ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Διπλωματική Εργασία

ΑΝΙΧΝΕΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΕ ΔΕΔΟΜΕΝΑ ΚΟΙΝΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΜΕΣΩ ΕΞΟΡΥΞΗΣ ΚΑΙ ΑΝΑΛΥΣΗΣ

της

ΓΕΩΡΓΙΑΣ ΧΑΡΙΣΗ

Θεσσαλονίκη, Σεπτέμβριος 2018

Ευχαριστίες

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα Αναπληρωτή Καθηγητή κ. Νικόλαο Πρωτόγερο για την πολύτιμη βοήθεια και συνεχή καθοδήγηση που μου παρείχε σε όλα τα στάδια εκπόνησης της διπλωματικής μου εργασίας.

Τέλος, θα ήθελα να εκφράσω ένα μεγάλο ευχαριστώ στην οικογένειά μου και τους φίλους για την διαρκή υποστήριξη και συνεχή ενθάρρυνση καθ' όλη τη διάρκεια συγγραφής της εργασίας χωρίς τους οποίους δεν θα ήταν εφικτή η υλοποίησή της.

Περίληψη

Κατά την διάρκεια των τελευταίων ετών, η ευρεία χρήση των μέσων κοινωνικής δικτύωσης έχει οδηγήσει στην ταχεία παραγωγή ενός μεγάλου όγκου δεδομένων, η αξιοποίηση των οποίων παρουσιάζει ιδιαίτερο ενδιαφέρον. Στόχος της παρούσας διπλωματικής εργασίας, είναι να εξετάσει μέσω τεχνικών εξόρυξης και ανάλυσης εάν είναι εφικτό να αποδοθούν στοιχεία συναισθήματος στα δεδομένα αυτά.

Για την επίτευξη του στόχου, αναπτύχθηκε με τη χρήση του οικοσυστήματος Hadoop, ένα πρωτότυπο σύστημα το οποίο εστιάζει στην πλατφόρμα του Twitter και θα χρησιμοποιηθεί για να διερευνηθεί η δυνατότητα αναγνώρισης συναισθημάτων σε αναρτήσεις χρηστών.

Αρχικά, μελετήθηκε η διεθνής βιβλιογραφία σχετικά με τα ερευνητικά πεδία της εργασίας όπως οι τεχνικές εξόρυξης δεδομένων, τα μέσα κοινωνικής δικτύωσης και η ανάλυση συναισθήματος. Στη συνέχεια, αναπτύχθηκε η μεθοδολογία έρευνας όπου και καταγράφηκαν ο τρόπος ανάπτυξης των αλγορίθμων ανίχνευσης του συναισθήματος και τα κριτήρια αξιολόγησής τους. Προσδιορίστηκαν οι προδιαγραφές τις οποίες θα πρέπει να πληροί το πρωτότυπο σύστημα ώστε να είναι εφικτή η υλοποίηση και αξιολόγηση των αλγορίθμων και αναζητήθηκαν τα κατάλληλα εργαλεία λογισμικού για να υλοποιηθεί.

Τα αποτελέσματα της αξιολόγησης του συστήματος επιβεβαιώνουν τον ισχυρισμό, ότι με την χρήση τεχνικών εξόρυξης και ανάλυσης μπορεί να αποδοθεί συναίσθημα σε δεδομένα χρηστών από κοινωνικά δίκτυα.

Λέξεις Κλειδιά: Κοινωνικά δίκτυα, Εξόρυξη Δεδομένων, Ανάλυση Συναισθήματος, Twitter, Twitter API, Apache Hadoop, Apache Flume, Apache Hive

Abstract

Over the last few years, widespread use of social media has led to the rapid production of a large volume of data, the exploitation of which is of particular interest. The aim of this thesis is to examine through mining and analysis techniques whether it is feasible to attribute sentiment to these data.

To achieve this goal, an original system was developed using the Hadoop ecosystem that focuses on the Twitter platform and will be used to explore the potential for recognizing sentiment in users' posts.

