ΠΑΝΤΕΙΟΝ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΟΙΝΩΝΙΚΩΝ ΚΑΙ ΠΟΛΙΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

PANTEION UNIVERSITY OF SOCIAL AND POLITICAL SCIENCES

ΣΧΟΛΗ ΔΙΕΘΝΩΝ ΣΠΟΥΔΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΚΑΙ ΠΟΛΙΤΙΣΜΟΥ

ΤΜΗΜΑ ΕΠΙΚΟΙΝΩΝΙΑΣ, ΜΕΣΩΝ ΚΑΙ ΠΟΛΙΤΙΣΜΟΥ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΠΟΛΙΤΙΣΤΙΚΗ ΔΙΑΧΕΙΡΙΣΗ, ΕΠΙΚΟΙΝΩΝΙΑ ΚΑΙ ΜΕΣΑ»

ΚΑΤΕΥΘΥΝΣΗ: ΚΟΙΝΩΝΙΑ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ, ΜΕΣΑ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑ

«Ανάλυση Λόγου του twitter προφίλ @atsipras και όσων έκαναν αναφορά σε αυτόν»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΙΑΣΩΝ - ΔΗΜΗΤΡΙΟΣ ΤΣΕΤΣΟΣ

Αθήνα, 2023

Τριμελής Επιτροπή

Γιάννης Σκαρπέλος, Καθηγητής Παντείου Πανεπιστημίου (Επιβλέπων)

Παντελής Βατικιώτης, Επίκουρος Καθηγητής Παντείου Πανεπιστημίου

Κώστας Καρπούζης, Επίκουρος Καθηγητής Παντείου Πανεπιστημίου

Ευχαριστίες

Ευχαριστώ τον Καθηγητή μου, Γιάννη Σκαρπέλο, που μέσα από τα μαθήματά του, με έκανε να ανακαλύψω τον μαγευτικό κόσμο του Data Mining.

Ευχαριστώ τους συγγενείς και φίλους μου, για την τεράστια υπομονή και κατανόηση που έδειξαν όλο αυτό το διάστημα.

Περιεχόμενα

Εικόνες

Περίληψη

Η παρούσα έρευνα, αποσκοπεί στη διερεύνηση του Λόγου που εκφράζεται στα Μέσα Κοινωνικής Δικτύωσης, μεταξύ πολιτικών και χρηστών. Ως δείγματα για την έρευνά μας, χρησιμοποιήθηκαν όλες οι δημοσιεύσεις του Αλέξη Τσίπρα στην πλατφόρμα του Twitter, από τις 13 Ιουλίου του 2011 έως τις 31 Δεκεμβρίου του 2022 και όλες οι αναφορές που έγιναν προς το λογαριασμό αυτό από τις 2 Μαΐου του 2012 έως τις 31 Δεκεμβρίου του 2022. Αφού αποκτήσαμε αυτά τα δεδομένα με τη χρήση εντολών στη γλώσσα προγραμματισμού Python, στη συνέχεια αφαιρέσαμε εκείνα τα στοιχεία που δεν μας ήταν χρήσιμα και επεξεργαστήκαμε κατάλληλα τα δεδομένα, ώστε να μπορέσουμε να τα χρησιμοποιήσουμε με το λογισμικό του Orange. Ακολούθως, εφαρμόσαμε σε αυτά, ειδικές μεθόδους Εξόρυξης Δεδομένων, όπως την Ανάλυση Συναισθημάτων, τη Συσταδοποίηση Κειμένου, τη Σημασιολογική Ανάλυση και τη Θεματική Μοντελοποίηση. Με τη χρήση αυτών των μεθόδων, ο στόχος μας ήταν να αποκτήσουμε ακριβή, μετρήσιμα στοιχεία σχετικά με το περιεχόμενο των δύο συνόλων δεδομένων που συλλέξαμε, να ανακαλύψουμε τυχόν υπάρχοντα μοτίβα και να οδηγηθούμε σε χρήσιμα συμπεράσματα. Συγκεκριμένα, με την Ανάλυση Συναισθημάτων, εντοπίσαμε εκείνα τα συναισθήματα που εκφράζει το κάθε αρχείο κειμένου, χρησιμοποιώντας το μοντέλο κατάταξης συναισθημάτων που επινόησε ο Ekman. Με την εφαρμογή της Ιεραρχικής Συσταδοποίησης, δώσαμε αριθμητικά χαρακτηριστικά γνωρίσματα στα κείμενα, υπολογίσαμε τις αποστάσεις μεταξύ τους και τα χωρίσαμε σε μικρές ομοιογενείς ομάδες, τις συστάδες, έχοντας ως σκοπό τον εντοπισμό τυχόν ενδιαφερόντων ευρημάτων, όπως για παράδειγμα, πανομοιότυπων κειμένων μεταξύ των χρηστών. Χρησιμοποιώντας τεχνικές που εντάσσονται στη Σημασιολογική Ανάλυση, ανακαλύψαμε ομάδες γειτνίασης κειμένων, ανάλογα με την κατανομή τους στο δισδιάστατο χάρτη t-SNE και εξάγαμε λέξεις - κλειδιά από αυτές, έχοντας ως σκοπό να ανακαλύψουμε το κατά πόσο τείνουν να χαρακτηρίσουν τα κείμενα στα οποία ανήκουν. Τέλος, με τη Θεματική Μοντελοποίηση και τη χρήση του αλγορίθμου της Λανθάνουσας Κατανομής Dirichlet, βρήκαμε τις κρυμμένες θεματικές για τα κείμενα των δύο συνόλων δεδομένων που αποκτήσαμε, καθώς και τις λέξεις που χαρακτηρίζουν την κάθε θεματική, ώστε να κατανοήσουμε καλύτερα το περιεχόμενο των tweets του Αλέξη Τσίπρα, όσο και των χρηστών που αναφέρονται σε αυτόν. Η παρούσα Διπλωματική εργασία, εκτός του ότι αποτελεί μια μελέτη περίπτωσης στο πλαίσιο ενός μεταπτυχιακού προγράμματος σπουδών, φιλοδοξεί να

εξοικειώσει το ελληνικό κοινό με απλές μεθόδους εύρεσης πληροφοριών μέσα από μεγάλο όγκο κειμενικών δεδομένων, καθώς και να αποτελέσει πηγή έμπνευσης για μελλοντικές έρευνες Ανάλυσης Δεδομένων που αφορούν δημόσια πρόσωπα στην Ελλάδα.
Λέξεις-κλειδιά: Εξόρυξη Δεδομένων, Ανάλυση Συναισθημάτων, Συσταδοποίηση Κειμένου, Σημασιολογική Ανάλυση, Θεματική Μοντελοποίηση.

Abstract

The present study, aims to explore the Speech expressed in Social Media Networks, between politicians and citizens. As samples for our study, we used all of Alexis Tsipras' posts on

Twitter, from 13 of July 2011 until 31 December 2022 and all of the mentions to him from 2 May 2012 until 31 December 2022. After we obtained these datasets using Python language, then we removed those components which were useless and processed the data, in order to use them with Orange. Next, we applied to them, special methods of Data Mining, such as Sentiment Analysis, Text Clustering, Semantic Analysis and Topic Modeling. Using these methods, our aim was to obtain precise, countable information about the content of the two datasets we collected, to discover patterns and to be led into convenient conclusions. Specifically, with Sentiment Analysis, we detected those emotions expressed in each document, using the emotion classification model invented by Ekman. With the implementation of Text Clustering, we gave numeric characteristics to texts, we calculated the distances between them and divided them in small homogenous groups, called clusters, in order to discover interesting findings, such as, identical texts between the users. Using Semantic Analysis techniques, we discovered neighboring groups of texts, based on their distribution on the 2D t-SNE map and we extracted their keywords, in order to examine how much they tend to characterize the texts they belong to. Finally, with Topic Modelling and the use of Latent Dirichlet Allocation algorithm, we found the latent topics of the two datasets which we obtained, as well as the words that tend to characterize each topic, in order to understand better the content of Alexis Tsipras' tweets and the user's mentions referred to him. The current Thesis, despite the fact that it is a study created in the context of a postgraduate programme of studies, aspires to familiarize the greek public with simple methods of information retrieval from big text data, as well as to become a source of inspiration for future Data Analysis surveys about public figures in Greece.
Keywords: Data Mining, Sentiment Analysis, Text Clustering, Semantic Analysis, Topic Modeling.

Εισαγωγή

Data Analysis και Data Mining

Ζούμε στην εποχή της Κοινωνίας της Πληροφορίας[1], στην οποία τα Δεδομένα έχουν πρωταρχικό ρόλο. Η κυριαρχία του Διαδικτύου σε κάθε πτυχή της ζωής των ανθρώπων,

[1] Hanna, K. T. (2022, November). What is an information society? *TechTarget.* Ανάκτηση από: https://www.techtarget.com/whatis/definition/Information-Society Τελευταία πρόσβαση: 31/5/2023.

επηρεάζει καθημερινά εκατομμύρια συμπολίτες μας, με τη δημιουργία, διανομή και χρήση πληροφοριών για οικονομικές, πολιτικές και πολιτιστικές δραστηριότητες.

Έρευνα του Εθνικού Κέντρου Κοινωνικών Ερευνών[2] (ΕΚΚΕ), που δημοσιεύτηκε το 2020 από τη διαΝΕΟσις στο πλαίσιο του διεθνούς έργου "*World Internet Project*", δείχνει ότι στην Ελλάδα 7 στους 10 πολίτες έχουν πρόσβαση στο διαδίκτυο και ότι το 100% των νέων κάτω των 35 ετών το χρησιμοποιούν.

Κάθε μέρα, παγκοσμίως, παράγεται ένας τεράστιος όγκος Δεδομένων από τους χρήστες του Διαδικτύου. Σύμφωνα με τη διεθνή εταιρεία ανάλυσης δεδομένων, SG Analytics[3], το μέγεθος των δεδομένων που παράγονταν καθημερινά κατά το έτος 2020, υπολογίστηκε στα 2.5 πεντάκις εκατομμύρια bytes, τα οποία δημιουργήθηκαν από την χρήση διαφόρων πηγών παραγωγής και άντλησης πληροφορίας, όπως πλατφόρμες κοινωνικής δικτύωσης, emails, video κλπ.

Η επίδραση των Μέσων Κοινωνικής Δικτύωσης (ΜΚΔ) στην ανθρώπινη επικοινωνία είναι καταλυτική. Έρευνα που έγινε το 2022 από την εταιρεία ψηφιακού μάρκετινγκ Humble[4] και με συμμετέχοντες πάνω από 4.000 χρήστες εφαρμογών ΜΚΔ στην Ελλάδα, έδειξε ότι ο κάθε ένας κάτοχος λογαριασμού ΜΚΔ, διαθέτει κατά μέσο όρο λογαριασμούς σε 5 με 7 διαφορετικές πλατφόρμες κοινωνικής δικτύωσης.

Τα ευρήματα από τη συγκεκριμένη, αλλά και από άλλες έρευνες που έχουν γίνει κατά τα τελευταία χρόνια, καταδεικνύουν το βαθμό που το Διαδίκτυο και τα ΜΚΔ, έχουν εισχωρήσει στον ιστό της ελληνικής κοινωνίας, επηρεάζοντας έτσι τις ζωές εκατομμυρίων συμπολιτών μας. Τί είναι όμως αυτά τα δεδομένα που παράγονται και πώς μπορούν να χρησιμοποιηθούν για την άντληση Γνώσης;

Από τον τεράστιο όγκο των Δεδομένων που παράγεται κάθε μέρα, ένα μεγάλο μέρος του μπορεί να χρησιμοποιηθεί, αφού πρώτα γίνουν οι κατάλληλες ενέργειες καθαρισμού και μορφοποιησης του. Οι ενέργειες αυτές αναφέρονται συνοπτικά ως Ανάλυση Δεδομένων και ανήκουν στη σφαίρα του επιστημονικού πεδίου της Επιστήμης των Δεδομένων, που έχει ως

---

[2] Τσέκερης, Χ., Δεμερτζής, Ν. κ.ά. (2020). Το Διαδίκτυο στην Ελλάδα: Η έρευνα του ΕΚΚΕ για το World Internet Project. Αθήνα: διαΝΕΟσις. Ανάκτηση από: https://www.dianeosis.org/wp-content/uploads/2020/06/wip_greece_final_2-6_2020.pdf Τελευταία πρόσβαση: 31/5/2023.

[3] Durairaj, J. (2020, Aug. 14). 2.5 Quintillion Bytes of Data Generated Everyday - Top Data Science Trends 2020. *SG Analytics.* Ανάκτηση από: https://us.sganalytics.com/blog/2-5-quintillion-bytes-of-data-generated-everyday-top-data-science-trends-2020/ Τελευταία πρόσβαση: 31/5/2023.

[4] Humble Team. (2022, Νοέμβριος 9). Greeks & Social Media Research: An up-to-date Whitepaper for the Greek market - 4.000+ Responses (2022). *Humble.* Ανάκτηση από: https://humble.gr/blog/insights/humble-research/ Τελευταία πρόσβαση: 31/5/2023.

σκοπό την εξαγωγή χρήσιμων πληροφοριών, με τη χρήση διαφόρων τεχνικών από πολλά επιστημονικά πεδία, όπως των Μαθηματικών, της Στατιστικής, του Προγραμματισμού, της Γλωσσολογίας, της Τεχνητής Νοημοσύνης (AI) και της Μηχανικής Μάθησης (ML)[5].

Όπως είχε δηλώσει το 2006 σε ομιλία του ο Clive Humby, "*τα δεδομένα είναι το νέο πετρέλαιο*"[6]. Αυτό σημαίνει ότι τα δεδομένα αφού εξαχθούν και μορφοποιηθούν κατάλληλα, μπορούν να παράγουν χρήσιμες πληροφορίες με σκοπό τη λήψη αποφάσεων από εταιρείες, οργανισμούς ή κυβερνήσεις, αλλά και το σχεδιασμό στρατηγικών για την επίλυση συγκεκριμένων προβλημάτων.

Τι είναι όμως ακριβώς η Ανάλυση Δεδομένων και με ποιους τρόπους επιτυγχάνεται; Η Ανάλυση των Δεδομένων που συλλέγονται από το Διαδίκτυο, περιλαμβάνει πολλές δραστηριότητες, όπως τη συλλογή, τον καθαρισμό και την οργάνωση των δεδομένων[7]. Αυτές οι διαδικασίες, οι οποίες συνήθως γίνονται με ειδικά λογισμικά, είναι απαραίτητες για την προετοιμασία των δεδομένων και την μετατροπή τους σε πληροφορίες, για επιχειρηματικούς ή άλλους σκοπούς.

Υπάρχουν δύο βασικές μέθοδοι Ανάλυσης των Δεδομένων που εξάγονται, οι ποσοτικές και οι ποιοτικές. Οι ποσοτικές μέθοδοι αναλύουν ποσοτικά δεδομένα ή αλλιώς δομημένα δεδομένα. Τα δεδομένα αυτά αποτελούνται από μετρήσιμα στοιχεία και μπορούν συνήθως να εντοπιστούν μέσα σε ένα τυπικό σύνολο δεδομένων, με σειρές και στήλες. Οι τεχνικές ανάλυσής τους, χρησιμοποιούνται, για να εξηγήσουν συγκεκριμένες καταστάσεις ή να πραγματοποιήσουν προβλέψεις και τα ευρήματα τους έχουν αντικειμενική ερμηνεία. Από την άλλη, οι ποιοτικές μέθοδοι, ασχολούνται με την ανάλυση ποιοτικών ή αλλιώς αδόμητων δεδομένων. Τα δεδομένα που αναλύονται με τις ποιοτικές μεθόδους, δεν μπορούν να ιδωθούν αντικειμενικά και έτσι τείνουν προς υποκειμενικές ερμηνείες. Χαρακτηριστικά παραδείγματα αδόμητων δεδομένων είναι κείμενα, απαντήσεις σε συνεντεύξεις και αναρτήσεις στα ΜΚΔ[8].

---

[5]IBM. What is Data Science? Ανάκτηση από: https://www.ibm.com/topics/data-science Τελευταία πρόσβαση: 31/5/2023.

[6]LaRock, T. (2022, Oct 6). "Data is the New Oil", But That Also Means it Can be Risky. *database.* Ανάκτηση από: https://www.dbta.com/Columns/Next-Gen-Data-Management/Data-is-the-New-Oil-But-That-Also-Means-it-Can-be-Risky-155275.aspx Τελευταία πρόσβαση: 31/5/2023.

[7] Maryville University. Top 4 Data Analysis Techniques That Create Business Value. Ανάκτηση από: https://online.maryville.edu/blog/data-analysis-techniques/#what-is Τελευταία πρόσβαση: 31/5/2023.

[8]Stevens, E. (2023, May 10). The 7 Most Useful Data Analysis Methods and Techniques. *CAREERFOUNDRY.* Ανάκτηση από: https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/ Τελευταία πρόσβαση: 31/5/2023.

Για να αντλήσουμε γνώση από ένα σύνολο δεδομένων, πρέπει να ακολουθήσουμε μία σειρά από ενέργειες, που ονομάζονται συνοπτικά Εξόρυξη Δεδομένων (Data Mining). Με τον όρο αυτό, εννοούμε τη διαδικασία εξαγωγής χρήσιμης πληροφορίας, μοτίβων και τάσεων μέσα από αδόμητα δεδομένα. Σύμφωνα με την εγκυκλοπαίδεια Britannica[9], Εξόρυξη Δεδομένων στην υπολογιστική επιστήμη, ονομάζεται η διαδικασία της ανακάλυψης ενδιαφερόντων και χρήσιμων μοτίβων και σχέσεων, μέσα σε μεγάλο όγκο δεδομένων. Το πεδίο αυτό, συνδυάζει μεθόδους από τη Στατιστική και την Τεχνητή Νοημοσύνη, μαζί με τη Διαχείριση Βάσεων Δεδομένων, με στόχο την ανάλυση μεγάλων συνόλων δεδομένων. Η Εξόρυξη των Δεδομένων χρησιμοποιείται ευρέως στις επιχειρήσεις, την επιστημονική έρευνα και την κρατική ασφάλεια.

Σε αυτό το σημείο, θα πρέπει αρχικά να γίνει κατανοητή η διαφορά μεταξύ των όρων Εξόρυξη Δεδομένων και Ιστοσυγκομιδή (Web Scraping)[10]. Ο τελευταίος, αναφέρεται στη διαδικασία που ακολουθείται για την εξαγωγή δεδομένων από το Διαδίκτυο και την μετατροπή τους σε εύχρηστη μορφή, ενώ δεν περιλαμβάνει κάποια διαδικασία επεξεργασίας ή ανάλυσης. Από την άλλη, η Εξόρυξη Δεδομένων, επικεντρώνεται στην ανάλυση μεγάλων συνόλων δεδομένων και δεν ασχολείται με την συλλογή τους. Με λίγα λόγια, το Web Scraping προετοιμάζει το έδαφος για το Data Mining, δηλαδή την ανακάλυψη της πληροφορίας.

Κάποιες από τις διαδεδομένες τεχνικές Εξόρυξης Δεδομένων, είναι η Συσταδοποίηση (Clustering), οι Κανόνες Συσχέτισης (Association Rules), η Κατηγοριοποίηση (Classification), η Σημασιολογική Ανάλυση (Semantic Analysis), η Ανάλυση Συναισθημάτων (Sentiment Analysis), η Μηχανική Μάθηση (Machine Learning), τα Νευρωνικά Δίκτυα (Neural Networks), ο Καθαρισμός Δεδομένων (Data Cleaning) και η Οπτικοποίηση Δεδομένων (Data Visualization)[11].

Η διαδικασία που ακολουθείται για την Ανάλυση των Δεδομένων, περιλαμβάνει συγκεκριμένα βήματα. Υπάρχουν διάφορες προσεγγίσεις για το ποιά είναι αυτά τα βήματα που πρέπει να ακολουθηθούν, αλλά μπορούν να συνοψιστούν σε 5 βασικές διαδικασίες: α)

[9]Clifton, C. data mining. *Encyclopaedia Britannica*. Ανάκτηση από: https://www.britannica.com/technology/data-mining Τελευταία πρόσβαση: 31/5/2023.
[10] parsehub. (2020, March 2). Web Scraping vs Data Mining: What's the Difference?. Ανάκτηση από: https://www.parsehub.com/blog/web-scraping-vs-data-mining/ Τελευταία πρόσβαση: 31/5/2023.
[11]Simplilearn. (2023, May 4). Types of Data Mining Techniques. *simplilearn.* Ανάκτηση από: https://www.simplilearn.com/types-of-data-mining-techniques-article Τελευταία πρόσβαση: 31/5/2023.

στην Ανίχνευση του ερωτήματος που θέλουμε να απαντηθεί ή του προβλήματος που θέλουμε να λύσουμε, β) τη Συλλογή των δεδομένων, γ) τον Καθαρισμό τους με την αποβολή άχρηστων στοιχείων καθώς και τη μετατροπή των δεδομένων σε διαχειρίσιμα, δ) την Ανάλυση των δεδομένων με τη χρήση διαφόρων τεχνικών ανάλυσης και εργαλείων με σκοπό την εξαγωγή και μετατροπή των χρήσιμων πληροφοριών σε εύληπτες γραφικές αναπαραστάσεις και ε) την Ερμηνεία των αποτελεσμάτων της Ανάλυσης για την απάντηση στο ερώτημα ή το πρόβλημα που ανιχνεύσαμε στο πρώτο βήμα[12].

Data Mining και ΜΚΔ

Σύμφωνα με τους Obar και Wildman (2015)[13], για να ορίσουμε τα ΜΚΔ, θα πρέπει να συνδυάσουμε τους ορισμούς που δίνονται στη βιβλιογραφία και να ανιχνεύσουμε κάποια κοινά χαρακτηριστικά που μοιράζονται μεταξύ τους οι υφιστάμενες υπηρεσίες κοινωνικής δικτύωσης, όπως το ότι βασίζονται σε διαδικτυακές εφαρμογές του Web 2.0, δίνεται η δυνατότητα δημιουργίας περιεχομένου στους χρήστες, άτομα και ομάδες μπορούν να δημιουργήσουν προφίλ και τέλος, διευκολύνεται η ανάπτυξη κοινωνικών δικτύων μέσω της σύνδεσης του προφίλ ενός χρήστη με άλλα προφίλ ή και ομάδες.

Αυτή την στιγμή υπάρχουν δεκάδες πλατφόρμες και εφαρμογές κοινωνικής δικτύωσης, κάποιες από τις οποίες είναι ιδιαίτερα δημοφιλείς στην χώρα μας. Έρευνα που δημοσιεύτηκε τον Φεβρουάριο του 2022 στον ιστότοπο Datareportal[14] για τη χρήση των κυριότερων εφαρμογών ΜΚΔ στην Ελλάδα, έδειξε ότι οι πλατφόρμες που προτιμούν οι Έλληνες είναι το Instagram (35,99%), το TikTok (23,03%), το YouTube (15,13%), το Facebook (12,22%) και το Pinterest (5,07%).

Πρόσφατα στατιστικά στοιχεία για τον ακριβή αριθμό των χρηστών της πλατφόρμας Twitter στην Ελλάδα δεν υπάρχουν, ωστόσο η ίδια έρευνα αναφέρει ότι σύμφωνα με στοιχεία που δημοσίευσε το Twitter για τα διαφημιστικά του έσοδα, κατά τους πρώτους μήνες του 2022 στην Ελλάδα, υπήρχαν 706.7 χιλιάδες χρήστες του Μέσου.

[12] Coursera. (2023, April 12). What is Data Analysis? (With Examples). Ανάκτηση από: https://www.coursera.org/articles/what-is-data-analysis-with-examples Τελευταία πρόσβαση: 31/5/2023.
[13]Obar, J. & Wildman, S. (2015, July 22). Social Media Definition and the Governance Challenge: An Introduction to the Special Issue. *Telecommunications Policy,* 39(9), 745-750., Quello Center Working Paper No. 2647377, http://dx.doi.org/10.2139/ssrn.2647377 Ανάκτηση από: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2647377 Τελευταία πρόσβαση: 31/5/2023.
[14]Kemp, S. (2022, February 15). DIGITAL 2022: GREECE. *Datareportal.* Ανάκτηση από: https://datareportal.com/reports/digital-2022-greece Τελευταία πρόσβαση: 31/5/2023.

Το Twitter είναι μία υπηρεσία κοινωνικής δικτύωσης, στην οποία οι χρήστες παράγουν περιεχόμενο και αλληλεπιδρούν μεταξύ τους με αναρτήσεις, γνωστές ως "tweets". Με τον όρο "tweet", εννοούμε ένα μήνυμα δημοσιευμένο στο Twitter. Το μήνυμα αυτό, μπορεί να περιλαμβάνει κείμενο, συνδέσμους, φωτογραφίες ή βίντεο[15]. Συνεπώς, κάθε tweet περιέχει αδόμητα δεδομένα, τα οποία με τις κατάλληλες ενέργειες μπορούν να εξαχθούν και να μετατραπούν σε χρήσιμες, αξιοποιήσιμες πληροφορίες και γνώση.

Τα δεδομένα που αντλούνται από την πλατφόρμα του Twitter, μπορούν να χρησιμοποιηθούν για πολλούς σκοπούς, όπως επιχειρηματικούς, διαφημιστικούς, πολιτικούς, ψυχαγωγικούς, καθώς και για την ανίχνευση των απόψεων μεταξύ των χρηστών του, γύρω από διάφορα θέματα της επικαιρότητας. Με την ανάλυση των tweets, μπορούν να εντοπιστούν συναισθήματα, δομές κοινωνικών δικτύων, θέματα και τάσεις που φέρονται να απασχολούν τους πολίτες και ως εκ τούτου, να καθοριστούν πολιτικές και να αναληφθούν δράσεις από άτομα, κυβερνήσεις και οργανώσεις[16].

Όσον αφορά τη χρήση του Twitter για πολιτικούς σκοπούς, διεξάγεται ένας διαρκής και εκτενής, δημόσιος και ακαδημαϊκός διάλογος, σχετικά με τη χρήση του από πολιτικούς και ομάδες πίεσης ή με ποιόν τρόπο λειτουργούν οι αλγόριθμοι, ενισχύοντας ή μειώνοντας την απήχηση της κάθε ομάδας[17].

Χαρακτηριστικό παράδειγμα της χρήσης του Twitter ως εργαλείο πολιτικής επικοινωνίας, ήταν η δραστηριότητα του λογαριασμού @realDonaldTrump κατά τη διάρκεια των Προεδρικών Εκλογών στις Η.Π.Α. το 2016[18]. Σχετικά με το λογαριασμό του πρώην Προέδρου των Η.Π.Α. και τη συνολική του παρουσία στο Twitter, έχουν γίνει αρκετές έρευνες και αποτελεί κλασσικό case study για το πώς μπορεί η ρητορική ενός πολιτικού, ιδιαίτερα

[15]Dictionary.com. Tweet Definition & Meaning. Ανάκτηση από: https://www.dictionary.com/browse/tweet Τελευταία πρόσβαση: 31/5/2023.

[16]Azam, N., Jahiruddin, Abulaish, M. and Haldar, N. Al-H. (2015). *Twitter Data Mining for Events Classification and Analysis,* 2nd International Conference on Soft Computing and Machine Intelligence (ISCMI'15), Hong Kong, IEEE CPS, Nov. 23-24, 2015. Ανάκτηση από: https://www.researchgate.net/publication/297453149_Twitter_Data_Mining_for_Events_Classification_and_Analysis Τελευταία πρόσβαση: 31/5/2023.

[17]Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022, January). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences of the United States of America*, 119(1), e2025334119. https://doi.org/10.1073/pnas.2025334119. Ανάκτηση από: 3/5/2023. Τελευταία πρόσβαση: 31/5/2023.

[18]Muller, K., Schwarz, C. & Fujiwara, T. (2020, 30 October). How Twitter affected the 2016 presidential election. *VoxEu.* Ανάκτηση από: https://cepr.org/voxeu/columns/how-twitter-affected-2016-presidential-election Τελευταία πρόσβαση: 31/5/2023.

ενεργού στα ΜΚΔ, να επηρεάσει μερίδα των πολιτών και να τους κατευθύνει σε δράσεις στο φυσικό κόσμο.

