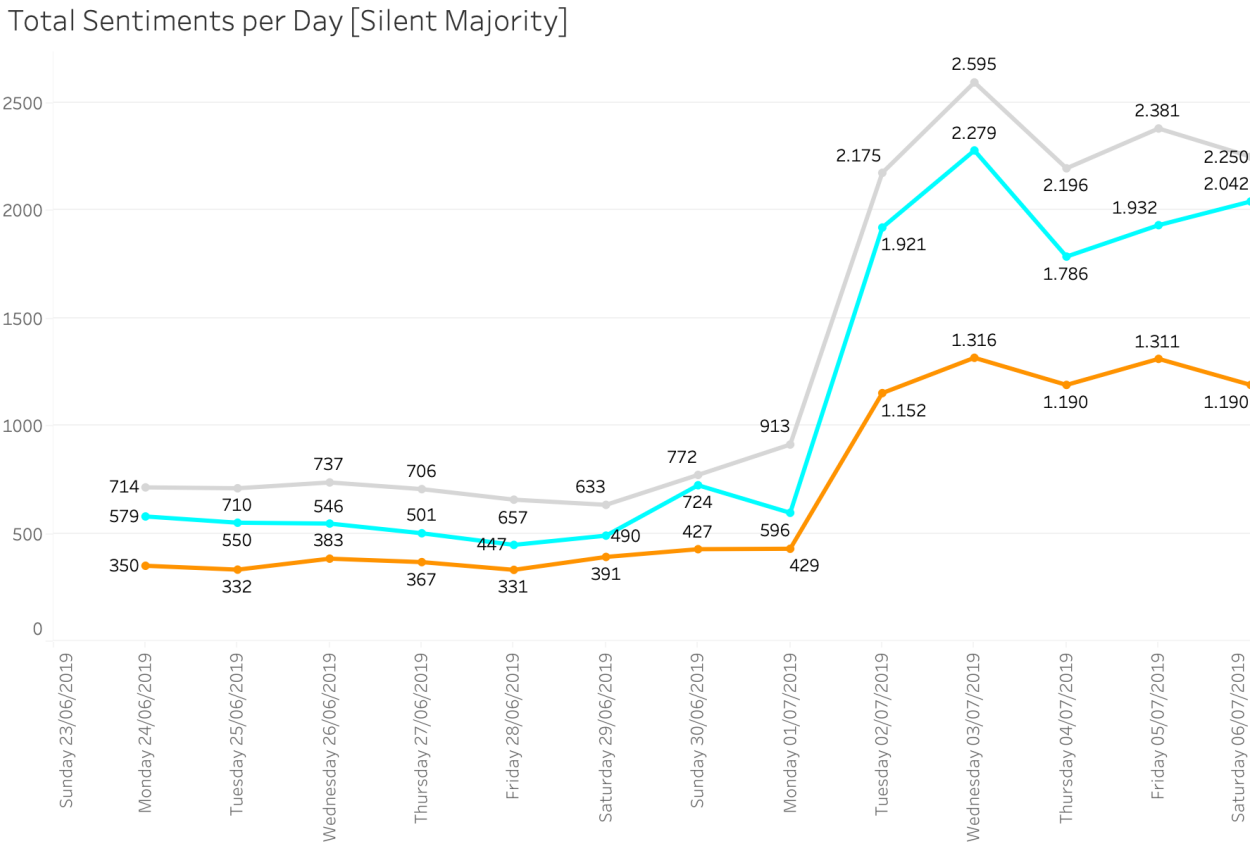



**Figure 54 Total Sentiments per day [Silent majority]**

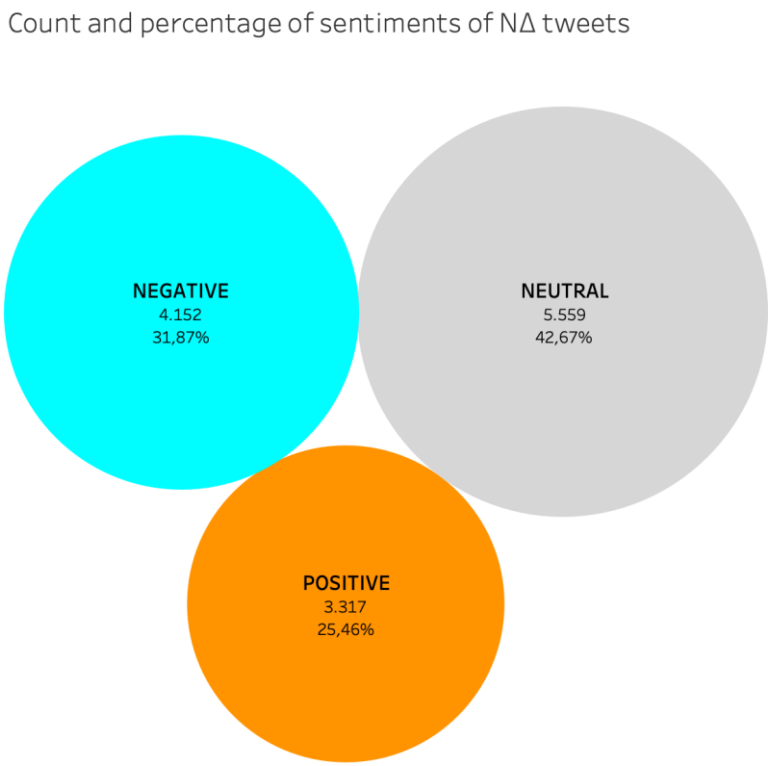



**Figure 55 Count and percentage of sentiments of ND tweets [Silent majority]**

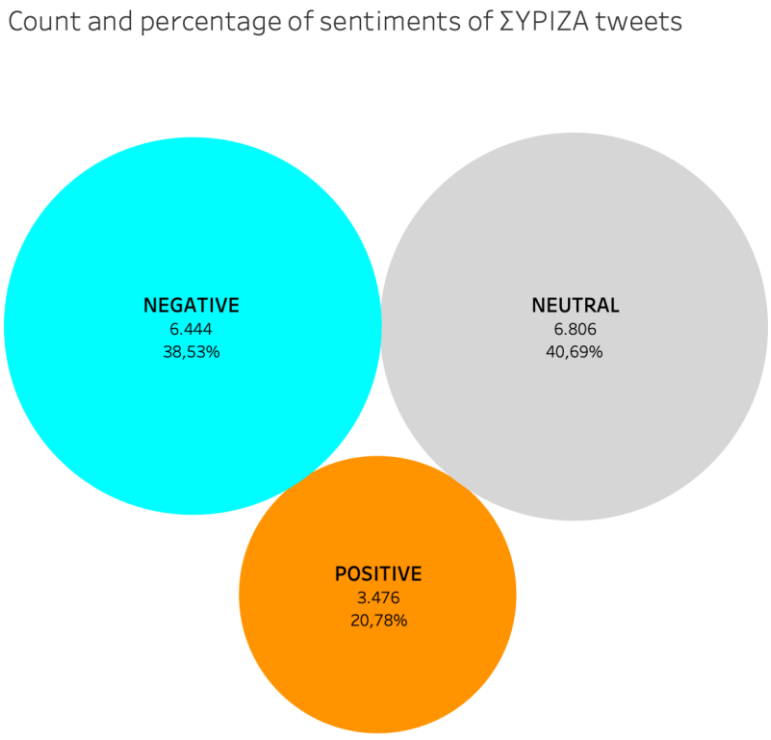



**Figure 56 Count and percentage of sentiments of SYRIZA tweets [Silent majority]**

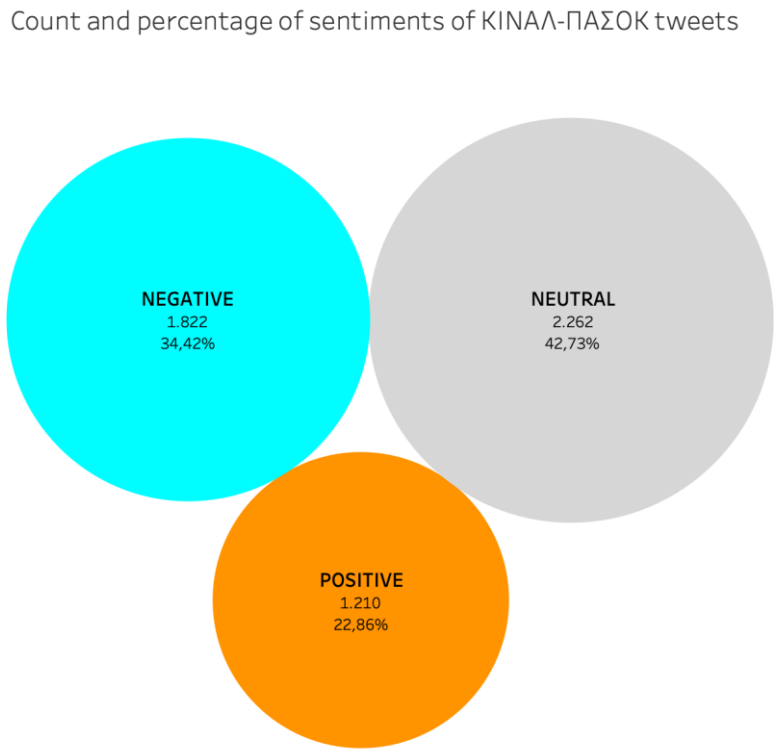



**Figure 57 Count and percentage of sentiments of KINAL-PASOK tweets [Silent majority]**

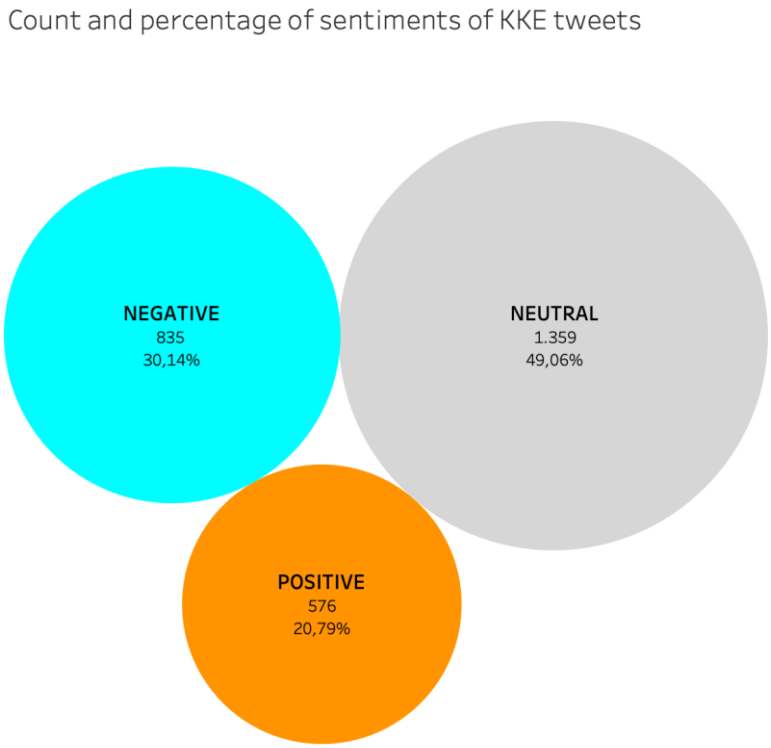



**Figure 58 Count and percentage of sentiments of KKE tweets [Silent majority]**

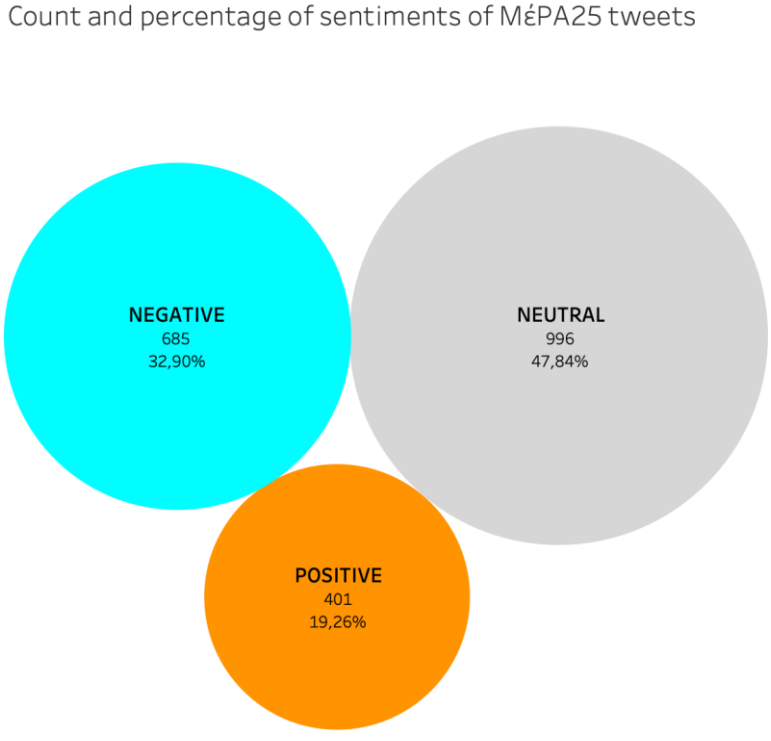