Initially, is presented a review of the international literature on research fields such as social media networks, data mining and sentiment analysis techniques. The research methodology is then introduced containing the way of developing the sentiment detection algorithms and their evaluation criteria. Afterwards, were identified the specifications that the original system should meet to allow for the implementation and evaluation of the algorithms as well the appropriate software tools to be realized.

The results of the system's assessment confirm the claim that it is possible to attribute sentiment to user data from social networks, using mining and analysis techniques.

Key words: Social Media, Data Mining, Sentiment Analysis, Twitter, Twitter API, Apache Hadoop, Apache Flume, Apache Hive

Πίνακας Περιεχομένων

Ευχαριστίες ii

Περίληψη iii

Abstract iv

Πίνακας Περιεχομένων v

Πίνακας Εικόνων viii

1 Εισαγωγή 1

1.1 Κίνητρο – Περιγραφή προβλήματος 1

1.2 Αντικείμενο και Στόχος διπλωματικής εργασίας 1

1.3 Διάρθρωση διπλωματικής εργασίας 2

2 Βιβλιογραφική Ανασκόπηση 4

2.1 Κοινωνικά δίκτυα 4

2.2 Εξόρυξη δεδομένων 5

2.2.1 Διαδικασία εξόρυξης δεδομένων 6

2.2.2 Τεχνικές εξόρυξης δεδομένων 6

2.3 Εξόρυξη δεδομένων από κοινωνικά δίκτυα 8

2.3.1 Κίνητρα εξόρυξης δεδομένων από κοινωνικά δίκτυα 8

2.4 Ανάλυση συναισθήματος και εφαρμογές 10

2.4.1 Τεχνικές ανάλυσης συναισθήματος 10

2.4.2 Αξιολόγηση τεχνικών ανάλυσης συναισθήματος 13

2.4.3 Ανασκόπηση εφαρμογών ανάλυσης συναισθήματος 14

3 Μεθοδολογία 17

3.1 Πλαίσιο έρευνας 17

3.2 Χαρακτηριστικά πηγής δεδομένων 17

3.3 Χαρακτηριστικά συστήματος 17

3.4 Μεθοδολογία ανάλυσης συναισθήματος 18

3.5 Αξιολόγηση αποτελεσμάτων 18
4 Αποτελέσματα 19
4.1 Επιλογή τεχνολογιών 19
4.1.1 Twitter API 20
4.1.2 Apache Hadoop 21
4.1.2.1 Hadoop Distributed File System – HDFS 22
4.1.2.2 Hadoop MapReduce 24
4.1.2.3 Hadoop YARN 25
4.1.2.4 Οικοσύστημα Hadoop 28
4.1.2.5 Apache Flume 28
4.1.2.6 Apache Hive 30
4.2 Σχεδίαση συστήματος 31
4.2.1 Δομές δεδομένων 34
4.2.2 Υλικό – Λογισμικό 36
4.3 Υλοποίηση συστήματος 36
4.3.1 Δημιουργία Twitter App 36
4.3.2 Διαμόρφωση Flume για εξόρυξη δεδομένων 38
4.3.3 Δημιουργία πινάκων για προ-επεξεργασία δεδομένων 41
4.3.4 Δημιουργία πινάκων για προετοιμασία και καθαρισμό δεδομένων 42
4.3.5 Κατασκευή αλγορίθμου για απόδοση συναισθήματος 44
4.4 Εκτέλεση – Αποτελέσματα συστήματος 45
4.5 Αξιολόγηση αποτελεσμάτων συστήματος 47
4.6 Αποτελέσματα έρευνας 47
5 Συμπεράσματα και Προτάσεις 48
5.1 Συμπεράσματα 48
5.2 Προτάσεις – Μελλοντικές επεκτάσεις 48
Βιβλιογραφία 50

