ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Η ΕΠΑΝΑΣΤΑΣΗ ΤΩΝ BIG DATA: ΑΝΙΧΝΕΥΟΝΤΑΣ ΣΥΝΑΙΣΘΗΜΑΤΑ ΣΤΑ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

Διπλωματική Εργασία

της

Τακτικού Ελένης

Θεσσαλονίκη, Φεβρουάριος 2021

Η ΕΠΑΝΑΣΤΑΣΗ ΤΩΝ BIG DATA: ΑΝΙΧΝΕΥΟΝΤΑΣ ΣΥΝΑΙΣΘΗΜΑΤΑ ΣΤΑ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

Τακτικού Ελένη

Εφαρμοσμένη Πληροφορική

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπων Καθηγητής/τρια
Κωνσταντίνος Ψάννης

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26/2/21

Ονοματεπώνυμο 1	Ονοματεπώνυμο 2	Ονοματεπώνυμο 3
Κωνσταντίνος Ψάννης	Μαντάς Μιχαήλ	Φούσκας Κωνσταντίνος
.....	.....	.....

Τακτικού Ελένη

.....

Περίληψη

Τα Big Data έχουν επηρεάσει τον κόσμο γύρω μας σε πολλούς τομείς. Ένας από αυτούς είναι τα social media. Πλέον οι ενδιαφερόμενοι μπορούν να ανιχνεύουν συναισθήματα μέσω των social media και να ανταποκρίνονται στις επιθυμίες του κοινού τους πριν καν αυτό συνειδητοποιήσει τις ανάγκες του. Τα εργαλεία και οι τεχνικές στο συγκεκριμένο τομέα είναι πολλά και ανά τα χρόνια συνεχώς εξελίσσονται. Η μηχανική μάθηση και οι αλγόριθμοί της χρησιμοποιούνται κατά βάση. Για την ανάλυση όμως των αναρίθμητων σχολίων και αλληλεπιδράσεων στα social media δεν αρκεί μόνο η συλλογή και μελέτη τους. Ένα πολύ σημαντικό στοιχείο που δυσκολεύει την διαδικασία είναι ο εντοπισμός και ανάλυση των συναισθημάτων που περιέχουν. Η ανάλυση συναισθημάτων στα social media παρέχει πληροφορίες για το πώς αισθάνονται οι άνθρωποι για μια επιχείρηση, επωνυμία,τάση ή γεγονός στο διαδίκτυο. Αντί για ένα απλό πλήθος αναφορών ή σχολίων, η ανάλυση συναισθημάτων λαμβάνει υπόψη τα συναισθήματα και τις απόψεις. Στην παρούσα εργασία παρουσιάζονται κάποια από τα πιο γνωστά εργαλεία και τεχνικές που χρησιμοποιούνται. Τέλος, ένα εργαλείο που χρησιμοποιείται κατά κόρον στην συλλογή και ανάλυση των big data είναι το Apache Spark και επιλέχθηκε να παρουσιαστεί στην εργασία, μιας και επεξεργάζεται τα δεδομένα σε πραγματικό χρόνο. Σαν μελέτη περίπτωσης χρησιμοποιούνται σχόλια χρηστών στο Twitter

Λέξεις-Κλειδιά

Big data, διαδίκτυο των πραγμάτων, υπολογιστική νέφους, ανάλυση συναισθημάτων, μηχανική μάθηση, spark, spark streaming, twitter

Abstract

Big Data has affected the world around us in many fields. In Social media, nowadays, marketers and businesses can detect emotions and respond to the needs of their audience before it even realizes them due to Big Data Analysis. The tools and techniques in this field are countless and are constantly evolving over the years. Machine learning and its algorithms are usually used. However, for the analysis of the innumerable comments and interactions on social media, to collect and study them is not enough. A very important aspect that complicates the process is the identification and analysis of the emotions of these comments. Sentiment analysis on social media provides information on how people feel about a business, brand, trend or hot internet topic. Instead of just a few comments, sentiment analysis takes into consideration emotions and opinions. This paper presents some of the most common tools and methods used for this purpose. Finally, one tool that is widely used in the collection and analysis of big data is Apache Spark. In this paper you choose to present the Spark Streaming, a context of Apache Spark which can deal with big data in real time. As a case study, we use twitter comments.

Keywords

Big data, internet of things, cloud computing, sentiment analysis, machine learning, spark, spark streaming, twitter

Πρόλογος – Ευχαριστίες

Η παρούσα διπλωματική εκπονήθηκε το 2020 στα πλαίσια του μεταπτυχιακού προγράμματος σπουδών της Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.. Αντικείμενο της εργασίας αποτελεί η βιβλιογραφική ανασκόπηση των τεχνικών και εργαλείων που περικλείονται στο πλαίσιο των Big Data και η εφαρμογή σε παραδειγμάτα ανάλυσης δεδομένων από το Twitter σε πραγματικό χρόνο με το πρόγραμμα Spark Streaming και την γλώσσα προγραμματισμού Scala.

Ευχαριστώ πολύ τον Επιβλέποντα καθηγητή μου κ. Κωνσταντίνο Ψάννη για την πολύτιμη βοήθεια και καθοδήγησή του στην εκπόνηση της παρούσας εργασίας. Επίσης, ευχαριστώ τον κ.Στεργίου Χρήστο, υποψήφιο Διδάκτορα, για την βοήθεια που μου προσέφερε στο ερευνητικό κομμάτι. Χωρίς τη συμπαράσταση και συνεχή βοήθειά τους, η ολοκλήρωση αυτής της εργασίας δεν θα ήταν δυνατή.

Περιεχόμενα

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΕΦΑΛΑΙΟ 1 : ΕΙΣΑΓΩΓΗ

1.1 Σημαντικότητα θέματος:

Η ανάλυση των Big Data είναι μία από τις πιο σύγχρονες και επιτακτικές διαδικασίες στις μέρες μας. Το διαδίκτυο των πραγμάτων και οι καταιγιστικές εξελίξεις στην τεχνολογία μας κατακλίζουν καθημερινά με όλο και περισσότερα δεδομένα, που αν μελετηθούν σωστά, μπορούν να κάνουν την ζωή μας πολύ πιο εύκολη και εξατομικευμένη. Τι θα λέγατε τώρα, αν ξέρατε πως τα δεδομένα δεν μπορούν να αναλυθούν μόνο ως προς το περιεχόμενο τους αλλά και ως προς τα συναισθήματα που κρύβουν; Σίγουρα κάνουμε λόγο για επανάσταση στον τρόπο έρευνας και μελέτης πληροφοριών που έρχονται από το διαδίκτυο και για έναν τομέα που χρήζει συνεχής μελέτης.

1.2 Σκοπός – Στόχοι:

Η παρούσα εργασία πραγματοποιήθηκε με σκοπό την καταγραφή όλων των στοιχείων που περικλείονται στο πλαίσιο των Big Data. Από σύγχρονες μεθόδους και τεχνικές, μέχρι εργαλεία και υποδομές που χρησιμοποιούνται για την ανάλυση των Big Data κυρίως στα πλαίσια του επιχειρηματικού κόσμου και των social media.

Έπειτα από έρευνα καταλήξαμε πως η φύση των δεδομένων επιτάσσει ταχύτητα, ακρίβεια και πολλούς υπολογιστικούς πόρους. Το τελευταίο από τα παραπάνω επιλύεται με την χρήση της υπολογιστικής νέφους. Για τα άλλα δύο, καταλήξαμε πως σκόπιμη είναι η περεταίρω μελέτη του εργαλείου Apache Spark και συγκεκριμένα της σινιστώσας του Spark Streaming. Πρόκειται για ένα σύγχρονο εργαλείο που αναλύει δεδομένα σε πραγματικό χρόνο με μεγάλη ταχύτητα και ευκολία. Στην παρούσα μελέτη τα δεδομένα αντλούνται από το twitter. Ο προγραμματισμός έγινε με τη χρήση της γλώσσας Scala μιας και είναι πιο εύκολη και σύγχρονη από την python για τον συγκεκριμένο τομέα χρήσης.

Σκοπός είναι η διαπίστωση της ευκολίας χρήσης του προγράμματος και η διερεύνηση, του γιατί πολλές σημαντικές εταιρείες έχουν καταλήξει στη χρήση του αντικαθιστώντας άλλα εργαλεία.

1.3 Συνεισφορά:

Η καταγραφή των τεχνικών και των επιμέρους θεωρητικών στοιχείων έχει ως στόχο να εξυπηρετήσει μελλοντικά ερευνητές στον εντοπισμό των κατάλληλων εργαλείων για την ανάλυση των Big Data.

Επιπλέον , η παρούσα εργασία μπορεί να χρησιμοποιηθεί ως εγχειρίδιο για την εγκατάσταση αλλά και εισαγωγή στο Apache Spark. Επιπλέον παρουσιάζεται όλη η διαδικασία για την σύνδεση με το twitter API και τα διεπιστευτήρια που απαιτούνται για την άντληση στοιχείων. Τα παραδείγματα που παρατίθονται μπορούν να διευρυνθούν και να αποτελέσουν πολύ χρήσιμη βάση για την ανάλυση δεδομένων μέσω του twitter σε πραγματικό χρόνο από tweets.

1.4 Χρήσιμη Ορολογία:

Στο υποκεφάλαιο αυτό αναφέρονται οι βασικοί ορισμοί για ευκολότερη κατανόηση των εννοιών που παρουσιάζονται στα μετέπειτα κεφάλαια.

Big Data:

Τα big data αναφέρονται στα μεγάλα, ποικίλα σύνολα πληροφοριών που αναπτύσσονται με συνεχώς αυξανόμενα ποσοστά. Περιλαμβάνουν τον όγκο των πληροφοριών και την ταχύτητα με την οποία δημιουργούνται και συλλέγονται, αλλά και την ποικιλία ή την εμβέλεια των σημείων δεδομένων που καλύπτουν. Τα big data προέρχονται συχνά από πολλές πηγές και φτάνουν σε πολλαπλές μορφές. *(Troy Segal, Jan 2021)*

Τα big data είναι ένας συνδυασμός δομημένων, ημιδομημένων και μη δομημένων δεδομένων που συλλέγονται από οργανισμούς, και μπορούν να αξιοποιηθούν για πληροφορίες και να χρησιμοποιηθούν σε προγράμματα μηχανικής μάθησης, πρόβλεψης μοντέλων και άλλων προηγμένων εφαρμογών ανάλυσης. *(Bridget Botelho; Stephen J. Bigelow, Oct 2019)*

Μια δημοφιλής ερμηνεία των big data αναφέρεται σε εξαιρετικά μεγάλα σύνολα δεδομένων. Τα big data αποτελούνται από "εκτεταμένα σύνολα δεδομένων - κυρίως στα χαρακτηριστικά όγκου, ταχύτητας και / ή μεταβλητότητας - που απαιτούν κλιμακωτή αρχιτεκτονική για αποτελεσματική αποθήκευση, χειρισμό και ανάλυση". *(University of Wisconsin)*

Τα big data είναι πληροφορίες υψηλού όγκου, υψηλής ταχύτητας και υψηλής ποικιλίας που απαιτούν οικονομικά αποδοτικές, καινοτόμες μορφές επεξεργασίας πληροφοριών για βελτιωμένη διορατικότητα και λήψη αποφάσεων. *(Gartner IT Glossary,n.d.)*

Τα big data είναι ένας όρος που περιγράφει μεγάλους όγκους υψηλής ταχύτητας, σύνθετων και μεταβλητών δεδομένων που απαιτούν προηγμένες τεχνικές και τεχνολογίες για τη σύλληψη, αποθήκευση, διανομή, διαχείριση και ανάλυση τους. *(Tech America Foundation)*

Είναι μια μεγάλη ποσότητα δεδομένων που εξετάζονται χρησιμοποιώντας υπολογισμούς με σκοπό την εύρεση τάσεων ή μοτίβων στα δεδομένα αυτά, ιδίως Αυτών που σχετίζονται με τη δημόσια συμπεριφορά όπως παρατηρείται για παράδειγμα στα κοινωνικά μέσα. Δεδομένα που πληρούν διάφορους περιορισμούς σχετικά με την ταχύτητα, την ποικιλία, τον όγκο και την ακρίβεια μπορούν να χαρακτηριστούν ως big data.

1) ΟΓΚΟΣ: Ο όγκος είναι το μέγεθος/ η ποσότητα δεδομένων που παράγονται, συλλέγονται και αποθηκεύονται. Η ποσότητα δεδομένων διασαφηνίζει εάν τα δεδομένα μπορούν να κληθούν ως big data ή όχι.

2) ΠΟΙΚΙΛΙΑ: Η ποικιλία αναφέρεται στον τύπο ή στη μορφή των δεδομένων. Τα big data μπορούν να ταξινομηθούν ως εικόνα, ήχος, κείμενο ή βίντεο.

3) ΤΑΧΥΤΗΤΑ: Το συγκεκριμένο χαρακτηριστικό αφορά την ταχύτητα με την οποία δημιουργούνται / αναλύονται τα δεδομένα. Τα big data δημιουργούνται κυρίως σε πραγματικό χρόνο, όπως τα δεδομένα στα κοινωνικά δίκτυα που αλλάζουν κάθε δευτερόλεπτο.

4) ΑΚΡΙΒΕΙΑ: Η ακρίβεια αφορά την ποιότητα των δεδομένων. Η ποιότητα των δεδομένων δηλαδή, καθορίζει την ακρίβεια της ανάλυσης που θα γίνει.

(Muhammad Nouman Noor, Farah Haneef ,2020)

Μηχανική Μάθηση:

Η μηχανική μάθηση είναι ο κλάδος της τεχνητής νοημοσύνης στον οποίο η μηχανή ή το σύστημα μπορούν να μάθουν από δεδομένα, να βρουν τα μοτίβα και λαμβάνουν αποφάσεις βάσει αυτής της μάθησης.

Μη επισημασμένα Δεδομένα:

Τα μη επισημασμένα δεδομένα αποτελούνται από δείγματα χωρίς σημαντικές ετικέτες που να σχετίζονται με αυτά και να μπορούν να τα περιγράψουν. Για παράδειγμα, αν μια εικόνα δεν έχει περιγραφή/λεζάντα η μηχανή δε μπορεί να καταλάβει τι απεικονίζει. Συνήθως τα δεδομένα αυτά περιλαμβάνουν φωτογραφίες, βίντεο και ήχους χωρίς λεζάντα ή εξήγηση.

Επισημασμένα Δεδομένα:

Τα επισημασμένα δεδομένα είναι μια συλλογή δεδομένων με ετικέτα με κάποια σημαντική λεζάντα ή πληροφορία με τη μορφή "Κατηγορίας" ή "Ετικέτας (tag)".

Εποπτευόμενη μάθηση:

Η εποπτευόμενη μάθηση είναι μια μέθοδος μηχανικής μάθησης που γίνεται σε δεδομένα εκπαίδευσης που έχουν συσχετιστεί με τάξεις ή ετικέτες (είναι επισημασμένα) και παράγουν αντικειμενική συνάρτηση που ταξινομεί άλλα αόρατα δεδομένα.

Μη εποπτευόμενη μάθηση:

Η μη εποπτευόμενη μάθηση είναι μια μέθοδος μηχανικής μάθησης που γίνεται σε δεδομένα εκπαίδευσης που δεν έχουν τάξεις ή ετικέτες που να συσχετίζονται με αυτά, και σχηματίζουν ομάδες με βάση την ομοιότητα.

Βαθιά Μάθηση:

Η βαθια μηχανική μάθηση είναι μια εξέλιξη της μάθησης μηχανής που βασίζεται σε τεχνητά νευρωνικά δίκτυα και είναι πολύ δημοφιλής για:

- μοντελοποίηση,
- ταξινόμηση
- και αναγνώριση πολύπλοκων δεδομένων όπως εικόνες, ομιλία και κείμενο.

(Muhammad Nouman Noor, Farah Haneef ,2020)

Internet of Things – Διαδίκτυο των πραγμάτων:

Το Διαδίκτυο των πραγμάτων (IoT) είναι μια νέα τεχνολογία που αναπτύσσεται ραγδαία στον τομέα των τηλεπικοινωνιών και ειδικά στον σύγχρονο τομέα των ασύρματων τηλεπικοινωνιών. Ο κύριος στόχος της τεχνολογίας αυτής είναι η αλληλεπίδραση και η συνεργασία μεταξύ αντικειμένων μέσω των ασύρματων δικτύων.

Το IoT αποτελείται από τρία κύρια μέρη:

1. τα «πράγματα» (αντικείμενα).
2. τα δίκτυα επικοινωνίας που τα συνδέουν
3. τα συστήματα υπολογιστών που χρησιμοποιούν ροή δεδομένων από και προς τα αντικείμενα.

(Christos Stergiou Kostas E. Psannis, 2016)

Υπολογιστική Νέφους:

Το Cloud computing ή αλλιώς Υπολογιστική Νέφους ως υποδομή παρέχει υπολογισμούς, αποθήκευση, υπηρεσίες και εφαρμογές μέσω Διαδικτύου. Σε γενικές γραμμές, για να καταστούν τα smartphone ενεργειακά αποδοτικά και υπολογιστικά ικανά, απαιτούνται σημαντικές αλλαγές στο επίπεδο υλικού και λογισμικού τους. Στην βιβλιογραφία αναφέρεται και ως Κινητή Υπολογιστική Νέφους. *(Andreas Plageras; Kostas E. Psannis, 2017)*

Κοινωνικά δίκτυα – Social Media:

Ένα κοινωνικό δίκτυο είναι μια δομή που αποτελείται από σύνολα κοινωνικών, δυαδικών δεσμών και άλλων κοινωνικών αλληλεπιδράσεων μεταξύ ανθρώπων. Η προοπτική των κοινωνικών δικτύων προσφέρει ένα σύνολο μεθόδων για την ανάλυση της δομής ολόκληρων κοινωνικών οντοτήτων, καθώς και μια ποικιλία θεωριών που εξηγούν τα πρότυπα που παρατηρούνται σε αυτές τις δομές.*(Christos Stergiou, Kostas E. Psannis, Andreas P. Plageras, Theofanis Xifilidis, and B.B. Gupta, 2018)*

ΚΕΦΑΛΑΙΟ 2 : ΟΙ ΤΡΟΠΟΙ ΠΟΥ ΤΑ BIG DATA ΕΠΗΡΕΑΖΟΥΝ ΤΟΝ ΚΟΣΜΟ ΜΑΣ

2.1 Εισαγωγή στην ανάλυση δεδομένων

Παρά τις αναφορές που συναντάμε στα μέσα της δεκαετίας του '90 για τα big data ο όρος έγινε ευρέως γνωστός πολύ πιο πρόσφατα, και συγκεκριμένα το 2011. Πολλοί ορισμοί προσπάθησαν να αποδώσουν τί είναι ενώ άλλοι τί κάνουν και δημιουργούνταν συνεχώς σύγχιση. Ερχόμαστε στο τώρα, και παρατηρούμε πως στον σημερινό ανταγωνιστικό κόσμο, οι αποφάσεις που παίρνονται βάσει δεδομένων διαδραματίζουν πολύ σημαντικό ρόλο. Η ζήτηση για επεξεργασία μεγάλου όγκου δεδομένων είναι αυξημένη γιατι συμβάλει στη μελλοντική πρόβλεψη σε διάφορους τομείς και έτσι δίνει ανταγωνιστικά πλεονεκτήματα. Ουσιαστικά τα big data εξυπηρετούν ακριβώς αυτό το σκοπό.

Ζούμε στην εποχή της πληροφόρησης, όπου bytes δεδομένων παράγονται καθημερινά. Πολλά από τα δεδομένα αυτά είναι ημιδομημένα ή μη δομημένα, και παράγονται με πολύ γρήγορους ρυθμούς. Σύμφωνα με *(Farzana Shaikh ,Afsha Khan, Firdaus Rangrez, Uzma Shaikh, 2018)*, τα big data δεν αφορούν μόνο ένα πολύ μεγάλο κομμάτι δεδομένων αλλά και τον τρόπο που τα αποθηκεύουμε, επεξεργαζόμαστε κλπ.

Το Twitter είναι μια από τις μεγαλύτερες πηγές big data εξαιτίας των πολλών tweets που παρατηρούνται καθημερινά. Μπορεί να είναι μια πολύ πλούσια πηγή για την ανάλυση της κοινής γνώμης, συγχρόνως όμως, πολύ δύσκολη στο να επεξεργαστούν τα δεδομένα με παραδοσιακούς τρόπους γιατί παράγονται σε πραγματικό χρόνο. Μια άνευ προηγουμένου κλίμακα κοινωνικών σχέσεων διαχέεται γενικότερα στα social media, και επηρεάζει τις συμπεριφορές σε δημόσιο επίπεδο και τη δημιουργία γνώσης. Η εξαγωγή πληροφοριών από αυτά τα δεδομένα έχει γίνει μια πολύ διευρυμένη πολυεπιστημονική περιοχή που απαιτεί τη συνεργασία των επιστημονικών εργαλείων. Τα big data εξυπηρετούν το σκοπό αυτό, μιας και έχουν δημιουργηθεί πολλά σύγχρονα εργαλεία και μέθοδοι για την άρτια επεξεργασία τους.

Η ανάλυση της κοινής γνώμης είναι υψίστης σημασίας για την δημιουργία μοντέλων και στρατηγικών marketing αλλά επίσης είναι χρήσιμη πηγή πληροφοριών και σε πολλούς ακόμη τομείς, ακόμα και στον εκπαιδευτικό. Υπάρχει ανάγκη για γρήγορη ανταπόκριση στις ανάγκες του κοινού, και αυτή μπορεί να διασφαλίσει την εμπιστοσύνη και την πίστη του. Ένα ακόμη σημαντικό θέμα είναι η οπτικοποίηση των δεδομένων αυτών για να είναι εύκολα στην κατανόηση, και όπως είπαμε οι παραδοσιακές τεχνικές δεν επαρκούν.

Οι βασικές πρακτικές και τεχνικές ανάλυσης περιλαμβάνουν ανάλυση κοινωνικών δικτύων, συναισθημάτων, τάσεων και συστάσεων συνεργασίας. Το γεγονός ότι αναφερόμαστε σε έναν σχετικά νέο τομέα, και η επιστήμη βρίσκεται σε αρχικό στάδιο επεξεργασίας δεδομένων που παράγονται από ανθρώπους, προκαλεί

την ανάγκη για μια επικαιροποιημένη και κατανοητή ταξινόμηση της σχετικής έρευνας.

2.2 Η κοινωνική επιρροή των big data

Οι πρόσφατες εξελίξεις στα συστήματα πληροφοριών που βασίζονται σε δεδομένα, τοποθετούν την έρευνα των big data και τα εργαλεία ανάλυσης που χρησιμοποιούν οι επιχειρήσεις στον πυρήνα της πληροφορικής και της κοινωνικής επιστήμης. Σήμερα μιλάνε όλοι για τα big data, και όμως το αντίκτυπό τους στα άτομα, τις κοινωνίες και τη ζωή μας γενικότερα, παραμένει υποθετικό παρά πλήρως κατανοητό. (*Farzana Shaikh ,Afsha Khan, Firdaus Rangrez, Uzma Shaikh, 2018)*

Με βάση το άρθρο των *(Miltiades D. Lytras, Anna Visvizi, 2019),* είναι σαφές ότι τα big data και η χρήση τους έχουν ήδη δημιουργήσει πολλές ερωτήσεις, συμπεριλαμβανομένων εκείνων για το πώς τα δεδομένα μπορούν να συλλεχθούν και να χρησιμοποιηθούν με ηθικούς και κοινωνικά ευαίσθητους τρόπους. Η αξία των δεδομένων και το κατά πόσο η χρήση τους είναι ηθική σχετίζεται με τρεις παράγοντες:

- την πρόθεση κοινοποίησης προσωπικών δεδομένων,
- τις ανησυχίες του ατόμου και
- το κοινωνικό αντίκτυπο των big data.

2.3 Big Data, επιχειρήσεις και διαδίκτυο των πραγμάτων

Τα τελευταία χρόνια, σύμφωνα με το άρθρο *(Andreas Plageras; Kostas E. Psannis, 2017)* η εμφάνιση νέων τεχνολογιών και του Διαδικτύουο των πραγμάτων (IoT) οδήγησαν σε μια εκρηκτική αύξηση δεδομένων. Ένας τεράστιος αριθμός δικτυωμένων συσκευών (αισθητήρες, ενεργοποιητές κ.λπ.) σε όλο τον κόσμο συλλέγουν διαφορετικούς τύπους δεδομένων (περιβαλλοντικά, γεωγραφικά, λογιστικά κ.λπ.). Στη συνέχεια, οι συσκευές IoT μεταδίδουν τα συλλεγόμενα δεδομένα ώστε να μπορούν να αποθηκευτούν, να υποβληθούν σε επεξεργασία και να αναλυθούν. Επίσης, έως το 2030 περίπου ένα τρισεκατομμύριο αισθητήρες θα συνδεθούν, θα συλλέγουν και θα μεταφέρουν μεγάλες ποσότητες δεδομένων. Επομένως, η ανάγκη για υιοθέτηση εφαρμογών big data και IoT συνδυαστικά, αλληλεξαρτώνται και πρέπει να αναπτυχθούν από κοινού.

Οι άνθρωποι στην καθημερινή τους ζωή έρχονται σε επαφή με αμέτρητες συσκευές που συνδέονται μεταξύ τους και σχηματίζουν δίκτυο, που μας οδηγεί στο IoT. Τα δεδομένα που αναπτύσσονται είναι τεράστια και η συλλογή και επεξεργασία τους ανεκτίμητης αξίας. Δυστυχώς υπάρχουν προβλήματα και ενδυασμοί σχετικά με την ασφάλεια και την διασφάλιση των προσωπικών δεδομένων.

Αν έρθουμε τώρα στον επιχειρηματικό κόσμο, η πρόσβαση σε άπειρα δεδομένα, χάρις στο Διαδίκτυο των πραγμάτων είναι μεγάλη πρόκληση και ευκαιρία! Η ακρίβεια στην επεξεργασία των big data μπορεί να οδηγήσει σε πιο σίγουρη λήψη αποφάσεων και οι καλύτερες αποφάσεις μπορούν να οδηγήσουν σε μεγαλύτερη λειτουργική αποδοτικότητα, μείωση κόστους και μειωμένο κίνδυνο στη λήψη αποφάσεων. Από το παραπάνω, συνειδητοποιούμε ότι τα big data είναι τώρα εξίσου σημαντικά για τις επιχειρήσεις όσο και για το διαδίκτυο. Αυτό συμβαίνει επειδή περισσότερες πληροφορίες οδηγούν σε πιο ακριβείς αναλύσεις. Το πρόβλημα που παρουσιάζεται εδώ όμως είναι πως υπάρχουν μεν μεγάλες ποσότητες δεδομένων, αλλά δεν γνωρίζουμε αν έχουν κάποια ουσιαστική αξία. *(Christos Stergiou Kostas E. Psannis, 2016)*

Αν θεωρήσουμε πως οι επιχειρήσεις μπορούν να λάβουν δεδομένα από οποιαδήποτε πηγή, επιτυγχάνεται το εξής:

(1) μείωση του κόστους

(2) μείωση του χρόνου

(3) ανάπτυξη νέων προϊόντων και βελτιστοποίηση των προσφορών τους και

(4) λήψη πιο έξυπνων αποφάσεων.

Ο σημερινός ψηφιακός κόσμος παρουσιάζει ένα ζήτημα που δεν μας είχε απασχολήσει ποτέ πριν. Κάθε συσκευή στο σπίτι μας είτε είναι, είτε πρόκειται πολύ σύντομα να είναι συνδεδεμένη στο Διαδίκτυο των Πραγμάτων (IoT), κάτι που σημαίνει πως θα μπορεί να συλλέγει δεδομένα!

Η εισροή των δεδομένων που συλλέγονται επιτρέπει στις επιχειρήσεις να κατανοούν καλύτερα τα πρότυπα συμπεριφοράς και αγορών των πελατών, αλλά τα big data πηγαίνουν ένα βήμα πιο πέρα. Βοηθάνε τους επιστήμονες να χειριστούν παγκόσμια ζητήματα, παρέχοντας παράλληλα στους εμπόρους πληροφορίες που απαιτούνται για τη σωστή λήψη αποφάσεων. Και ενώ η πολυπλοκότητα του θέματος απαιτεί μεγάλη συζήτηση, κυρίως λόγω του θέματος της προστασίας των προσωπικών δεδομένων, το κύριο θέμα των big data είναι μάλλον απλό. Εμείς, ως κοινωνία, απλά δεν έχουμε τα κατάλληλα εργαλεία για να επεξεργαστούμε των μεγάλο τους όγκο σωστά.