Συλλογή Δεδομένων στο Twitter

Με ποιον όμως τρόπο, μπορεί κάποιος να συλλέξει δεδομένα από μία πλατφόρμα ΜΚΔ, όπως το Twitter, ώστε μετά να τα αναλύσει; Η νόμιμη και πιο γνωστή μέθοδος, είναι μέσω του Twitter API, το οποίο επιτρέπει την προγραμματική πρόσβαση στην πλατφόρμα του Twitter και μπορεί να χρησιμοποιηθεί για την εμφάνιση και συλλογή ενός συγκεκριμένου αριθμού tweets πάνω σε ένα θέμα ή δημόσια διαθέσιμων πληροφοριών (public information) που αφορούν χρήστες. Για τη χρήση του Twitter API, είναι απαραίτητη η δημιουργία λογαριασμού και η απόκτηση των κατάλληλων πιστοποιητικών (credentials)[19].

Αφού αποκτηθεί πρόσβαση στις υπηρεσίες του Twitter API, μπορεί να γίνει χρήση διαφόρων βιβλιοθηκών της Python για τον εντοπισμό και τη συλλογή των tweets, όπως το tweepy και το twarc, οι οποίες όμως έχουν rate limit στη συλλογή των δεδομένων, σύμφωνα με τους περιορισμούς που θέτει το Twitter API και έτσι είναι δύσκολη έως αδύνατη η εξαγωγή μεγάλου όγκου δεδομένων και tweets που δημοσιεύτηκαν πριν από συγκεκριμένο χρονικό διάστημα[20].

Ωστόσο, είναι εφικτό να συλλεχθούν ιστορικά tweets ή ολόκληρα σύνολα δεδομένων με τις αναρτήσεις ενός χρήστη; Αυτό, μπορεί να γίνει με τη χρήση συγκεκριμένων βιβλιοθηκών της Python, οι οποίες παρακάμπτουν τους περιορισμούς στην ανεύρεση των πληροφοριών, κάνοντας scraping[21] σε δεδομένα Υπηρεσιών Κοινωνικής Δικτύωσης (Social Networking Services).

Μια τέτοια βιβλιοθήκη, είναι το snscrape, το οποίο μπορεί να χρησιμοποιηθεί για την εξαγωγή δεδομένων και από άλλες πλατφόρμες ΜΚΔ, πέραν του Twitter, όπως το Facebook, το Instagram, το Reddit κλπ. Το snscrape λειτουργεί με την έκδοση 3.8 της Python ή και νεότερες και μπορεί να εξάγει τα δεδομένα σε αρχεία μορφής .csv, διευκολύνοντας τη μετέπειτα ανάλυση και οπτικοποίηση τους με τα κατάλληλα εργαλεία.

[19]Fontanella, C. (2021, June 21). How to Get, Use & Benefit From Twitter's API. *HubSpot.* Ανάκτηση από: https://blog.hubspot.com/website/how-to-use-twitter-api Τελευταία πρόσβαση: 31/5/2023.
[20] Twitter API. Ανάκτηση από: https://developer.twitter.com/en/docs/twitter-api Τελευταία επίσκεψη: 31/5/2023.
[21]Bettenbuk, Z. (2022, October). How to Scrape Twitter Data Using Python Without Using Twitter's API. *scrapperapi.* Ανάκτηση από: https://www.scraperapi.com/blog/scrape-twitter-data/ Τελευταία πρόσβαση: 31/5/2023.

Ερευνητικό Ερώτημα

Με την παρούσα διπλωματική εργασία, επιχειρήσαμε να διερευνήσουμε τα tweets του λογαριασμού του @atsipras στο Twitter, από την ημέρα που πραγματοποίησε την πρώτη του ανάρτηση, στις 13 Ιουλίου του 2011, έως την ημέρα που ολοκληρώσαμε την καταγραφή για αυτή την έρευνα, στις 31 Δεκεμβρίου του 2022. Παράλληλα, μελετάμε τα mentions προς τον λογαριασμό αυτό, από τις 2 Μαΐου του 2012, έως τις 31 Δεκεμβρίου του 2022.

Τα δεδομένα αυτά, αφού τα συλλέξαμε μέσω της βιβλιοθήκης του Snscrape, τα καθαρίσαμε και τα τροποποιήσαμε με συγκεκριμένες διαδικασίες, ώστε να καταστούν εύχρηστα και διαχειρίσιμα.

Για την ανάλυση των κειμενικών αυτών δεδομένων, χρησιμοποιήσαμε το ειδικό λογισμικό ανοικτού κώδικα, Orange, με το οποίο επιχειρήσαμε ενέργειες Ανάλυσης Συναισθημάτων (Sentiment Analysis), Συσταδοποίησης Κειμένου (Text Clustering), Σημασιολογικής Ανάλυσης (Semantic Analysis) και Θεματικής Μοντελοποίησης (Topic Modeling).

Με την Ανάλυση Συναισθημάτων, εντοπίσαμε τα συναισθήματα που εκφράζονται στα κείμενα των δύο συνόλων δεδομένων και σε συνδυασμό με τις ημερομηνίες που αναρτήθηκαν, καταφέραμε να εξάγουμε χρήσιμες πληροφορίες. Με την εφαρμογή της Συσταδοποίησης Κειμένου, χωρίσαμε τα κείμενα σε μικρές ομάδες, ανάλογα με το βαθμό ομοιότητά τους και είδαμε ποιές συστάδες είναι συμπαγείς ως προς το περιεχόμενό τους και ποιές όχι. Μέσω της Σημασιολογικής Ανάλυσης, αναλύσαμε ομάδες κειμένων που παρουσιάζουν παρόμοιο σημασιολογικό περιεχόμενο και εμφανίσαμε τις λέξεις - κλειδιά που τις χαρακτηριζουν, σε σχέση με την παρουσία τους στο συνολικό Σώμα Κειμένου (Corpus). Τέλος, με τη Θεματική Μοντελοποίηση, εντοπίσαμε τις κρυμμένες θεματικές για τα δύο σύνολα δεδομένων που αποκτήσαμε και με τις λέξεις - κλειδιά που τις χαρακτηρίζουν, προσπαθήσαμε να κατανοήσουμε το περιεχόμενό τους.

Τα ερευνητικά ερωτήματα που τίθενται στην παρούσα Διπλωματική Εργασία είναι:
α) Τί είδους συναισθήματα μπορούν να δημιουργηθούν κατά την έκφραση λόγου στο Twitter, από και προς τον λογαριασμό του @atsipras, σε συνάρτηση με τις συγκυρίες κατά τις οποίες γράφτηκαν;

β) Μπορούν να ομαδοποιηθούν, ανάλογα την ομοιότητά τους, τα κείμενα πολιτικού περιεχομένου που δημιουργήθηκαν και αν ναι, τι ευρήματα προκύπτουν από την ομαδοποίηση;

γ) Υπάρχουν λέξεις - κλειδιά που εμφανίζονται στις αναρτήσεις του @atsipras και των ονομαστικών αναφορών προς αυτόν και αν ναι, είναι χαρακτηριστικές των corpora;

δ) Υπάρχουν θεματικές που αναδύονται από τις αναρτήσεις του @atsipras και τις αναφορές χρηστών προς αυτόν;

Δομή και Διάρθρωση της Έρευνας

Η παρούσα ερευνητική μελέτη, αποτελείται από έξι κεφάλαια. Στο πρώτο κεφάλαιο, με τίτλο "*Βιβλιογραφική Ανασκόπηση*", αναφέρουμε τις διάφορες θεωρίες με βάση τις οποίες πραγματοποιήσαμε στη συνέχεια τις ενέργειές μας. Συγκεκριμένα, επιχειρείται μια ουσιώδης αναφορά στις θεωρίες της Ανάλυσης Συναισθημάτων, Συσταδοποίησης Κειμένων, Σημασιολογικής Ανάλυσης και Θεματικής Μοντελοποίησης. Σκοπός αυτού του κεφαλαίου είναι να πληροφορήσει τον αναγνώστη για μερικές από τις πολλές εφαρμογές της Εξόρυξης Δεδομένων πάνω σε κειμενικά δεδομένα.

Το δεύτερο κεφάλαιο, με τίτλο "*Μεθοδολογία*", αρχικά παρουσιάζει τις ενέργειες που ακολουθήσαμε για την εξόρυξη, τον καθαρισμό και την προετοιμασία των προς ανάλυση δεδομένων και έπειτα, αναλύει τις μεθόδους που εφαρμόσαμε πάνω σε αυτά, με σκοπό να εξάγουμε ενδιαφέρουσες και αξιοποιήσιμες πληροφορίες.

Στο τρίτο κεφάλαιο, με τίτλο "*Ευρήματα*", παρουσιάζονται τα αποτελέσματα των μεθόδων που εφαρμόσαμε προηγουμένως. Συγκεκριμένα, παρατίθενται τα ευρήματα των ενεργειών Ανάλυση Συναισθημάτων, Συσταδοποίηση Κειμένων, Σημασιολογική Ανάλυση και Θεματική Μοντελοποίηση. Σκοπός αυτού του κεφαλαίου, είναι να παρουσιάσει μια πρώτη εικόνα για το ποιά είναι τα αποτελέσματα της εφαρμογής των παραπάνω μεθόδων και να ενημερώσει τον αναγνώστη για μερικά από τα ευρήματα που εξάγαμε.

Το επόμενο κεφάλαιο, με τίτλο "*Επεξήγηση των Ευρημάτων*", αναλύει τα ευρήματα που παρουσιάσαμε στο προηγούμενο κεφάλαιο, ώστε να μπορέσουμε να εξάγουμε κάποια χρήσιμα συμπεράσματα για τα δύο σύνολα δεδομένων που αναλύσαμε και να εντοπίσουμε ενδιαφέροντα μοτίβα που υποδεικνύουν σχέσεις μεταξύ των δεδομένων.

Στο επόμενο κεφάλαιο, "*Συζήτηση και Συμπεράσματα*", επιχειρούμε να απαντήσουμε αναλυτικά στα 4 βασικά ερευνητικά ερωτήματα που θέσαμε στην Εισαγωγή. Με τον τρόπο αυτό, κάνουμε μια ανακεφαλαίωση και σταχυολογούμε εκείνα τα ευρήματά που αποτελούν χρήσιμη πληροφορία και άρα, Γνώση.

Κλείνοντας, με το έκτο κεφάλαιο "*Περιορισμοί και Προτάσεις για Μελλοντική Έρευνα*", αναφέρουμε συνοπτικά τα τυχόν προβλήματα που συναντήσαμε κατά τη διεκπεραίωση της παρούσας Διπλωματικής έρευνας και κάνουμε μια πρόταση για μελλοντική έρευνα πάνω στο πεδίο της Εξόρυξης Δεδομένων.

Κεφάλαιο 1: Βιβλιογραφική Ανασκόπηση

1.1. Θεωρία της Ανάλυσης Συναισθημάτων

Η Ανάλυση Συναισθημάτων[22], είναι μια τεχνική που χρησιμοποιεί μεθόδους Επεξεργασίας Φυσικής Γλώσσας (NLP), για να εξάγει, μετατρέψει και να ερμηνεύσει στοιχεία ενός κειμένου και στη συνέχεια να τα ταξινομήσει σε θετικά, αρνητικά ή ουδέτερα συναισθήματα. Αποτελεί το πεδίο της επιστήμης που αναλύει τις γνώμες, τα συναισθήματα και τις συμπεριφορές των ανθρώπων προς οντότητες, οι οποίες μπορεί να είναι προϊόντα, υπηρεσίες, οργανώσεις, θεσμοί, άνθρωποι ή γεγονότα. Το συγκεκριμένο ερευνητικό πεδίο αντιπροσωπεύει ένα μεγάλο φάσμα ενεργειών, με πολλές, παρόμοιες ονομασίες και ελάχιστα διαφορετικές εργασίες, όπως για παράδειγμα τις μεθόδους: Ανάλυση Συναισθήματος, Εξόρυξη Γνώμης (Opinion Mining), Ανάλυση Γνώμης (Opinion Analysis), Εξόρυξη Συναισθήματος (Sentiment Mining), Εντοπισμό Συναισθήματος (Emotion Detection), Υποκειμενική Ανάλυση (Subjectivity Analysis), Ανάλυση Επίδρασης (Affect Analysis) και Εξόρυξη Αξιολόγησης (Review Mining). Όλες αυτές οι μέθοδοι, ανήκουν στην σφαίρα της Ανάλυσης Συναισθημάτων.

Το πρώτο επιστημονικό άρθρο που αναφέρεται σε μέθοδο Εξόρυξης Γνώμης, δημοσιεύτηκε από τον Stagner (1940)[23] και αφορούσε την ανάπτυξη τεχνικής για τον

[22]Liu, B. (2020). Introduction. In Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. *Studies in Natural Language Processing*, pp. 1-17. Cambridge: Cambridge University Press. doi:10.1017/9781108639286.002. Ανάκτηση από: https://www.cambridge.org/core/books/sentiment-analysis/introduction/563742A639EEE9F5AB3F29CB2387E41C Τελευταία πρόσβαση: 31/5/2023.

[23]Stagner R. (1940). The cross-out technique as a method in public opinion analysis. *Journal of Social Psychology.* 11(1):79–90. doi: 10.1080/00224545.1940.9918734. Ανάκτηση από:

εντοπισμό ψυχολογικών προβλημάτων. Κατά τη δεκαετία του '90, εμφανίζονται οι πρώτες μελέτες που αφορούν στην Ανάλυση Συναισθημάτων με τη χρήση Τεχνητής Νοημοσύνης, όπως για παράδειγμα, η πρόταση του Wiebe (1990)[24] για την ανάπτυξη λογισμικού ανίχνευσης υποκειμενικών χαρακτήρων μέσα σε ένα αφηγηματικό κείμενο.

Το 2002, στο πλαίσιο συνεδρίου για τη χρήση εμπειρικών μεθόδων στην Επεξεργασία Φυσικής Γλώσσας, δημοσιεύτηκε επιστημονικό άρθρο[25] με αξιολογήσεις ταινιών, οι οποίες κατηγοριοποιούνταν με βάση το συναίσθημα που προκαλούσαν (positive ή negative) και όχι το περιεχόμενο τους.

Οι έρευνες των τελευταίων ετών, τείνουν πρός μια ταξινόμηση πολλαπλών ετικετών (multilabel classification) των συναισθημάτων. Το 2014, στο 6ο Διεθνές Συνέδριο Νεφοϋπολογιστικής Τεχνολογίας και Επιστήμης, παρουσιάστηκε πρόταση για την κατηγοριοποίηση των tweet με θετικό πρόσημο, σε μία θετική κατηγορία και αυτών με αρνητικό πρόσημο, σε μία αρνητική κατηγορία[26].

Η Ανάλυση Συναισθημάτων με τη χρήση Τεχνητής Νοημοσύνης, χωρίζεται σε τρεις βασικές κατηγορίες μεθόδων: α) στις Μεθόδους που βασίζονται στην χρήση Λεξικών (Lexicon-based Methods), β) στις Μεθόδους Μηχανικής Μάθησης (Machine Learning Methods) και γ) τις Υβριδικές Μεθόδους (Hybrid Methods). Οι μέθοδοι που βασίζονται στην χρήση Λεξικών, υπολογίζουν τις τιμές των συναισθημάτων μέσα σε προτάσεις και κείμενα, ανάλογα με τις τιμές που δόθηκαν στις λεξικογραφικές μονάδες του λεξικού. Με αυτόν τον τρόπο, ένα κείμενο βαθμολογείται με βάση το άθροισμα των τιμών των λέξεων που το αποτελούν, σε τιμή πολικότητας με θετικό ή αρνητικό πρόσημο. Οι μέθοδοι Μηχανικής Μάθησης, περιλαμβάνουν αλγόριθμους που κατηγοριοποιούν το συναίσθημα σύμφωνα με στατιστικά μοντέλα, αφού πρώτα μετατρέψουν τις προτάσεις σε διανυσματικούς χώρους

https://www.tandfonline.com/doi/abs/10.1080/00224545.1940.9918734?journalCode=vsoc20 Τελευταία πρόσβαση: 31/5/2023.

[24]Wiebe, J. (1990). Recognizing subjective sentences: a computational investigation of narrative text. State University of New York, Buffalo. Ανάκτηση από: https://aclanthology.org/C90-2069.pdf Τελευταία πρόσβαση: 31/5/2023.

[25]Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques.* Proceedings of the ACL-02 conference on empirical methods in natural language processing, Morristown, NJ, USA. Association for Computational Linguistics, pp 79–86. Ανάκτηση από; https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf Τελευταία πρόσβαση: 31/5/2023.

[26]Wang Z, Joo V, Tong C, Xin X, Chin HC (2014). *Anomaly detection through enhanced sentiment analysis on social media data*. In: 2014 IEEE 6th international conference on cloud computing technology and science. IEEE, pp 917–922. Ανάκτηση από: https://www.researchgate.net/publication/275220299_Anomaly_Detection_through_Enhanced_Sentiment_Analysis_on_Social_Media_Data Τελευταία πρόσβαση: 31/5/2023.

(vector spaces). Μερικοί από τους αλγόριθμους Μηχανικής Μάθησης που χρησιμοποιούνται για την Ανάλυση Συναισθημάτων, είναι οι Naive Bayes, Support Vector Machine και Word Embedding. Τέλος, οι Υβριδικές Μέθοδοι, αποτελούν μια μίξη των Μεθόδων Μηχανικής Μάθησης και των Μεθόδων που βασίζονται στη χρήση Λεξικών[27].

Η Ανάλυση των Συναισθημάτων για αρχεία κειμένου, χωρίζεται σε τρία επίπεδα[28]. Σε επίπεδο κειμένου, πρότασης ή λέξης. Στο επίπεδο του κειμένου, δίνεται μία τιμή συναισθήματος, θετική ή αρνητική, για ολόκληρο το κείμενο, η οποία υπολογίζεται από το άθροισμα των τιμών των λέξεων ή προτάσεων που το αποτελούν. Στο επίπεδο της πρότασης ή φράσης, η τιμή του συναισθήματος, υπολογίζεται με βάση το άθροισμα των τιμών που έχει δοθεί για την κάθε επιμέρους λέξη. Για να υπολογιστεί το συναίσθημα και στα τρία επίπεδα, χρησιμοποιούνται οι δύο κατηγορίες μεθόδων που αναφέραμε παραπάνω: οι μέθοδοι που βασίζονται στη χρήση Λεξικού και αυτοί που υπολογίζουν το μοτίβο εμφάνισης των λέξεων με τη χρήση αλγορίθμων Μηχανικής Μάθησης.

Οι προσεγγίσεις της Ανάλυσης Συναισθημάτων που βασίζονται στη χρήση Λεξικών, δίνουν θετικές, αρνητικές ή ουδέτερες τιμές συναισθημάτων σε μία λέξη, πρόταση ή κείμενο, ενώ οι μέθοδοι Μηχανικής Μάθησης, προσδίδουν αριθμητικές τιμές πολικοτητας με θετικό ή αρνητικό πρόσημο[29].

Ενώ η Ανάλυση Συναισθημάτων, καταχωρεί τα κείμενα σε θετικά, αρνητικά και ουδέτερα ή τους προσδίδει αριθμητικές τιμές πολικότητας, ο Εντοπισμός Συναισθήματος (Emotion Detection), μπορεί και ανιχνεύει τα διακριτά ανθρώπινα συναισθήματα που βρίσκονται μέσα στα κείμενα, τα οποία και κατηγοριοποιεί βάσει αυτών[30].

Υπάρχουν διάφορα μοντέλα Εντοπισμού Συναισθήματος που χρησιμοποιούνται, με ένα από τα πιό γνωστά, το μοντέλο κατηγοριοποίησης συναισθημάτων του Ekman.

---

[27]Yilmaz, B. (2023, January 15). Sentiment Analysis Methods in 2023: Overview, Pros & Cons. *AIMultiple.* Ανάκτηση από: https://research.aimultiple.com/sentiment-analysis-methods/ Τελευταία πρόσβαση: 31/5/2023.

[28]Kharde, V. & Sonawane, S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications.* 139. 5-15. 10.5120/ijca2016908625. Ανάκτηση από: https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf Τελευταία πρόσβαση: 31/5/2023.

[29]Kharde & Sonawane. (2016), ό.π.

[30]Nandwani, P. & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining.* 11.10.1007/s13278-021-00776-6. Ανάκτηση από: https://www.researchgate.net/publication/354196437_A_review_on_sentiment_analysis_and_emotion_detection_from_text Τελευταία πρόσβαση: 31/5/2023.

Ο Αμερικανός ψυχολόγος Paul Ekman, μελετώντας προηγούμενες έρευνες πάνω στην έκφραση των συναισθημάτων, κατέληξε στη θεωρία ότι υπάρχουν 6 βασικά συναισθήματα, τα οποία απαντώνται σε όλους τους ανθρώπινους πολιτισμούς και εκφράζονται μέσω των κινήσεων και των εκφράσεων του προσώπου. Τα συναισθήματα αυτά, είναι ο Φόβος (Fear), ο Θυμός (Anger), η Χαρά (Joy), η Λύπη (Sadness), η Απέχθεια (Disgust) και η Έκπληξη (Surprise)[31].

Το μοντέλο κατηγοριοποίησης του Ekman, βασίζεται στη θεωρία του Charles Darwin ότι τα ανθρώπινα Συναισθήματα είναι πρωτόγονα ή θεμελιώδη και απαντώνται σε όλους τους πολιτισμούς[32]. Ο Ekman πρότεινε το παραπάνω μοντέλο για να διαχωρίσει ξεχωριστά συναισθήματα και συναισθηματικά φαινόμενα που παρουσιάζονται στους ανθρώπους και τα υπόλοιπα πρωτεύοντα[33].

1.2. Θεωρία της Συσταδοποίησης Κειμένων

Η Συσταδοποίηση Κειμένων[34], είναι μια μέθοδος μάθησης χωρίς επίβλεψη, που χρησιμοποιείται για τη συγκέντρωση κειμενικών αρχείων σε συστάδες, με βάση την ομοιότητα του περιεχομένου τους, έτσι ώστε όσα κείμενα ανήκουν στην ίδια συστάδα, να ομοιάζουν μεταξύ τους και να παρουσιάζουν διαφορές σε σχέση με τα κείμενα των υπολοίπων συστάδων. Διαφέρει από άλλες μεθόδους μάθησης με επίβλεψη, (π.χ. Κατηγοριοποίηση), διότι η μέθοδος αυτή, δεν προσδίδει από πριν κάποιο χαρακτηριστικό γνώρισμα στα κειμενικά δεδομένα, αλλά τα καταχωρεί - έπειτα από ειδικές μετρήσεις - σε περισσότερο ή λιγότερο ομοιογενή υποσύνολα, τις συστάδες[35]. Η μέθοδος της

[31]Ekman, P. (2003, December). *Emotions inside out. 130 Years after Darwin's "The Expression of the Emotions in Man and Animal".* Ann N Y Acad Sci. 2003 Dec;1000:1-6. doi: 10.1196/annals.1280.002. Ανάκτηση από: https://pubmed.ncbi.nlm.nih.gov/15045707/ Τελευταία πρόσβαση: 31/5/2023.

[32]Darwin, C. (1872). *The expression of the emotions in man and animals*. Oxford University Press, USA. Ανάκτηση από: https://pure.mpg.de/rest/items/item_2309885/component/file_2309884/content Τελευταία πρόσβαση: 31/5/2023.

[33]Cai, Y., Li, X., & Li, J. (2023). Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. *Sensors*, *23*(5), 2455. MDPI AG. http://dx.doi.org/10.3390/s23052455 Ανάκτηση από: https://www.mdpi.com/1424-8220/23/5/2455 Τελευταία επίσκεψη: 31/5/2023.

[34]Yoo, I., & Hu, X. (2006). *A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE*. Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06*)*, 220-229. Ανάκτηση από: https://www.semanticscholar.org/paper/A-comprehensive-comparison-study-of-document-for-a-Yoo-Hu/08eafbff40575455e0c89a76245320026af76f14 Τελευταία πρόσβαση: 31/5/2023.

[35]Backer, E. & Jain, A.K. (1981, Jan) A Clustering Performance Measure Based on Fuzzy Set Decomposition In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-3, no. 1, pp. 66-75, Jan. 1981, doi: 10.1109/TPAMI.1981.4767051. Ανάκτηση από: https://ieeexplore.ieee.org/abstract/document/4767051 Τελευταία πρόσβαση: 31/5/2023.

Συσταδοποίησης Κειμένων, η οποία ανήκει στη σφαίρα της Επεξεργασίας Φυσικής Γλώσσας, είναι ιδιαίτερα χρήσιμη, καθώς έχει τη δυνατότητα να αναλύει μεγάλο όγκο κειμένου και να το καταχωρεί σε ομάδες, ανάλογα με το περιεχόμενο και τον βαθμό που μοιάζουν μεταξύ τους, ανακαλύπτοντας έτσι κρυμμένα μοτίβα και σχέσεις μεταξύ των κειμένων.

Η θεωρητική βάση πάνω στην οποία στηρίχθηκε και αναπτύχθηκε η μέθοδος της Συσταδοποίησης των Κειμένων, είναι η Θεωρία των Συνόλων, η οποία αποτελεί κλάδο της Μαθηματικής Λογικής. Η Θεωρία των Συνόλων, αφορά στη μελέτη των αντικειμένων που ανήκουν σε σύνολα (sets), γνωστά και ως μέλη ή στοιχεία των συνόλων. Σύμφωνα με αυτή τη θεωρία, τα στοιχεία που ομαδοποιούνται και χαρακτηρίζονται (labeled) με βάση την ομοιότητά τους, εντάσσονται σε κατηγορίες (classes) ή συστάδες (clusters)[36].

Οι μετρήσεις των Αποστάσεων / Ομοιότητας, είναι ο πυρήνας των μεθόδων της Συσταδοποίησης Κειμένων και της Κατηγοριοποίησης. Αυτές οι μετρήσεις λειτουργούν λαμβάνοντας υπόψη την παρουσία ή την απουσία των λέξεων και αξιολογώντας την απόσταση μεταξύ των διανυσμάτων δύο κειμενικών αρχείων. Εφόσον αποδειχθεί ότι τα δύο αρχεία παρουσιάζουν ομοιότητες, συγκροτούν ομάδα. Κάποιες από τις μετρήσεις ομοιότητας που χρησιμοποιούνται ευρέως, είναι η Ευκλείδεια Απόσταση (Euclidean Distance), η Απόσταση Συνημιτόνου (Cosine Distance) και η Απόσταση Manhattan[37].

Υπάρχουν δύο βασικές μέθοδοι, με τις οποίες γίνεται η Συσταδοποίηση ενός Σώματος Κειμένου, η αυστηρή (hard clustering) και η ελαστική συσταδοποίηση (soft clustering)[38]. Στην πρώτη περίπτωση, το κάθε αντικείμενο ανήκει σε ακριβώς μία συστάδα, ενώ στη δεύτερη περίπτωση, το ίδιο αντικείμενο μπορεί να ανήκει σε παραπάνω από μία συστάδες.

[36]Vidhya, K.A. & Geetha, T.V. (2017, Jan. 1). Rough Set Theory For Document Clustering: A Review. *Journal of Intellingent & Fuzzy Systems,* vol. 32, no. 3, pp. 2165-2185, 2017. Ανάκτηση από: https://www.researchgate.net/profile/Ka-Vidhya/publication/314125074_Rough_set_theory_for_document_clustering_A_review/links/60c83a7f299bf108abd975e1/Rough-set-theory-for-document-clustering-A-review.pdf Τελευταία πρόσβαση: 31/5/2023.

[37]Amer, A.A. & Abdalla, H.I. (2020) A set theory based similarity measure for text clustering and classification. *Journal of Big Data* 7, 74 (2020). https://doi.org/10.1186/s40537-020-00344-3 Ανάκτηση από: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00344-3#citeas Τελευταία πρόσβαση: 31/5/2023.

[38]Koch, K., (2020, March 26). A Friendly Introduction to Text Clustering. *Towards Data Science.* 26 Mar. 2020. Ανάκτηση από: https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04 Τελευταία πρόσβαση: 31/5/2023.

Για να εφαρμοστεί Συσταδοποίηση Κειμένων, μπορούν να χρησιμοποιηθούν διάφοροι αλγόριθμοι, όπως αυτός της Ιεραρχικής Συσταδοποίησης (Hierarchical Clustering)[39].

Η Ιεραρχική Συσταδοποίηση, αναπαρίσταται με τη χρήση ενός Δενδρογράμματος, στο οποίο τα φύλλα είναι οι συστάδες και οι αλγόριθμοι που χρησιμοποιεί χωρίζονται σε συγκεντρωτικούς (agglomerative) και διχαστικούς (divisive). Στους συγκεντρωτικούς αλγόριθμους, υπάρχει αρχικά ένας Ν αριθμός συστάδων, οι οποίες επαναλαμβανόμενα συγχωνεύονται έως ότου δημιουργηθεί μία κοινή συστάδα. Αντιθέτως, στους διχαστικούς αλγόριθμους, υπάρχει μία αρχική, κοινή συστάδα, η οποία με τον υπολογισμό κάποιων πράξεων, υποδιαιρείται σε μικρότερες συστάδες, μέχρις ότου η κάθε συστάδα να περιέχει ακριβώς ένα στοιχείο[40].