Παράρτημα 57

Πίνακας Εικόνων

Εικόνα 1: Δημιουργία Δεδομένων σε Ιστοτόπους Κοινωνικής Δικτύωσης – 2017 4
Εικόνα 2: Κατάταξη μέσων κοινωνικής δικτύωσης με βάση τον αριθμό (σε εκατομμύρια) των ενεργών χρηστών - 2017 5
Εικόνα 3: Διαδικασία εξόρυξης δεδομένων (R.Nayak, 2002) 6
Εικόνα 4: HDFS Architecture 23
Εικόνα 5: MapReduce Architecture 25
Εικόνα 6: YARN Architecture 27
Εικόνα 7: Apache Hadoop Ecosystem 28
Εικόνα 8: Flume Agent 29
Εικόνα 9: Αρχιτεκτονική Συστήματος 32
Εικόνα 10: Δημιουργία Twitter Application 37
Εικόνα 11: Δημιουργία Access Token 37
Εικόνα 12: Twitter Application 38
Εικόνα 13: Flume Agent 39
Εικόνα 14: Δημιουργία path στον HDFS 40
Εικόνα 15: Αρχείο Διαμόρφωσης Flume Agent 40
Εικόνα 16: Δημιουργία πίνακα twitter_data 42
Εικόνα 17: Δημιουργία view tweet_simple 42
Εικόνα 18: Δημιουργία view split_words 43
Εικόνα 19: Δημιουργία view explode_word 43
Εικόνα 20: Δημιουργία πίνακα stopwords 43
Εικόνα 21: Δημιουργία view clean_word1 43
Εικόνα 22: Δημιουργία view clean_word2 44
Εικόνα 23: Δημιουργία view clean_word3 44
Εικόνα 24: Δημιουργία view clean_word4 44
Εικόνα 25: Δημιουργία πίνακα dictionary 44
Εικόνα 26: Δημιουργία view word_join 45
Εικόνα 27: Δημιουργία πίνακα tweet_sentiment 45
Εικόνα 28: Εκτέλεση Flume Agent 45
Εικόνα 29: Hadoop Administration Web UI 46
Εικόνα 30: Εξαγωγή αποτελεσμάτων 46
Εικόνα 31: Αποτελέσματα συστήματος 47

1 ΕΙΣΑΓΩΓΗ

1.1 Κίνητρο – Περιγραφή προβλήματος

Η ακμάζουσα δυναμικότητα και η ευρεία χρήση των μέσων κοινωνικής δικτύωσης που παρατηρείται κατά τη διάρκεια των τελευταίων ετών, έχουν ως συνέπεια την παραγωγή δεδομένων μεγάλου όγκου τα οποία επιχειρήσεις διάφορων κλάδων επιχειρούν να αξιοποιήσουν για την εξαγωγή χρήσιμων συμπερασμάτων αναφορικά με το πελατειακό κοινό, τα προϊόντα και τις υπηρεσίες τους.

Τα μέσα κοινωνικής δικτύωσης, μεταξύ άλλων, επιτρέπουν στους χρήστες να αναρτούν, να τροποποιούν και να βαθμολογούν περιεχόμενο καθώς επίσης και να εκφράζουν τις απόψεις τους πάνω σε συγκεκριμένα θέματα. Στο πλαίσιο αυτό, μέσα κοινωνικής δικτύωσης όπως το Facebook, το Twitter και το Instagram, παρέχουν στα στελέχη επιχειρήσεων την δυνατότητα να εξερευνούν τις προτιμήσεις, τις απόψεις και τη συμπεριφορά των καταναλωτών μέσω της εξέτασης του περιεχομένου που καθημερινά παράγεται από τους χρήστες (User Generated Content – UGC).

Στον επιχειρηματικό κόσμο, όπου το διαδίκτυο και τα μέσα κοινωνικής δικτύωσης κυριαρχούν, η εξερεύνηση, η ανάλυση και η κατανόηση της γνώμης των καταναλωτών μέσω της εξόρυξης των κοινωνικών δικτύων και η δυνατότητα μετατροπής της εξαγόμενης πληροφορίας σε δυναμική νοημοσύνη η οποία αυξάνει την αποτελεσματικότητα των προσπαθειών μάρκετινγκ μιας επιχείρησης, αποτελούν διαδικασίες ιδιαίτερα σημαντικές.