2.4 Οι τρόποι που τα big data επηρεάζουν τα Social Media.

Στον ορισμό των social media αναφερθήκαμε και παραπάνω. Πρόκειται για ένα σύνολο αλληλεπιδράσεων και διαπροσωπικών σχέσεων οργανωμένα σε μια πλατφόρμα στο διαδίκτυο. Επιτρέπουν την διεπαφή ανάμεσα στους χρήστες τους π.χ. με σχόλια, φωτογραφίες, μηνύματα κ.α Οι ιστότοποι αυτοί αποτελούν εικονικές κοινότητες όπου οι χρήστες μπορούν να επικοινωνούν και να αναπτύσσουν επαφές.

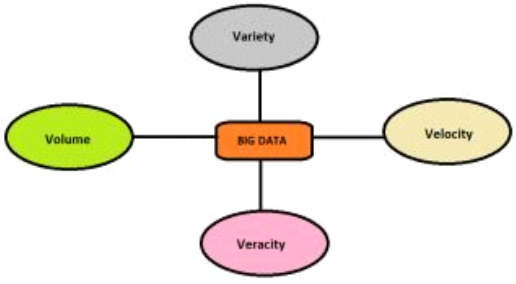
Εικόνα 2.1: Οι πιο γνωστές πλατφόρμες κοινωνικών δικτύων

Μέσα σε όλες αυτές τις αλληλεπιδράσεις χρηστών λοιπόν παράγονται τεράστιοι όγκοι δεδομένων με εξαιρετικό ενδιαφέρον για μελέτη. Οι προκλήσεις που αφορούν δεδομένα σήμερα συχνά κατηγοριοποιούνται ως «Μεγάλες» επειδή ασχολούνται με ένα ή περισσότερα από τα ακόλουθα:

- "Μεγάλος" όγκος,
- Ταχύτητα
- Ποικιλία.
- Ακρίβεια

Οι προκλήσεις της ανάλυσης τέτοιων «μεγάλων δεδομένων» που συζητούνται πιο συχνά είναι η αύξηση του όγκου, η ταχύτητα και η ποικιλία των δεδομένων που παράγεται στα κοινωνικά μέσα αλλά και το πόσο ακριβή είναι τα δεδομένα αυτά. Η αυξανόμενη χρήση των κοινωνικών δικτύων, όπως το Facebook και το Twitter παράγουν τεράστιο όγκος δεδομένων. Το Twitter δημοσιεύει περισσότερα από 500 εκατομμύρια tweets κάθε μέρα. Πολλές επιχειρήσεις προσπαθούν να ανακαλύψουν νέες επιχειρήσεις με διορατικότητα μέσω αυτών των δεδομένων. (*Miltiades D. Lytras, Anna Visvizi, 2019)*

Όπως φαίνεται και στην εικόνα παρακάτω από το άρθρο των *(Muhammad Nouman Noor, Farah Haneef, 2020),* αυτά είναι τα κύρια χαρακτηριστικά που καθορίζουν τα δεδομένα ως big data:



Εικόνα 2.2: Χαρακτηριστικά των Big Data

Όπως αναφέρθηκε και προηγουμένως, η ανάπτυξη του Διαδικτύου έχει κάνει τα κοινωνικά μέσα μέρος της ζωής μας. Αλλά δεν είναι μόνο το γεγονός πως τα social media χρησιμοποιούνται από άτομα για να συνδεθούν με άλλα άτομα, αλλά το σημαντικότερο είναι πως χρησιμοποιούνται και από εταιρείες για να προσεγγίσουν το κοινό τους ή και νέο κοινό. Με την άνοδο των big data, οι τεχνικές marketing έχουν αλλάξει, το ίδιο και το social media marketing. Σε αυτό το πλαίσιο, οι social media marketers έχουν αρχίσει να εξερευνούν big data για να κατανοήσουν τι αρέσει και τι πρέπει να μοιράζονται με τους πελάτες τους, και έτσι σχεδιάζουν τις καμπάνιες τους στα μέσα κοινωνικής δικτύωσης.

Τα social media αποτελούν μία από τις σπουδαιότερες πηγές big data. Κάθε τάση, συμβάν ή είδηση ξεδιπλώνεται στον χώρο τους σε πραγματικό χρόνο προκαλώντας χειμαρρώδη ροή απόψεων. Ιδιαίτερα πρωτοφανές είναι το μέγεθος των δημόσιων δεδομένων κοινωνικής συμπεριφοράς και επικοινωνίας τα οποία ανοίγουν νέους δρόμους ερμηνείας για τον τρόπο λειτουργίας της κοινωνίας. Ένα από τα πιο hot θέματα που αφορούν τα big data για το 2020 είναι τα εργαλεία ανάλυσής τους. Το Analytics παρέχουν ένα ανταγωνιστικό πλεονέκτημα για τις επιχειρήσεις. Η Gartner, Inc πρόβλεψε ότι οι εταιρείες που δεν επένδυσαν σε μεγάλο βαθμό στη συλλογή στοιχείων μέσω ανάλυσης δεδομένων έως τα τέλη του 2020, ενδεχομένως σήμερα να μην είναι πλέον επιχειρήσεις.(μικρές επιχειρήσεις, όπως αυτοαπασχολούμενοι και πολλοί καλλιτέχνες, δεν περιλαμβάνονται σε αυτή τη πρόβλεψη.)

Τα εργαλεία είναι θεμελιώδη στην προσπάθεια των ανθρώπων να απαντήσουν ερευνητικά ερωτήματα μέσω της ανάλυσης δεδομένων. Ωστόσο πενιχρή πρόοδος έχει διατυπωθεί, αφενός διότι ο επιστημονικός τομέας ανάλυσης κοινωνικών big data βρίσκεται σε νηπιακή ηλικία και αφετέρου ανησυχίες σχετιζόμενες με το πλαίσιο των social media όπως η ποιότητά και η ιδιωτικότητα των δεδομένων περιπλέκουν περεταίρω το παρόν εγχείρημα.

Παρόλα αυτά, υπάρχουν πολλά θετικά αντίκτυπα της επίδρασης αυτής και σίγουρα μέσα σε αυτά περιλαμβάνονται:

<u>Ταχύτερες αποφάσεις:</u> Με την ανάλυση big data, οι social media marketers μπορούν να εντοπίσουν τις τελευταίες τάσεις στα κοινωνικά μέσα και να λάβουν αναλόγως καλύτερες και ταχύτερες αποφάσεις. Επίσης, διευκολύνουν την παρακολούθηση των δημογραφικών στοιχείων για να ληφθούν αποφάσεις σχετικά με το ποιά πλατφόρμα κοινωνικών μέσων να στοχεύσουν.

Τα big data βοηθούν τους επαγγελματίες του marketing να κατανοήσουν τις σκέψεις των ανθρώπων σχετικά με ένα brand, το περιεχόμενο που προτιμούν να δουν σε αυτό, αλλά και ποιά είναι η καλύτερη πλατφόρμα για να την προσέγγισή τους.

Οι επιχειρήσεις μπορούν εύκολα να κατανοήσουν τα συναισθήματα των καταναλωτών μέσω big data, επιτρέποντάς τες να αναπτύξουν στρατηγικές νίκης. Αντί να βασίζονται αποκλειστικά σε προηγούμενες επιδόσεις για να διαπιστώσουν ποιες βελτιώσεις απαιτούνται, τα big data βοηθούν στην λήψη τεκμηριωμένων

αποφάσεων για την καλύτερη κάλυψη των μελλοντικών αναγκών και προσδοκιών των καταναλωτών.

Εξατομίκευση: Μέσω της εξέτασης των big data, οι επιχειρηματίες και οι επαγγελματίες του marketing, μπορούν να προσεγγίσουν τους πελάτες τους με έναν πιο εξατομικευμένο τρόπο με βάση τις επιθυμίες τους. Σε αυτό το πλαίσιο, μπορούν να δημιουργήσουν εξατομικευμένες διαφημίσεις που λαμβάνουν υπόψη αυτό που αρέσει στους χρήστες. Με τα big data, θα είναι ευκολότερο για τα brands να εμφανίζουν μόνο εκείνες τις διαφημίσεις που ενδιαφέρουν τους καταναλωτές, μετατρέποντας τες σε μια μη ενοχλητική εμπειρία. Με εξατομικευμένες διαφημίσεις, οι επιχειρηματίες θα μπορούν να ενισχύσουν τις σχέσεις τους με τους χρήστες των κοινωνικών μέσων και να τους μετατρέψουν σε πελάτες αφού εντοπίσουν την πιο αποτελεσματική πλατφόρμα, την ώρα και τη μορφή των διαφημίσεών τους.

Πληροφορίες προϊόντος: Τα big data μπορούν να βοηθήσουν στην κατανόηση των μελλοντικών τάσεων και της συμπεριφοράς των καταναλωτών, ώστε να αναπτυχθούν κοινωνικές δράσεις και προϊόντα που ενδέχεται να υπερβαίνουν τις προσδοκίες τους.

Αποτελεσματική αξιολόγηση εκστρατειών κοινωνικών μέσων: Με την ανάλυση big data, οι marketers μπορούν να παρακολουθούν την αποτελεσματικότητα των εκστρατειών τους στα μέσα κοινωνικής δικτύωσης πριν, κατά τη διάρκεια και μετά την δημοσίευσή τους. Αποκτώντας ευαίσθητες πληροφορίες από big data, οι επιχειρήσεις λαμβάνουν μια ιδέα σχετικά με τις ώρες αιχμής των πελατών, τις προτιμήσεις τους, τη συμπεριφορά τους κ.λπ., κάτι που οδηγεί σε αυξημένη αποτελεσματικότητα της εκστρατείας τους. Οι επιχειρηματίες μπορούν να αποκτήσουν σημαντικές πληροφορίες σχετικά με όλη τη διαδικασία από τη στιγμή που ένας πελάτης είδε τη διαφήμιση, μέχρι την αλληλεπίδραση και την μετατροπή (αγορά, επίσκεψη σε site κλπ). Με αυτόν τον τρόπο, μπορούν να ανακαλύψουν τις σταδιακές αλλαγές στην απόδοση επένδυσης (ROI) καθώς και να δημιουργήσουν δοκιμαστικές καμπάνιες προτού τις δημοσιεύσουν. Επίσης, τα αναλυτικά εργαλεία πρόβλεψης δίνουν τη δυνατότητα στις επιχειρήσεις να λαμβάνουν αποφάσεις σχετικά με το πότε πρέπει να σταματήσει η εκστρατεία τους για να αποφευχθούν απώλειες.

Προγραμματισμός μελλοντικών γεγονότων: Μέσω της μελέτης των big data, οι αρμόδιοι μπορούν να καταλάβουν τι λειτούργησε στο παρελθόν και τι όχι, οπότε μπορούν να αποφασίσουν των πιο αποδοτικό σχεδιασμό των μελλοντικών εκστρατειών τους.

Τα big data χρησιμοποιούνται από τους επαγγελματίες του χώρου των social media ως εργαλείο τροφοδότησης των ψηφιακών εκστρατειών προς την επιτυχία. Μέσω της ανάλυσης τους, κατανοούνται καλύτερα οι διαδικτυακές κοινότητες που

διαθέτουν και προβλέπουν τη συμπεριφορά του κοινού τους, ώστε να μπορούν να παρέχουν εξατομικευμένες υπηρεσίες και να επιλύουν γρήγορα οποιοδήποτε πρόβλημα. *(Roberta Nicora, 2019)*

Στόχευση μέσω διαφημίσεων στα Social media

Ο πιο εμφανής τρόπος που τα big data αλλάζουν τις ζωές μας είναι ότι παρέχουν στους επιχειρηματίες τη δυνατότητα να στοχεύσουν ατομικά μέσα σε μία συγκεκριμένη ομάδα ατόμων. Δεδομένου ότι τα δεδομένα είναι διαχωρισμένα μέσω αλγορίθμων μηχανικής μάθησης, οι πλατφόρμες των social media είναι σε θέση να επιτρέπουν στις εταιρείες να απευθύνονται άμεσα σε άτομα βάσει λεπτομερών προδιαγραφών πιο συγκεκριμένων από τη φυλή, την ηλικία, το φύλο ή την κοινωνική θέση.*(Keith D. Foote, 2020)*

Στην πραγματικότητα, σήμερα είμαστε σε θέση να διαφημίσουμε σε κάποιους μια υπηρεσία, μόνο και μόνο επειδή τους αρέσει μια συγκεκριμένη ταινία. Και ενώ τα παραδείγματα στόχευσης που βλέπουμε στα social media είναι γενικά τρομακτικά από άποψη προσωπικών δεδομένων και της εκμετάλλευσης του συστήματος, η αλήθεια είναι ότι η στόχευση δεν είναι αρνητική τακτική marketing.

Στην πραγματικότητα, με τη βοήθεια της τεχνητής νοημοσύνης, τα big data επιτρέπουν στις επιχειρήσεις να συνδεθούν άμεσα με το κοινό τους, επιτρέποντάς τους να αυξήσουν τα κέρδη τους μειώνοντας το κόστος της διαφήμισης. Κατά μία έννοια, όσο περισσότεροι άνθρωποι ενδιαφέρονται για το προϊόν ή την υπηρεσία τους και βλέπουν τη διαφήμιση, τόσο λιγότερα χρήματα μια επιχείρηση δαπανά από τη διαφήμιση σε άτομα που δεν ενδιαφέρονται για το θέμα.
Ταυτόχρονα, η στόχευση βοηθά και για τον πελάτη. Με την εμφάνιση στοχευμένων διαφημίσεων στη ροή των social media, μπορεί να λαμβάνει πληροφορίες σχετικά με προϊόντα ή / και υπηρεσίες που ενδέχεται να τον ενδιαφέρουν. Έτσι οι διαφημίσεις πλέον δεν είναι μια ενοχλητική διαδικασία.

Big data στην ανάλυση των social media:

Σχεδόν κάθε εταιρεία προσπαθεί να πάρει μερίδιο στα κοινωνικά μέσα. Η δημιουργία προφίλ σε πολλές πλατφόρμες κοινωνικών μέσων με την ελπίδα να αυξηθεί το branding είναι πάντα μια καλή ιδέα. Η πρόσβαση σε πολλές μετρήσεις και μετρικές, όπως τα likes, οι αντιδράσεις, οι απαντήσεις και άλλα επιτρέπουν σε κάθε επιχείρηση να κατανοήσει καλύτερα την ίδια τη φύση της αλληλεπίδρασης μεταξύ των πελατών της και του περιεχομένου της. Η ανάλυση των social media παρέχει στις επιχειρήσεις την ευκαιρία να βελτιώσουν το marketing τους και τη στρατηγική τους, παρέχοντας παράλληλα καλύτερη υποστήριξη και διαφάνεια μέσα από σχετικές και έγκαιρες πληροφορίες. Αλλά όπως και με οποιοδήποτε άλλο θέμα που σχετίζεται με big data, χωρίς έναν σωστό τρόπο συγκέντρωσης και ταχείας εξέτασης, η ανάλυση

δεν ανταποκρίνεται στην πραγματικότητα. Γι αυτό και απαιτούνται εξειδικευμένα εργαλεία.

Με τα κατάλληλα εργαλεία ανάλυσης big data για social media μπορούμε να δούμε γρήγορα και εύκολα τις πιο σημαντικές μετρήσεις της απόδοσης ενός brand. Για παράδειγμα, με ένα γράφημα αύξησης του κοινού παρατηρούμε τον αριθμό των νέων likes/follows σε ένα προφίλ κοινωνικών μέσων σε καθημερινή βάση, ενώ διάγραμμα συνολικής αλληλεπίδρασης μας δίνει πρόσβαση σε πληροφορίες σχετικά με τον τρόπο που το κοινό αλληλεπιδρά με κάποιο συγκεκριμένο περιεχόμενο. Οι λεπτομερείς αναλύσεις, όπως τα δημογραφικά στοιχεία, βοηθούν στη δημιουργία καλύτερης εικόνας για το κοινό μιας σελίδας τη συγκεκριμένη στιγμή. Αυτό, από μόνο του, μπορεί να βοηθήσει έναν digital marketer να καθορίσει εάν οι στόχοι και / ή το μήνυμα του brand μιας επιχείρησης πρέπει να προσαρμοστούν ώστε να ταιριάζουν καλύτερα στην υπάρχουσα πελατειακή βάση ή είναι σωστά εξαρχής.

2.5 Το είδους των δεδομένων που μπορούμε να επεξεργαστούμε, οι συνθήκες και οι σύγχρονες προκλήσεις του χώρου

Ο τύπος και η ποσότητα των προσωπικών δεδομένων που μπορεί να επεξεργαστεί μια εταιρεία / οργανισμός εξαρτάται από τον λόγο της επεξεργασίας και την προβλεπόμενη χρήση.

Κανόνες που πρέπει να τηρούνται:

- Τα προσωπικά δεδομένα πρέπει να υποβάλλονται σε επεξεργασία με νόμιμο δίκαιο και διαφανή τρόπο.
- Πρέπει να υπάρχουν συγκεκριμένοι σκοποί για την επεξεργασία των δεδομένων και η εταιρεία / οργανισμός πρέπει να αναφέρει τους σκοπούς αυτούς στα άτομα κατά τη συλλογή των προσωπικών τους δεδομένων. Μια εταιρεία / οργανισμός δεν μπορεί απλώς να συλλέξει προσωπικά δεδομένα για απροσδιόριστους σκοπούς.
- Πρέπει να συλλέγονται και να επεξεργάζονται μόνο τα προσωπικά δεδομένα που είναι απαραίτητα για την εκπλήρωση του εκάστοτε σκοπού («ελαχιστοποίηση δεδομένων»).
- Πρέπει να διασφαλίζεται ότι τα προσωπικά δεδομένα είναι ακριβή και ενημερωμένα.
- Δεν μπορούν να χρησιμοποιηθούν περαιτέρω τα προσωπικά δεδομένα για άλλους σκοπούς που δεν είναι συμβατοί με τον αρχικό σκοπό και δεν πρέπει να αποθηκεύονται μετά το πέρας της διαδικασίας.

- Πρέπει να χρησιμοποιούνται τεχνικές και οργανωτικές διασφαλίσεις που υπόσχονται ασφάλεια προσωπικών δεδομένων, συμπεριλαμβανομένης της προστασίας από μη εξουσιοδοτημένη ή παράνομη επεξεργασία και από τυχαία απώλεια, καταστροφή ή ζημία, χρησιμοποιώντας την κατάλληλη τεχνολογία. *(European Commission, 2018)*

Υπάρχουν όμως και πολλές προκλήσεις που πρέπει να αντιμετωπιστούν ώστε η ανάλυσή τους να επιφέρει τα σωστά αποτελέσματα. Σύμφωνα με το άρθρο των *(Andreas P. Plageras; Christos Stergiou; George Kokkonis; Kostas E. Psannis; Yutaka Ishibashi; Byung-Gyu Kim; B. Brij Gupta, 2017),* οι προκλήσεις που επιφέρουν τα big data είναι οι εξείς:

- Ο τρόπος παρουσίασής τους
- Η μείωση του πλεονασμού
- Η εύρεση κατάλληλης ποιότητας και ποικιλίας
- Η διαχείριση του κύκλου ζωής τους,
- Το απόρρητο και η ασφάλεια
- Η δυνατότητά τους
- Η διαχείρισή τους
- Η ετερογένεια τους
- Η ταχύτητα που δημιουργούνται
- Η ακρίβειά τους
- Η αποθήκευσή τους,
- Η εξαγόμενη γνώση από αυτά,
- Η δημιουργία και η ανάπτυξη εργαλείων και αλγορίθμων ανάλυσης τους

και πολλά ακόμη στοιχεία που χρειάζονται βελτίωση.

..και τελικά η αξία της ανάλυσης κοινωνικών δικτύων:

Ποιά είναι όμως τελικά το οφέλη της ανάλυσης; Η ανάλυση κοινωνικών δικτύων είναι χρήσιμη για σχεδόν κάθε τομέα για τη λήψη αποφάσεων βάσει των κοινωνικών τάσεων. Οι επιχειρήσεις χρειάζονται αναλυτικά στοιχεία κοινωνικών δικτύων για παρακολούθηση των κριτικών στα προϊόντα τους, οι πολιτικοί για την ενίσχυση της δημόσιας φήμης τους, οι κυβερνήσεις για την ενημέρωση σχετικά με την κοινή γνώμη πάνω σε ορισμένες εκστρατείες και ο τομέας της υγείας για τον έλεγχο της δημοτικότητας των φαρμάκων και της αποτελεσματικότητάς του.

Μερικά από τα κύρια οφέλη χρήσης εργαλείων ανάλυσης στα κοινωνικά δίκτυα *(A Review on big data and Social Network Analytics Techniques)*:

- Καλύτερη αξιολόγηση των αναγκών των καταναλωτών.

- Συμβάλλει στην διαμόρφωση target group και στην ανίχνευση κρυφών πελατών.
- Βοηθά στην ανάλυση της απόκρισης του κοινού σχετικά με τα μέτρα που λαμβάνονται κατά τη διάρκεια της κρίσης.
- Γνώση αντίδρασης κοινού σε νέες ιδέες ή προιόντα.
- Κατά την περίοδο των εκλογών, βοηθά στην πρόβλεψη της πρόθεσης των ψηφοφόρων.
- Παρακολούθηση τάσεων
- Επίβλεψη του πόσο "υγιής" είναι μια επιχείρηση ή ένας οργανισμός
- Προσδιορισμός ατόμων που λειτουργούν ως "influencers" σε άλλα άτομα.
- Εξοικονόμηση χρόνου για τη λήψη κρίσιμων αποφάσεων.
- Παρέχουν συγκριτικό πλεονέκτημα στις επιχειρήσεις έναντι των ανταγωνιστών τους.
- Βοηθά στη βελτίωση της απόδοσης επένδυσης
- Εντοπίζει τις προκλήσεις ανταγωνιστών και κερδίζει τους δυσαρεστημένους πελάτες τους.
- Βοηθά στη δημιουργία αποτελεσματικής στρατηγικής.
- Έγκαιρη λήψη και αξιολόγηση ανατροφοδότησης.
- Πρόβλεψη πορείας ενός προϊόντος ή μιας υπηρεσίας.
- Εύρεση δημοτικότητας συγκεκριμένης προσωπικότητας.
- Καθορίζει εάν ένα προϊόν ή υπηρεσία κερδίζει δημοτικότητα ή αυτή μειώνεται με την πάροδο του χρόνου.
- Εντοπίζει απάτες εταιρειών στον τομέα της ασφάλισης.
- Βοηθά στην ανάλυση εγκλημάτων.
- Βοηθά στην καταπολέμηση της τρομοκρατίας.
- Προσδιορισμός των εξτρεμιστών.
- Έρευνα χρηματοοικονομικής διαφθοράς.
- Παρέχει προτάσεις και ενδιαφέρον περιεχόμενο / προϊόν στο κατάλληλο κοινό.
- Παροχή καθοδήγησης για καριέρα σε μαθητές.

ΚΕΦΑΛΑΙΟ 3 : ΤΟ ΦΑΣΜΑ ΤΩΝ ΕΔΡΑΙΩΜΕΝΩΝ ΠΡΑΚΤΙΚΩΝ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

Τα κοινωνικά μέσα και τα κοινωνικά δίκτυα βοηθούν τους ανθρώπους να εκφέρουν την άποψή τους και να κάνουν κριτικές και ανατροφοδότηση για ένα προϊόν ή μια υπηρεσία, κάτι που επηρεάζει σε σημαντικό βαθμό τις επιχειρήσεις. Χρησιμοποιώντας τέτοια δεδομένα σωστά, οι επιχειρήσεις μπορεί να αυξήσουν τα κέρδη τους. Τα δεδομένα που δημιουργούνται σε ένα κοινωνικό δίκτυο ποικίλουν και δεν μπορούν να αναλυθούν με παραδοσιακές προσεγγίσεις.

3.1 Οι υποδομές που απαιτούνται για την επεξεργασία big data

Αρχικά για την χρήση των big data δύο βασικές υποδομές που εμπλέκονται ώστε να γίνεται με όσο το δυνατόν μεγαλύτερη ασφάλεια και αλλά και ταχύτητα, και φυσικά με μικρότερη χρήση πόρων είναι η λήψη δεδομένων μέσα του Διαδικτύου των Πραγμάτων και με την υποδομή της Υπολογιστικής Νέφους. Η Υπολογιστική Νέφους αναφέρεται σε μια υποδομή στην οποία τόσο η αποθήκευση δεδομένων όσο και η επεξεργασία τους πραγματοποιούνται εκτός της κινητής συσκευής *(C. L. Stergiou; A. P. Plageras; K. E. Psanni, 2018)*. Ή με άλλα λόγια, η τεχνολογία του Cloud Computing αναφέρεται στη δύναμη επεξεργασίας των δεδομένων στην «άκρη» ενός δικτύου (*Christos Stergiou; Kostas E.Psannis; Brij B.Gupta; Yutaka Ishibashi ,2018)*. Η τεχνολογία «Cloud Computing» ή απλά "Cloud" παρέχει καλύτερη χωρητικότητα αποθήκευσης, χαμηλό κόστος, εύκολη επεκτασιμότητα, ευελιξία, αποδοτικότητα, ανθεκτικότητα και αξιοπιστία.

Το Internet of Things (IoT) ή αλλιώς Διαδίκτυο των Πραγμάτων αναδύεται στον τομέα των δικτύων και των τηλεπικοινωνιών με ιδιαίτερη απήχηση στον «σύγχρονο» τομέα των ασύρματων τηλεπικοινωνιακών συστημάτων. Υποστηρίζεται πως το μεγαλύτερο μέρος των Big Data προκύπτει από το διαδίκτυο των πραγμάτων και τις αμέτρητες συνδεδεμένες συσκευές (*Andreas P. Plageras, Kostas E. Psannis, Christos Stergiou, Haoxiang Wang, and B. B. Gupta*). Ό κύριος στόχος της αλληλεπίδρασης και συνεργασίας μεταξύ πραγμάτων και αντικειμένων που αποστέλλονται μέσω των ασύρματων δικτύων είναι να εκπληρώσει τον στόχο που τους έχει οριστεί ως συνδυασμένη οντότητα, δηλαδή η επίτευξη ένος καλύτερου περιβάλλοντος για τη χρήση Big Data (BD). Επιπλέον, με βάση την τεχνολογία των ασύρματων δικτύων, τόσο η Υπολογιστική Νέφους όσο και το IoT θα μπορούσαν να αναπτυχθούν γρήγορα και μαζί. Οι ερευνητές εξέτασαν τις τεχνολογίες IoT και Cloud Computing με έμφαση σε προβλήματα ασφάλειας. Συμπερασματικά, η έρευνα καταλήγει πως και οι δύο υποδομές αυτές συμβάλλον σε μεγάλο βαθμό στην δημιουργία ενός ασφαλούς περιβάλλοντος για την ανάλυση των Big Data.

Επιπλέον μιας και κάνουμε λόγο για Big Data, απαιτείται πολύ μεγάλη χρήση ενέργειας για την επεξεργασία τους. Ένας τρόπος να μειωθούν οι πόροι που απαιτούνται είναι η χρήση μιας Πράσινης Υπολογιστικής Νέφους υποδομής. Οι ερευνητές *(Christos Stergiou; Kostas E. Psannis; Yutaka Ishibashi, 2020)* προτείνουν τη χρήση ενός πράσινου περιβάλλοντος που βασίζεται σε ενοποιημένο δίκτυο νέφους με τη χρήση του CloudSim και καταλήγουν πως είναι ένα πολύ αποδοτικό και φιλικό ως προς το περιβάλλον μοντέλο.