Η Συσταδοποίηση Κειμένων, μπορεί να γίνει σε τρία επίπεδα: κειμένου, πρότασης και λέξης. Ως μέθοδος ταξινόμησης και κατανόησης κειμενικού περιεχομένου, ακολουθεί συγκεκριμένα στάδια μέχρι την ολοκλήρωσή της, όπως για παράδειγμα, την προεπεξεργασία του κειμένου, την εξαγωγή χαρακτηριστικών, τη μέτρηση της ομοιότητας μεταξύ των κειμένων, την εφαρμογή του αλγόριθμου συσταδοποίησης και τέλος, την αξιολόγηση των συστάδων. Στόχος αυτών των βημάτων, είναι η μετατροπή των μη επεξεργασμένων κειμενικών δεδομένων σε αριθμητικές αναπαραστάσεις, οι οποίες μπορούν να μετρηθούν και στην συνέχεια να χρησιμοποιηθούν για τον υπολογισμό των αποστάσεων μεταξύ των κειμένων, με τελικό αποτέλεσμα την κατανομή αυτών σε συστάδες[41].

Με βάση τα παραπάνω, το πρώτο στάδιο της Συσταδοποίησης Κειμένου, αποτελείται από την προεπεξεργασία των δεδομένων, με την απομάκρυνση άχρηστων στοιχείων, όπως αριθμών και συμβόλων, την αφαίρεση τονισμού και τη μετατροπή του κειμένου σε πεζούς χαρακτήρες. Έπειτα, στα επεξεργασμένα δεδομένα, δίδονται αριθμητικά χαρακτηριστικά, που αφορούν π.χ. τον αριθμό εμφάνισης των λέξεων ή τη συχνότητά τους εντός ενός σώματος κειμένου. Αφού εξαχθούν τα χαρακτηριστικά αυτά, μετρώνται οι αποστάσεις

[39]Suyal, H., Panwar, A. & Negi, A. S. (2014). Text Clustering Algorithms: A Review. *International Journal of Computer Applications* (0975 – 8887) Volume 96 – No.24. June 2014. Ανάκτηση από: https://research.ijcaonline.org/volume96/number24/pxc3897075.pdf Τελευταία πρόσβαση: 31/5/2023.
[40]Day, W.H. and Edelsbrunner, H. (1984) Efficient Algorithms for Agglomerative Hierarchical Clustering Methods. *Journal of Classification*, 1, 1-24. http://dx.doi.org/10.1007/BF01890115 Ανάκτηση από: https://link.springer.com/article/10.1007/bf01890115 Τελευταία πρόσβαση: 31/5/2023.
[41]Miroshnyk, O. (2023, Feb 2). What is Text Clustering? *oneai.* 2 Feb. 2023. Ανάκτηση από: https://www.oneai.com/learn/text-clustering Τελευταία πρόσβαση: 31/5/2023.

μεταξύ των κειμένων και στο τέλος γίνεται η χρήση του κατάλληλου αλγόριθμου Συσταδοποίησης[42]. Για τις ανάγκες της παρούσας έρευνας, επιλέχθηκε ο αλγόριθμος της Ιεραρχικής Συσταδοποίησης, έναντι άλλων μεθόδων, για λόγους που εξηγούνται στη συνέχεια.

Ο αλγόριθμος της Ιεραρχικής Συσταδοποίησης, έχει περισσότερα πλεονεκτήματα σε σχέση με τους υπόλοιπους, καθώς μπορεί και παρέχει επαρκείς πληροφορίες σχετικά με το ποιές συστάδες κειμένων ομοιάζουν με άλλες, δίνει παρόμοια αποτελέσματα όσες φορές και αν τρέξει ο ίδιος αλγόριθμος υπολογισμού, οι ομαδοποιήσεις των κειμένων γίνονται εύκολα ανιχνεύσιμες με μια απλή παρατήρηση, ανεξάρτητα από την πληθώρα των κειμένων που μελετώνται και των συστάδων που διαμορφώνονται, αποτελεί μια πολύ καλά μελετημένη από τους ερευνητές μέθοδος υπολογισμού και τέλος, μπορεί εύκολα να προσαρμοστεί για να δεχθεί μεταβλητές κατηγοριοποίησης[43].

Τα βασικά μειονεκτήματα της Ιεραρχικής Συσταδοποίησης, είναι ότι είναι ευαίσθητη στο θόρυβο και τις ακραίες τιμές, ενώ δεν υπάρχουν κάποια κριτήρια για τα όρια του τερματισμού της[44].

1.3. Θεωρία της Σημασιολογικής Ανάλυσης

Η Σημασιολογική Ανάλυση[45], είναι μια μέθοδος ανάλυσης κειμένου, η οποία χρησιμοποιείται για την εξαγωγή νοήματος. Αποτελεί σημαντικό εργαλείο της Φυσικής Επεξεργασίας Γλώσσας και έχει ευρεία χρήση στις νέες τεχνολογίες, όπως τις μηχανές αναζήτησης, τα chatbots κλπ.

Με τη μέθοδο αυτή, δίνεται η δυνατότητα στους υπολογιστές να κατανοήσουν και να αντιληφθούν προτάσεις, παραγράφους ή ολόκληρα αρχεία κειμένων, αναλύοντας τη γραμματική τους δομή και αναγνωρίζοντας τις σχέσεις μεταξύ ξεχωριστών λέξεων, μέσα σε ένα συγκεκριμένο πλαίσιο. Ωστόσο, η Σημασιολογική Ανάλυση διαφέρει από τη Λεξιλογική

[42]Miroshnyk, O., ό.π.
[43]Ellis, C., When to use hierarchical clustering. *Crunching the Data.* Ανάκτηση από: https://crunchingthedata.com/when-to-use-hierarchical-clustering/ Τελευταία πρόσβαση: 31/5/2023.
[44]Xu, R. & Wunsch, D. (2005, June). Survey of Clustering Algorithms. *Neural Networks, IEEE Transactions* on. 16. 645-678. DOI: 10.1109/TNN.2005.845141. Ανάκτηση από: https://www.researchgate.net/publication/3303538_Survey_of_Clustering_Algorithms Τελευταία πρόσβαση: 31/5/2023.
[45]Liza, U., Semantic Analysis: What Is It, How It Works + Examples. *QuestionPro*. Ανάκτηση από: https://www.questionpro.com/blog/semantic-analysis/ Τελευταία πρόσβαση: 31/5/2023.

(Lexical Analysis), καθώς βασίζεται στην ανάλυση μεγαλύτερων μορφών κειμένου, ενώ η δεύτερη στην ανάλυση των λέξεων σε επίπεδο μονάδας[46].

Καθώς οι υπολογιστές δεν διαθέτουν κάποιο σύστημα κατανόησης λέξεων, προτάσεων ή κειμένων, η Σημασιολογική Ανάλυση έρχεται να καλύψει αυτό το κενό, με μια σειρά αναλύσεων και ενεργειών.

Αρχικά, η Λεξιλογική Ανάλυση, διαβάζει τους χαρακτήρες των λέξεων και τις καταχωρεί ως λεξικογραφικές μονάδες. Στη συνέχεια, η Γραμματική Ανάλυση συσχετίζει την ακολουθία των λέξεων και εφαρμόζει γραμματικούς κανόνες που χρησιμοποιούνται στην κάθε γλώσσα, ώστε να πραγματοποιηθεί επισήμανση των μερών του λόγου (pos tagging). Τέλος, η Συντακτική Ανάλυση, αναλύει τη σύνταξη και τροποποιεί το κείμενο, ώστε να υπάρχει κάποια συνοχή σε επίπεδο λέξεων, προτάσεων και κειμένου. Όλες οι παραπάνω ενέργειες, βοηθούν τα υπολογιστικά συστήματα στο να κατανοούν το κάθε φορά δοθέν κείμενο, με τον ίδιο τρόπο που το αντιλαμβάνονται και οι άνθρωποι[47].

Κάποια από τα σημαντικά στοιχεία της Σημασιολογικής Ανάλυσης είναι η Υπωνυμία (Hyponymy), η Ομωνυμία (Homonymy), η Πολυσημία (Polysemy), η Συνωνυμία (Synonymy) και η Αντωνυμία (Antonymy). Υπωνυμία ονομάζεται η σχέση μεταξύ ενός γενικού όρου και των στιγμών που συγκροτούν αυτό τον όρο (για παράδειγμα, η λέξη "*κόκκινο"* είναι υπώνυμο και αποτελεί στιγμή του όρου "*χρώμα*", που είναι υπέρνυμο). Η Ομωνυμία χαρακτηρίζει λέξεις που έχουν την ίδια μορφή, αλλά διαφορετική σημασία. Ο όρος Πολυσημία χρησιμοποιείται για λέξεις που έχουν διαφορετικό νόημα, αλλά σχετίζονται. Τέλος, η Συνωνυμία αφορά διαφορετικές λέξεις που έχουν το ίδιο νόημα, ενώ η Αντωνυμία αναφέρεται σε λέξεις που είναι αντίθετες μεταξύ τους. Με βάση τα παραπάνω, η Σημασιολογική Ανάλυση είναι η μέθοδος που ενώνει μεταξύ τους οντότητες, έννοιες, σχέσεις και προβλέψεις, προκειμένου να περιγράψει μια κατάσταση και να αντλήσει νόημα, μέσα από μια ακολουθία λέξεων[48].

---

[46]Goyal, C., (2021, June 23). Part 9: Step by Step Guide to Master NLP - Semantic Analysis. *Analytics Vidhya.* Ανάκτηση από: https://www.analyticsvidhya.com/blog/2021/06/part-9-step-by-step-guide-to-master-nlp-semantic-analysis/ Τελευταία πρόσβαση: 31/5/2023.

[47]Expert.ai Team. (2022, April 26). What Semantic Analysis Means to Natural Language Processing. *expert.ai.* Ανάκτηση από: https://www.expert.ai/blog/natural-language-process-semantic-analysis-definition/ Τελευταία πρόσβαση: 31/5/2023.

[48]tutorialspoint. *Natural Language Processing - Semantic Analysis.* Ανάκτηση από: https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_semantic_analysis.htm Τελευταία πρόσβαση: 31/5/2023.

Η Σημασιολογική Ανάλυση έχει ευρεία χρήση στις νέες τεχνολογίες, με διαρκώς αναβαθμιζόμενο ρόλο, μετά και την καθιέρωση των ΜΚΔ στην καθημερινή ζωή και την οικονομία. Τα αδόμητα δεδομένα αποτελούν συχνά πρόβλημα για επιχειρήσεις και φορείς, μιας και η ακατέργαστη μορφή τους δεν μπορεί να παρέχει επαρκείς πληροφορίες για τον αντίκτυπο των υπηρεσιών που αυτές προσφέρουν. Κάθε αλληλεπίδραση με το κοινό είναι πολύτιμη και το περιεχόμενο των πληροφοριών που εξάγονται θα πρέπει να γίνεται κατανοητό. Εκεί έρχεται η Σημασιολογική Ανάλυση, προσφέροντας λύσεις, με εργαλεία που μετατρέπουν τα αδόμητα δεδομένα, σε εύληπτες, μετρήσιμες και διαχειρίσιμες πληροφορίες[49].

1.4. Θεωρία της Θεματικής Μοντελοποίησης

Με τον ολοένα και αυξανόμενο όγκο των αδόμητων δεδομένων που παράγονται καθημερινά από τη χρήση των νέων τεχνολογιών και ιδιαίτερα των ΜΚΔ, οργανισμοί, εταιρείες και κυβερνήσεις κρίνουν ως αναγκαία την εξαγωγή πληροφοριών από αυτά, προκειμένου να τις χρησιμοποιήσουν για οικονομικούς ή άλλους σκοπούς. Μια από τις τεχνικές εξόρυξης δεδομένων που χρησιμοποιείται ευρέως, είναι η Θεματική Μοντελοποίηση[50], δηλαδή η διαδικασία ανίχνευσης των κρυμμένων θεματικών που υπάρχουν εντός ενός κειμένου και η αξιοποίησή τους για την εξαγωγή νοήματος, η οποία θα οδηγήσει στην λήψη των κατάλληλων αποφάσεων.

Βασικός σκοπός της Θεματικής Μοντελοποίησης είναι ο καθορισμός των δομών που κρύβονται μέσα σε συλλογές κειμένων. Με αυτό το σκεπτικό, αναπτύχθηκαν από νωρίς διάφορα μοντέλα εύρεσης πληροφοριών, όπως το μοντέλο διανυσματικού χώρου αυτοματοποιημένης καταχώρησης που πρότειναν οι Salton, κ.α. (1975)[51], ώστε να καταστήσουν την περιήγηση σε ένα μεγάλο όγκο κειμένων πιό εύκολη.

Ως μέθοδος κειμενικής ανάλυσης χωρίς επιτήρηση, η Θεματική Μοντελοποίηση, αναλύει συλλογές κειμένων με βάση τις χαρακτηριστικές τους λέξεις, με σκοπό να εξάγει το

[49]O' Neal, H. (2022, June 23). Semantics NLP. *MetaDialog.* Ανάκτηση από: https://www.metadialog.com/blog/semantic-analysis-in-nlp/ Τελευταία πρόσβαση: 31/5/2023.
[50]Bansal, S., (2016, August 24). Beginners Guide to Topic Modelling in Python. *Analytics Vidhya*. Ανάκτηση από: https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/ Τελευταία πρόσβαση: 31/5/2023.
[51]Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing, *Communications of the ACM* 18 (11), pp. 613-620. Ανάκτηση από: https://dl.acm.org/doi/pdf/10.1145/361219.361220 Τελευταία πρόσβαση: 31/5/2023.

νόημά τους, σε συνάρτηση με το ευρύτερο πλαίσιο της χρήσης των λέξεων αυτών στη φυσική γλώσσα.

Παρά την ομοιότητά της με άλλες μεθόδους εξόρυξης δεδομένων, όπως π.χ. τη Συσταδοποίηση Κειμένου, η Θεματική Μοντελοποίηση διαφέρει από αυτές, κυρίως στο ότι χρησιμοποιεί μία μίξη μεθόδων, οι οποίες δεν καταχωρούν ένα κείμενο σε μία μόνο θεματική, αλλά προβλέπουν την πιθανότητα να ανήκει σε περισσότερες[52].

Βασικές εφαρμογές της Θεματικής Μοντελοποίησης, είναι η ταξινόμηση, η κατηγοριοποίηση και η περίληψη των κειμένων. Οι πιό γνωστοί αλγόριθμοι που χρησιμοποιούνται για τις ενέργειες της Θεματικής Μοντελοποίησης, είναι η Λανθάνουσα Σημασιολογική Ανάλυση (LSA), η Πιθανότητα Λανθάνουσας Σημασιολογικής Ανάλυσης (PLSA), η Λανθάνουσα Κατανομή Dirichlet (LDA) και η Ιεραρχική Διαδικασία Dirichlet (HDP)[53].

Από τους πιό εφαρμοσμενους αλγόριθμους Θεματικής Μοντελοποίησης, είναι η Λανθάνουσα Κατανομή Dirichlet[54]. Ο αλγόριθμος αυτός, χρησιμοποιεί την κατανομή Dirichlet, η οποία του δίνει τη δυνατότητα να προβλέπει τις πιθανότητες ενός κειμένου να ανήκει σε κάποια θεματική, με βάση τις λέξεις που το αποτελούν. Η μέθοδος LDA, υποθέτει, ότι κάθε κείμενο περιέχει ένα μείγμα θεμάτων και κάθε θέμα αποτελείται από ένα μείγμα λέξεων - κλειδιών που το χαρακτηρίζουν.

Καθώς με τη χρήση της Θεματικής Μοντελοποίησης γίνεται εύκολη η εξαγωγή νοήματος μέσα από μεγάλα σώματα κειμένων, με αντικειμενικό τρόπο και εξοικονόμηση χρόνου, πολλοί ερευνητές, επιχειρήσεις και φορείς επιλέγουν τη μέθοδο αυτή για να εξάγουν και να αναλύσουν αδόμητα δεδομένα, καθώς και για να εντοπίσουν τις κρυμμένες θεματικές που βρίσκονται εντός τους. Μερικοί από τους τομείς στους οποίους η Θεματική Μοντελοποίηση έχει ευρεία εφαρμογή, είναι οι επιστημονικές έρευνες, η δημοσιογραφία, η καταγραφή των ΜΚΔ και οι διαφημίσεις[55].

---

[52]Bail, C., Topic Modeling. *Github.* Ανάκτηση από: https://cbail.github.io/SICSS_Topic_Modeling.html Τελευταία πρόσβαση: 31/5/2023.

[53]cogitotech, (2021, September 27). Topic Modeling: Algorithms, Techniques, and Application. *Data Science Central.* Ανάκτηση από: https://www.datasciencecentral.com/topic-modeling-algorithms-techniques-and-application/ Τελευταία πρόσβαση: 31/5/2023.

[54]Ipshita. (2021, Jul. 16). Topic Modeling using LDA. *Analytics Vidhya*. Ανάκτηση από: https://medium.com/analytics-vidhya/topic-modelling-using-lda-aa11ec9bec13 Τελευταία πρόσβαση: 31/5/2023.

[55]Peddireddi, Y., (2021, May 1). Topic Modelling in Natural Language Processing. *Analytics Vidhya.* Ανάκτηση από: https://www.analyticsvidhya.com/blog/2021/05/topic-modelling-in-natural-language-processing/ Τελευταία πρόσβαση: 31/5/2023.

Κεφάλαιο 2: Μεθοδολογία

2.1. Εξαγωγή, Καθαρισμός και Προετοιμασία των Δεδομένων

Η εξαγωγή των δύο συνόλων δεδομένων που χρησιμοποιήθηκαν στην παρούσα έρευνα, έγινε με τη χρήση της βιβλιοθήκης Snscrape, η οποία λειτουργεί με τη γλώσσα προγραμματισμού Python.

Να σημειωθεί, ότι το Snscrape μπορεί να εγκατασταθεί και να "τρέξει" με δύο μεθόδους, είτε μέσω τερματικού είτε μέσω του σημειωματαρίου Jupyter. Για την εξαγωγή των δεδομένων, χρησιμοποιήσαμε τη δεύτερη επιλογή και συγκεκριμένα το Jupyter σημειωματάριο Google Colab. Τέλος, ο κώδικας που συντάξαμε, βασίστηκε στο αποθετήριο του JustAnotherArchivist[56] στο GitHub.

Με τις εντολές αυτού του κώδικα, όπως εμφανίζεται στο Παράρτημα της Εργασίας, αποκτήσαμε τα σύνολα δεδομένων που αναλύονται στην παρούσα έρευνα. Έτσι λοιπόν, εξάγαμε δύο αρχεία, με τις ονομασίες atsipras_tweets_Original.csv και mentions_to_atsipras_Original.csv. Τα αρχεία αυτά, τα αποθηκεύσαμε σε έναν υποφάκελο του Tweet_Folder των παραδοτέων αρχείων, με την ονομασία Original_Data. Στο επόμενο βήμα, επεξηγείται η διαδικασία ενός αρχικού καθαρισμού τους, ώστε να γίνουν αξιοποιήσιμα για ανάλυση.

Ο καθαρισμός των δεδομένων αποτελεί ένα σημαντικό βήμα για τη διαδικασία της Εξόρυξης Δεδομένων, καθώς αφαιρεί μη αξιοποιήσιμα στοιχεία από αυτά, βελτιώνοντας έτσι την ποιότητά τους και προλειαίνοντας τον δρόμο για το επόμενο στάδιο, αυτό της Ανάλυσης. Σε ένα σύνολο δεδομένων, υπάρχουν πολλά περιττά στοιχεία που πρέπει να καθαριστούν, όπως τυπογραφικά λάθη, ατέλειες στη δομή, ελλειπή ή διπλότυπα δεδομένα

[56]JustAnotherArchivist. (2020). snscrape. *GitHub.* Ανάκτηση από: https://github.com/JustAnotherArchivist/snscrape Τελευταία πρόσβαση: 31/5/2023.

κλπ[57]. Ο καθαρισμός των δεδομένων, μπορεί να γίνει χειροκίνητα μέσω εντολών ή με την χρήση ειδικού λογισμικού. Στην προκειμένη περίπτωση, χρησιμοποιήσαμε απλές εντολές, αφού πρώτα εισάγαμε τα δύο σύνολα δεδομένων στο Google Sheets για επισκόπηση και επεξεργασία.

Αρχικά, ας δούμε από τι στοιχεία αποτελούνται αυτά τα δύο σύνολα δεδομένων. Το atsipras_tweets_Original.csv, που αποκτήσαμε με τη χρήση της βιβλιοθήκης Snscrape, αποτελείται από 8.467 σειρές, δηλαδή από 8.466 tweets και 13 στήλες, με τις ονομασίες Datetime, Tweet Id, Text, Lang, Source, Media, Hashtags, Reply Count, Retweet Count, Like Count, Quote Count, Mentioned Users και Username.

Στη στήλη Datetime, διαπιστώνουμε οτι τα tweets φτάνουν μέχρι τις 31 Μαρτίου του 2023, ημέρα που αποκτήσαμε αυτό το σύνολο δεδομένων. Καθώς θέλουμε να κρατήσουμε μόνο τα tweets που αναρτήθηκαν έως τις 31 Δεκεμβρίου του 2022, διαγράψαμε τις σειρές που περισσεύουν και αφαιρέσαμε από την στήλη Datetime την ώρα, κρατώντας μόνο την ημερομηνία.

Τα περιεχόμενα των στηλών Tweet Id, Source, Media, Hashtags, Reply Count, Retweet Count, Like Count, Quote Count και Mentioned Users δεν μας απασχολούν στην παρούσα έρευνα, γι' αυτό και διαγράφηκαν. Το "καθαρό" αρχείο, το ονομάσαμε atsipras_tweets_Cleaned.csv και το αποθηκεύσαμε στο φάκελο Cleaned_Tweets του Tweet_Folder, των παραδοτέων με τη Διπλωματική αρχείων.

Το αρχείο mentions_to_atsipras_Original.csv, αποτελείται από 21.483 σειρές, δηλαδή 21.482 mentions και 7 στήλες, με τις ονομασίες Datetime, Tweet Id, Text, Lang, Hashtags, Mentioned Users και Username.

Όπως έγινε και με το προηγούμενο σύνολο δεδομένων, με βάση τη στήλη Datetime, διαγράψαμε όλα εκείνα τα mentions που έγιναν προς τον @atsipras μετά τις 31 Δεκεμβρίου του 2022 και κρατήσαμε μόνο τις ημερομηνίες.

Οι στήλες Tweet Id, Hashtags και Mentioned Users από αυτό το σύνολο δεδομένων δεν αφορούν την έρευνα μας, γι' αυτό και τις αφαιρέσαμε. Το νέο αρχείο, το ονομάσαμε mentions_to_atsipras_Cleaned.csv και με τη σειρά του το αποθηκεύσαμε στο φάκελο Cleaned_Tweets.

[57]Faraz, M. (2022, May 16). Data Cleaning in Data Mining Simplified 101. *HEVO.* Ανάκτηση από: https://hevodata.com/learn/data-cleaning-in-data-mining-simplified-101/ Τελευταία πρόσβαση: 31/5/2023.

Προτού αναλύσουμε αυτά τα δεδομένα με το λογισμικό του Orange, χρειάστηκε να τα επεξεργαστούμε για ακόμα μια φορά, με σκοπό την αφαίρεση εκείνων των στοιχείων που μπορούν να αποτελέσουν "θόρυβο" κατά τη διαδικασία της ανάλυσης.

Παρατηρόντας τα δύο σύνολα δεδομένων, atsipras_tweets_Cleaned και mentions_to_atsipras_Cleaned, εντοπίσαμε με βάση τη στήλη Lang, ότι πολλά tweets και mentions είναι γραμμένα σε διάφορες γλώσσες πέραν της ελληνικής και της αγγλικής, συμπεριλαμβανόμενων των greeklish, άγνωστων γλωσσών, συνδέσμων, καθώς και συνδυασμών των ανωτέρω.

Το λογισμικό του Orange, όπως θα δούμε παρακάτω, χρησιμοποιεί λεξικά για ενέργειες Μηχανικής Μάθησης (ML) και Επεξεργασίας Φυσικής Γλώσσας, τα οποία περιέχουν λέξεις από υπαρκτές γλώσσες. Συνεπώς, δεν μπορούν να κατανοηθούν με επιτυχία κείμενα που είναι εξ' ολοκλήρου γραμμένα π.χ. σε greeklish.

Για να ξεπεράσουμε αυτό το εμπόδιο, σε κάθε σύνολο δεδομένων, αφαιρέσαμε όσα κείμενα ήταν γραμμένα σε άλλες γλώσσες, πέραν της ελληνικής και της αγγλικής.

Τα τελικά σύνολα δεδομένων που αξιοποιήσαμε στην έρευνά μας, τα ονομάσαμε atsipras_tweets_Orange και mentions_to_atsipras_Orange, αντίστοιχα και τα αποθηκεύσαμε στο φάκελο Spreadsheets_for_Orange του φακέλου Orange_Folder, των παραδοτέων αρχείων της παρούσας εργασίας.

2.2. Πραγματοποίηση της Ανάλυσης Συναισθημάτων

Για να εντοπίσουμε τα Συναισθήματα στα δύο σύνολα δεδομένων που αποκτήσαμε, χρησιμοποιήσαμε το μοντέλο κατηγοριοποίησης συναισθημάτων που πρότεινε ο Ekman, το οποίο και αναλύσαμε στο υποκεφάλαιο "*Θεωρία της Ανάλυσης Συναισθημάτων*".

Το Orange διαθέτει ένα ειδικό εργαλείο για τον εντοπισμό των Συναισθημάτων μέσα από κείμενα tweets, το Tweet Profiler[58]. Το εργαλείο αυτό, στέλνει τα tweets σε ένα server, ο οποίος υπολογίζει την πιθανότητα ή και την τιμή του Συναισθήματος για το κάθε ένα από αυτά τα κείμενα και με βάση αυτό τον υπολογισμό, τα κατηγοριοποιεί σε ένα από τα έξι βασικά συναισθήματα που διέκρινε ο Ekman.

[58]Orange3 Text Mining. *Tweet Profiler.* Ανάκτηση από: https://orange3-text.readthedocs.io/en/latest/widgets/tweetprofiler.html Τελευταία πρόσβαση: 3/5/2023.

Για να αναλύσουμε τα tweets του @atsipras και τα mentions που έγιναν πρός αυτόν, δημιουργήσαμε δύο ροές εργασιών, μια για το κάθε σύνολο δεδομένων. Πριν προχωρήσουμε στον εντοπισμό των Συναισθημάτων, επεξεργαστήκαμε τα κειμενικά δεδομένα με το εργαλείο της Προεπεξεργασίας Κειμένου, κατατμίζοντας τις λεξικογραφικές μονάδες ανά tweet, μετατρέποντας τους χαρακτήρες σε πεζούς και αφαιρώντας τονισμό, αριθμούς, διευθύνσεις urls, ετικέτες html και κανονικές εκφράσεις (regex).

Έπειτα, τα αποτελέσματα των κατηγοριοποιήσεων με βάση το Συναίσθημα, τα εμφανίσαμε σε ένα Πίνακα Δεδομένων και τα αποθηκεύσαμε σε αρχείο μορφής .xlsx, ώστε αργότερα να μπορούμε να οπτικοποιήσουμε τις διακυμάνσεις στην κατανομή τους, ανάλογα με την ημερομηνία κατά την οποία αναρτήθηκαν στο Twitter.

2.3. Πραγματοποίηση της Συσταδοποίησης Κειμένων

Η Συσταδοποίηση των Κειμένων στην παρούσα έρευνα, έγινε με τη χρήση του αλγορίθμου της Ιεραρχικής Συσταδοποίησης. Για την εφαρμογή της, ακολουθήσαμε συγκεκριμένες διαδικασίες, που περιλαμβάνουν μετρήσεις και χρησιμοποιήσαμε ειδικά εργαλεία του Orange, όπως το Bag of Words (BoW) και οι Αποστάσεις (Distances).