Ως εκ τούτου, οι τεχνικές εξόρυξης και ανάλυσης δεδομένων κοινωνικών δικτύων, έχουν γνωρίσει ιδιαίτερη άνθηση. Ωστόσο, ο μεγάλος όγκος που τα χαρακτηρίζει, καθιστά αδύνατη τη χειροκίνητη ανάλυσή τους και προϋποθέτει τη χρήση ισχυρών εργαλείων τα οποία ενδείκνυνται για την ανάλυση και τον μετασχηματισμό δεδομένων μεγάλου όγκου.

1.2 Αντικείμενο και Στόχος διπλωματικής εργασίας

Αντικείμενο της παρούσας διπλωματικής εργασίας αποτελεί η εξόρυξη και ανάλυση δεδομένων από κοινωνικά δίκτυα. Το ερευνητικό ερώτημα στο οποίο καλείται να απαντήσει είναι: «Είναι εφικτό να ανιχνευτεί συναίσθημα σε δημοσιεύσεις χρηστών σε ένα κοινωνικό δίκτυο με την αξιοποίηση τεχνικών εξόρυξης και ανάλυσης δεδομένων;»

Για την απάντηση του ερωτήματος αυτού, επιχειρείται η ανάπτυξη ενός συστήματος, το οποίο επικεντρώνεται στην εξόρυξη δεδομένων μεγάλου όγκου προερχόμενα από κοινωνικά δίκτυα σε μη δομημένη μορφή. Στο σύστημα αυτό, θα εφαρμοστούν και θα δοκιμαστούν αλγόριθμοι που θα αναλύσουν δεδομένα κειμένου προκειμένου να ανιχνεύσουν το συναίσθημα των χρηστών που μπορεί να εμπεριέχεται στις αναρτήσεις τους.

Κύρια πηγή δεδομένων του συστήματος, αποτελούν τα μη δομημένα δεδομένα που προέρχονται από την πλατφόρμα του Twitter. Το σύνολο των δεδομένων, συλλέγονται με το εργαλείο ανοιχτού κώδικά Apache Flume και εν συνεχεία διοχετεύονται στο κατανεμημένο σύστημα αρχείων HDFS (Hadoop Distributed File System) του Apache Hadoop.

Η υλοποίηση των αλγορίθμων για την επεξεργασία και ανάλυση των δεδομένων, γίνεται με την βοήθεια του λογισμικού Apache Hive το οποίο επιτρέπει την κατανεμημένη επεξεργασία δεδομένων σε πραγματικό χρόνο. Στο στάδιο αυτό, με τη χρήση εντολών Hive, τα μη δομημένα δεδομένα μετασχηματίζονται ώστε να μπορούν να επεξεργαστούν και με τη βοήθεια ενός λεξικού αναφοράς, σε κάθε tweet προσδίδεται μια πολικότητα ανάλογα με το συναίσθημα που εκφράζει.

Τέλος, παρουσιάζεται η αξιολόγηση των αποτελεσμάτων του συστήματος το οποίο κατάφερε να αποδώσει σωστά το συναίσθημα στο 71% των δεδομένων που αναλύθηκαν απαντώντας θετικά στο ερευνητικό ερώτημα.

1.3 Διάρθρωση διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία αποτελείται από πέντε (5) κεφάλαια. Έπειτα από την εισαγωγή όπου αναφέρονται τα κίνητρα και οι στόχοι της εργασίας, ακολουθεί το δεύτερο κεφάλαιο στο οποίο παρουσιάζεται το θεωρητικό υπόβαθρο της εξόρυξης και ανάλυσης δεδομένων από κοινωνικά δίκτυα και γίνεται μια βιβλιογραφική ανασκόπηση των εφαρμογών ανάλυσης συναισθήματος.

Στο τρίτο κεφάλαιο, παρατίθεται η μεθοδολογία και οι επιλογές που έγιναν για την ανάπτυξη του συστήματος. Ταυτόχρονα, περιγράφεται ο τρόπος αξιολόγησης των αποτελεσμάτων.

Στο τέταρτο κεφάλαιο, παρουσιάζονται τα εργαλεία και οι τεχνολογίες που χρησιμοποιήθηκαν και υλοποιείται το σύστημα που αναπτύχθηκε. Στη συνέχεια,

διατίθενται τα αποτελέσματα και η αξιολόγησή του και τέλος απαντάται το ερευνητικό ερώτημα.

