ΕΛΛΗΝΙΚΟ ΜΕΣΟΓΕΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

SOCIAL ANALYTICS: ΑΝΑΛΥΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΔΗΜΟΦΙΛΩΝ ΥΠΗΡΕΣΙΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Εισηγητής: Δέσποινα Αικατερίνη Δοκμετζή Α.Μ. 615
Ηλιάνα Στενάκη Α.Μ. 646

Επιβλέπων: Εμμανουήλ Περακάκης, Επίκουρος καθηγητής

HELLENIC MEDITERRANEAN UNIVERSITY

SCHOOL OF MANAGEMENT AND ECONOMICS SCIENCE

DEPARTMENT OF MANAGEMENT SCIENCE AND TECHNOLOGY

SOCIAL ANALYTICS FROM POPULAR SOCIAL NETWORKING SERVICES

DIPLOMA THESIS

Student: Despoina Aikaterini Dokmetzi 615
Iliana Stenaki 646
Supervisor: Emmanouil Perakakis, Assistant Professor

<u>Υπεύθυνη Δήλωση</u> *: Βεβαιώνουμε ότι είμαστε συγγραφείς αυτής της πτυχιακής εργασίας και ότι κάθε βοήθεια την οποία είχαμε για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην πτυχιακή εργασία. Επίσης έχουμε αναφέρει τις όποιες πηγές από τις οποίες κάναμε χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Τέλος βεβαιώνουμε ότι αυτή η πτυχιακή εργασία προετοιμάστηκε από εμάς προσωπικά ειδικά για τις απαιτήσεις του προγράμματος σπουδών του Τμήματος Διοικητικής Επιστήμης και Τεχνολογίας του ΕΛ.ΜΕ.ΠΑ.*

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία εστιάζει στην ανάλυση των στατιστικών(Analytics) που αντλούμε από τα Μέσα κοινωνικής δικτύωσης (Social Media). Όπως είναι γνωστό τα τελευταία χρόνια η χρήση των μέσων κοινωνικής δικτύωσης έχει αυξηθεί σημαντικά. Εκατομμύρια χρήστες από όλο τον κόσμο επικοινωνούν μεταξύ τους, αλληλεπιδρούν και χρησιμοποιούν τα μέσα είτε για προσωπική χρήση είτε για επαγγελματική. Αυτός είναι ο λόγος που αυξάνεται συνεχώς το ενδιαφέρον για την ανάλυση των στατιστικών. Για να γίνει λοιπόν μια ανάλυση των στατιστικών στα μέσα κοινωνική δικτύωσης απαιτούνται κάποιες αρχικές ενέργειες όπως είναι η έρευνα, η μεθοδολογία και κριτική. Εξίσου σημαντικό ρόλο παίζει ο καθαρισμός δεδομένων και η ανάλυση τους συναισθήματος. Τέλος για να έχουμε μια ολοκληρωμένη ανάλυση απαιτούνται συγκεκριμένες τεχνικές, επιστημονικά εργαλεία και εργαλεία παρακολούθησης καθώς και πλατφόρμες ειδήσεων και μέσων κοινωνικής δικτύωσης. Η εργασία κλείνει με την παρουσίαση μιας μικρής επιχείρησης στην Ελλάδα που δραστηριοποιείται μέσω των μέσων κοινωνικής δικτύωσης. Γίνεται μια ολοκληρωμένη αναφορά στα στατιστικά και την ανάλυση τους, καθώς και στα συμπεράσματα και τις προτάσεις βελτίωσης για το μέλλον.

Λέξεις – κλειδιά: ανάλυση στατιστικών, ψηφιακό μάρκετινγκ, κοινωνικά μέσα δικτύωσης

ABSTRACT

This paper focuses on the analysis of the statistics (Analytics) that we derive from Social Media. As we all know, the use of social media has increased significantly in recent years. Millions of users from all over the world communicate, interact and use the media either for personal or professional use. This is why there is a growing interest in the analysis of statistics. Therefore, to make an analysis of statistics in social media requires some initial actions such as research, methodology and criticism. Data cleaning and sentiment analysis play an equally important role. Finally, a comprehensive analysis requires specific techniques, scientific and monitoring tools, as well as news and social media platforms. The paper concludes with the presentation of a small business in Greece that operates through social media. A comprehensive report on the statistics and their analysis, as well as the conclusions and suggestions for improvement for the future is given.

Keywords: analytics, digital marketing, social media

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1. ΕΙΣΑΓΩΓΗ

Τα μέσα κοινωνικής δικτύωσης ορίζονται ως εφαρμογές Διαδικτύου που βασίζονται στον ιστό και σε κινητές συσκευές που επιτρέπουν τη δημιουργία, την πρόσβαση , την ανταλλαγή περιεχομένου και την αλληλεπίδραση μεταξύ των χρηστών και είναι από όλους και από παντού προσβάσιμο. Εκτός από τα μέσα κοινωνικής δικτύωσης (π.χ. Twitter και Facebook), για διευκόλυνση θα χρησιμοποιήσουμε επίσης τον όρο «κοινωνικά μέσα» που συμπεριλαμβάνουν τις απλές ροές διανομής (RSS), τα ιστολόγια, τα wikis και τις ειδήσεις, τα οποία είναι προσβάσιμα μέσω του ιστού και παράγουν μη δομημένο κείμενο. Τα μέσα κοινωνικής δικτύωσης είναι ιδιαίτερα σημαντικά για την έρευνα στην υπολογιστική κοινωνική επιστήμη που διερευνά ερωτήματα χρησιμοποιώντας ποσοτικές τεχνικές όπως υπολογιστικές στατιστικές, μηχανική μάθηση και πολυπλοκότητα και τα λεγόμενα μεγάλα δεδομένα για εξόρυξη δεδομένων και μοντελοποίηση προσομοίωσης (Cioffi- Revilla, 2010).

Αυτό έχει οδηγήσει σε πολυάριθμες υπηρεσίες δεδομένων, εργαλεία και πλατφόρμες ανάλυσης. Ωστόσο, αυτή η εύκολη διαθεσιμότητα δεδομένων μέσων κοινωνικής δικτύωσης για ακαδημαϊκή έρευνα μπορεί να αλλάξει σημαντικά λόγω εμπορικών πιέσεων. Επιπλέον, όπως συζητείται στο δεύτερο κεφάλαιο, τα εργαλεία που έχουν στη διάθεσή τους οι ερευνητές κάθε άλλο παρά ιδανικά. Είτε παρέχουν επιφανειακή πρόσβαση στα ακατέργαστα δεδομένα είτε (για μη επιφανειακή πρόσβαση) απαιτούν από τους ερευνητές να προγραμματίσουν αναλυτικά στοιχεία σε μια γλώσσα όπως η Java.

1.1. Βασικές τεχνικές

Ξεκινάμε με τους ορισμούς μερικών βασικών τεχνικών που σχετίζονται με την ανάλυση μη δομημένων δεδομένων κειμένου:

- **Η επεξεργασία φυσικής γλώσσας (NLP):** Είναι ένας τομέας της επιστήμης των υπολογιστών, της τεχνητής νοημοσύνης και της γλωσσολογίας που ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των

ανθρώπινων (φυσικών) γλωσσών. Συγκεκριμένα, είναι η διαδικασία κατά την οποία ένας υπολογιστής εξάγει σημαντικές πληροφορίες από την είσοδο φυσικής γλώσσας και παράγει έξοδο φυσικής γλώσσας.

- **Αναλυτικά στοιχεία ειδήσεων**: Είναι η μέτρηση διαφόρων ποιοτικών και ποσοτικών χαρακτηριστικών των κειμενικών (μη δομημένων δεδομένων) ειδήσεων. Μερικά από αυτά τα χαρακτηριστικά είναι το συναίσθημα και η συνάφεια.

- **Εξόρυξη απόψεων (συναισθημάτων,απόψεων):** Είναι ο τομέας έρευνας που επιχειρεί να δημιουργήσει αυτόματα συστήματα για τον προσδιορισμό της ανθρώπινης γνώμης από κείμενο γραμμένο σε φυσική γλώσσα.

- **Scraping(Απόξεση):** Είναι η συλλογή δεδομένων στο διαδίκτυο από μέσα κοινωνικής δικτύωσης και άλλους ιστότοπους με τη μορφή μη δομημένου κειμένου γνωστή και ως scraping ιστότοπου, συλλογή ιστού και εξαγωγή δεδομένων ιστού.

- **Ανάλυση συναισθήματος:** Η οποία αναφέρεται στην εφαρμογή της επεξεργασίας φυσικής γλώσσας, της υπολογιστικής γλωσσολογίας και της ανάλυσης κειμένου για τον εντοπισμό και την εξαγωγή υποκειμενικών πληροφοριών στο υλικό πηγής.

- **Αναλύσεις κειμένου**: Περιλαμβάνει ανάκτηση πληροφοριών (IR), λεξιλογική ανάλυση για τη μελέτη κατανομών συχνότητας λέξεων, αναγνώριση προτύπων, επισήμανση, σχολιασμό, εξαγωγή πληροφοριών, τεχνικές εξόρυξης δεδομένων, όπως ανάλυση συνδέσμων και συσχετισμών, οπτικοποίηση και προγνωστική ανάλυση.

1.2. Προκλήσεις της έρευνας

Η ανάλυση των μέσων κοινωνικής δικτύωσης παρέχουν μια πλούσια πηγή ακαδημαϊκής έρευνας προκλήσεων για κοινωνικούς επιστήμονες, επιστήμονες υπολογιστών και φορείς χρηματοδότησης. Οι προκλήσεις περιλαμβάνουν:

- **Scraping (Απόξεση)**: Παρόλου που τα δεδομένα κοινωνικών μέσων είναι προσβάσιμα μέσω API, λόγω της εμπορικής αξίας των δεδομένων, οι περισσότερες από τις κύριες πηγές όπως το Facebook και η Google καθιστούν όλο και πιο δύσκολο για τους ακαδημαϊκούς να αποκτήσουν ολοκληρωμένη πρόσβαση στα «ακατέργαστα» δεδομένα τους. Πολύ λίγες πηγές κοινωνικών δεδομένων παρέχουν οικονομικά προσιτές προσφορές δεδομένων στον ακαδημαϊκό κόσμο και τους ερευνητές. Οι ειδησεογραφικές υπηρεσίες όπως η Thomson Reuters και το Bloomberg χρεώνουν συνήθως ένα ασφάλιστρο για την πρόσβαση στα δεδομένα τους. Αντίθετα, το Twitter ανακοίνωσε πρόσφατα το πρόγραμμα Twitter Data Grants, όπου οι ερευνητές μπορούν να κάνουν αίτηση για πρόσβαση στα δημόσια tweets και ιστορικά δεδομένα του Twitter, προκειμένου να λάβουν πληροφορίες από το τεράστιο σύνολο δεδομένων του (το Twitter έχει περισσότερα από 500 εκατομμύρια tweets την ημέρα).

- **Η εκκαθάριση δεδομένων**: Ο καθαρισμός μη δομημένων δεδομένων κειμένου (π.χ. κανονικοποίηση κειμένου), ειδικά δεδομένων υψηλής συχνότητας που μεταδίδονται σε πραγματικό χρόνο, εξακολουθεί να παρουσιάζει πολλά προβλήματα και ερευνητικές προκλήσεις.

- **Ολιστικές πηγές δεδομένων**: Οι ερευνητές ολοένα και περισσότερο συγκεντρώνουν και συνδυάζουν νέες πηγές δεδομένων: δεδομένα μέσων κοινωνικής δικτύωσης, δεδομένα αγοράς και πελατών σε πραγματικό χρόνο και γεωγραφικά δεδομένα για ανάλυση.

- **Προστασία δεδομένων**: Μόλις δημιουργηθεί ο πόρος "μεγάλων δεδομένων" τα δεδομένα πρέπει να διασφαλιστούν, να επιλυθούν ζητήματα ιδιοκτησίας και IP (ένας μοναδικός αριθμός που χρησιμοποιείται από συσκευές σε ένα δίκτυο υπολογιστών για τη μεταξύ τους αναγνώριση και συνεννόηση) και να

παρέχονται διαφορετικά επίπεδα σε κάθε χρήστη ώστε να μην υπάρξει υποκλοπή πολύτιμων δεδομένων από τη βάση δεδομένων.

- **Η ανάλυση δεδομένων**: ανάλυση δεδομένων μέσων κοινωνικής δικτύωσης για εξόρυξη γνώμης (π.χ. ανάλυση συναισθήματος) εξακολουθεί να εγείρει πολλές προκλήσεις λόγω ξένων γλωσσών η λέξεων, ορθογραφικών λαθών και της φυσικής εξέλιξης της γλώσσας.

- **Πίνακες εργαλείων Analytics**: Αρκετές πλατφόρμες μέσων κοινωνικής δικτύωσης για την πρόσβαση σε αυτές απαιτείται από τους χρήστες να γράφουν API ώστε να εισέρχονται σε ροές δεδομένων ή μοντέλα αναλυτικών προγραμμάτων σε μια γλώσσα προγραμματισμού, όπως η Java. Αν και για τους επιστήμονες της πληροφορικής είναι εύκολες αυτές οι δεξιότητες , για κάποιους ερευνητές είναι πάνω από τις δυνατότητες τους. Απαιτούνται διεπαφές χωρίς προγραμματισμό για την παροχή αυτού που θα μπορούσε να χαρακτηριστεί ως "βαθιά" πρόσβαση σε "ακατέργαστα" δεδομένα π.χ. διαμόρφωση API, συγχώνευση ροών κοινωνικής δικτύωσης, συνδυασμός ολιστικών πηγών και ανάπτυξη αναλυτικών μοντέλων.

- **Οπτικοποίηση δεδομένων:** Είναι η οπτική αναπαράσταση δεδομένων όπου οι πληροφορίες έχουν αφαιρεθεί σε κάποια σχηματική μορφή ώστε να είναι αποτελεσματικότερες στην επικοινωνία των πληροφοριών μέσω των γραφικών μέσων. Έτσι η οπτικοποιήση αποκτά ολοένα και μεγαλύτερη σημασία.

1.3. Έρευνα και εφαρμογές μέσων κοινωνικής δικτύωσης

Τα δεδομένα των μέσων κοινωνικής δικτύωσης αποτελούν πλέον τη μεγαλύτερη, πλουσιότερη και πιο δυναμική βάση δεδομένων της ανθρώπινης συμπεριφοράς, προσφέροντας νέες ευκαιρίες για την κατανόηση των ατόμων, των ομάδων και της κοινωνίας. Οι καινοτόμοι επιστήμονες και οι επαγγελματίες του κλάδου βρίσκουν συνεχώς νέους τρόπους για την αυτόματη συλλογή, τον συνδυασμό και την ανάλυση πολλαπλών δεδομένων. Φυσικα όσο να μιλάμε για τις πρωτοποριακές εφαρμογές που χρησιμοποιεί αρκετό. Ενδεικτικά τρεις τομείς είναι:

- οι επιχειρήσεις

- οι βιοεπιστήμες
- οι κοινωνικές επιστήμες

Οι πρώτες **επιχειρήσεις** που υιοθέτησαν την ανάλυση των κοινωνικών μέσων ήταν εταιρείες του λιανικού εμπορίου και του χρηματοοικονομικού τομέα. Οι εταιρείες λιανικού εμπορίου χρησιμοποιούν τα μέσα κοινωνικής δικτύωσης για να αξιοποιήσουν την αναγνωρισιμότητα του εμπορικού τους σήματος, να διαφημίσουν και να βελτιώσουν τα προϊόντα ή τις υπηρεσίες τους ακόμη και να τα συγκρίνουν και άλλα προϊόντα/υπηρεσίες, καθώς και να ανιχνεύσουν τυχόν απάτες. Στα χρηματοοικονομικά, τα μέσα κοινωνικής δικτύωσης χρησιμοποιούνται για τη μέτρηση του κλίματος της αγοράς και τα δεδομένα των ειδήσεων για τις συναλλαγές.

Στις **βιοεπιστήμες**, τα μέσα κοινωνικής δικτύωσης χρησιμοποιούνται για τη συλλογή δεδομένων σε μεγάλες ομάδες για πρωτοβουλίες αλλαγής συμπεριφοράς και παρακολούθησης των επιπτώσεων, όπως π.χ. Αντιμετώπιση των κρουσμάτων COVID, Ένα παράδειγμα είναι οι βιολόγοι του Πανεπιστημίου Penn State (Salathé et al. 2012), οι οποίοι έχουν αναπτύξει καινοτόμα συστήματα και τεχνικές για την παρακολούθηση της εξάπλωσης μολυσματικών ασθενειών, με τη βοήθεια ειδησεογραφικών ιστοτόπων, ιστολογίων και μέσων κοινωνικής δικτύωσης.

Οι εφαρμογές των υπολογιστικών **κοινωνικών επιστημών** μέσω των μέσων κοινωνικής δικτύωσης όπως π.χ. το Twitter έχουν την δυνατότητα να παρακολουθούν τις αντιδράσεις που έχουν οι χρήστες τόσο σε κοινωνικά θέματα αλλα κυρίως σε πολιτικά όπου γίνονται μέσα από ανακοινώσεις, εκδηλώσεις, ομιλίες ή δημοσκοπήσεις σε ομάδες (όπου είναι δυσκολότερο να επικοινωνήσει κανείς) Για παράδειγμα, οι Lerman et al. (2008) χρησιμοποιούν την υπολογιστική γλωσσολογία για να προβλέψουν αυτόματα τι αντίκτυπο θα έχει αυτό στις ειδήσεις και θα ποια θα είναι η αντίληψη του κοινού για τους υποψηφίους. Ένα ακόμη παράδειγμα είναι οι Yessenov και Misailovic (2009) οι οποίοι χρησιμοποιώντας σχόλια και κριτικές από ταινίες μελετούν την επίδραση διαφόρων προσεγγίσεων στην εξαγωγή χαρακτηριστικών κειμένου και στην ακρίβεια τεσσάρων μεθόδων μηχανικής μάθησης Naive Bayes, Decision Trees, Maximum Entropy και K-Means clustering.

Επισκόπηση των μέσων κοινωνικής δικτύωσης

Για αυτήν την εργασία, ομαδοποιούμε τα εργαλεία κοινωνικής δικτύωσης σε:

- **Δεδομένα μέσων κοινωνικής δικτύωσης:** Τύποι δεδομένων κοινωνικών μέσων (π.χ. social media, wikis, ιστολόγια, ροές δεδομένων RSS κ.λπ.) και μορφές (π.χ. XML). Αυτό περιλαμβάνει σύνολα δεδομένων και περισσότερες σημαντικές ροές δεδομένων σε πραγματικό χρόνο, όπως τηλεπικοινωνιακά δεδομένα, οικονομικά δεδομένα κ.λ.π.

- **Προγραμματική πρόσβαση στα μέσα κοινωνικής δικτύωσης**: Δηλαδή υπηρεσίες δεδομένων και εργαλεία για την προμήθεια και τη συλλογή δεδομένων (κειμενικού) από μέσα κοινωνικής δικτύωσης, wiki, ροές RSS, ειδήσεις κ.λπ. Αυτά μπορούν να υποδιαιρεθούν σε:

    - ***Πηγές δεδομένων, υπηρεσίες και εργαλεία***: Όπου η πρόσβαση στα δεδομένα γίνεται με εργαλεία που προστατεύουν τα πρωτογενή δεδομένα ή παρέχουν απλές αναλύσεις. Παραδείγματα περιλαμβάνουν: Google Trends, SocialMention, SocialPointer και SocialSeek, τα οποία συγκεντρώνουν πληροφορίες από διάφορες ροές μέσων κοινωνικής δικτύωσης.

    - ***Πηγές δεδομένων* μέσω *API*** : Όπου ένα σύνολο δεδομένων και οι ροές τους είναι προσβάσιμα μέσω των API όπου βασίζονται σε HTTP και επιστρέφουν δεδομένα με ετικέτα όπως XML ή JSON κ.λπ. Παραδείγματα περιλαμβάνουν το Facebook, το Twitter , το Wikipedia κ.λ.π.

- **Εργαλεία καθαρισμού και αποθήκευσης κειμένου:** Το Google Refine και το Data Wrangler είναι παραδείγματα καθαρισμού δεδομένων.

- **Εργαλεία ανάλυσης κειμένου:** Υπάρχουν είτε μεμονωμένα είτε σε βιβλιοθήκες εργαλείων για την ανάλυση δεδομένων μέσων κοινωνικής δικτύωσης αφού πρώτα έχει γίνει scraping (απόξεση) και έχουν καθαριστεί. Είναι κυρίως εργαλεία επεξεργασίας, ανάλυσης και ταξινόμησης φυσικής γλώσσας, τα οποία παρουσιάζονται παρακάτω:

- ***Εργαλεία μετασχηματισμού:*** Απλά εργαλεία που έχουν τη δυνατότητα να μετατρέψουν τα δεδομένα εισαγωγής κειμένου σε πίνακες,γραφήματα (όπως πίτα, γραμμή, ράβδος κ.λπ.), χάρτες ακόμη και σε κίνηση (cartoon) όπως Google Fusion Tables, Zoho Reports, Tableau Public ή της IBM's Many Eyes.

- ***Εργαλεία ανάλυσης***: Προσεγμένα εργαλεία ανάλυσης για την ανάλυση δεδομένων κοινωνικής δικτύωσης, τον εντοπισμό συνδέσεων και τη δημιουργία δικτύων, όπως το πρόσθετο Excel NodeXL.

- **Πλατφόρμες μέσων κοινωνικής δικτύωσης**: Παρέχουν ολοκληρωμένα δεδομένα από τα μέσα κοινωνικής δικτύωσης και απο βιβλιοθήκες εργαλείων για αναλυτικά στοιχεία. Μερικά παραδείγματα που περιλαμβάνουν: Thomson Reuters, Radian 6 και Lexalytics.

    - ***Πλατφόρμες μέσων κοινωνικής δικτύωσης* :** Πλατφόρμες όπως το Facebook η στο Twitter που παρέχουν εξόρυξη δεδομένων και αναλύσεις σε άλλες πηγές μέσων κοινωνικής δικτύωσης.

    - ***Πλατφόρμες ειδήσεων:*** Πλατφόρμες όπως η Thomson Reuters που παρέχουν ροές ειδήσεων και εμπορικά στοιχεία και συναφή αναλυτικά στοιχεία.

2. ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΚΡΙΤΙΚΗ ΤΩΝ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Τα δύο βασικά εμπόδια στη χρήση των μέσων κοινωνικής δικτύωσης για ακαδημαϊκή έρευνα είναι πρώτον η πρόσβαση σε ολοκληρωμένα σύνολα δεδομένων και δεύτερον εργαλεία που επιτρέπουν βαθιά ανάλυση δεδομένων χωρίς την ανάγκη προγραμματισμού σε μια γλώσσα όπως η Java. Η πλειονότητα των πόρων των μέσων κοινωνικής δικτύωσης είναι εμπορικές και οι εταιρείες προσπαθούν να δημιουργήσουν έσοδα από τα δεδομένα τους. Όπως συζητήθηκε, είναι σημαντικό οι ερευνητές να έχουν πρόσβαση σε μεγάλα σύνολα δεδομένων του ανοιχτού κώδικα και εγκαταστάσεις για πειραματισμό. Διαφορετικά, η έρευνα στα μέσα κοινωνικής δικτύωσης θα έπρεπε να γίνεται αποκλειστικά από μεγάλες εταιρείες με τον τομέα πληροφορικής, από κυβερνητικούς φορείς και από ακαδημαϊκούς επιστήμονες και ερευνητές που έχουν πρόσβαση στα ιδιωτικά δεδομένα και μπορούν να παράγουν εργασίες που δεν θα κριθούν η θα αναπαραχθούν. Πρόσφατα, υπήρξε μια μέτρια ανταπόκριση, καθώς το Twitter και το Gnip εφαρμόζουν πιλοτικά ένα νέο πρόγραμμα για πρόσβαση σε δεδομένα.

2.1. Μεθοδολογία

Οι ερευνητικές απαιτήσεις μπορούν να ομαδοποιηθούν σε: δεδομένα, αναλυτικά στοιχεία και εγκαταστάσεις.

2.1.1. Δεδομένα

Για να γίνει ολοκληρωμένη μελέτη στα μέσα κοινωνικής δικτύωσης θα πρέπει οι ερευνητές να έχουν πλήρη πρόσβαση στα ιστορικά αρχεία σε πραγματικό χρόνο, ειδικά στις κύριες πηγές, για την διεξαγωγη Παγκοσμίων κορυφαίων ερευνών.

- **Μέσα κοινωνικής δικτύωσης**: Πρόσβαση σε ολοκληρωμένα σύνολα ιστορικών δεδομένων και πρόσβαση σε πηγές σε πραγματικό χρόνο μια μικρή χρονική καθυστέρηση περίπου 15 λεπτών, όπως συμβαίνει με τα οικονομικά δεδομένα Thomson Reuters.

- **Δεδομένα ειδήσεων:** Πρόσβαση σε δεδομένα ειδήσεων σε πραγματικό χρόνο και πρόσβαση σε ιστορικά δεδομένα μέσω άδειας χρήσης λογισμικού.

- **Δημόσια δεδομένα**: Πρόσβαση σε αποκομμένα και αρχειοθετημένα δημόσια δεδομένα. Διατίθεται μέσα ροών RSS, ιστολογίων ή ανοιχτών κυβερνητικών βάσεων δεδομένων.

- **Προγραμματιζόμενες διεπαφές**: Οι ερευνητές χρειάζονται επιπλέον πρόσβαση σε απλές διεπαφές προγραμματισμού εφαρμογών (API) ώστε να συλλέξουν η να αποκόψουν πληροφορίες που ενδέχεται να μην συλλέγονται αυτόματα.

2.1.2. Analytics

Τα δεδομένα των μέσων κοινωνικής δικτύωσης είναι συνήθως διαθέσιμα είτε μέσω απλών γενικών ρουτινών είτε από ερευνητές που προγραμματίζουν τα στοιχεία σε μια γλώσσα όπως το MATLAB, η Java ή η Python. Όπως αναφέρθηκε παραπάνω, οι ερευνητές απαιτούν:

- **Πίνακες εργαλείων ανάλυσης:** Απαιτούνται μη προγραμματιστικές διεπαφές για την πρόσβασης σε «ακατέργαστα» δεδομένα σε αυτό που θα μπορούσε να χαρακτηριστεί ως «βαθιά».

- **Ολιστική ανάλυση δεδομένων**: Απαιτούνται εργαλεία όπου θα λειτουργήσουν σε συνδυασμό με διεξαγωγή αναλυτικών στοιχείων σε πολλαπλά μέσα κοινωνικής δικτύωσης και άλλα σύνολα δεδομένων.

- **Οπτικοποίηση δεδομένων**: Οι ερευνητές απαιτούν εργαλεία οπτικοποίησης, με τα οποία οι πληροφορίες που έχουν αφαιρεθεί μπορούν να οπτικοποιηθούν σε κάποια σχηματική μορφή με στόχο την σαφή και αποτελεσματική επικοινωνία πληροφοριών μέσω γραφικών μέσων.

2.1.3. Εγκαταστάσεις

Ο τεράστιος όγκος των δεδομένων μέσων κοινωνικής δικτύωσης που παράγονται έχει την ικανότητα να στηρίζει τη δημιουργία εθνικών και διεθνών εγκαταστάσεων για την υποστήριξη της έρευνας στα μέσα κοινωνικής δικτύωσης

- **Αποθήκευση δεδομένων**: Λόγο ότι ο όγκος των δεδομένων των μέσων κοινωνικής δικτύωσης (τρέχοντες και προβλεπόμενοι) είναι υπερβολικός δεν μπορεί να διαχειριστεί από μεμονωμένα πανεπιστήμια άλλα από Ιδρύματα Εθνικών επιστημών τα οποία έχουν την ικανότητα να αποθηκεύουν κύριες πηγές δεδομένων όπως το Twitter άλλα και πηγές που συλλέγονται από μεμονωμένα έργα και αρχειοθετούνται για μελλοντική χρήση από άλλους ερευνητές.

- **Υπολογιστική διευκόλυνση**: απαιτούνται επίσης απομακρυσμένες υπολογιστικές εγκαταστάσεις για:

    - Προστασία της πρόσβασης στα αποθηκευμένα δεδομένα
    - Φιλοξενία των εργαλείων ανάλυσης και οπτικοποίησης
    - Παροχή υπολογιστικών πόρων όπως δίκτυα και GPU που απαιτούνται για την επεξεργασία των δεδομένων στην εγκατάσταση αντί για τη μετάδοσή τους σε ένα δίκτυο.

2.2. Κριτική

Για την έρευνα στα μέσα κοινωνικής δικτύωσης χρειάζεται αρκετή προσπάθεια. καθώς όπως αναφερθήκαμε και πριν τα μέσα κοινωνικής δικτύωσης η πλειονότητα τους είναι εμπορική, ακριβή και δύσκολη κυρίως για τους ακαδημαϊκούς να αποκτήσουν πλήρη πρόσβαση.

2.2.1. Δεδομένα

Γενικά, η πρόσβαση σε σημαντικές πηγές δεδομένων μέσων κοινωνικής δικτύωσης δεν είναι τόσο εύκολη καθώς υπάρχει περιορισμός και η πλήρης εμπορική πρόσβαση είναι ακριβή.

- **Δεδομένα αποσιωπημένα:** Πηγές δεδομένων όπως το Facebook η το Twitter έχουν απομονωμένες πληροφορίες πράγμα που καθιστά δύσκολο τον συνδυασμό με άλλες πηγές.

- **Ολιστικά δεδομένα** : Οι ερευνητές ενδιαφέρονται ολο και περισσότερο για την πρόσβαση, την αποθήκευση και το συνδυασμό νέων πηγών δεδομένων όπως:
    1. Δεδομένα μέσων κοινωνικής δικτύωσης.
    2. Δεδομένα χρηματοοικονομικής αγοράς και πελατών σε πραγματικό χρόνο.
    3. Γεωχωρικά δεδομένα για ανάλυση.

Όμως είναι εξαιρετικά δύσκολη διαδικασία ακόμα και για τα τμήματα της Πληροφορικής.

2.2.2. Analytics

Τα αναλυτικά εργαλεία που παρέχονται από προμηθευτές συνδέονται συχνά με ένα ενιαίο σύνολο δεδομένων. Μερικές φορές έχουν περιορισμένη ικανότητα ανάλυσης και η χρέωση δεδομένων τα καθιστά ακριβά στη χρήσης τους.

2.2.3. Εγκαταστάσεις

Εμπορικές πλατφόρμες όπως αυτές που παρέχονται από τη SAS και την Thomson Reuters παρόλου που έχουν αυξημένο αριθμό , οι χρεώσεις τους είναι σχεδόν απαγορευτικές για την ακαδημαϊκή έρευνα. Η επίλυση του προβλήματος αυτού είναι

είτε να παρέχονται συγκρίσιμες εγκαταστάσεις από εθνικά επιστημονικά ιδρύματα είτε οι πωλητές να πειστούν να εισάγουν την έννοια της «εκπαιδευτικής άδειας».

3. ΔΕΔΟΜΕΝΑ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Με βάση τα δεδομένα υπάρχει αυξημένος αριθμός εμπορικών υπηρεσιών που έχουν πρόσβαση σε μέσα κοινωνικής δικτύωσης (π.χ. Twitter, Facebook) όπως επίσης και σε υπηρεσίες ειδήσεων (π.χ. Thomson Reuters Machine Readable News). Οι ισοδύναμες κύριες ακαδημαϊκές υπηρεσίες είναι σπάνιες. Παρακάτω οι τύποι δεδομένων και οι μορφές που παράγονται από τις υπηρεσίες.

3.1. Τύποι δεδομένων

Παρόλου που αναφερόμαστε στα μέσα κοινωνικής δικτύωσης, υπάρχουν και άλλοι παράμετροι που παίζουν το ρόλο τους. Μια παράμετρος είναι αυτή τον ερευνητών όπου συνεχώς ανακαλύπτουν νέες και καινοτόμες πηγές δεδομένων όπου τις συλλέγουν και τις αναλύουν. Επομένως, όταν εξετάζουμε την ανάλυση δεδομένων κειμένου, θα πρέπει να λαμβάνουμε υπόψη πολλαπλές πηγές (π.χ. μέσα κοινωνικής δικτύωσης, ροές RSS, ιστολόγια και ειδήσεις) που συμπληρώνονται από αριθμητικά (οικονομικά) δεδομένα, δεδομένα τηλεπικοινωνιών, ενδεχομένως δεδομένα ομιλίας και βίντεο.

Η χρήση πολλαπλών πηγών δεδομένων αποτελεί σίγουρα το μέλλον της ανάλυσης.

Τα δεδομένα υποδιαιρούνται σε:

- **Σύνολα ιστορικών δεδομένων** : Συσσωρευμένα και αποθηκευμένα από το παρελθόν δεδομένα όπως κοινωνικά,οικονομικά,πολιτικά, χρηματοοικονομικα κτλ.

- **Τροφοδοσίες σε πραγματικό χρόνο:** Ζωντανές ροές δεδομένων από ροή κοινωνικών μέσων, υπηρεσίες ειδήσεων, χρηματοοικονομικές ανταλλαγές, υπηρεσίες τηλεπικοινωνιών, συσκευές GPS και ομιλία.

Επίσης σε:

- **Μη επεξεργασμένα δεδομένα**: Ανεπεξέργαστα δεδομένα υπολογιστή απευθείας από την πηγή που μπορεί να μην έχουν αναλυθεί είτε να περιέχουν σφάλματα.

- **Καθαρισμένα δεδομένα**: Εκκαθάριση, διόρθωση ή αφαίρεση λανθασμένων δεδομένων που προκαλούνται από διαφορές, σφάλματα πληκτρολόγησης, ελλείποντα bits κ.λπ.

- **Δεδομένα προστιθέμενης αξίας:** Δεδομένα όπου έχουν περάσει από την επεξεργασία καθαρισμού και ανάλυσης και έχουν αυξηθεί με "έξτρα" γνώση.

3.2. Μορφές δεδομένων κειμένου

Οι τέσσερις πιο κοινές μορφές που χρησιμοποιούνται για τη σήμανση κειμένου είναι: HTML, XML, JSON και CSV.

- **HTML**: Είναι η γλώσσα σήμανσης για ιστοσελίδες και άλλες πληροφορίες που μπορούν να προβληθούν σε ένα πρόγραμμα περιήγησης ιστού. Το HTML γράφεται υπό μορφή στοιχείων HTML, τα οποία περιλαμβάνουν ετικέτες που περικλείονται σε αγκύλες <html> εντός του περιεχομένου της ιστοσελίδας. Συνήθως λειτουργούν ανα ζεύγη <h1> ετικέτα έναρξης και </h1> κλείσιμο ετικέτας.

- **XML (Extensible Markup Language)**: Είναι η γλώσσα σήμανσης που περιέχει ένα σύνολο κανόνων για τη δόμηση δεδομένων κειμένου χρησιμοποιώντας <tag> και <\tag> για τον ορισμό στοιχείων.

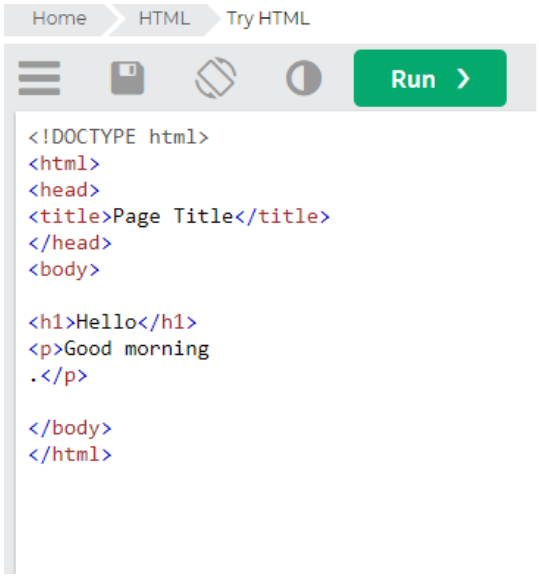
- **JSON (JavaScript Object Notation):** Είναι ένα ανοιχτό πρότυπο που βασίζεται σε κείμενο και μπορεί να διαβαστεί απο τον άνθρωπο. Έχει σχεδιαστεί για ανταλλαγή δεδομένων αναγνώσιμων και προέρχεται από JavaScript.

- **CSV**: Είναι ένα αρχείο τιμών διαχωρισμένων με κόμματα όπου περιέχει τις τιμές σε έναν πίνακα ως μια σειρά γραμμών κειμένου ASCII οργανωμένες έτσι

ώστε κάθε τιμή στήλης να διαχωρίζεται με κόμμα από την τιμή της επόμενης στήλης και κάθε σειρά ξεκινά μια νέα γραμμή.

Παρακάτω απεικονίζονται σε εικόνες παραδείγματα από τύπους δεδομένων HTML, XML και JSON:

Εικ.1



Εικ.2

Εικ.3

Αυτό το παράδειγμα είναι μια συμβολοσειρά JSON:

```
'{"name":"John", "age":30, "car":null}'
```

Ορίζει ένα αντικείμενο με 3 ιδιότητες:

- όνομα
- ηλικία
- αυτοκίνητο

Κάθε ακίνητο έχει μια αξία.

Τόσο η HTML όσο και η XML είναι οι λεγόμενες γλώσσες σήμανσης όπου ορίζουν ένα σύνολο απλών συντακτικών κανόνων για την κωδικοποίηση εγγράφων σε μορφή όπου μπορεί να την διαβάσει τόσο ένας άνθρωπος όσο και μια μηχανή.

Πολλές ροές χρησιμοποιούν σημειογραφία αντικειμένου JavaScript (JSON), την ελαφριά μορφή ανταλλαγής δεδομένων, που βασίζεται σε ένα υποσύνολο της γλώσσας προγραμματισμού JavaScript. Το JSON είναι μια ανεξάρτητη μορφή κειμένου από τη γλώσσα που χρησιμοποιεί ,ενώ χρησιμοποιεί συμβάσεις γνώστες στου προγραμματιστές σαν οικογένεια γλωσσών C (C, C++, C#) καθώς και Java, JavaScript, Python, Perl κ.π.λ.Οι βασικοί τύποι του JSON είναι: Number, String, Boolean, Array (μια ακολουθία τιμών, διαχωρισμένη με κόμμα όπου μπαίνει μέσα σε αγκύλες) και Object (μια μη ταξινομημένη συλλογή ζευγών κλειδιών:τιμών). Η μορφή JSON απεικονίζεται στην Εικ. 4 για ένα ερώτημα στο API του Twitter στη συμβολοσειρά "UCL", η οποία επιστρέφει δύο αποτελέσματα "κειμένου" από τον χρήστη Twitter "uclnews".

Εικ. 4

Παράδειγμα JSON

Οι τιμές διαχωρισμένες με κόμματα δεν είναι μια ενιαία, καλά καθορισμένη μορφή, αλλά αναφέρονται σε οποιοδήποτε αρχείο κειμένου που: (α) είναι απλό κείμενο χρησιμοποιώντας ένα σύνολο χαρακτήρων όπως ASCII, Unicode ή EBCDIC. (β) αποτελείται από εγγραφές κειμένου (π.χ. μία εγγραφή ανά γραμμή). (γ) με εγγραφές χωρισμένες σε πεδία διαχωρισμένα με οριοθέτες (π.χ. κόμμα, ερωτηματικό και καρτέλα). και (δ) όπου κάθε εγγραφή έχει την ίδια ακολουθία πεδίων.

4. ΠΑΡΟΧΟΙ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Οι πόροι δεδομένων μέσων κοινωνικής δικτύωσης υποδιαιρούνται γενικά σε αυτούς που παρέχουν:

- **Ελεύθερα διαθέσιμες βάσεις δεδομένων**: Όπου μπορούν να μεταφορτωθούν ελεύθερα και το σύνολο δεδομένων ηλεκτρονικού ταχυδρομείου της Enron.

- **Πρόσβαση σε δεδομένα μέσω εργαλείων**: Πηγές που παρέχουν ελεγχόμενη πρόσβαση στα δεδομένα των μέσων κοινωνικής δικτύωσης μέσω ειδικών εργαλείων, τόσο για τη διευκόλυνση της διερεύνηση όσο και για την εμπόδιση των χρηστών να "τραβάνε" όλα τα δεδομένα από το χώρο αποθήκευσης. Χαρακτηριστικό παράδειγμα το Google Trends.Αυτά υποδιαιρούνται περαιτέρω σε:

    - ***Δωρεάν πηγές*:** Δεδομένα ελεύθερα προσβάσιμα όμως με κάποιους περιορισμούς καθώς τα εργαλεία προστατεύουν ή περιορίζουν την πρόσβαση σε ακατέργαστα δεδομένα, όπως κάποια εργαλεία που παρέχει η Google.

    - ***Εμπορικές πηγές*:** Δεδομένα τα οποία έχουν μεταπωληθεί και χρεώνουν για την πρόσβαση δεδομένα των μέσων κοινωνικής δικτύωσης. Το DataSift παρέχουν εμπορική πρόσβαση στα δεδομένα του Twitter μέσω μιας συνεργασίας και η Thomson Reuters σε δεδομένα ειδήσεων.

- **Πρόσβαση σε δεδομένα μέσω API**: Δεδομένα κοινωνικών μέσων που παρέχουν προγραμματιζόμενη πρόσβαση βάσει HTTP στα δεδομένα μέσω API (π.χ. Twitter, Facebook και Wikipedia).

4.1. Βάσεις δεδομένων ανοιχτού κώδικα

Μια σημαντική ανοιχτή πηγή κοινωνικών μέσων είναι η Wikipedia, η οποία προσφέρει δωρεάν αντίγραφα όλου του διαθέσιμου περιεχομένου στους ενδιαφερόμενους χρήστες. Αυτές οι βάσεις δεδομένων μπορούν να χρησιμοποιηθούν για κατοπτρισμό, ερωτήματα βάσης δεδομένων και αναλύσεις μέσων κοινωνικής δικτύωσης.

Ένα άλλο παράδειγμα ελεύθερα διαθέσιμων δεδομένων για έρευνα είναι τα δεδομένα της Παγκόσμιας Τράπεζας. Η Τράπεζα Δεδομένων της Παγκόσμιας Τράπεζας παρέχει περισσότερες από 40 βάσεις δεδομένων, όπως Παγκόσμιες οικονομικές προοπτικές, Δείκτες Παγκόσμιας Ανάπτυξης όπως επίσης στατιστικά πληθυσμού, φύλου, υγείας, διατροφής κ.π.λ. Οι περισσότερες από τις βάσεις δεδομένων μπορούν να φιλτραριστούν ανά χώρα η περιοχή, σε σειρά,θέματα ή χρόνο. Επιπλέον, παρέχονται εργαλεία που επιτρέπουν την προσαρμογή και την εμφάνιση των αναφορών σε μορφή πίνακα, γραφήματος ή χάρτη.

4.2. Πρόσβαση στα δεδομένα μέσω εργαλείων

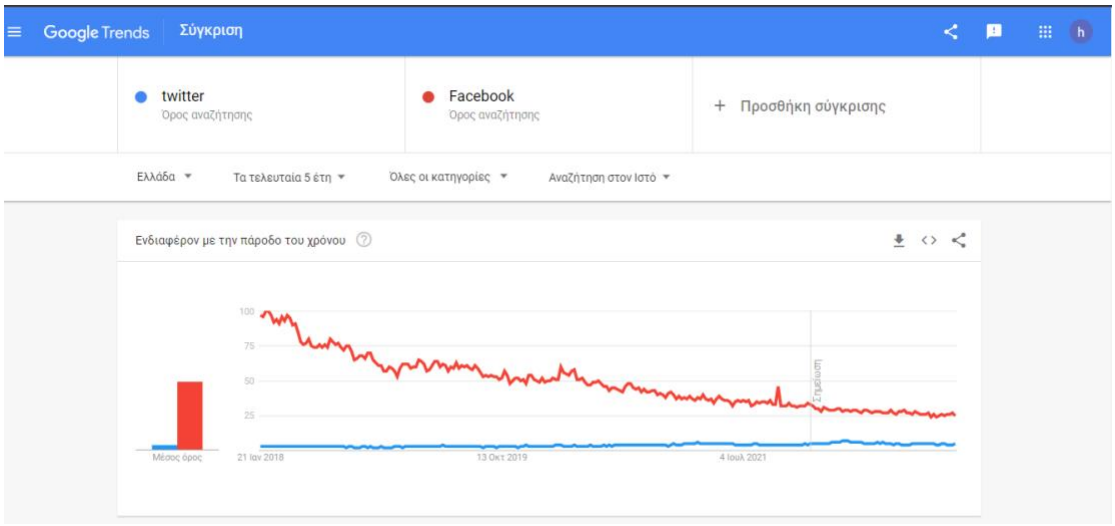
Όπως αναφέρθηκε, οι περισσότερες εμπορικές υπηρεσίες παρέχουν πρόσβαση σε δεδομένα κοινωνικών μέσων μέσω διαδικτυακών εργαλείων, τόσο για τον έλεγχο της πρόσβασης στα ακατέργαστα δεδομένα όσο και για τη δημιουργία εσόδων από τα δεδομένα ολοένα και περισσότερο.

4.2.1. Ελεύθερα προσβάσιμες πηγές

Η Google με εργαλεία όπως το Trends και το InSights είναι ένα καλό παράδειγμα αυτής της κατηγορίας. Η Google είναι η μεγαλύτερος "scraper" στον κόσμο, αλλά καταβάλλει κάθε προσπάθεια για να αποθαρρύνει το “scraping” δικών της σελίδων. Η στρατηγική της Google είναι να παρέχει ένα ευρύ φάσμα πακέτων, όπως το Google Analytics, αντί από την άποψη των ερευνητών τα πιο χρήσιμα προγραμματιζόμενα API που βασίζονται σε HTTP.