Το μοντέλο Bag of Words[59], είναι μία μέθοδος κατηγοριοποίησης κειμένου, η οποία μετράει τις φορές που εμφανίζεται η κάθε λέξη μέσα σε ένα Σώμα Κειμένου. Χρησιμοποιείται ευρέως στην Επεξεργασία Φυσικής Γλώσσας και την Εύρεση Πληροφοριών (IR), όπου τα κείμενα αναπαρίστανται ως τσάντες λέξεων και οι λέξεις μετριούνται ως “weights”, ανάλογα τον αριθμό εμφάνισής τους.

Στο Orange, το BoW, δέχεται ένα σώμα κειμένων, στα οποία επισυνάπτει αριθμητικά χαρακτηριστικά, μετά από μετρήσεις. Μερικές από τις παραμέτρους που χρησιμοποιεί αυτό το εργαλείο, είναι η Συχνότητα Όρων (Term Frequency), η Συχνότητα Αρχείων (Document Frequency) και η Ομαλοποίηση (Reguralization). Η Συχνότητα των Όρων, ενδέχεται να υπολογιστεί με βάση τον αριθμό εμφάνισης των λέξεων (Count), να είναι δυαδική (Binary) ή σχεδόν γραμμική (Sublinear). Η Συχνότητα Αρχείων δέχεται τρεις επιλογές, το none, την

[59]Qader, W.A., Ameen, M.M. & Ahmed, B.I. (2019) *An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges,* International Engineering Conference (IEC), Erbil, Iraq, 2019, pp. 200-204, doi: 10.1109/IEC47844.2019.8950616. Ανάκτηση από: https://www.researchgate.net/publication/338511771_An_Overview_of_Bag_of_WordsImportance_Implementation_Applications_and_Challenges Τελευταία πρόσβαση: 3/5/2023.

Αντίστροφη Συχνότητα Αρχείων (IDF) και την Ομαλή Αντίστροφη Συχνότητα Αρχείων (Smooth IDF). Τέλος, η Ομαλοποίηση, χωρίζεται σε none, L1 και L2[60].

Για την Συσταδοποίηση των Κειμένων, το BoW υπολογίζει τις ομοιότητες μεταξύ των κειμενικών αρχείων που εισάγαμε, με βάση τον αριθμό εμφάνισης των λέξεων, και τους προσδίδει αριθμητικά χαρακτηριστικά, προετοιμάζοντας έτσι τον μετέπειτα υπολογισμό των αποστάσεων μεταξύ τους από το αντίστοιχο εργαλείο[61].

Έπειτα, το εργαλείο των Αποστάσεων[62], υπολογίζει τις αποστάσεις μεταξύ των κειμένων, με βάση τα αριθμητικά χαρακτηριστικά που τους προσέδωσε το BoW. Για να υπολογιστούν οι αποστάσεις σε μεγάλα σώματα κειμένου, συνήθως χρησιμοποιούνται τα μετρικά συστήματα Euclidean, Manhattan και Cosine.

Τέλος, το εργαλείο της Ιεραρχικής Συσταδοποίησης[63], κατανέμει τα κειμενικά αντικείμενα σε συστάδες, χρησιμοποιώντας τον ομώνυμο αλγόριθμο. Αφού εισαχθούν στο εργαλείο οι διανυσματικές αποστάσεις που υπολογίστηκαν, τα κείμενα βάσει αυτών και με τη χρήση του αλγόριθμου της Ιεραρχικής Συσταδοποίησης, κατανέμονται σε συστάδες κειμένων και εμφανίζονται σε μορφή Δενδρογράμματος.

Η Ιεραρχική Συσταδοποίηση χρησιμοποιεί διάφορες μεθόδους σύνδεσης, οι οποίες καταχωρούν τα κειμενικά αντικείμενα σε συστάδες, ανάλογα με το βαθμό ομοιότητάς τους. Οι μέθοδοι σύνδεσης που μπορούν να εφαρμοστούν σε αυτό το εργαλείο, είναι: η Απλή Σύνδεση (Simple Linkage), η Μέση Σύνδεση (Average Linkage), η Απόλυτη Σύνδεση (Complete Linkage) και η Σύνδεση Ward (Ward Linkage)[64]. Για την εφαρμογή της Ιεραρχικής Συσταδοποίησης στα δύο σύνολα δεδομένων που αποκτήσαμε, επιλέξαμε τη Σύνδεση Ward, η οποία ενώνει δύο συστάδες σύμφωνα με το τετραγωνικό σφάλμα του αθροίσματος των τιμών τους.

[60]Orange. *Bag of Words.* Ανάκτηση από: https://orangedatamining.com/widget-catalog/text-mining/bagofwords-widget/ Τελευταία πρόσβαση: 3/5/2023.
[61]Tilbe, A. (2022, Jul. 11). 3 Critical Reasons Why Bag of Words is Implemented Before Sentiment Analysis. *Towards AI.* Ανάκτηση από: https://pub.towardsai.net/3-critical-reasons-why-bag-of-words-is-implemented-before-sentiment-analysis-24407e815491 Τελευταία πρόσβαση: 3/5/2023.
[62]Orange. *Distances.* Ανάκτηση από: https://orangedatamining.com/widget-catalog/unsupervised/distances/ Τελευταία πρόσβαση: 3/5/2023.
[63]Orange. *Hierarchical Clustering.* Ανάκτηση από: https://orangedatamining.com/widget-catalog/unsupervised/hierarchicalclustering/ Τελευταία πρόσβαση: 3/5/2023.
[64] ZLP, J. (2019, Nov. 17). Distance Measures and Linkage Methods in Hierarchical Clustering. *Level Up Coding.* Ανάκτηση από: https://levelup.gitconnected.com/distance-measures-and-linkage-methods-in-hierarchical-clustering-8b7d488d7ebc Τελευταία πρόσβαση: 3/5/2023.

Γενικά, για τον υπολογισμό των συστάδων σε μεγάλα σώματα κειμένου, προτιμάται η μέθοδος της σύνδεσης Ward[65], καθώς αναλύει την απόκλιση μεταξύ των συστάδων και διευκολύνει την κατανόηση των σχέσεων που αναπτύσσονται, με αποτέλεσμα την καλύτερη διαχείριση των κειμενικών δεδομένων.

Για να εφαρμόσουμε τη μέθοδο της Συσταδοποίηση Κειμένων, δημιουργήσαμε μια ξεχωριστή ροή εργασιών για κάθε σύνολο δεδομένων.

Αφού πραγματοποιήσαμε ενέργειες προεπεξεργασίας του κειμένου, μεταφέραμε τα διαμορφωμένα δεδομένα στο BoW, όπου επιλέξαμε τον υπολογισμό της Συχνότητας των Όρων με βάση τη Μέτρηση, τη Συχνότητα των Αρχείων με το μοντέλο της Αντίστροφης Συχνότητας Αρχείου, ενώ δεν κάναμε καμία Ομαλοποίηση. Με τις ενέργειες αυτές, μετατρέψαμε τα δεδομένα σε αριθμητικές αναπαραστάσεις και τα προετοιμάσαμε για το επόμενο στάδιο, αυτό της μέτρησης των Αποστάσεων.

Οι αριθμητικές αναπαραστάσεις που εξάγαμε από το BoW, χρησιμοποιήθηκαν από το εργαλείο των Αποστάσεων, στο οποίο επιλέξαμε τη μέθοδο της Ομοιότητας Συνημίτονου (Cosine), για να μετρηθούν οι αποστάσεις μεταξύ των κειμενικών αντικειμένων.

Τέλος, τα αποτελέσματα του υπολογισμού των αποστάσεων, τα εισάγαμε στο εργαλείο της Ιεραρχικής Συσταδοποίησης, όπου πραγματοποιήσαμε τη Σύνδεση των κειμένων με τη μέθοδο του Ward και επιλέξαμε η κοπή του δενδρογράμματος να γίνει χειροκίνητα, με διαφορετικό σημείο κοπής για κάθε σύνολο δεδομένων.

2.4. Πραγματοποίηση της Σημασιολογικής Ανάλυσης

Για την εξαγωγή νοήματος μέσα από λέξεις, προτάσεις και κείμενα με τη χρήση της Σημασιολογικής Ανάλυσης, το Orange, διαθέτει διάφορα εργαλεία, όπως την Ενσωμάτωση Αρχείων (Document Embedding), την Εξαγωγή Λέξεων - Κλειδιών (Extract Keywords), το δισδιάστατο χάρτη t-SNE και το Σημασιολογικό Προβολέα (Semantic Viewer).

Κάνοντας χρήση των παραπάνω εργαλείων, μπορούμε να εντοπίσουμε τις λέξεις - κλειδιά μέσα σε κειμενικά αρχεία, να τις βαθμολογήσουμε με βάση τη συχνότητα της παρουσίας τους μέσα σε ένα Σώμα Κειμένου και να υπολογίσουμε τις ομάδες γειτνίασης

[65]Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.2307/2282967 Ανάκτηση από: https://www.jstor.org/stable/2282967 Τελευταία πρόσβαση: 3/5/2023.

κειμένων μέσα σε ένα χάρτη δύο διαστάσεων, ανακαλύπτοντας έτσι τις μεταξύ τους σχέσεις[66].

Συγκεκριμένα, η Ενσωμάτωση Αρχείων, διαθέτει την ικανότητα ανάλυσης των λεξικογραφικών μονάδων που βρίσκονται σε κάθε αρχείο του συνολικού σώματος κειμένου, προσδίδει ενσωματώσεις σε κάθε λέξη χρησιμοποιώντας κατάλληλα προεκπαιδευμένα μοντέλα Μεταγλωττιστών[67] και δημιουργεί αριθμητικά διανύσματα για κάθε κειμενικό αρχείο. Το εργαλείο της Ενσωμάτωσης Αρχείων στο Orange, υποστηρίζει τους Μεταγλωττιστές fastText και Multilingual Sbert και μπορεί να συγκεντρώσει τις λεξικογραφικές ενσωματώσεις με βάση το Νόημα (Mean), Σύνολο (Sum), Μέγιστο (Max) και Ελάχιστο (Min)[68].

Με τη χρήση του t-SNE[69], τα κείμενα αναπαρίστανται ως κουκίδες και διανέμονται στο δισδιάστατο χάρτη, ανάλογα με την πιθανότητα κατανομής τους. Το εργαλείο αυτό, αποτελεί μία στατιστική μέθοδο εξερεύνησης και οπτικοποίησης πολυδιάστατων δεδομένων και έχει τη δυνατότητα να προσφέρει μια εικόνα ή πρόβλεψη για τον τρόπο με τον οποίον κατατάσσονται τα δεδομένα σε έναν πολυδιάστατο χώρο. Επινοήθηκε από τους Maatens και Hinton (2008)[70] και αποτελεί ένα χρήσιμο εργαλείο οπτικοποίησης του Orange, καθώς αποτυπώνει σε ένα χάρτη τις αποστάσεις μεταξύ των κειμένων και τις ομάδες που διαμορφώνονται, με βάση τους υπολογισμούς που πραγματοποίησε προηγουμένως η Ενσωμάτωση Αρχείων. Αφού επιλέξουμε μία διακριτή ομάδα πάνω στο χάρτη, μπορούμε να βρούμε τις λέξεις - κλειδιά που τη χαρακτηρίζουν.

Το εργαλείο της Εξαγωγής των Λέξεων - Κλειδιών[71], δέχεται ένα σώμα κειμένου και βρίσκει τις λέξεις - κλειδιά που χαρακτηρίζουν το κείμενο αυτό. Υπάρχουν διάφοροι

---

[66]Pretnar, A., (2021, Sep. 17). Semantic Analysis of Documents. *Orange.* Ανάκτηση από: https://orangedatamining.com/blog/2021/2021-09-17-semantic-analysis/ Τελευταία πρόσβαση: 31/5/2023.

[67]Grave, E., Bojanowski, P., Gupta, P., Joulin, A. & Mikolov, T.. (2018, May). *Learning Word Vectors for 157 Languages.* Proceedings of the International Conference on Language Resources and Evaluation, 2018. Ανάκτηση από: https://aclanthology.org/L18-1550/ Τελευταία πρόσβαση: 31/5/2023.

[68]Orange. *Document Embedding*. Ανάκτηση από: https://orangedatamining.com/widget-catalog/text-mining/documentembedding/ Τελευταία πρόσβαση: 31/5/2023.

[69]Violante, A., (2018, August 29). An Introduction to t-SNE with Python Example. *Towards Data Science.* 29 Aug. 2018. Ανάκτηση από: https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1 Τελευταία πρόσβαση: 31/5/2023.

[70]Maaten, L.V., & Hinton, G.E. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605. Ανάκτηση από: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf Τελευταία πρόσβαση: 31/5/2023.

[71]Orange. *Extract Keywords.* Ανάκτηση από: https://orangedatamining.com/widget-catalog/text-mining/keywords/ Τελευταία πρόσβαση: 31/5/2023.

αλγόριθμοι που χρησιμοποιούνται για την εξαγωγή των λέξεων - κλειδιών, μεταξύ των οποίων η Συχνότητα Λέξεων - Αντίστροφη Συχνότητα Αρχείων (TF-IDF), η οποία και μετρά τη συχνότητα εμφάνισης των λέξεων, υπολογίζοντας τη σχέση τους με το συνολικό Σώμα Κειμένου.

Τέλος, ο Σημασιολογικός Προβολέας[72], αφού δεχτεί μια Λίστα από λέξεις - κλειδιά, υπολογίζει και στην συνέχεια εμφανίζει την παρουσία τους στο Σώμα Κειμένου. Ως επιλογές, διαθέτει το φιλτράρισμα με βάση το Κατώφλι (Threshold), και μπορεί να παρουσιάσει είτε το αρχείο, είτε το τμήμα του κειμένου, είτε την πρόταση στην οποία εμφανίζονται αυτές οι λέξεις - κλειδιά. Για να οπτικοποιήσει τα αποτελέσματα των ευρημάτων, εμφανίζει έναν πίνακα με την Αντιστοίχιση (Match), δηλαδή το σύνολο των εμφανίσεων των λέξεων - κλειδιών μέσα στο κάθε ένα κείμενο, τις Τιμές (Score), οι οποίες υπολογίζονται σε επίπεδο πρότασης μεταξύ των SBERT επισυνάψεων της πρότασης και των λέξεων - κλειδιών που εισάγαμε και τέλος, το Αρχείο (Document), δηλαδή το κείμενο του κάθε tweet ή mention.

Με την εφαρμογή της Σημασιολογικής Ανάλυσης, μπορούμε να αποκτήσουμε κάποιες χρήσιμες πληροφορίες σχετικά με το περιεχόμενο ενός κειμένου. Στην παρούσα ανάλυση, μελετήσαμε τα σύνολα δεδομένων με τα tweets που ανήρτησε ο @atsipras και τα mentions που έγιναν προς αυτόν, ώστε να εντοπίσουμε ποιές είναι αυτές οι λέξεις - κλειδιά που χρησιμοποιούνται και το ενδεχόμενο να χαρακτηρίζουν το υπόλοιπο Σώμα Κειμένου.

Ανακαλύπτοντας τις ομάδες γειτνίασης που συγκροτούν τα κείμενα και εξάγοντας τις λέξεις - κλειδιά τους, μπορούμε να ανιχνεύσουμε τις σχέσεις μεταξύ των λέξεων αυτών και του συνολικού Σώματος Κειμένου, δηλαδή το κατά πόσο αντιπροσωπεύουν το κείμενο, εξάγουμε κάποια χρήσιμα συμπεράσματα για τα σύνολα δεδομένων που αναλύουμε.

Πρώτο βήμα για να πραγματοποιήσουμε ενέργειες Σημασιολογικής Ανάλυσης, είναι η Προεπεξεργασία του Κειμένου, όπου κατατμήσαμε τις λεξικογραφικές μονάδες με βάση τις κανονικές εκφράσεις, μετατρέψαμε τους χαρακτήρες σε πεζούς, αφαιρέσαμε περιττά στοιχεία όπως αριθμούς και χρησιμοποιήσαμε αρχείο με Stopwords.

Έπειτα, με το εργαλείο της Ενσωμάτωσης Αρχείων, υπολογίσαμε τις ενσωματώσεις για κάθε λέξη του Σώματος Κειμένου, χρησιμοποιώντας τον Μεταγλωττιστή Multilingual Sbert και μετατρέψαμε τα κείμενα σε διανυσματικές αναπαραστάσεις. Με αυτό τον τρόπο, μπορούμε στη συνέχεια να χρησιμοποιήσουμε τις διανυσματικές αναπαραστάσεις των

[72]Orange. *Semantic Viewer.* Ανάκτηση από: https://orangedatamining.com/widget-catalog/text-mining/semanticviewer/ Τελευταία πρόσβαση: 31/5/2023.

κειμένων με το εργαλείο t-SNE, προκειμένου να υπολογίσουμε τη θέσης τους στον χάρτη και να τα κατανείμουμε σε ομάδες.

Ο δισδιάστατος χάρτης t-SNE, κατατάσει τα κείμενα σε ομάδες γειτνίασης, ανάλογα με την πιθανότητα κατανομής τους. Για τα δύο σύνολα δεδομένων που εντάξαμε στο t-SNE, κάναμε τις ίδιες επιλογές: κρατήσαμε την τιμή της Σύγχυσης (Perplexity) στο 30, τοποθετήσαμε τη μπάρα της Υπερβολής (Exaggeration) στο 1 και επιλέξαμε τα PCA Components να είναι 2. Τέλος, εφαρμόσαμε ομαλοποίηση των δεδομένων. Οι επιλογές αυτές, έγιναν μετά από πολλές δοκιμές, και έχουν ως σκοπό να καταστήσουν ευδιάκριτες τις κειμενικές ομάδες που δημιουργήθηκαν.

Για να παρατηρήσουμε το περιεχόμενο κάποιας t-SNE ενσωμάτωσης, επιλέγουμε την αντίστοιχη ομάδα από το χάρτη και οπτικοποιούμε το κειμενικό της περιεχόμενο με το Σύννεφο Λέξεων και τον Προβολέα Σώματος Κειμένου.

Εάν θέλουμε να εξάγουμε τις λέξεις - κλειδιά αυτής της ομάδας, χρησιμοποιούμε το εργαλείο της Εξαγωγής των Λέξεων - Κλειδιών και επιλέγουμε τον κατάλληλο αλγόριθμο. Για τις ανάγκες της παρούσας έρευνας, επιλέξαμε στο εργαλείο της Εξαγωγής των Λέξεων - Κλειδιών, τη Συχνότητα Λέξεων - Αντίστροφη Συχνότητα Αρχείων, τη συγκέντρωση των λέξεων με βάση το Νόημα και την επιλογή των πρώτων 10 λέξεων που εμφανίζουν τη μεγαλύτερη τιμή.

Τέλος, εξάγαμε αυτές τις λέξεις με το εργαλείο Λίστα Λέξεων και τις μεταφέραμε στο Σημασιολογικό Προβολέα, όπου εμφανίστηκαν οι Αντιστοιχίσεις και οι Τιμές τους για το κάθε ένα κείμενο. Στόχος αυτών των ενεργειών ήταν να εντοπίσουμε το κατά πόσον αυτές οι λέξεις εμφανίζονται στο Σώμα Κειμένου, δηλαδή εάν είναι αντιπροσωπευτικές αυτού και να εξάγουμε χρήσιμα συμπεράσματα για το σημασιολογικό περιεχόμενο των κειμενικών ομάδων στις οποίες ανήκουν.

2.5. Πραγματοποίηση της Θεματικής Μοντελοποίησης

Για να πραγματοποιήσουμε τη μέθοδο της Θεματικής Μοντελοποίησης, χρησιμοποιήσαμε τον αλγόριθμο LDA. Η Λανθάνουσα Κατανομή Dirichlet, χαρακτηρίζεται ως *"ένα πιθανολογικό μοντέλο που εφαρμόζεται σε συλλογές δεδομένων, όπως τα κείμενα. Αποτελεί ένα ιεραρχικό Μπεϋζιανό μοντέλο τριών επιπέδων, στο οποίο το κάθε αντικείμενο μιας συλλογής, αναπαρίσταται ως μία πεπερασμένη μίξη πάνω σε ένα κρυμμένο σύνολο*

θεματικών πιθανοτήτων. Στην περίπτωση της Μοντελοποίησης κειμένου, οι θεματικές πιθανότητες προσφέρουν μία ξεκάθαρη αναπαράσταση ενός κειμένου" (Blei et al., 2003)[73].

Με ποιον όμως τρόπο δουλεύει ο αλγόριθμος της Λανθάνουσας Κατανομής Dirichlet; Για να βρούμε τις λέξεις - κλειδιά που χαρακτηρίζουν μια θεματική και να κατανεμηθεί ένα κείμενο σε κάποιες από αυτές τις κρυμμένες θεματικές, θα πρέπει πρώτα να ακολουθήσουμε συγκεκριμένες ενέργειες, καθώς ο αλγόριθμος εξ' αρχής υποθέτει ότι ένα κείμενο δημιουργήθηκε από μία στατιστική παραγωγική διαδικασία, με βάση την οποία το κάθε κείμενο είναι μια μίξη θεματικών και η κάθε θεματική είναι μια μίξη λέξεων[74].

Βασικά εργαλεία που μπορούμε να χρησιμοποιήσουμε στο Orange για τη διενέργεια Θεματικής Μοντελοποίησης με το LDA και την οπτικοποίηση των αποτελεσμάτων της, είναι το BoW, η Θεματική Μοντελοποίηση και το LDAvis.

Το BoW, όπως αναφέραμε προηγουμένως, μετράει τις εμφανίσεις των λέξεων μέσα σε ένα Σώμα Κειμένου. Καθώς ο βασικός σκοπός της Θεματικής Μοντελοποίησης, είναι η απόκτηση γνώσης σχετικά με το περιεχόμενο ενός μεγάλου όγκου κειμένων και ο εντοπισμος των θεματικών που κρύβονται μέσα σε αυτά, το BoW είναι ένα χρήσιμο εργαλείο, ιδιαίτερα για το μοντέλο της Λανθάνουσας Κατανομής Dirichlet, καθώς προσδίδει αριθμητικά χαρακτηριστικά γνωρίσματα στις λέξεις, ανάλογα με τις φορές που εμφανίζονται στο συνολικό Σώμα Κειμένου. Έτσι, όσο περισσότερα weights έχει μία λέξη, δηλαδή αριθμό εμφανίσεων, τόσο πιο πιθανό είναι να αποτελεί χαρακτηριστική λέξη για κάποια/ες από τις 10 θεματικές που εμφανίζει το εργαλείο της Θεματικής Μοντελοποίησης[75].

Το εργαλείο της Θεματικής Μοντελοποίησης μόλις δεχθεί τα κειμενικά δεδομένα από το Σώμα Κειμένου και τα χαρακτηριστικά γνωρίσματα που υπολόγισε το BoW, εφαρμόζει κάποιον από τους αλγόριθμους που υποστηρίζει το Orange, και στην περίπτωσή μας τον LDA, εμφανίζοντας έναν πίνακα με 10 θεματικές και τις 10 λέξεις - κλειδιά που τις χαρακτηρίζουν. Στην συνέχεια, τα ευρήματα της Θεματικής Μοντελοποίησης, μπορούν να οπτικοποιηθούν με εργαλεία όπως το LDAvis, για την εξαγωγή χρήσιμων συμπερασμάτων.

---

[73]Blei, D., Andrew, Ng. & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research.* 3. 601-608. Ανάκτηση από: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf Τελευταία πρόσβαση: 31/5/2023.

[74]Great Learning Team. (2022, Oct 31). Understanding Latent Dirichlet Allocation (LDA). Ανάκτηση από: https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/ Τελευταία πρόσβαση: 31/5/2023.

[75]Tufts, C. (2018, January 31). Bag of Words. In *The Little Book of LDA.* Ανάκτηση από: https://miningthedetails.com/LDA_Inference_Book/word-representations.html#bag-of-words Τελευταία πρόσβαση: 31/5/2023.

Το LDAvis, προέρχεται από ένα πακέτο της R που αναπτύχθηκε από τους Sievert & Shirey, το 2014[76]. Λειτουργεί ως το εργαλείο οπτικοποίησης των θεματικών που εξάγονται αποκλειστικά από τον αλγόριθμο της Λανθάνουσας Κατανομής Dirichlet, εμφανίζοντας τις 20 πιό ψηλά βαθμολογημένες λέξεις για την κάθε θεματική. Όσο πιό ψηλά εμφανίζεται μια λέξη στην λίστα, τόσο περισσότερο χαρακτηρίζει αυτή την θεματική. Το εργαλείο, επίσης, μπορεί και εμφανίζει τη συχνότητα των ίδιων λέξεων και σε άλλες θεματικές.

Όπως μπορούμε να αντλήσουμε σημαντικές πληροφορίες μέσα από έναν μεγάλο όγκο κειμένων με την εφαρμογή της Θεματικής Μοντελοποίησης, έτσι είναι εξίσου δυνατό να ανακαλύψουμε τις θεματικές που υπάρχουν μέσα στα κείμενα που αναρτούν οι χρήστες των ΜΚΔ, όπως το Twitter. Στην περίπτωση της παρούσας έρευνας, επιχειρήσαμε να εντοπίσουμε τις κρυμμένες θεματικές εντός των συνόλων δεδομένων που αφορούν στα tweets του @atsipras και στα mentions που έγιναν προς αυτόν.

Αρχικά, δημιουργήσαμε δύο ροές εργασιών για το κάθε ένα από τα σύνολα δεδομένων που έχουμε στην διάθεσή μας. Αφού πραγματοποιήσαμε τις κατάλληλες ενέργειες προεπεξεργασίας του κειμένου, μεταφέραμε τα δεδομένα στο BoW, όπου με τη μέθοδο της Αντίστροφης Συχνότητας Αρχείων, υπολογίσαμε την παρουσία των λέξεων μέσα στο Σώμα Κειμένου και τους δώσαμε αριθμητικά χαρακτηριστικά γνωρίσματα.

Στο εργαλείο της Θεματικής Μοντελοποίησης, επιλέξαμε τον αλγόριθμο της Λανθάνουσας Κατανομής Dirichlet και εμφανίσαμε έναν πίνακα με τις 10 θεματικές που εντοπίστηκαν και τις 10 λέξεις - κλειδιά που είναι χαρακτηριστικές για την κάθε μία θεματική.

Διαλέγοντας μία από τις 10 θεματικές, μπορούμε να παρατηρήσουμε το περιεχόμενό της με τα εργαλεία Προβολέας Κειμένου, Πίνακας Δεδομένων, LDAvis και Σύννεφο Λέξεων. Με αυτόν τον τρόπο, μπορούμε να βρούμε χρήσιμα ευρήματα, όπως την πιθανότητα ενός κειμένου να ανήκει σε μία ή παραπάνω θεματικές, τα weights των λέξεων που ανήκουν σε μια θεματική και τις κορυφαίες λέξεις - κλειδιά που χαρακτηρίζουν την κάθε θεματική, σε σχέση με το συνολικό Σώμα Κειμένου.

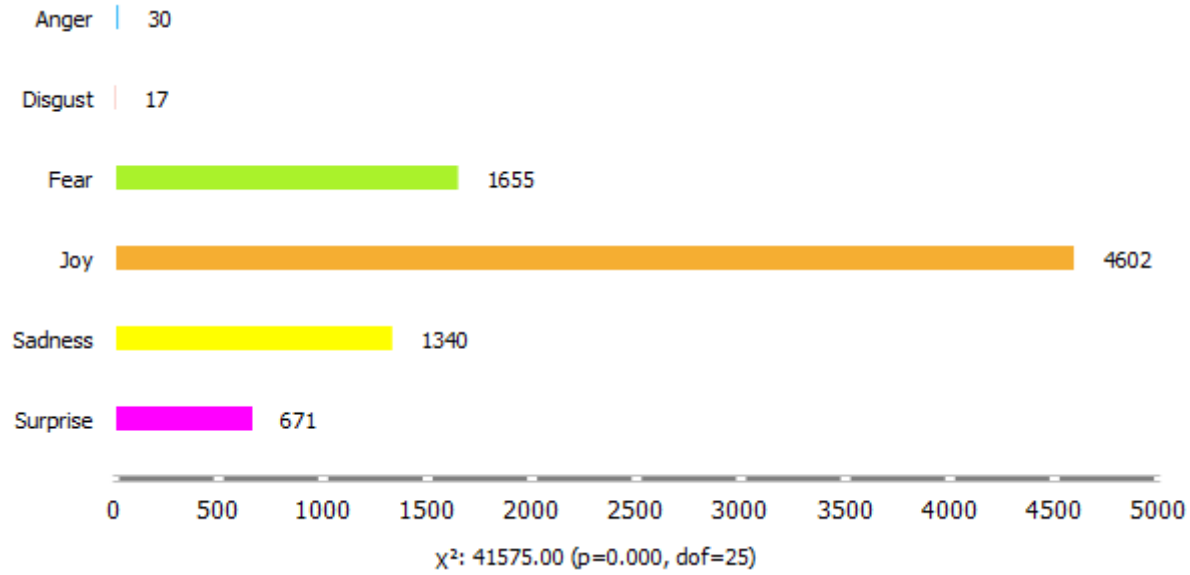
[76]Sievert, C. & Shirley, K. (2014). *LDAvis: A method for visualizing and interpreting topics.* In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics DOI: 10.3115/v1/W14-3110 Ανάκτηση από: https://aclanthology.org/W14-3110/ Τελευταία πρόσβαση: 31/5/2023.