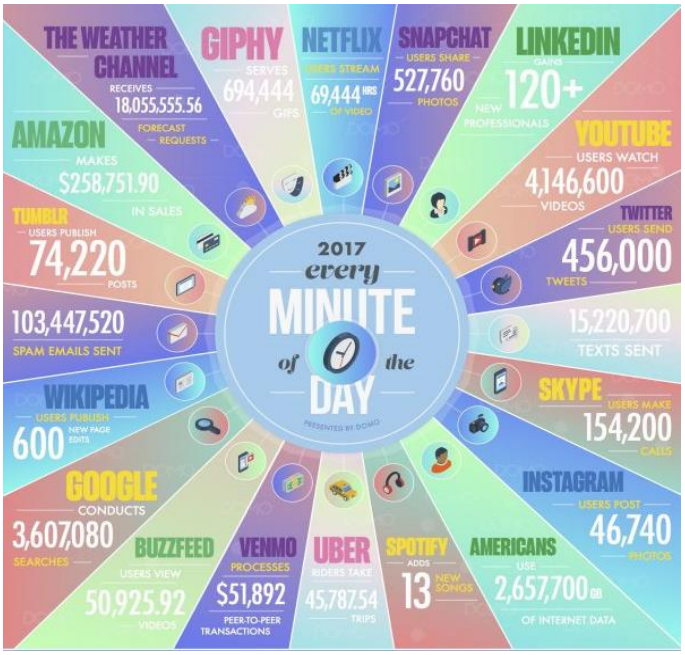
Στο πέμπτο και τελευταίο κεφάλαιο, παρατίθενται τα συμπεράσματα της εργασίας καθώς και προτάσεις για μελλοντική επέκταση του συστήματος.

2 ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

2.1 Κοινωνικά δίκτυα

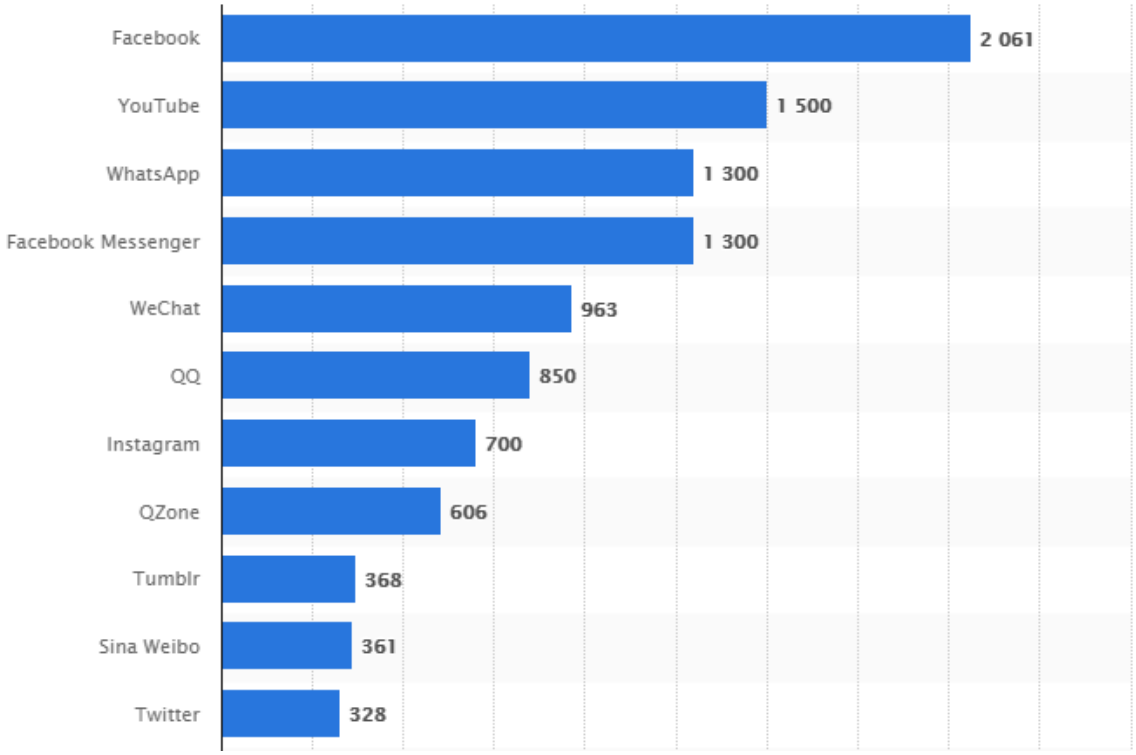
Τα μέσα κοινωνικής δικτύωσης, αποτελούν ένα δυναμικά αναπτυσσόμενο πεδίο το οποίο παρέχει πολύπλευρες δυνατότητες στους χρήστες τόσο σε προσωπικό όσο και επιχειρησιακό επίπεδο. Αποτελούν μια ομάδα διαδικτυακών εφαρμογών που βασίζονται στα ιδεολογικά και τεχνολογικά θεμέλια του Web 2.0 και επιτρέπουν τη δημιουργία και την ανταλλαγή περιεχομένου που δημιουργεί ο χρήστης (Kaplan & Haenlein, 2010).