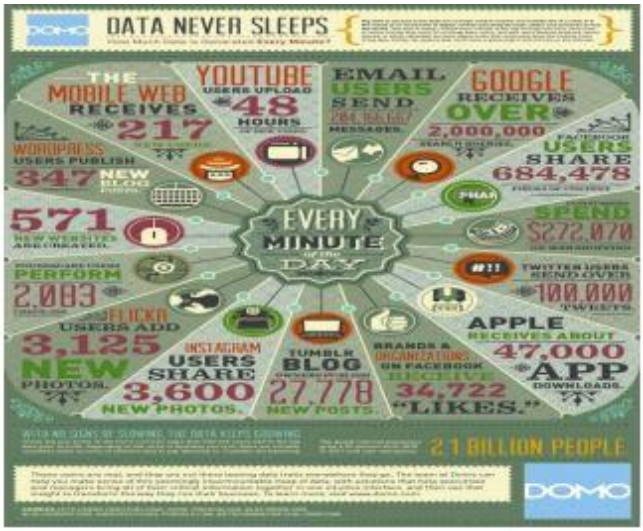
3.2 Τεχνικές που χρησιμοποιούνται για την επεξεργασία big data

Μία από τις πλέον εδραιωμένες τεχνικές που χρησιμοποιούνται για την ανάλυση δεδομένων είναι η Μηχανική Μάθηση. Οι τεχνικές μηχανικής μάθησης βοηθούν τις επιχειρήσεις να ωριμάσουν. Η μηχανική μάθηση είναι ένας τομέας που παρέχει σχήματα για την εξαγωγή σημαντικών μοτίβων από τα δεδομένα.

Οι εταιρείες και οι βιομηχανίες συλλέγουν τεράστιες ποσότητες πρωτογενών δεδομένων, τα οποία διαθέτουν πληροφορίες σε κρυφά επίπεδα. Πλούσιες πηγές δεδομένων βρίσκονται στις κινητές συσκευές που χρησιμοποιούμε καθημερινά, είτε πρόκειται για τηλέφωνα, tablet είτε για αυτοκίνητα. Αυτές οι συσκευές είναι εξοπλισμένες με πληθώρα διαφορετικών αισθητήρων ικανών να παράγουν τεράστιες ποσότητες δεδομένων *(A.Nilsson, S.Smith, G.Ulm, E.Gustavsson, M. Jirstrand, 2018).* Η μορφή, το μέγεθος, η ποικιλία και η ταχύτητα που αυτά παράγονται δημιουργούν δυσκολίες ώστε να χρησιμοποιηθούν αποδοτικά από τις εταιρείες. Έτσι, είναι επιτακτική η ανάγκη ανάπτυξης προηγμένων εργαλείων και τεχνικών για να ξεπεραστούν οι δυσκολίες διαχείρισης ακατέργαστων δεδομένων. Για την ανάλυση big data δηλαδή χρησιμοποιούνται τεχνικές μηχανικής μάθησης και μέθοδοι βαθιάς μηχανικής μάθησης *(Muhammad Nouman Noor, Farah Haneef , 2020)* .

Με τη χρήση προηγμένων αναλυτικών εργαλείων, αναλύονται big data και γίνονται γνωστές οι σχέσεις που υπάρχουν στα κοινωνικά δίκτυα και χαρακτηρίζουν την κοινωνική συμπεριφορά των ατόμων και των ομάδων. Χρησιμοποιώντας δεδομένα που περιγράφουν τις σχέσεις αυτές, εντοπίζονται "κοινωνικοί ηγέτες" που επηρεάζουν τη συμπεριφορά των άλλων σε κάποιο δίκτυο, όπως και τα άτομα που επηρεάζονται περισσότερο.

Ένα κοινωνικό δίκτυο είναι ένα σύνολο που αποτελείται από κόμβους και τους συνδέσμους μεταξύ κάθε δύο κόμβων. Οι κόμβοι αντιπροσωπεύουν κοινωνικούς φορείς και οι σύνδεσμοι είναι οι σχέσεις. Ο όρος Κοινωνικό Δίκτυο είναι γνωστός σχεδόν σε κάθε άτομο και περιγράφεται καλύτερα ως μια διαδικτυακή υπηρεσία που επιτρέπει τη δημιουργία ατομικού προφίλ για την επικοινωνία και την κοινή χρήση βίντεο, εικόνων, απόψεων και ενδιαφερόντων με τους άλλους χρήστες του δικτύου. Τα κοινωνικά δίκτυα έχουν τεράστια επιτυχία επειδή δίνουν την ευκαιρία σε άτομα να συνδεθούν με άλλα άτομα με παρόμοια ενδιαφέροντα, να μοιράζονται απόψεις και πολλά ακόμη. Είναι ζωτικές πηγές για διαδικτυακές αλληλεπιδράσεις, κοινή χρήση περιεχομένου και αξιολογήσεων. Τα κοινωνικά δίκτυα παρέχουν μια πλατφόρμα στους χρήστες για ταχεία ανταλλαγή πληροφοριών ανεξάρτητα από την τοποθεσία τους. Ένα κοινωνικό δίκτυο αντιπροσωπεύει κοινωνικές σχέσεις με μορφή γραφήματος με συνδέσμους και κόμβους. Οι οντότητες αντιπροσωπεύονται με τη μορφή κόμβων και οι σχέσεις μεταξύ τους με συνδέσμους. Πιο σημαντικό απ όλα όμως είναι η ανατροφοδότηση που μπορούν να λάβουν οι εταιρείες, οι οργανισμοί και άλλοι ενδιαφερόμενοι σχετικά με το αν αρέσει το περιεχόμενο και τα προϊόντα τους στο κοινό τους. Στην εικόνα παρακάτω βλέπουμε τα εκτιμώμενα δεδομένα που δημιουργούνται σε κοινωνικά δίκτυα κάθε λεπτό σύμφωνα με το άρθρο των *(Muhammad Nouman Noor, Farah Haneef , 2020)*



Εικόνα 3.1: Δεδομένα που δημιουργούνται στα κοινωνικά δίκτυα κάθε δευτερόλεπτο

Η επεξεργασία των big data είναι πολύ διαφορετική από αυτή των συμβατικών δεδομένων. Απαιτεί ειδικές τεχνικές και εργαλεία για την λήψη σωστών και ολοκληρωμένων αποφάσεων.

Η Μηχανική μάθηση είναι μία τέτοια τεχνική που παρέχει στο σύστημα τη δυνατότητα να μαθαίνει αυτόματα μέσω δεδομένων και να εξαγάγει σημαντικές πληροφορίες χωρίς να έχει προγραμματιστεί ρητά γι αυτό. Η εξαγωγή σημαντικών πληροφοριών βασίζεται στην εύρεση προτύπων στα δεδομένα και στη δημιουργία προβλέψεων βάση των δεδομένων αυτών. Στον σύγχρονο κόσμο, οι τεχνικές μηχανικής μάθησης έχουν χρησιμοποιηθεί σε διάφορα πολύπλοκα πεδία όπως η βιολογία, η φαρμακοβιομηχανία, μέσα κοινωνικής δικτύωσης, αστρονομία και πολλά ακόμη για την εύρεση κρυφών πληροφοριών μέσα στα δεδομένα.

Η διαδικασία της μηχανικής μάθησης γίνεται σε διάφορα επίπεδα και με διαφορετικές ρυθμίσεις και μπορεί να είναι ενισχυμένη, με επίβλεψη ή και χωρίς επίβλεψη. Η ενισχυμένη μηχανική μάθηση είναι ένας τομέας στον οποίο η μάθηση γίνεται με την αλληλεπίδραση με το σύστημα, τη χρήση διαφορετικών ενεργειών και τη λήψη των αποτελεσμάτων τους. Στην μη εποπτευόμενη μηχανική μάθηση η γνώση έρχεται από δεδομένα εισόδου που δεν φέρουν ετικέτα. Αντίθετα στην εποπτευόμενη τα δεδομένα έχουν ετικέτα (*Muhammad Nouman Noor, Farah Haneef , 2020*) .

3.3 Σύγχρονα Εργαλεία ανάλυσης

Ο έλεγχος των δεδομένων τόσο των δομημένων όσο και των μη δομημένων άνοιξε το δρόμο προς την ανάλυση κοινωνικών δικτύων. Οι τεχνικές ανάλυσης big data περιλαμβάνουν: Προβλεπτική Ανάλυση, Στατιστική Ανάλυση, Εξόρυξη

Δεδομένων, Οπτικοποίηση Δεδομένων, complex SQL, Τεχνική Διαδικασία Φυσικής Γλώσσας και Τεχνητή Νοημοσύνη.

Τα big data μπορούν να αναλυθούν μέσω ανάλυσης κειμένων, προγνωστικών analytics, εξόρυξης γνώσης και στατιστικών analytics.

Παρακάτω, στους πίνακες φαίνονται τα πιο γνωστά εργαλεία ανάλυσης των Big Data και οι αναλυτικές τεχνικές ανάλογα με το είδος της επεξεργασίας που καλούμαστε να κάνουμε.

Εργαλεία	Περιγραφή
Apache Hadoop	ένα εργαλείο ανοιχτού κώδικα, αξιόπιστο και ισχυρό. Αποτελείται από στοιχεία γνωστά ως Map Reduced Framework & Hadoop Distrbuted File System.
Apache Storm και Apache Spark	χρησιμοποιούνται κυρίως για κατανεμημένους υπολογισμούς σε πραγματικό χρόνο.
Apache Hive	χρησιμοποιείται για ανάλυση, έρευνα και σύνοψη δεδομένων.
Jaql	χρήση για συναρτησιακή επεξεργασία δεδομένων.
Nosql	κυρίως χρησιμοποιείται για ανάκτηση και αποθήκευση δεδομένων.
Hama and Spark Tools	τα εργαλεία Hama και Spark κερδίζουν επίσης δημοτικότητα στην έρευνα ανάλυσης κοινωνικών δικτύων.

Εικόνα 3.2: Πίνακας με τα εργαλεία ανάλυσης big data

Τεχνικές	Περιγραφή
Prescriptive Analytics	Βοηθητικά για την επιλογή της ενέργειας που θα λάβει χώρα.
Predictive Analytics	Προβλεπτικά που σκοπό έχουν την πρόβλεψη του μέλλοντος.
Diagnostic Analytics	Διαγνωστικά, για ανάλυση της προηγούμενης κατάστασης, γιατί συνέβη και πώς θα ξεπεραστεί.
Descriptive Analytics	Περιγραφικά για ανάλυση της τρέχουσας κατάστασης και πρόβλεψη του εγγύς μέλλοντος.

Εικόνα 3.3: Πίνακας με τις αναλυτικές τεχνικές big data

Με τη βοήθεια κατάλληλης αναλυτικής τεχνικής, οι επιχειρήσεις μπορούν να προβλέπουν το μέλλον προκειμένου να κερδίσουν περισσότερα ή να βελτιώσουν τις υπηρεσίες τους. Τα Predictive analytics βοηθούν τις επιχειρήσεις να κάνουν ταχύτερες και καλύτερες αποφάσεις.

Στο άρθρο τους οι ερευνητές *(Muhammad Aslam Jarwar, Rabeeh Abbasi, Mubashar Mushtaq, Onaiza Maqbool, 2017)*, ένα πλαίσιο που προτάθηκε από τους συγγραφείς του και ονομάζεται «CommuniMents», μπορεί να υπολογίσει τα συναισθήματα του κοινού για μια συγκεκριμένη περίσταση. Η CommuniMents χρησιμοποίησε «αυτοματοποιημένη δειγματοληψία χιονοστιβάδας» για την εύρεση των συμμετεχόντων και μετά παίρνει το περιεχόμενο που αυτοί οι συμμετέχοντες-

χρήστες κοινωνικών δικτύων - δημοσίευσαν για την συγκεκριμένη περίπτωση. Στη συνέχεια επεξεργάζεται τις αναρτήσεις/ τα tweets τους και υπολογίζει τα συναισθήματα τους.

Στη συνέχεια, στο άρθρο τους οι ερευνητές *(S. Magesh, 2016)* αναφέρει πως τα κοινωνικά δίκτυα μπορούν να αναλυθούν με 2 προσεγγίσεις. Με την προσέγγιση γραφικής βάσης δεδομένων ή με την προσέγγιση του παραλληλισμού.(graph database approach και parallelization approach). Το επίκεντρο της προσοχής στην προσέγγιση του παραλληλισμού είναι να χωριστεί ένα τεράστιο σύνολο δεδομένων σε μικρά υποσύνολα και να υπολογιστούν τα δεδομένα αυτά ταυτόχρονα μέσω cloud computing. Η προσέγγιση της γραφικής βάσης δεδομένων χρησιμοποιείται κυρίως για δίκτυα όπου η δομή των δεδομένων αντιπροσωπεύει άμεσα το βασικό χαρακτηριστικό του προβλήματος. Τα παρακείμενα στοιχεία σε αυτές τις βάσεις δεδομένων αλληλοσυνδέονται. Είναι εύκολη η κλιμάκωση τους για τεράστιους όγκους δεδομένων γι αυτό και είναι εξαιρετική επιλογή για την ανάλυση των κοινωνικών δικτύων. Επίσης οι σύνδεσμοι μεταξύ των κόμβων είναι εξαιρετικά χρήσιμοι για την ανάλυση και την δημιουργία συμπερασμάτων. Κάποιες από τις γραφικές βάσεις δεδομένων που χρησιμοποιούνται σε κοινωνικά δίκτυα:
ArrangoDB, Allegrograph, NoSQL, MongoDB, Cassandra, Neo4j.

3.4 Βασική τεχνική ανάλυσης - Μηχανική Μάθηση:

Η διαδικτυακή άποψη οποιουδήποτε ατόμου μπορεί να επηρεάσει τη δημόσια εικόνα μιας επωνυμίας και την κερδοφορία της, επομένως οι επιχειρήσεις ενδιαφέρονται και ανησυχούν για το τι λένε οι χρήστες για το προϊόν ή τις υπηρεσίες τους στα κοινωνικά δίκτυα. Όταν χρειαζόμαστε τα δεδομένα για τη λήψη πολύ σημαντικών αποφάσεων, είναι απαραίτητες για την επεξεργασία τους τεχνικές μηχανικής μάθησης.

Η μηχανική μάθηση είναι η διαδικασία που παρέχει στη μηχανή τη δυνατότητα να μάθει τροφοδοτώντας την με δεδομένα ως είσοδο. Ο ερχομός των big data βελτίωσε κατά πολύ τις τεχνικές μηχανικής μάθησης, σε τέτοιο βαθμό που πολύπλοκοι υπολογισμοί μπορούν να εφαρμοστούν ακόμη και σε τεράστιο όγκο δεδομένων.

Η μηχανή μαθαίνει με τη βοήθεια τριών τεχνικών:

- Εποπτευόμενη μάθηση
- Μη εποπτευόμενη μάθηση
- Ενισχυμένη μάθηση

Τα κοινωνικά δίκτυα ως επί το πλείστον αναλύονται με την βοήθεια μη εποπτευόμενης μάθησης γιατί τα δεδομένα τους συνήθως δεν είναι δομημένα. Οι τεχνικές που χρησιμοποιούνται έχουν κυρίως τον σκοπό δημιουργίας προβλέψεων.

Οι τεχνικές μηχανικής μάθησης μπορούν να ταξινομηθούν σε διαφορετικούς τύπους που είναι:

- Ταξινόμηση,
- Γραμμική παλινδρόμηση,
- Λογιστική Παλινδρόμηση
- Νευρωνικά Δίκτυα.

Οι τεχνικές/αλγόριθμοι μηχανικής μάθησης που διαδραματίζουν σημαντικό ρόλο στην ανάλυση κοινωνικών δικτύων περιλαμβάνουν τους: Naive Bayes, Learning tree, Μέγιστη μέθοδο εντροπίας, ταξινομητής Nearest Neighbor, Dynamic Language Model classifier, Μηχανή υποστηρικτικού διανύσματος, NPL, γραμμική παλινδρόμηση & λογιστική παλινδρόμηση, Πολυστρωματικές Αντιλήψεις και Bayes Net (*Muhammad Nouman Noor, Farah Haneef , 2020*) .

Ο Naïve Bayes είναι ένας ταξινομητής που χρησιμοποιείται κυρίως για την ταξινόμηση κειμένου με σκοπό να βρει την κατηγορία του εκάστοτε κειμένου/ εγγράφου, π.χ. Για ανίχνευση σεξουαλικού ακατάλληλου περιεχομένου ή ανεπιθύμητων μηνυμάτων. Οι αλγόριθμοι Naïve Bayes, Support Vector Machines και Maximum Entropy Method χρησιμοποιούνται για εποπτευόμενη μάθηση κυρίως και εκεί αποδίδουν καλύτερα. Τα δέντρα αποφάσεων χρησιμοποιούνται για ταξινόμηση κειμένου. Ο ταξινομητής Nearest Neighbor είναι μια εποπτευόμενη μέθοδος μάθησης για την αναγνώριση προτύπων από δεδομένα. Η τεχνική χρησιμοποιείται στην ανάλυση κοινωνικών δικτύων όταν δεν υπάρχουν καθόλου ή υπάρχουν ελάχιστες πληροφορίες σχετικά με τα κοινόχρηστα δεδομένα. Ο Support Vector Machine (SVM) είναι ένας μαθησιακός αλγόριθμος που ασκεί εποπτευόμενη μάθηση (Labeled Data) και αναλύει τα δεδομένα που θα χρησιμοποιηθούν για την ανάλυση παλινδρόμησης και την ταξινόμηση. Ο Language model classifier χρησιμοποιείται για την ταξινόμηση δεδομένων σε δύο κλάσεις και είναι επέκταση του Naïve Bayes καθώς και οι δύο χρησιμοποιούν το ίδιο υποθετικό πλαίσιο. Οι Γραμμικές και οι λογιστικές παλινδρομήσεις χρησιμοποιούνται για προβλέψεις. Η γραμμική παλινδρόμηση βρίσκει τη μοναδική τιμή εξόδου ενώ η λογιστική παλινδρόμηση χρησιμοποιεί πιθανότητες για να δείξει την τιμή εξόδου. Στην αρχή για την ανάλυση big data από κοινωνικά δίκτυα χρησιμοποιήθηκε η λογιστική παλινδρόμηση που βασίζεται σε γραφικά μοντέλα. Ο Multilayer Perceptron λειτουργεί ακριβώς όπως ο απλός Perceptron για την ταξινόμηση δεδομένων αλλά έχει πολλαπλά επίπεδα, στα οποία κάθε προηγούμενο επίπεδο συνδέεται πλήρως με το επόμενο.

Στον πίνακα φαίνονται όλοι οι αλγόριθμοι αλλά και η χρήση τους σχετικά με τα κοινωνικά δίκτυα *(Muhammad Nouman Noor, Farah Haneef , 2020)*

Τεχνική	Μέθοδος	Σκοπός Χρήσης
Naive Bayes	Εποπτευόμενη και μη εποπτευόμενη μάθηση	είναι ένας ταξινομητής που χρησιμοποιείται κυρίως για την ταξινόμηση κειμένου με σκοπό να βρει την κατηγορία του εκάστοτε κειμένου/εγγράφου, π.χ. Για ανίχνευση σεξουαλικού ακατάλληλου περιεχομένου ή ανεπιθύμητων μηνυμάτων
Δέντρα Αποφάσεων	Εποπτευόμενη και μη εποπτευόμενη μάθηση	χρησιμοποιούνται για ταξινόμηση κειμένου.
Ταξινομητής - Ο κοντινότερος γείτονας	Εποπτευόμενη μάθηση	χρήση για την αναγνώριση προτύπων από δεδομένα. Η τεχνική χρησιμοποιείται στην ανάλυση κοινωνικών δικτύων όταν δεν υπάρχουν καθόλου ή υπάρχουν ελάχιστες πληροφορίες σχετικά με τα κοινόχρηστα δεδομένα
Μηχανή διανυσμάτων υποστήριξης- SVM	Εποπτευόμενη μάθηση	αναλύει τα δεδομένα που θα χρησιμοποιηθούν για την ανάλυση παλινδρόμησης και την ταξινόμησης
NPL- Αλγόριθμος φυσικής γλώσσας	Εποπτευόμενη και μη εποπτευόμενη μάθηση	χρησιμοποιείται για την ταξινόμηση δεδομένων σε δύο κλάσεις και είναι επέκταση του Naïve Bayes. Και οι δύο χρησιμοποιούν το ίδιο υποθετικό πλαίσιο.
Γραμμική παλινδρόμηση & λογιστική παλινδρόμηση	Εποπτευόμενη μάθηση	χρησιμοποιούνται για προβλέψεις. Η γραμμική παλινδρόμηση βρίσκει τη μοναδική τιμή εξόδου ενώ η λογιστική παλινδρόμηση χρησιμοποιεί πιθανότητες για να δείξει την τιμή εξόδου.
Πολυστρωματικές Αντιλήψεις	Εποπτευόμενη μάθηση	χρησιμοποιείται για την ταξινόμηση δεδομένων αλλά έχει πολλαπλά επίπεδα, στα οποία κάθε προηγούμενο επίπεδο συνδέεται πλήρως με το επόμενο.

Εικόνα 3.4: Αλγόριθμοι Μηχανικής Μάθησης

Τεράστια ποικιλία τεχνικών μηχανικής μάθησης έχουν εφαρμοστεί στο πλαίσιο των κοινωνικών δικτύων, αλλά υπάρχουν και πολλές ακόμη που δεν έχουν εφαρμοστεί και υπάρχει γόνιμο έδαφος για περαιτέρω έρευνα.

3.5 Μπαίνουμε στα βαθιά - Βαθιά μηχανική Μάθηση:

Πολλές φορές οι έννοιες μηχανική μάθηση και βαθιά μηχανική μάθηση μπερδεύονται. Ο πιο εύκολος τρόπος να διαχωριστούν είναι η κατανόηση πως η βαθιά μηχανική μάθηση είναι μηχανική μάθηση. Η βαθια μηχανική μάθηση είναι μια εξέλιξη της μάθησης μηχανής που βασίζεται σε τεχνητά νευρωνικά δίκτυα και είναι πολύ δημοφιλής για μοντελοποίηση, ταξινόμηση και αναγνώριση πολύπλοκων δεδομένων όπως εικόνες, ομιλία και κείμενο. Η άνευ προηγουμένου ακρίβεια της μεθόδου της βαθιάς μάθησης την καθιστούν σε θεμέλια τεχνική για υπηρεσίες τεχνητής νοημοσύνης στο διαδίκτυο. Οι εταιρείες που συλλέγουν δεδομένα χρηστών σε μεγάλη κλίμακα την έχουν υιοθετήση, καθώς η επιτυχία των τεχνικών βαθιάς μάθησης είναι άμεσα ανάλογη με τον αριθμό των διαθέσιμων δεδομένων που μπορούν να μελετηθούν. Η βαθιά μάθηση έχει ενεργοποιήση πολλές πρακτικές εφαρμογές της μηχανικής μάθησης και κατ 'επέκταση της τεχνητής νοημοσύνης.

Η διαφορά της με την μηχανική μάθηση είναι πως οι αλγόριθμοι της βαθιάς μηχανικής μάθησης μπορούν μόνοι τους να καθορίσουν αν μια πρόβλεψη είναι ακριβής ή όχι μέσω του δικού της νευρωνικού δικτύου. Στην μηχανική μάθηση πολλές φορές απαιτείται η ανθρώπινη παρέμβαση αν και τα τελευταία χρόνια έχουν εξελιγχθεί πολύ. Η λογική που λειτουργούν οι αλγόριθμοι βαθιάς μηχανικής μάθησης είναι παρόμοια με αυτή του ανθρώπινου εγκεφάλου, γι αυτό το λόγο χρησιμοποιούν τεχνητά νευρωνικά δίκτυα. Με λίγα λόγια η διαφορά των δύο τεχνικών είναι πως η μηχανική μάθηση αναλύει συνεχώς δεδομένα με σκοπό να μάθει και παίρνει αποφάσεις με βάση αυτό, ενώ η βαθιά μηχανική μάθηση δημιουργεί ένα τεχνητό νευρωνικό δίκτυο το οποίο μπορεί να μάθει και να αποφασίζει αυτόνομα.

3.6 Εργαλεία για ανάλυση κοινωνικών δικτύων:

Σε αυτήν την ενότητα αναφέρονται τα διάφορα διαθέσιμα εργαλεία για την ανάλυση κοινωνικών δικτύων όπως παρουσιάζουν στο άρθρο τους οι (*Muhammad Nouman Noor, Farah Haneef , 2020).* Η επιλογή του σωστού εργαλείου ανάλυσης είναι πολύ σημαντική για τις επιχειρήσεις και εξαρτάται κυρίως από την ευκολία χρήσης του εργαλείου, τα υποστηριζόμενα εργαλεία από τα κοινωνικά δίκτυα, τα υποστηριζόμενα δεδομένα, την ταχύτητα, την ακρίβεια και από το πόσο συχνά η εταιρεία ενημερώνει το λογισμικό της. Πρώτα παρουσιάζονται τα εργαλεία που παρέχονται από τα ίδια τα κοινωνικά δίκτυα και έπειτα τα διαθέσιμα εργαλεία της αγοράς.

FACEBOOK INSIGHTS

Αυτή η πλατφόρμα που παρέχεται από το Facebook μπορεί να χρησιμοποιηθεί για την εμφάνιση πληροφοριών και την προσιτότητα των οργανικών αναρτήσεων αλλά και των χορηγούμενων. Βοηθά επίσης στον έλεγχο των δημογραφικών πληροφοριών σχετικά με το κοινό-στόχο. Το εργαλείο δεν παρέχει ανάλυση συναισθημάτων.

PINTEREST ANALYTICS

Αυτό το εργαλείο παρέχεται από το Pinterest. Βοηθά στη μέτρηση της επισκεψιμότητας του ιστότοπου και επίσης μεταδίδει πληροφορίες σχετικά με το περιεχόμενο που αποθηκεύτηκε στο προφίλ των χρηστών. Αυτή η πλατφόρμα παρέχει δημογραφικές πληροφορίες σχετικά με το κοινό και τα ενδιαφέροντα του. Γίνεται επίσης σύγκριση του κοινού-στόχου με τους γενικούς χρήστες του Pinterest. Αυτό το εργαλείο επίσης δεν παρέχει ανάλυση συναισθημάτων.

TWITTER ANALYTICS

Αυτή η πλατφόρμα παρέχεται από το Twitter, και βοηθά τους χρήστες να αποκτήσουν γνώσεις σχετικά με τα tweets τους ή τις τάσεις για το αν είναι δημοφιλή στο κοινό ή όχι. Αυτό το εργαλείο δίνει επίσης μια ιδέα για τις δημογραφικές πληροφορίες του κοινού που είναι αντίθετο σε κάποιο tweet. Πρόσθετο εργαλείο Analytics που παρέχεται από την πλατφόρμα είναι η παρακολούθηση συνομιλιών και πληροφορίες για θεάσεις βίντεο. Δεν παρέχει δυνατότητα ανταλλαγής μηνυμάτων για στόχευση κοινού.

LINKEDIN ANALYTICS

Τα LinkedIn Analytics μας παρέχουν την ευκαιρία να δούμε δημογραφικά στοιχεία του κοινού σχετικά με την εργασία (αναζήτηση εργασίας, προσόντα, βιογραφικό κλπ). Δείχνει επίσης ενημερώσεις σχετικά με τους επισκέπτες, τις αναρτήσεις και τους ακόλουθους κάθε χρήστη.

INSTAGRAM INSIGHTS

Το Instagram Insights είναι το πιο ευέλικτο ενσωματωμένο εργαλείο ανάλυσης κοινωνικών δικτύων. Παρέχει αναλυτικά στοιχεία από δεδομένα που αυξάνονται κάθε δευτερόλεπτο όπως οι αναρτήσεις, οι πρόσφατες δραστηριότητες, οι ιστορίες, οι επισκέψεις προφίλ ή τα clicks. Αυτό το εργαλείο μπορεί επίσης να ενσωματωθεί στην πλατφόρμα του Pinterest για πιο αποτελεσματική ανάλυση. Μας δίνει επίσης δημογραφικές πληροφορίες από τους ακολούθους κάθε προφιλ.

Τι γίνεται αν το βασικό κοινό είναι ενεργό σε πλατφόρμα χωρίς ενσωματωμένο εργαλείο ανάλυσης; Χρησιμοποιείται κάποιο άλλο από τα διαθέσιμα εργαλεία που κυκλοφορούν στην αγορά.

Κάποια από αυτά τα εργαλεία που δεν είναι ενσωματωμένα παρουσιάζονται παρακάτω:

BOARDREADER

Το BoardReader είναι ένα εργαλείο ανάλυσης κοινωνικών δικτύων που βοηθά να αναλυθούν οι αναρτήσεις και οι κριτικές σε forums, τα οποία αποτελούν μια παλιά μορφή κοινωνικών δικτύων. Αυτό το εργαλείο παρέχει τα γραφήματα με τα οποία μπορούν να συγκριθούν με τον όγκο των συνομιλιών για συγκεκριμένο προϊόν ενάντια σε ανταγωνιστικά προϊόντα, αλλά δεν παρέχει τη δυνατότητα να ελεγχθεί αν η συνομιλία σχετικά με το προϊόν

Είναι θετική ή αρνητική. Είναι χρήσιμο για μικρές επιχειρήσεις, ώστε να πάρουν μια ιδέα σχετικά με τα εργαλεία ανάλυσης κοινωνικών δικτύων. Αυτό το εργαλείο μπορεί επίσης να ενσωματωθεί με άλλα εργαλεία για προηγμένη ανάλυση.