Κεφάλαιο 3: Ευρήματα

3.1. Ευρήματα της Ανάλυσης Συναισθημάτων

Στη ροή εργασιών για τα tweets του @atsipras, αναλύσαμε συνολικά 8.315 tweets, με βάση το μοντέλο εντοπισμού Συναισθήματος που επινόησε ο Ekman και υποστηρίζεται από το εργαλείο Tweet Profiler.

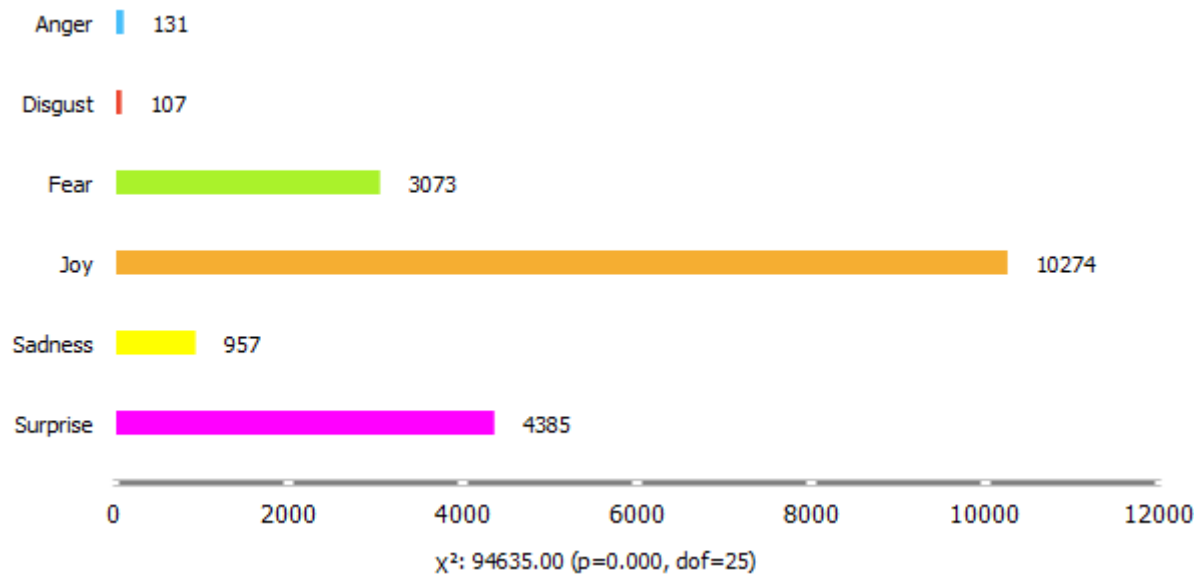
Τα κύρια Συναισθήματα που εντοπίσαμε (Εικόνα 1), είναι η Χαρά (4.602 tweets), ο Φόβος (1.655 tweets), η Λύπη (1.340 tweets), η Έκπληξη (671 tweets), ο Θυμός (30 tweets) και η Απέχθεια (17 tweets).



Εικόνα 1: Η κατανομή των Συναισθημάτων για τα tweets του @atsipras.

Αντίστοιχα, στη ροή εργασιών για τα mentions προς τον @atsipras, αναλύσαμε 18.927 mentions, με χρήση του ίδιου μοντέλου.

Τα κύρια Συναισθήματα που εντοπίσαμε (Εικόνα 2), είναι η Χαρά (10.274 mentions), η Έκπληξη (4.385 mentions), ο Φόβος (3.073 mentions), η Λύπη (957 mentions), ο Θυμός (131 mentions) και η Απέχθεια (107 mentions).



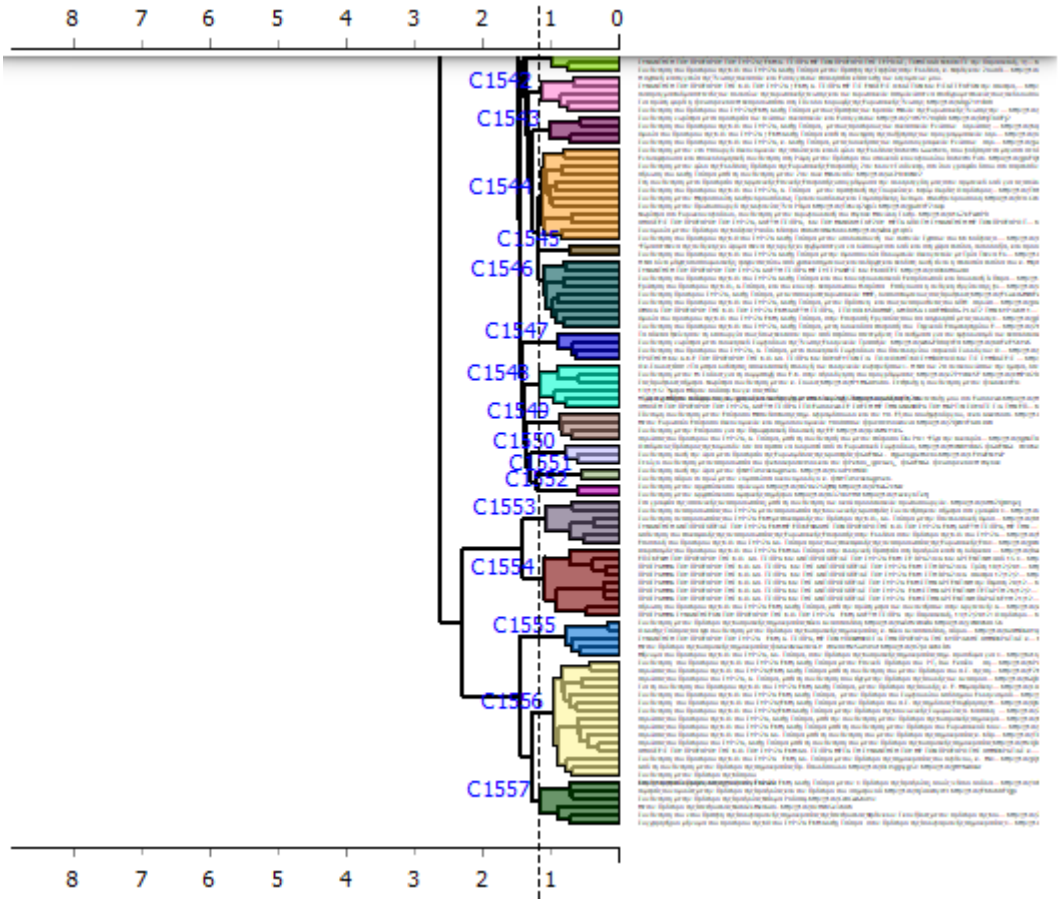
Εικόνα 2: Η κατανομή των Συναισθημάτων για τα mentions προς τον @atsipras.

Οι πληροφορίες που εξάγαμε από την κατανομή των Συναισθημάτων στα δύο σύνολα δεδομένων, μπορούν να χρησιμοποιηθούν για να αποκτήσουμε περεταίρω γνώση, όπως για παράδειγμα να εντοπίσουμε εκείνες τις λέξεις που μπορούν να κατηγοριοποιηθούν από ένα συναίσθημα με βάση το μοντέλο του Ekman ή να ανακαλύψουμε τυχόν συσχετισμούς μεταξύ της περιόδου κατά την οποία γράφτηκαν αυτά τα tweets και τα mentions και εκείνων των Συναισθημάτων που εκφράζουν. Σε κάθε περίπτωση, η Ανάλυση των Συναισθημάτων, είναι μία μέθοδος που έχει ευρεία χρήση σε πολλούς τομείς και μπορεί να χρησιμοποιηθεί σε συνδυασμό με άλλες μεθόδους Εξόρυξης Δεδομένων, όπως η Κατηγοριοποίηση Κειμένων ή η Σημασιολογική Ανάλυση.

3.2. Ευρήματα της Συσταδοποίησης Κειμένων

Όπως αναφέραμε στο υποκεφάλαιο "*Πραγματοποίηση της Συσταδοποίησης Κειμένων*", αφού δώσαμε αριθμητικά χαρακτηριστικά στα κείμενα μέσω του BoW και μετρήσαμε τις αποστάσεις μεταξύ των αρχείων με το εργαλείο Distances, έπειτα εφαρμόσαμε τη μέθοδο της Συσταδοποίησης Κειμένων, για να καταχωρήσουμε τα κείμενα σε ομάδες, ανάλογα το βαθμό ομοιότητάς τους.

Η Συσταδοποίηση Κειμένων για τα tweets του @atsipras, πραγματοποιήθηκε με τον αλγόριθμο της Ιεραρχικής Συσταδοποίησης, την επιλογή της Σύνδεσης Ward και την ελεύθερη κοπή του Δενδρογράμματος στο 13%. Με βάση αυτές τις ρυθμίσεις, υπολογίσαμε 1.557 συστάδες, σε σύνολο 8.315 tweets (Εικόνα 3).



Εικόνα 3: Η Ιεραρχική Συσταδοποίηση, για τα tweets του @atsipras.

Αφού εφαρμόσαμε τη Συσταδοποίηση, παρατηρούμε ότι δημιουργήθηκαν πολλές συστάδες, με μικρό αριθμό tweets. Αυτό σημαίνει, ότι πραγματοποιήσαμε αυστηρή συσταδοποίηση και πώς πρέπει να αναμένουμε ότι τα κείμενα θα εμφανίζουν παρόμοιο περιεχόμενο μεταξύ τους, ανά συστάδα, εφόσον κατατμήθηκαν σε ομάδες με βάση την απόσταση. Επιλέγοντας μια από αυτές τις συστάδες, μπορούμε να δούμε το περιεχομενο της με τον Προβολέα του Σώματος Κειμένου και τις λέξεις από τις οποίες αποτελούνται τα κειμενά της, με το Σύννεφο των Λέξεων.

Για παράδειγμα, η πρώτη συστάδα (C1) των tweets του @atsipras, παρατηρούμε ότι αποτελείται από μόλις 11 κείμενα (Εικόνα 4) και λέξεις που αφορούν στην οικονομία (Εικόνα 5).

1	Το κόμμα που χρωστάει ...
2	Πρωθυπουργός και ...
3	Επίσης προτείνουμε τη ...
4	Με τις ενισχύσεις που ...
5	Την περίοδο 2015-19 ...
6	Στην Περιφέρεια Δυτικής ...
7	Οι προηγούμενες ...
8	Από το 2015 μέχρι σήμερ...
9	Ενισχύσαμε τη Μεσσηνία ...
10	Στο πλαίσιο αυτό, το ...
11	Το ποσό που έλαβε η ΝΔ ...

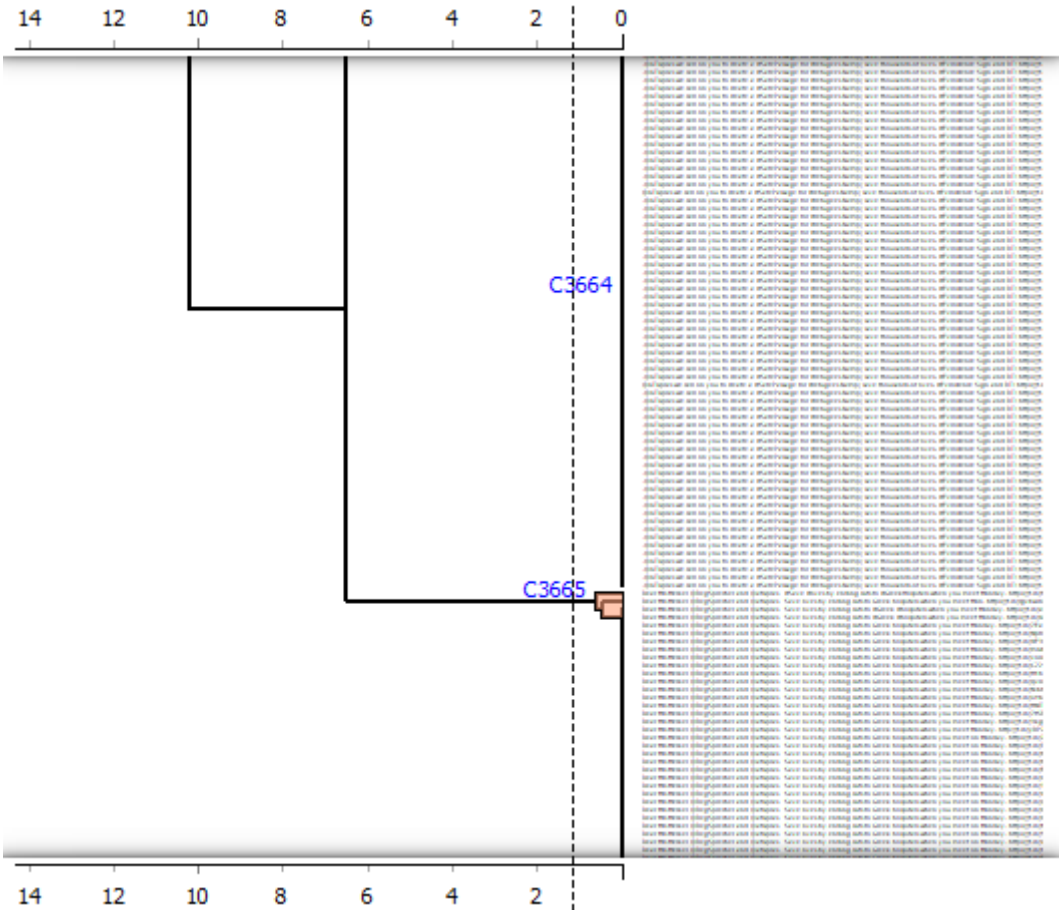
Text: Το κόμμα που χρωστάει 390 εκατ. ευρώ στις τράπεζες, που το υπουργικό συμβούλιο έχει 12 εκατ. χρέη συνολικά, έχει βγάλει στην αγορά Πάτσηδες να κυνηγούν νοικοκύρηδες.

Cluster: C1

Εικόνα 4: Ο Προβολέας του Σώματος Κειμένου με τα κείμενα της συστάδας C1, για τα tweets του @atsipras.

Εικόνα 5: Το Σύννεφο Λέξεων της συστάδας C1, για τα tweets του @atsipras.

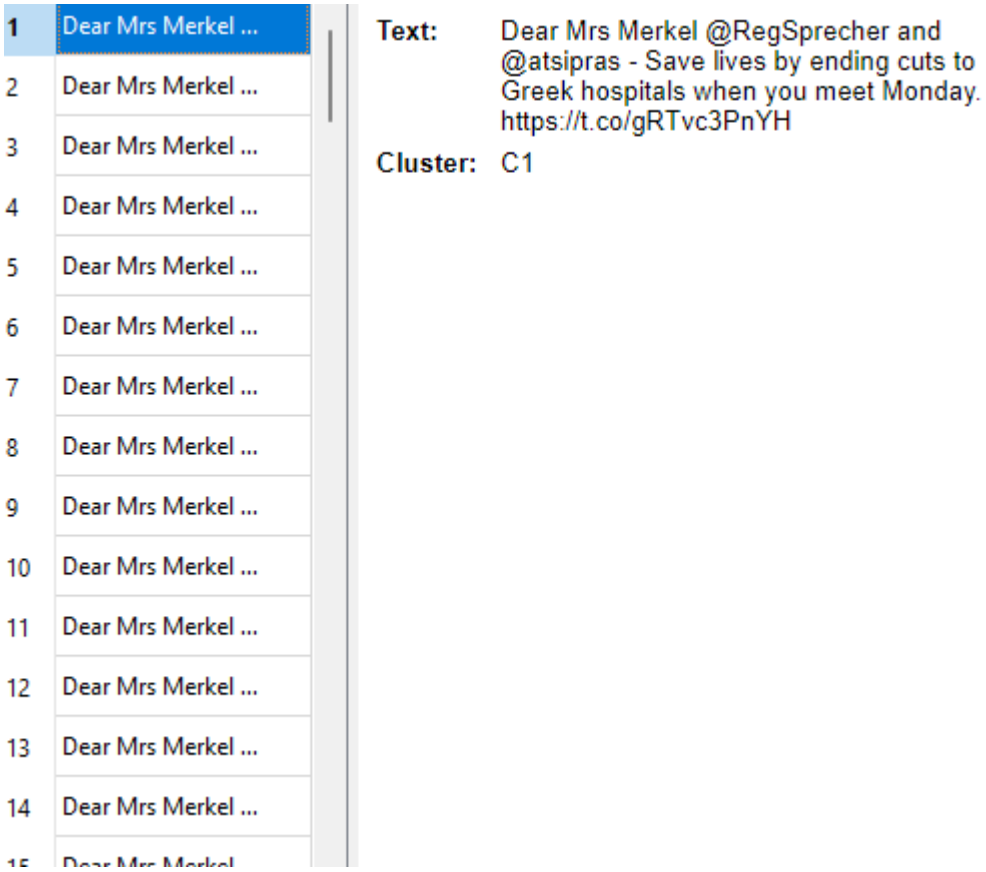
Η Συσταδοποίηση Κειμένων για τα mentions προς τον @atsipras, έγινε με τις ίδιες ρυθμίσεις και κοπή του Δενδρογράμματος στο 8%. Έτσι, υπολογίσαμε 3.665 συστάδες σε σύνολο 18.927 mentions (Εικόνα 6).



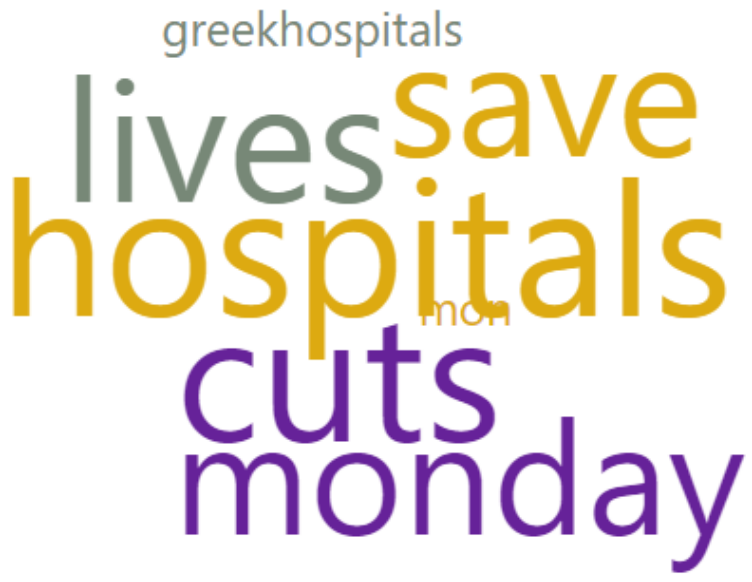
Εικόνα 6: Η Ιεραρχική Συσταδοποίηση, για τα mentions προς τον @atsipras.

Και σε αυτή την περίπτωση, παρατηρούμε ότι εμφανίστηκαν πολλές συστάδες, με μικρό αριθμό mentions. Εξαίρεση αποτελούν κάποιες λίγες συστάδες, που, όπως θα δούμε παρακάτω, διαθέτουν μεγάλο αριθμό πανομοιότυπων κειμένων.

Αφού επιλέξουμε για παράδειγμα την τελευταία συστάδα (C3665) των mentions, διαπιστώνουμε ότι αποτελείται από 122 πανομοιότυπα κείμενα (Εικόνα 7) και λίγες λέξεις που αφορούν σε περικοπές δαπανών στα Νοσοκομεία (Εικόνα 8).



Εικόνα 7: Ο Προβολέας του Σώματος Κειμένου με τα κείμενα της συστάδας C3665, για τα mentions προς τον @atsipras.



Εικόνα 8: Το Σύννεφο Λέξεων της συστάδας C3665, για τα mentions προς τον @atsipras.

Το γεγονός ότι τα κείμενα αυτής της συστάδας είναι πανομοιότυπα, κατά μεγάλη πιθανότητα μπορεί να σημαίνει ότι αποτελούν προϊόν copy - paste ή ότι δημιουργήθηκαν

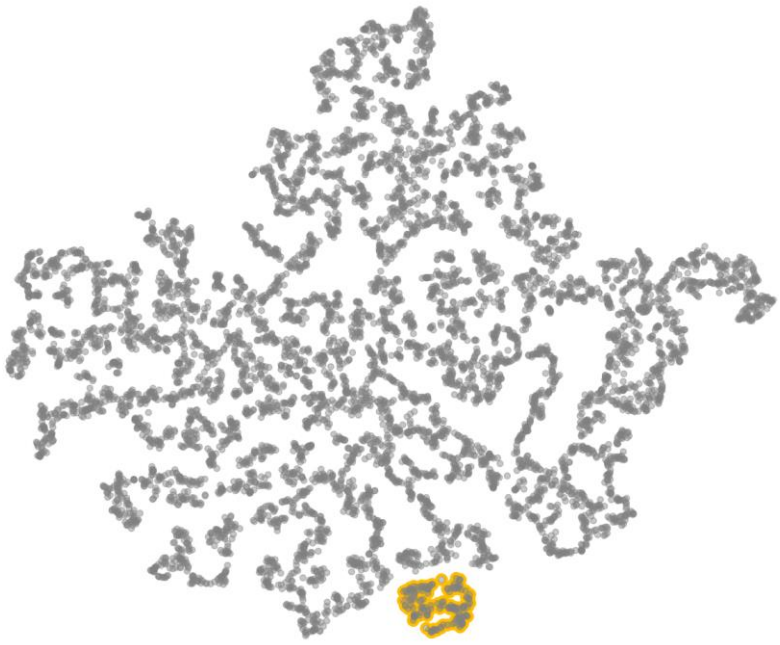
από λογαριασμούς bots, εάν λάβουμε υπόψη τον σχεδόν απόλυτο βαθμό ομοιότητάς τους και ότι αναρτήθηκαν μέσα σε λίγες μέρες από πολλούς και διαφορετικούς λογαριασμούς.

Με βάση τα αποτελέσματα της Συσταδοποίησης Κειμένων γι' αυτά τα δύο σύνολα δεδομένων, μπορούμε να εξάγουμε χρήσιμες πληροφορίες σχετικά με τον τρόπο χρήσης του γραπτού λόγου μεταξύ των χρηστών ή ακόμα και να συνεχίσουμε την ανάλυση των δεδομένων, προσδίδοντας χαρακτηριστικά γνωρίσματα (attributes) στις συστάδες ανάλογα με το θέμα στο οποίο αναφέρονται ή τις λέξεις που εντοπίζονται εντός τους. Σίγουρα, πρόκειται για μια χρήσιμη μέθοδο, που μπορούμε να τη χρησιμοποιήσουμε ακόμα και σε συνδυασμό με άλλες, για να εντοπίσουμε ανωμαλίες κατά την εκφορά του γραπτού λόγου, καθώς και ομοιότητες ή διαφορές μεταξύ των κειμένων.

3.3. Ευρήματα της Σημασιολογικής Ανάλυσης

Αφού υπολογίσαμε τις ενσωματώσεις των κειμένων με τον Μεταγλωττιστή Multilingual SBERT του εργαλείου της Ενσωμάτωσης Αρχείων, στη συνέχεια μεταφέραμε τα δεδομένα στο εργαλείο t-SNE, όπου ομαδοποιήσαμε τα κείμενα με βάση την πιθανότητα κατανομής τους στο δισδιάστατο χάρτη.

Για τα tweets του @atsipras και με βάση τις ρυθμίσεις που εφαρμόσαμε, παρατηρούμε ότι σχηματίζονται αρκετές ομαδοποιήσεις κειμένων στο χάρτη t-SNE, με σχετικά μεγάλες αποστάσεις μεταξύ τους (Εικόνα 9).



Εικόνα 9: Ο χάρτης t-SNE, για τα tweets του @atsipras.

Επιλέγοντας οποιαδήποτε από τις διακριτές ομάδες που εμφανίζονται στο χάρτη, μπορούμε να δούμε το περιεχόμενό της με τον Προβολέα του Σώματος Κειμένου, να εντοπίσουμε τις λέξεις από τις οποίες αποτελείται με το Σύννεφο των Λέξεων, να εξάγουμε τις λέξεις - κλειδιά με το εργαλείο Εξαγωγή των Λέξεων - Κλειδιών, να τις μετατρεψουμε σε Λίστα Λέξεων και τέλος, να τις εντάξουμε στο Σημασιολογικό Προβολέα, για να δούμε εάν είναι αντιπροσωπευτικές για τα κείμενα της ομάδας αυτής. Όσο πιο υψηλός είναι ο αριθμός των Αντιστοιχίσεων και των Τιμών, τόσο πιό αντιπροσωπευτικές είναι αυτές οι λέξεις.

Αφού επιλέξαμε μία ομάδα που εμφανίζεται μακριά από τις υπόλοιπες, διαπιστώσαμε από τις 10 κορυφαίες λέξεις - κλειδιά που την χαρακτηρίζουν (Εικόνα 10), ότι αυτά τα 161 κείμενα αφορούν αποκλειστικά την κοινοβουλευτική παρουσία του @atsipras.

Word	TF-IDF
βουλη	0.034
πρωθυπουργο	0.018
ομιλια	0.017
ερωτηση	0.017
δευτερολογια	0.016
συζητηση	0.016
προεδρου	0.016
συριζα	0.015
υπουργειου	0.015
επικαιρη	0.014
νομοσχεδιο	0.014
τσιπρα	0.014
κυβερνηση	0.012
κοινοβουλευτι...	0.011
ομαδας	0.011
μητσοτακη	0.011
νδ	0.010
κυβερνησης	0.010
νομου	0.009

Εικόνα 10: Οι λέξεις - κλειδιά της ομάδας που επιλέξαμε στο t-SNE, για τα tweets του @atsipras.

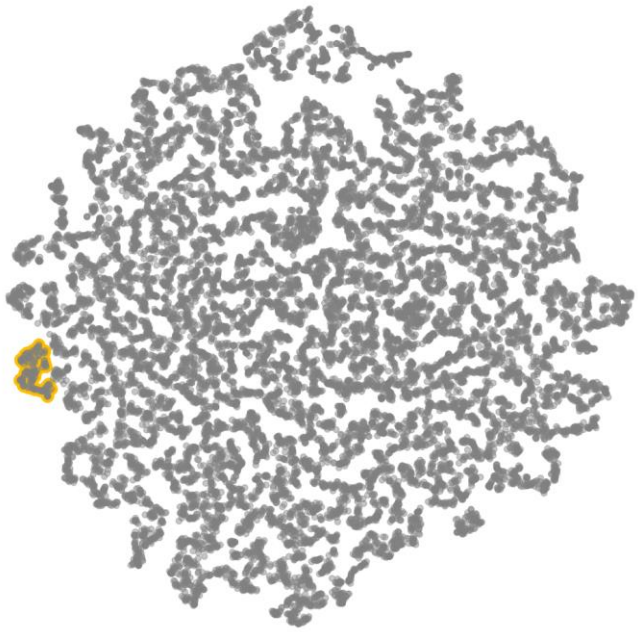
Μεταφέροντας αυτές τις λέξεις στο Σημασιολογικό Προβολέα, παρουσιάζονται οι Αντιστοιχίσεις τους και οι Τιμές που σημειώνουν για το κάθε κείμενο. Συγκεκριμένα, κάποιες από τις παραπάνω λέξεις - κλειδιά, μπορούν να εμφανιστούν σε κάθε tweet από 0 έως 6 φορές και με τιμές που κυμαίνονται από -0.049 έως 0.753 (Εικόνα 11). Αυτό σημαίνει, ότι αυτές οι λέξεις - κλειδιά, τείνουν να χαρακτηρίσουν σε μεγάλο βαθμό την ομάδα των κειμένων στην οποία εντάσσονται με βάση την κατανομή τους.