**Figure 59 Count and percentage of sentiments of Mera25 tweets [Silent majority]**

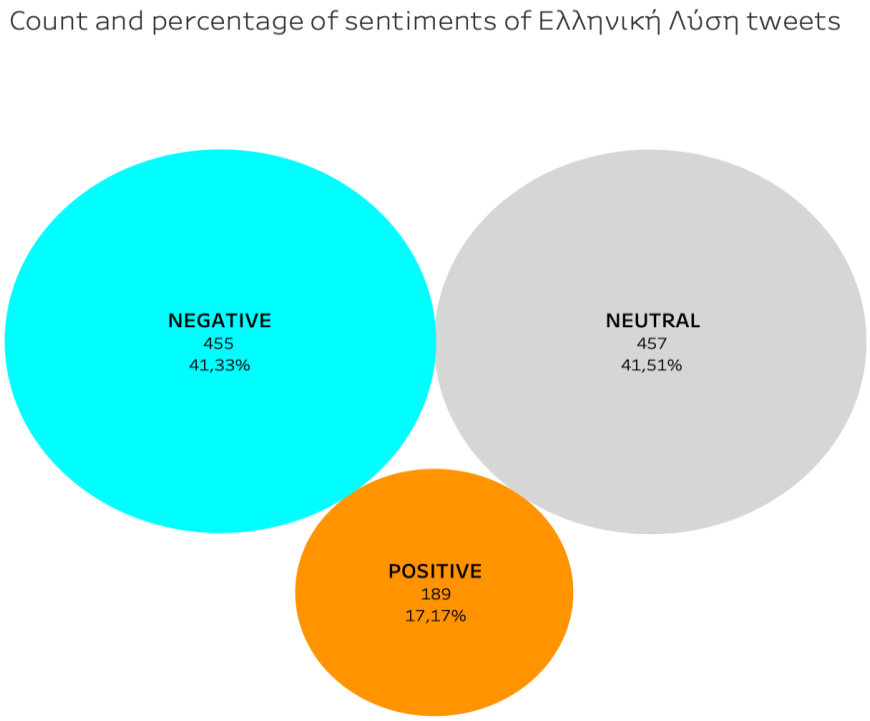



**Figure 60 Count and percentage of sentiments of Elliniki Lysi tweets [Silent majority]**

**Table 13 Percentages of different sentiments on tweets per political party [Silent majority]**

| Political Party | Positive | Negative | Neutral |
|---|---|---|---|
| **ΝΔ** | 25,46 | 31,87 | 42,67 |
| **ΣΥΡΙΖΑ** | 20,78 | 38,53 | 40,69 |
| **ΚΙΝΑΛ-ΠΑΣΟΚ** | 22,86 | 34,42 | 42,73 |
| **ΚΚΕ** | 20,79 | 30,14 | 49,06 |
| **Ελληνική Λύση** | 17,17 | 41,33 | 41,51 |
| **ΜέΡΑ25** | 19,26 | 32,9 | 47,84 |

All political parties, regardless of the number of tweets they collect, show a relatively similar quota of positive, negative and neutral