SQUARELOVIN

Το SquareLovin είναι ένα εργαλείο ανάλυσης κοινωνικών δικτύων που εστιάζει στο Instagram. Αναλύει το ρυθμό ανάπτυξης των δημοσιεύσεων, παρέχει μηνιαίο report ανάλυσης και ιστορικό δημοσιεύσεων ανά ώρα, ημέρα, μήνα και έτος. Αυτό το εργαλείο βοηθά επίσης στην επιλογή του σωστού χρόνου δημοσίευσης και παρέχει δεδομένα σχετικά με το στοχευμένο κοινό και το πεδίο των ενδιαφερόντων του.

SUMALL

Είναι ευέλικτο και επιτρέπει τη σύνδεση διαφορετικών κοινωνικών πλατφορμών όπως το Twitter, το Instagram και το Facebook. Βοηθά επίσης στην αυτόματη επικοινωνία με το κοινό.

UNIONMETRICS

Το Union Metrics μπορεί να παρέχει αναλυτικά στοιχεία των Instagram, Facebook και Twitter. Επιτρέπει την παρακολούθηση όλων των θεμάτων και των κοινωνικών λογαριασμών που παρουσιάζουν ενδιαφέρον για την εκάστοτε περίπτωση, και συμβάλλει στην ανάπτυξη στρατηγικών για κοινωνικά μέσα ή λογαριασμούς.

VIZIA

Αυτό το εργαλείο χρησιμοποιείται για την ανάλυση περιεχομένου από βίντεο από κοινωνικά δίκτυα όπως το YouTube. Συμβάλλει στην αύξηση των views και στο να γίνει το περιεχόμενο ενός λογαριασμού viral. Λειτουργεί τέλεια ενσωματωμένο σε άλλες κοινωνικές πλατφόρμες κα επιτρέπει την ανάλυση συναισθημάτων για την προώθηση ή την προώθηση βίντεο στο σωστό κοινό. Βοηθά στη βελτίωση του περιεχομένου παρέχοντας τη δυνατότητα λήψης σχολίων από το κοινό με τη μορφή ερωτηματολογίων ή δημοσκοπήσεων.

WISELYTICS

Προς το παρόν χρησιμοποιείται μόνο στο Facebook. Βοηθάει στην παρακολούθηση των δημοσιεύσεων με την περισσότερη απήχηση αλλά και διάδραση με το κοινό. Επίσης βοηθά στον εντοπισμό και ανάλυση της ομάδας χρηστών που αφήνουν αρνητικά σχόλια.

Στον παρακάτω πίνακα παρατίθενται όλα τα διαθέσιμα εργαλεία, ενσωματωμένα σε πλατφόρμες αλλά και μη με τα βασικά τους χαρακτηριστικά.

Εργαλεία	Υποστηριζόμενες Πλατφόρμες	Περιορισμοί (Λειτουργίες που δεν υποστηρίζουν)	Διαθεσιμότητα
Facebook Insights	Facebook	Sentiment Analytics	Ενσωματωμένο στο Facebook
Pinterest Analytics	Pinterest	Sentiment Analytics	Ενσωματωμένο στο Pinterest
Twitter Analytics	Twitter	Δυνατότητα ανταλλαγής μηνυμάτων με στοχευμένο κοινό	Ενσωματωμένο στο Twitter
LinkedIn Analytics	LinkedIn	Sentiment Analytics	Ενσωματωμένο στο LinkedIn
Instagram Insights	Instagram	Sentiment Analytics	Ενσωματωμένο στο Instagram
BoardReader	Facebook Pinterest LinkedIn Instagram Twitter	Αν τα σχόλια του κοινού είναι θετικά ή αρνητικά	Διαθέσιμο για αγορά
SquareLovin	Instagram	Αναλυτικά στοιχεία του πρώτου μήνα	Διαθέσιμο για αγορά
Sumall	Instagram Twitter Facebook	Λεπτομερής ανάλυση παρέχεται μόνο για το Twitter	Διαθέσιμο για αγορά
UnionMetrics	Instagram Twitter Facebook	Οι πολλαπλοί λογαριασμό ί είναι δύσκολοι στη διαχείριση	Διαθέσιμο για αγορά
Vizia	Youtube	Πολλές αξιολογήσεις απορρίπτονται αυτόματα	Διαθέσιμο για αγορά
Wiselytics	Facebook Twitter	Τμηματοποίηση καμπάνιας	Διαθέσιμο για αγορά

Εικόνα 3.5: Εργαλεία ανάλυσης κοινωνικών δικτύων

3.7 Η συλλογή των δεδομένων πριν την ανάλυση

Υπάρχουν πολλές διαθέσιμες εφαρμογές για τη συλλογή δεδομένων από Facebook, Twitter, Linked-in ή οποιοδήποτε άλλο κοινωνικό δίκτυο όπως το Twitter API και το Facebook API. Το Linked-in παρέχει δύο τύπους εφαρμογών για τη συλλογή δεδομένων. Εφαρμογή τύπου JavaScript και τύπου REST. Χρησιμοποιούνται επίσης συνδρομητικές εφαρμογές συλλογής δεδομένων όπως το Gardenhose και το Firehose για τη λήψη δεδομένων από το twitter. Μετά τη χρήση αυτών των εφαρμογών συνόλων δεδομένων, οι τεχνικές μηχανικής μάθησης όπως ο αλγόριθμος Naïve Bays, τα δέντρα αποφάσεων και η μηχανή διανυσμάτων υποστήριξης, χρησιμοποιούνται για την ανάλυση τους (Muhammad *Nouman Noor, Farah Haneef , 2020).*

ΚΕΦΑΛΑΙΟ 4 ΑΝΑΚΑΛΥΠΤΟΝΤΑΣ ΣΥΝΑΙΣΘΗΜΑΤΑ ΣΤΑ SOCIAL MEDIA

4.1 Εισαγωγή -Αναλύοντας απόψεις και ανιχνεύοντας συναισθήματα

Η λήψη ενός σαρκαστικού σχολίου με emoticons στα κοινωνικά δίκτυα είναι μια διαδικασία που προβληματίζει. Δεν μπορούμε να πούμε με σιγουριά ότι είναι σαρκαστικό. Τι είναι στη πραγματικότητα; Χαρούμενο; Θυμωμένο ή ουδέτερο; Το παραπάνω γεγονός κάνει την ανάλυση συναισθημάτων έναν πολύ ενδιαφέρον αλλά και πολύπλοκο τομέα έρευνας. Στη παρούσα φάση, ο σαρκασμός είναι ένα συναίσθημα που δεν μπορεί να εντοπιστεί από την διαδικασία εφόσον πολλές φορές και εμείς οι άνθρωποι δυσκολευόμαστε να τον εντοπίσουμε. Η ανάλυση συναισθημάτων συχνά αναφέρεται και ως εξόρυξη γνώμης είναι ουσιαστικά η διαδικασία του ορισμού και της κατηγοριοποίησης ενός κειμένου με άποψη ανάλογα με το συναίσθημα που περικλείει *(Adam Coombs, 2017*).

Καθώς η τεχνολογία εξελίσσεται η ανάλυση συναισθημάτων γίνεται ένα από τα χρησιμότερα εργαλεία για τις επιχειρήσεις. Χρησιμοποιείται ευρέως από τις υπηρεσίες ηλεκτρονικού ταχυδρομείου για να κρατήσουν το ανεπιθύμητο περιεχόμενο εκτός των εισερχομένων των χρηστών τους και από τον έλεγχο ιστότοπων για να προτείνει νέο περιεχόμενο όπως ταινίες ή τηλεοπτικές εκπομπές με βάση τις κριτικές που προγενέστερα έχει δώσει κάποιος συγκεκριμένος χρήστης. Ωστόσο, έχει χρησιμοποιηθεί σε πιο σκοτεινές περιστάσεις. Το Facebook, για παράδειγμα, δέχτηκε πυρά όταν ανακαλύφθηκε ότι χρησιμοποίησε ανάλυση συναισθημάτων για να δει αν θα μπορούσε να χειραγωγήσει τα συναισθήματα των ανθρώπων αλλάζοντας τους αλγόριθμους του ώστε να εισάγουν συχνότερα αρνητικές ή θετικές αναρτήσεις στις ροές ειδήσεων των χρηστών τους. Τότε, είχε διαπιστωθεί, παρά το "παράνομο" του πράγματος εφόσον δεν είχαν ενημερωθεί οι συμμετέχοντες για τη συμμετοχή τους στο πείραμα, πως μπορούσε να επηρεάσει αποφασιστικά και υποσυνείδητα τις απόψεις των χρηστών. Τα όρια είναι πολύ λεπτά και η χρήση της ανάλυσης συναισθημάτων μια διαδικασία που δεν πρέπει να τα ξεπερνά *(Galen Panger , 2015).*

Η ανάλυση συναισθημάτων στα μέσα κοινωνικής δικτύωσης παρέχει πληροφορίες για το πώς αισθάνονται οι άνθρωποι για ένα brand στο διαδίκτυο. Αντί για ένα απλό πλήθος αναφορών ή σχολίων, η ανάλυση συναισθημάτων λαμβάνει υπόψη τα συναισθήματα και τις απόψεις. Περιλαμβάνει επίσης, τη συλλογή και ανάλυση πληροφοριών από αναρτήσεις που μοιράζονται οι χρήστες. Η μέτρηση του συναισθήματος είναι πάρα πολύ σημαντική σε κάθε πλάνο ελέγχου των κοινωνικών δικτύων.

Ο υπολογισμός των αναφορών ενός brand στα κοινωνικά δίκτυα δείχνει απλά πόσες φορές οι χρήστες μιλάνε για την επωνυμία αυτή. Τι λένε όμως; Με μια πρώτη ματιά η συγκέντρωση ενός μεγάλου αριθμού αναφορών μπορεί να συνεπάγεται επιτυχία όμως δεν είναι πάντα έτσι.

Σύμφωνα με το άρθρο *(Christina Newberry, 2020)* κάθε επωνυμία χρειάζεται ανάλυση συναισθημάτων γιατί:

- Συμβάλει στην κατανόηση του κοινού: Η κατανόηση του πως αισθάνεται το κοινό για ένα brand, τις αναρτήσεις μια επιχείρησης στα κοινωνικά δίκτυα ή τις καμπάνιες που δημιουργούνται, είναι πολύ πιο ουσιαστικής σημασίας από τη γνώση πως απλά αναφέρεται σε αυτά. Επίσης η άμεση ανάλυση συναισθημάτων κοινωνικών μέσων μπορεί να ειδοποιήσει μια επιχείρηση γρήγορα όταν αλλάξουν οι προτιμήσεις και οι επιθυμίες των πελατών της. Ένα πολύ χαρακτηριστικό παράδειγμα είναι πως σύμφωνα με την ειδική έκθεση της *Edelman Trust Barometer* σχετικά με την εμπιστοσύνη σε ένα brand και την πανδημία Coronavirus διαπίστωσαν ότι το 57% των ανθρώπων ήθελαν από τις εταιρείες να σταματήσουν το μάρκετινγκ που ήταν «χιουμοριστικό ή πολύ ελαφρύ» σχετικά με την κατάσταση.
- Βελτιώνει την εξυπηρέτηση πελατών: Πρώτον, μπορεί να ειδοποιήσει τις ομάδες εξυπηρέτησης και υποστήριξης για τυχόν νέα ζητήματα που πρέπει να γνωρίζουν. Στη συνέχεια, η εταιρεία μπορεί να προετοιμάσει σωστά τη στρατηγική της για απάντηση. Ακόμη μπορεί να υπάρξει ενημέρωση που αφορά κάποιο συγκεκριμένο θέμα προϊόντος. Η παρακολούθηση κοινωνικών αναφορών με αρνητικό συναίσθημα επιτρέπει σε μια επιχείρηση να προσεγγίσει και τα πιο απαιτητικά άτομα. Μια απλή απάντηση ή παρακολούθηση μπορεί να συμβάλει πολύ στην επίλυση των παραπόνων των πελατών.
- Στοχευμένα μηνύματα και ανάπτυξη προϊόντων: Με την πάροδο του χρόνου φαίνεται πως τα μηνύματα μιας επιχείρησης μπορούν να επηρεάσουν το κοινό της για το πως σκέφτεται σχετικά με αυτή. Ακολουθώντας τις τάσεις αυτό είναι όλο και πιο πιθανό. Θα μπορούσε ακόμα και να αποκτηθούν με αυτό τον τρόπο πληροφορίες που θα διαμορφώσουν εκ νέου την στρατηγική μιας εταιρείας.
- Βοηθάει στο να θέτονται ρεαλιστικοί στόχοι: Δεν μπορούν όλα τα brands να ταιριάζουν σε όλους. Η ανάλυση συναισθημάτων μπορεί να βοηθήσει μια επιχείρηση να βρει το κοινό στο οποίο πρέπει να απευθύνεται, να χρησιμοποιεί τα κατάλληλα μηνύματα και την κατάλληλη στιγμή. Επίσης σε ποιον τομέα μια επιχείρηση είναι καλή και που πρέπει να βελτιωθεί. Για παράδειγμα, χρησιμοποιώντας ανάλυση συναισθημάτων σε κοινωνικά δίκτυα, οι ερευνητές διαπίστωσαν ότι το αεροδρόμιο Heathrow είναι γνωστό για καλό wifi, τουαλέτες, εστιατόρια και σαλόνια. Ωστόσο, οι χρήστες των μέσων κοινωνικής δικτύωσης δεν ήταν ευχαριστημένοι με τη στάθμευση του αεροδρομίου, τους χρόνους αναμονής, τις διαδικασίες ελέγχου διαβατηρίων και το προσωπικό. Με αυτές τις γνώσεις, το Heathrow θα μπορούσε να στοχεύει στη βελτίωση των τομέων που οι πελάτες δεν

είναι ικανοποιημένοι. Ή θα μπορούσε να επιλέξει να επικεντρωθεί στους τομείς όπου ήδη λειτουργεί καλά και να προωθείται ως ένα άνετο αεροδρόμιο με ανέσεις.

- Αντιμετωπίζει τις κρίσεις στα αρχικά στάδια: Η ανάλυση συναισθημάτων βοηθά στον εντοπισμό λαθών σε αρχικό στάδιο πριν δημιουργηθεί ανεπανόρθωτο πρόβλημα. Είναι πολύ σημαντικό για τις εταιρείες να ακούσουν τις ανάγκες των πελατών τους. Η ειδική έκθεση της *Edelman* διαπίστωσε ότι μέχρι το τέλος του Μαρτίου, 33% των ερωτηθέντων είχαν ήδη «πείσει άλλους ανθρώπους να σταματήσουν να χρησιμοποιούν μια μάρκα [που] ένιωθαν ότι δεν ενεργούσε κατάλληλα ως απάντηση στην πανδημία». Η παρακολούθηση του κοινωνικού συναισθήματος θα βοηθούσε αυτές τις εταιρείες να διορθώσουν την πορεία τους εγκαίρως για να σταματήσουν τις απώλειες πελατών.

Η χρήση ενός αξιόπιστου εργαλείου μπορεί να κάνει την διαδικασία πιο εύκολη, ακριβή στο αποτέλεσμα και φυσικά γρήγορη. Το πρώτο βήμα είναι να βρεθούν οι συνομιλίες και τοποθετήσεις που κάνουν οι χρήστες για μια επωνυμία στο διαδίκτυο. Η πρόκληση σε αυτό είναι ότι πολλές φορές δεν υπάρχει αναφορά (mention) στην εταιρεία σε πολλά από τα σχόλια που βρίσκονται στο διαδίκτυο *(Christina Newberry, 2020).*

Πως αναλύεται το συναίσθημα από τις αναφορές;

Ερευνώνται οι όροι που μπορεί να υποδηλώνουν συναισθήματα μέσα στην εκάστοτε αναφορά των χρηστών σε μια επωνυμία. Είτε είναι θετικά, είτε αρνητικά φορτισμένοι.

Για παράδειγμα:

- Θετικά φορτισμένες λέξεις: αγαπω, καταπληκτικό, υπέροχο, καλύτερο, τέλειο.
- Αρνητικά φορτισμένες λέξεις: κακό, απαίσιο, τρομερό, χειρότερο, μισώ.

Πιθανότατα θα υπάρχουν άλλοι όροι πιο συγκεκριμένοι για ένα συγκεκριμένο προϊόν, επωνυμία ή κλάδο. Οι επιχειρήσεις μπορούν να δημιουργούν μια λίστα θετικών και αρνητικών λέξεων και να σαρώνουν τις αναφορές που περιέχουν τους όρους αυτούς.

Πρέπει να έχουμε στο μυαλό μας πως πρέπει να μελετάμε τα δεδομένα "εκτος πλαισίου" μιας και κάποιος μπορεί να λέει για μια επωνυμία πως είναι η καλύτερη και να το λέει ειρωνικά.

Μόλις τα δεδομένα συγκεντρώνονται έρχεται η ώρα της ανάλυσης.

Η αναφορά συναισθημάτων των κοινωνικών μέσων πρέπει να περιλαμβάνει τουλάχιστον τα εξής:

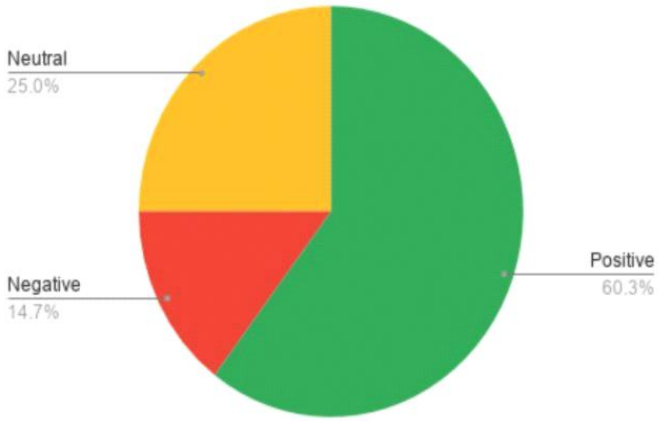
- Το σύνολο των αλληλεπιδράσεων με την επωνυμία σε μια συγκεκριμένη χρονική περίοδο.
- Τις συνολικές αναφορές της επωνυμίας.
- Τον αριθμό ή το ποσοστό των θετικών αναφορών.
- Τον αριθμό ή το ποσοστό των αρνητικών αναφορών
- Υπολογισμό της βαθμολογίας των συναισθημάτων στα κοινωνικά μέσα ως ποσοστό.
- Ένα γράφημα της βαθμολογίας των συναισθημάτων των κοινωνικών μέσων με την πάροδο του χρόνου.

Ο υπολογισμός της βαθμολογίας των συναισθημάτων στα κοινωνικά μέσα γίνεται με 2 τρόπους:

- Θετικές αναφορές ως ποσοστό των συνολικών αναφορών.
- Θετικές αναφορές ως ποσοστό αναφορών που περιλαμβάνουν συναίσθημα (αφαίρεση ουδέτερων αναφορών)

Η 2η μέθοδος έχει πάντα μεγαλύτερο σκορ.

Είναι χρήσιμο να συμπεριλαμβάνεται ένα γραφικό σχήμα απεικόνισης την αναλογία θετικών, ουδέτερων και αρνητικών αναφορών. Για να φαίνονται οι λεπτομέρειες συναισθημάτων με μια ματιά σαν το παρακάτω.



Εικόνα 4.1: Παράδειγμα γραφήματος απεικόνισης αναλογίας θετικών, αρνητικών και ουδέτερων στοιχείων

Όμως πρέπει οι επιχειρήσεις να βλέπουν και τα σχόλια στα οποία δεν γίνεται αναφορά στην επωνυμία τους (που δεν είναι ταγκαρισμένες). Στο Instagram μια καλή πρακτική είναι ο έλεγχος των hashtags που σχετίζονται με την επιχείρηση, τη δράση ή τα προιόντα της. Στο twitter αντίστοιχα hashtags και λέξεις κλειδιά.

Εκτός από το θετικό και αρνητικό συναίσθημα, παρακολουθούνται συγκεκριμένα συναισθήματα, όπως θυμός και χαρά. Αυτό επιτρέπει την αναζήτηση ξαφνικών αλλαγών ή των τρεχουσών τάσεων. Μπορεί επίσης να φιλτράρει το συναίσθημα με βάση την τοποθεσία ή βάση δημογραφικών στοιχείων.

Πως μπορούν να βελτιωθούν τα συναισθήματα που προκαλεί μια επιχείρηση στα κοινωνικά δίκτυα;

- Πρέπει να γνωρίσει το κοινό της σε βάθος. Η ανάλυση των συναισθημάτων βοηθάει αλλά δουλεύει και αμφίδρομα, δηλαδή η γνώση του κοινού βοηθά στο να έχει καλύτερα συναισθηματικά αποτελέσματα και να τα διατηρεί. Όταν υπάρχει γνώση του κοινού η εκάστοτε επιχείρηση μπορεί να δημιουργεί μηνύματα που προκαλούν χαρά και να αποφεύγει τη δημιουργία μηνυμάτων που μπορεί να προκαλέσουν συναισθήματα όπως ο θυμός. Πρέπει να δίνεται βάση στις επιθυμίες του κοινού και στα προβλήματα που αντιμετωπίζουν και να χρησιμοποιείται η ανάλυση συναισθημάτων με σκοπό την λύση τους. Με λίγα λόγια οι επιχειρήσεις πρέπει να ακούν το κοινό τους!
- Να αλληλεπιδρά. Η αλληλεπίδραση στα κοινωνικά δίκτυα χωρίζεται σε δύο κατηγορίες:
    - Αντιδραστική αλληλεπίδραση (Όταν η επιχείρηση απαντά σε μηνύματα σχόλια κλπ)
    - Προοληπτική αλληλεπίδραση(Όταν η επιχείρηση κάνει την πρώτη κίνηση να επικοινωνήσει με το κοινό της)

Κυρίως για την ανάλυση συναισθημάτων χρησιμοποιείται ένας από τους τρεις παρακάτω αλγορίθμους:

- Naive Bayes
- Δέντρα Αποφάσεων
- Μηχανές διανυσμάτων υποστήριξης

Κάθε αλγόριθμος έχει τα πλεονεκτήματα και τα μειονεκτήματα του και σαφώς σε κάθε περίπτωση κάποιος φέρει καλύτερα αποτελέσματα από τους υπόλοιπους. Σε πολλές έρευνες ωστόσο διαπιστώνεται πως ο Naive-bayes είναι ο πιο αποδοτικός στην πλειοψηφία των περιπτώσεων.

Όταν η μέθοδος που χρησιμοποιείται βασίζεται σε λεξικό, υπάρχουν δύο αλγόριθμοι που απαιτείται να χρησιμοποιηθούν. Ο Corpus και ο Dictionary. Πιο ακριβή αποτελέσματα παρέχει ο συνδυασμός τους.

Ο κόσμος μέρα με τη μέρα ψηφιοποιείται. Πολλά δεδομένα δημιουργούνται από τους χρήστες των κοινωνικών μέσων και διαδραματίζουν έναν από τους βασικότερους ρόλους στη λήψη αποφάσεων. Γι' αυτό και από πολλούς τα κοινωνικά μέσα χαρακτηρίζονται ως "θησαυροί πληροφοριών". Πολλές φορές όμως είναι αδύνατο να διαβαστούν ολόκληρα τα κείμενα που δημιουργούνται. Εκεί έρχεται η ανάλυση συναισθημάτων, που καθιστά εύκολη την ανάλυση καθώς παρέχει πολικότητα στο κείμενο και το ταξινομεί ανάλογα με το αν είναι θετικό ή αρνητικό. Η ταξινόμηση μπορεί να εκτελεστεί χρησιμοποιώντας διαφορετικούς αλγόριθμους με αποτέλεσμα διαφορετικό επίπεδο ακρίβειας. Επίσης υπάρχουν πάρα πολλές μέθοδοι για την ανάλυση συναισθημάτων. Παρακάτω θα αναλυθούν και θα συγκριθούν κάποιες από αυτές.

Η ανάλυση συναισθημάτων διαδραματίζει σημαντικό ρόλο στη λήψη αποφάσεων και στα συστήματα προτάσεων. Η λήψη αποφάσεων περιλαμβάνει την αγορά ενός προϊόντος και την πραγματοποίηση μιας επένδυσης. Πάντα οι χρήστες ενδιαφέρονται πριν την αγορά ενός προϊόντος ή πριν κάνουν μια επένδυση, να δουν κριτικές και απόψεις γι αυτό από άλλους χρήστες. Υπάρχει τόσο μεγάλος όγκος σχολίων όμως που είναι αδύνατο να διαβαστούν όλα. Εκεί η ανάλυση συναισθημάτων κάνει την διαδικασία αυτή πολύ πιο εύκολη, βάζοντας θετικό ή αρνητικό πρόσημο στο σύνολο των κριτικών. Μπορεί ο ενδιαφερόμενος να γνωρίζει επίσης αν μια κριτική είναι θετική ή αρνητική χωρίς να την διαβάσει.

Για την ταξινόμηση συναισθημάτων απαιτείται η ακολουθία μιας σειράς βημάτων που είναι η συλλογή δεδομένων, η προεπεξεργασία τους, η εξαγωγή χαρακτηριστικών, η ταξινόμηση και τέλος η αξιολόγηση τους. Τα δεδομένα συλλέγονται από διάφορες πηγές και είναι σε αρχική, ανεπεξέργαστη μορφή. Για να ανακαλυφθεί το συναίσθημα πρέπει τα δεδομένα να είναι σε δομημένη μορφή. Γι αυτό και γίνεται η προεπεξεργασία των δεδομένων. Μετά την προεπεξεργασία, πραγματοποιείται εξαγωγή χαρακτηριστικών. Μόλις το χαρακτηριστικό είναι έτοιμο για εξαγωγή, πρέπει να εκτελεστεί η ταξινόμηση των συναισθημάτων. Για την εκτέλεση της ταξινόμησης μπορούν να χρησιμοποιηθούν διαφορετικές προσεγγίσεις ή μέθοδοι ταξινόμησης συναισθημάτων όπως:

- Μέθοδος που βασίζεται σε λεξικό
- Μηχανική μάθηση
- Υβριδική μέθοδος

4.2 Ανάλυση συναισθημάτων – τα βασικά στοιχεία

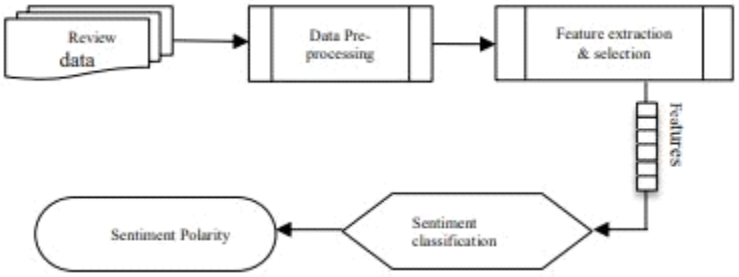
Η ανάλυση συναισθημάτων (SA), η οποία είναι κοινώς γνωστή ως εξόρυξη γνώμης, χρησιμοποιείται στην επεξεργασία φυσικής γλώσσας (NLP), στην υπολογιστική γλωσσολογία, στην ανάλυση κειμένου που βοηθά στον εντοπισμό και σε πολλές άλλες διαδικασίες. Η ανάλυση συναισθημάτων χρησιμοποιείται ευρέως για την ανάδειξη της "φωνής" των καταναλωτών όπως κριτικές και απαντήσεις. Η

ανάλυση συναισθημάτων χρησιμοποιεί τρεις όρους για να καθορίσει το συναίσθημα. Ένας είναι το αντικείμενο για το οποίο γίνεται η ανάλυση, τα χαρακτηριστικά του αντικειμένου και αυτός που δίνει τη γνώμη του για το αντικείμενο. Η ανάλυση συναισθημάτων χειρίζεται διάφορες προκλήσεις όπως η ταυτοποίηση του αντικειμένου, η εξαγωγή χαρακτηριστικών και βρίσκει τον προσανατολισμό της γνώμης.