Match	Score	Documer
6	0.655	Ομιλία του ..
6	0.735	Ομιλία του ..
6	0.453	ΟΜΙΛΙΑ ΤΟΥ
5	0.542	Επίκαιρη ...
5	0.532	Ομιλία του ..
5	0.593	Επίκαιρη ...
5	0.667	Επίκαιρη ...
5	0.611	Επίκαιρη ...
5	0.630	Ομιλία του ..
5	0.565	Επίκαιρη ...
5	0.589	Επίκαιρη ...
5	0.517	Ομιλία του ..
5	0.710	Ομιλία του ..
5	0.650	Επίκαιρη ...
5	0.457	Επίκαιρη ...
5	0.650	Ομιλία του ..
5	0.478	ΕΠΙΚΑΙΡΗ ...
5	0.582	Επίκαιρη ...

Ομιλία του Προέδρου της Κ.Ο. του ΣΥΡΙΖΑ, Αλέξη Τσίπρα, σε επίκαιρη ερώτηση προς τον Πρωθυπουργό «Ακύρωση Μεσοπρό... http://t.co/MmeceTs

Εικόνα 11: Ο Σημασιολογικός Προβολέας της ομάδας που επιλέξαμε στο t-SNE, για τα tweets του @atsipras.

Για τα mentions προς τον @atsipras, παρατηρούμε ότι σχηματίζονται πολλές ομάδες στο χάρτη t-SNE, με μικρές όμως αποστάσεις μεταξύ τους (Εικόνα 12).



Εικόνα 12: Ο χάρτης t-SNE, για τα mentions προς τον @atsipras.

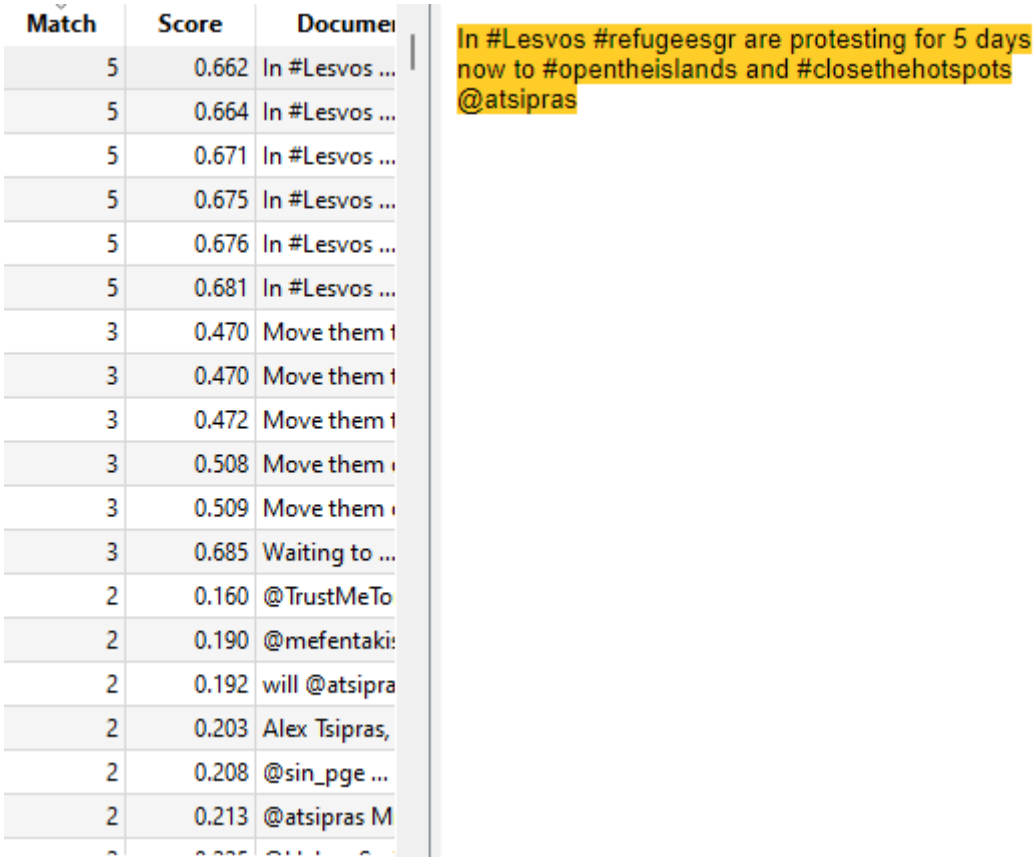
Μπορούμε να χρησιμοποιήσουμε τα ίδια εργαλεία, όπως έγινε παραπάνω, για να δούμε το περιεχόμενο μιας ομάδας.

Αφού επιλέξαμε μια διακριτή ομάδα, παρατηρούμε από τις λέξεις - κλειδιά που την χαρακτηρίζουν (Εικόνα 13), ότι τα 146 κείμενα που την αποτελούν, αφορούν, ενδεχομένως τις διαμαρτυρίες προσφύγων και μεταναστών στη Λέσβο, κατά τη διάρκεια της ανθρωπιστικής κρίσης του 2015.

Word	TF-IDF
time	0.016
refugeesgr	0.013
lesvos	0.013
closethehotspots	0.012
days	0.012
protesting	0.012
continue	0.012
media	0.010
forget	0.010
alex	0.010
obama	0.009
nnamdi	0.009
wrong	0.008
speak	0.008
learn	0.008
killing	0.008
kanu	0.008
refugees	0.008
unga	0.008

Εικόνα 13: Οι λέξεις - κλειδιά της ομάδας που επιλέξαμε στο t-SNE, για τα mentions προς τον @atsipras.

Εντάσσοντας αυτές τις λέξεις στο Σημασιολογικό Προβολέα, παρατηρούμε ότι μπορούν να εμφανιστούν ανά mention από 0 έως 6 φορές και με τιμές που κυμαίνονται από 0.016 έως 0.751 (Εικόνα 14).



Match	Score	Documer
5	0.662	In #Lesvos ...
5	0.664	In #Lesvos ...
5	0.671	In #Lesvos ...
5	0.675	In #Lesvos ...
5	0.676	In #Lesvos ...
5	0.681	In #Lesvos ...
3	0.470	Move them t
3	0.470	Move them t
3	0.472	Move them t
3	0.508	Move them
3	0.509	Move them
3	0.685	Waiting to ...
2	0.160	@TrustMeTo
2	0.190	@mefentaki
2	0.192	will @atsipra
2	0.203	Alex Tsipras,
2	0.208	@sin_pge ...
2	0.213	@atsipras M

Εικόνα 14: Ο Σημασιολογικός Προβολέας της ομάδας που επιλέξαμε στο t-SNE, για τα mentions προς τον @atsipras.

Τα αποτελέσματα της Σημασιολογικής Ανάλυσης γι' αυτά τα δύο σύνολα δεδομένων, μας βοηθούν να αποκτήσουμε μία εικόνα για το δημόσιο λόγο που αναπτύσσεται στο Twitter γύρω από θέματα της τρέχουσας επικαιρότητας και όχι μόνο. Βρίσκοντας, για παράδειγμα, τα tweets του @atsipras που αφορούν στην κοινοβουλευτική του παρουσία ή τα mentions κάποιων χρηστών γύρω από κοινωνικά θέματα, εντοπίζουμε τις λέξεις - κλειδιά που είναι χαρακτηριστικές γι' αυτές τις κειμενικές ομάδες και μπορούμε επιπλέον να μελετήσουμε το κατά πόσον είναι αντιπροσωπευτικές για το συνολικό Σώμα Κειμένου στο οποίο ανήκουν. Με τον τρόπο αυτό, εντοπίζουμε ενδιαφέροντα μοτίβα σχετικά με την εμφάνιση κάποιων λέξεων στα κείμενα και με βάση αυτές τις λέξεις, μπορούμε να προχωρήσουμε σε επόμενες ενέργειες, όπως για παράδειγμα την Κατηγοριοποίηση των Κειμένων.

3.4. Ευρήματα της Θεματικής Μοντελοποίησης

Όπως έχουμε αναφέρει, για να εφαρμόσουμε τη Θεματική Μοντελοποίηση, χρησιμοποιήσαμε τον αλγόριθμο της Λανθάνουσας Κατανομής Dirichlet, η οποία εντόπισε

τις κρυμμένες θεματικές και τις χαρακτηριστικές τους λέξεις, για τα δύο σύνολα δεδομένων που μελετάμε.

Αφού επιλέξουμε μια από τις 10 θεματικές που ανακαλύφθηκαν για τα tweets του @atsipras, μπορούμε να δούμε το περιεχόμενό της. Για παράδειγμα, εάν διαλέξουμε την πρώτη θεματική, παρατηρούμε ότι αρκετές από τις λέξεις που την χαρακτηρίζουν, πιθανότατα αφορούν εργασιακά ζητήματα (Εικόνα 15).

Topic	Topic keywords
1	ομοσπονδια, κεντρικη, εργαζομενων, πανελληνια, δεη, τελευταιες, ανακοινωση, εργασιας, προτασεως, λαρισας
2	αλεξη, συνεντευξη, κο, θεμα, ερωτησης, συριζα, επικαιρης, αποσπασματα, συζητηση, συγκεντρωση
3	τσιπρας, θεμα, συνεντευξη, τηλεοραση, ομαδας, συνεδριαση, νετ, κεα, κοινοβουλευτικης, συριζα
4	real, προεδρων, μνημονιο, ελλαδα, σαμαρας, ενωτικο, σημερ, ευρωπη, εκτελεστικη, κρατικου
5	ευρωπαικης, συνοδου, αριστερας, κομματος, εξελιξεις, πρεσβη, τυπο, πολιτικες, αγωνα, κουβας
6	κορυφης, εκτελεστικης, συνοδο, μερκελ, αποτελεσματα, αυτοδιοικησης, αποτυχια, σχετικα, ελλαδα, αναπτυξη
7	αντιπροεδρου, τεταρτη, ζωντανα, κεντρο, νεολαιας, γερμανικης, χαιρετισμος, οψει, ευρωζωνης, αριστερη
8	ανοιχτη, κουρουμπλη, δημοσιου, αντιπροεδρος, πανελλαδικης, εισηγηση, λειτουργιας, ελληνικου, κοζανη, αναπτυξης
9	αλεξης, επικεφαλης, σαββατο, ευρωβουλευτη, μετωπο, παπανδρεου, κοινωνικο, κυβερνηση, πανελλαδικη, αυγη
10	προεδρου, τσιπρα, συριζα, δηλωσεις, συναντηση, ερωτηση, δηλωση, πρωθυπουργο, ομιλια, επικαιρη

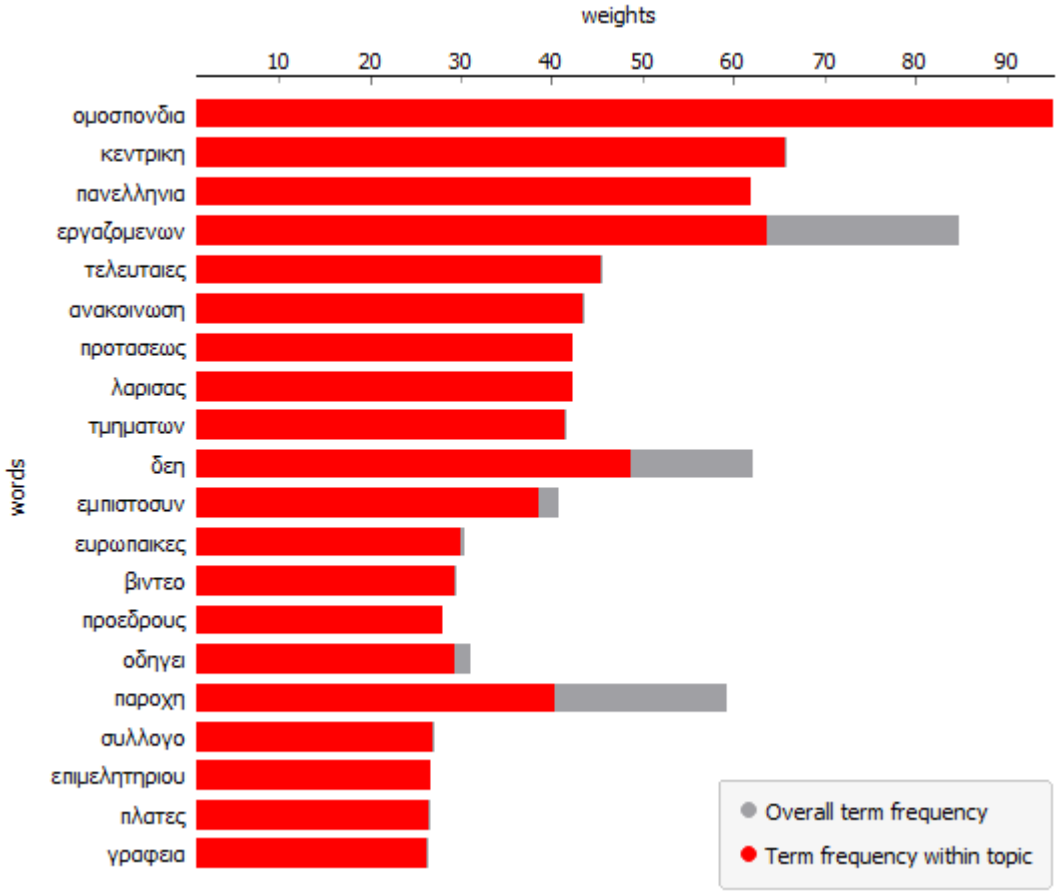
Εικόνα 15: Ο πίνακας με τις θεματικές, για τα tweets του @atsipras.

Με το Σύννεφο Λέξεων (Εικόνα 16), βλέπουμε τις λέξεις που έχουν τις περισσότερες εμφανίσεις σε αυτή την θεματική και έτσι, μπορούμε να επαληθεύσουμε την παραπάνω διαπίστωση για το περιεχόμενό της.

Εικόνα 16: Το Σύννεφο Λέξεων της πρώτης θεματικής, για τα tweets του @atsipras.

Χρησιμοποιώντας το εργαλείο LDAvis, εμφανίζουμε τις 20 λέξεις αυτής της θεματικής, ταξινομημένες ανάλογα το βαθμό σχετικότητάς τους. Για παράδειγμα, η

κορυφαία ταξινομημένη λέξη, "*ομοσπονδια*", εμφανίζεται σχεδόν αποκλειστικά σε αυτή τη θεματική. Από την άλλη, λέξεις όπως "*εργαζομένων*", "*δεη*" και "*παροχη*", αν και συγκεντρώνουν υψηλό βαθμό σχετικότητας σε αυτή την θεματική, εμφανίζονται και σε άλλες (Εικόνα 17).



Εικόνα 17: Το LDAvis με τις κορυφαίες λέξεις της πρώτης θεματικής, για τα tweets του @atsipras.

Ας εξετάσουμε το περιεχόμενο μιας θεματικής από το σύνολο δεδομένων των mentions που γράφτηκαν προς τον @atsipras. Για παράδειγμα, από τις λέξεις που χαρακτηρίζουν την έκτη θεματική, καταλαβαίνουμε ότι ίσως αφορούν διαπραγματεύσεις μεταξύ Ελλάδας και Ευρωπαϊκής Ένωσης, σχετικά με το ελληνικό χρέος (Εικόνα 18).

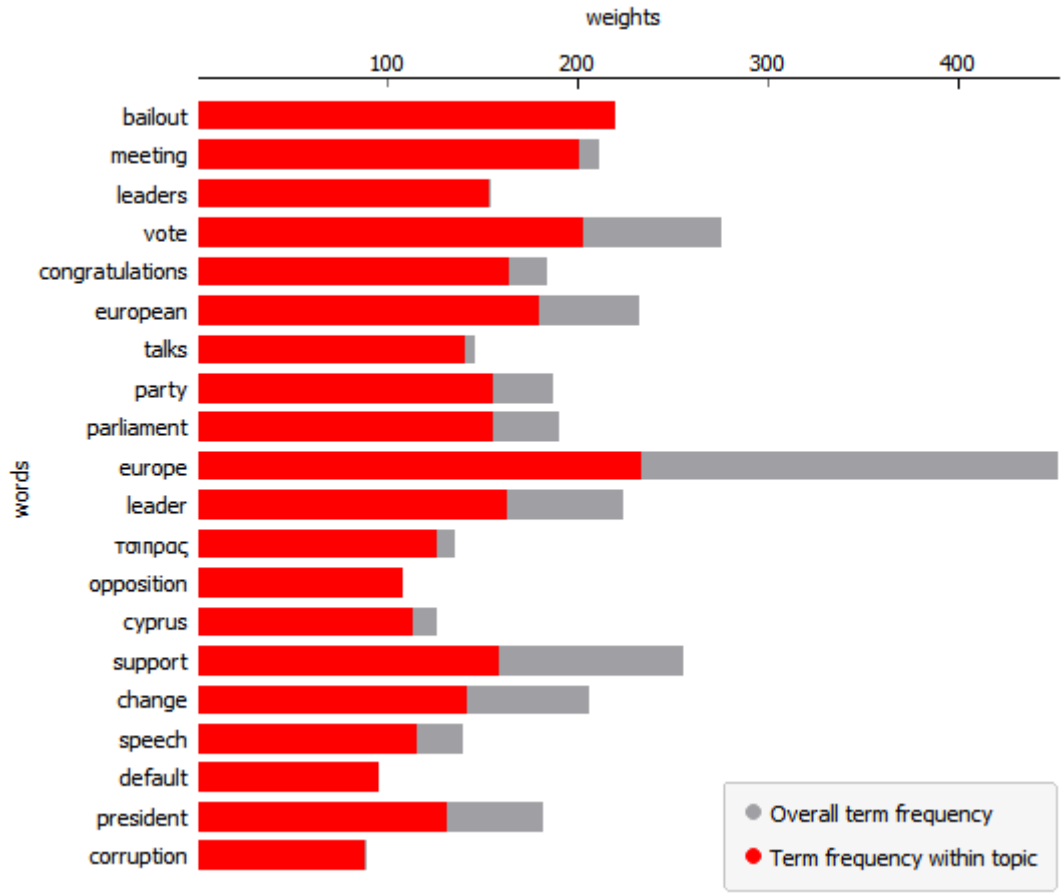
Topic	Topic keywords
1	eurogroup, brussels, election, tax, tomorrow, varoufakis, debt, ψευτης, elections, resign
2	minister, prime, imf, debt, austerity, ecb, victory, return, pay, plan
3	creditors, spain, voters, offer, promises, congrats, democracy, makedni, juncker, νεο
4	monday, letter, ξερει, lagarde, καλη, καλο, bank, markets, tweets, solution
5	λεει, κυβερνηση, κανει, ειπε, ξυπνησε, μεγαλη, εδω, αξιοπρεπεια, μελλον, ζωη
6	europe, bailout, vote, meeting, european, congratulations, leader, support, parliament, party
7	save, solution, lives, cuts, hospitals, νδ, sign, mandate, euronews, dignity
8	συριζα, αλεξη, eurozone, banks, λαου, ζητω, message, αγωνα, russia, avoid
9	ελλαδα, εκλογες, ομιλια, κυριακη, ευρωπη, interview, decide, πεισμα, game, happy
10	ψευτη, μαυριστε, παθολογικα, taxes, elected, ξεμπερδευουμε, νικησουμε, create, ισοδυναμος, καινουργιος

Εικόνα 18: Ο πίνακας με τις θεματικές, για τα mentions προς τον @atsipras.

Πράγματι, παρατηρώντας τις λέξεις που εμφανίζονται στο Σύννεφο Λέξεων (Εικόνα 19) για την έκτη θεματική, μπορούμε να επαληθεύσουμε την παραπάνω άποψη, καθώς εμφανίζονται οικονομικοί και πολιτικοί όροι που ταιριάζουν με γεγονότα που συνέβησαν στη χώρα μας κατά το πρόσφατο παρελθόν.

Εικόνα 19: Το Σύννεφο Λέξεων της έκτης θεματικής, για τα mentions προς τον @atsipras

Στο LDAvis, κοιτάζοντας τις 20 λέξεις που παρουσιάζουν τη μεγαλύτερη συχνότητα, μπορούμε να παρατηρήσουμε, ότι η κορυφαία σε βαθμολογία λέξη, "*bailout*", εμφανίζεται αποκλειστικά σε αυτή την θεματική, ενώ κάποιες άλλες λέξεις, όπως "*europe*", "*vote*", "*european*" και "*support*", ίσως να εμφανίζονται σε περισσότερες από μια θεματικές (Εικόνα 20).



Εικόνα 20: Το LDAvis με τις κορυφαίες λέξεις της έκτης θεματικής, για τα mentions προς τον @atsipras.

Η μέθοδος της Θεματικής Μοντελοποίησης, ουσιαστικά λειτουργεί ως "περίληψη" ενός Σώματος Κειμένου, εννοώντας ότι διαθέτει την ικανότητα εντοπισμού των κρυμμένων θεματικών, τις οποίες αναπαριστά με μια σειρά από λέξεις - κλειδιά. Όσες λέξεις εμφανίζονται στον πίνακα του ομώνυμου εργαλείου του Orange, χαρακτηρίζουν αυτές τις θεματικές και μπορεί να εμφανίζονται και σε άλλες. Ομοίως, ένα μικρό κειμενικό αντικείμενο, όπως ένα tweet ή mention, μπορεί να κατανέμεται ποσοστιαία σε παραπάνω από μια θεματικές. Όπως και οι υπόλοιπες μέθοδοι που χρησιμοποιήσαμε στην παρούσα έρευνα, έτσι και η Θεματική Μοντελοποίηση μπορεί να χρησιμοποιηθεί σε συνδυασμό με άλλες μεθόδους Εξόρυξης Δεδομένων, για την ανακάλυψη χρήσιμων πληροφοριών.

Κεφάλαιο 4: Επεξήγηση των Ευρημάτων

4.1. Επεξήγηση των Ευρημάτων της Ανάλυσης Συναισθημάτων

Από τα ευρήματα της Ανάλυσης των Συναισθημάτων για τα δύο σύνολα δεδομένων που μελετήσαμε, είναι δύσκολο να οδηγηθούμε σε κάποιο συμπέρασμα, δίχως να λάβουμε υπόψη τις κοινωνικές, οικονομικές και πολιτικές συγκυρίες κατά τις οποίες αναρτήθηκαν αυτά τα tweets και mentions. Γι' αυτόν τον λόγο, προτείνουμε τα ευρήματα των Συναισθημάτων, να ιδωθούν ανά έτος, σύμφωνα με το ιστορικό πλαίσιο στο οποίο δημιουργήθηκαν.

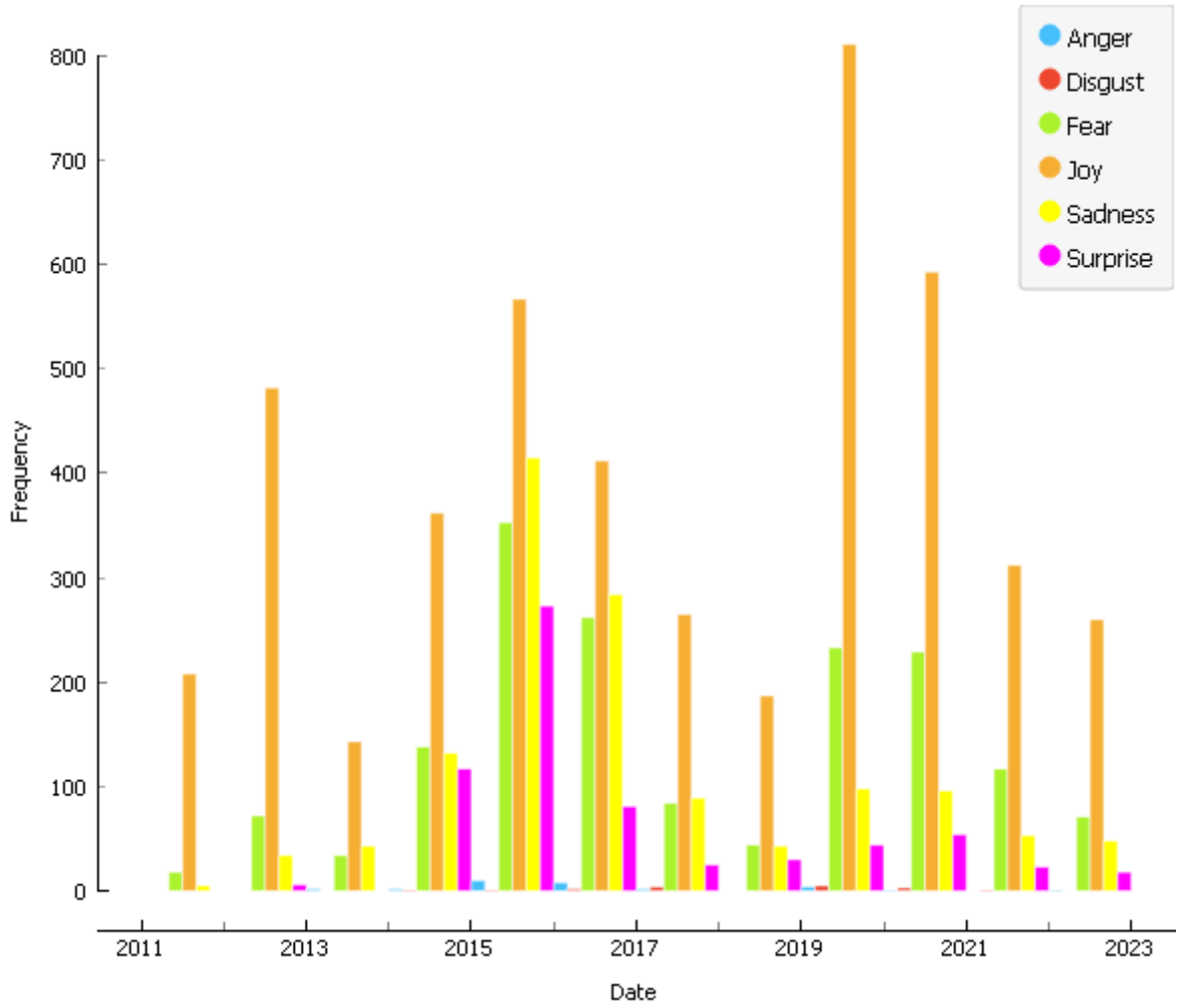
Παρατηρώντας τις κατανομές των tweets του @atsipras, βλέπουμε ότι το 2015 ήταν η χρονιά που ανήρτησε τα περισσότερα (19,47% επί του συνόλου). Να σημειωθεί ότι στις 25 Ιανουαρίου του 2015, ο ΣΥ.ΡΙΖ.Α. κέρδισε στις εθνικές εκλογές με ποσοστό 35,4% και σχημάτισε κυβέρνηση συνεργασίας. Κατά τους επόμενους μήνες, ξεκίνησε διαπραγματεύσεις με τους δανειστές για την αποπληρωμή του ελληνικού χρέους και με την αποτυχία αυτών των διαπραγματεύσεων, ανακοίνωσε στις 27 Ιουνίου τη διεξαγωγή δημοψηφίσματος για τις 5 Ιουλίου, με θέμα τη ψήφιση ή την απόρριψη των προτάσεων των ευρωπαίων για τη συμφωνία. Αποτέλεσμα του δημοψηφίσματος, ήταν το "ΟΧΙ", σε ποσοστό 61,31%. Η πολιτική αβεβαιότητα σε συνδυασμό τα capital controls που επιβλήθηκαν, οδήγησαν τον τότε πρωθυπουργό, Αλέξη Τσίπρα, να προκηρύξει πρόωρες εκλογές για τις 20 Σεπτεμβρίου του 2015[77]. Εκείνη την περίοδο, η Ελλάδα βρέθηκε στο επίκεντρο διεθνών και ευρωπαϊκών Μέσων Ενημέρωσης, αλλά αποτέλεσε επίσης και αντικείμενο συζητήσεων στα ΜΚΔ και ειδικά στο Twitter[78].

Δεύτερο έτος σε αριθμούς, για τα tweets του @atsipras, ήταν το 2019 (14,37% επί του συνόλου). Τον Ιανουάριο του 2019, οι ΑΝ.ΕΛ. αποχώρησαν από το κυβερνητικό σχήμα, λόγω της στάσης που κράτησε η κυβέρνηση στο "Μακεδονικό", με αποτέλεσμα ο Αλέξης Τσίπρας να ζητήσει ψήφο εμπιστοσύνης την οποία και έλαβε. Μετά την ήττα του ΣΥ.ΡΙΖ.Α.

[77] Τζανάκης, Σ. (2022, 4 Ιουλίου). Το πικρό καλοκαίρι του 2015: Όταν η Ελλάδα βρέθηκε να χορεύει στο χείλος του γκρεμού. *Πρώτο Θέμα*. Ανάκτηση από: https://www.protothema.gr/greece/article/1261180/to-pikro-kalokairi-tou-2015-otan-i-ellada-vrethike-horeuodas-sto-heilos-tou-gremou/ Τελευταία πρόσβαση: 31/5/2023.
[78] Harding, S. (2015, Jul 9). Greek Crisis: Twitter actively shapes fast-moving events. *The Guardian.* Ανάκτηση από: https://www.theguardian.com/world/2015/jul/09/greek-crisis-twitter-social-media Τελευταία πρόσβαση: 31/5/2023.