tweets and in the case of the "silent majority"(Figures 53, 55, 56, 57, 58, 59, 60). Neutral tweets are first in number and percentage, followed by negative and finally positive tweets in this case as well. From this data we can draw the following conclusions-estimates:

- The ND party is first in percentage terms in terms of positive tweets. This data confirms the social upward trends that ND had in the 2019 elections.
- The KKE party loses the 1st place it had in the general approach and the 2nd place it had in the "vocal minority" approach and ends up in 3rd place after the KINAL-PASOK party.
- The SYRIZA party continues to have a high percentage of negative tweets, coming in one place ahead of the last party.

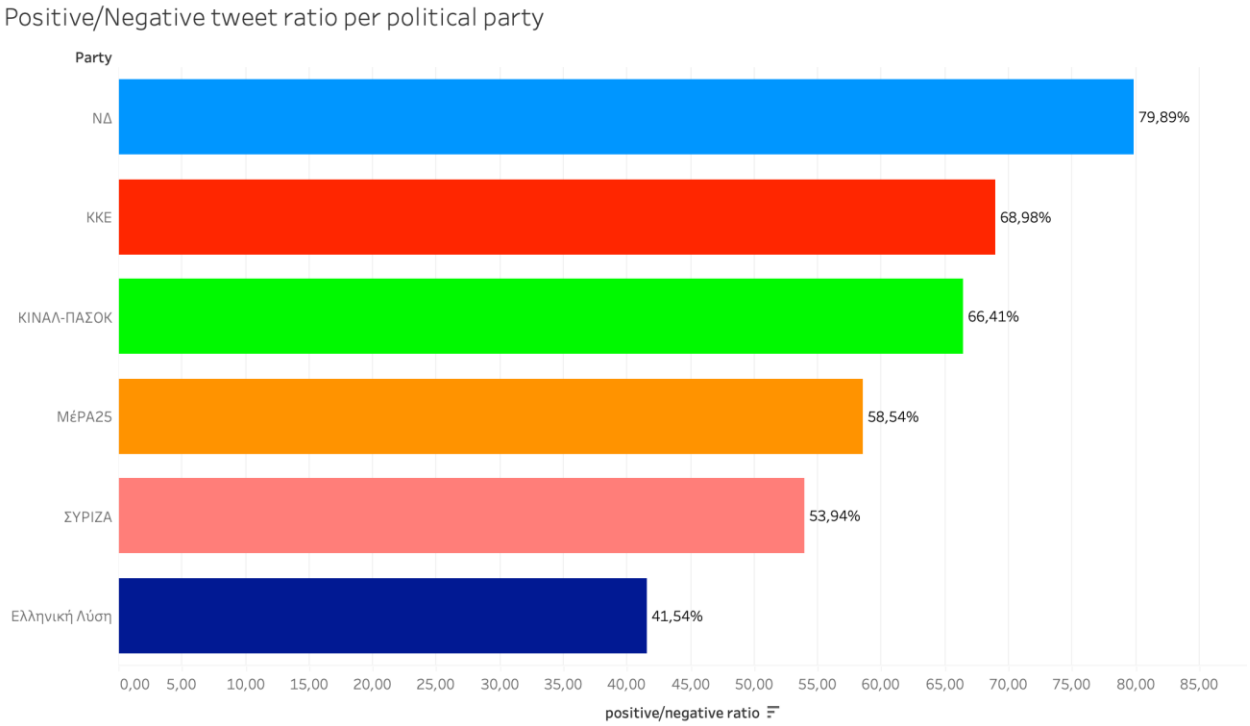



**Figure 61 Positive/Negative tweets ratio per political party [Silent majority]**

In the case of silent majority, it has some differences in terms of comparison based on the positive/negative ratio metric (Figure 61). In particular, we proceed to the following conclusions-observations:

- The ND party emerges 1st in this approach with a slightly smaller percentage, taking this position from the KKE. Approximately for every 8 positive tweets, we have 10 negative tweets.
- The KKE party loses the first place it had in the 2 other approaches, reaching 68.98%.
- The SYRIZA party continues to be in second last place, maintaining a ratio close to 55%.