Όπως αναφέρθηκε και παραπάνω, η ταξινόμηση στην ανάλυση συναισθημάτων γίνεται με τα εξής βήματα:

Επίπεδο εγγράφου
Επίπεδο πρότασης
Επίπεδο χαρακτηριστικών ή επίπεδο άποψης

Στην παρακάτω εικόνα φαίνεται η διαδικασία της ανάλυσης συναισθημάτων βήμα βήμα *(Ankur Goel; Jyoti Gautam; Sitesh Kumar , 2016).*



Εικόνα 4.2: Η διαδικασία της ανάλυσης συναισθημάτων

Το επίπεδο ταξινόμησης του εγγράφου χρησιμοποιείται όταν ο στόχος είναι η εύρεση της συνολικής πολικότητας ενός θέματος ανεξάρτητα από το ποιός είναι ο κάτοχος της γνώμης. Η ανάλυση συναισθημάτων σε επίπεδο εγγράφου προϋποθέτει ότι η γνώμη για τη μεμονωμένη οντότητα/αντικείμενο εκφράζεται από το έγγραφο. Αυτό ισχύει στην περίπτωση κριτικής προϊόντος, κριτικής ταινίας κ.λπ. Όπου το έγγραφο εκφράζει τη γνώμη για μία συγκεκριμένη ταινία ή ένα συγκεκριμένο προϊόν.
Η πρόταση είναι μια συντομότερη μορφή εγγράφου καθώς μια συλλογή προτάσεων δημιουργεί ένα έγγραφο. Η ταξινόμηση στο επίπεδο προτάσεων προϋποθέτει κάθε μία πρόταση να έχει μια ενιαία άποψη. Εδώ η ταξινόμηση περιλαμβάνει δύο δευτερεύουσες εργασίες: ανίχνευση υποκειμενικότητας και ανίχνευση γνώμης.
Στο επίπεδο χαρακτηριστικού ή στο επίπεδο άποψης, η ανάλυση εκτελείται για τα διάφορα χαρακτηριστικά ενός αντικειμένου. Παράδειγμα: Ας υποθέσουμε ότι ένας καταναλωτής αγοράζει ένα κινητό τηλέφωνο Samsung και, στη συνέχεια, παρατηρεί ότι η ποιότητα της κάμερας του κινητού είναι εξαιρετική αλλά η ποιότητα του ήχου δεν είναι. Έτσι, για την ανάλυση, διαφόρων πτυχών του προϊόντος πραγματοποιείται ανάλυση επιπέδου χαρακτηριστικών/απόψεων.

Εξαγωγή χαρακτηριστικών στην ανάλυση συναισθημάτων

Η εξαγωγή χαρακτηριστικών από κείμενο είναι μια πολύ βασική διαδικασία στην ανάλυση συναισθημάτων. Σε αυτήν την τεχνική, το κείμενο πρέπει να μετατραπεί σε διάνυσμα χαρακτηριστικού με τη βοήθεια της προσέγγισης που βασίζεται σε δεδομένα. Μερικά χαρακτηριστικά που χρησιμοποιούνται συχνά στην ανάλυση συναισθημάτων: Παρουσία όρου vs Συχνότητα όρου, Χαρακτηριστικά Ngram, Μέρη ομιλίας, Θέση όρου.

Παρουσία όρου vs Συχνότητα όρου:

Η «Συχνότητα όρου» χρησιμοποιείται για την εύρεση του αριθμού όρων που εμφανίζονται. Η «Παρουσία όρου» είναι στην πραγματικότητα ένα δυαδικό εκτιμημένο διάνυσμα χαρακτηριστικών, το οποίο δείχνει ότι ο όρος υπάρχει στην πρόταση ή όχι. Το 1 αντιπροσωπεύει την παρουσία όρου και το 0 την απουσία του. Οι *(Pang-Lee et al.)* στο άρθρο τους δείχνουν ότι η "παρουσία όρου" είναι πιο σημαντική από την "συχνότητα όρου" στην ανάλυση συναισθημάτων. Επίσης παρατηρείται ότι η εμφάνιση σπάνιων λέξεων περιέχει πιο αξιόπιστες πληροφορίες σε σύγκριση με την εμφάνιση πιο συχνών λέξεων. Το φαινόμενο που χρησιμοποιείται σε αυτήν τη διαδικασία είναι γνωστό ως Hapax Legomena.

Χαρακτηριστικά N-gram:

Τα χαρακτηριστικά N-gram χρησιμοποιούνται ευρέως στον αλγόριθμο NLP (Επεξεργασία Φυσικής Γλώσσας). Ο αριθμός των όρων που εμφανίζονται μαζί σε ένα κείμενο είναι γνωστός ως n-gram. Όταν μόνο ένας όρος λαμβάνεται ως χαρακτηριστικό είναι γνωστός ως unigram, δύο όροι ως bigram. Στο ίδιο άρθρο, οι Pang et.al ανακάλυψαν πως τα unigrams έχουν μεγαλύτερη απόδοση με την πολικότητα συναισθημάτων απ ότι τα bigrams ενώ αντίθετα οι Dave et al. στο “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews” ότι τα bigrams και τα trigrams είναι καλύτερα.

Μέρη Ομιλίας:

Τα ρήματα, τα επίθετα και τα επιρρήματα περιέχουν κυρίως τη γνώμη ενός ατόμου στην Αγγλική γλώσσα. Η προσθήκη ετικετών “Μέρη Ομιλίας” βοηθά στην εύρεση αυτών των λέξεων με ετικέτα σε ένα αρχείο. Τα επίθετα, τα επιρρήματα και τα ρήματα μπορούν να θεωρηθούν ως χαρακτηριστικά και άσχετες λέξεις μπορούν να αφαιρεθούν από το αρχείο ώστε να μειωθεί ο όγκος του.

Αρνητικότητα:

Όταν αρνητικές λέξεις συνοδεύουν μία θετική κριτική αντιστρέφουν την πολικότητα από θετική σε αρνητική. Για παράδειγμα «όχι καλή ταινία» έχει το «καλή» με θετική πολικότητα αλλά το «όχι» την αλλάζει σε αρνητική.

4.3 Αλγόριθμοι και Μέθοδοι Ανάλυσης συναισθημάτων

Οι μέθοδοι ανάλυσης συναισθημάτων είναι είτε βασισμένες στη μηχανική μάθηση, είτε στην ανάλυση με λεξικό, είτε στην υβριδική μέθοδο. Στη μέθοδο μηχανικής μάθησης το σύνολο των δεδομένων με ετικέτα χρησιμοποιείται όπου αναφέρεται ήδη η πολικότητα μιας πρότασης. Από αυτό το σύνολο δεδομένων, εξάγουμε χαρακτηριστικά και αυτά τα χαρακτηριστικά βοηθούν στην ταξινόμηση της πολικότητας της άγνωστης πρότασης εισόδου. Οι μέθοδοι Μηχανικής μάθησης χωρίζονται σε εποπτευόμενη μάθηση και μη εποπτευόμενη μάθηση.

Εποπτευόμενη μάθηση:

Αυτή η προσέγγιση χρησιμοποιείται όταν υπάρχουν διαθέσιμα δεδομένα για την εκπαίδευση του μοντέλου. Δύο βήματα χρησιμοποιούνται στην εποπτευόμενη μάθηση: το πρώτο είναι να εκπαιδεύσει το μοντέλο και το άλλο είναι η πρόβλεψη. Κατά τη διάρκεια της εκπαίδευσης, το σύνολο δεδομένων με τις ετικέτες του τροφοδοτείται στον αλγόριθμο ταξινόμησης που δίνει ένα μοντέλο ως έξοδο. Μετά από αυτό τα δεδομένα δοκιμής τροφοδοτούνται στο μοντέλο για να προβλέψουν την κατηγορία. Υπάρχουν διάφοροι αλγόριθμοι εποπτευόμενης ταξινόμησης:

Naïve Bayes:

Ο Naive Bayes είναι ένας αλγόριθμος ταξινόμησης μηχανικής μάθησης που θέτει μια ανεξάρτητη τιμή για κάθε λειτουργία ή στοιχείο σε ένα σύνολο δεδομένων. Δηλαδή, κάθε στοιχείο αποτιμάται ξεχωριστά για να προσδιοριστεί η πιθανότητα ότι το άθροισμα αυτών των τιμών θα αποτελεί μια προκαθορισμένη ετικέτα ή αποτέλεσμα. Για παράδειγμα, ένα φρούτο μπορεί να θεωρηθεί μήλο εάν είναι κόκκινο, στρογγυλό και περίπου 7 εκατοστά σε διάμετρο. Ακόμα κι αν αυτά τα χαρακτηριστικά εξαρτώνται το ένα από το άλλο, όλες αυτές οι ιδιότητες συμβάλλουν ανεξάρτητα στην πιθανότητα ότι αυτός ο καρπός είναι μήλο. Στην ποσοτικοποίηση των αποτελεσμάτων εμπλέκονται πολλά μαθηματικά.

Ουσιαστικά, θεωρεί κάθε λέξη ανεξάρτητη καθώς δεν λαμβάνει υπόψη τη θέση ενός όρου στην πρόταση. βασίζεται στο θεώρημα Bayes για τον υπολογισμό της πιθανότητας κάθε όρου που αντιστοιχεί σε μια ετικέτα.

$$\text{p(label|features)} = p(label) * p(features|label)\ p(features)$$

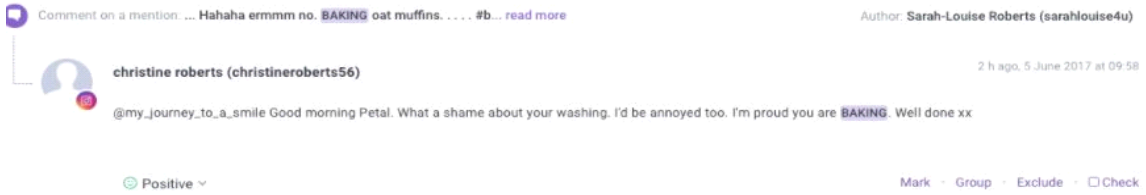
Η μεταβλητή p (label) είναι η προηγούμενη πιθανότητα της ετικέτας στο σύνολο δεδομένων.

Η p(label|features) είναι η προηγούμενη πιθανότητα ενός χαρακτηριστικού που σχετίζεται με μια ετικέτα.
Η $p(features)$ είναι η προηγούμενη πιθανότητα ενός χαρακτηριστικού που έχει συμβεί.

Οι (*Ankur Goel; Jyoti Gautam; Sitesh Kumar, 2016)* χρησιμοποίησαν το SentiWordNet Lexicon σε συνδυασμό με τον αλγόριθμο Naïve Bayes για την βελτίωση της ταξινόμησης συνόλου δεδομένων από το twitter καθώς το SentiWordNet Lexicon παρέχει βαθμολογία θετικών και αρνητικών tweets.

Για την ανάλυση συναισθημάτων στα κοινωνικά δίκτυα, ο Naive Bayes όπως και οι υπόλοιποι αλγόριθμοι ανάλυσης συναισθημάτων πρέπει να αποφασίσουν εάν μια κριτική έχει αρνητική, θετική ή ουδέτερη χροιά. Με τον Naive Bayes αρχικά πρέπει να προκαθοριστεί ένα σύνολο δεδομένων. Στην ανάλυση συναισθηματικών κειμένων, αυτό το σύνολο δεδομένων συνήθως έχει τη μορφή ενός εκπαιδευτικού συνόλου που έχει ήδη ταξινομηθεί σε θετικές ή αρνητικές κατηγορίες. Μια θετική λέξη μπορεί να έχει βαθμολογία +1, ενώ μια αρνητική λέξη θα έχει βαθμολογία -1. Επίσης η βαθμολογία αυτή μπορεί να είναι υψηλότερη σε κάποιες λέξεις σε σχέση με άλλες αν τα συναισθήματα είναι πιο έντονα.

Ένα παράδειγμα φαίνεται παρακάτω από το άρθρο *(Adam Coombs , 2017)*



Εικόνα 4.3: Παράδειγμα ανάλυσης συναισθημάτων α

Χωρίζουμε τη φράση σε λέξεις, και σε αυτές που μπορεί να δοθεί, εκχωρείται θετική ή αρνητική πολικότητα.

Word/Phrase	Positive	Negative
Good morning	✓	
Shame		✓
Annoyed		✓
Proud	✓	
Well done	✓	

Εικόνα 4.3: Παράδειγμα ανάλυσης συναισθημάτων β

Κάθε λέξη εμφανίζεται μόνο μία φορά, έτσι για χάρη του χρόνου, δεν χρειάζεται να δημιουργηθεί πίνακας συχνοτήτων. Εάν αποδοθεί σε κάθε θετική και αρνητική τιμή το «1», τότε μπορεί απλά να διαιρεθούν οι θετικές και αρνητικές λέξεις με τον αριθμό των λέξεων (19) σε ολόκληρη την αναφορά.

Positive words: 3/19 = 0.16

Negative words: 2/19 = 0.11

(P)0.16 – (N)0.11 = +0.05

Εικόνα 4.3: Παράδειγμα ανάλυσης συναισθημάτων γ

Δεδομένου ότι το σύνολο της αναφοράς είναι θετικό, θεωρείται ότι το συναίσθημα της παραπάνω αναφοράς είναι θετικό. Πρόκειται για μια αρκετά ξεκάθαρη περίπτωση. Δεν υπάρχουν λέξεις ουδέτερες, ούτε κάποια με μεγαλύτερη βαθμολογία, είτε θετική είτε αρνητική, ώστε να περιπλέξει το αποτέλεσμα.

Λίγα λόγια και για τους υπόλοιπους αλγορίθμους:

Bayesian Network:

Καθώς ο Naïve Bayes αντιμετωπίζει κάθε λέξη ως ανεξάρτητη, δεν μπορεί να βρει μια σημασιολογική σχέση μεταξύ των λέξεων, ενώ το δίκτυο Bayesian μπορεί. Ουσιαστικά λαμβάνει σοβαρά υπόψη την εξάρτηση των λέξεων μεταξύ τους. Έχει τη μορφή ενός κατευθυνόμενου γραφήματος που δεν είναι κυκλικό, και κάθε κόμβος αντιπροσωπεύει τη λέξη ως μεταβλητή και τα άκρα αντιπροσωπεύουν την εξάρτηση μεταξύ των μεταβλητών αυτών. Σε διάφορες μελέτες όπως και στο άρθρο των *(ohammadAl-Smadi; MahmoudAl Ayyoub; YaserJararweh; OmarQawasmeh ,2019)* φαίνεται πως το δίκτυο Bayesian είναι αρκετά ανταγωνιστικό ως προς την απόδοση και πολλές φορές υπερέχει έναντι άλλων ταξινομητών.

Support Vector Machine (SVM) - Μηχανή διανυσμάτων υποστήριξης:

Η SVM χρησιμοποιήθηκε για πρώτη φορά για την επίλυση των προβλημάτων δυαδικής ταξινόμησης. Επικεντρώνεται στον προσδιορισμό καλύτερων υπερπλάνων που λειτουργούν ως διαχωριστές για να περιγράψουν τα όρια απόφασης μεταξύ σημείων δεδομένων τα οποία είναι σε διαφορετικές κατηγορίες.
Η SVM έχει τη δυνατότητα διαχείρισης γραμμικών και μη γραμμικών εργασιών ταξινόμησης.

Τεχνητό νευρωνικό δίκτυο:

Το τεχνητό νευρικό δίκτυο-Artificial Neural Network(ANN) μιμείται τη δομή του νευρώνα του ανθρώπινου εγκεφάλου. Η βασική μονάδα για το ANN είναι αυτός ο νευρώνας. περιλαμβάνει ένα επίπεδο εισόδου, ένα κρυφό επίπεδο και ένα επίπεδο εξόδου. Ένα διάνυσμα «a (i)» δίνεται ως είσοδος στο δίκτυο και δηλώνει τη συχνότητα μιας λέξης σε ένα έγγραφο. Υπάρχει ένα βάρος «A», που αντιστοιχεί σε κάθε νευρώνα που χρησιμοποιείται για τον υπολογισμό της συνάρτησης. Η

γραμμική συνάρτηση είναι: x (i) =A. (a (i)) Το σύμβολο του x (i) χρησιμοποιείται για την ταξινόμηση της κατηγορίας.

Στα τεχνητά νευρωνικά δίκτυα η εκπαίδευση του μοντέλου αποτελείται από δύο βήματα: διάδοση προς τα εμπρός και προς τα πίσω. Στην προς τα εμπρός διάδοση, η είσοδος δίνεται στο επίπεδο εισόδου των νευρώνων το οποίο πολλαπλασιάζεται με τα βάρη που είναι τυχαίοι αριθμοί.

Οι συναρτήσεις χρησιμοποιούνται για την κανονικοποίηση της τιμής εξόδου μεταξύ 0 και 1. Στη συνέχεια, η έξοδος συγκρίνεται με την τιμή-στόχο και εάν υπάρχει διαφορά (σφάλμα) μεταξύ των δύο τιμών, τότε εκτελείται διάδοση προς τα πίσω. Κατά την διάρκεια της διάδοσης προς τα πίσω η είσοδος πολλαπλασιάζεται με την τιμή σφάλματος έτσι ώστε το βάρος να μπορεί να ρυθμιστεί. Ως εκ τούτου η μάθηση εξαρτάται από το σφάλμα. Οι *(Sachdeva K., Kaur A. & M. Sabharwal*) στο άρθρο τους χρησιμοποίησαν νευρωνικό δίκτυο για ταξινόμηση προσώπων με υψηλό ποσοστό ακριβείας.

Οι (*Vega, L. & Mendez- Vazquez* πρότειναν ένα μοντέλο Δυναμικού Νευρωνικού Δικτύου(DNN) με ανταγωνιστική μάθηση για τη μαθησιακή διαδικασία. Είδαν την απόδοση και κατέληξαν πως είναι πιο αποδοτικό από τις βασικές μεθόδους που χρησιμοποιούνται. Οι *(Patil, S. et aï.)* από την άλλη, στο άρθρο πρότειναν μια τεχνική που χρησιμοποιεί λανθάνουσα σημασιολογική ανάλυση (LSA) με ένα νευρωνικό δίκτυο συνέλιξης (CNN). Η LSA είναι μια τεχνική για τη μετατροπή λέξεων σε διάνυσμα. Η στάθμιση σε LSA πραγματοποιήθηκε με TF-IDF αλγόριθμο. Το μοντέλο αυτό παρέχει 87% ακρίβεια.

Δέντρο Αποφάσεων:

Είναι μια δομή που μοιάζει με δέντρο, όπου οι μη τερματικοί κόμβοι αντιπροσωπεύουν ένα χαρακτηριστικό και ο τερματικός κόμβος αντιπροσωπεύει την ετικέτα. Η διαδρομή ακολουθείται βάση μιας συνθήκης. Είναι μια αναδρομική διαδικασία και τελικά φτάνει σε έναν τερματικό κόμβο ώστε να καταχωρήσει μια ετικέτα σε μια είσοδο. Η κύρια πρόκληση στο δέντρο αποφάσεων είναι να βρεθεί ποιο χαρακτηριστικό πρέπει να επιλεγεί ως κόμβος ρίζας. Ένα δέντρο αποφάσεων είναι μια καλή μέθοδος ανάλυσης συναισθημάτων, διότι παρέχει καλό αποτέλεσμα ακόμα και σε μεγάλο αριθμό δεδομένων.

Το δέντρο αποφάσεων διαιρεί τα δεδομένα εκπαίδευσης ιεραρχικά. Για τη διαίρεση δεδομένων, χρησιμοποιείται μια συνθήκη που βρίσκεται στην τιμή του χαρακτηριστικού. Βασική προϋπόθεση είναι το εάν μια λέξη είναι παρούσα ή απουσιάζει. Η διαδικασία διαίρεσης συνεχίζεται έως ότου οι κόμβοι τερματικού αντιπροσωπεύσουν μικρούς αριθμούς χαρακτηριστικών που χρησιμοποιούνται για την εργασία ταξινόμησης. Οι *(Igor Kotenko; Andrey Chechulin; Dmitry Komashinsky ,2015)* χρησιμοποιούν ένα δέντρο αποφάσεων για να αποκλείσουν το ψευδές περιεχόμενο σε έναν ιστότοπο.

Ταξινομητής βάσει κανόνα - Rule-Based Classifier:

Το μοντέλο ταξινομητή που δημιουργείται με βάση κάποιο κανόνα είναι μια ομάδα από κανόνες. Με βάση αυτούς τους κανόνες καθορίζεται η πρόβλεψη για νέες πληροφορίες. Οι κανόνες είναι πάντα με τη μορφή προγενέστερος και επακόλουθος. Στην αριστερή πλευρά, ο προγενέστερος κανόνας δηλαδή, αντιπροσωπεύει τις συνθήκες, ενώ δεξιά ο επακόλουθος αντιπροσωπεύει την πρόβλεψη κατηγορίας. Παρακάτω φαίνεται ένας τέτοιος κανόνας:

{w1 /\ w2/\w3} {+|-}

Η λέξη σε έναν κανόνα εκφράζει το συναίσθημα όπως φαίνεται παρακάτω.

{Good} {+} {Bad} {-}

Μη εποπτευόμενη μάθηση:

Αυτή η μέθοδος χρησιμοποιείται όταν η αξιοπιστία των επισημασμένων δεδομένων (με ετικέτα) δεν είναι βέβαιη. Η συλλογή δεδομένων χωρίς ετικέτα είναι πιο εύκολη από τη συλλογή δεδομένων με ετικέτα. Η πρόταση κατηγοριοποιείται με βάση κάποια λίστα λέξεων-κλειδιών κάθε κατηγορίας. Για την ανάλυση των δεδομένων που εξαρτώνται από τον τομέα, είναι ευκολότερη η χρήση της προσέγγισης για μη εποπτευόμενη μάθηση. Οι *(Muqtar Unnisa; Ayesha Ameen; Syed Raziuddin, 2016)* πραγματοποίησαν ανάλυση συναισθημάτων χρησιμοποιώντας τη μη εποπτευόμενη προσέγγιση κατά την οποία τα tweets συγκεντρώθηκαν στο θετικό και αρνητικό σύμπλεγμα χρησιμοποιώντας προσέγγιση φασματικής ομαδοποίησης. Η φασματική ομαδοποίηση υπερτερεί των Naïve Bayes, SVM και Maximum Entropy.

Μέθοδος βασισμένη σε λεξικό:

Οι λέξεις που εκφράζουν άποψη είναι οι πιο σημαντικές για την ανάλυση συναισθημάτων. Οι θετικές απόψεις είναι οι επιθυμητές ετικέτες ενώ οι αρνητικές οι ανεπιθύμητες για μια οντότητα, είτε αυτή είναι brand, είτε αντικείμενο, είτε υπηρεσία είτε ακόμα και πρόσωπο. Το Λεξικό είναι μια συλλογή προκαθορισμένων λέξεων όπου η βαθμολογία πολικότητας σχετίζεται με κάθε λέξη. Είναι η ευκολότερη προσέγγιση για την ταξινόμηση συναισθημάτων. Ο ταξινομητής ουσιαστικά χρησιμοποιεί ένα λεξικό και εκτελεί αντιστοίχιση λέξεων για την κατηγοριοποίηση μιας πρότασης. Η απόδοση αυτής της προσέγγισης ταξινόμησης εξαρτάται από το μέγεθος του λεξικού. Υπάρχουν 2 προσεγγίσεις που χρησιμοποιούν την μέθοδο με λεξικό και είναι:

- Μέθοδος Dictionary-Based

Στη μέθοδο αυτή επιλέγονται συγκεκριμένες λέξεις ως λέξεις-σπόροι και αυτές οι λέξεις χρησιμοποιούνται για να βρεθούν συνώνυμα και να διευρυνθεί το μέγεθος του συνόλου λέξεων. Τα διαδικτυακά λεξικά χρησιμοποιούνται για την επέκταση του μεγέθους. Οι λέξεις-σπόροι είναι οι λέξεις άποψης που είναι μοναδικές και σημαντικές για μια οντότητα. Αυτές οι αρχικές λέξεις- σπόροι σε συνδυασμό με τις νέες διευρυμένες λέξεις χρησιμοποιούνται ως χαρακτηριστικά για την εκτέλεση ανάλυσης συναισθημάτων. Υπάρχουν διάφορα λεξικά όπως τα: WordNet, SentiWordNet, SentiFul, SenticNet.

Στο άρθρο τους οι *(Seongik Park, Yanggon Kim, 2016)* πρότειναν μια μέθοδο για την δημιουργία ενός λεξικού - θησαυρού με την προσέγγιση Dictionary-Based. Χρησιμοποίησαν τρία διαδικτυακά λεξικά για την κατασκευή ενός "θησαυρού" και αποθήκευσαν μόνο τις λέξεις που συνυπάρχουν στα 3 λεξικά κάτι που ενίσχυσε την αξιοπιστία του λεξικού-θησαυρού που δημιούργησαν. Η επέκταση του λεξικού γίνεται με συνώνυμα και αντώνυμα των αρχικών λέξεων σπόρων. Για την επιλογή των λέξεων σπόρων χρησιμοποιήθηκε η μέθοδος TF-IDF, και αν και ο θησαυρός που δημιουργήθηκε βελτίωσε κατά πολύ τη διαδικασία της ταξινόμησης, ήταν μια χρονοβόρα διαδικασία.

- Μέθοδος Corpus-Based

Στην μέθοδο corpus - based, δηλαδή στη μέθοδο που δε βασίζεται στο λεξικό αλλά στο "σώμα", σκοπός δεν είναι να βρεθεί μόνο η ετικέτα μιας λέξης αλλά ο προσανατολισμός του γενικότερου πλαισίου. Σε αυτήν την προσέγγιση καταρχάς προετοιμάζεται μια λίστα λέξεων σπόρων και στη συνέχεια χρησιμοποιείται το συντακτικό μοτίβο αυτών των λέξεων για τη δημιουργία νέων υποκειμενικών λέξεων στο σώμα.

Αυτή η προσέγγιση λειτουργεί περαιτέρω με δύο τρόπους:

- Στατιστική προσέγγιση
- Σημασιολογική προσέγγιση

4.4 Ψυχογραφικά χαρακτηριστικά χρηστών και τμηματοποίηση

Έρευνες για το πώς τα big data συμβάλλουν στην ενημέρωση των ψυχολογικών πτυχών των καταναλωτών δεν έχουν λάβει την πρέπουσα σημασία. Τα ψυχογραφικά στοιχεία έχει αποδειχθεί ότι αποτελούν πολύτιμο εργαλείο για την τμηματοποίηση της αγοράς και για την κατανόηση των προτιμήσεων των καταναλωτών *(Hui Liu; Yinghui Huang; Zichao Wang; Kai Liu; Xiangen Hu; Weijun Wang, 2019)* .

Στο πλαίσιο του ηλεκτρονικού εμπορίου, ως συστατικό της ψυχογραφικής κατάτμησης, η προσωπικότητα των καταναλωτών έχει αποδειχθεί αποτελεσματική για την πρόβλεψη των προτιμήσεων τους. Τα big data που προκύπτουν από καταναλωτές υπόσχονται να αλλάξουν το παιχνίδι στον τομέα του marketing.

Ειδικότερα, οι πληροφορίες που αντλούνται από ψυχολογικές πτυχές της συμπεριφοράς των χρηστών, είναι κρίσιμες για την κατανόηση και όχι απλώς την πρόβλεψη των προτιμήσεων τους, και συμβάλλουν τελικά στην έξυπνη λήψη αποφάσεων στον τομέα του marketing. Μελέτες έχουν δείξει ότι οι ψυχολογικές μεταβλητές, όπως οι αξίες και η προσωπικότητα των χρηστών, είναι τα κύρια εργαλεία για την ψυχογραφική τμηματοποίηση τους, ωστόσο, αν και η προγνωστική και επεξηγηματική δύναμη των χαρακτηριστικών της προσωπικότητας για τη διαδικτυακή συμπεριφορά των χρηστών δεν τίθεται αμφισβήτησης, η χρήση τους είναι αμφιλεγόμενη. Επίσης δεν είναι γνωστό κατά πόσο η προγνωστική και επεξηγηματική δύναμη της συναισθηματικής ανάλυσης των δεδομένων διαφέρει από κατηγορία σε κατηγορία και από τομέα σε τομέα.