στις ευρωεκλογές και τις αυτοδιοικητικές εκλογές του Μαίου, κηρύχθηκαν πρόωρες εκλογές για τις 7 Ιουλίου του 2019, όπου και ηττήθηκε από τη Νέα Δημοκρατία με 31,53%[79].

Έχοντας υπόψη τα παραπάνω ιστορικά στοιχεία, παρατηρούμε στο γράφημα που δημιουργήσαμε ανά έτος (Εικόνα 21), ότι τα tweets του @atsipras για το 2015, έχουν υψηλές τιμές στα συναισθήματα Χαράς (567 tweets), Λύπης (415 tweets), αλλά και Φόβου (353 tweets). Για το 2019, δηλαδή το έτος που ολοκληρώθηκε η τετραετία της διακυβέρνησης του ΣΥ.ΡΙΖ.Α., τα συναισθήματα που παρουσιάζουν τις πιό υψηλές τιμές είναι η Χαρά (811 tweets), ο Φόβος (233 tweets) και η Λύπη (98 tweets).

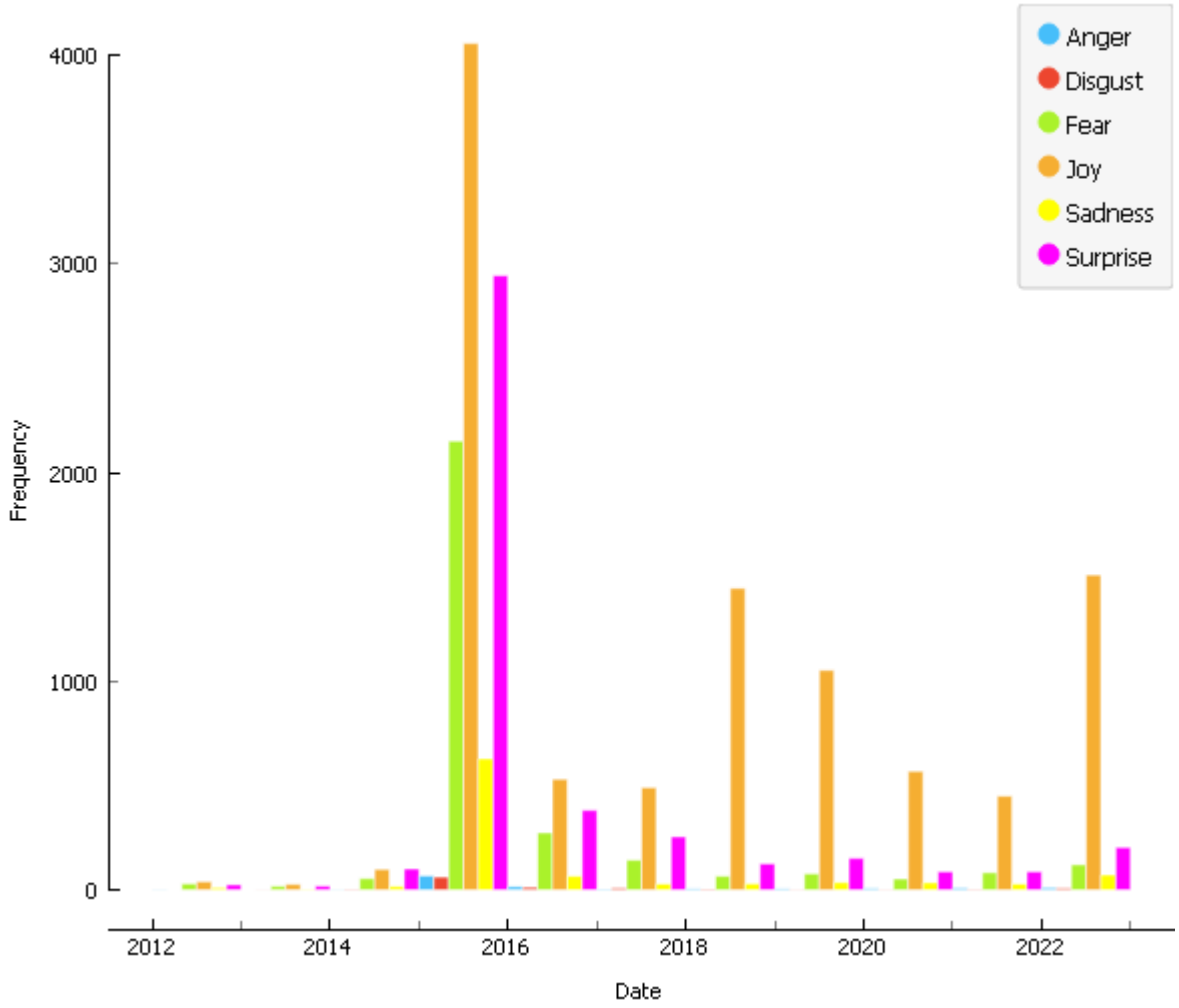


Εικόνα 21: Η κατανομή των Συναισθημάτων ανά έτος, για τα tweets του @atsipras.

Για τα mentions προς τον @atsipras, παρατηρούμε ότι μέχρι το 2015, είχε ελάχιστα mentions ανά έτος (Εικόνα 22). Το 2015, σημειώθηκε μεγάλη άνοδος στον αριθμό των mentions (52,38% επί του συνόλου), με πολύ υψηλές συγκεντρώσεις στα συναισθήματα

[79] Δεμέτης, Χ. (2019, 7 Ιουλίου). Εθνικές εκλογές 2019: Μπροστά η ΝΔ με 38,9%, στο 31,5% ο ΣΥΡΙΖΑ. *news247.* Ανάκτηση από: https://www.news247.gr/ekloges/ethnikes-ekloges-2019-mprosta-i-nd-me-39-8-sto-teliko-apotelesma.7471185.html Τελευταία πρόσβαση: 31/5/2023.

Χαράς (4.058 mentions), Έκπληξης (2.945 mentions) και Φόβου (2.152 mentions). Αξίζει να σημειωθεί ότι τα mentions προς τον @atsipras είναι στην μεγάλη τους πλειοψηφία γραμμένα στην αγγλική γλώσσα (90,11% επί του συνόλου) και κατά τη διάρκεια του έτους 2015, εύρημα που αποτελεί ένδειξη για το εύρος της δημοσιότητας που έλαβε αυτό το πολιτικό πρόσωπο, εκείνη τη συγκεκριμένη περίοδο από χρήστες της πλατφόρμας Twitter.



Εικόνα 22: Η κατανομή των Συναισθημάτων ανά έτος, για τα mentions προς τον @atsipras.

Ταξινομώντας τα tweets και τα mentions, ανάλογα με το Συναίσθημα που προκαλούν, μπορούμε να αντλήσουμε ενδιαφέρουσες πληροφορίες σχετικά με τη μεταβολή των συναισθημάτων κατά την πάροδο του χρόνου, τόσο για τα συναισθήματα που εξέφρασε ο @atsipras στο Twitter καθ' όλη τη διάρκεια της παρουσίας του στο Μέσο, όσο και για τα συναισθήματα εκείνων των χρηστών, που έκαναν ονομαστική αναφορά προς αυτό τον πολιτικό, με αφορμή γεγονότα που απασχόλησαν την επικαιρότητα.

4.2. Επεξήγηση των Ευρημάτων της Συσταδοποίησης Κειμένων

Με βάση τα ευρήματα της Συσταδοποίησης Κειμένων, μπορούμε να πούμε ότι οι διαφορές μεταξύ των συστάδων τείνουν να είναι πιο ευδιάκριτες όταν τα κείμενα γράφονται από ένα άτομο, καθώς συνήθως χρησιμοποιεί συγκεκριμένες λέξεις, όπως είδαμε στην περίπτωση του @atsipras. Έτσι, σε κάθε συστάδα, εντοπίζονται αρκετές κοινές λέξεις ανάμεσα στα κείμενα που την αποτελούν.

Από την άλλη, στα κείμενα που έχουν γραφτεί από πολλούς χρήστες, όπως δηλαδή έγινε με τα mentions, ο διαχωρισμός των συστάδων γίνεται δύσκολα διακριτός από τον ερευνητή, καθώς εντός της κάθε συστάδας υπάρχουν κείμενα που έχουν διαφορετικό τρόπο γραφής ή μπορεί να περιέχουν λέξεις από άλλη γλώσσα. Έτσι, μπορεί ακόμα να παρατηρηθούν συστάδες, που τα κείμενά τους δεν έχουν καμμία λέξη κοινή. Εξαίρεση αποτελούν, όπως είδαμε στο υποκεφάλαιο με τα "*Ευρήματα της Συσταδοποίησης Κειμένων*", τα πανομοιότυπα κείμενα, που μπορεί είτε να είναι αποτέλεσμα copy-paste μεταξύ των χρηστών ή ακόμα και να έχουν αναρτηθεί από bots[80].

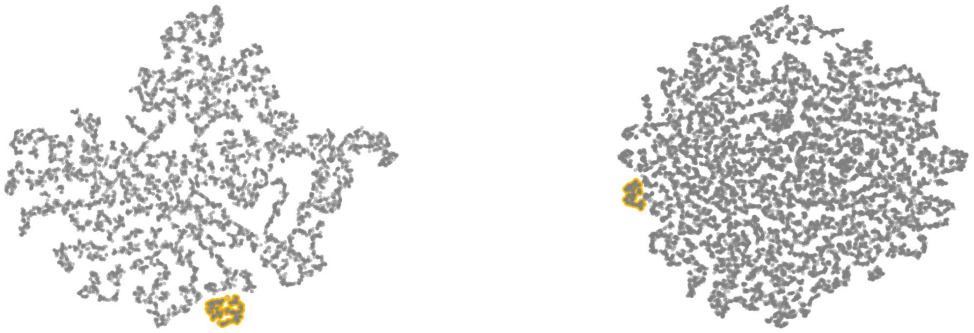
Για να εφαρμόσουμε τη Συσταδοποίηση Κειμένων, διαχωρίσαμε τα κείμενα των δύο συνόλων δεδομένων, σε συστάδες, σύμφωνα με την μέτρηση των αποστάσεων μεταξύ των αριθμητικών χαρακτηριστικών που τους δόθηκαν. Με αυτό τον τρόπο, τα κείμενα που ανήκουν σε μία συστάδα, παρουσιάζουν ομοιότητα μεταξύ τους και μπορούν να χρησιμοποιηθούν ως σύνολο για την περαιτέρω επεξεργασία των κειμενικών δεδομένων, με σκοπό την ανακάλυψη ενδιαφέροντων μοτίβων και χρήσιμων πληροφοριών.

4.3. Επεξήγηση των Ευρημάτων της Σημασιολογικής Ανάλυσης

Όπως αναφέραμε, διαβάζοντας τις λέξεις - κλειδιά μιας ομάδας κειμένων, όπως αυτή κατατάσσεται στον χάρτη του t-SNE, μπορούμε να κατανοήσουμε το σημασιολογικό περιεχόμενο της και με τη χρήση του Σημασιολογικού Προβολέα, να εξετάσουμε εάν αυτές οι λέξεις εμφανίζονται στο συνολικό Σώμα Κειμένου και κατά πόσον το χαρακτηρίζουν, με βάση τις τιμές τους.

[80]Aljabri, M., Zagrouba, R., Shaahid, A. *et al.* (2023, January 5). Machine learning-based social media bot detection: a comprehensive literature review. In *Social Network Analysis and Mining.* 13, 20 (2023). https://doi.org/10.1007/s13278-022-01020-5 Ανάκτηση από: https://link.springer.com/article/10.1007/s13278-022-01020-5#citeas Τελευταία πρόσβαση: 31/5/2023.

Εάν συγκρίνουμε τους χάρτες t-SNE των δύο συνόλων δεδομένων, παρατηρούμε ότι οι κατανομές των κειμένων στο χώρο, διαφέρουν αισθητά μεταξύ τους, παρόλο που χρησιμοποιήσαμε τις ίδιες ρυθμίσεις και ακολουθήσαμε τα ίδια βήματα. Συγκεκριμένα, ο χάρτης των tweets, εμφανίζει αρκετά διακριτές ομάδες με μεγάλες αποστάσεις μεταξύ τους, ενώ ο χάρτης με τα mentions φαίνεται να είναι πιό συγκεντρωτικός και να αποτελείτσαι από μη-διακριτές ομάδες, με κοντινές αποστάσεις (Εικόνα 23).



Εικόνα 23: Οι χάρτες t-SNE, για τα δύο σύνολα δεδομένων.

Από τις ομάδες που επιλέξαμε στο κάθε σύνολο δεδομένων, εξάγαμε τις χαρακτηριστικές τους λέξεις και εξετάσαμε εάν αυτές εμφανίζονται στα κείμενα των υπολοίπων tweets ή mentions. Όσο υψηλότερες είναι οι Αντιστοιχίσεις και οι Τιμές των λέξεων στο Σημασιολογικό Προβολέα, τόσο πιό πολύ τείνουν να χαρακτηρίζουν το κείμενο που αναλύουμε.

Ποιές είναι όμως οι χαρακτηριστικές λέξεις - κλειδιά, για το κάθε ένα σύνολο δεδομένων;

Επιλέγοντας ολόκληρο το σύνολο δεδομένων με τα tweets του @atsipras, βρήκαμε ότι οι 10 χαρακτηριστικές λέξεις αυτών των κειμένων, είναι: “*συριζα*”, “*τσιπρα*”, “*προεδρου*”, “*αλεξη*”, “*ομιλια*”, “*κυβερνηση*”, “*ελλαδα*”, “*βουλη*”, “*εκμ*” και “*κυβερνηση*”. Παράλληλα, αυτές οι λέξεις, μπορούν να εντοπιστούν ανά tweet από 0 έως 8 φορές και με τιμές από -0.046 έως 0.739.

Οι 10 λέξεις - κλειδιά των mentions προς τον @atsipras, είναι: “*monday*”, “*solution*”, “*time*”, “*europe*”, “*country*”, “*greeks*”, “*save*”, “*deal*”, “*government*” και “*turkey*” και μπορούν να εντοπιστούν ανά mention από 0 έως 8 φορές και με τιμές από 0.018 έως 0.797.

Η εφαρμογή της μεθόδου της Σημασιολογικής Ανάλυσης, μας βοήθησε να αποκτήσουμε μία πρώτη εικόνα για το περιεχόμενο αυτών των δύο συνόλων δεδομένων και

να εντοπίσουμε ποιές είναι αυτές οι λέξεις που χαρακτηρίζουν τα Σώματα των Κειμένων στα οποία ανήκουν.

4.4. Επεξήγηση των Ευρημάτων της Θεματικής Μοντελοποίησης

Όπως αναφέραμε στο υποκεφάλαιο "*Πραγματοποίηση της Θεματικής Μοντελοποίησης*", κάθε κείμενο tweet ή mention, μπορεί να ανήκει ποσοστιαία, σε πάνω από μια θεματικές. Με τη Θεματική Μοντελοποίηση, εντοπίζουμε τις κρυμμένες θεματικές που αφορούν ένα κείμενο, καθώς και τις λέξεις - κλειδιά που τις χαρακτηρίζουν. Ποιές είναι όμως οι θεματικές που ανακάλυψε ο αλγόριθμος της Λανθάνουσας Κατανομής Dirichlet για τα δύο σύνολα δεδομένων;

Δεν είναι πάντα εύκολο να εξάγουμε το ακριβές νόημα μιας θεματικής. Για παράδειγμα, η πρώτη θεματική των tweets του @atsipras που αναλύσαμε παραπάνω, αποτελείται από τις λέξεις - κλειδιά: "*ομοσπονδια*", "*κεντρικη*", "*εργαζομενων*", "*πανελληνια*", "*δεη*", "*τελευταιες*", "*ανακοινωση*", "*εργασιας*", "*προτασεως*" και "*λαρισας*". Όπως είδαμε, βάσει αυτών των λέξεων, μπορούμε να υποθέσουμε ότι η συγκεκριμένη θεματική αναφέρεται σε κινητοποιήσεις εργαζομένων και αυτό το επιβεβαιώσαμε με το Σύννεφο Κειμένων και το LDAvis. Αυτό ισχύει και για άλλες θεματικές των tweets που αναρτήθηκαν από τον @atsipras.

Τί συμβαίνει όμως, όταν οι συγγραφείς των κειμένων είναι περισσότεροι και μάλιστα χιλιάδες; Οι θεματικές που ανακαλύπτονται από τον αλγόριθμο LDA, μπορούν να μας βοηθήσουμε να κατανοήσουμε το περιεχόμενο ενός κειμένου;

Κοιτάζοντας την ίδια θεματική από τα mentions προς τον @atsipras που χρησιμοποιήσαμε ως παράδειγμα στο υποκεφάλαιο "*Ευρήματα της Θεματικής Μοντελοποίησης*", παρατηρούμε ότι οι λέξεις - κλειδιά της είναι: "*europe*", "*bailout*", "*vote*", "*meeting*", "*european*", "*congratulations*", "*leader*", "*support*", "*parliament*" και "*party*". Πράγματι, πιθανότατα αναφέρεται στις διαπραγματεύσεις μεταξύ Ευρωπαϊκής Ένωσης και Ελλάδας, με αφορμή το ελληνικό χρέος.

Ανάλογα ευρήματα μπορούν να εντοπιστούν και να αναλυθούν σε όλες τις θεματικές των δύο συνόλων δεδομένων που αποκτήσαμε. Βλέποντας τις λέξεις από τις οποίες χαρακτηρίζεται μία θεματική, ανακαλύπτουμε χρήσιμα στοιχεία για το περιεχόμενό της και

μπορούμε να προβούμε σε περαιτέρω ενέργειες ανάλυσης, όπως αυτή της Κατηγοριοποίησης Κειμένων.

Κεφάλαιο 5: Συζήτηση και Συμπεράσματα

Στο κεφάλαιο αυτό, θα μελετήσουμε εάν τελικά απαντώνται τα 4 βασικά ερευνητικά ερωτήματα που θέσαμε στην Εισαγωγή της παρούσας Διπλωματικής εργασίας. Σκοπός μας, εξ' αρχής, ήταν να εξαχθούν χρήσιμα συμπεράσματα από αυτή την έρευνα και να αποτελέσει αυτή, ει δυνατόν, πηγή έμπνευσης για μελλοντικές έρευνες Ανάλυσης Δεδομένων που αφορούν δημόσια πρόσωπα στην Ελλάδα.

Το πρώτο ερευνητικό ερώτημα που θέσαμε, ήταν: "Τί είδους συναισθήματα μπορούν να δημιουργηθούν κατά την έκφραση λόγου στο Twitter, από και προς τον λογαριασμό του @atsipras, σε συνάρτηση με τις συγκυρίες κατά τις οποίες γράφτηκαν;".

Αναλύοντας τα tweets του πρώην πρωθυπουργού και νυν αρχηγού της αξιωματικής αντιπολίτευσης, Αλέξη Τσίπρα, ανακαλύψαμε πώς υπάρχουν πολλοί τρόποι που αυτά μπορούν να ιδωθούν επιστημονικά. Ανάλογα με τη δομή των δεδομένων, μπορούμε να χρησιμοποιήσουμε και το κατάλληλο μοντέλο Ανάλυσης Συναισθημάτων. Αυτό είναι κάτι που αφορά όλα τα κείμενα, όχι μόνο αυτά που δημιουργούνται στα ΜΚΔ, και σίγουρα δεν αφορά αποκλειστικά τον συγκεκριμένο λογαριασμό.

Τα συγκεκριμένα μοντέλα που χρησιμοποιεί το λογισμικό του Orange, έχουν μία ιδιαιτερότητα: δε διαθέτουν ικανοποιητικό Λεξικό για τα ελληνικά, που να βρίσκει πάντα και να καταχωρεί τιμές συναισθήματος. Για το λόγο αυτό, δεν χρησιμοποιήσαμε κάποιο μοντέλο ανάλυσης που δίνει αριθμητικές τιμές, με θετικό ή αρνητικό πρόσημο, όπως η Πολυγλωσσική ανάλυση, ούτε κάποιο μοντέλο που να κατηγοριοποιεί τα κείμενα σε θετικά, αρνητικά και ουδέτερα.

Οπότε, επιλέξαμε τη μέθοδο της κατηγοριοποίησης των συναισθημάτων με βάση κάποιο μοντέλο, όπως αυτό του Ekman, το οποίο μπορεί και διακρίνει τα tweets, με βάση τα 6 βασικά Συναισθήματα. Η επιλογή αυτού του μοντέλου έναντι των άλλων, έγινε επειδή το θεωρήσαμε ως το πιο κατάλληλο για την εξαγωγή συναισθήματος από κειμενικά αρχεία του Twitter.

Όπως είδαμε, στο σύνολο των 8.315 tweets που γράφτηκαν, από τις 13 Ιουλίου του 2011 έως τις 31 Δεκεμβρίου του 2022, από τον @atsipras, εντοπίστηκαν 4.602 tweets με αίσθημα Χαράς, 1.655 Φόβου, 1.340 Λύπης, 671 Έκπληξης, 30 Θυμού και 17 Απέχθειας. Στο υποκεφάλαιο "*Επεξήγηση των Ευρημάτων της Ανάλυσης Συναισθημάτων*", αναφέραμε ότι εμφανίζονται κάποια ενδιαφέροντα μοτίβα στην κατανομή των Συναισθημάτων ανά έτος, που αφορούν συγκεκριμένες κοινωνικές, οικονομικές και πολιτικές συγκυρίες.

Ας δούμε τα συναισθήματα που εκφράστηκαν από κάποιους χρήστες προς τον λογαριασμό του @atsipras, από τις 2 Μαΐου του 2012 έως τις 31 Δεκεμβρίου του 2022. Αρχικά, θα πρέπει να αναφέρουμε ότι πρόκειται για mentions, δηλαδή αναφορές χρηστών του Twitter. Κάθε mention που αναφέρει το όνομα ενός χρήστη π.χ. του @atsipras, δεν αναφέρει συνήθως μόνο αυτό, αλλά και ονόματα άλλων χρηστών. Επίσης, σε αντίθεση με τα replies (απαντήσεις), τα mentions δεν αντιστοιχούν σε κάποιο tweet, οπότε είναι δύσκολο να καταλάβουμε εάν τα συναισθήματα που εντοπίζονται, αφορούν κάποιο συγκεκριμένο tweet που περιέχει μια δήλωση ή που αναφέρει μία δράση που πραγματοποίησε ο συγκεκριμένος πολιτικός π.χ. συνέντευξη, επίσκεψη κλπ.

Τα 18.926 mentions των χρηστών προς το λογαριασμό του @atsipras, με την χρήση του μοντέλου του Ekman, κατηγοριοποιήθηκαν σε συναισθήματα: 10.274 Χαράς, 4.385 Έκπληξης, 3.073 Φόβου, 956 Λύπης, 131 Θυμού και 107 Απέχθειας.

Παρατηρώντας επίσης και σε αυτό το σύνολο δεδομένων τη χρονική περίοδο που γράφτηκαν, μπορούμε να αντλήσουμε σημαντικά στοιχεία για τα συναισθήματα που εντοπίζονται στα κείμενά τους σε συνάρτηση με τις τότε συγκυρίες.

Το δεύτερο ερώτημα που θέσαμε, ήταν: "Μπορούν να ομαδοποιηθούν, ανάλογα την ομοιότητά τους, τα κείμενα πολιτικού περιεχομένου που δημιουργήθηκαν και αν ναι, τι ευρήματα προκύπτουν από την ομαδοποίηση;".

Η απάντηση σε αυτό το ερώτημα είναι σχετικά απλή. Εφόσον τα κείμενα αυτά δεν έχουν ταξινομηθεί προηγουμένως με βάση κάποιο χαρακτηριστικό τους γνώρισμα, μέσω μιας μεθόδου Κατηγοριοποίησης Κειμένου, μπορούν να ομαδοποιηθούν ανάλογα με τις ομοιότητες που παρουσιάζουν, με τη μέθοδο της Συσταδοποίησης Κειμένου. Με την εφαρμογή αυτής της μεθόδου, τα κείμενα που παρουσιάζουν μεταξύ τους ομοιότητες εντάσσονται στην ίδια συστάδα και ταυτόχρονα διαφέρουν από τα κείμενα που βρίσκονται σε άλλες συστάδες.

Τα σύνολα δεδομένων που αναλύθηκαν στην παρούσα έρευνα, έχουν την ιδιαιτερότητα να είναι αρκετά μεγάλα και ανομοιογενή ως προς το περιεχόμενό τους. Γι' αυτόν το λόγο, επιλέχθηκε μια αυστηρή μέθοδος ομαδοποίησής τους, αυτή της Ιεραρχικής Συσταδοποίησης, με χρήση της μεθόδου του Ward, η οποία θεωρείται κατάλληλη για τέτοιου είδους αρχεία και ελεύθερη κοπή του Δενδρογράμματος σε σημείο που να γίνονται εμφανείς οι διακριτές συστάδες.

Έτσι, εφαρμόζοντας τις συγκεκριμένες ενέργειες της Ιεραρχικής Συσταδοποίησης, δημιουργήσαμε πολλές και μικρές συστάδες, για τα tweets και για τα mentions προς τον @atsipras. Με αυτόν τον τρόπο, έχουμε μικρές και συμπαγείς ομάδες κειμένων, που μοιράζονται κάποια κοινά χαρακτηριστικά. Για παράδειγμα, τα tweets που αφορούν στην προεκλογική ομιλία του @atsipras σε μία πόλη ή μια συνέντευξη σε ένα κανάλι, θα συγκροτούν τις αντίστοιχες ομάδες με τα ανάλογα κείμενα. Ομοίως, συμπαγείς ομάδες δημιουργούν και τα mentions που αναφέρονται π.χ. σε ένα κοινωνικό θέμα και χρησιμοποιούν συγκεκριμένες λέξεις.

Οπότε, απαντώντας σε αυτό το ερώτημα, τα κείμενα ομαδοποιούνται ανάλογα το βαθμό ομοιότητάς τους και κάποιες κοινές λέξεις που μοιράζονται μεταξύ τους.

Η τρίτη ερώτηση αφορά το εάν "Υπάρχουν λέξεις - κλειδιά που εμφανίζονται στις αναρτήσεις του @atsipras και των ονομαστικών αναφορών προς αυτόν και αν ναι, είναι χαρακτηριστικές των corpora;".

Σε αυτό το ερώτημα απαντήσαμε στο υποκεφάλαιο "*Επεξήγηση των Ευρημάτων της Σημασιολογικής Ανάλυσης*".

Είτε εξετάσουμε τις λέξεις που εμφανίζονται στο χάρτη του t-SNE σε μία ομάδα κειμένων, είτε επιλέξουμε όλα τα κείμενα, για το κάθε ένα από τα δύο σύνολα δεδομένων, μπορούμε να εξάγουμε τις χαρακτηριστικές λέξεις - κλειδιά και έτσι να κατανοήσουμε το σημασιολογικό περιεχόμενο των κειμένων στα οποία ανήκουν.

Συνεπώς, επιλέγοντας όλα τα κείμενα που γράφτηκαν από τον @atsipras, εντοπίζουμε τις λέξεις - κλειδιά: "*συριζα*", "*τσιπρα*", "*προεδρου*", "*αλεξη*", "*ομιλια*", "*κυβερνηση*", "*ελλαδα*", "*βουλη*", "*εκμ*" και "*κυβερνηση*" και από αυτά που γράφτηκαν από τους υπόλοιπους χρήστες, τις λέξεις: "*monday*", "*solution*", "*time*", "*europe*", "*country*", "*greeks*", "*save*", "*deal*", "*government*" και "*turkey*", συνοδευόμενες όλες από τις TF - IDF τιμές τους, που εμφανίζουν τη συχνότητα της παρουσίας τους εντός του Σώματος Κειμένου. Έτσι, μπορούμε να δούμε εάν είναι χαρακτηριστικές των δύο corpora.

Το τέταρτο ερευνητικό ερώτημα αφορά το εάν “Υπάρχουν θεματικές που αναδύονται από τις αναρτήσεις του @atsipras και τις αναφορές χρηστών προς αυτόν;”

Σαφώς εντοπίζονται θεματικές μέσα σε ένα Σώμα Κειμένου και ο τρόπος για να εξαχθούν αυτές, είναι με τη χρήση της μεθόδου της Θεματικής Μοντελοποίησης, που εμφανίζει έναν αριθμό θεματικών και τις χαρακτηριστικές τους λέξεις. Μέσα από τις λέξεις αυτές, αντιλαμβανόμαστε και το νόημα της θεματικής στην οποία εντάσσονται.