Οι ιστότοποι κοινωνικής δικτύωσης, είναι κοινώς γνωστοί για τη διάδοση πληροφοριών, τη δημοσίευση προσωπικών δραστηριοτήτων, τον σχολιασμό προϊόντων και υπηρεσιών, τον διαμοιρασμό φωτογραφιών, τη δημιουργία επαγγελματικών προφίλ, τη διαφήμιση και την έκφραση απόψεων και συναισθημάτων (M. Adedoyin-Olowe et al. 2011). Το γεγονός αυτό, τα καθιστά μια σημαντική πηγή αλληλεπίδρασης (B. Liu, 2012) που επιτρέπουν την ταχεία παραγωγή και ανταλλαγή πληροφοριών μεταξύ των χρηστών ανεξαρτήτως τοποθεσίας.



Εικόνα 1: Δημιουργία Δεδομένων σε Ιστοτόπους Κοινωνικής Δικτύωσης – 2017

Τα πιο δημοφιλή σε παγκόσμιο επίπεδο δίκτυα με βάση τον αριθμό των ενεργών λογαριασμών είναι το Facebook, το YouTube, το WhatsApp και το Facebook Messenger (Statista, 2017) .



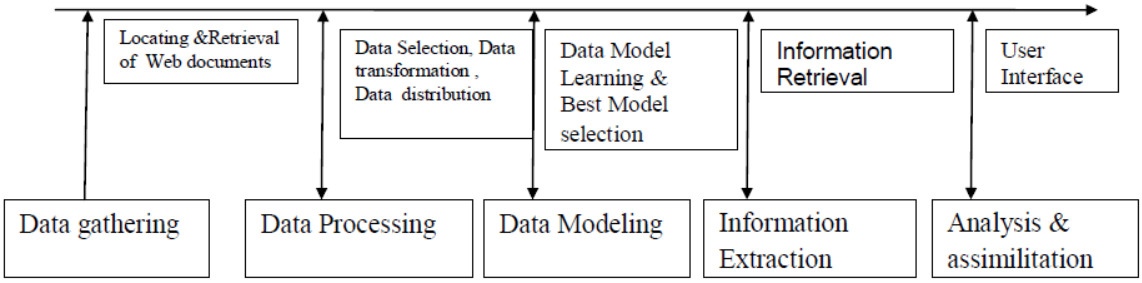
Εικόνα 2: Κατάταξη μέσων κοινωνικής δικτύωσης με βάση τον αριθμό (σε εκατομμύρια) των ενεργών χρηστών - 2017