# 5. CONCLUSIONS AND FURTHER RESEARCH

## 5.1 Summary

Sentiment analysis is a rapidly growing field in natural language processing that has gained significant attention in recent years. With the extensive use of social media platforms and online communication channels, sentiment analysis has become an essential tool for businesses, governments, and individuals to understand and gain insights to public opinion on a wide range of topics.

Sentiment analysis involves analyzing text data, such as tweets, reviews, or news articles, to determine the overall sentiment or attitude conveyed by the text. The analysis can range from identifying basic polarity, such as positive or negative, to more nuanced emotions, such as happiness, sadness, or anger.

The present Thesis targets to implement a lexicon-based sentiment analysis approach for a series of datasets. These datasets concern the 2019 general elections in Greece and the top six (6) parties in terms of electoral share in these elections. They are written in the Greek language. The present research may also be considered as a first attempt to answer the following research question: "Is there a correlation between the sentiment expressed on twitter and the social trends expressed in the elections? And if so, to what extent does this exist?".

In general, the presented implementation can be summarized with the following steps:

1. Removal of the "noise" from the datasets with appropriate preprocessing.
2. Detecting the existence of negation in the text.
3. Application of a lexicon-based approach, in which for each tweet, we collect the positively and negatively charged words. In this way, we compute the polarity and the overall sentiment of each tweet. The sentiments we estimate are three: positive, negative and neutral.
4. For each political party, we update its dataset with the corresponding values for polarity and sentiment.

In addition to the sentiment analysis of the tweets, we also utilize other data to evaluate each tweet such as the total number of retweets, “likes”, and the quota of positive/negative/neutral tweets and the ratio of positive/negative tweets.

In summary, we conclude that there is a relative correlation between the overall sentiment, as expressed on twitter about the election issue, and social trends and attitudes about the same issue.

In other words, the main target of the present Thesis is to achieve a robust lexicon-based implementation approach in sentiment analysis on datasets related to the Greek political scene.

## 5.2 Conclusions

From the above-presented data and results, specific conclusions can be reached in regard to the way political sentiment and opinion in reflected in Twitter –at least in respect to the data for the 2019 Greek general elections. In particular, we can draw the following conclusions:

1. Correlation between feelings expressed on social media about elections and the social trends expressed in the same elections:

   The use of twitter and social media in general reflects and expresses to a large extent social trends around various phenomena, such as in this case elections. We can confidently answer that there is a correlation between the feelings expressed on social media about elections and the social trends expressed in the same elections. A number of results prove it. For example, in the majority of approaches, ND is 1st in the number of positive tweets (in absolute number and percentage), but also in the top positions in the ratio of positive/negative tweets. In the 2019 general election, ND achieves a significant victory and regains the 1st position after its defeat in the 2015 elections. In contrast, SYRIZA manages to be in 1st place in the number of negative tweets in all 3 approaches, but also has a fairly low positive/negative tweet ratio. SYRIZA suffered a defeat in the 2019 elections and found itself in 2nd place, losing the 1st place it had gained in the 2015 elections. The KKE manages to gain space in the results by showing very good percentages in the positive/negative ratio. KKE is a party that managed to have the smallest percentage losses in the 2019 general election.

2. Relation between tweet metrics and votes:

   In some results, such as tweets of positive sentiment (we estimate that positive tweets constitute a "vote" for the respective party), the result is the same as in the elections for the top 4 parties in the majority of approaches.

3. Tweet metrics and dataset comparison:

   Some metrics such as the positive/negative ratio and the quota between positive/negative and neutral tweets allow us to compare datasets of disparate size with each other and draw safe conclusions.

4. Limitations of Twitter and Social Media for predicting election outcome:

   Finally, we conclude that sentiment expressed in social media, such as twitter, expresses existing social trends. However, it cannot be fully exploited for predicting election outcomes / results for a number of reasons:

   - Social media is not fully utilized by all citizens who make up the electorate. As a result, the data collected covers relatively and not fully the electoral preferences of citizens.
   - Social media accounts, like twitter, do not show the same activity. For example, there are simple users on one side and large news agencies on the other posting on the same topics.

## 5.3 Limitations and challenges

For the implementation of the above-presented research, various limitations and obstacles were confronted, which could be summarized in the following two (2) areas:

- The nature and content of the Greek language
- The nature and content of the language and the nature of the content of a text on Twitter

It is generally considered that the richer, in linguistic and syntactic features, a language is, the more demanding is the effort to process it by a machine. In particular, for the Greek language, the following practical (I) and language-related (II) problems were confronted:

(I)

1. Sentiment Lexicons: The dictionaries with emotional terms (either positive or negative) must be as rich as possible and trained based on the corresponding datasets of the application. The richer a dictionary is, the more emotional terms are detected by the application. However, none of the existing sentiment lexicons in the Greek language are ideal. For this reason, their enrichment process needs to be performed manually, a process that is quite time-consuming, especially in the case of large texts.

2. Entries in existing dictionaries: The absence of dictionaries for emotionally charged phrases. Apart from the existence of emotionally-charged words - terms, there is a wide variety of phrases in Greek that express emotion. A complete Greek language processing application for sentiment analysis should include them. For example, the phrase "πήγε περίπατο" is used with a negative intent, indicating a negative sentiment as opposed to its constituent terms. Similar phrases are "πέφτω από τα σύννεφα", "πάω περίπατο", "περιορισμένης ευθύνης", "κλαίω από τα γέλια", "τον έκανα γιο γιο", etc.