Με βάση τα ευρήματα της ανάλυσής μας, κάποιες από τις θεματικές των tweets που έγραψε ο @atsipras, αφορούν τα εργασιακά, τη δημόσια ή κοινοβουλευτική παρουσία του Αλέξη Τσίπρα, διεθνή θέματα, συνεδριάσεις και επισκέψεις στο εξωτερικό. Οι θεματικές που ανακαλύπτονται στα mentions των χρηστών, αφορούν στις σχέσεις μεταξύ Ελλάδας και Ευρωπαϊκής Ένωσης, θέματα της πολιτικής επικαιρότητας, διεθνή ή σε κάποιες, λίγες περιπτώσεις δεν φαίνεται να βγάζουν κάποιο νόημα, καθώς αυτό το σύνολο δεδομένων αποτελείται από λέξεις σε αγγλικά και ελληνικά, γεγονός που ίσως αποτρέπει τον αλγόριθμο από το να πραγματοποιήσει τους κατάλληλους συσχετισμούς μεταξύ των λέξεων και έτσι να εξαχθεί κάποιο συμπέρασμα.

Κεφάλαιο 6: Περιορισμοί και Προτάσεις για Μελλοντική Έρευνα

Ουσιαστικά, δεν υπήρξαν κάποιοι ιδιαίτεροι περιορισμοί ή δυσκολίες κατά την εκπόνηση της παρούσας έρευνας. Στο Διαδίκτυο είναι αναρτημένοι όλοι οι τρόποι με τους οποίους μπορούμε να αποκτήσουμε δεδομένα από μια πλατφόρμα ΜΚΔ, το ίδιο και για τα επόμενα βήματα που αφορούν στην επεξεργασία και ανάλυσή τους.

Οι περιορισμοί έγκεινται κυρίως στις μεθόδους εξόρυξης που υπάρχουν. Αφού αποφασίσουμε το περιεχόμενο των δεδομένων που θέλουμε να εξάγουμε, εάν δηλαδή επιθυμούμε, για παράδειγμα μια real - time άντληση δεδομένων που καταχωρούνται με κάποιο hashtag ή ακόμα και ολόκληρο το χρονολόγιο (timeline) ενός χρήστη, όπως έγινε στην προκειμένη περίπτωση, μπορούμε να διαλέξουμε και το ανάλογο εργαλείο εξόρυξης δεδομένων. Για τις ανάγκες της έρευνας μας, επιλέξαμε την βιβλιοθήκη Snscrape, έναντι άλλων, καθώς μέχρι τη στιγμή που αποκτήσαμε τα δεδομένα, δεν έθετε περιορισμούς στην εξαγωγή των δεδομένων από τα ΜΚΔ.

Αφού αντλήσαμε τα δεδομένα από το Διαδίκτυο, επόμενο στάδιο ήταν η προεπεξεργασία τους. Ο σωστός καθαρισμός των δεδομένων, παράγει στην συνέχεια "καθαρά" αποτελέσματα. Ένα σχετικά μικρό σύνολο δεδομένων, σε μορφή .csv ή .xlsx, με λίγες και απαλλαγμένες από "θόρυβο" στήλες, μπορεί στη συνέχεια να αναλυθεί με μεγάλη ευκολία χρησιμοποιώντας τα κατάλληλα εργαλεία.

Όσον αφορά το λογισμικό ανοιχτού κώδικα Orange, το οποίο και χρησιμοποιήσαμε για τις αναλύσεις μας, αυτό χρησιμοποιείται διεθνώς από Πανεπιστήμια και data analysts, για την ανάλυση και οπτικοποίηση δεδομένων. Μάλιστα, υπάρχει αναρτημένη πληθώρα οπτικοακουστικού υλικού και επεξηγηματικών κειμένων, που αναλύουν με επεξηγηματικό τρόπο τις λειτουργίες και τα εργαλεία του λογισμικού αυτού.

Επίσης, αν και σχετικά πρόσφατος ο κλάδος της Ανάλυσης Δεδομένων, υπάρχει διαθέσιμο πλούσιο βιβλιογραφικό υλικό με το θεωρητικό υπόβαθρο πίσω από τις ενέργειες που ακολουθούνται, με πληθώρα πληροφοριών σχετικά με τεχνικές, μεθόδους και αλγόριθμους. Στην παρούσα έρευνα, παρουσιάστηκε ένα μόνο μέρος των μεθόδων που υπάρχουν για την απόκτηση γνώσης μέσα από κειμενικά δεδομένα.

Μια πρόταση για μελλοντική έρευνα, θα ήταν να διεξαχθούν και άλλες μελέτες - ακαδημαϊκές και μη - για την παρουσία πολιτικών και ευρύτερα δημόσιων προσώπων στα ΜΚΔ και όχι μόνο στο Twitter. Τα στοιχεία που θα εξήγαγε μία τέτοια έρευνα, θα ήταν

δυνατό να εξοικειώσουν το ευρύτερο κοινό με την Επιστήμη των Δεδομένων και τις χρήσεις που μπορεί αυτή να έχει για την Κοινωνία, την Οικονομία, την Επικοινωνία και την Πολιτική.

Βιβλιογραφία

Ελληνόγλωσση

Δεμέτης, Χ. (2019, 7 Ιουλίου). Εθνικές εκλογές 2019: Μπροστά η ΝΔ με 38,9%, στο 31,5% ο ΣΥΡΙΖΑ. *news247.* Ανάκτηση από: https://www.news247.gr/ekloges/ethnikes-ekloges-2019-mprosta-i-nd-me-39-8-sto-teliko-apotelesma.7471185.html Τελευταία πρόσβαση: 31/5/2023.

Τζανάκης, Σ. (2022, 4 Ιουλίου). Το πικρό καλοκαίρι του 2015: Όταν η Ελλάδα βρέθηκε να χορεύει στο χείλος του γκρεμού. *Πρώτο Θέμα*. Ανάκτηση από: https://www.protothema.gr/greece/article/1261180/to-pikro-kalokairi-tou-2015-otan-i-ellada-vrethike-horeuodas-sto-heilos-tou-gremou/ Τελευταία πρόσβαση: 31/5/2023.

Τσέκερης, Χ., Δεμερτζής, Ν. κ.ά. (2020). *Το Διαδίκτυο στην Ελλάδα: Η έρευνα του ΕΚΚΕ για το World Internet Project.* Αθήνα: διαΝΕΟσις. Ανάκτηση από: https://www.dianeosis.org/wp-content/uploads/2020/06/wip_greece_final_2-6_2020.pdf Τελευταία πρόσβαση: 31/5/2023.

Ξενόγλωσση

Aljabri, M., Zagrouba, R., Shaahid, A. *et al.* (2023, January 5). Machine learning-based social media bot detection: a comprehensive literature review. In *Social Network Analysis and Mining.* 13, 20 (2023). https://doi.org/10.1007/s13278-022-01020-5 Ανάκτηση από: https://link.springer.com/article/10.1007/s13278-022-01020-5#citeas Τελευταία πρόσβαση: 31/5/2023.

Amer, A.A. & Abdalla, H.I. (2020) A set theory based similarity measure for text clustering and classification. *Journal of Big Data* 7, 74 (2020). https://doi.org/10.1186/s40537-020-00344-3 Ανάκτηση από: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00344-3#citeas Τελευταία πρόσβαση: 31/5/2023.

Azam, N., Jahiruddin, Abulaish, M. and Haldar, N. Al-H. (2015). *Twitter Data Mining for Events Classification and Analysis,* 2nd International Conference on Soft Computing and Machine Intelligence (ISCMI'15), Hong Kong, IEEE CPS, Nov. 23-24, 2015. Ανάκτηση από: https://www.researchgate.net/publication/297453149_Twitter_Data_Mining_for_Events_Classification_and_Analysis Τελευταία πρόσβαση: 31/5/2023.

Backer, E. & Jain, A.K. (1981, Jan) A Clustering Performance Measure Based on Fuzzy Set Decomposition In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-3, no. 1, pp. 66-75, Jan. 1981, doi: 10.1109/TPAMI.1981.4767051. Ανάκτηση από: https://ieeexplore.ieee.org/abstract/document/4767051 Τελευταία πρόσβαση: 31/5/2023.

Bail, C., Topic Modeling. *Github.* Ανάκτηση από: https://cbail.github.io/SICSS_Topic_Modeling.html Τελευταία πρόσβαση: 31/5/2023.

Bansal, S., (2016, August 24). Beginners Guide to Topic Modelling in Python. *Analytics Vidhya*. Ανάκτηση από: https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/ Τελευταία πρόσβαση: 31/5/2023.

Bettenbuk, Z. (2022, October). How to Scrape Twitter Data Using Python Without Using Twitter's API. *scrapperapi.* Ανάκτηση από: https://www.scraperapi.com/blog/scrape-twitter-data/ Τελευταία πρόσβαση: 31/5/2023.

Blei, D., Andrew, Ng. & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research.* 3. 601-608. Ανάκτηση από:

https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf Τελευταία πρόσβαση: 31/5/2023.

Cai, Y., Li, X., & Li, J. (2023). Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. *Sensors*, *23*(5), 2455. MDPI AG. http://dx.doi.org/10.3390/s23052455 Ανάκτηση από: https://www.mdpi.com/1424-8220/23/5/2455 Τελευταία επίσκεψη: 31/5/2023.

Clifton, C. data mining. *Encyclopaedia Britannica*. Ανάκτηση από: https://www.britannica.com/technology/data-mining Τελευταία πρόσβαση: 31/5/2023.

cogitotech, (2021, September 27). Topic Modeling: Algorithms, Techniques, and Application. *Data Science Central.* Ανάκτηση από: https://www.datasciencecentral.com/topic-modeling-algorithms-techniques-and-application/ Τελευταία πρόσβαση: 31/5/2023.

Coursera. (2023, April 12). What is Data Analysis? (With Examples). Ανάκτηση από: https://www.coursera.org/articles/what-is-data-analysis-with-examples Τελευταία πρόσβαση: 31/5/2023.

Darwin, C. (1872). *The expression of the emotions in man and animals*. Oxford University Press, USA. Ανάκτηση από: https://pure.mpg.de/rest/items/item_2309885/component/file_2309884/content Τελευταία πρόσβαση: 31/5/2023.

Day, W. H. & Edelsbrunner, H. (1984) Efficient Algorithms for Agglomerative Hierarchical Clustering Methods. *Journal of Classification*, 1, 1-24. http://dx.doi.org/10.1007/BF01890115 Ανάκτηση από: https://link.springer.com/article/10.1007/bf01890115 Τελευταία πρόσβαση: 31/5/2023.

Dictionary.com. Tweet Definition & Meaning. Ανάκτηση από: https://www.dictionary.com/browse/tweet Τελευταία πρόσβαση: 31/5/2023.

Durairaj, J. (2020, Aug. 14). 2.5 Quintillion Bytes of Data Generated Everyday - Top Data Science Trends 2020. *SG Analytics.* Ανάκτηση από: https://us.sganalytics.com/blog/2-5-quintillion-bytes-of-data-generated-everyday-top-data-science-trends-2020/ Τελευταία πρόσβαση: 31/5/2023.

Ekman, P. (2003, December). *Emotions inside out. 130 Years after Darwin's "The Expression of the Emotions in Man and Animal".* Ann N Y Acad Sci. 2003 Dec;1000:1-6. doi: 10.1196/annals.1280.002. Ανάκτηση από: https://pubmed.ncbi.nlm.nih.gov/15045707/ Τελευταία πρόσβαση: 31/5/2023.

Ellis, C., When to use hierarchical clustering. *Crunching the Data.* Ανάκτηση από: https://crunchingthedata.com/when-to-use-hierarchical-clustering/ Τελευταία πρόσβαση: 31/5/2023.

Expert.ai Team. (2022, April 26). What Semantic Analysis Means to Natural Language Processing. *expert.ai.* Ανάκτηση από: https://www.expert.ai/blog/natural-language-process-semantic-analysis-definition/ Τελευταία πρόσβαση: 31/5/2023.

Faraz, M. (2022, May 16). Data Cleaning in Data Mining Simplified 101. *HEVO.* Ανάκτηση από: https://hevodata.com/learn/data-cleaning-in-data-mining-simplified-101/ Τελευταία πρόσβαση: 31/5/2023.

Fontanella, C. (2021, June 21). How to Get, Use & Benefit From Twitter's API. *HubSpot.* Ανάκτηση από: https://blog.hubspot.com/website/how-to-use-twitter-api Τελευταία πρόσβαση: 31/5/2023.

Goyal, C., (2021, June 23). Part 9: Step by Step Guide to Master NLP - Semantic Analysis. *Analytics Vidhya.* Ανάκτηση από:

https://www.analyticsvidhya.com/blog/2021/06/part-9-step-by-step-guide-to-master-nlp-semantic-analysis/ Τελευταία πρόσβαση: 31/5/2023.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A. & Mikolov, T.. (2018, May). *Learning Word Vectors for 157 Languages*. Proceedings of the International Conference on Language Resources and Evaluation, 2018. Ανάκτηση από: https://aclanthology.org/L18-1550/ Τελευταία πρόσβαση: 31/5/2023.

Great Learning Team. (2022, Oct 31). Understanding Latent Dirichlet Allocation (LDA). Ανάκτηση από: https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/ Τελευταία πρόσβαση: 31/5/2023.

Hanna, K. T. (2022, November). What is an information society? *TechTarget.* Ανάκτηση από: https://www.techtarget.com/whatis/definition/Information-Society Τελευταία πρόσβαση: 31/5/2023.

Harding, S. (2015, Jul 9). Greek Crisis: Twitter actively shapes fast-moving events. *The Guardian.* Ανάκτηση από: https://www.theguardian.com/world/2015/jul/09/greek-crisis-twitter-social-media Τελευταία πρόσβαση: 31/5/2023.

Humble Team. (2022, Νοέμβριος 9). Greeks & Social Media Research: An up-to-date Whitepaper for the Greek market - 4.000+ Responses (2022). *Humble.* Ανάκτηση από: https://humble.gr/blog/insights/humble-research/ Τελευταία πρόσβαση: 31/5/2023.

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022, January). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences of the United States of America*, 119(1), e2025334119. https://doi.org/10.1073/pnas.2025334119. Ανάκτηση από: 3/5/2023. Τελευταία πρόσβαση: 31/5/2023.

IBM. What is Data Science? Ανάκτηση από: https://www.ibm.com/topics/data-science Τελευταία πρόσβαση: 31/5/2023.

Ipshita. (2021, Jul. 16). Topic Modeling using LDA. *Analytics Vidhya*. Ανάκτηση από: https://medium.com/analytics-vidhya/topic-modelling-using-lda-aa11ec9bec13 Τελευταία πρόσβαση: 31/5/2023.

JustAnotherArchivist. (2020). snscrape. *GitHub.* Ανάκτηση από: https://github.com/JustAnotherArchivist/snscrape Τελευταία πρόσβαση: 31/5/2023.

Kemp, S. (2022, February 15). DIGITAL 2022: GREECE. *Datareportal.* Ανάκτηση από: https://datareportal.com/reports/digital-2022-greece Τελευταία πρόσβαση: 31/5/2023.

Kharde, V. & Sonawane, S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications.* 139. 5-15. 10.5120/ijca2016908625. Ανάκτηση από: https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf Τελευταία πρόσβαση: 31/5/2023.

Koch, K., (2020, March 26). A Friendly Introduction to Text Clustering. *Towards Data Science.*26 Mar. 2020. Ανάκτηση από: https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04 Τελευταία πρόσβαση: 31/5/2023.

LaRock, T. (2022, Oct 6). “Data is the New Oil”, But That Also Means it Can be Risky. *database.* Ανάκτηση από: https://www.dbta.com/Columns/Next-Gen-Data-Management/Data-is-the-New-Oil-But-That-Also-Means-it-Can-be-Risky-155275.aspx Τελευταία πρόσβαση: 31/5/2023.

Liu, B. (2020). Introduction. In Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. *Studies in Natural Language Processing*, pp. 1-17. Cambridge: Cambridge University Press. doi:10.1017/9781108639286.002. Ανάκτηση από:

https://www.cambridge.org/core/books/sentiment-analysis/introduction/563742A639EEE9F5AB3F29CB2387E41C Τελευταία πρόσβαση: 31/5/2023.

Liza, U., Semantic Analysis: What Is It, How It Works + Examples. *QuestionPro*. Ανάκτηση από: https://www.questionpro.com/blog/semantic-analysis/ Τελευταία πρόσβαση: 31/5/2023.

Maaten, L.V., & Hinton, G.E. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605. Ανάκτηση από: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf Τελευταία πρόσβαση: 31/5/2023.

Maryville University. Top 4 Data Analysis Techniques That Create Business Value. Ανάκτηση από: https://online.maryville.edu/blog/data-analysis-techniques/#what-is Τελευταία πρόσβαση: 31/5/2023.

Miroshnyk, O. (2023, Feb 2). What is Text Clustering? *oneai.* 2 Feb. 2023. Ανάκτηση από: https://www.oneai.com/learn/text-clustering Τελευταία πρόσβαση: 31/5/2023.

Muller, K., Schwarz, C. & Fujiwara, T. (2020, 30 October). How Twitter affected the 2016 presidential election. *VoxEu.* Ανάκτηση από: https://cepr.org/voxeu/columns/how-twitter-affected-2016-presidential-election Τελευταία πρόσβαση: 31/5/2023.

Nandwani, P. & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining.* 11.10.1007/s13278-021-00776-6. Ανάκτηση από: https://www.researchgate.net/publication/354196437_A_review_on_sentiment_analysis_and_emotion_detection_from_text Τελευταία πρόσβαση: 31/5/2023.

O' Neal, H. (2022, June 23). Semantics NLP. *MetaDialog.* Ανάκτηση από: https://www.metadialog.com/blog/semantic-analysis-in-nlp/ Τελευταία πρόσβαση: 31/5/2023.

Obar, J. & Wildman, S. (2015, July 22). Social Media Definition and the Governance Challenge: An Introduction to the Special Issue. *Telecommunications Policy*, 39(9), 745-750., Quello Center Working Paper No. 2647377, http://dx.doi.org/10.2139/ssrn.2647377 Ανάκτηση από: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2647377 Τελευταία πρόσβαση: 31/5/2023.

Orange. *Bag of Words.* Ανάκτηση από: https://orangedatamining.com/widget-catalog/text-mining/bagofwords-widget/ Τελευταία πρόσβαση: 3/5/2023.

Orange. *Distances.* Ανάκτηση από: https://orangedatamining.com/widget-catalog/unsupervised/distances/ Τελευταία πρόσβαση: 3/5/2023.

Orange. *Document Embedding*. Ανάκτηση από: https://orangedatamining.com/widget-catalog/text-mining/documentembedding/ Τελευταία πρόσβαση: 31/5/2023.

Orange. *Extract Keywords.* Ανάκτηση από: https://orangedatamining.com/widget-catalog/text-mining/keywords/ Τελευταία πρόσβαση: 31/5/2023.

Orange. *Hierarchical Clustering.* Ανάκτηση από: https://orangedatamining.com/widget-catalog/unsupervised/hierarchicalclustering/ Τελευταία πρόσβαση: 3/5/2023.

Orange. *Semantic Viewer.* Ανάκτηση από: https://orangedatamining.com/widget-catalog/text-mining/semanticviewer/ Τελευταία πρόσβαση: 31/5/2023.

Orange3 Text Mining. *Tweet Profiler.* Ανάκτηση από: https://orange3-text.readthedocs.io/en/latest/widgets/tweetprofiler.html Τελευταία πρόσβαση: 3/5/2023.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. In: Proceedings of the ACL-02 conference on empirical methods in natural language processing, Morristown, NJ, USA. Association for Computational Linguistics, pp 79–86. Ανάκτηση από; https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf Τελευταία πρόσβαση: 31/5/2023.

parsehub. (2020, March 2). Web Scraping vs Data Mining: What's the Difference? Ανάκτηση από: https://www.parsehub.com/blog/web-scraping-vs-data-mining/ Τελευταία πρόσβαση: 31/5/2023.

Peddireddi, Y., (2021, May 1). Topic Modelling in Natural Language Processing. *Analytics Vidhya.* Ανάκτηση από: https://www.analyticsvidhya.com/blog/2021/05/topic-modelling-in-natural-language-processing/ Τελευταία πρόσβαση: 31/5/2023.

Pretnar, A., (2021, Sep. 17). Semantic Analysis of Documents. *Orange.* Ανάκτηση από: https://orangedatamining.com/blog/2021/2021-09-17-semantic-analysis/ Τελευταία πρόσβαση: 31/5/2023.

Qader, W.A., Ameen, M.M. & Ahmed, B.I. (2019) *An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges,* International Engineering Conference (IEC), Erbil, Iraq, 2019, pp. 200-204, doi: 10.1109/IEC47844.2019.8950616. Ανάκτηση από: https://www.researchgate.net/publication/338511771_An_Overview_of_Bag_of_WordsImportance_Implementation_Applications_and_Challenges Τελευταία πρόσβαση: 3/5/2023.

Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing, *Communications of the ACM* 18 (11), pp. 613-620. Ανάκτηση από: https://dl.acm.org/doi/pdf/10.1145/361219.361220 Τελευταία πρόσβαση: 31/5/2023.

Sievert, C. & Shirley, K. (2014). *LDAvis: A method for visualizing and interpreting topics.* In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics DOI: 10.3115/v1/W14-3110 Ανάκτηση από: https://aclanthology.org/W14-3110/ Τελευταία πρόσβαση: 31/5/2023.

Simplilearn. (2023, May 4). Types of Data Mining Techniques. *simplilearn.* Ανάκτηση από: https://www.simplilearn.com/types-of-data-mining-techniques-article Τελευταία πρόσβαση: 31/5/2023.

Stagner R. (1940). The cross-out technique as a method in public opinion analysis. *Journal of Social Psychology.* 11(1):79–90. doi: 10.1080/00224545.1940.9918734. Ανάκτηση από: https://www.tandfonline.com/doi/abs/10.1080/00224545.1940.9918734?journalCode=vsoc20 Τελευταία πρόσβαση: 31/5/2023.

Stevens, E. (2023, May 10). The 7 Most Useful Data Analysis Methods and Techniques. *CAREERFOUNDRY.* Ανάκτηση από: https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/ Τελευταία πρόσβαση: 31/5/2023.

Suyal, H., Panwar, A. & Negi, A. S. (2014). Text Clustering Algorithms: A Review. *International Journal of Computer Applications* (0975 – 8887) Volume 96 – No.24. June 2014. Ανάκτηση από: https://research.ijcaonline.org/volume96/number24/pxc3897075.pdf Τελευταία πρόσβαση: 31/5/2023.

Tilbe, A. (2022, Jul. 11). 3 Critical Reasons Why Bag of Words is Implemented Before Sentiment Analysis. *Towards AI.* Ανάκτηση από: https://pub.towardsai.net/3-critical-reasons-why-bag-of-words-is-implemented-before-sentiment-analysis-24407e815491 Τελευταία πρόσβαση: 3/5/2023.

Tufts, C. (2018, January 31). Bag of Words. In *The Little Book of LDA.* Ανάκτηση από: https://miningthedetails.com/LDA_Inference_Book/word-representations.html#bag-of-words Τελευταία πρόσβαση: 31/5/2023.

tutorialspoint. *Natural Language Processing - Semantic Analysis.* Ανάκτηση από: https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_semantic_analysis.htm Τελευταία πρόσβαση: 31/5/2023.

Twitter API. Ανάκτηση από: https://developer.twitter.com/en/docs/twitter-api Τελευταία επίσκεψη: 31/5/2023.

Vidhya, K.A. & Geetha, T.V. (2017, Jan. 1). Rough Set Theory For Document Clustering: A Review. *Journal of Intellingent & Fuzzy Systems,* vol. 32, no. 3, pp. 2165-2185, 2017. Ανάκτηση από: https://www.researchgate.net/profile/Ka-Vidhya/publication/314125074_Rough_set_theory_for_document_clustering_A_review/links/60c83a7f299bf108abd975e1/Rough-set-theory-for-document-clustering-A-review.pdf Τελευταία πρόσβαση: 31/5/2023.

Violante, A., (2018, August 29). An Introduction to t-SNE with Python Example. *Towards Data Science.* 29 Aug. 2018. Ανάκτηση από: https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1 Τελευταία πρόσβαση: 31/5/2023.

Wang Z, Joo V, Tong C, Xin X, Chin HC (2014). *Anomaly detection through enhanced sentiment analysis on social media data*. In: 2014 IEEE 6th international conference on cloud computing technology and science. IEEE, pp 917–922. Ανάκτηση από: https://www.researchgate.net/publication/275220299_Anomaly_Detection_throug

h Enhanced Sentiment Analysis on Social Media Data Τελευταία πρόσβαση: 31/5/2023.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.2307/2282967 Ανάκτηση από: https://www.jstor.org/stable/2282967 Τελευταία πρόσβαση: 3/5/2023.

Wiebe, J. (1990). Recognizing subjective sentences: a computational investigation of narrative text. State University of New York, Buffalo. Ανάκτηση από: https://aclanthology.org/C90-2069.pdf Τελευταία πρόσβαση: 31/5/2023.

Xu, R. & Wunsch, D. (2005, June). Survey of Clustering Algorithms. *Neural Networks, IEEE Transactions* on. 16. 645-678. DOI: 10.1109/TNN.2005.845141. Ανάκτηση από: https://www.researchgate.net/publication/3303538_Survey_of_Clustering_Algorithms Τελευταία πρόσβαση: 31/5/2023.

Yilmaz, B. (2023, January 15). Sentiment Analysis Methods in 2023: Overview, Pros & Cons. *AIMultiple.* Ανάκτηση από: https://research.aimultiple.com/sentiment-analysis-methods/ Τελευταία πρόσβαση: 31/5/2023.

Yoo, I., & Hu, X. (2006). *A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE*. Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06*)*, 220-229. Ανάκτηση από: https://www.semanticscholar.org/paper/A-comprehensive-comparison-study-of-document-for-a-Yoo-Hu/08eafbff40575455e0c89a76245320026af76f14 Τελευταία πρόσβαση: 31/5/2023.

ZLP, J. (2019, Nov. 17). Distance Measures and Linkage Methods in Hierarchical Clustering. *Level Up Coding.* Ανάκτηση από: https://levelup.gitconnected.com/distance-measures-and-linkage-methods-in-hierarchical-clustering-8b7d488d7ebc Τελευταία πρόσβαση: 3/5/2023.

Παράρτημα

Ο κώδικας που χρησιμοποιήθηκε για την άντληση των tweets του @atsipras:

```
pip install snscrape
import snscrape.modules.twitter as sntwitter
import pandas as pd
```

```
    maxTweets = 10000000000000
    tweets_list1 = []
for i, tweets in enumerate (sntwitter.TwitterSearchScraper('from:atsipras').get_items()):
    if i > maxTweets:
        break
    tweets_list1.append([tweet.date, tweet.id, tweet.content, tweet.lang, tweet.source,
tweet.media, tweet.hashtags, tweet.replyCount, tweet.retweetCount, tweet.likeCount,
tweet.quoteCount, tweet.mentionedUsers, tweet.user.username])
tweets_df1 = pd.DataFrame (tweets_list1, columns=['Datetime', 'Tweet Id', 'Text', 'Lang',
'Source', 'Media', 'Hashtags', 'Reply Count', 'Like Count', 'Quote Count', 'Mentioned Users',
'Username'])
tweets_df1.to_csv('atsipras_tweets.csv', sep=',', encoding='utf-8', index=False)
```

Ο κώδικας που χρησιμοποιήθηκε για την άντληση των mentions προς τον @atsipras:

```
pip install snscrape
import snscrape.modules.twitter as sntwitter
import pandas as pd
maxTweets = 10000000000000
tweets_list1 = []
for i, tweet in enumerate (sntwitter.TwitterSearchScraper ('to:atsipras').get_items()):
    if i > maxTweets:
        break
    tweets_list1.append([tweet.date, tweet.id, tweet.content, tweet.lang,
tweet.hashtags, tweet.mentionedUsers, tweet.user.username])
tweets_df1 = pd.DataFrame (tweets_list1, columns=['Datetime', 'Tweet Id', 'Text', 'Lang',
'Hashtags', 'Mentioned Users', 'Username'])
tweets_df1.to_csv ('mentions_to_atsipras.csv', sep= ',', encoding='utf-8', index=False)
```