2.2 Εξόρυξη δεδομένων

Η εξόρυξη δεδομένων μπορεί να οριστεί ως μια διαδικασία ανακάλυψης χρήσιμης ή ενεργής γνώσης από δεδομένα μεγάλης κλίμακας (Tan et al. 2006). Βασικός στόχος κατά την ανάπτυξη τεχνικών εξόρυξης δεδομένων, αποτελεί η διαχείριση μεγάλων βάσεων δεδομένων για τον εντοπισμό χρήσιμων μοτίβων (Gürsoy et al. 2017). Η εξόρυξη δεδομένων παρέχει ένα ευρύ φάσμα τεχνικών για την ανίχνευση χρήσιμης γνώσης από μαζικά σύνολα δεδομένων όπως τάσεις, πρότυπα και κανόνες. Αυτά τα πρότυπα και οι τάσεις απασχολούν τις επιχειρήσεις και μπορούν να αξιοποιηθούν κατά το σχεδιασμό της επιχειρησιακής στρατηγικής ή την εισαγωγή νέων προγραμμάτων (Dr. B. Umadeni & P. Surya, 2017).

2.2.1 Διαδικασία εξόρυξης δεδομένων

Η εξόρυξη δεδομένων είναι επίσης γνωστή ως ανακάλυψή γνώσης από δεδομένα (Knowledge Discovery from Data – KDD) καθώς περιγράφει την τυπική διαδικασία εξαγωγής χρήσιμων πληροφοριών από ακατέργαστα δεδομένα (J. Han et al. 2006). Σύμφωνα με τον R. Nayak (2002) ο εντοπισμός ενός προβλήματος σύμφωνα με τους στόχους και τα ενδιαφέροντα του αναλυτή αποτελεί την εκκίνηση μιας τυπικής διαδικασίας εξόρυξης γνώσης. Στη συνέχεια, ακολουθεί ο εντοπισμός των πηγών πληροφόρησης και από τα συσσωρευμένα δεδομένα παράγεται ένα υποσύνολο στο οποίο εφαρμόζονται οι τεχνικές εξόρυξης. Για να διασφαλιστεί η ποιότητα, το σύνολο των δεδομένων προ-επεξεργάζεται ώστε να αφαιρεθεί ο θόρυβος και να εντοπιστούν τυχόν ελλείψεις και μετασχηματίζεται ώστε να λάβει την κατάλληλη μορφή. Σε επόμενο στάδιο, μια τεχνική εξόρυξης δεδομένων ή ένας συνδυασμός τεχνικών κατάλληλες για τον τύπο της γνώσης που πρόκειται να ανακαλυφθεί εφαρμόζεται στο σύνολο δεδομένων. Η εξαγόμενη γνώση αξιολογείται και ερμηνεύεται συνήθως με ορισμένες τεχνικές απεικόνισης. Τέλος, οι πληροφορίες παρουσιάζονται στο χρήστη ώστε να ενσωματωθούν στις επιχειρησιακές στρατηγικές της εταιρείας.



Εικόνα 3: Διαδικασία εξόρυξης δεδομένων (R.Nayak, 2002)

2.2.2 Τεχνικές εξόρυξης δεδομένων

Οι τεχνικές εξόρυξης δεδομένων μπορούν να ταξινομηθούν σε διάφορες κατηγορίες εκ των οποίων οι πλέον διαδεδομένες είναι οι κατηγορίες επιβλεπόμενης και μη επιβλεπόμενης μάθησης. Η επιβλεπόμενη προσέγγιση βασίζεται σε γνώση η οποία είναι γνωστή εκ των προτέρων (π.χ. ετικέτες κλάσης) ενώ οι αλγόριθμοι μη επιβλεπόμενης μάθησης χρησιμοποιούνται για τον χαρακτηρισμό δεδομένων χωρίς προηγούμενη γνώση ως προς τα είδη των μοτίβων που πρόκειται να αποκαλυφθούν (G. Barbier & H. Liu, 2011; P. Gundecha & H. Liu, 2012).