3. Data annotation: The absence of annotated (labeled) data for Sentiment Analysis for the Greek language. The existence of numerous and rich labeled datasets for Greek could play a decisive role for the accurate estimation of sentiment in Greek texts.

(II)

4. Polysemy: The polysemy of the Greek language. Polysemy is the phenomenon where a word or phrase has multiple and related meanings. This phenomenon confuses the machine processing of the Greek language. Still there may be words that belong to both syntactic lexicons and mean something different in each case. For example, the word "consistency" in Greek expresses either the behavior of a consistent person or consequence. Greek language processing applications need to take the specific cases seriously and carefully.

5. Contexts of words: The context of use of a word. The process of determining the context of use of a word is quite complicated. This context determines the final meaning of a word, which is a crucial point for processing its emotional content. For example, in the sentence "η κατάσταση αυτή μου κάνει να γελάω" the word "γελάω" has a positive emotional content. However, in the sentence "με αυτά που κάνεις, γελάω μαζί σου ", it has the opposite.

6. Irony: The existence of irony. Irony, as we have already mentioned, is a difficult phenomenon to detect in natural language processing. There are instances of irony in datasets that either contain emotionally charged terms (and thus the polarity of the words should be reversed) or contain no such terms at all.

7. Negation: The complexity of handling negation instances. In the application I implemented, a remarkable effort was made to manage the most frequent negation instances. However, this effort should be enriched by managing more complex negation instances such as double negation.

8. Target of positive/negative sentiment: Identifying the entity that becomes the recipient of positive or negative terms. In a complete and effective processing of Greek for sentiment analysis at sentence level, it should be clarified who is the receiver of the positive or negative sentiment. Otherwise, there is the possibility that there are tweets that refer to another entity (and another dataset respectively), but characterize another dataset. For example, the tweet "Ο ΣΥΡΙΖΑ είναι ανίκανος, η ΝΔ τώρα πρέπει να πιέσει" expresses a negative sentiment about Syriza. If this tweet belongs to the dataset for SYRIZA, then our estimation is correct. But if it belongs to the dataset for ND, then our estimate is wrong. In other words, we need to verify which entity each emotionally charged word refers to.

9. Translation issues: The inability to correctly translate Greek into English. My first approach to implement the application was to translate the tweets into English first and then use various off-the-shelf libraries for Sentiment Analysis. However, the inability to accurately translate the texts (from all the free translation tools available) of the tweets was an obstacle to continue this approach.

Furthermore, the tweets themselves displayed a variety of characteristics and related issues that had to be dealt with, as listed below:

1. The existence of spelling errors. There were many instances where a number of emotionally charged words had spelling errors, resulting in an inability to detect them.

2. The frequent use of slang and neologisms. The frequent use of slang words and phrases, as well as the creation of new ones in a corresponding linguistic style requires manual editing of the texts to identify them. But this process is quite time-consuming.

3. The discourse form of a tweet text. A tweet has more characteristics of informal everyday speech, which makes it difficult to process.

4. The nature of twitter users. A twitter account, could be owned by users with different purposes. For example tweets can be produced by news agencies, channels, individual journalists, political parties, official state bodies, celebrities, targeted bots and even ordinary citizens. Depending on the nature of the subject, we have corresponding tweets. For example, a news site posts tweets with objective criteria, without taking the position of a political party. This contrasts with the tweets of politicians or ordinary citizens who post emotionally charged tweets without being bound by ethics.

5. The nature of the content of a tweet. In a tweet a user has the possibility to post beyond pure text, photos, videos, etc. This makes sentiment analysis of corresponding tweets impossible.

## 5.4 Further Research

For lexicon-based approaches in Greek there is still much to be desired for development and improvement. Further research geared towards this direction can contribute with the following:

1. The development of a complete, rich and dynamic lemmatizer for the Greek language. Existing lemmatizers either have a limited number of words or produce word lemmas in an incorrect way. At the same time, most of them do not have the ability to be manually enriched. Proper rendering of word lemmas offers better performance prospects in lexicon based approaches.
2. The development of a tool for correcting spelling errors for the Greek language. Errors, which are observed in spelling in tweet texts, are quite common among individual users. The ability to correct them reduces any errors in sentiment analysis of tweets.
3. The development of an effective and complete method for negation / denial management. The existence of negation is a situation that is difficult to manage by sentiment analysis tools, as it reverses the polarity of emotionally charged terms.
4. The development of an effective and complete method for managing irony and sarcasm. Irony is considered a complex yet re-occurring issue for sentiment analysis applications and, especially for the case of the Greek language, approaches for managing irony are still quite limited

## ABBREVIATIONS-ACRONYMS

| | |
|---|---|
| SA | Sentiment Analysis |
| TSA | Twitter Sentiment Analysis |
| API | Application Programming Interface |
| GPU | Graphical Processing Units |
| SVM | Support Vector Machine |
| BoW | Bag of Words |
| TCP/IP | Transmission Control Protocol/ Internet Protocol |