Η Εικ. 5 δείχνει πώς το Google Trends εμφανίζει έναν συγκεκριμένο όρο αναζήτησης, σε αυτήν την περίπτωση έχει ως παράδειγμα το Twitter και το Facebook τα πέντε τελευταία χρόνια στην Ελλάδα. Η σύγκριση μπορεί να γίνει έως και με πέντε θέματα κάθε φορά και επίσης να δείτε πόσο συχνά αναφέρονται σε αυτά τα θέματα και σε ποιές γεωγραφικές περιοχές έχει γίνει μεγαλύτερη αναζήτηση.

Εικ. 5



Google Trends

4.2.2. Εμπορικές πηγές

Πολλές εμπορικές υπηρεσίες διαγράφουν μέσα κοινωνικής δικτύωσης ώστε να να παρέχουν πρόσβαση επί πληρωμή μέσω απλών εργαλείων ανάλυσης. Επιπλέον, εταιρείες όπως το Twitter περιορίζουν την ελεύθερη πρόσβαση στα δεδομένα τους και χορηγούν άδεια για τα δεδομένα τους σε εμπορικούς μεταπωλητές δεδομένων, όπως το Gnip και το DataSif22t.

Το Gnip είναι η μεγαλύτερη εταιρεία συγκέντρωσης κοινωνικών δεδομένων στο ,κόσμο. μεγαλύτερος πάροχος κοινωνικών δεδομένων στον κόσμο. Αγοράστηκε απο το Twitter (2014) ώστε να διαθέτει δεδομένα και έκτοτε Η Gnip ήταν η πρώτη που

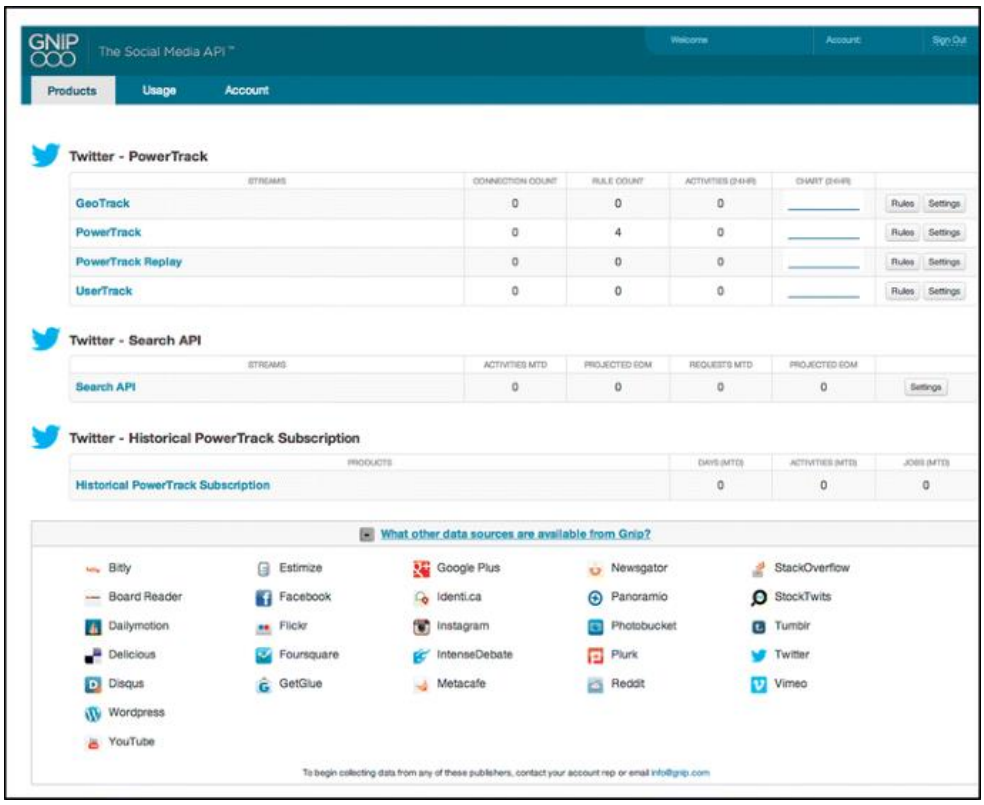
συνεργάστηκε με το Twitter για να διαθέσει τα δεδομένα και έκτοτε ήταν η πρώτη που συνεργάστηκε με τα Tumblr, Foursquare, WordPress και άλλες κορυφαίες πλατφόρμες κοινωνικής δικτύωσης. Η Gnip παρέχει δεδομένα κοινωνικής δικτύωσης σε πελάτες σε περισσότερες από 40 χώρες και οι πελάτες της, Gnip παρέχουν αναλύσεις κοινωνικών μέσων σε περισσότερο από το 95% του Fortune 500. Τα δεδομένα της Gnip μπορούν να παραδοθούν είτε ως Firehose για κάθε δραστηριότητα είτε μέσω PowerTrack, ένα ιδιόκτητο εργαλείο φιλτραρίσματος όπου επιτρέπει στους χρήστες να δημιουργούν ερωτήματα μόνο για τα δεδομένα που χρειάζονται και να λαμβάνουν μόνο τα δεδομένα για τα οποία ενδιαφέρονται αυτοί ή οι πελάτες τους. Οι κανόνες PowerTrack μπορούν να φιλτράρουν ροές δεδομένων με βάση τις λέξεις-κλειδιά, τα γεωγραφικά όρια, τις αντιστοιχίσεις φράσεων και ακόμη και τον τύπο περιεχομένου ή μέσων στη δραστηριότητα. Στη συνέχεια, η εταιρεία εμπλουτίζει τις ροές δεδομένων όπως το Profile Geo που προσθέτει σημαντικά στοιχεία όπως γεωγραφικά δεδομένα για το Twitter, την επέκταση URL και τον εντοπισμό γλώσσας για να βελτιώσει περαιτέρω την αξία των δεδομένων που παραδίδονται. Εκτός από την πρόσβαση σε δεδομένα σε πραγματικό χρόνο, η εταιρεία προσφέρει επίσης πρόσβαση στο Historical PowerTrack και στο Search API για το Twitter που δίνουν στους πελάτες τη δυνατότητα να κάνουν οποιοδήποτε Tweet. Η χρήση κανόνων PowerTrack για το φιλτράρισμα του Tweet διασφαλίζει ότι οι πελάτες λαμβάνουν όλα τα δεδομένα και μόνο τα δεδομένα που χρειάζονται για την εφαρμογή τους.

Το Gnip παρέχει πρόσβαση σε premium πηγές "Πηγές πλήρης πρόσβασης" όπου εκδότες έχουν συνάψει συμφωνία με το Gnip για να μεταπωλήσουν τα δεδομένα τους όπως επίσης και σε δωρεάν πηγές "Πηγές Διαχειριζόμενης δημόσιας πρόσβασης API" όπου παρέχουν πρόσβαση σε κανονικοποιημένα και ενοποιημένα δωρεάν δεδομένα από τα API τους. Βέβαια επί πληρωμή υπηρεσίες του Gnip για τους "Συλλέκτες Δεδομένων" μέσω του πίνακα ελέγχου του (βλ. Εικ. 6).

Αρχικά ο χρήστης βλέπει μόνο τις ροές στον πίνακα εργαλείων για τις οποίες πληρώθηκε βάσει συμφωνίας πώλησης. Για να επιλέξει μια ροή, ο χρήστης κάνει κλικ σε έναν εκδότη και στη συνέχεια επιλέγει μια συγκεκριμένη ροή από αυτόν τον εκδότη. Διαφορετικοί τύποι ροών εξυπηρετούν διαφορετικούς τύπους περιπτώσεων χρήσης και αντιστοιχούν σε διαφορετικούς τύπους ερωτημάτων και τερματικά σημεία API στο API πηγής του εκδότη. Μετά την επιλογή της ροής, ο χρήστης παίρνει βοήθεια από την

Gnip να το διαμορφώσει τις απαιτούμενες παραμέτρους προτού αρχίσει να συλλέγει δεδομένα. Αυτό περιλαμβάνει την προσθήκη τουλάχιστον ενός κανόνα. Στην περιοχή «Λήψη δεδομένων» – > «Σύνθετες ρυθμίσεις» μπορούμε επίσης να διαμορφώσουμε πόσο συχνά η ροή μας ρωτάει το API προέλευσης για δεδομένα (το «ποσοστό ερωτημάτων»). Επιλέγουμε μεταξύ της εγγενούς μορφής δεδομένων του εκδότη και της μορφής Activity Streams του Gnip (ροές XML για Enterprise Data Collector).

Εικ. 6



Gnip Dashboard, Publishers and Feeds

4.3. Πρόσβαση στη ροή δεδομένων μέσω API

Για τους ερευνητές, οι πιο χρήσιμες πηγές δεδομένων κοινωνικών μέσων είναι εκείνες που παρέχουν προγραμματιζόμενη πρόσβαση μέσω API, συνήθως χρησιμοποιώντας πρωτόκολλα που βασίζονται σε HTTP. Δεδομένης της σημασίας τους για τους

ακαδημαϊκούς, εδώ, εξετάζουμε μεμονωμένα wiki, μέσα κοινωνικής δικτύωσης, ροές RSS, ειδήσεις κ.λπ.

4.3.1. Wiki media

Η Wikipedia (και τα wikis γενικά) παρέχει στους ακαδημαϊκούς μεγάλα αποθετήρια ανοιχτού κώδικα περιεχομένου που δημιουργείται από χρήστες (πληθυσμός). Η Wikipedia παρέχει API που βασίζονται σε HTTP που επιτρέπουν προγραμματιζόμενη πρόσβαση και αναζήτηση (δηλαδή, scraping) που επιστρέφει δεδομένα σε διάφορες μορφές, συμπεριλαμβανομένης της XML. Στην πραγματικότητα, το API δεν είναι μοναδικό στη Wikipedia αλλά μέρος της εργαλειοθήκης ανοιχτού κώδικα του MediaWiki και ως εκ τούτου μπορεί να χρησιμοποιηθεί με οποιαδήποτε wiki που βασίζεται στο MediaWiki.

Το API που βασίζεται σε wiki HTTP λειτουργεί αποδεχόμενοι αιτήματα που περιέχουν ένα ή περισσότερα ορίσματα εισόδου και επιστρέφοντας συμβολοσειρές, συχνά σε μορφή XML, που μπορούν να αναλυθούν και να χρησιμοποιηθούν από τον αιτούντα πελάτη. Άλλες μορφές που υποστηρίζονται περιλαμβάνουν JSON, WDDX, YAML κ.π.λ. Το αίτημα HTTP πρέπει να περιέχει:

α) Την ζητούμενη «ενέργεια», όπως λειτουργία ερωτήματος, επεξεργασίας ή διαγραφής.

β) αίτημα ελέγχου ταυτότητας·

γ) τυχόν άλλες υποστηριζόμενες ενέργειες.

4.3.2. Μέσα κοινωνικής δικτύωσης

Όπως και με τη Wikipedia, τα δημοφιλή κοινωνικά δίκτυα, όπως το Facebook, το Twitter και το Foursquare, κάνουν ένα μέρος των δεδομένων τους προσβάσιμο μέσω API.

Παρόλο που πολλοί ιστότοποι μέσων κοινωνικής δικτύωσης παρέχουν API, δεν παρέχουν όλοι οι ιστότοποι (π.χ. LinkedIn) πρόσβαση API για τη συλλογή δεδομένων. Ενώ όλο και περισσότερα κοινωνικά δίκτυα μετατοπίζονται σε δημόσια διαθέσιμο περιεχόμενο, πολλά κορυφαία δίκτυα περιορίζουν την ελεύθερη πρόσβαση, ακόμη και σε ακαδημαϊκούς. Για παράδειγμα, η Foursquare ανακοίνωσε τον Δεκέμβριο του 2013 ότι δεν θα επιτρέπει πλέον τα ιδιωτικά check-in στο iOS 7 και πλέον έχει συνεργαστεί με τη Gnip για να παρέχει μια συνεχή ροή ανώνυμων δεδομένων check-in. Τα δεδομένα είναι διαθέσιμα σε δύο πακέτα:

- Το πλήρες επίπεδο πρόσβασης Firehose
- Μια φιλτραρισμένη έκδοση μέσω της υπηρεσίας PowerTrack της Gnip.

API δηλαδή τα οποία παρέχονται από το Twitter και το Facebook.

4.3.3. Twitter

Η προεπιλεγμένη ρύθμιση του λογαριασμού των Tweets δίνει τη δυνατότητα στους χρήστες να προστατεύουν τα Tweets τους και να τα κάνουν ορατά μόνο σε επιλεγμένους χρήστες στο Twitter. Ωστόσο, λιγότερο από το 10% όλων των λογαριασμών Twitter είναι ιδιωτικοί. Τα tweets από δημόσιους λογαριασμούς (συμπεριλαμβανομένων των απαντήσεων και των αναφορών) είναι διαθέσιμα σε μορφή JSON μέσω του API αναζήτησης του Twitter για ομαδικά αιτήματα προηγούμενων δεδομένων και του API ροής για δεδομένα σχεδόν σε πραγματικό χρόνο.

- **Αναζήτηση API** : Ερωτήσεις στο Twitter για πρόσφατα Tweets που περιέχουν συγκεκριμένες λέξεις-κλειδιά. Αποτελεί μέρος του Twitter REST API v1.1 και απαιτεί μια εξουσιοδοτημένη εφαρμογή όπως την oAuth πριν από την ανάκτηση οποιουδήποτε αποτελέσματος από την API.

- **Streaming API** : Μια ροή Tweet σε πραγματικό χρόνο, φιλτραρισμένη κατά αναγνωριστικό χρήστη, λέξη-κλειδί, γεωγραφική τοποθεσία ή τυχαία δειγματοληψία.

Κάποιος μπορεί να ανακτήσει πρόσφατα Tweets που περιέχουν συγκεκριμένες λέξεις-κλειδιά μέσω του Search API του Twitter (μέρος του REST API v1.1) και με την κλήση ροής API να ανακτήσεις χρονικά δεδομένα σε πραγματικό χρόνο.

Το Streaming API του Twitter επιτρέπει την πρόσβαση στα δεδομένα μέσω φιλτραρίσματος (κατά λέξεις-κλειδιά, αναγνωριστικά χρηστών ή τοποθεσία) ή με δειγματοληψία όλων των ενημερώσεων από επιλεγμένο αριθμό χρηστών. Το προεπιλεγμένο επίπεδο πρόσβασης «Spritzer» επιτρέπει τη δειγματοληψία περίπου του 1 % όλων των δημόσιων καταστάσεων, με την επιλογή ανάκτησης του 10 % όλων των καταστάσεων μέσω του επιπέδου πρόσβασης «Gardenhose».

Στα μέσα κοινωνικής δικτύωσης, τα API ροής ονομάζονται συχνά Firehose μια ροή που δημοσιεύει όλες τις δημόσιες δραστηριότητες που συμβαίνουν στο Twitter. Το Twitter δημιούργησε ένα πρόγραμμα το Twitter Data Grants, όπου οι ερευνητές μπορούν να κάνουν αίτηση για πρόσβαση στα δημόσια tweets και ιστορικά δεδομένα του Twitter, προκειμένου να λάβουν πληροφορίες από το τεράστιο σύνολο δεδομένων του (περισσότερα από 500 εκατομμύρια tweets ημερησίως). Όμως ερευνητικά ιδρύματα και ακαδημαϊκοί δεν θα λάβουν το επίπεδο πρόσβασης Firehose. Αντίθετα, θα λάβουν μόνο το σύνολο δεδομένων που απαιτούνται για το ερευνητικό τους έργο.

4.3.4. Facebook

Τα ζητήματα απορρήτου του Facebook είναι πιο περίπλοκα από αυτά του Twitter, πράγμα που σημαίνει ότι πολλά μηνύματα κατάστασης είναι πιο δύσκολο να ληφθούν από τα Tweets, απαιτώντας την κατάσταση «ανοιχτής εξουσιοδότησης» από τους χρήστες. Το Facebook αποθηκεύει όλα τα δεδομένα ως αντικείμενα και διαθέτει μια σειρά από API, που κυμαίνονται από το Graph και το Public Feed API έως το Keyword Insight API. Για να έχει ο χρήστης πρόσβαση στις ιδιότητες ενός αντικειμένου, θα πρέπει να γνωρίζει το ID για να πραγματοποιήσει την κλήση API. Η λεπτομερής μορφή ερωτήματος API φαίνεται στην Εικ. 7. Εδώ, το "QUERY" μπορεί να αντικατασταθεί

από οποιονδήποτε όρο αναζήτησης σαν “σελίδα”, “χρήστης”, ομάδα”, “τοποθεσία”. Τα αποτελέσματα αυτής της αναζήτησης θα περιέχουν ID για κάθε αντικείμενο. Όταν επιστρέφεται το ατομικό ID για ένα συγκεκριμένο αποτέλεσμα αναζήτησης, μπορούμε να χρησιμοποιήσουμε το https://graph.facebook.com/ID για να λάβουμε περισσότερες λεπτομέρειες σελίδας, όπως τον αριθμό των "μου αρέσει". Αυτού του είδους οι πληροφορίες ενδιαφέρουν τις εταιρείες όσον αφορά την αναγνωρισιμότητα της επωνυμίας και την παρακολούθηση του ανταγωνισμού.

Εικ. 7

```
GET graph.facebook.com
  /search?
    q={your-query}&
    [type={object-type}](#searchtypes)
```

Μορφή ερωτήματος αναζήτησης API Graph Facebook

Τα ερωτήματα αναζήτησης του Facebook Graph API απαιτούν ένα διακριτικό πρόσβασης που περιλαμβάνεται στο αίτημα. Η αναζήτηση σελίδων και τοποθεσιών απαιτεί «διακριτικό πρόσβασης εφαρμογής», ενώ η αναζήτηση για άλλους τύπους απαιτεί διακριτικό πρόσβασης χρήστη.

Η αντικατάσταση της «σελίδας» με τη «ανάρτηση» στην προαναφερθείσα διεύθυνση URL αναζήτησης θα επιστρέψει όλες τις δημόσιες καταστάσεις που περιέχουν αυτόν τον όρο αναζήτησης. Το Facebook επιστρέφει επίσης δεδομένα σε μορφή JSON και έτσι μπορούν να ανακτηθούν και να αποθηκευτούν χρησιμοποιώντας τις ίδιες μεθόδους που χρησιμοποιούνται με τα δεδομένα από το Twitter, αν και τα πεδία διαφέρουν ανάλογα με τον τύπο αναζήτησης, όπως φαίνεται στην Εικ. 8

Εικ. 8

```
{
  "id":"96184651725",
  "name":"Centrica",
  "picture":"http:\/\/profile.ak.fbcdn.net\/hprofile-ak-snc4\/71177_96184651725_7616434_s.jpg",
  "link":"http:\/\/www.facebook.com\/centricaplc",
  "likes":427,
  "category":"Energy\/utility",
  "website":"http:\/\/www.centrica.com",
  "username":"centricaplc",
  "about":"We're Centrica, meeting our customers' energy needs now...and in the future. As a leading integrated energy company, we're investing more now than ever in new sources of gas and power. http:\/\/www.centrica.com",
  "location":{
    "street":"Millstream, Maidenhead Road",
    "city":"Windsor",
    "country":"United Kingdom",
    "zip":"SL4 5GD ",
    "latitude":51.485694848812,
    "longitude":-0.63927860415725
  },
  "phone":"+44 (0)1753 494000",
  "checkins":228,
  "talking_about_count":5
}
```


Facebook Graph API Αποτελέσματα αναζήτησης για q='Centrica' and type='page'

4.3.5. Τροφοδοσίες RSS

Το RSS είναι μια μέθοδος όπου περιγράφει ειδήσεις/τάσεις ή άλλου είδους περιεχόμενα στο Παγκόσμιο ιστό και είναι διαθέσιμα για "τροφοδοσία" , δηλαδή διανομή από έναν εκδότη σε χρήστες του Διαδικτύου.

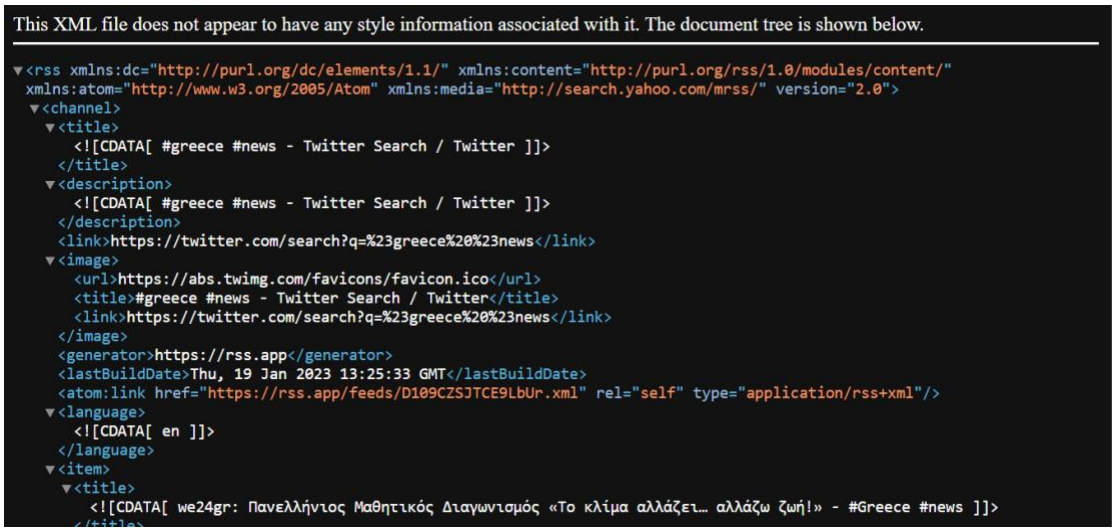
Με ένα πρόγραμμα περιήγησης Web δίνει τη δυνατότητα στους χρήστες να έχουν πρόσβαση σε περιεχόμενα μέσω ροών RSS. Αυτό είναι το πρότυπο διανομής για τη δημοσίευση τακτικών ενημερώσεων σε περιεχόμενο που βασίζεται στον ιστό και κατ επέκταση σε έναν τύπο αρχείου XML που βρίσκεται σε διακομιστή Διαδικτύου. Για τοποθεσίες Web, οι ροές RSS μπορούν να δημιουργηθούν χειροκίνητα ή αυτόματα.

Ένας αναγνώστης τροφοδοσίας RSS διαβάζει το αρχείο τροφοδοσίας RSS, βρίσκει ότι είναι νέο το μετατρέπει σε HTML και το εμφανίζει.

Χαρακτηριστικό παράδειγμα η Εικ. 9 όπου δημιουργήσαμε ένα κώδικα όπου δείχνει μέσα από το Twitter και τα hashtags #greece και #news τη συνολική ροή μέσω τροφοδοσίας RSS σε μετατροπή xml.

https://rss.app/feeds/D109CZSJTCE9LbUr.xml

Εικ. 9



Παράδειγμα ελέγχου ροής RSS στο Twitter.