Ο βασικός όμως λόγος που συλλογή δεδομένων ψυχογραφικής τμηματοποίησης από χρήστες ηλεκτρονικού εμπορίου είναι δύσκολη σε μεγάλη κλίμακα είναι γιατί απαιτεί την συμπλήρωση ερωτηματολογίων από τους ίδιους. Πρακτικά, οι τεχνικές επεξεργασίας φυσικής γλώσσας (NLP) μπορούν να εφαρμοστούν για τον εντοπισμό ψυχογραφικών μεταβλητών, όπως η χρήση λέξεων στο διαδίκτυο από χρήστες ηλεκτρονικού εμπορίου, για την κατανόηση και την πρόβλεψη των συμπεριφορών και των προτιμήσεων τους σε μεγάλη κλίμακα. Οι ιστότοποι ηλεκτρονικού εμπορίου έχουν συσσωρεύσει μεγάλο αριθμό περιεχομένου από μη δομημένα δεδομένα που δημιουργείται από χρήστες, το οποίο παρέχει τη βάση για την παρατήρηση της ψυχοσύνθεσής τους και την πρόβλεψη των προτιμήσεών τους άμεσα.

Πρόκειται για μια πολλά υποσχόμενη μέθοδο, γιατί η χρήση των ψυχογραφικών στοιχείων των καταναλωτών σε συνδυασμό με τα μεγάλα δεδομένα, φαίνεται να δίνει λύση στο πρόβλημα της χρήσης ψυχογραφικών προσεγγίσεων σε μεγάλη κλίμακα.

Η τμηματοποίηση της αγοράς είναι μια τεχνική που χρησιμοποιείται ήδη από το 1956 όταν την πρότεινε ο Smith ως στρατηγική διαφοροποίησης προϊόντων για την αύξηση της ανταγωνιστικότητας. Πλέον, έχει αναγνωριστεί ευρέως ως θεμελιώδες εργαλείο για την κατανόηση των συμπεριφορών των πελατών. Οι *(Wedel και Kamakura, 2012)* ορίζουν την τμηματοποίηση ως ένα σύνολο μεταβλητών ή χαρακτηριστικών που χρησιμοποιούνται για την εκχώρηση δυνητικών πελατών σε ομοιογενείς ομάδες.

Η τμηματοποίηση είναι κρίσιμη για τις επιχειρήσεις επειδή μια εταιρεία έχει περιορισμένους πόρους και πρέπει να επικεντρωθεί στον καλύτερο τρόπο αναγνώρισης και εξυπηρέτησης των πελατών της με όσο το δυνατόν λιγότερη σπατάλη των πόρων αυτών. Αρχικά, η τμηματοποίηση της αγοράς βασίστηκε κυρίως σε προφίλ προσωπικοτήτων. Η πιο συχνά χρησιμοποιούμενη κλίμακα για τη μέτρηση γενικών πτυχών της προσωπικότητας ως τρόπος καθορισμού ομοιογενών

υποκατηγοριών της αγοράς είναι το *Edwards Personal Preference Schedule (EPPS).* Σε γενικές γραμμές όμως, όλες οι μελέτες που βασίστηκαν στο πρόγραμμα αυτό παρατηρείται πως είχαν χαμηλές και ασυνεπείς συσχετίσεις σχετικά με την

Πλέον χρησιμοποιείται το *Big Five Factor* στη θέση του Edwards Personal Preference Schedule (EPPS) με πολύ καλύτερα και ακριβή αποτελέσματα.

Ο τρόπος ζωής ορίζεται ως ένα σύνολο συμπεριφορών που αντικατοπτρίζουν ατομικές ψυχολογικές σκέψεις και κοινωνιολογικές συνέπειες και είναι πιο έγκυρος παράγοντας στη διαμόρφωση της ατομικής συμπεριφοράς σε σύγκριση με τα χαρακτηριστικά της προσωπικότητας. Η δομή του τρόπου ζωής που χρησιμοποιείται στην τμηματοποίηση της αγοράς βασίζεται σε έρευνα σχετικά με τα κίνητρα και τις προσωπικές αξίες. Οι δραστηριότητες, τα ενδιαφέροντα και οι απόψεις (AIO) ήταν από τα πιο ευρέως χρησιμοποιούμενα εργαλεία μέτρησης. Βέβαια, γιατί πολλές φορές τέτοια στοιχεία δεν είναι εύκολο να βρεθούν, χρησιμοποιούνται πιο απλές διαισθητικές αξίες για την σκιαγράφηση του τρόπου ζωής.

Κάποιες από τις μετρήσεις που χρησιμοποιήθηκαν εναλλακτικά από τη μέτρηση AIO: αξία, στάση και τρόπος ζωής (VALS). λίστα αξιών (LOV) και η Έρευνα Αξίας Rokeach (RVS). Μία από τις πιο ευρέως χρησιμοποιούμενες μεθόδους τμηματοποίησης αξιών είναι η τυπολογία συμβουλών αξιών και τρόπου ζωής (VALS2) του *Stanford Research Institute*, η οποία περιέχει 35 ψυχογραφικές και τέσσερις δημογραφικές ερωτήσεις που συνδέουν τα δημογραφικά στοιχεία και τα πρότυπα αγοράς με ψυχολογικές στάσεις.

Στον παρακάτω πίνακα παρουσιάζονται λέξεις που όταν εντοπίζονται σε κείμενα και ανάλογα με την πολικότητα τους τμηματοποιούνται στις κατηγορίες Big Five Factors. Οι κατηγορίες είναι εξωστρέφεια, νευρωτισμός, ανοικτότητα, ευχαρίστηση και ευσυνείδησια και η λέξεις είναι από κριτικές ταινιών.

Big Five Factors	σχετικές λέξεις με θετική πολικότητα	σχετικές λέξεις με αρνητική πολικότητα
Νευρωτισμός	terrible, irony,lazy	visited, odest,
Εξωστρέφια	drinks, dancing, bars, pools, grilled	computer, minor
Ανοικτότητα	humans, art,films, blues, poets	giveawasy
Ευχαρίστηση	wonderful, morning, visiting, beautiful, spring, share, joy	Porn, Fuck
Ευσυνειδησία	adventure, enjoing	Stupid, boring, desperate

Εικόνα 4.4: Τμηματοποίηση λέξεων με βάση την πολικότητά τους με τη μέθοδο BBF

Το WordNet είναι μια μεγάλη αγγλική βάση λέξεων που ομαδοποιεί ουσιαστικά, ρήματα, επίθετα και επιρρήματα σε σύνολα γνωστικών συνωνύμων που το καθένα εκφράζει μια ξεχωριστή έννοια. Τα συνώνυμα αλληλοσυνδέονται μέσω εννοιολογικών σημασιολογικών και λεξικών σχέσεων. Με βάση το WordNet και τις ψυχογραφικές λέξεις-σπόρους BFF, εξάγονται οι σχετικές λέξεις ακολουθώντας τον κανόνα των συνωνύμων: για κάθε θετική λέξη-σπόρο, το συνώνυμό της θεωρείται ότι έχει το ίδι BFF. Επομένως, λαμβάνεται το ψυχογραφικό του υποψήφιου καταναλωτή που αποτελείται από λέξεις-σπόρους και τα συνώνυμα τους από το WordNet.

Κατά την ανάλυση συναισθημάτων όσο περισσότεροι χρήστες λένε την άποψη τους τόσο πιο εύκολο είναι να βγει ασφαλές συμπέρασμα για το αν η γενική εικόνα έχει θετική ή αρνητική πολικότητα. Όσο περισσότερο μια "μηχανη" εντοπίζει σωστά τις κριτικές για ένα προϊόν ή υπηρεσία, τόσο λιγότερο απαιτείται και ο ανθρώπινος παράγοντας.

4.5 Νέοι δρόμοι στην ανάλυση κοινωνικών δικτύων με ανάλυση συναισθημάτων:

Στο άρθρο των ερευνητών (*Hyoji Ha; Sang Woo Han; Seongmin Mun, 2019).* οι συγγραφείς προτείνουν μια μέθοδο οπτικοποίησης συναισθημάτων στα κοινωνικά μέσα. Για το σκοπό αυτό, σχεδιάστηκε ένας μηχανισμός δικτύου οπτικοποίησης συναισθημάτων πολλαπλών επιπέδων που βασίζεται σε συναισθηματικές λέξεις στον τομέα της κριτικής ταινιών.

Προτείνονται 3 τεχνικές για την οπτικοποίηση:

- Μια απεικόνιση με χάρτη θερμότητας των σημασιολογικών/συναισθηματικών λέξεων κάθε κόμβου.
- Ένας δισδιάστατος χάρτης κλίμακας δεδομένων από συναισθηματικές λέξεις.
- Μια απεικόνιση αστερισμού για κάθε σύμπλεγμα του δικτύου.

Οι προτεινόμενες απεικονίσεις χρησιμοποιούνται ως σύστημα συστάσεων που προτείνει ταινίες που προκαλούν παρόμοια συναισθήματα με αυτές που παρακολουθήθηκαν προηγουμένως οι χρήστες και τους άρεσαν. Αυτή η ιδέα συστάσεων περιεχομένου που βασίζεται σε παρόμοια συναισθηματικά μοτίβα, μπορεί να εφαρμοστεί και σε άλλα κοινωνικά δίκτυα.

Το άρθρο των *(Hannah Kim; Young-Seob Jeong, 2019)* ασχολείται με τα προβλήματα και τις προκλήσεις της ταξινόμησης συναισθηματικών από κείμενα. Προτείνουν ένα συμβατικό νευρωνικό δίκτυο (CNN) που αποτελείται από ένα επίπεδο ενσωμάτωσης, δύο συμβατικά επίπεδα, ένα επίπεδο συγκέντρωσης και ένα πλήρως συνδεδεμένο. Το μοντέλο αξιολογείται σε τρία σύνολα δεδομένων (δεδομένα κριτικών ταινίας, δεδομένα επισκόπησης χρηστών και δεδομένα *Stanford Sentiment Treebank*) και συγκρίνεται με τα παραδοσιακά μοντέλα μηχανικής

μάθησης αλλά και τα σύγχρονα μοντέλα βαθιάς μάθησης. Το κύριο συμπέρασμα είναι ότι η χρήση διαδοχικών και συνεχών επιπέδων είναι αποτελεσματική για σχετικά μεγάλα κείμενα.

Στο άρθρο *τους* οι συγγραφείς *(Xingliang Mao;Shuai Chang; Jinjing Shi; Fangfang Li; Ronghua Shi, 2019)* προτείνουν τη χρήση μιας ενσωμάτωσης λέξεων με συναισθήματα για τη βελτίωση της συναισθηματικής ανάλυσης. Η προτεινόμενη μέθοδος δημιουργεί μια υβριδική αναπαράσταση που συνδυάζει συναισθηματικές ενσωματώσεις λέξεων που βασίζονται σε λεξικό, με σημασιολογικές ενσωματώσεις λέξεων που βασίζονται στο Word2Vec. Χρησιμοποιεί το συναισθηματικό λεξικό DUTIR, το οποίο είναι μια κινεζική οντολογική πηγή από το *Dalian University of Technology Information Retrieval Laboratory*. Το DUTIR υποσημειώνει στις καταχωρήσεις λεξικού με ένα μοντέλο επτά συναισθημάτων (ευτυχία, εμπιστοσύνη, θυμός, θλίψη, φόβος, αηδία και έκπληξη) το εκάστοτε συναίσθημα κάθε φορά. Η αξιολόγηση γίνεται με δεδομένα από το Weibo, έναν δημοφιλή ιστότοπο κοινωνικής δικτύωσης της Κίνας.

Το άρθρο αξιολογεί δύο μεθόδους, τον άμεσο συνδυασμό και την προσθήκη για την δημιουργία της υβριδικής αναπαράστασης σε διάφορα σύνολα δεδομένων. Το συμπέρασμα από τα παραπάνω πειράματα είναι ότι αποδεικνύεται πως η χρήση υβριδικών διανυσμάτων λέξεων είναι αποτελεσματική για την εποπτεία της ταξινόμησης συναισθημάτων, βελτιώνοντας σημαντικά την ακρίβεια της ταξινόμησης.

Μέρος 2ο

ΚΕΦΑΛΑΙΟ 5 APACHE SPARK

5.1 Τι είναι το Apache Spark;

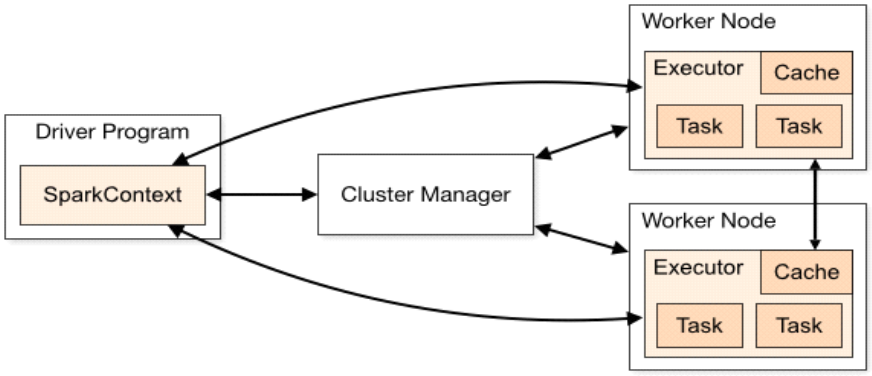
Το Apache Spark ™ είναι μια μηχανή ανάλυσης ανοιχτού κώδικα που δημιουργήθηκε το 2009, για επεξεργασία δεδομένων μεγάλης κλίμακας που επιτυγχάνει υψηλή απόδοση και ταχύτητα τόσο για δεδομένα δέσμης, όσο και για δεδομένα ροής. Χρησιμοποιεί έναν υπερσύγχρονο προγραμματιστή DAG, έναν βελτιστοποιητή ερωτημάτων και μια μηχανή φυσικής εκτέλεσης. Η δημιουργία εφαρμογών είναι αρκετά εύκολη και μπορεί να γίνει σε Java, Scala, Python, R και SQL.

Το Spark προσφέρει πάνω από 80 λειτουργίες υψηλού επιπέδου που διευκολύνουν την κατασκευή παράλληλων εφαρμογών.

Εικόνα 5.1: Apache Spark Logo

Συνδυάζει SQL, Streamming και σύνθετα εργαλεία ανάλυσης. Ενεργοποιεί μια στοίβα βιβλιοθηκών, συμπεριλαμβανομένων των SQL και DataFrames, τις MLlib για μηχανική εκμάθηση, και των GraphX και Spark Streaming. Οι βιβλιοθήκες αυτές μπορούν να χρησιμοποιηθούν και στην ίδια εφαρμογή.

Ένα από τα μεγαλύτερα πλεονεκτήματά του είναι ότι τρέχει παντού! Στο Hadoop, Apache Mesos, Kubernetes, αυτόνομα ή σε cloud. Μπορεί να έχει πρόσβαση σε διαφορετικές πηγές δεδομένων όπως: HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive και εκατοντάδες άλλες πηγές δεδομένων.



Εικόνα 5.2: Πως λειτουργεί το Spark

Οι εφαρμογές του Spark εκτελούνται ως ανεξάρτητα σύνολα διαδικασιών σε ένα σύμπλεγμα, που συντονίζονται από το αντικείμενο SparkContext στο κύριο πρόγραμμά. Ουσιαστικά, ακόμα και όταν μια εφαρμογή του Spark τρέχει τοπικά σε έναν υπολογιστή, υπάρχει ένα script που αναφέρει πως πώς εισάγονται τα δεδομένα, πώς μετατρέπονται, πώς επεξεργάζονται και πώς εξάγονται συμπεράσματα. Αυτό το script βρίσκεται στον οδηγό του προγράμματος (driver), όπως φαίνεται και στην παραπάνω εικόνα. Ένα άλλο πολύ ενδιαφέρον χαρακτηριστικό του Spark που του προσδίδει στη ταχύτητα είναι πως τίποτα δεν συμβαίνει αν δεν γίνει κλήση μιας εντολής.

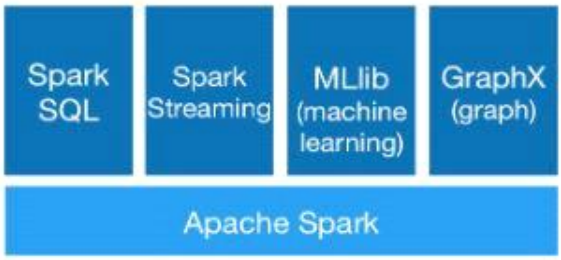
Συγκεκριμένα, για να τρέξει σε ένα σύμπλεγμα, το SparkContext μπορεί να συνδεθεί με διάφορους τύπους διαχειριστών συμπλέγματος (τον δικό του αυτόνομα, Mesos, YARN), οι οποίοι κατανέμουν πόρους σε όλες τις εφαρμογές. Μόλις συνδεθεί, το Spark αποκτά εκτελεστές σε κόμβους του συμπλέγματος, οι οποίοι είναι διαδικασίες που εκτελούν υπολογισμούς και αποθηκεύουν δεδομένα για την εκάστοτε εφαρμογή. Στη συνέχεια, στέλνει τον κωδικό της εφαρμογής(ορίζεται από αρχεία JAR ή Python που έχουν μεταφερθεί στο SparkContext) στους εκτελεστές. Τέλος, το SparkContext στέλνει εργασίες στους εκτελεστές.

Το Spark έχει σχεδιαστεί να καλύπτει ευρύ φάσμα εργασιών που προηγουμένως απαιτούσαν ξεχωριστά κατανεμημένα συστήματα.

Από τι αποτελείται το Spark:

- **Spark Sql**: Η Spark SQL επιτρέπει την δημιουργία προβολών από ένα σύνολο δεδομένων (RDD).
- **Spark Streaming**: Το Spark Streaming είναι ένα από τα βασικά συστατικά του Spark. Είναι με λίγα λόγια, η τεχνολογία του Spark που ασχολείται με την επεξεργασία ροών από big data που έρχονται σε πραγματικό χρόνο από διάφορες πηγές δεδομένων.
- **MLlib**: Η MLlib είναι βιβλιοθήκη που περιλαμβάνει μια συλλογή από αλγορίθμους μηχανικής μάθησης που απαιτούνται.
- **GraphX**: Συστατικό που αφορά την θεωρία γραφημάτων. Η GraphX είναι μια βιβλιοθήκη για χειρισμό γράφων και εφαρμογής σε αυτούς

παράλληλων υπολογισμών. Δηλαδή, για παράδειγμα στα κοινωνικά δίκτυα υπάρχουν περίπλοκα δίκτυα μεταξύ χρηστών, αισθητήρων κ.α. που περιλαμβάνονται στο GraphX. Επίσης περιλαμβάνονται και κάποια βοηθητικά προγράμματα που είναι απαραίτητα για την ανάλυση τέτοιων γραφημάτων.



Εικόνα 5.3: Από τι αποτελείται το Spark

Δομή Δεδομένων RDD

Τα RDDs είναι ανθεκτικά σε σφάλματα, διανεμημένα, σύνολα δεδομένων. Ένα από τα βασικότερα στοιχεία ενός RDD είναι πως είναι αμετάβλητο. Με άλλα λόγια, μπορεί να ειπωθεί πως ένα RDD είναι ένα τεράστιο, ογκώδες σύνολο δεδομένων, που δεν μπορεί να χωρέσει σε καμία μηχανή. Ένας χρήστης δεν μπορεί να προσθέσει δεδομένα σε ένα RDD, αλλά μπορεί να εφαρμόσει σε αυτό ενέργειες και μετατροπές με σκοπό να σχηματίσει ένα νέο RDD. Οι λειτουργίες που μπορούν να πραγματοποιηθούν σε ένα RDD είναι οί map, flatmap filter, reduce, distinct, sample, count, collect κ.α. Στο Spark όλες οι ενέργειες που γίνονται έχουν να κάνουν είτε με δημιουργία νέων RDDs, είτε με την μετατροπή υπαρχόντων RDDs, είτε τέλος με την εφαρμογή πράξεων σε ένα RDD για τον υπολογισμό ενός αποτελέσματος.

Παρέχονται δύο πολύ βασικές εντολές οι map και reduce. Η map στο στάδιο των μετατροπών και η reduce των πράξεων (Σαν το MapReduce του Hadoop.
Στον προγραμματισμό με Scala στο Spark δημιουργούμε ένα αντικείμενο Spark Context το οποίο είναι αυτό που δημιουργεί τα RDDs. Μια τέτοια παραλλαγή κάνουμε και στο Spark Streaming με την εντολή:

```
        val ssc = new StreamingContext("local[*]", "PrintTweets",
Seconds(1))
```

Είναι το αντικείμενο που ουσιαστικά ασχολείται με τα δεδομένα σε πραγματικό χρόνο. Με λίγα η εντολή δηλώνει πως η επεξεργασία (PrintTweets) γίνεται τοπικά στον υπολογιστή και τα σύνολο δεδομένα παίρνονται κάθε 1 δευτερόλεπτο.

Αντί να δημιουργήσουμε δεδομένα από μια ροή, μπορούμε να τα δημιουργήσουμε ακόμα και τοπικά (μια λίστα αριθμών πχ. Num = sc.parallelize [1,2,3,4,5,6]). Τό παραπάνω βέβαια δεν είναι καθόλου πρακτικό και χρήσιμο, εφόσον είτε τα δεδομένα δεν θα είναι Big Data ώστε να μπορούμε να τα

διαχειριστούμε με αυτόν το τρόπο, και άρα, δεν θα χρειάζεται το Spark, είτε αναφερόμαστε σε Big Data όντως και ο σκοπός είναι να μην καταλαμβάνουμε πόρους από την μνήμη τοπικά για τόσο ογκώδη δεδομένα.
Μπορούμε επίσης να τα αντλήσουμε από κάποιο αρχείο από τον δίσκο του υπολογιστή (π.χ.textfile) ή και από κάποια βάση δεδομένων (HBase, Cassandra) κ.α

Apache Spark vs Handoop

Γίνεται λόγος πως το Spark είναι ο αντικαταστάτης του Handoop. Αυτό που πραγματικά εννοείται από την παραπάνω δήλωση είναι ότι το Spark είναι ο αντικαταστάτης της διαδικασίας Map Reduce και όχι του Hadoop στο σύνολό του. Η MapReduce είναι ανεπαρκής για επαναληπτικές και διαδραστικές υπολογιστικές εργασίες. Το Spark είναι 100 φορές πιο γρήγορο από τη Map Reduce στη μνήμη αλλά και 10 φορές πιο γρήγορο στο δίσκο. Αυτό συμβαίνει γιατί με τον τρόπο που λειτουργεί, τίποτα δεν συμβαίνει μέχρι να δοθεί εντολή από το χρήστη να γίνει. Υπάρχει έλεγχος της ροής δεδομένων, την βελτιστοποιεί, ώστε η επεξεργασία τους να είναι η καλύτερη που μπορεί.

Επομένως, το Spark δεν αποκλείει το Hadoop. Το Spark μπορεί να χρησιμοποιηθεί συνδυαστικά με το Hadoop (σύμπλεγμα YARN) και να γίνει εκμετάλλευση της επεκτασιμότητας και ανθεκτικότητας του Hadoop, αλλά και όχι, μιας και έχει και δικό του σύμπλεγμα και επιπλέον υπάρχουν και άλλες σύγχρονες εναλλακτικές όπως το Mesos.

Κάποιες από τις εταιρείες που χρησιμοποιούν το Spark είναι η Amazon, η Yahoo, το Ebay, Tripadvisor και πολλές ακόμη.

Στην παρούσα εργασία θα ασχοληθούμε με το Spark Streaming και θα προχωρήσουμε σε κάποια παραδείγματα ώστε να κατανοήσουμε τη χρήση του, αντλώντας δεδομένα σε πραγματικό χρόνο από το Twitter.

Παρακάτω παρουσιάζονται αναλυτικά όλα τα βήματα:

5.2 Εγκατάσταση του Spark:

Το Spark Streaming είναι ένα από τα βασικά συστατικά του Spark. Είναι με λίγα λόγια, η τεχνολογία του Spark που ασχολείται με την επεξεργασία ροών από big data που έρχονται σε πραγματικό χρόνο.

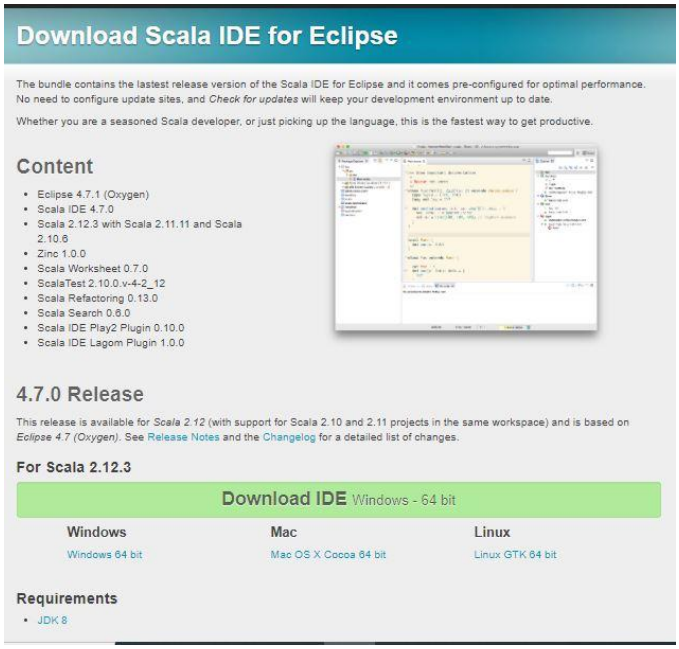
Στην παρούσα ενότητα θα αναλυθεί η διαδικασία εγκατάστασης του Apache Spark σε Windows.

Αρχικά, αν δεν υπάρχει ήδη εγκατεστημένη στον υπολογιστή, πρέπει να πραγματοποιηθεί εγκατάσταση της Java. Για το παρόν εγχείρημα κατέβηκε η έκδοση 8 από το site java.com.

Εικόνα 5.4: Java Download

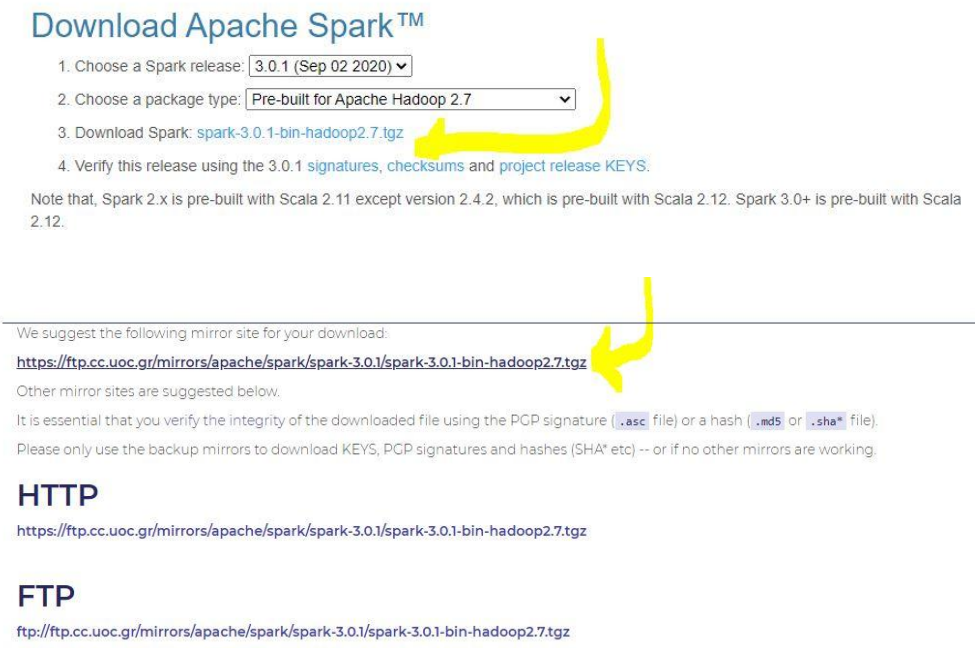
Προσοχή. Πρέπει η έκδοση της Java να είναι συμβατή με την έκδοση της Scala με την οποία θα πραγματοποιηθεί ο προγραμματισμός του κώδικα στο Eclipse. Για την παρούσα εργασία εγκαταστάθηκε η Scala IDE 4.7.1.. Παρατηρούμε πως στα προαπαιτούμενα είναι η JDK 8 την οποία και κατεβάσαμε. Μόλις κατέβουν οι εφαρμογές πρέπει να εγκατασταθούν τοπικά στο σύστημα.

http://scala-ide.org/download/sdk.html



Εικόνα 5.5: Scala Download

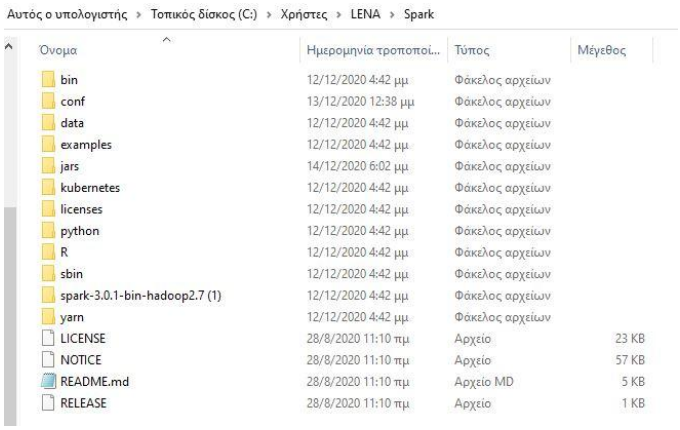
Μετά την εγκατάσταση της Java και της Scala προχωράμε στην εγκατασταση του Spark το οποίο υπάρχει εδώ: https://spark.apache.org/downloads.html . Επιλέγεται η τελευταία έκδοση και μια εκδοχή pre-built για το Apache Hadoop.