# APPENDIX I

# APPENDIX II

# 6. REFERENCES

[1] Bing Liu, Sentiment analysis and opinion mining. San Rafael: Morgan And Claypool, 2012.
[2] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.
[3] “What is Sentiment Polarity | IGI Global,” www.igi-global.com. https://www.igi-global.com/dictionary/sentiment-polarity/69751
[4] K. Ravi and V. Ravi, “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications,” Knowledge-Based Systems, vol. 89, pp. 14–46, Nov. 2015, doi: https://doi.org/10.1016/j.knosys.2015.06.015.
[5] R. Agrawal, “Must Known Techniques for Text Preprocessing in NLP,” *Analytics Vidhya*, Jun. 14, 2021. https://www.analyticsvidhya.com/blog/2021/06/must-known-techniques-for-text-preprocessing-in-nlp/
[6] Tableau, “Data visualization beginner’s guide: a definition, examples, and learning resources,” Tableau Software, 2018. https://www.tableau.com/learn/articles/data-visualization
[7] M. Wankhade, A. C. S. Rao, and C. Kulkarni, “A survey on sentiment analysis methods, applications, and challenges,” Artificial Intelligence Review, Feb. 2022, doi: https://doi.org/10.1007/s10462-022-10144-1.
[8] “What is Sentiment Polarity | IGI Global,” www.igi-global.com. https://www.igi-global.com/dictionary/sentiment-polarity/69751
[9] “1.9. Naive Bayes — scikit-learn 0.21.3 documentation,” Scikit-learn.org, 2019. https://scikit-learn.org/stable/modules/naive_bayes.html
[10] “1.4. Support Vector Machines — scikit-learn 0.20.3 documentation,” Scikit-learn.org, 2018. https://scikit-learn.org/stable/modules/svm.html
[11] Wikipedia Contributors, “Logistic regression,” Wikipedia, Apr. 12, 2019. https://en.wikipedia.org/wiki/Logistic_regression
[12] “Machine Learning Tutorial: The Max Entropy Text Classifier,” blog.datumbox.com. https://blog.datumbox.com/machine-learning-tutorial-the-max-entropy-text-classifier/
[13] Wikipedia Contributors, “AdaBoost,” Wikipedia, Nov. 01, 2019. https://en.wikipedia.org/wiki/AdaBoost
[14] Merriam-Webster, “Definition of SARCASM,” Merriam-webster.com, 2019. https://www.merriam-webster.com/dictionary/sarcasm
[15] R. P. Rao, S. Dayanand, K. R. Varshitha, and K. Kulkarni, “Sarcasm Detection for Sentiment Analysis: A RNN-Based Approach Using Machine Learning,” Lecture Notes in Electrical Engineering, pp. 47–56, 2022, doi: https://doi.org/10.1007/978-981-16-9885-9_4.
[16] A. Alsaeedi and M. Zubair, “A Study on Sentiment Analysis Techniques of Twitter Data,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 2, 2019, doi: https://doi.org/10.14569/ijacsa.2019.0100248.
[17] “What is a microblog?,” Sprout Social. https://sproutsocial.com/glossary/microblog/
[18] Wikipedia Contributors, “Twitter,” Wikipedia, Jan. 27, 2019. https://en.wikipedia.org/wiki/Twitter
[19] S. Kemp, “Digital 2022: Greece,” DataReportal – Global Digital Insights, Feb. 15, 2022. https://datareportal.com/reports/digital-2022-greece

[20] A. Alsaeedi and M. Zubair, “A Study on Sentiment Analysis Techniques of Twitter Data,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 2, 2019, doi: 10.14569/ijacsa.2019.0100248.
[21] A. Barhan and A. Shakhomirov, “Methods for Sentiment Analysis of twitter messages,” presented at the 12th Conference of FRUCT Association, Saint Petersburg.
[22] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” in LREc 2010, vol. 10, no. 2010.
[23] E. Kouloumpis, T. Wilson, and J. Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!,” Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, no. 1, pp. 538–541, Aug. 2021, doi: 10.1609/icwsm.v5i1.14185.
[24] H. Saif, Y. He, and H. Alani, “Semantic sentiment analysis of twitter,” in International semantic web conference 2012, pp. 508–524.
[25] M. Anjaria and R. Guddeti, “Influence factor based opinion mining of Twitter data using supervised learning,” in Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014, pp. 1–8.
[26] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences, vol. 181, no. 6, pp. 1138–1152, Mar. 2011, doi: 10.1016/j.ins.2010.11.023.
[27] J. Lin and A. Kolcz, “Large-scale machine learning at twitter,” in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012, pp. 793–804.
[28] N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, “Tweet sentiment analysis with classifier ensembles,” Decision Support Systems, vol. 66, pp. 170–179, Oct. 2014, doi: 10.1016/j.dss.2014.07.003.
[29] M. Hagen, M. Potthast, M. Büchner, and B. Stein, “Webis: An ensemble for twitter sentiment detection,” in Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015, 2015, pp. 582–589.
[30] T. Chalothom and J. Ellman, “Simple Approaches of Sentiment Analysis via Ensemble Learning,” Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.
[31] G. Paltoglou and M. Thelwall, “Twitter, MySpace, Digg,” ACM Transactions on Intelligent Systems and Technology, vol. 3, no. 4, pp. 1–19, Sep. 2012, doi: 10.1145/2337542.2337551.
[32] F. M. Kundi, A. Khan, and M. Z. Asghar, “Lexicon-based sentiment analysis in the social web,” ournal of Basic and Applied Scientific Research, vol. 4, no. 6, 2014.
[33] M. A. Asghar, A. Khan, S. Ahmad, M. Qasim, and I. A. Khan, “Lexicon- enhanced sentiment analysis framework using rule-based classification scheme,” PloS one, vol. 12, no. 2, 2017.
[34] P. Balage Filho and T. Pardo, “NILC_USP: A hybrid system for sentiment analysis in twitter messages,” in Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, vol. 2, pp. 568–572.
[35] M. Ghiassi, J. Skinner, and D. Zimbra, “Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network,” Expert Systems with Applications, vol. 40, no. 16, pp. 6266–6282, Nov. 2013, doi: 10.1016/j.eswa.2013.05.057.
[36] M. Ghiassi, J. Skinner, and D. Zimbra, “Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network,” Expert Systems with Applications, vol. 40, no. 16, pp. 6266–6282, Nov. 2013, doi: 10.1016/j.eswa.2013.05.057.
[37] F. H. Khan, S. Bashir, and U. Qamar, “TOM: Twitter opinion mining framework using hybrid classification scheme,” Decision Support Systems, vol. 57, pp. 245–257, Jan. 2014, doi: 10.1016/j.dss.2013.09.004.