4.3.6. Blogs, ομάδες ειδήσεων και υπηρεσίες συνομιλίας

- Blogs: Ο ορισμός αυτού του εργαλείου δεν έχει γίνει σαφής καθώς δεν υπάρχει συγκεκριμένη χρήση και ο κάθε χρήστης μπορεί να το χρησιμοποιήσει όπως εκείνος επιθυμεί χωρίς περιορισμούς. Παρόλα αυτά είναι ένας τρόπος επικοινωνίας και αλληλεπίδρασης μεταξύ χρηστών.
- Ομάδες ειδήσεων/συζήτησης: Είναι ένα σύστημα το οποίο επιτρέπει στους χρήστες να διαβάζουν και να δημοσιεύουν μηνύματα σε μια η περισσότερες ομάδες ειδήσεων/συζήτησης.

- Υπηρεσίες συνομιλίας: Επιτρέπουν στους χρήστες, τους πελάτες και τους επισκέπτες να εισέρχονται σε μια ιστοσελίδα και να συνομιλούν σε πραγματικό χρόνο με κάποιον εκπρόσωπο (π.χ. Live chat στην εφαρμογή της Cosmote).

Η απόξεση ιστολογίου είναι η διαδικασία σάρωσης ενός μεγάλου όγκου ιστολογίων, αναζήτησης και αντιγραφής περιεχομένου. Αυτή η διαδικασία διεξάγεται μέσω αυτοματοποιημένου λογισμικού. Η Εικ. 10 απεικονίζει παράδειγμα κώδικα για την απόξεση ιστολογίου. Αυτό περιλαμβάνει τη λήψη του πηγαίου κώδικα μιας τοποθεσίας Web μέσω της κλάσης URL της Java, η οποία μπορεί τελικά να αναλυθεί μέσω Κανονικών εκφράσεων για την καταγραφή του περιεχομένου-στόχου.

Εικ. 10

```
// Use Java's URL, InputStream and DataInputStream classes to read in the content of the supplied URL.
URL url;
InputStream inputStream = null;
DataInputStream dataInputStream;
String line;
scrapedContent = "";
try {
   // Attempt to open the URL (if valid):
   url = new URL("http://blog.wordpress.com/");
   inputStream = url.openStream();  // throws an IOException
   dataInputStream = new DataInputStream(new BufferedInputStream(inputStream));
   // Read the content line by line and store it in the scrapedContent variable:
   while ((line = dataInputStream.readLine()) != null) {
      scrapedContent += line + "\n";
   }
} catch (MalformedURLException exception) {
    exception.printStackTrace();
} catch (IOException exception) {
    exception.printStackTrace();
} finally {
   try {
      inputStream.close();
   } catch (IOException exception) {
   }
}
[...]
// Use regular expressions (RE) to parse the desired content from the scrapedContent. RE will attempt to delimit text between
some unique tags.
```

Παράδειγμα κώδικα για απόξεση ιστολογίου

4.3.7. Ροές ειδήσεων

Οι ροές ειδήσεων παραδίδονται σε ποικιλία μορφών κειμένου, συχνά ως αναγνώσιμα από μηχανή έγγραφα XML, αρχεία JSON ή CSV. Περιλαμβάνουν αριθμητικές τιμές, ετικέτες και άλλες ιδιότητες που τείνουν να αντιπροσωπεύουν υποκείμενες

ειδήσεις. Για σκοπούς δοκιμής, οι ιστορικές πληροφορίες συχνά παραδίδονται μέσω επίπεδων αρχείων, ενώ τα ζωντανά δεδομένα για παραγωγή υποβάλλονται σε επεξεργασία και παραδίδονται μέσω άμεσων ροών δεδομένων ή API. Η Εικ. 11 δείχνει ένα απόσπασμα των κλήσεων λογισμικού για την ανάκτηση φιλτραρισμένων άρθρων των NY Times.

Εικ. 11

```
nyTimesArticles = GET http://api.nytimes.com/svc/search/v1/article?query=(field:)keywords (facet:[value])(&params)&api-key=your-API-key
parse_JSON(nyTimesArticles)
```

Απόξεση άρθρων των New York Times

4.3.8. Γεωχωρικές τροφοδοσίες

Πολλά από τα «γεωχωρικά» δεδομένα μέσων κοινωνικής δικτύωσης προέρχονται από κινητές συσκευές που παράγουν δεδομένα ευαίσθητα σε τοποθεσία και χρόνο. Υπάρχουν τέσσερις τύποι ροών μέσων κοινωνικής δικτύωσης για κινητές συσκευές:

- **Ευαίσθητη τοποθεσία και ώρα**: Ανταλλαγή μηνυμάτων με συνάφεια για μια συγκεκριμένη τοποθεσία σε ένα συγκεκριμένο χρονικό σημείο (π.χ. AthensBook).

- **Μόνο ευαίσθητη τοποθεσία**: Ανταλλαγή μηνυμάτων με συνάφεια για μια συγκεκριμένη τοποθεσία, τα οποία έχουν επισημανθεί σε ένα συγκεκριμένο μέρος και διαβάζονται αργότερα από άλλους (π.χ. Yelp)

- **Μόνο ευαίσθητο στον χρόνο** : Μεταφορά παραδοσιακών εφαρμογών μέσων κοινωνικής δικτύωσης σε κινητές συσκευές για αύξηση της αμεσότητας (π.χ. ανάρτηση μηνυμάτων Twitter)

- **Ούτε η τοποθεσία ούτε ο χρόνος είναι ευαίσθητος**: Μεταφορά παραδοσιακών εφαρμογών κοινωνικών μέσων σε κινητές συσκευές (π.χ. παρακολούθηση βίντεο YouTube)

Στις κινητές συσκευές όπως τα smartphones πέρα από το κλασικό περιεχόμενα (φωτογραφίες, μηνύματα κ.λ.π.) έχει προστεθεί και ο γεωγραφικος προσδιορισμός ο οποίος ονομάζεται "geotagged" . Ο γεωγραφικός προσδιορισμός αποτελεί δεδομένα όπως τις συντεταγμένες γεωγραφικού πλάτους και μήκους, το υψόμετρο, την απόσταση κ.λ.π. Το GeoRSS είναι ένα πρότυπο για την κωδικοποίηση της γεωγραφικής θέσης σε μια διαδικτυακή ροή με δύο κύριες κωδικοποιήσεις:

- GML (Geography Markup Language)
- GeoRSS Simple

Παραδείγματα εργαλείων είναι το GeoNetwork Opensource μια δωρεάν ολοκληρωμένη εφαρμογή όπου καταγράφει γεωγραφικές πληροφορίες με γεωγραφικη αναφορά και το FeedBurner ένας παράχοσος διαδικτυακών τροφοδοσιών που μια απο τις δυνατότητες του είναι να παρέχει τροφοδοσίες με γεωγραφικές ετικέτες, φυσικά αν και οι προκαθορισμένες ρυθμίσεις το επιτρέπουν. Ενδεικτικά στην Εικ.12 παρουσιάζεται ο ψευδοκώδικας για την ανάλυσης μια γεωχωρικής τροφοδοσίας.

Εικ. 12

```
// Attempt to get the web site geotags by scraping the web page source code:
try getIcbmTags() // attempt to get ICBM tags, such as <meta name='ICBM' content="latitude, longitude" />
try getGeoStructureTags() // attempt to get tags such as <meta name="geo.position" content="coord1;coord2" />, <meta
name="geo.region" content="region">, <meta name="geo.placename" content="Place name">
// Attempt to get the web site's RSS geotags by scraping the RSS feeds, where the RSS source or each article can have their
own geotags.
// Attempt to get Resource Description Framework (RDF) tags, such as
<rdf:RDF><geo:Point><geo:lat>latitude</geo:lat><geo:long>longitude</geo:long><geo:alt>altitude</geo:alt></geo:Point></rdf:
RDF>
try getRdfRssTags()
// Attempt to get RSS article-specific geotags, e.g.: <rss
version="2.0"><item><title>title</title>[...]<icbm:latitude>latitude</icbm:latitude><icbm:longitude>longitude</icbm:longitude>[..
.]</item>
try getIcbmRssTags()
```

Ψευδο-κώδικας για την ανάλυση μιας γεωχωρικής τροφοδοσίας

5. ΚΑΘΑΡΙΣΜΟΣ ΚΕΙΜΕΝΟΥ, ΠΡΟΣΘΗΚΗ ΕΤΙΚΕΤΩΝ ΚΑΙ ΑΠΟΘΗΚΕΥΣΗ

Ο καθαρισμός δεδομένων είναι ένας σημαντικός τομέας στην ανάλυση των μέσων κοινωνικής δικτύωσης. Η διαδικασία αυτή περιλαμβάνει αφαίρεση τυπογραφικών σφαλμάτων, διόρθωση τιμών κτλ. Πιο συγκεκριμένα ένα κείμενο μπορεί να περιέχει ορθογραφικά λάθη, κενά, αλλαγές γραμμής, ξένες λέξει, συμβολισμούς και ειδικούς χαρακτήρες. Για να επιτευχθεί εξόρυξη κειμένου υψηλής ποιότητας είναι απαραίτητο να γίνει καθαρισμός δεδομένων με δυο απλά βήματα.

Πρώτο βήμα: ορθογραφικός έλεγχος, έλεγχος διπλότυπων, καθαρισμός ημερομηνίας και ώρας, διάταξη γραμμών και στηλών, διόρθωση αριθμών κ.π.λ.

Δεύτερο βήμα: Μόλις γίνει η παραπάνω επεξεργασία προχωράμε στον καθαρισμό δεδομένων για την αφαίρεση εσφαλμένων, ασυνεπών ή ελλιπών πληροφοριών.

Όμως πριν τον καθαρισμό δεδομένων μπορεί να υπάρξουν πιθανά προβλήματα δεδομένων (Narang 2009) όπως:

- **Ελλειπή δεδομένα** : Όταν μια πληροφορία υπάρχει αλλά δεν συμπεριλαμβάνονται ακατέργαστα δεδομένα. Μπορεί να υπάρξουν προβλήματα με: *α) αριθμητικά δεδομένα* όταν το "κενό" ή μια τιμή που λείπει αντικαθίσταται λανθασμένα από το "μηδέν" που στη συνέχεια λαμβάνεται (για παράδειγμα) ως η τρέχουσα τιμή. και *β) δεδομένα κειμένου* όταν μια λέξη που λείπει μπορεί να αλλάξει ολόκληρο το νόημα μιας πρότασης.

- **Λανθασμένα δεδομένα** : Όταν μια πληροφορία προσδιορίζεται εσφαλμένα (όπως λάθος λέξη σε δεδομένα κειμένου) ή ερμηνεύεται εσφαλμένα (όπως ένα σύστημα που υποθέτει ότι η τιμή του νομίσματος είναι σε € ενώ στην πραγματικότητα είναι σε $).

- **Ασυνεπή δεδομένα** : Όταν μια πληροφορία προσδιορίζεται με ασυνέπεια. Για παράδειγμα, με αριθμητικά δεδομένα, αυτό μπορεί να χρησιμοποιεί ένα μείγμα μορφών για ημερομηνίες: 2012/10/14 ή 14/10/2012. Για δεδομένα κειμένου,

μπορεί να είναι τόσο απλό όπως: η χρήση της ίδιας λέξης σε ένα μείγμα περιπτώσεων, η ανάμειξη αγγλικών και γαλλικών σε ένα μήνυμα κειμένου

5.1. Δεδομένα καθαρισμού

Μια παραδοσιακή προσέγγιση για τον καθαρισμό δεδομένων κειμένου είναι να «τραβήξετε» δεδομένα σε ένα υπολογιστικό φύλλο ή έναν πίνακα που μοιάζει με υπολογιστικό φύλλο και στη συνέχεια να διαμορφώσετε ξανά το κείμενο. Για παράδειγμα, *το Google Refine* είναι μια αυτόνομη εφαρμογή επιφάνειας εργασίας για καθαρισμό δεδομένων και μετατροπή σε διάφορες μορφές. Οι εκφράσεις μετασχηματισμού είναι γραμμένες σε αποκλειστική γλώσσα Google Refine Expression Language (GREL) ή JYTHON (μια υλοποίηση της γλώσσας προγραμματισμού Python γραμμένη σε Java). Η Εικ. 13 απεικονίζει τον καθαρισμό κειμένου.

Εικ. 13

```
cleanseText(blogPost) {
  // Remove any links from the blog post:
  blogPost['text'] = handleLinks(blogPost['text'])
  // Remove unwanted ads inserted by Google Ads etc. within the main text body:
  blogPost['text'] = removeAds(blogPost['text'])
  // Normalize contracted forms, e.g. isn't becomes is not (so that negation words are explicitly specified).
  blogPost['text'] = normalizeContractedForms(blogPost['text'])
  // Remove punctuation; different logic rules should be specified for each punctuation mark
  // You might not want to remove a hyphen surrounded by alphanumeric characters.
  // However you might want to remove a hyphen surrounded by at least one white space.
  blogPost['text'] = handlePunctuation(blogPost['text'])
  // Tokenize the text on white space, i.e. create an array of words from the original text.
  tokenizedText = tokenizeStatusOnWhiteSpace(blogPost['text'])
  // For each word, attempt to normalize it if it doesn't belong to the WordNet lexical database.
  for word in tokenizedStatus:
    if word not in WordNet dictionary:
      word = normalizeAcronym(word)
      // Further Natural Language Processing, POS Tagging
      ...
  return tokenizedText
}
```

Ψευδοκώδικας καθαρισμού κειμένου

5.2. Προσθήκη ετικετών σε μη δομημένα δεδομένα

Δεδομένου ότι τα περισσότερα από τα δεδομένα των μέσων κοινωνικής δικτύωσης παράγονται από ανθρώπους και ως εκ τούτου είναι αδόμητα (δηλαδή, δεν διαθέτουν προκαθορισμένη δομή ή μοντέλο δεδομένων), απαιτείται ένας αλγόριθμος για να τα μετατρέψει σε δομημένα δεδομένα για να αποκτήσει κανείς πληροφορίες. Επομένως,

τα μη δομημένα δεδομένα πρέπει να υποβληθούν σε προεπεξεργασία, να επισημανθούν και στη συνέχεια να αναλυθούν προκειμένου να ποσοτικοποιηθούν/αναλυθούν τα δεδομένα των μέσων κοινωνικής δικτύωσης.

Η προσθήκη επιπλέον πληροφοριών στα δεδομένα (δηλαδή η προσθήκη ετικετών στα δεδομένα) μπορεί να πραγματοποιηθεί χειροκίνητα ή μέσω μηχανών κανόνων, που αναζητούν μοτίβα ή ερμηνεύουν τα δεδομένα χρησιμοποιώντας τεχνικές όπως η εξόρυξη δεδομένων και η ανάλυση κειμένου. Οι αλγόριθμοι εκμεταλλεύονται τη γλωσσική, ακουστική και οπτική δομή που είναι εγγενής σε όλες τις μορφές ανθρώπινης επικοινωνίας. Η προσθήκη ετικετών στα μη δομημένα δεδομένα συνήθως περιλαμβάνει την προσθήκη ετικετών στα δεδομένα με ετικέτες μεταδεδομένων ή τμήματος του λόγου (POS). Σαφώς, η αδόμητη φύση των δεδομένων των μέσων κοινωνικής δικτύωσης οδηγεί σε ασάφεια και παρατυπία όταν υποβάλλονται σε επεξεργασία από ένα μηχάνημα με αυτόματο τρόπο.

Η χρήση ενός ενιαίου συνόλου δεδομένων μπορεί να προσφέρει μερικές ενδιαφέρουσες πληροφορίες. Ωστόσο, ο συνδυασμός περισσότερων συνόλων δεδομένων και η επεξεργασία των μη δομημένων δεδομένων μπορεί να οδηγήσει σε πιο πολύτιμες πληροφορίες, επιτρέποντάς μας να απαντήσουμε σε ερωτήσεις που ήταν αδύνατες εκ των προτέρων.

5.3. Αποθήκευση δεδομένων

Όπως συζητήθηκε, η φύση των δεδομένων των μέσων κοινωνικής δικτύωσης έχει μεγάλη επιρροή στη σχεδίαση της βάσης δεδομένων και πιθανώς του υποστηρικτικού υλικού. Θα ήταν επίσης πολύ σημαντικό να σημειωθεί ότι κάθε πλατφόρμα κοινωνικής δικτύωσης έχει πολύ συγκεκριμένους (και στενούς) κανόνες σχετικά με τον τρόπο αποθήκευσης και χρήσης των αντίστοιχων δεδομένων της. Αυτά μπορούν να βρεθούν στους Όρους Παροχής Υπηρεσιών για κάθε πλατφόρμα.

Για πληρότητα, οι βάσεις δεδομένων περιλαμβάνουν:

- **Επίπεδο αρχείο** — ένα επίπεδο αρχείο είναι μια δισδιάστατη βάση δεδομένων (κάπως σαν υπολογιστικό φύλλο) που περιέχει εγγραφές που δεν έχουν δομημένη αλληλεπίδραση και μπορούν να αναζητηθούν διαδοχικά.

- **Σχεσιακή βάση δεδομένων** —μια βάση δεδομένων που οργανώνεται ως ένα σύνολο επίσημα περιγραφόμενων πινάκων για την αναγνώριση των σχέσεων μεταξύ αποθηκευμένων στοιχείων πληροφοριών, επιτρέποντας πιο σύνθετες σχέσεις μεταξύ των στοιχείων δεδομένων. Παραδείγματα είναι βάσεις δεδομένων SQL που βασίζονται σε γραμμές και kdb + που βασίζονται σε στήλες που χρησιμοποιούνται στη χρηματοδότηση.

- **Βάσεις δεδομένων noSQL** — μια κατηγορία συστημάτων διαχείρισης βάσεων δεδομένων (DBMS) που προσδιορίζεται από τη μη τήρηση του ευρέως χρησιμοποιούμενου μοντέλου συστήματος διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS). Οι βάσεις δεδομένων noSQL/newSQL χαρακτηρίζονται ως: μη σχεσιακές, κατανεμημένες, ανοιχτού κώδικα και οριζόντια κλιμακούμενες.

5.3.1. Βάσεις και εργαλεία Apache (noSQL)

Η ανάπτυξη πολύ μεγάλων τοποθεσιών Web, όπως το Facebook και η Google, οδήγησε στην ανάπτυξη βάσεων δεδομένων noSQL ως τρόπου για να ξεπεραστούν οι περιορισμοί ταχύτητας που επιβάλλονται στις σχεσιακές βάσεις δεδομένων. Ένας βασικός οδηγός ήταν το MapReduce της Google, δηλαδή το πλαίσιο λογισμικού που επιτρέπει στους προγραμματιστές να γράφουν προγράμματα που επεξεργάζονται τεράστιες ποσότητες αδόμητων δεδομένων παράλληλα σε ένα κατανεμημένο σύμπλεγμα επεξεργαστών ή αυτόνομων υπολογιστών (Chandrasekar and Kowsalya 2011). Αναπτύχθηκε στη Google για την ευρετηρίαση ιστοσελίδων και αντικατέστησε τους αρχικούς αλγόριθμους ευρετηρίασης και ευρετικές μεθόδους το 2004. Το μοντέλο είναι εμπνευσμένο από τις συναρτήσεις «Χάρτης» και «Μείωση» που χρησιμοποιούνται συνήθως στον λειτουργικό προγραμματισμό. Το MapReduce (εννοιολογικά) λαμβάνει ως είσοδο μια λίστα εγγραφών και ο υπολογισμός «Χάρτης» τις χωρίζει στους διαφορετικούς υπολογιστές σε ένα σύμπλεγμα. Το αποτέλεσμα του υπολογισμού του χάρτη είναι μια λίστα ζευγών κλειδιών/τιμών. Ο αντίστοιχος υπολογισμός «Μείωση» λαμβάνει κάθε σύνολο τιμών που έχει το ίδιο κλειδί και τις συνδυάζει σε μια ενιαία τιμή. Ένα πρόγραμμα MapReduce αποτελείται από μια διαδικασία 'Map()' για φιλτράρισμα και ταξινόμηση και μια διαδικασία 'Reduce()' για μια συνοπτική λειτουργία (π.χ. μέτρηση και ομαδοποίηση).

Η Εικ. 14. παρέχει ένα κανονικό παράδειγμα εφαρμογής του MapReduce. Αυτό το παράδειγμα είναι μια διαδικασία μέτρησης των εμφανίσεων κάθε διαφορετικής λέξης σε ένα σύνολο εγγράφων (MapReduce 2011).

Εικ. 14

```
void map(String name, String document):
  // name: document name
  // document: document contents
  // Split the input amongst the various computers within the cluster.
  for each word w in document:
    EmitIntermediate(w, "1"); // Output key-value pairs as the map function processes the data in its input file.
void reduce(String word, Iterator partialCounts):
  // word: a word
  // partialCounts: a list of aggregated partial counts
  // Take each set of values with the same key and combines them into a single value.
  int sum = 0;
  for each pc in partialCounts:
    sum += ParseInt(pc);
  Emit(word, AsString(sum));
```

Το Κανονικό Παράδειγμα Εφαρμογής του MapReduce

5.3.2. Λογισμικό ανοιχτού κώδικα Apache

Η ερευνητική κοινότητα χρησιμοποιεί όλο και περισσότερο λογισμικό Apache για αναλύσεις μέσων κοινωνικής δικτύωσης. Στο πλαίσιο του Apache Software Foundation, τρία επίπεδα λογισμικού είναι σχετικά:

- **Βάσεις δεδομένων Cassandra**: Το Apache Cassandra είναι ένα κατανεμημένο DBMS ανοιχτού κώδικα (noSQL) που παρέχει ένα δομημένο χώρο αποθήκευσης «κλειδιού-τιμής». Τα καταστήματα βασικών τιμών επιτρέπουν σε μια εφαρμογή να αποθηκεύει τα δεδομένα της με τρόπο χωρίς σχήμα. Τα σχετικά προϊόντα βάσης δεδομένων noSQL περιλαμβάνουν: Apache Hive, Apache Pig και MongoDB, μια επεκτάσιμη και υψηλής απόδοσης βάση δεδομένων ανοιχτού κώδικα που έχει σχεδιαστεί για να χειρίζεται την αποθήκευση προσανατολισμένη στα έγγραφα. Δεδομένου ότι οι βάσεις δεδομένων noSQL είναι "χωρίς δομή", είναι απαραίτητο να έχουμε μια συνοδευτική βάση δεδομένων SQL για να διατηρεί και να χαρτογραφεί τη δομή των αντίστοιχων δεδομένων.

- **Η πλατφόρμα Hadoop**: Είναι ένα πλαίσιο προγραμματισμού βασισμένο σε Java που υποστηρίζει την επεξεργασία μεγάλων συνόλων δεδομένων σε ένα

κατανεμημένο υπολογιστικό περιβάλλον. Μια εφαρμογή αναλύεται σε πολλά μικρά μέρη (ονομάζονται επίσης θραύσματα ή μπλοκ) που μπορούν να εκτελεστούν σε συστήματα με χιλιάδες κόμβους που περιλαμβάνουν χιλιάδες terabyte αποθήκευσης.