Εικόνα 5.6: Download Apache Spark

Μόλις ολοκληρώθηκε το κατέβασμα του Spark, γίνεται αποσυμπίεση με το winrar (https://www.win-rar.com/start.html?&L=0) και μεταφέρεται στον τοπικό

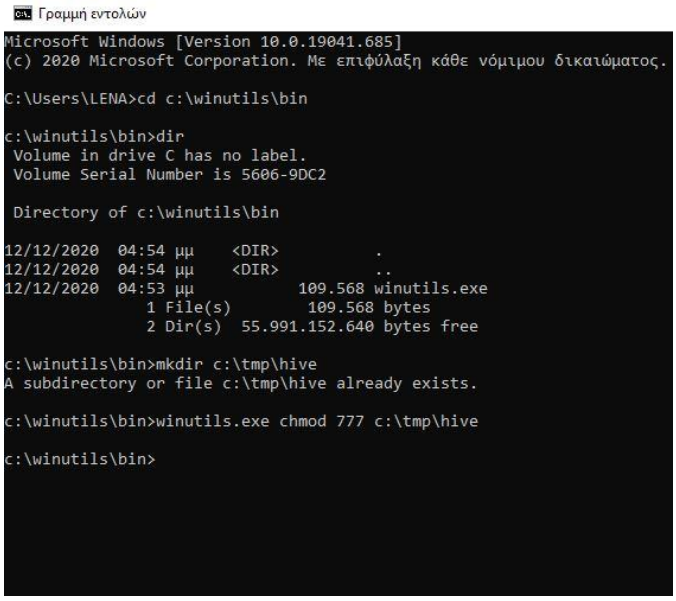
δίσκο C. Για την ακρίβεια, η διαδρομή τοποθέτησης και τα περιεχόμενα τπυ προγράμματος φαίνονται στην παρακάτω εικόνα:



Εικόνα 5.7: Περιεχόμενα Apache Spark

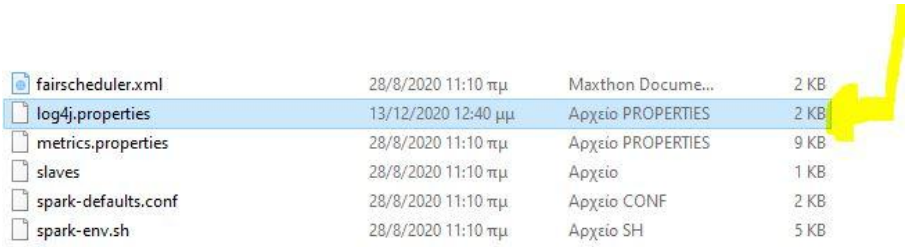
Το Spark έχει δημιουργηθεί για μια συγκεκριμένη έκδοση του Hadoop, παρόλα αυτά δεν χρειάζεται να τρέξουμε το Hadoop τοπικά στον υπολογιστή μας. Με το αρχείο winutils.exe (βοηθητικά προγράμματα για τα Windows), τα Windows θα «πιστεύουν» πως υπάρχει το Hadoop. Απαιτείται δημιουργία του φακέλου winutils\bin στον τοπικό δίσκο C και εκεί να γίνει τοποθέτηση του αρχείου winutils.exe. Ουσιαστικά σκοπός είναι η δημιουργία ενός προσωρινού hive directory ώστε να φαίνεται στα windows πως τρέχουμε το πρόγραμμα hadoop μέσω linux, χωρίς αυτό στη πραγματικότητα να συμβαίνει.

Ανοίγουμε το παράθυρο της γραμμής εντολών και ακολουθούμε τα εξής βήματα:

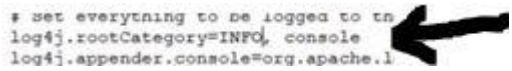


Εικόνα 5.8: Γραμμη Εντολών

Έπειτα, το Spark όσο εκτελούνται εργασίες είναι ρυθμισμένο να εμφανίζει πολλά μηνύματα με πληροφορίες. Ώστε να μην προκαλείται σύγχιση καλό θα ήταν να μην γίνεται. Για να λυθεί το εξής θέμα πάμε στο φάκελο της εγκατάστασης του προγράμματος, στον υποφάκελο conf και στο αρχείο που φαίνεται στην εικόνα:

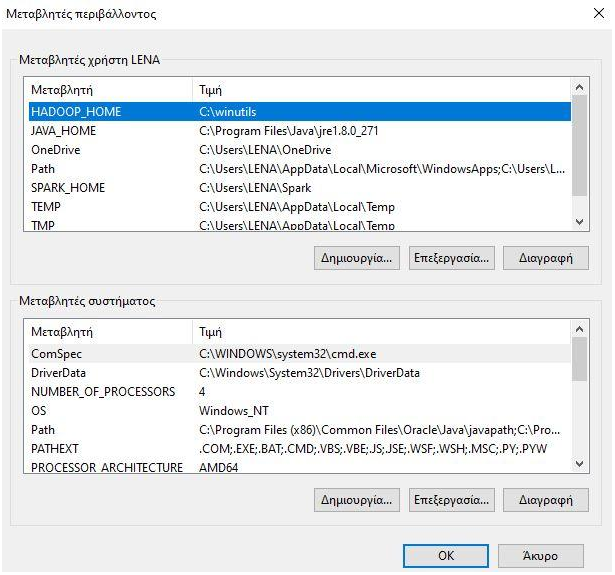


Εικόνα 5.9: Properties



Βρίσκουμε την παραπάνω σειρά και αλλάζουμε το INFO σε ERROR.

Το τελευταίο που πρέπει να ρυθμιστεί ώστε να μπορέσουμε να χρησιμοποιήσουμε το Spark είναι να ρυθμιστούν κάποιες από τις μεταβλητές περιβάλλοντος ώστε να ενημερωθούν τα Windows πως το Spark έχει εγκατασταθεί.



Εικόνα 5.10: Μεταβλητές περιβάλλοντος

Επιλέγουμε δημιουργία και προσθέτουμε τις παρακάτω μεταβλητές:
Όνομα μεταβλητής: SPARK_HOME
Τιμή μεταβλητής: C:\Users\LENA\Spark

Όνομα μεταβλητής: JAVA_HOME
Τιμή μεταβλητής: C:\Program Files\Java\jre1.8.0_271

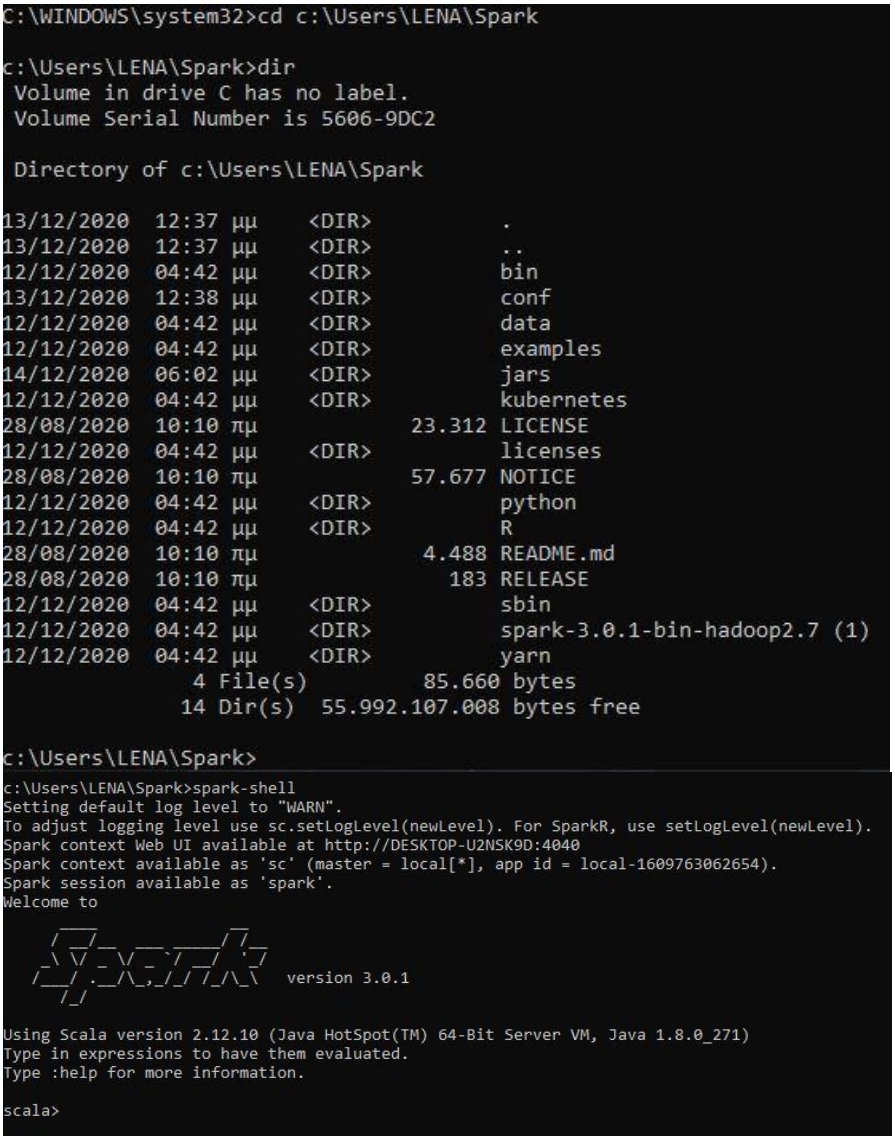
Όνομα μεταβλητής: HADOOP_HOME
Τιμή μεταβλητής: C:\winutils

Έπειτα πρέπει να τροποποιηθεί και το path directory (επεξεργασία-> νέο) και προσθέτουμε τα εξής:

%SPARK_HOME%\bin
%HADOOP_HOME%\bin
C:\Program Files\Java\jdk1.8.0_271

Μέσα στην εγκατάσταση της Scala που πραγματοποιήθηκε προηγουμένως υπάρχει το Eclipse το οποίο θα χρησιμοποιήσουμε για να τρέξουμε τα έργα μας. Καλό θα ήταν να μεταφέρουμε τον φάκελο του στο τοπικό δίσκο C ώστε να έχου με πιο άμεση πρόσβαση και να δημιουργήσουμε μία συντόμευση του προγράμματος.

Τέλος, ανοίγουμε και πάλι την γραμμή εντολών, αυτή την φορά ως διαχειριστής, για να ελέγξουμε αν όλα έχουν πάει καλά.



Εικόνα 5.11: Επιβεβαίωση πως όλα πήγαν σωστά

Όπως βλέπουμε είμαστε έτοιμοι να προγραμματίσουμε σε Scala.

- **Πρόσβαση στο Twitter for Developers**

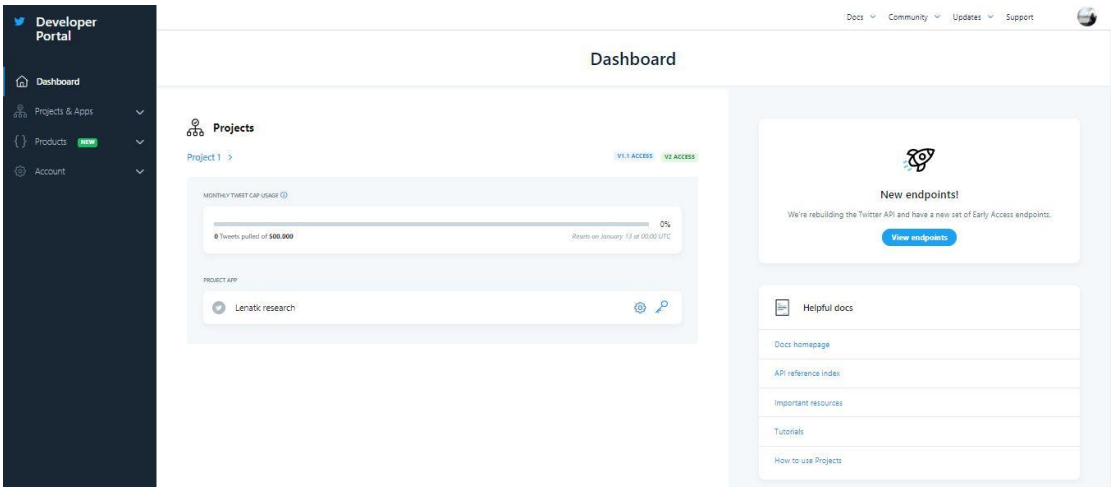
Για να αντλήσουμε δεδομένα από το Twitter απαιτείται η δημιουργία λογαριασμού developer.

Δημιουργούμε το λογαριασμό στο παρακάτω σύνδεσμο:

https://developer.twitter.com/en

Εκεί απαιτείται να δηλωθεί η χρήση που θα γίνει στα δεδομένα που ζητάμε πρόσβαση, ποιοι θα τα δουν και τι ακριβώς είναι τα δεδομένα αυτά (tweets, προσωπικά στοιχεία κ.α.)

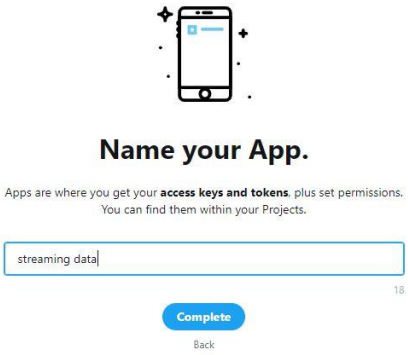
Εφόσον δημιουργήσουμε τον λογαριασμό έχουμε πρόσβαση στη δημιουργία εφαρμογών. Ο πίνακας ελέγχου φαίνεται στην παρακάτω εικόνα.



Εικόνα 5.12: Twitter API

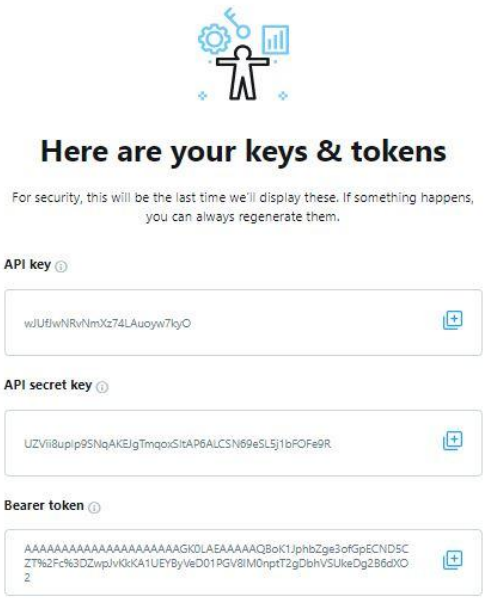
Δημιουργία νέας εφαρμογής:

Από την επισκόπηση επιλέγουμε create an app και δίνουμε όνομα.



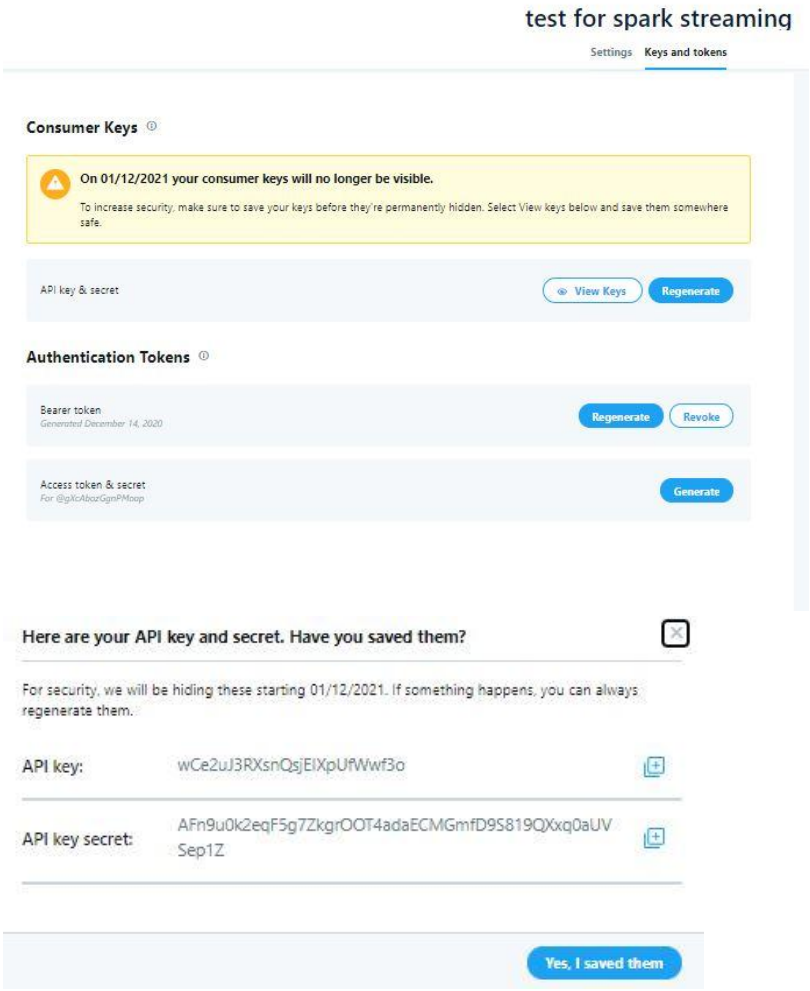
Εικόνα 5.13: Όνομα Εφαρμογής

Αυτά που χρειαζόμαστε για το streaming είναι τα keys και τα access tokens. Για παράδειγμα:



Εικόνα 5.14: Κλειδιά που απαιτούνται για την άντληση δεδομένων

Έπειτα επιλέγουμε ρυθμισεις → Keys and Tokens και εκεί υπάρχουν και τα υπόλοιπα στοιχεία που χρειαζόμαστε.



Εικόνα 5.15: Κλειδιά που απαιτούνται για την άντληση δεδομένων β

Δημιουργούμε ένα txt αρχείο twitter.txt που μέσα περιέχει όλα τα κλειδιά και τις άδειες. Έπειτα θα το προσθέσουμε στον κώδικα ώστε να μπορέσουμε να κάνουμε streaming. Για παράδειγμα: (Τα συγκεκριμένα κλειδιά δεν είναι πλέον ενεργά)
accessToken 1338077958070820875-E0Q932XaA1aCikb1k5IkzKLXBfDBC3
accessTokenSecret CR13ktWTagpdxWPw3UJuCpiuFBSEl8WP7daoXdLJYHQfP
consumerKey QP2Bve5nVJqNqs15Wf1Qkqz6D
consumerSecret 2BnEP3EzDPN31hvnIpwPlQ5sdQrH9aATGwyYRTHIfq5sYcmcMg

*Το Consumer Key είναι το Api Key

Για την διαδικασία χρησιμοποιήθηκε το eclipse ώστε να γράφετε ο κώδικας και να εκτελείται το πρόγραμμα. Για το παρόν εγχείρημα, όπως αναφέρθηκε και προγηουμένως θα χρησιμοποιηθεί η γλώσσα Scala.

Μέσα στο project που δημιουργήθηκε στο eclipse προστέθηκαν όλες οι απαραίτητες βιβλιοθήκες (Spark Libraries) στο Java Build Path, για να μπορέσουν να εκτελεστούν όλα τα παραδείγματα. Η εξωτερικές αυτές βιβλιοθήκες είναι τα αρχεία .jar που υπάρχουν μέσα στον φάκελο της εγκατάστασης του προγράμματος.

Πέρα από τις βιβλιοθήκες για το Spark χρειαζόμαστε και τις βιβλιοθήκες για το Twitter Streaming. Οπότε προσθέτουμε και αυτές στο ίδιο σημείο.

Εικόνα 5.16: Βιβλιοθήκες για Twitter Streaming

Με την παρακάτω εντολή (θα παρουσιαστεί πιο αναλυτικά στα παραδείγματα που ακολουθούν), συνδέουμε τα διεπιστευτήρια που έχουμε από το Twiiter στο αρχείο Twitter.txt.

```
setupTwitter()
```

5.3 Παραδείγματα με χρήση του Eclipse

Στην ενότητα αυτή θα δούμε 3 παραδείγματα επεξεργασίας δεδομένων σε πραγματικό χρόνο με το Spark Streaming προγραμματίζοντας στο Eclipse με την γλώσσα προγραμματισμού Scala.

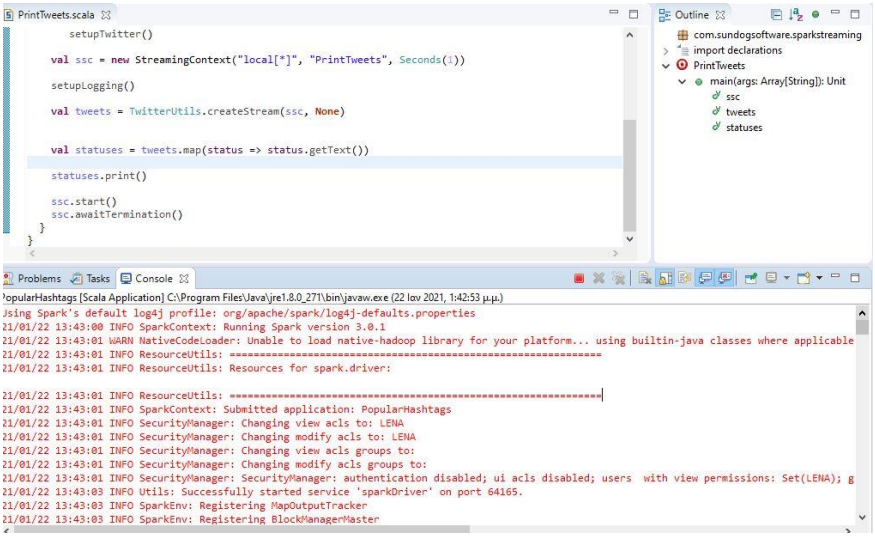
- **Εμφάνιση των tweets τοπικά στο δίσκο**

```scala
import org.apache.spark._

/** Simple application to listen to a stream of Tweets and print them out */
object PrintTweets {

  /** Our main function where the action happens */
  def main(args: Array[String]) {

      setupTwitter()

    val ssc = new StreamingContext("local[*]", "PrintTweets", Seconds(1))

    setupLogging()

    val tweets = TwitterUtils.createStream(ssc, None)


    val statuses = tweets.map(status => status.getText())

    statuses.print()

    ssc.start()
    ssc.awaitTermination()
  }
}
```

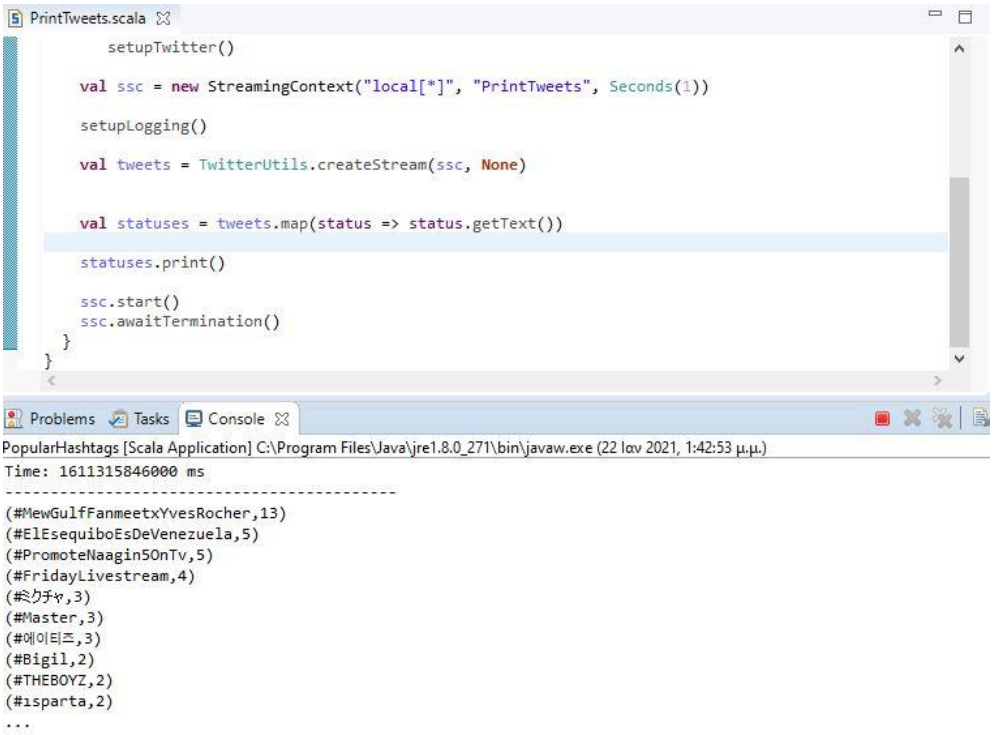
Εικόνα 5.17: Κώδικας για εμφάνιση tweets τοπικά στο δίσκο

Στην παραπάνω εικόνα βλέπουμε τον κώδικα που απαιτείται. Η πρώτη εντολή μέσα στη main είναι αυτή που εγκαθιστά στο πρόγραμμα τα διαπιστευτήρια από το twitter. Αυτό γίνεται για να μπορούμε να έχουμε πρόσβαση στα tweets σε πραγματικό χρόνο. Η επόμενη αμέσως εντολή είναι αυτή που ορίζει πως θα χρησιμοποιήσουμε όλες τις δυνατότητες που έχει το σύστημά μας τοπικά. Επίσης καλούμε το σύστημα να σώζει σύνολα tweets κάθε 1 δευτερόλεπτο. Με την επόμενη εντολή απαλασσόμαστε από τα log spams. Έπειτα δημιουργούμε το stream από το twiiter. Με την επόμενη εντολή λέμε στο πρόγραμμα να παίρνει το κείμενο από τα tweets και έπειτα να τα εκτυπώνει. Με την συνάρτηση map. (εφαρμόζουμε αυτό που λέει η συνάρτηση map σε κάθε αντικείμενο του προγράμματος. Εν προκειμένω στα tweets.

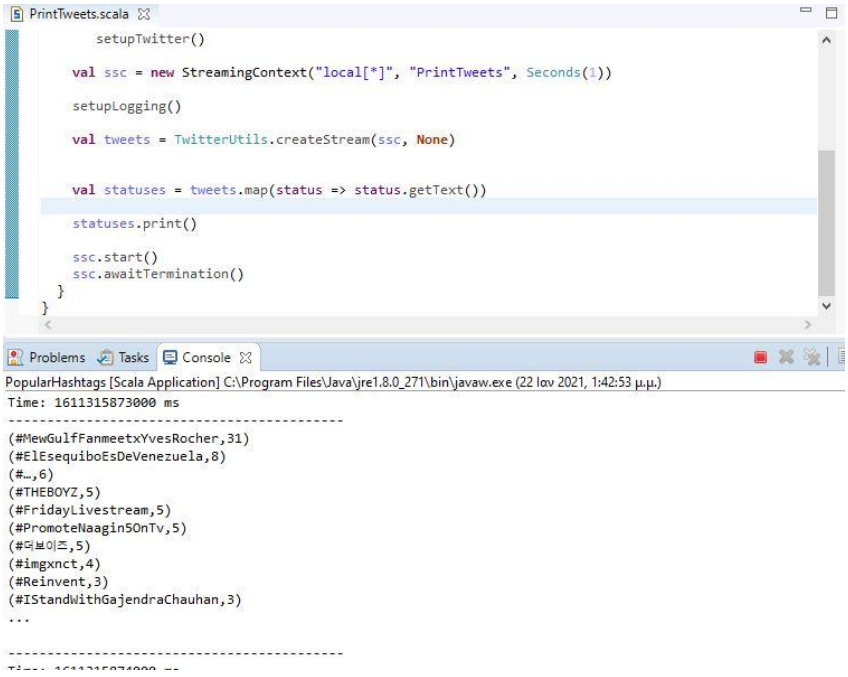
Τρέχουμε τον κώδικα και παρακάτω βλέπουμε στιγμιότυπα από τα αποτελέσματα. Πλέον μπορούμε να εκτυπώνουμε tweets χρηστών σε πραγματικό χρόνο τοπικά!