[38] I. Sabuncu, M. A. Balci, and O. Akguller, “Prediction of USA November 2020 Election Results Using Multifactor Twitter Data Analysis Method,” Oct. 2020, doi: https://doi.org/10.48550/arXiv.2010.15938.
[39] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment,” Citeseer, 2010, pp. 455–479.
[40] A. Bermingham and A. F. Smeaton, “On Using Twitter to Monitor Political Sentiment and Predict Election Results,” Chiang Mai, Thailand, 2011, pp. 2–10.
[41] E. T. K. Sang and J. Bos, “Predicting the 2011 dutch senate election results with twitter,” 2012, pp. 53–60.
[42] M. Choy, M. L. F. Cheong, N. L. Ma, and P. S. Koo, “US Presidential Election 2012 Prediction using Census Corrected Twitter Model,” 2012, pp. 1–12. [Online]. Available: http://arxiv.org/abs/1211.0938.
[43] T. Mahmood, T. Iqbal, F. Amin, W. Lohanna, and A. Mustafa, “Mining Twitter big data to predict 2013 Pakistan election winner,” 2013, pp. 49–54.
[44] A. Makazhanov, D. Rafiei, and M. Waqar, “Predicting political preference of Twitter users,” Social Network Analysis and Mining, vol. 4, no. 1, May 2014, doi: 10.1007/s13278-014-0193-5.
[45] J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, “Election result prediction using Twitter sentiment analysis,” 2016, pp. 1–5.
[46] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, “140 characters to victory?: Using Twitter to predict the UK 2015 General Election,” Electoral Studies, vol. 41, pp. 230–233, Mar. 2016, doi: 10.1016/j.electstud.2015.11.017.
[47] Andy Januar Wicaksono, Suyoto, and Pranowo, “A proposed method for predicting US presidential election by analyzing sentiment in social media,” 2016 2nd International Conference on Science in Information Technology (ICSITech), Oct. 2016, doi: 10.1109/icsitech.2016.7852647.
[48] Wikipedia Contributors, “Python (programming language),” Wikipedia, May 04, 2019. https://en.wikipedia.org/wiki/Python_(programming_language)
[49] Wikipedia Contributors, “Tableau Software,” Wikipedia, Jan. 05, 2020. https://en.wikipedia.org/wiki/Tableau_Software
[50] M. Beck, “How to Scrape Tweets With snscrape,” Medium, Jan. 05, 2022. https://betterprogramming.pub/how-to-scrape-tweets-with-snscrape-90124ed006af
[51] “WordCloud for Python documentation — wordcloud 1.8.1 documentation,” amueller.github.io. https://amueller.github.io/word_cloud/
[52] “Sentiment Lexicons for 81 Languages,” www.kaggle.com. https://www.kaggle.com/datasets/rtatman/sentiment-lexicons-for-81-languages (accessed Jan. 24, 2023)
[53] N.Krystallis,“Greek-SentimentAnalysis, ”www.github.com.https://github.com/NKryst/Greek-Sentiment-Analysis/blob/master/Files/Greek%20Sentiment%20Lexicon/Fixed_Greek_Lexicon.xlsx.
[54] “A Neural NLP toolkit for Greek.” http://nlp.ilsp.gr/nws/
[55] P. Prokopidis and S. Piperidis, “A Neural NLP toolkit for Greek,” 11th Hellenic Conference on Artificial Intelligence, Sep. 2020, doi: 10.1145/3411408.3411430.
[56] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas, “Vocal Minority Versus Silent Majority: Discovering the Opionions of the Long Tail,” 2011 IEEE Third Int’l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int’l Conference on Social Computing, 2011, doi: https://doi.org/10.1109/passat/socialcom.2011.188.