- **Mahout**: Παρέχει υλοποιήσεις κατανεμημένων ή αλλιώς κλιμακούμενων αλγορίθμων αναλυτικών στοιχείων (μηχανικής εκμάθησης) που εκτελούνται στην πλατφόρμα Hadoop. Υποστηρίζει τέσσερις κατηγορίες αλγορίθμων: α) ομαδοποίηση (π.χ. K-Means, Fuzzy C-Means) που ομαδοποιεί κείμενο σε σχετικές ομάδες. β) ταξινόμηση (π.χ. Συμπληρωματικός ταξινομητής Naive Bayes) που χρησιμοποιεί εποπτευόμενη μάθηση για την ταξινόμηση κειμένου. γ) η συχνή εξόρυξη συνόλων στοιχείων παίρνει ένα σύνολο ομάδων αντικειμένων και προσδιορίζει ποια μεμονωμένα στοιχεία εμφανίζονται συνήθως μαζί. και δ) εξόρυξη προτάσεων (π.χ. συστάσεις που βασίζονται σε χρήστες και στοιχεία) που λαμβάνει υπόψη τη συμπεριφορά των χρηστών και από αυτήν προσπαθεί να βρει στοιχεία που μπορεί να αρέσουν στους χρήστες.

6. ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Η εξόρυξη γνώμης ή η ανάλυση συναισθημάτων είναι μια προσπάθεια αξιοποίησης των τεράστιων ποσοτήτων κειμένων και ειδήσεων που δημιουργούνται από χρήστες στο διαδίκτυο. Ένα από τα κύρια χαρακτηριστικά αυτού του περιεχομένου είναι η αταξία του κειμένου και η υψηλή ποικιλομορφία του. Εδώ, η επεξεργασία φυσικής γλώσσας, η υπολογιστική γλωσσολογία και η ανάλυση κειμένου αναπτύσσονται για τον εντοπισμό και την εξαγωγή υποκειμενικών πληροφοριών από το κείμενο πηγής. Ο γενικός στόχος είναι να προσδιοριστεί η στάση ενός συγγραφέα (ή ομιλητή) σε σχέση με κάποιο θέμα ή τη συνολική πολικότητα των συμφραζομένων ενός εγγράφου.

6.1. Τεχνικές Υπολογιστικής Επιστήμης

Η αυτοματοποιημένη ανάλυση συναισθήματος ψηφιακών κειμένων χρησιμοποιεί στοιχεία από τη μηχανική μάθηση, όπως λανθάνουσα σημασιολογική ανάλυση, μηχανές υποστήριξης διανυσμάτων, μοντέλο bag of words και σημασιολογικό προσανατολισμό. Με λίγα λόγια οι τεχνικές χρησιμοποιούν τρεις μεγάλους τομείς:

- **Υπολογιστική στατιστική**: Παρουσιάζει στατικές μεθόδους εντατικής υπολογιστικής επεξεργασίας, μέθοδοι επαναδειγματοληψίας, τοπικής παλινδρόμηση εκτίμηση πυκνότητας και ανάλυση κύριων συνιστωσών.

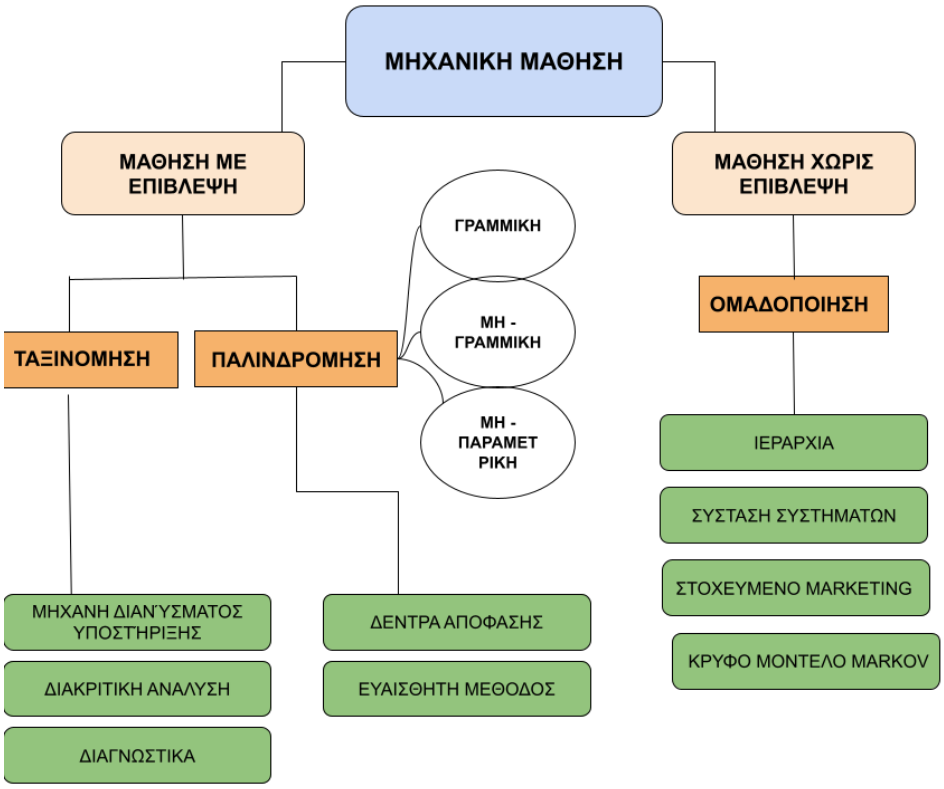
- **Μηχανική μάθηση** : Ουσιαστικά είναι ένα υποσύνολο τεχνητής νοημοσύνης. Δηλαδή ένα σύστημα αυτόνομης απόκτησης και ενσωμάτωση της γνώσης που αποκτάται μέσα από εμπειρία και αναλυτική παρατήρηση. Αυτά τα συστήματα υποδιαιρούνται περαιτέρω σε:

    - ***Μάθηση με επίβλεψη:*** Δέντρα παλινδρόμησης, ανάλυση συνάρτησης διακριτικής λειτουργίας, μηχανές υποστήριξης διανυσμάτων.

    - ***Μάθηση χωρίς επίβλεψη:*** Αυτοοργάνωση στους χάρτες (SOM).

Κύριος σκοπός της Μηχανικής Μάθησης είναι η επίλυση προβλήματος της ύπαρξης μεγάλων όγκων δεδομένων με πολλές μεταβλητές. Χρησιμοποιείται κατά κύριο λόγο σε τομείς όπως η αναγνώριση προτύπων (π.χ. εικόνες), οικονομικοί αλγόριθμοι (συναλλαγές με αλγόριθμους) , πρόβλεψη ενέργειας (φορτίο ή τιμή) και την βιολογία (ανίχνευση ασθενειών). Η Εικ. 15 απεικονίζει τους δύο τύπους εκμάθησης μηχανικής μάθησης και τις κατηγορίες αλγορίθμων τους.

Εικ. 15



Επισκόπηση μηχανικής μάθησης

- **Επιστήμη πολυπλοκότητας** : Είναι σύνθετα μοντέλα τα οποία δεν είναι εύκολα προβλέψιμα καθώς περιέχει θεωρητικές έννοιες, τεχνικά στοιχεία και πρότυπα. Συνδυάζουν δηλαδή την στατιστική φυσική, τη θεωρία της πληροφορίας και τη μη γραμμική δυναμική. Θα λέγαμε ότι είναι το βασίλειο των φυσικών και των μαθηματικών.

Οι παραπάνω τεχνικές αναπτύσσονται με δύο τρόπους:

- **Εξόρυξη δεδομένων:** Ανάπτυξη γνώσης που εξάγει κρυφά μοτίβα από τεράστιο όγκο δεδομένων χρησιμοποιώντας ενέργειες όπως: στατιστικούς διαχωριστές, διαφορικές εξισώσεις, μεθόδους επίλυσης και τεχνικές μηχανικής μάθησης τεχνητής νοημοσύνης.

- **Μοντελοποίηση προσομοίωσης**: Είναι η ανάλυση βασισμένη σε προσομοίωση που μπορεί να ελέγχει υποθέσεις. Η προσομοίωση χρησιμοποιείται για να προβλέψει την δυναμική των συστημάτων ώστε να μπορεί να ελεγχθεί άμεσα η εγκυρότητα της υπόθεσης.

6.2. Επεξεργασία ροής

Τέλος, θα αναφερθούμε σε κάτι πολύ σημαντικό και αυτό είναι η επεξεργασία ροής. Πλέον όλο και περισσότερες εφαρμογές που καταναλώνουν σε πραγματικό χρόνο δεδομένα από τα μέσα κοινωνικής δικτύωσης, τα χρηματοοικονομικά δίκτυα κ.π.λ. πρέπει να επεξεργάζονται μεγάλο όγκων χρονικών δεδομένων με μικρή χρονική καθυστέρηση. Οι εφαρμογές αυτές απαιτούν υποστήριξη για την online ανάλυση των ταχέως μεταβαλλόμενων ροών δεδομένων. Τα παραδοσιακα συστήματα βάσεων δεδομένων δεν έχουν αυτή τη δυνατότητα καθως δεν μπορούν να διαχειριστούν δεδομένα σε πραγματικό χρόνο. Αυτό οδήγησε στην ανάπτυξη συστημάτων διαχείρισης ροών δεδομένων όπου μπορούν να χειρίζονται online τις μεταβατικές ροές δεδομένων και να επεξεργάζονται συνεχείς ερωτήσεις σε αυτές τις ροές δεδομένων. Ένα παράδειγμα εμπορικού συστήματος είναι το Streaminsight της Microsoft.

6.3. Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος ή η εξόρυξη γνώμης εφαρμόζεται κυρίως σε οντότητες (π.χ. συγγραφέας, ομιλητής, εταιρεία ή οργανισμός κ.π.λ), στοχεύει στον προσδιορισμό της στάσης που μπορεί να εκφράζει ένας συγγραφέας κειμένου σε σχέση με το θέμα ή τη συνολική πολικότητα των συμφραζόμενων ενός εγγράφου.

Σύμφωνα με τον Dave et al. [69] το ιδανικό εργαλείο ανάλυσης συναισθήματος ή εξόρυξης απόψεων θα μπορεί να επεξεργάζεται ένα σύνολο αποτελεσμάτων αναζήτησης για ένα δεδομένο στοιχείο δημιουργώντας μια λίστα για αυτό με τα χαρακτηριστικά του (ποιότητα, χαρακτηριστικά κ.λ.π.) και έπειτα να συγκεντρώνει τις απόψεις

6.3.1. Ταξινόμηση συναισθημάτων

- **Συναισθηματικό πλαίσιο**: Για να έχει κάποιος απόλυτη άποψη θα πρέπει να γνωρίζει το "πλαίσιο" του κειμένου, το οποίο δεν είναι συνέχεια το ίδιο και ποικίλλει ως προς τις ροές τροφοδοσίας ειδικών κριτικών και μέσα από διάφορα forum όπου καλύπτουν ένα συγκεκριμένο θέμα.

- **Επίπεδο συναισθήματος** : Εν ολίγοις η ανάλυση ενός κειμένου μπορεί να διεξαχθεί σε επίπεδο πρόταση, εγγράφου ή χαρακτηριστικού.

- **Υποκειμενικότητα συναισθήματος:** Να λαμβάνει αποφάσεις εαν τα δεδομένα κειμένου εκφράζουν μια άποψη ή μια γνώμη χωρίς να εκφράζουν αν είναι θετικά η αρνητικά.

- **Προσανατολισμός συναισθήματος/πολικότητα**: Αποφασίζει εάν μια γνώμη σε ένα κείμενο είναι θετική, αρνητική ή ουδέτερη.

- **Δύναμη συναισθήματος**: Αποφασίζει τη "δύναμη" μια γνώση ενός κειμένου και κατά πόσο αυτό θα είναι: ισχυρό, αδύναμο ή ήπιο.

Ίσως, η πιο δύσκολη ανάλυση είναι ο προσδιορισμός του προσανατολισμού/πολικότητας και της δύναμης του συναισθήματος όπου χωρίζεται σε:

- *θετικό* (τέλειο, υπέροχο)
- *ουδέτερο*(μέτριο,καλό,αρκετό)

- *αρνητικό*(χάλια,καθόλου,κακό)

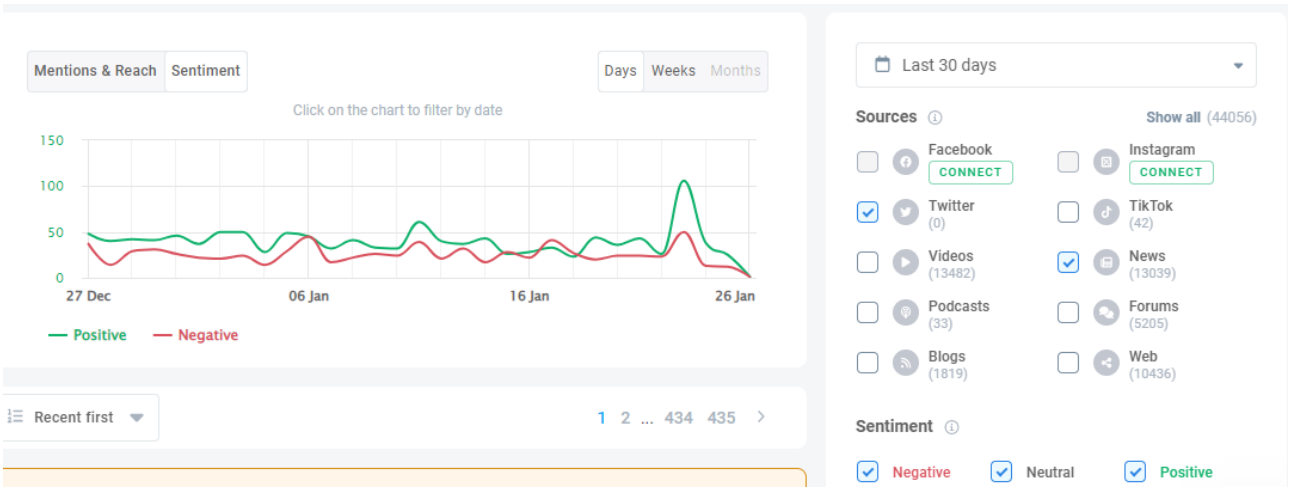
Μια δημοφιλής προσέγγιση είναι να εκχωρούνται βαθμολογίες προσανατολισμού/πολικότητας (+1, 0, −1) σε όλες τις λέξεις δηλαδή:

- *θετική* γνώμη (+1)
- *ουδέτερη* γνώμη (0)
- *αρνητική* γνώμη(−1).

Η συνολική βαθμολογία προσανατολισμού/πολικότητας του κειμένου είναι το άθροισμα των βαθμολογιών προσανατολισμού όλων των λέξεων «άποψη» που βρέθηκαν.

Παρακάτω σας παραθέτουμε τα στατιστικά του τελευταίου μήνα σχετικά με το Twitter και τις ειδήσεις του και κατά πόσο αντέδρασαν θετικά, ουδέτερα ή αρνητικά οι χρήστες.

Εικ. 16



6.3.2. Τεχνικές μάθησης

Για την ανάλυση συναισθήματος χρησιμοποιούνται δημοφιλείς υπολογιστικές στατιστικές και τεχνικές μάθησης. Οι τεχνικές περιλαμβάνουν:

- **Naïve Bayes (NB)** : Είναι ένας απλό μοντέλο ταξινόμησης που βασίζεται στην εφαρμογή του θεωρήματος Bayes με την "αφελή" υπόθεση ανεξαρτησίας , όταν δηλαδή τα χαρακτηριστικά είναι ανεξάρτητα το ένα από το άλλο σε κάθε τάξη.

- **Μέγιστη εντροπία (ΜΕ)** : Δηλώνει ότι η κατανομή πιθανοτήτων που αντιπροσωπεύει καλύτερα την τρέχουσα κατάσταση γνώσης είναι αυτή με τη μεγαλύτερη θεωρητική εντροπία πληροφοριών.

- **Μηχανές διανυσμάτων υποστήριξης (SVM)** : Είναι μια ομάδα αλγόριθμων όπου η χρήση τους αρχικά ήταν η ανάλυση ταξινόμησης, ανάλυση δεδομένων και η αναγνώριση μοτίβων και έπειτα χρησιμοποιήθηκαν και σε προβλήματα παλινδρόμησης.

- **Μοντέλο λογιστικής παλινδρόμησης (LR)** : Είναι ένας τύπος ανάλυσης παλινδρόμησης που χρησιμοποιείται για την πρόβλεψη του αποτελέσματος μιας μεταβλητής που μπορεί να λάβει περιορισμένο αριθμό κατηγοριών σε μια η περισσότερες προγνωστικές μεταβλητές. Είναι εύκολα αντιληπτό καθώς το αντικείμενο της είναι η Λογιστική Παλινδρόμησης η οποία βρίσκει εφαρμογή σε πλήθος επιστημονικών πεδίων. Π.χ. Μπορεί να προβλέψει αν ένας πελάτης ενδιαφέρεται η οχι για να αγοράσει ένα προϊόν ή μια υπηρεσία.

- **Λανθάνουσα σημασιολογική ανάλυση**: Η βασική τεχνική που χρησιμοποιεί είναι μια μαθηματική τεχνική που ονομάζεται αποσύνθεση μοναδιαίας αξίας (SVD) και χρησιμοποιείται για να εντοπίσει μοτίβα στις σχέσεις μεταξύ όρων - εγγράφων που περιέχονται σε μια αδόμητη συλλογή κειμένου.

Το μοντέλο bag of words είναι ένα παραδοσιακό μοντέλο που χάρη στην απλότητα του εφαρμόζεται για την ανάλυση συναισθήματος. Χρησιμοποιείται στην επεξεργασίας φυσικής γλώσσας και στην IR, όπου μια πρόταση ή ένα έγγραφο αναπαρίσταται ως μη ταξινομημένη συλλογή λέξεων, αδιαφορώντας για την γραμματική ή την σειρά των λέξεων.

6.3.3. Ταξινομητής Naïve Bayes (NBC)

Ο ταξινομητής Naive Bayes (Murphy 2006) είναι γενικής χρήσης, είναι εύκολος στην εφαρμογή και λειτουργεί καλα για μια σειρά εφαρμογών. Παρακάτω παρουσιάζεται ένα παράδειγμα ανάλυσης συναισθήματος:

- **Βήμα εκπαίδευσης** : Με την χρήση των δειγμάτων εκπαίδευσης, η μέθοδος εκτιμά τις παραμέτρους μιας κατανομής πιθανοτήτων, υποθέτοντας ότι τα χαρακτηριστικά είναι ανεξάρτητα υπό όρους λόγου της κλάσης.
- **Βήμα ανάλυσης/ελέγχου**: Για τα δείγματα τα οποία δεν είναι ορατά η μέθοδος υπολογίζει την εκ των υστέρων πιθανότητα του δείγματος να ανήκει σε κάθε κατηγορία και στη συνέχεια ταξινομεί το δείγμα δοκιμής.

  Η χρήση λοιπόν του ταξινομητή είναι να υπολογίζει την πιθανότητα εάν ένα κείμενο ανήκει σε κάθε μια από τις κατηγορίες που εξετάζουμε. Η κατηγορία με τις υψηλότερες πιθανότητες για το δεδομένο κείμενο κερδίζει.

Η Εικ. 17 περιέχει ένα παράδειγμα ταξινόμησης συναισθημάτων χρησιμοποιώντας έναν ταξινομητή Naïve Bayes στην Python.

Εικ. 17

```
for (tweet, label) in trainingSetMessage:
    // Normalize words, handle punctuation, tokenize on white space etc.
    preProcessMessage(tweet)
    for tweetWord in tweet:
        // Tokenize each Tweet, assign the label to each word and store it in the training set
        trainingSet += (tweetWord, label)
classifier = NaiveBayesClassifier.train(trainingSet)
predictedLabel = classifier.classify(getFeatures(preProcessMessage(trainingSet)))
```

7. ΕΡΓΑΛΕΙΑ ΑΝΑΛΥΣΗΣ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Στο προηγούμενο κεφάλαιο αναφερθήκαμε στη σημασία της εξόρυξη γνώμης. Σε αυτό το σημείο θα θέλαμε να αναφέρουμε και τα εργαλεία εξόρυξη γνώμης τα οποία είναι

γεμάτα με εμπορικούς παρόχους, όπου ολοένα και περισσότεροι από αυτούς στρέφονται προς την ανάλυση των συναισθημάτων σχετικά με τα σχόλια των πελατών που αφορούν ένα προϊόν ή μια υπηρεσία.

Υπάρχει μια τεράστια γκάμα εργαλείων που χρησιμοποιούνται για την ανάλυση κειμένου τα οποία κυμαίνονται από τα τα πιο απλά εργαλεία ανοιχτού κώδικα έως και σε εμπορικές εργαλειοθήκες πολλαπλών λειτουργιών, πλατφόρμες κ.π.λ.

7.1. Επιστημονικά εργαλεία Analytics των κοινωνικών δικτύων

Μερικά από τα καλύτερα εργαλεία ανάλυσης των μέσων κοινωνικής δικτύωσης προσφέρονται δωρεάν και έχουν το πλεονέκτημα να μπορουν να μετρήσουν μέσω των social media την αποτελεσματικότητα και την άνοδο τους με την πάροδο του χρόνου.Μια επιχείρηση μπορεί να εξάγει συμπεράσματα ανάλογα με τα δημογραφικά στοιχεία των πελατών ,τις αντιδράσεις , τις επισκέψεις στους ιστότοπους τους και έτσι να προσδιορίσουν σε συγκεκριμένο χρονικό διάστημα την σύγκριση που θα συμπίπτει π.χ με την ολοκλήρωση μιας καμπάνιας. Επίσης , υπάρχει η δυνατότητα να έχεις συγκριτικά στοιχεία για ημερήσια , εβδομαδιαία , μηνιαία ή και ετήσια ακόμα στοιχεία και να συλλέξεις πληροφορίες για τον στόχο σου.

Ένα από αυτά είναι και το Twitter Analytics.