Εικόνα 5.18: Οθόνες εκτέλεσης του κώδικα για εμφάνιση tweets τοπικά α



Εικόνα 5.18: Οθόνες εκτέλεσης του κώδικα για εμφάνιση tweets τοπικά β



Εικόνα 5.19: Οθόνες εκτέλεσης του κώδικα για εμφάνιση tweets τοπικά γ

Για να σταματήσει η διαδικασία πατάμε το terminate.

- **Ποιος είναι ο μέσος όρος του μεγέθους των tweets και ποιό το μεγαλύτερο σε χαρακτήρες;**

Οι πρώτες εντολές είναι ίδιες με το προηγούμενο παράδειγμα. Πάλι αντλούμε το κείμενο απο τα tweets. Εκτός από το κείμενο θέλουμε και το μήκος από κάθε tweet. Χρησιμοποιούμε τη κλάση atomiclog της java για να βεβαιωθούμε πως τα σύνολα που χρειαζόμαστε είναι ασφαλ΄ή, γιατί εκτελούνται παράλληλα πολλές διαδικασίες.

Στην 2η εικόνα βλέπουμε τη διαδικασία με την οποία βρίσκουμε τον μέσο όρο του μεγέθους των tweets όπως και το ποιό είναι το μεγαλύτερο σε χαρακτήρες με την βοήθεια της συνάρτησης reduce.

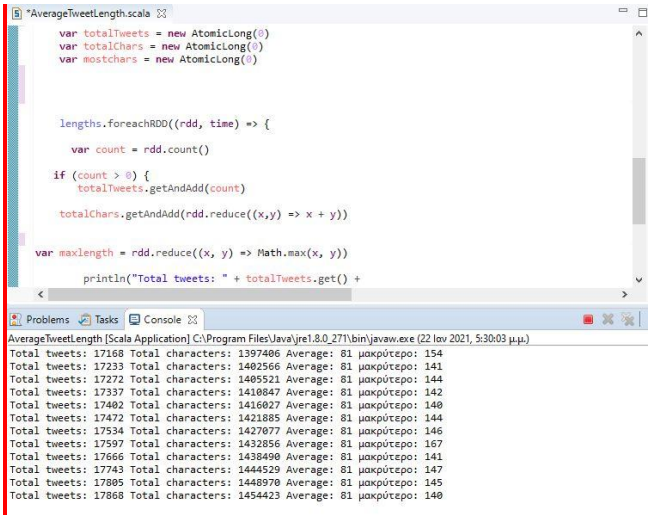
Ουσιαστικά με τη χρήση του μετρητή count βρίσκουμε τα συνολικά tweets, μετά τους συνολικούς χαρακτήρες και κάνουμε μια διαίρεση ώστε να βρούμε το μέσο μέγεθος. Για να βρούμε το μεγαλύτερο καλούμε με την βοήθεια της reduce της συνάρτηση math.max

```
import org.apache.spark._


/** Uses thread-safe counters to keep track of the average length of
object AverageTweetLength {

  /** Our main function where the action happens */
  def main(args: Array[String]) {

    setupTwitter()


    val ssc = new StreamingContext("local[*]", "AverageTweetLength", Seconds(1))

    setupLogging()

    val tweets = TwitterUtils.createStream(ssc, None)

    val statuses = tweets.map(status => status.getText())

    val lengths = statuses.map(status => status.length())
```

Εικόνα 5.20: Μέσος όρος μεγέθους tweets και μεγαλύτερο σε χαρακτήρες κώδικας α

```
*AverageTweetLength.scala
    var totalTweets = new AtomicLong(0)
    var totalChars = new AtomicLong(0)
    var mostchars = new AtomicLong(0)



    lengths.foreachRDD((rdd, time) => {

      var count = rdd.count()

    if (count > 0) {
        totalTweets.getAndAdd(count)

    totalChars.getAndAdd(rdd.reduce((x,y) => x + y))

var maxlength = rdd.reduce((x, y) => Math.max(x, y))

        println("Total tweets: " + totalTweets.get() +
            " Total characters: " + totalChars.get() +
            " Average: " + totalChars.get() / totalTweets.get() +
            " μακρύτερο: " + maxlength
            )

}
    })

    ssc.checkpoint("C:/checkpoint/")
    ssc.start()
    ssc.awaitTermination()
  }
}
```

Εικόνα 5.21: Μέσος όρος μεγέθους tweets και μεγαλύτερο σε χαρακτήρες κώδικας β



Εικόνα 5.22: Μέσος όρος μεγέθους tweets και μεγαλύτερο σε χαρακτήρες κώδικας γ

- **Τα 10 πιο δημοφιλή hashtags που χρησιμοποιούν οι χρήστες**

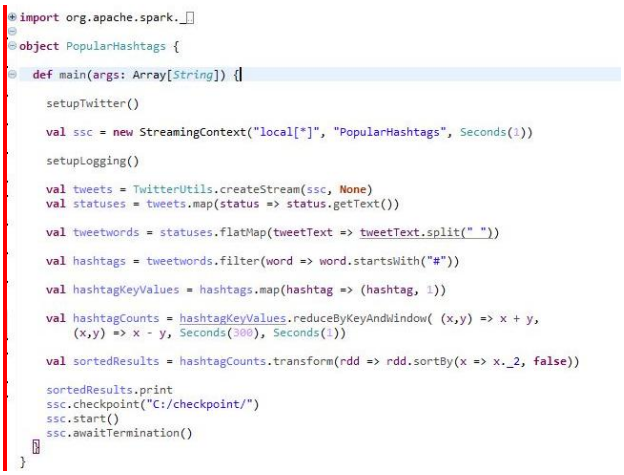
Στο παράδειγμα αυτό χρησιμοποιούμε την συνάρτηση flatmap για να διαχωρίσουμε κάθε λέξη που περιέχει ένα tweet.

Έπειτα, επιλέγουμε να κρατήσουμε μόνο τις λέξεις που αρχίζουν με # δηλαδή τα hashtags.

Τρέξαμε το πρόγραμμα για 5 λεπτά και παίρνουμε σύνολα με hashtags κάθε 1 δευτερόλεπτο.

Έχουμε μετρήσει πόσες φορές εμφανίζεται κάθε hashtag και τα έχουμε ταξινομήσει.

Ετσι επιλέγουμε να εμφανίσουμε τα 10 πρώτα.



Εικόνα 5.23: Δημοφιλή Hashtags Κώδικας α

Εικόνα 5.24: Μέσος Δημοφιλή Hashtags κώδικας β

5.4 Ερωτήματα και περεταίρω διερεύνηση:

Είδαμε κάποια λίγες μόνο από τις δυνατότητες που προσφέρει το Spark Streaming. Στην παρούσα μελέτη όπως προαναφέρθηκε τα δεδομένα αντλούνται από το Twitter. Η πηγή θα μπορούσε να είναι διαφορετική. Αρχικά βλέπουμε πως μπορούμε να εκτυπώσουμε σχόλια χρηστών σε πραγματικό χρόνο, έπειτα πως μπορούμε να βρούμε το μέσο και το μέγιστο μήκος των σχολίων που εκτυπώνουμε, και τέλος να βρούμε τα 10 πιο δημοφιλή hashtags ανα δευτερόλεπτο.

Οι δυνατότητες της διαδικασίας όμως είναι πολύ περισσότερες. Θα μπορούσαμε να βρούμε και να εμφανίσουμε τα σχόλια των χρηστών που περιέχουν κάποιο συγκεκριμένο hashtag. Για παράδειγμα να βρούμε όλα τα σχόλια των χρηστών που περιέχουν το hashtag #covid19. Επίσης θα μπορούσαμε να αναζητήσουμε σχόλια με βάση λέξεις που εκδηλώνουν συναισθήματα χρηστών σε σχέση με τον Κορωνοιό ώστε να κάνουμε ανάλυση συναισθημάτων. Οι λέξεις μπορούν να βρεθούν από κατάλληλο λεξικό συναισθημάτων και μπορεί να υποδηλώνουν συναισθήματα όπως θυμός, φόβος, αγωνία, αμφισβήτηση, ελπίδα κ.α.

ΚΕΦΑΛΑΙΟ 6 ΣΥΜΠΕΡΑΣΜΑΤΑ - ΠΡΟΤΑΣΕΙΣ

Χωρίς τα big data το social media marketing δεν θα ήταν στο σημείο που είναι σήμερα. Όσο βελτιώνονται τα εργαλεία ανάλυσης οι marketers είναι σε θέση να πάρουν πιο στοχευμένες αποφάσεις σχετικά με την κατεύθυνση που θα ακολουθήσει μια επιχείρηση. Τα αποτελέσματα είναι πιο κερδοφόρα και επιτρέπουν στους πελάτες να έχουν μια πιο ευχάριστη εμπειρία. Από μια άποψη τα big data ανοίγουν το δρόμο για μια συναρπαστική νέα εμπειρία για επαγγελματίες και πελάτες. Με την πρόσβαση σε big data κοινωνικών δικτύων και με τη χρήση των κατάλληλων εργαλείων για την ανάλυση, η σκηνή ψηφιακού marketing θα αλλάξει ακόμη πάρα πολύ γρήγορα. Με την πρόσβαση σε στατιστικά στοιχεία σχετικά με την απόδοση των social media ενός brand, υπάρχει η δυνατότητα να δούμε διαφορετικά μοτίβα.

Οι αναλύσεις των social media είναι από τα πιο στοχευμένα παραδείγματα για το πως τα big data επηρεάζουν την καθημερινότητα μας. Ωστόσο, η αλήθεια είναι ότι τα big data είναι ένα ζωτικό εργαλείο για σχεδόν κάθε πτυχή της ζωής μας. Ακόμα και αν σε κάποια βιομηχανία δεν χρησιμοποιούνται σήμερα, δεν είναι απίθανο να χρησιμοποιηθούν πολύ πρόσφατα στο μέλλον. Με την άνθιση της τεχνολογίας, που φαίνεται από τα έξυπνα σπίτια και τις συσκευές που συλλέγουν στοιχεία και γίνονται όλο και μικρότερες και πιο αποτελεσματικές, είναι θέμα χρόνου να δούμε αλλαγή στην αντίληψή μας για τον κόσμο. Ένα μελλοντικό εργαλείο σε μια άλλη βιομηχανία μπορεί να συμβάλει στην ανακάλυψη από κάτι εκπληκτικό που δεν έχει περάσει ακόμη από το μυαλό μας σήμερα. Όλα βασίζονται στη σωστή συλλογή, την εύκολη παρουσίαση και τελικά στη σωστή ανάλυση των δεδομένων που μας παρουσιάζουν τα social media.

Τα analytics των social media τώρα, καλύπτουν ένα διεπιστημονικό πεδίο και χρησιμοποιούνται για διαφορετικούς σκοπούς από διάφορους ερευνητές. Υπάρχουν συνεπώς πολλές ερευνητικές προκλήσεις ακόμα αλλά και πολλές έχουν απαντηθεί και αφορούν τόσο την ανάλυση όσο και τη συλλογή των δεδομένων. Όλες οι έρευνες καταλήγουν πως κοινή πρόκληση είναι ο όγκος των δεδομένων. Είναι πολύ σημαντικό να αναφερθεί πως ο όγκος προκύπτει κυρίως εξαιτίας της φύσης των δεδομένων να δημιουργούνται σε πραγματικό χρόνο. Γι αυτό με τη σειρά του είναι πολύ σημαντικός και ο καθαρισμός των δεδομένων αυτών που συνήθως έχουν αδόμητο και αβέβαιο χαρακτήρα.

Σίγουρα, υπάρχουν κάποιες τεχνικές μηχανικής μάθησης που διαχειρίζονται τα δεδομένα αυτά πολύ καλύτερα από άλλες τεχνικές και τα ερμηνεύουν με πολύ μεγαλύτερη ακρίβεια. Απαιτείται περαιτέρω εργασία και μελέτη για τον προσδιορισμό των τεχνικών μηχανικής μάθησης που απαιτούν τροποποιήσεις για την ανάλυση κοινωνικών δικτύων και για τον συνδυασμό τους με τεχνικές εξόρυξης δεδομένων για καλύτερη απόδοση στην ανάλυση των κοινωνικών μέσων.

Με λίγα λόγια, από εταιρείες κολοσσούς έως πολιτικούς και ακαδημαϊκούς φορείς, τα big data αποτελούν αντικείμενο προσοχής και σε κάποιο βαθμό φόβου. Η

ξαφνική άνοδος τους έχει αφήσει πολλούς απροετοίμαστους. Πλέον μπορούμε να πούμε πως ο χώρος σιγά σιγά ωριμάζει αν και υπάρχει ακόμα περιθώριο βελτίωσης.

Η ανάλυση συναισθημάτων, όπως φάνηκε παραπάνω, έχει γίνει βασική τεχνολογία για την απόκτηση πληροφοριών από τα κοινωνικά δίκτυα. Το πεδίο έχει φτάσει σε ένα επίπεδο ωριμότητας που ανοίγει το δρόμο για την εκμετάλλευσή του σε πολλούς διαφορετικούς τομείς όπως το μάρκετινγκ, η υγεία, οι τραπεζικές συναλλαγές ή η πολιτική. Οι τελευταίες τεχνολογικές εξελίξεις, όπως η χρήση τεχνικών βαθιάς μάθησης, έχουν λύσει ορισμένες από τις παραδοσιακές προκλήσεις όπως η έλλειψη λεξικών πόρων. Οι τεχνολογίες ανάλυσης συναισθημάτων κάνουν δυνατή την αυτόματη ανάλυση των πληροφοριών που διανέμονται μέσω των μέσων κοινωνικής δικτύωσης για τον εντοπισμό της πολικότητας των δημοσιευμένων απόψεων (αν είναι αρνητικές ή όχι). Αυτές οι τεχνολογίες επεκτάθηκαν τα τελευταία χρόνια για να αναλύσουν άλλες πτυχές, όπως τη στάση ενός χρήστη απέναντι σε ένα θέμα ή τα συναισθήματα των χρηστών, ακόμη και για το συνδυασμό εργαλείων ανάλυσης κειμένων με άλλες εισόδους, όπως εργαλεία ανάλυσης πολυμέσων και κοινωνικών δικτύων.

Παραδοσιακά, η ανάλυση συναισθημάτων έχει επικεντρωθεί στην ανάλυση κειμένου χρησιμοποιώντας επεξεργασία φυσικής γλώσσας και τεχνικές μηχανικής μάθησης βασισμένες σε χαρακτηριστικά οντολογιών. Οι εξελίξεις σε κλάδους όπως οι τεχνολογίες των Big Data και της βαθιάς μάθησης έχουν επηρεάσει και ωφελήσει την εξέλιξη του πεδίου της ανάλυσης συναισθημάτων.

Τέλος το Spark προσφέρει μια πληθώρα δυνατοτήτων για την ανάλυση και ερμηνεία των Big Data. Το Spark Streaming και η επεξεργασία δεδο”μένων σε πραγματικό χρόνο λύνει τα χέρια στους ερευνητές και έρχεται να κουμπώσει τέλεια στην ίδια την φύση των δεδομένων που είναι να δημιουργούνται σε πραγματικό χρόνο. Το θέμα σίγουρα απαιτεί περεταίρω διερεύνηση.

Πηγές - Αναφορές:

1. Farzana Shaikh ,Afsha Khan, Firdaus Rangrez, Uzma Shaikh (2018). 'Social Media Analytics Based on big data' International Conference on Intelligent Computing and Control (I2C2) DOI: 10.1109/I2C2.2017.8321806

2. Christos Stergiou Kostas E. Psannis (2016). 'Recent advances delivered by Mobile Cloud Computing and Internet of Things for Big Data applications: a survey' 27(3) https://doi.org/10.1002/nem.1930

3. C. L. Stergiou; A. P. Plageras; K. E. Psanni (2018). 'Secure Machine Learning Scenario from Big Data in Cloud Computing via Internet of Things Networks' Handbook of Computer Networks and Cyber Security pp 525-554

4. Christos Stergiou; Kostas E. Psannis; Yutaka Ishibashi(2020). 'Green Cloud Communication System for Big Data Management' The 3rd World Symposium on Communication Engineering (WSCE 2020) DOI: 10.1109/WSCE51339.2020.9275579

5. Andreas Plageras; Kostas E. Psannis (2017). 'Algorithms for Big Data Delivery over the Internet of Things' 19th IEEE Conference on Business Informatics 2017 (CBI2017) DOI: 10.1109/CBI.2017.27

6. Avishek Saha, Young-Woon Lee, Young-Sup Hwang, Kostas E. Psannis, Byung-Gyu Kim, Context-aware Block-based Motion Estimation Algorithm for Multimedia Internet of Things (IoT) Platform, Personal and Ubiquitous Computing (Springer), February 2018, Volume 22, Issue 1, pp 163–172

7. Andreas P. Plageras, Kostas E. Psannis, Christos Stergiou, Haoxiang Wang, and B. B. Gupta, Efficient Sensor Big Data Collection-Processing and Analysis in Smart Buildings, Future Generation Computer Systems. Volume 82, May 2018, Pages 349-357

8. Andreas P. Plageras, Kostas E. Psannis, Christos Stergiou, Haoxiang Wang, and B. B. Gupta, Efficient Sensor Big Data Collection-Processing and Analysis in Smart Buildings, Future Generation Computer Systems. Volume 82, May 2018, Pages 349-357

9. Christos Stergiou; Kostas E.Psannis; Brij B.Gupta; Yutaka Ishibashi (2018) 'Security, privacy & efficiency of sustainable Cloud Computing for Big Data

& IoT' Sustainable Computing: Informatics and Systems Vol19, pp 174-184 https://doi.org/10.1016/j.suscom.2018.06.003

10. Andreas P. Plageras; Christos Stergiou; George Kokkonis; Kostas E. Psannis; Yutaka Ishibashi; Byung-Gyu Kim; B. Brij Gupta (2017). 'Efficient Large-scale Medical Data (eHealth Big Data) Analytics in Internet of Things' IEEE 19th Conference on Business Informatics (CBI), DOI**:** 10.1109/CBI.2017.3

11. Christos Stergiou, Kostas E. Psannis, Andreas P. Plageras, Theofanis Xifilidis, and B.B. Gupta, Security and Privacy of Big Data for Social Networking Services in Cloud, in Proceedings of IEEE conference on Computer Communications (IEEE INFOCOM 2018), Workshop on CCSNA: Cloud Computing Systems, Networks, and Applications, 15-20 April 2018, Honolulu, HI, USA

12. C. Stergiou, K. E. Psannis, Algorithms for Big Data in Advanced Communication Systems and Cloud Computing, 19th IEEE Conference on Business Informatics, Thessaloniki, Greece 24-26 July, 2017. [Award] & [link] [DOI: 10.1109/CBI.2017.28]

13. European Commission. 2018. What data can we process and under which conditions? Retrieved 14 March 2020, from https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/what-data-can-we-process-and-under-which-conditions_en
14. A. Nilsson, S. Smith, G. Ulm, E. Gustavsson, M. Jirstrand, "A Performance Evaluation of Federated Learning Algorithms", in Proceedings of DIDL '18: Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning, December 2018, pp. 1-8, Middleware '18: 19th International Middleware Conference Rennes France. [DOI: 10.1145/3286490.3286559]

15. Kostas E, Psannis, Christos Stergiou, and B. B. Gupta, Advanced Media-based Smart Big Data on Intelligent Cloud Systems, IEEE Transactions on Sustainable Computing (T-SUSC), June 2018.

16. Adrian Nilsson; Simon Smith; Gregor Ulm; Emil Gustavsson;Mats Jirstrand (2018). 'A Performance Evaluation of Federated Learning Algorithms' DIDL '18: Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning, DOI: 10.1145/3286490.3286559

17. Muhammad Nouman Noor, Farah Haneef (2020). ' A Review on big data and Social Network Analytics Techniques' Researchpedia Journal of Computing, Vol 1, Issue 1, Article 5, pp 39-49

18. Miltiades D. Lytras, Anna Visvizi (2019). 'Big data and Their Social Impact: Preliminary study' DOI: 10.3390/su11185067

19. Iman Raeesi Vanani, Setareh Majidian (2019), . 'Literature Review on big data Analytics Methods' Book: Social Media and Machine Learning DOI:10.5772/intechopen.86843

20. Agarwal, Ashish Sureka, Vikram Goyal (2015). 'Open Source Social Media Analytics for Intelligence and Security Informatics Applications' Springer, Cham ,Vol 9498, pp 21-37, https://doi.org/10.1007/978-3-319-27057-9_2

21. Muhammad Aslam Jarwar, Rabeeh Abbasi, Mubashar Mushtaq, Onaiza Maqbool (2017). 'CommuniMents: A Framework for Detecting Community Based Sentiments for Events' International journal on Semantic Web and information systems, Vol 13, pp 87-108

22. S. Magesh (2016). 'A Survey on Machine Learning Approaches to Social Media Analytics'

23. Vinay Kumar (2019). 'Sentiment Analysis Techniques for Social Media Data: A Review' https://www.researchgate.net/publication/336084873_Sentiment_Analysis_Techniques_for_Social_Media_Data_A_Review

24. Galen Panger (2015). 'Reassessing the Facebook Experiment: Critical Thinking About the Validity of Big Data Research' Information Communication and Society, 19(8):1108-1126 DOI: 10.1080/1369118X.2015.1093525

25. Ankur Goel; Jyoti Gautam; Sitesh Kumar (2016). 'Real time sentiment analysis of tweets using Naive Bayes' Publisher: IEEE International Conference on Next Generation Computing Technologies

26. MohammadAl-Smadi; MahmoudAl Ayyoub; YaserJararweh; OmarQawasmeh (2019). 'Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels reviews using morphological, syntactic and semantic features' 56(12) pp308-319

27. Hyoji Ha; Sang Woo Han; Seongmin Mun (2019). 'An Improved Study of Multilevel Semantic Network Visualization for Analyzing Sentiment Word of Movie Review Data' 9(12):2419, DOI: 10.3390/app9122419

28. Xingliang Mao;Shuai Chang; Jinjing Shi; Fangfang Li; Ronghua Shi;(2019). 'Sentiment-Aware Word Embedding for Emotion Classification' mdpi, *9*(7), 1334; https://doi.org/10.3390/app9071334

29. Mohammed Jabreel; Antonio Moreno (2019). 'A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets' *9*(6), 1123; https://doi.org/10.3390/app9061123

30. Seongik Park, Yanggon Kim (2016). "Building thesaurus lexicon using dictionary-based approach for sentiment classification" 2016 IEEE 14th International Conference, ISBN Information, DOI: 10.1109/SERA.2016.7516126

31. Hannah Kim; Young-Seob Jeong (2019). "Sentiment Classification Using Convolutional Neural Networks" MDPI 9(11), 2347; https://doi.org/10.3390/app9112347

32. Eline M. van den Broek-Altenburg; Adam J. Atherly (2019). 'Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season' MDPI 9(10), 2035 https://doi.org/10.3390/app9102035

33. Sunghee Park; Jiyoung Woo (2019); Gender Classification Using Sentiment Analysis and Deep Learning in a Health Web Forum' MDPI *9*(6), 1249; https://doi.org/10.3390/app9061249

34. Hui Liu; Yinghui Huang; Zichao Wang; Kai Liu; Xiangen Hu; Weijun Wang (2019) 'Personality or Value: A Comparative Study of Psychographic Segmentation Based on an Online Review Enhanced Recommender System' MDPI *9*(10), 1992; **https://doi.org/10.3390/app9101992**

35. Guadalupe Obdulia Gutiérrez-Esparza; Maite Vallejo-Allende; José Hernández-Torruco (2019). Classification of Cyber-Aggression Cases Applying Machine Learning MDPI *9*(9), 1828; https://doi.org/10.3390/app9091828

36. Hui Liu; Yinghui Huang; Zichao Wang; Kai Liu; Xiangen Hu; Weijun Wang (2019). 'Personality or Value: A Comparative Study of Psychographic Segmentation Based on an Online Review Enhanced Recommender System' MDPI *9*(10), 1992; **https://doi.org/10.3390/app9101992**

37. Muqtar Unnisa; Ayesha Ameen; Syed Raziuddin (2016) "Opinion Mining on Twitter Data using Unsupervised Learning Technique" International Journal of Computer Applications 148(12) DOI: 10.5120/ijca2016911317

38. Alexander Piazza, Christian Zagel, Julia Haeske, Freimut Bodendorf (2018). 'Do You Like According to Your Lifestyle? A Quantitative Analysis of the Relation Between Individual Facebook Likes and the Users' Lifestyle' Book: Advances in The Human Side of Service Engineering, Springer International Publishing

39. Luis Jimenez-Marquez; Gonzalez-Carrasco; Jose Luis Lopez Cuadrado; Belen Ruiz-Mezcua (2019); 'Towards a big data framework for analyzing social media content' Vol 44 pp1-12, https://doi.org/10.1016/j.ijinfomgt.2018.09.003

40. Stefan Stieglitz; Milad Mirbabaie; Björn Ross; Christoph Neuberger (2018); 'Social media analytics – Challenges in topic discovery, data collection, and data preparation' International Journal of Information Management Vol39, pp156-168 https://doi.org/10.1016/j.ijinfomgt.2017.12.002

41. Igor Kotenko; Andrey Chechulin; Dmitry Komashinsky (2015); 'Evaluation of text classification techniques for inappropriate web content blocking', 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), DOI: 10.1109/IDAACS.2015.7340769

42. Troy Segal, Big Data, Jan 2021
Available at: https://www.investopedia.com/terms/b/big-data.asp

43. Bridget Botelho; Stephen J. Bigelow, Big Data Definition, Oct 2019
Available at: https://searchdatamanagement.techtarget.com/definition/

44. University of Wisconsin, What is big data?
Available at: https://datasciencedegree.wisconsin.edu/data-science/what-is-big-data/

45. Roberta Nicora, How is big data impacting social media?, Dec 2019
Available at: https://medium.com/dative-io/how-is-big-data-impacting-social-media-df31aa3f66f6

46. Keith D. Foote, Big Data Trends in 2020, Jan 2020 Available at: https://www.dataversity.net/big-data-trends-in-2020/

47. Editorial Team, 4 Major Ways in Which Big Data is Impacting Social Media Marketing, Oct 2018 Available at: https://insidebigdata.com/2018/10/06/4-major-ways-big-data-impacting-social-media-marketing/

48. Big Data Analytics Industry Report 2020 - Rapidly Increasing Volume & Complexity of Data, Cloud-Computing Traffic, and Adoption of IoT & AI are Driving Growth, March 2020 Available at: https://www.globenewswire.com/news-release/2020/03/02/1993369/0/en/Big-Data-Analytics-Industry-Report-2020-Rapidly-Increasing-Volume-Complexity-of-Data-Cloud-Computing-Traffic-and-Adoption-of-IoT-AI-are-Driving-Growth.html

49. Adam Coombs, Understanding Sentiment Analysis in Social Media Monitoring, 2017, Available at: https://unamo.com/blog/social/sentiment-analysis-social-media-monitoring

50. Christina Newberry, How to Conduct a Social Media Sentiment Analysis (Tools + Free Template), Oct 2020, Available at: https://blog.hootsuite.com/social-media-sentiment-analysis-tools/

51. **https://developer.twitter.com/en**

52. **https://spark.apache.org/docs/latest/cluster-overview.html**

53. **https://www.freecodecamp.org/news/learning-scala-from-0-60-part-i-dc095d274b78/**

54. **https://www.tutorialspoint.com/scala/index.htm**

55. **https://www.udemy.com/course/taming-big-data-with-spark-streaming-hands-on/**

56. **https://www.udemy.com/course/taming-big-data-with-spark-streaming-hands-on/learn/lecture/5011010#questions/12066554**

57. **https://www.oreilly.com/library/view/learning-scala/9781449368814/**