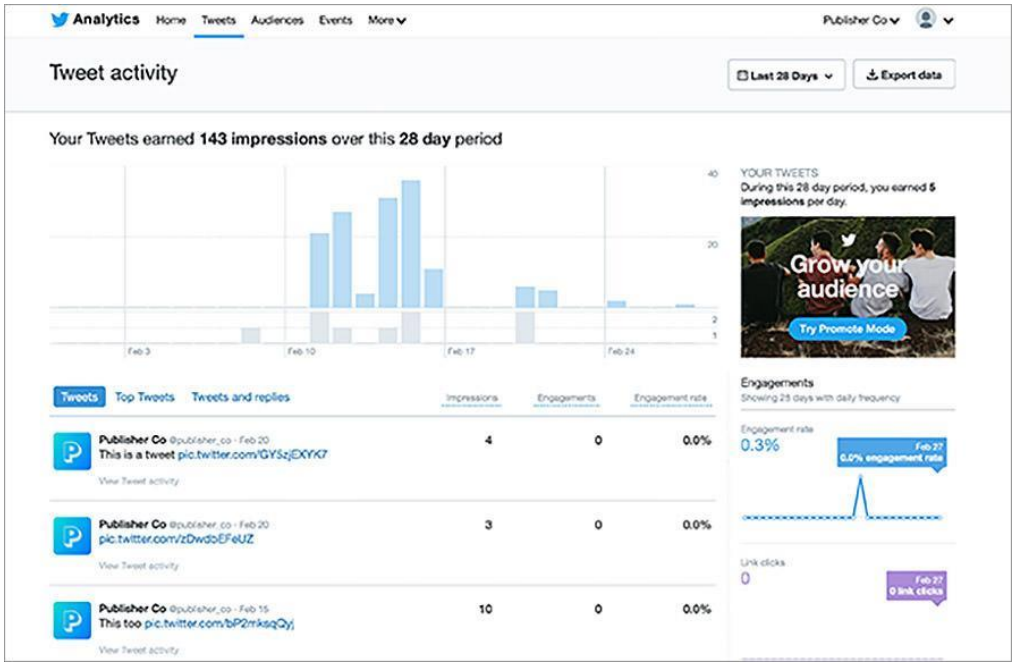
7.1.1. Τι είναι το Twitter Analytics

Είναι ένα εργαλείο του Twitter που μας επιτρέπει να έχουμε πρόσβαση στα στατιστικά του λογαριασμού μας και να γνωρίζουμε στοιχεία όπως αναφορές , εντυπώσεις, τις αλληλεπιδράσεις μας και την εξέλιξη τους. Μέσα από αυτό μπορούμε να δούμε

- Αριθμός Tweet.
- Εντυπώσεις από χρήστες στις δημοσιεύσεις μας.
- Επισκέψεις στο προφίλ μας από άλλους χρήστες, είτε είναι ακόλουθοι είτε όχι.
- Αναφέρει ότι μας έκαναν άλλοι άνθρωποι τον τελευταίο μήνα.

- Αριθμός ακολούθων

Εικ. 18



7.2. Επιστημονικά εργαλεία προγραμματισμού

Η βελτίωση που έχει γίνει στις στις δημοφιλείς βιβλιοθήκες και στα εργαλεία επιστημονικής ανάλυσης είναι αξιοσημείωτη καθώς παρέχει μεγάλη υποστήριξη για την προμήθεια, την αναζήτηση καθώς και την ανάλυση κειμένου. Χαρακτηριστικά παραδείγματα και η χρήση τους:

- R για στατιστικό προγραμματισμό
- MATLAB για αριθμητικό και επιστημονικό προγραμματισμό
- Mathematica για συμβολικό επιστημονικό προγραμματισμό (άλγεβρα υπολογιστών)

Αρχικά θα λέγαμε πως η Python είναι η κατάλληλη γλώσσα για τους αρχάριους λόγω αναγνωρισιμότητας και ευκολία στη χρήση. Διαθέτει αυτόματη διαχείριση μνήμης, χρησιμοποιείται πολλές φορές για την ανίχνευση γλώσσας, μπορεί να εξάγει τίτλους , περιεχόμα καθώς και αντιστοίχιση ερωτημάτων. Επιπλέον σε συνδυασμό με το scikit-learn μπορεί να εκπαιδευτεί ώστε να προσφέρει ανάλυση συναισθήματος.

Το MATLAB είναι ένα πρόγραμμα που βοηθά στην εκτέλεση μαθηματικών υπολογισμών, σχεδίασης, ανάλυσης και βελτιστοποίησης. Επίσης είναι ταχύτερο σε σχέση με τις παραδοσιακές γλώσσες προγραμματισμού. Έτσι η επεξεργασία και η μοντελοποίηση δεδομένων όπως πχ. Η ανάλυση παλινδρόμησης γίνεται ευκολότερα καθώς με τη χρήση παρέχει ανάλυση χρονολογικών σειρών, GUI και στατιστικά στοιχεία βασισμένα σε πίνακες. Τέλος αξίζει να σημειωθεί ότι οι εξαντλητικές ενσωματωμένες λειτουργίες που χρησιμοποιεί για τη σχεδίαση το καθιστούν ένα πολύπλοκο εργαλείο ανάλυσης. Σε συνδυασμό με πακέτα όπως το FastICA όπου χρησιμοποιείται για την εκτέλεση ανεξάρτητης ανάλυσης στοιχείων, μπορούν να αναπτυχθούν ισχυρότεροι αλγόριθμοι.

Ένα τελευταίο παράδειγμα είναι το Apache UIMA (Unstructured Information Management Applications) το οποίο είναι ένα έργο ανοιχτού κώδικα όπου αναλύει μεγάλο όγκο δεδομένων και ανακαλύπτει πληροφορίες που σχετίζονται με τον χρήστη.

7.3. Επιχειρηματικές εργαλειοθήκες

Τα Business Toolkits είναι μια συλλογή έμπειρων και προσαρμοσμένων εργαλείων όπου επιτρέπουν στους χρήστες να προμηθεύονται, να αναζητούν και να αναλύουν κείμενο για μια σειρά εμπορικών σκοπών.

Το SAS Sentiment Analysis Manager είναι μέρος του προγράμματος SAS Text Analytic και χρησιμοποιείται για να συλλέγει εσωτερικές πληροφορίες κειμένου ενός οργανισμού, να δημιουργεί αναφορές συναισθημάτων καταναλωτών, ανταγωνιστών

και πελατών και να συλλέγει πηγές περιεχομένων όπως αυτές που έχουν τα μέσα κοινωνικής δικτύωσης.

Το RapidMiner είναι μια συλλογή εργαλείων που προσφέρει εξόρυξη δεδομένων και κειμένων, προγνωστική ανάλυση , μοντελοποίηση, αξιολόγηση και ανάπτυξη. Είναι μια δωρεάν κοινοτική έκδοση υπό GNU AGPL καθώς και σε έκδοση Enterprise που προσφέρεται με εμπορική άδεια. Το λογισμικό είναι γραμμένο σε Java και χρησιμοποιεί σχήματα μάθησης και αξιολογητές χαρακτηριστικών από το περιβάλλον μηχανικής μάθησης Weka και σχήματα στατιστικής μοντελοποίησης από το έργο R.

Άλλο παράδειγμα είναι η Lexalytics η οποία περιέχει μια μηχανή ανάλυσης εμπορικού συναισθήματος για άμεσους πελάτες OEM. Τέλος το IBM SPSS Statistics είναι ένα από τα προγράμματα που χρησιμοποιούνται συχνότερα για τη στατιστική ανάλυση στις κοινωνικές επιστήμες.

7.4. Εργαλεία παρακολούθησης μέσων κοινωνικής δικτύωσης

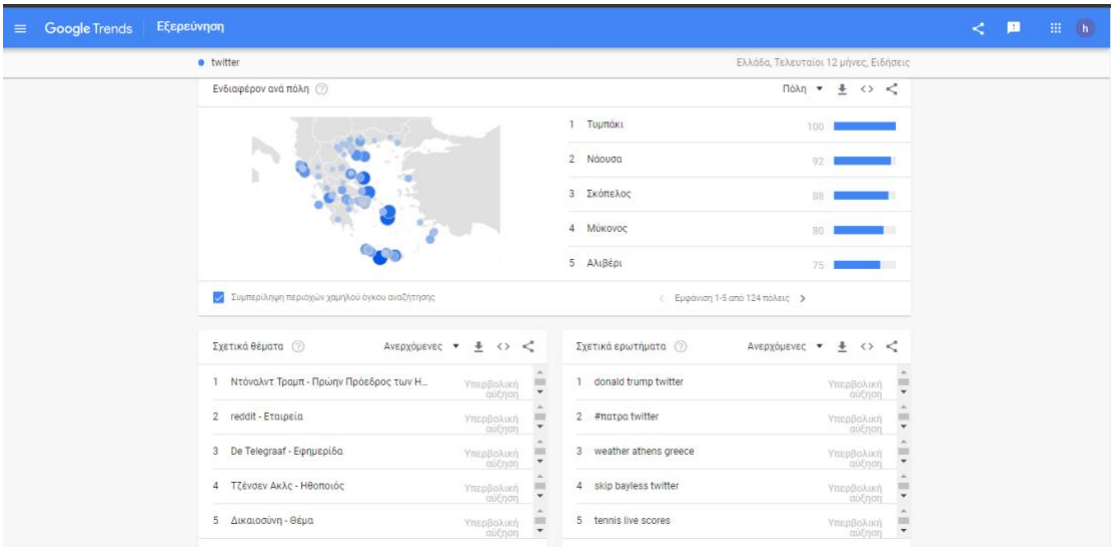
Τα εργαλεία παρακολούθησης μέσων κοινωνικής δικτύωσης χρησιμοποιούνται για την ανάλυση συναισθημάτων και μέτρηση απόψεων που έχουν τα άτομα για την εκάστοτε εταιρεία η και το προϊόν της και γενικότερα για θέματα που αφορούν τα μέσα κοινωνική δικτύωσης της εταιρείας.

Στον τομέα της παρακολούθησης μέσων κοινωνικής δικτύωσης περιλαμβάνονται και τα εξής παραδείγματα:

- Κοινωνική αναφορά: Η οποία παρέχει ειδοποιήσεις μέσων κοινωνικής δικτύωσης παρόμοιες με αυτές της Google.
- Ενισχυμένο Analytics το οποίο εστιάζει σε πληροφορίες του Marketing και σε κριτικές που αφορούν προϊόντα.
- Παρακολούθηση μέσων κοινωνικής δικτύωσης Lithium και Trackur, το οποίο είναι ένα διαδικτυακό εργαλείο συλλογής δεδομένων από το Διαδίκτυο όπου παρακολουθεί σε συνεχή ροή όσα λέγονται μέσα από σχόλια, like κτλ.

Η Google παρέχει επίσης μερικά χρήσιμα δωρεάν εργαλεία. Όπως το Google Trends συγκρίνει μια συγκεκριμένη εισαγωγή όρου αναζήτησης με τον συνολικό όγκο αναζητήσεων. Ένα άλλο εργαλείο είναι το Google Alerts όπου έχει δημιουργηθεί γύρω από την Αναζήτηση Google και ανιχνεύει αλλαγές που γίνονται στα περιεχόμενα και ειδοποιεί αυτόματα τον χρήστη.

Στην Εικ. 19 απεικονίζεται πόσοι χρήστες από την χώρα μας(ανα πόλη) αναζήτησαν τους τελευταίους 12 μήνες μέσω του Twitter διάφορες ειδήσεις.



https://trends.google.com/trends/explore?cat=16&geo=GR&q=twitter

Επίσης υπάρχουν και άλλα σημαντικά εργαλεία και πλατφόρμες παρακολούθησης τα οποία είναι τα εξής:

1. **Hootsuite**

Η οποία είναι μια πλατφόρμα διαχείρισης μέσων κοινωνικής δικτύωσης. Για την πρόσβαση και την χρήση της πλατφόρμας οι χρήστες/πελάτες πρέπει να πληρώσουν ώστε να μπορέσουν να επικοινωνήσουν, να αλληλεπιδράσουν με το κοινό ακόμα και να δημιουργήσουν μια επιχείρηση. Επίσης έχει ενσωματωθεί με τα κοινωνικά μέσα όπως είναι Facebook, Twitter, LinkedIn κ.π.λ. Στην Εικ. 20 απεικονίζονται τα στατιστικά ενός προφίλ χρήστη του Twitter.

Εικ. 20

2. Talkwalker

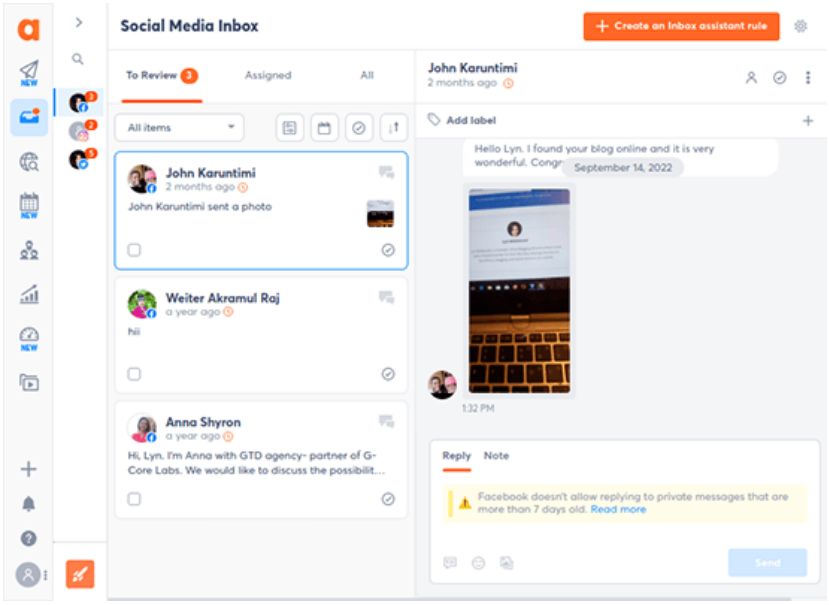
Είναι μια ολοκληρωμένη πλατφόρμα που εξυπηρετεί τις ανάγκες κάθε επιχειρήσεις. Μπορεί να παρακολουθήσει την φήμη μια επωνυμίας όπως και των μέσων κοινωνικής δικτύωσης καθώς επίσης και να μπορεί να τα συγκρίνει με άλλα. Στην Εικ. 21 απεικονίζεται η σύγκριση μεταξύ Facebook, Twitter και LinkedIn σε ότι αφορά τα posts που έχουν γίνει και από τα τρια κοινωνικά μέσα.

Εικ. 21

3. Agorapulse

Διαθέτει εργαλεία τα οποία είναι πιο προσβάσιμα για τις εταιρείες καθώς έχουν χαμηλό κόστος σε σύγκριση με άλλες πλατφόρμες. Αυτά τα εργαλεία επιτρέπουν στις εταιρείες να λαμβάνουν πληροφορίες από άτομα που τις παρακολουθούν έτσι ώστε να μπορούν να προσεγγίσουν και να εξυπηρετήσουν το κοινό στόχο τους. Στην Εικ. 22 βλέπουμε μια απλή διεπαφή μεταξύ χρηστών.

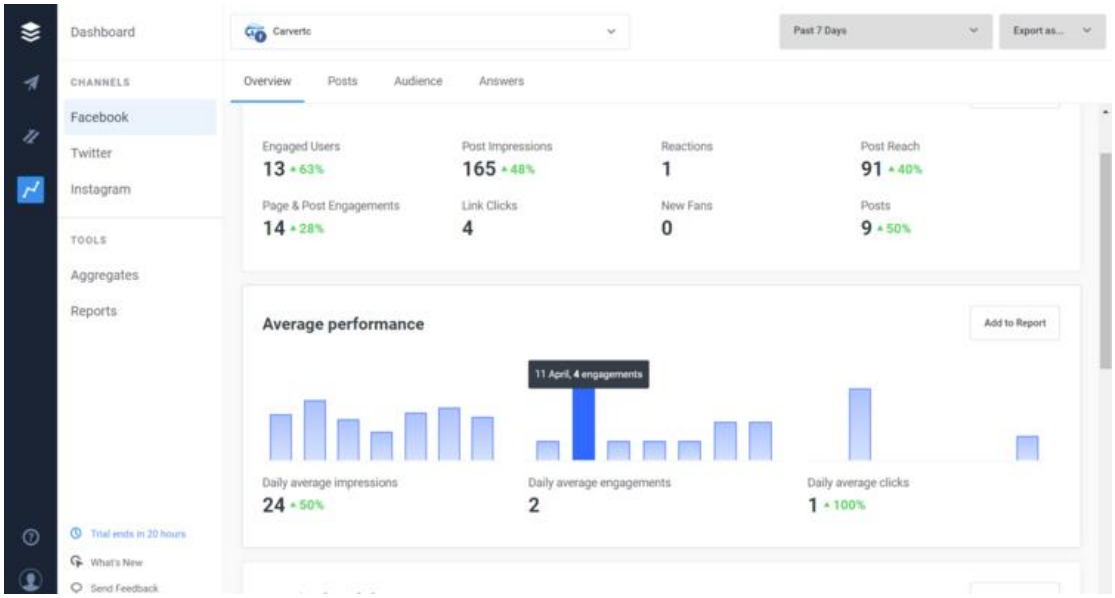
Εικ. 22



4. Buffer Analyze

Είναι μια πλατφόρμα λεπτομερής ανάλυσης αναρτήσεων στα μέσα κοινωνικής δικτύωσης. Για κάθε ανάρτηση που γίνεται υπάρχει η δυνατότητα μέτρησης ώστε να δούμε πως το κοινό αντέδρασε σε μια ανάρτηση. Στην Εικ. 23 απεικονίζεται η απόδοση του Facebook

Εικ. 23



5. BuzzSumo

Είναι ένα εργαλείο δημιουργίας και αναζήτησης περιεχομένων που μπορούμε να βλέπουμε τις κοινοποιήσεις, τα σχόλια, τους συνδέσμους κτλ.. Επίσης δίνει την δυνατότητα ανακάλυψης νέων ευκαιριών. Η Εικ. 24 περιλαμβάνει μια συλλογή σε γραφήματα που εμφανίζει το περιεχόμενο που δημοσιεύεται με την πάροδο του χρόνου, την μέση τιμή αφοσίωσης ανά τύπο περιεχομένου και την συνολική αφοσίωση.

Εικ. 24

7.5. Εργαλεία ανάλυσης κειμένου

Τα εργαλεία ανάλυσης κειμένου είναι εργαλεία για την επεξεργασία φυσικής γλώσσας και την ανάλυση κειμένου. Παραδείγματα εταιρειών ανάλυσης κειμένου είναι: OpenAmplify και Jodange των οποίων τα εργαλεία φιλτράρουν και συγκεντρώνουν αυτόματα τις σκέψεις, τα συναισθήματα και τι δηλώσεις από τα παραδοσιακά και τα μέσα κοινωνικής δικτύωσης.

Επίσης υπάρχει και ένας μεγάλος αριθμός από ελεύθερα διαθέσιμα εργαλεία τα οποία δημιουργούνται από ακαδημαϊκές ομάδες και μη κυβερνητικές οργανώσεις (ΜΚΟ) για την αναζήτηση και την επίλυση απόψεων. Μερικά παραδείγματα περιλαμβάνουν τα εργαλεία της ομάδας Stanford NLP και το LingPipe, μια σειρά βιβλιοθηκών Java για τη γλωσσική ανάλυση της ανθρώπινης γλώσσας (Teufl et al 2010).

Ένα ακόμα δημοφιλές εργαλείο είναι το Python NLTK—Natural Language Toolkit το οποίο περιλαμβάνει ενότητες Python και διαθέτει μια ποικιλία εργαλείων ανάλυσης κειμένου ανοιχτού κώδικα και ανάλυσης συναισθήματος. Ένα άλλο επίσης χρήσιμο εργαλείο είναι το GATE.

Το Lexalytics Sentiment Toolkit είναι επίσης ένα σημαντικό εργαλείο καθώς εκτελεί αυτόματη ανάλυση συναισθήματος σε έγγραφα εισόδου. Έχει επίσης τη δυνατότητα να χρησιμοποιείται σε μεγάλο αριθμών εγγράφων, όμως δεν μπορεί να χρησιμοποιήσει απόξεση δεδομένων

Άλλα εμπορικά λογισμικά για εξόρυξη κειμένου περιλαμβάνουν: AeroText, Attensity, Clarabridge, IBM LanguageWare, SPSS Text Analytics for Surveys, Language Computer Corporation, STATISTICA Text Miner και WordStat.

7.6. Εργαλεία οπτικοποίησης δεδομένων

Τα εργαλεία οπτικοποίησης δεδομένων δίνουν τη δυνατότητα στους χρήστες να αποκτήσουν πληροφορίες από τα «μεγάλα» δεδομένα. Μπορούν επίσης να πραγματοποιήσουν διερευνητική ανάλυση μέσω διαδραστικών διεπαφών που είναι που είναι πλέον διαθέσιμες και στη πλειονότητα των συσκευών, με πρόσφατη εστίαση στις κινητές συσκευές(smartphones). Τα εργαλεία οπτικοποίησης δεδομένων βοηθούν τους

χρήστες να εντοπίσουν μοτίβα, τάσεις και σχέσεις στα δεδομένα που ήταν προηγουμένως λανθάνοντα. Η γρήγορη ad hoc οπτικοποίηση στα δεδομένα μπορεί να αποκαλύψει μοτίβα και ακραίες τιμές και μπορεί να εκτελεστεί σε πλαίσια συνόλων δεδομένων μεγάλης κλίμακας, όπως το Apache Hadoop ή το Amazon Kinesis. Δύο αξιοσημείωτα εργαλεία οπτικοποίησης είναι το SAS Visual Analytics και το Tableau.

8. ΠΛΑΤΦΟΡΜΕΣ ΑΝΑΛΥΣΗΣ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Εδώ, εξετάζουμε ολοκληρωμένες πλατφόρμες μέσων κοινωνικής δικτύωσης που συνδυάζουν αρχεία μέσων κοινωνικής δικτύωσης, ροές δεδομένων, εξόρυξη δεδομένων και εργαλεία ανάλυσης δεδομένων. Με απλά λόγια, οι πλατφόρμες διαφέρουν από τα εργαλεία και τα κιτ εργαλείων, καθώς οι πλατφόρμες είναι πιο ολοκληρωμένες και παρέχουν τόσο εργαλεία όσο και δεδομένα.

Σε γενικές γραμμές υποδιαιρούνται σε:

- **Ειδησεογραφικές πλατφόρμες:** Πλατφόρμες όπως η Thomson Reuters που παρέχουν αρχεία/ροές ειδήσεων και συναφή αναλυτικά στοιχεία και στοχεύουν εταιρείες όπως χρηματοπιστωτικά ιδρύματα που επιδιώκουν να παρακολουθούν το κλίμα της αγοράς στις ειδήσεις.

- **Πλατφόρμες μέσων κοινωνικής δικτύωσης** : Πλατφόρμες που παρέχουν εξόρυξη δεδομένων και αναλύσεις στο Twitter, στο Facebook και σε ένα ευρύ φάσμα άλλων πηγών μέσων κοινωνικής δικτύωσης. Οι πάροχοι στοχεύουν συνήθως εταιρείες που επιδιώκουν να παρακολουθούν το συναίσθημα γύρω από τις μάρκες ή τα προϊόντα τους.

8.1. Πλατφόρμες ειδήσεων

Οι δύο πιο σημαντικοί πάροχοι επιχειρηματικών ειδήσεων είναι η Thomson Reuters και το Bloomberg.

Ο υπολογιστής διαβάζει ειδήσεις σε πραγματικό χρόνο και παρέχει αυτόματα βασικούς δείκτες και ουσιαστικές πληροφορίες. Οι ειδήσεις ανακτώνται αυτόματα, αναλύονται και ερμηνεύονται σε λίγα χιλιοστά του δευτερολέπτου. Οι μηχανικά αναγνώσιμοι δείκτες ειδήσεων μπορούν ενδεχομένως να βελτιώσουν τις ποσοτικές στρατηγικές, τη διαχείριση κινδύνου και τη λήψη αποφάσεων.

Παραδείγματα ειδήσεων με δυνατότητα ανάγνωσης από μηχανή περιλαμβάνουν: Ειδήσεις αναγνώσιμες από μηχανή Thomson Reuters, ροή συναλλαγών βάσει συμβάντων του Bloomberg και AlphaFlash (ροή ειδήσεων αναγνώσιμη από μηχανής της Deutsche Börse).

Το Thomson Reuters News Analytics χρησιμοποιεί τεχνικές Επεξεργασίας Φυσικής Γλώσσας (NLP) για τη βαθμολογία ειδήσεων σε δεκάδες χιλιάδες εταιρείες και σχεδόν 40 εμπορεύματα και θέματα ενέργειας. Τα στοιχεία μετρώνται στις ακόλουθες διαστάσεις:

- **Συναίσθημα συγγραφέα**: Μετρήσεις για το πόσο θετικός, αρνητικός ή ουδέτερος είναι ο τόνος του αντικειμένου, ειδικά για κάθε εταιρεία του άρθρου.
- **Συνάφεια**: Μετρήσεις για το πόσο σχετική ή ουσιαστική είναι η ιστορία για ένα συγκεκριμένο αντικείμενο.
- **Ανάλυση όγκου**: Μετρήσεις για το πόσα νέα συμβαίνουν σε μια συγκεκριμένη εταιρεία.
- **Μοναδικότητα**: Μετρήσεις για το πόσο νέο είναι το αντικείμενο σε διάφορες χρονικές περιόδους ή αν επαναλαμβάνεται.
- **Ανάλυση επικεφαλίδων:** Υποδηλώνει ειδικά χαρακτηριστικά όπως συνεντεύξεις, αποκλειστικότητες και περιλήψεις.

8.2. Πλατφόρμες μέσων κοινωνικής δικτύωσης

Το Attensity, το Brandwatch, το Salesforce Marketing Cloud (πρώην Radian6) και το Sysomos MAP (Media Analysis Platform) είναι παραδείγματα πλατφορμών παρακολούθησης μέσων κοινωνικής δικτύωσης, που μετρούν δημογραφικά στοιχεία, θέματα με επιρροή και συναισθήματα. Περιλαμβάνουν ανάλυση κειμένου και ανάλυση συναισθημάτων σε διαδικτυακές συνομιλίες καταναλωτών και παρέχουν φιλικές προς το χρήστη διεπαφές για την προσαρμογή του ερωτήματος αναζήτησης, των πινάκων

εργαλείων, των αναφορών και των δυνατοτήτων εξαγωγής αρχείων (π.χ. σε μορφή Excel ή CSV).

Οι πλατφόρμες ανάλυσης συναισθήματος χρησιμοποιούν δύο κύριες μεθοδολογίες. Η πρώτη περιλαμβάνει μια προσέγγιση στατιστικής ή βασισμένης σε μοντέλα, όπου το σύστημα μαθαίνει να αξιολογεί το συναίσθημα αναλύοντας μεγάλες ποσότητες υλικού προ-βαθμολογημένου. Η δεύτερη μέθοδος χρησιμοποιεί ένα μεγάλο λεξικό με προ-βαθμολογημένες φράσεις.

Το RapidMiner είναι μια πλατφόρμα που συνδυάζει την εξόρυξη δεδομένων και την ανάλυση δεδομένων, η οποία, ανάλογα με τις απαιτήσεις, μπορεί να είναι ανοιχτού κώδικα. Χρησιμοποιεί τη βιβλιοθήκη μηχανικής εκμάθησης WEKA και παρέχει πρόσβαση σε πηγές δεδομένων όπως Excel, Access, MySQL, PostgreSQL και αρχεία κειμένου.

Το DataSift παρέχει πρόσβαση σε δεδομένα κοινωνικής δικτύωσης σε πραγματικό χρόνο και σε ιστορικά δεδομένα από τα κορυφαία κοινωνικά δίκτυα και εκατομμύρια άλλες πηγές, επιτρέποντας στους πελάτες να συγκεντρώνουν, να φιλτράρουν , να αποκτούν πληροφορίες και να ανακαλύπτουν τάσεις από δημόσιες κοινωνικές συνομιλίες. Μόλις αυτά τα δεδομένα συγκεντρωθούν και υποβληθούν σε επεξεργασία όπως επεξεργασία γλώσσας, γεωγραφικά δεδομένα, δημογραφικά στοιχεία και κατηγοριοποίηση/ανίχνευση ανεπιθύμητης αλληλογραφίας κ.π.λ. οι πελάτες μπορούν να χρησιμοποιήσουν προκατασκευασμένες ενσωματώσεις με δημοφιλή εργαλεία BI, εργαλεία εφαρμογών και προγραμματιστών για την παράδοση των δεδομένων στις επιχειρήσεις τους ή τη χρήση των API του DataSift για τη ροή δεδομένων σε πραγματικό χρόνο στις εφαρμογές τους.

Η Thomson Reuters ανακοίνωσε πρόσφατα ότι ενσωματώνει τώρα ανάλυση συναισθήματος Twitter για την πλατφόρμα ανάλυσης και συναλλαγών αγοράς Thomson Reuters Eikon, παρέχοντας οπτικοποιήσεις και γραφήματα με βάση τα δεδομένα συναισθήματος. Το προηγούμενο έτος, το Bloomberg ενσωμάτωσε tweets που σχετίζονται με συγκεκριμένες εταιρείες σε μια ευρύτερη ροή δεδομένων.

8.3. Μελέτη περίπτωσης: Thomson Reuters News Analytics

Το Thomson Reuters News Analytics παρέχει ένα τεράστιο αρχείο ειδήσεων με αναλυτικά στοιχεία για την ανάγνωση και την ερμηνεία ειδήσεων, προσφέροντας σημαντικές πληροφορίες. Το Thomson Reuters News Analytics βαθμολογεί ειδήσεις σε περισσότερες από 25.000 μετοχές και σχεδόν 40 θέματα (εμπορεύματα και ενέργεια). Η πλατφόρμα αποκόπτει και αναλύει δεδομένα ειδήσεων σε πραγματικό χρόνο και τροφοδοτεί τα δεδομένα σε άλλα προγράμματα/έργα ή ποσοτικές στρατηγικές. Τέλος χρησιμοποιεί ένα σύστημα NLP από τη Lexalytics, έναν από τους ηγέτες της τεχνολογίας γλωσσολογίας, που μπορεί να παρακολουθεί το συναίσθημα των ειδήσεων με την πάροδο του χρόνου και να βαθμολογεί κείμενο σε διάφορες διαστάσεις.

Η βαθμολογία κειμένου και τα μεταδεδομένα της πλατφόρμας έχουν περισσότερα από 80 πεδία (Thomson Reuters 2010) όπως:

- **Τύπος στοιχείου**: Στάδιο της ιστορίας όπως ειδοποίηση, άρθρο, ενημερώσεις ή διορθώσεις.
- **Είδος αντικειμένου**: Ταξινόμηση της ιστορίας όπως π.χ. συνέντευξη, αποκλειστική και ολοκλήρωση.
- **Επικεφαλίδα**: Ειδοποίηση ή κείμενο επικεφαλίδας.
- **Συνάφεια**: Ποικίλλει από 0 έως 1,0.
- **Το κυρίαρχο συναίσθημα**: Συνήθως είναι 1, 0 ή −1.
- **Θετικό, ουδέτερο, αρνητικό**: Λεπτομερής ένδειξη συναισθήματος.
- **Ενέργεια μεσίτη:** Υποδηλώνει αναβάθμιση, υποβάθμιση, διατήρηση, απροσδιόριστο ή αν είναι ο ίδιος ο μεσίτης
- **Σχόλιο τιμής/αγοράς**: Χρησιμοποιείται για την επισήμανση στοιχείων που περιγράφουν σχολιασμό τιμολόγησης/αγοράς

- **Κωδικοί θεμάτων**: Περιγράφει τι αφορά η ιστορία. π.χ. RCH = Έρευνα, ΑΠΕ = Αποτελέσματα, RESF = Πρόβλεψη αποτελεσμάτων, MRG = Συγχωνεύσεις και Εξαγορές

Ένα απόσπασμα της ανάλυσης του συναισθήματος των ειδήσεων απεικονίζεται στην Εικ. 25.

Εικ. 25

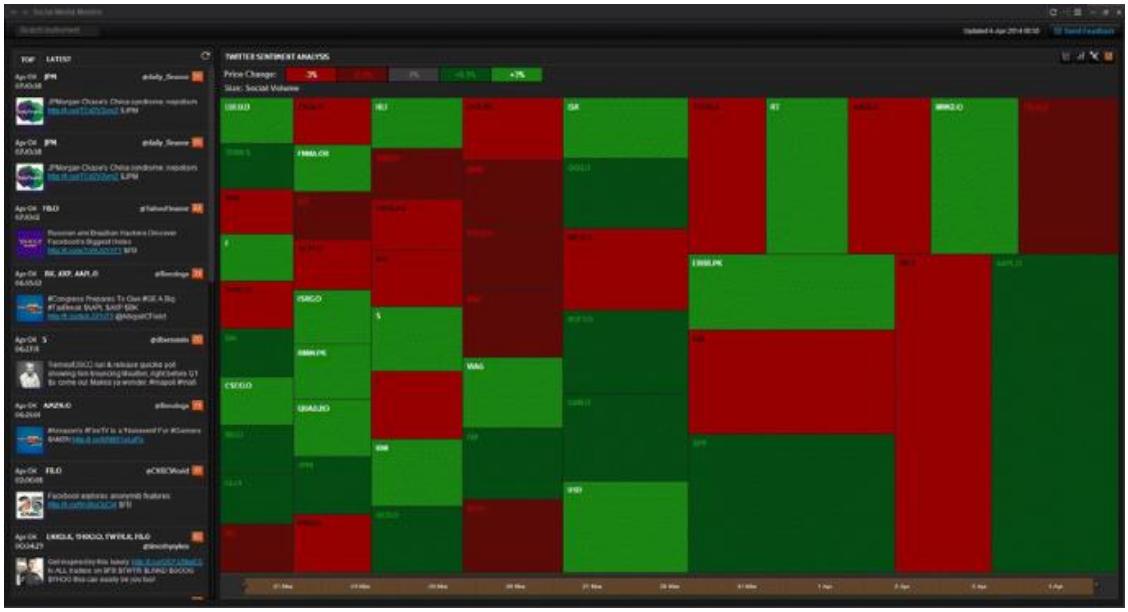
Thomson Reuters News Discovery Application with Sentiment Analysis

Η Thomson Reuters το 2012 επέκτεινε τις αναγνώσιμες από μηχανή ειδήσεις της ώστε να συμπεριλάβει την ανάλυση συναισθημάτων και τη βαθμολογία για τα μέσα κοινωνικής δικτύωσης. Η επέκταση του ονομάζεται Thomson Reuters News Analytics (TRNA) για Internet News και Social Media, η οποία συγκεντρώνει περιεχόμενο από περισσότερα από τέσσερα εκατομμύρια κανάλια μέσων κοινωνικής δικτύωσης και 50.000 ιστότοπους ειδήσεων στο Διαδίκτυο. Έπειτα, το περιεχόμενο αναλύεται από το TRNA σε πραγματικό χρόνο, δημιουργώντας ένα μετρήσιμο αποτέλεσμα σε διαστάσεις, όπως συναίσθημα, συνάφεια, όγκος καινοτομίας, κατηγορία και κατάταξη

πηγής. Αυτή η επέκταση χρησιμοποιεί την ίδια εκτεταμένη προσθήκη ετικετών μεταδεδομένων (σε περισσότερα από 80 πεδία).

Το TRNA για Internet News και Social Media είναι μια ισχυρή πλατφόρμα που αναλύει, προσθέτει ετικέτες και φιλτράρει εκατομμύρια δημόσιες και κορυφαίες πηγές περιεχομένου στο Διαδίκτυο, μετατρέποντας τα μεγάλα δεδομένα σε εφαρμόσιμες ιδέες. Παρέχει επίσης έναν τρόπο οπτικής ανάλυσης των μεγάλων δεδομένων. Μπορεί να συνδυαστεί με το λογισμικό Panopticon Data Visualization Software προκειμένου να καταλήξουμε σε ουσιαστικά συμπεράσματα πιο γρήγορα με οπτικά διαισθητικές οθόνες (Thomson Reuters 2012a, b, c), όπως φαίνεται στην Εικ. 26.

Εικ. 26



Συνδυασμός TRNA για ειδήσεις στο Διαδίκτυο και μέσα κοινωνικής δικτύωσης με λογισμικό οπτικοποίησης δεδομένων Panopticon

Η Thomson Reuters επέκτεινε επίσης την υπηρεσία News Analytics με MarketPsych Indices , η οποία επιτρέπει την ψυχολογική ανάλυση σε πραγματικό χρόνο των ειδήσεων και των μέσων κοινωνικής δικτύωσης. Η υπηρεσία Thomson Reuters MarketPsych Indices (TRMI) αποκτά άποψη για την ψυχολογία της αγοράς καθώς προσπαθεί να προσδιορίσει το ανθρώπινο συναίσθημα. Συμπληρώνει το TRNA και

χρησιμοποιεί επεξεργασία NLP που δημιουργήθηκε από την MarketPsych μια κορυφαία εταιρεία στη συμπεριφορική ψυχολογία στις χρηματοπιστωτικές αγορές.

Οι οικονομολόγοι της συμπεριφοράς έχουν ερευνήσει εκτενώς εάν τα συναισθήματα επηρεάζουν τις αγορές με προβλέψιμους τρόπους και το TRMI επιχειρεί να μετρήσει την κατάσταση των «συναισθημάτων» σε πραγματικό χρόνο προκειμένου να εντοπίσει τα μοτίβα καθώς αυτά εμφανίζονται. Το TRMI έχει δύο βασικούς τύπους δεικτών:

- **Συναισθηματικοί δείκτες (συναισθήματα):** Συναισθήματα όπως η χαρά, η εμπιστοσύνη, ο φόβος, ο θυμός, το άγχος κ.π.λ.
- **Μετρήσεις Buzz**: Υποδεικνύουν πόσο συζητείται κάτι στις ειδήσεις και τα μέσα κοινωνικής δικτύωσης και περιλαμβάνουν μακροοικονομικά θέματα όπως χρηματοοικονομικός τομέας, τομέας καθαρής τεχνολογίας κ.π.λ.

Τέλος η πλατφόρμα της Thomson Reuters επιτρέπει την εκμετάλλευση των ειδήσεων και των μέσων κοινωνικής δικτύωσης για τον εντοπισμό ευκαιριών και την αξιοποίηση της αποτελεσματικότητας της αγοράς (Thomson Reuters 2013).

9. ΕΡΕΥΝΑ - ΕΠΕΞΕΡΓΑΣΙΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

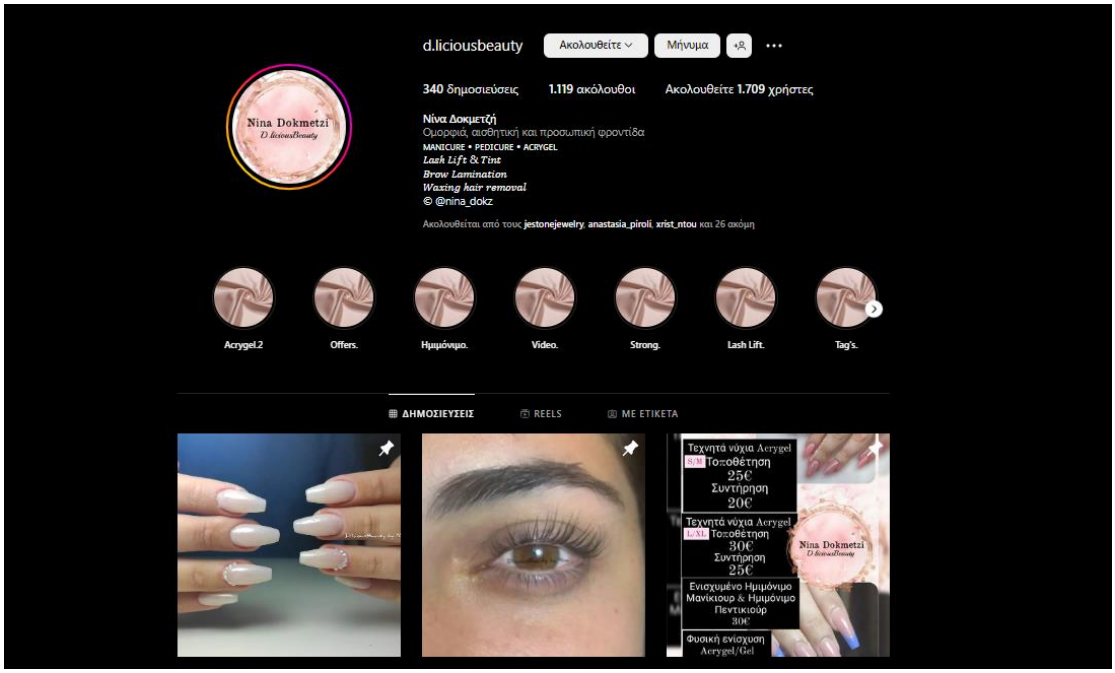
Όπως είναι πλέον γνωστό στις μέρες μας πολλές επιχειρήσεις, εταιρείες κ.τ.λ δραστηριοποιούνται μέσα από τα μέσα κοινωνικής δικτύωσης. Διαφημίζουν, προωθούν τα προϊόντα ή τις υπηρεσίες τους και προσεγγίζουν πελάτες.

Η έρευνα που κάναμε αφορά μια μικρή επιχείρηση στην Ελλάδα η οποία ανήκει και στην μια εκ των δυο εισηγητών της Πτυχιακής εργασίας. Η Δοκμετζή Δέσποινα - Αικατερίνη έχει δημιουργήσει έναν επαγγελματικό λογαριασμό στα μέσα κοινωνικής δικτύωσης και πιο συγκεκριμένα στο Instagram όπου ασχολείται με την ομορφιά, αισθητική και προσωπική φροντίδα.

Η έρευνα έγινε με σκοπό να μελετήσουμε τα στατιστικά της επιχείρησης των τελευταίο μήνα, πόσους χρήστες προσέγγισε, από που τους προσέγγισε και αν ήδη ήταν ακόλουθοι της επιχείρησης η ανακάλυψαν της επιχείρηση μέσα από δημοσιεύσεις, Video, Reels κ.π.λ.

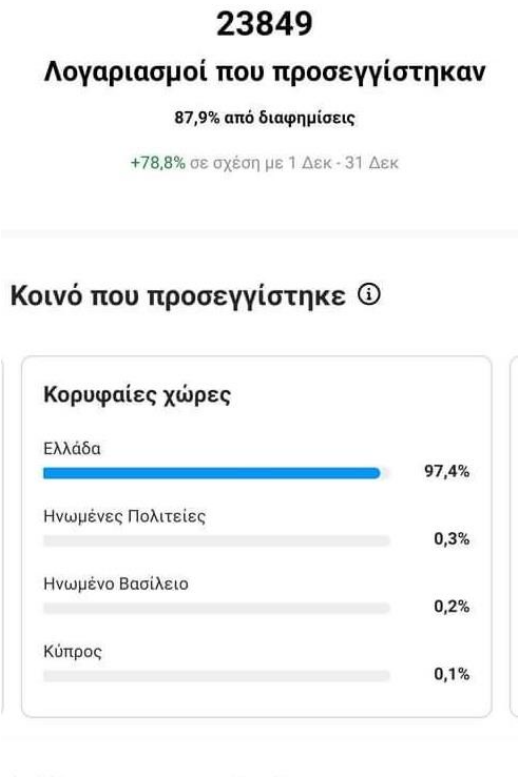
Τέλος θα αναφερθούμε στα αποτελέσματα της έρευνας, στα συμπεράσματα τα οποία προέκυψαν και στις προτάσεις που θα βοηθήσουν για καλύτερη επίδοση μελλοντικά.

Στην Εικ. 27 απεικονίζεται το προφίλ της επιχείρησης.



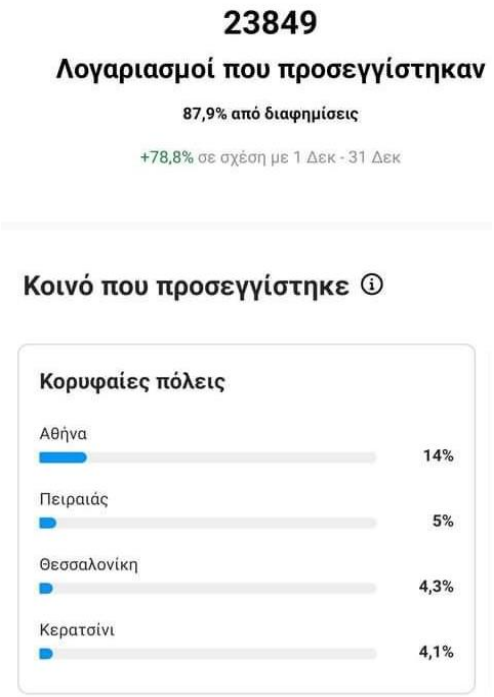
9.1. Επεξεργασία και ανάλυση έρευνας

1. Κοινό που προσεγγίστηκε (ανά χώρα)



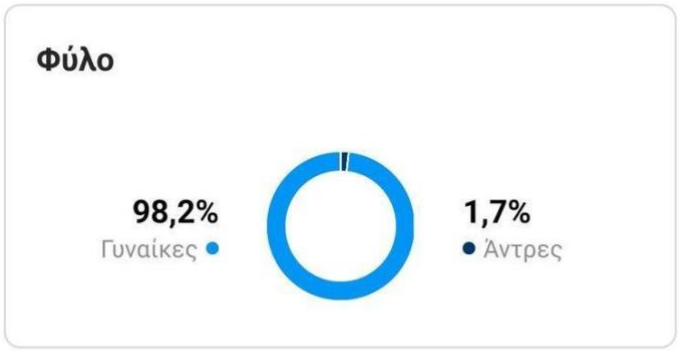
Από τους 23.849 λογαριασμούς που προσεγγίστηκαν (στο περίπου):	
Ελλάδα	23.228
Ηνωμένες Πολιτείες	71
Ηνωμένο Βασίλειο	47
Κύπρος	23

2. Κοινό που προσεγγίστηκε (ανά πόλη της Ελλάδος)



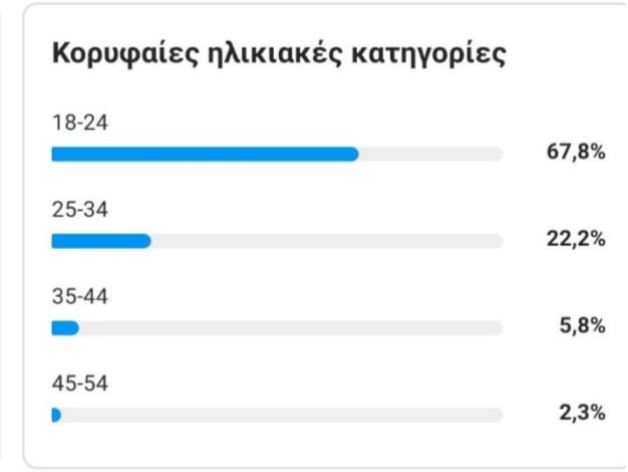
Από τους 23.849 λογαριασμούς που προσεγγίστηκαν (στο περίπου):	
Αθήνα	3.338
Πειραιάς	1.192
Θεσσαλονική	1.025
Κερατσίνι	977

3. Φύλο



(Από τους 23.849 χρήστες οι 23.443 είναι Γυναίκες ενώ το υπόλοιπο 406 είναι Άντρες).

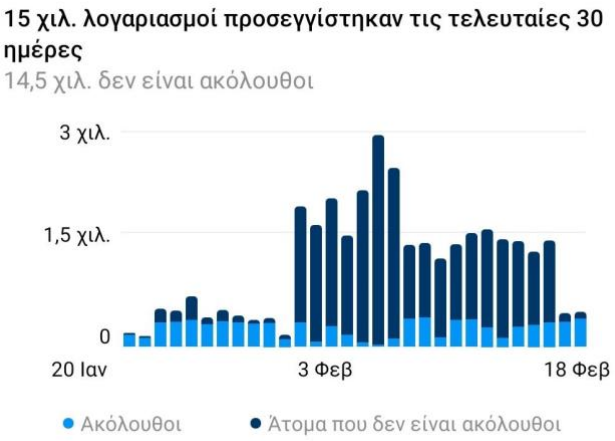
4. Ηλικιακή κατηγορία



Διαχωρισμός ηλικιακής κατηγορίας από τους 23.849 λογαριασμούς (στο περίπου):	
18 - 24	16.169
25 - 34	5.294
35 - 44	1.383
45 - 54	548

5. Απήχηση Περιεχομένου

5. Συνολική προσέγγιση



(Παρατηρούμε ότι η πλειοψηφία των χρηστών δεν είναι ακόλουθοι μας και σχεδόν μόνο το 3,5% είναι από άτομα τα οποία είναι ακόλουθοι).

9.2. Αποτελέσματα

Μετά την έρευνα που διεξήχθη παρατηρούμε τα εξής αποτελέσματα:

- Το μεγαλύτερο ποσοστό ανθρώπων είναι από την Ελλάδα καθώς είναι και εκεί η έδρα της επιχείρησης.
- Μεγαλύτερη απήχηση το νεανικό κοινό.
- Αυξημένο ενδιαφέρον κυρίως από άτομα τα οποία δεν είναι ακόλουθοι.
- Αυξητική τάση σε σύγκριση με τον προηγούμενο μήνα.
- Μεγάλη απήχηση κυρίως στις δημοσιεύσεις.

9.3. Συμπεράσματα

Παρατηρούμε ότι η πλειοψηφία των χρηστών δεν ακολουθεί την επιχείρηση μέσω "Follow". Δεν θα λέγαμε όμως ότι είναι κάτι αρνητικό για την επιχείρηση καθώς η σελίδα της επιχείρησης έχει προγραμματιστεί να εμφανίζεται σαν διαφήμιση σε προφίλ χρηστών τα οποία δεν είναι ακόλουθοι της.

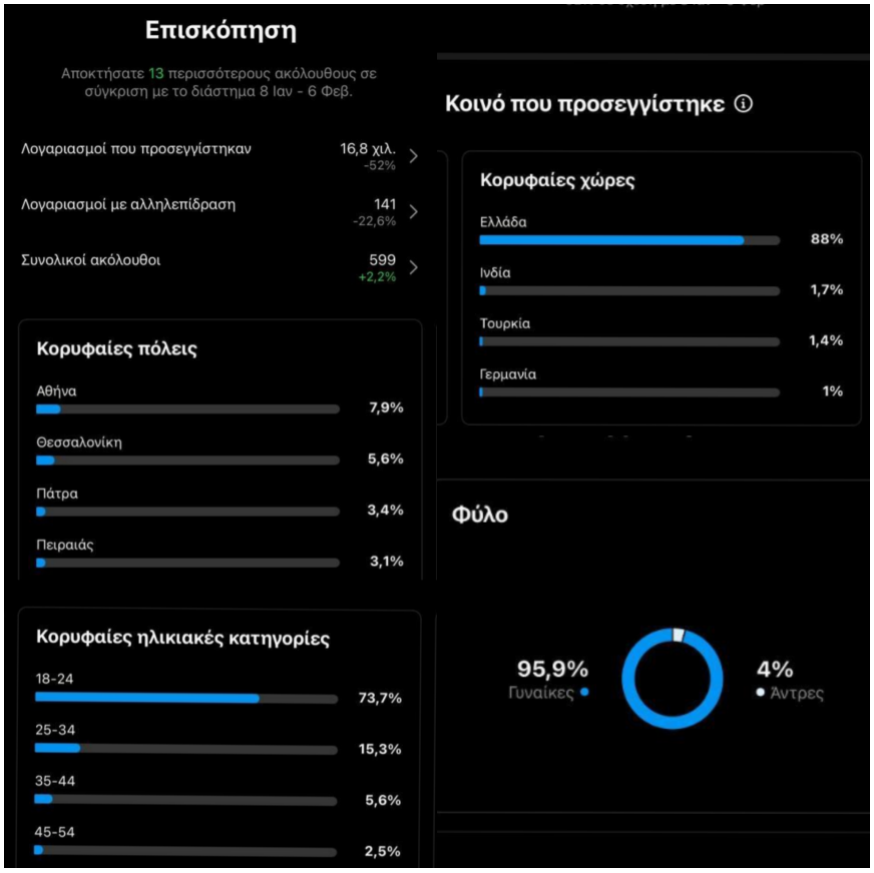
Επίσης η απήχηση που υπάρχει στα άτομα πιο νεανικής ηλικιακής ομάδας είναι κάτι το οποίο είχε προβλεφθεί εξ αρχής καθώς η νέα γενιά πλέον δραστηριοποιείται στα μέσα κοινωνικής δικτύωσης και είναι πιο εύκολο για αυτούς να επιλέξουν ένα προϊόν ή μια υπηρεσία μέσα από τα Social Media.

Τέλος να συμπληρώσουμε ότι μια επιχείρηση δεν την κάνει καλύτερη από μια άλλη αν μετράει μόνο τα Follow,τα Likes η τα Comments , σαφώς πολλοί κρίνουν μια επιχείρηση μέσα από αυτά. Όμως υπάρχουν και χρήστες που δεν τους αφορούν και εστιάζουν πραγματικά στα προϊόντα/υπηρεσίες που παρέχει η επιχείρηση.

Με αυτό το τρόπο θέλουμε να δείξουμε ότι δεν παίζει ρόλο πόσα άτομα έχουμε ακολούθους όταν από τα 15 χιλ. άτομα που προσεγγίσαμε τον τελευταίο μόνο μήνα, σχεδόν το 3,5% των χρηστών ήταν ακόλουθοι ενώ οι υπόλοιποι προσεγγίστηκαν μέσα από δημοσιεύσεις/προωθήσεις που έκανε η επιχείρηση.

9.4. Ανταγωνισμός

Το κομμάτι του ανταγωνισμού των επιχειρήσεων που δραστηριοποιούνται μέσα από τα μέσα κοινωνικής δικτύωσης είναι κάτι που αυξάνεται συνεχώς. Αυτό συμβαίνει διότι όπως αναφέραμε ολοένα και περισσότεροι άνθρωποι τείνουν να αγοράζουν μέσα από το Ίντερνετ και τα μέσα κοινωνικής δικτύωσης, έτσι η κάθε επιχείρηση προσπαθεί να κάνει κάτι καλύτερο από μια άλλη ώστε να έχει ανταγωνιστικό πλεονέκτημα. Ας δούμε συνοπτικά πως δραστηριοποιείται μια επιχείρηση ίδια με αυτή που αναλύσαμε.



9.4.1. Ανταγωνιστικό πλεονέκτημα

Σχετικα με την επιχείρηση την οποία αναλύσαμε έναντι μια επιχείρησης ίδιου τύπου που αναφέραμε συνοπτικά προκύπτει το εξής συμπέρασμα και ανταγωνιστικό μας πλεονέκτημα. Το πλεονέκτημά μας λοιπόν είναι ότι η επιχείρηση έχει προσεγγίσει περισσότερους χρήστες είτε είναι ακόλουθοι είτε όχι, μεγαλύτερη αυξητική τάση προσέγγισης σε σύγκριση με τον προηγούμενο μήνα καθώς επίσης και μεγαλύτερη αλληλεπίδραση στις δημοσιεύσεις. Αυτό συμβαίνει διότι έχει γίνει καλύτερη προώθηση και διαφήμιση των προϊόντων και των υπηρεσιών της κάτι το οποίο δεν έχει γίνει στην αντίστοιχη επιχείρηση και αυτό φαίνεται από τα στατιστικά.

9.4.2. Ανταγωνιστικό μειονέκτημα

Παρόλου που η επιχείρηση μας έχει περισσότερο κοινό παρατηρούμε ότι η ηλικιακή ομάδα καθώς επίσης και το φύλο που έχει προσεγγιστεί περισσότερο είναι ίδια με αυτή της ανταγωνιστικής επιχείρησης. Αυτό σαφέστατα συμβαίνει διότι και οι δυο επιχείρησεις απευθύνονται κυρίως σε γυναίκες, και η ηλικιακή ομάδα κυρίως μεταξύ 18 - 24 είναι εκείνη που θα δείξει μεγαλύτερο ενδιαφέρον για νέα προϊόντα και υπηρεσίες που αφορούν την αισθητική.

9.5. Προτάσεις για το μέλλον

Κύριος στόχος της επιχείρησης είναι κυρίως η αύξηση των στατιστικών, ώστε να καλύψει ένα ευρύ φάσμα κοινού. Για να επιτευχθεί αυτό πρέπει η επιχείρηση θα πρέπει λειτουργήσει διαφορετικά στο μέλλον ώστε να προσεγγίσει και άτομα διαφορετικού φύλου, καθώς και να συνδυάσει την ψηφιακή διαφήμιση με την παραδοσιακή για να προσεγγίσει και το κοινό το οποίο είτε λόγω ιδεολογίας, είτε λόγω ηλικιακής ομάδας δεν ασχολούνται/δραστηριοποιούνται μέσα από το Internet και τα Social Media και προτιμούν να μαθαίνουν για νέα προϊόντα και υπηρεσίες με τον παραδοσιακό τρόπο.

Παρακάτω παραθέτουμε μερικές προτάσεις σχετικα με την βελτίωση της επιχείρησης μελλοντικά:

- Χρήση υπαίθριας διαφήμισης με τις υπηρεσίες που διαθέτει πχ. αφισοκόλληση όπως επίσης
- Διανομή φυλλαδίων με προσφορά γνωριμίας και εκπτώσεις των υπηρεσιών.
- Δημιουργία σειράς προϊόντων με την επωνυμία της επιχείρησης Π.χ. Μια σειρά από κρέμες, μάσκες και λοσιόν περιποίησης προσώπου.
- Δημιουργία σειράς προϊόντων για την ανδρική περιποίηση, ώστε να προσεγγίσουμε και μια ομάδα ανδρών.
- Προώθηση και διανομή αυτών των προϊόντων σε ινστιτούτα αισθητικής για μεγαλύτερη αναγνωρισιμότητα.
- Δημιουργία ενός τηλεοπτικού spot και με τα προϊόντα με την επωνυμία όπως αναφέραμε και με τις υπηρεσίες της επιχείρησης.

Όσον αφορά των ψηφιακό τρόπο διαφήμισης οι προτάσεις είναι οι εξής:

- Δημιουργία Ιστοσελίδας στο Internet
- Διαφήμιση μέσω των μηχανών αναζήτησης όπως η Google μέσα από το Google Shopping Ads) όπου δίνει κιόλας την δυνατότητα μέτρησεις των αποτελεσμάτων προώθησης της υπηρεσίας.
- Διαφήμιση μέσω YouTube Π.χ. Ένα ολιγόλεπτο spot πριν την έναρξη ενός video που να αφορά κάτι σχετικό με υγεία, ομορφιά, περιποίηση κ.π.λ.
- Διεξαγωγή μαθημάτων/σεμιναρίων των υπηρεσιών της επιχείρησης εξ αποστάσεως ή δια ζώσης, είτε για να βοηθήσουμε ανθρώπους που είναι στο ξεκίνημα μιας παρόμοιας υπηρεσίας είτε για ανθρώπους που απλά τους ενδιαφέρει να μάθουν περαιτέρω πράγματα σχετικά με την περιποίηση και την ομορφιά.
- Συνεργασία με Influencers, όπου στο κοινό έχουν μεγάλη απήχηση και τους εμπιστεύονται να αγοράζουν τα προϊόντα ή τις υπηρεσίες που διαφημίζουν.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Αρχικά η ανάλυση των μέσων κοινωνικής δικτύωσης συνεπάγεται με την συλλογή των στατιστικών στοιχείων από πλατφόρμες και την αξιολόγηση δεδομένων που λαμβάνονται από στατιστικά στοιχεία. Όλα αυτά τα δεδομένα μπορούν να συλλεχθούν με την χρήση εργαλείων ανάλυσης μέσων κοινωνικής δικτύωσης.

Τα αναλυτικά μέσα κοινωνικής δικτύωσης χρησιμοποιούνται ως μέσο για την πρόσβαση στα συναισθήματα των χρηστών η των πελατών για ένα προϊόν ή υπηρεσία. Αυτή η μέθοδος ονομάζεται ανάλυση συναισθήματος και χρησιμοποιεί προηγμένες μεθόδους αλγορίθμων για την αξιολόγηση της γνώμης του χρήστη/πελάτη σχετικά με την επωνυμία και αναφέρει το συναίσθημα που εντοπίστηκε. Ωστόσο, αυτοί οι αλγόριθμοι διενεργούν συναισθηματική αξιολόγηση σε όλες τις αλληλεπιδράσεις που αφορούν το όνομα της επωνυμίας. Με την ανάλυση αυτή οι διαδικτυακές επωνυμίες έχουν την δυνατότητα να παρακολουθούν τις αντιδράσεις των πελατών/χρηστών στις προϊόν/υπηρεσίες τους.

Όπως αναφερθήκαμε η εύκολη διαθεσιμότητα των API που παρέχονται από τα μέσα κοινωνικής δικτύωσης και στις υπηρεσίες ειδήσεων έχουν οδηγήσει στην αύξηση των υπηρεσιών δεδομένων και των εργαλείων λογισμικού για την απόξεση (scrape) και την ανάλυση συναισθημάτων καθώς και των πλατφορμών ανάλυσης κοινωνικών μέσων.

Ένα από τα βασικά ερωτήματα και "προβλήματα" θα λέγαμε ότι είναι οι εταιρείες όπου ολοένα και περισσότερο περιορίζουν την πρόσβαση στα δεδομένα τους. Από την άλλη οι ερευνητές θα πρέπει να έχουν πρόσβαση στα δεδομένα των μέσων κοινωνικής δικτύωσης για πειραματισμό. Διαφορετικά η υπολογιστική κοινωνική δικτύωση θα πρέπει να γίνει τομέας μεγάλων εταιρειών/υπηρεσιών και ένα σύνολο ακαδημαϊκών που έχουν πρόσβαση σε ιδιωτικα δεδομένα όπου οι εργασίες τους δεν μπορούν να αναπαραχθούν ούτε να τους ασκηθεί κριτική. Για να γίνει όμως αυτό θα πρέπει να υπάρξει υπολογιστικό περιβάλλον και εγκαταστάσεις δεδομένων δημοσίου τομέα στο οποίο οι ερευνητές θα μπορούν να έχουν πρόσβαση μέσω μια εγκατάστασης βασισμένη στα υπολογιστικά δεδομένα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Botan I et al. (2010) SECRET: a model for analysis of the execution semantics of stream processing systems. Proc VLDB Endow 3(1–2):232–243

Salathé M et al. (2012) Digital epidemiology. PLoS Comput Biol 8(7):1–5

Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(3):1–8

Chandramouli B et al (2010) Data stream management systems for computational finance. IEEE Comput 43(12):45–52

Robert Malouf (2002) A comparison of algorithms for maximum entropy parameter estimation pp 52

Gail L. Rosen, Erin R. Reichenberger, Aaron M. Rosenfeld Author Notes (January 2011) NBC: the Naïve Bayes Classification tool webserver for taxonomic classification of metagenomic reads pp 128

Chandrasekar C, Kowsalya N (2011) Implementation of MapReduce Algorithm and Nutch Distributed File System in Nutch. Int J Comput Appl 1:6–11

Cioffi-Revilla C (2010) Computational social science. Wiley Interdiscip Rev Comput Statistics 2(3):259–271

Galas M, Brown D, Treleaven P (2012) A computational social science environment for financial/economic experiments. In: Proceedings of the Computational Social Science Society of the Americas, vol 1, pp 1–13

Hebrail G (2008) Data stream management and mining. In: Fogelman-Soulié F, Perrotta D, Piskorski J, Steinberger R (eds) Mining Massive Data Sets for Security. IOS Press, pp 89–102

Hirudkar AM, Sherekar SS (2013) Comparative analysis of data mining tools and techniques for evaluating performance of database system. Int J Comput Sci Appl 6(2):232–237

Mr. Huseyin (Mar 6, 2021) What is Matlab? Why we need it? Understanding the fundamentals of Matlab https://blog.devgenius.io/what-is-matlab-why-we-need-it-d61e405ef419

Παπαοικονόμου, Αθανάσιος (2015, Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ)), Τεχνικές ανάλυσης κοινωνικών δικτύων, με έμφαση σε γράφους εμπιστοσύνης http://hdl.handle.net/10442/hedi/39774

Original Articles What Is Complexity Science, Really? Steven E. Phelan Pages 120-136 | Published online: 15 Jun 2010 https://doi.org/10.1207/S15327000EM0301_08

Kaplan AM (2012) If you love something, let it go mobile: mobile marketing and mobile social media 4x4. Bus Horiz 55(2):129–139
Kaplan AM, Haenlein M (2010) Users of the world, unite! the challenges and opportunities of social media. Bus Horiz 53(1):59–68

Karabulut Y (2013) Can Facebook predict stock market activity? SSRN eLibrary, pp 1–58. http://ssrn.com/abstract=2017099.

Khan A, Baharudin B, Lee LH, Khan K (2010) A review of machine learning algorithms for text-documents classification. J Adv Inf Technol 1(1):4–20

Kobayashi M, Takeda K (2000) Information retrieval on the web. ACM Comput Surv CSUR 32(2):144–173

Lazer D et al (2009) Computational social science. Science 323:721–723

Lerman K, Gilder A, Dredze M, Pereira F (2008) Reading the markets: forecasting public opinion of political candidates by news analysis. In: *Proceedings of the 22nd international conference on computational linguistics* 1:473–480

MapReduce (2011) What is MapReduce?. http://www.mapreduce.org/what-is-mapreduce.php.

Mejova Y (2009) Sentiment analysis: an overview, pp 1–34. http://www.academia.edu/291678/Sentiment_Analysis_An_Overview.

Murphy KP (2006) Naive Bayes classifiers. University of British Columbia, pp 1–8. http://www.ic.unicamp.br/~rocha/teaching/2011s1/mc906/aulas/naivebayes.pdf

Murphy KP (2012) Machine learning: a probabilistic perspective. In: Chapter 1: Introduction. MIT Press, pp 1–26

Narang RK (2009) Inside the black box. Hoboken, New Jersey

Nuti G, Mirghaemi M, Treleaven P, Yingsaeree C (2011) Algorithmic trading. IEEE Comput 44(11):61–69

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135

SAS Institute Inc (2013) SAS sentiment analysis factsheet. http://www.sas.com/resources/factsheet/sas-sentiment-analysis-factsheet.pdf.

Hootsuite. https://www.websiteplanet.com/blog/guide-twitter-analytics/

Talkwalker. https://www.talkwalker.com/blog/5-free-twitter-analytics-tools-with-views-from-experts

Talkwalker. https://app.talkwalker.com/app/page#/FREE_SEARCH#cid=584f184f-a9f5-46a9-b3e4-02581f5cfa78&co=project&data=eyJyIjp7ImEiOnt9LCJpIjoiRlJFRV9TRUFSQ0giLCJzIjpbeyJhIjp7ImlkIjoibGYwNzYxcnRfZ3EweTJpN2MzZDZ2IiwidG9waWNfaWQiOiJsZjA3NjFydF9ncTB5Mmk3YzNkNncifSwiaSI6IlFVSUNLX1NFQVJDSF9NQUlOIiwicyI6W119XX19

Agorapulse. https://www.agorapulse.com/blog/category/social-media-management/

Buffer Analyze. https://buffer.com/analyze

Buffer Analyze. https://www.fulcrumforge.com/blog/2019/4/19/first-look-at-buffer-analyze-buffers-new-advanced-analytics-and-reporting-tool

BuzzSumo. https://buzzsumo.com/about/

BuzzSumo. https://influencermarketinghub.com/buzzsumo/

Teufl P, Payer U, Lackner G (2010) From NLP (natural language processing) to MLP (machine language processing). In: Kotenko I, Skormin V (eds) Computer network security, Springer, Berlin Heidelberg, pp 256–269

Thomson Reuters (2010). Thomson Reuters news analytics. http://thomsonreuters.com/products/financial-risk/01_255/News_Analytics_-_Product_Brochure-_Oct_2010_1_.pdf.

Thomson Reuters (2012) Thomson Reuters machine readable news. http://thomsonreuters.com/products/financial-risk/01_255/TR_MRN_Overview_10Jan2012.pdf.

Thomson Reuters (2012) Thomson Reuters MarketPsych Indices. http://thomsonreuters.com/products/financial-risk/01_255/TRMI_flyer_2012.pdf.

Thomson Reuters (2012) Thomson Reuters news analytics for internet news and social media. http://thomsonreuters.com/business-unit/financial/eurozone/112408/news_analytics_and_social_media.

Thomson Reuters (2013) Machine readable news. http://thomsonreuters.com/machine-readable-news/?subsector=thomson-reuters-elektron.

Turney PD (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* pp. 417–424

Vaswani V (2011) Hook into Wikipedia information using PHP and the MediaWiki API. http://www.ibm.com/developerworks/web/library/x-phpwikipedia/index.html.

Westerski A (2008) Sentiment analysis: introduction and the state of the art overview. Universidad Politecnica de Madrid, Spain, pp 1–9. http://www.adamwesterski.com/wpcontent/files/docsCursos/sentimentA_doc_TLAW.pdf.

Wikimedia Foundation (2014) Wikipedia:Database download. http://en.wikipedia.org/wiki/Wikipedia:Database_download.

Wolfram SMA (2010) Modelling the stock market using Twitter. Dissertation Master of Science thesis, School of Informatics, University of Edinburgh, pp 1–74. http://homepages.inf.ed.ac.uk/miles/msc-projects/wolfram.pdf.

Yessenov K, Misailovic S (2009) Sentiment analysis of movie review comments, pp 1–17. http://people.csail.mit.edu/kuat/courses/6.863/report.pdf.

Vedurumudi Priyanka (June 13, 2021)Twitter Sentiment Analysis using Deep Learning page 1. https://www.researchgate.net/publication/352780855_Twitter_Sentiment_Analysis_using_Deep_Learning

Wesley Mathew (November 29, 2022) The top 21 Social Media Monitoring Tools for 2023. https://www.meltwater.com/en/blog/top-social-media-monitoring-tools