Η κατηγοριοποίηση (Classification) αποτελεί μια κοινή επιβλεπόμενη προσέγγιση και είναι κατάλληλη όταν το σύνολο ή ένα μικρό μέρος των δεδομένων φέρει ετικέτες. Οι αλγόριθμοι κατηγοριοποίησης εκπαιδεύονται σε ένα σύνολο δεδομένων το οποίο περιλαμβάνει προκαθορισμένες κλάσεις για κάθε στοιχείο του συνόλου. Στη συνέχεια, δημιουργούν μοντέλα ταξινόμησης τα οποία αυτόματα κατατάσσουν νέα στοιχεία στις κατηγορίες που προέρχονται από τα δεδομένα εκπαίδευσης. Για την αξιολόγηση της απόδοσης, τα μοντέλα εφαρμόζονται σε δεδομένα δοκιμής ώστε να επιτευχθεί η ακρίβεια της κατηγοριοποίησης. Τα δέντρα αποφάσεων (Decision Trees), ο K-nearest Neighbors, ο αλγόριθμος Naive Bayes, η μηχανή διανυσματικής υποστήριξης (Support Vector Machine - SVM) και οι κανόνες κατάταξης (Classification Rules) αποτελούν παραδείγματα εποπτευόμενων τεχνικών κατηγοριοποίησης.

Η ομαδοποίηση (Clustering) αποτελεί μια τεχνική μη επιβλεπόμενης εξόρυξης δεδομένων που είναι χρήσιμη όταν αντιμετωπίζονται σύνολα δεδομένων χωρίς ετικέτες. Σε αντίθεση με τους αλγόριθμους κατηγοριοποίησης, οι αλγόριθμοι ομαδοποίησης δεν εξαρτώνται από δεδομένα εκπαίδευσης για την ανάπτυξη μοντέλων. Δεδομένου ενός πλαισίου, οι αλγόριθμοι ομαδοποίησης δημιουργούν μοντέλα με βάση την ομοιότητα ή την ανομοιότητα μεταξύ των δεδομένων οι οποίες μπορούν να υπολογιστούν με μέτρα εγγύτητας όπως η ευκλείδεια απόσταση. Ο K-means, η ιεραρχική ομαδοποίησης και η ομαδοποίηση με βάση την πυκνότητα αποτελούν τυπικά παραδείγματα μη επιβλεπόμενων τεχνικών ομαδοποίησης.

Η εξόρυξη δεδομένων εκτός από τις τεχνικές κατηγοριοποίησης και ομαδοποίησης περιλαμβάνει και άλλες τεχνικές όπως οι κανόνες συσχέτισης, η παλινδρόμηση και τα πρότυπα ακολουθιών.

Οι τεχνικές εξόρυξης κανόνων συσχέτισης (Association Rules) χρησιμοποιούνται για την ανεύρεση συχνών μοτίβων και συσχετίσεων μεταξύ των στοιχείων του συνόλου δεδομένων (M. Eirinaki, 2008). Στόχος αποτελεί ο εντοπισμός ισχυρών κανόνων από βάσεις δεδομένων μέσω της χρήσης ορισμένων μέτρων ενδιαφέροντος. Ένας κανόνας συσχέτισης λαμβάνει συνήθως τη μορφή X → Y όπου τα X και Y αντιπροσωπεύουν διαφορετικά σύνολα στοιχείων. Η ισχύς ενός κανόνα συσχέτισης μπορεί να μετρηθεί σε όρους υποστήριξης και εμπιστοσύνης. Η υποστήριξη καθορίζει το πόσο συχνά εφαρμόζεται ένας κανόνας σε ένα συγκεκριμένο σύνολο δεδομένων ενώ η εμπιστοσύνη καθορίζει το πόσο συχνά τα στοιχεία του συνόλου Y εμφανίζονται σε συναλλαγές που περιέχουν στοιχεία του συνόλου X (P. N. Tan et al., 2006).

Τα πρότυπα ακολουθιών (Sequential Patterns) αποτελούν μια εξέλιξη των τεχνικών εξόρυξης κανόνων συσχέτισης υπό την έννοια ότι ενσωματώνουν την ιδέα της χρονικής ακολουθίας (M. Eirinaki, 2008). Οι τεχνικές εξόρυξης μοτίβων χρησιμοποιούνται για την ανακάλυψη χρήσιμων και απροσδόκητων προτύπων από βάσεις δεδομένων στις περιπτώσεις όπου η χρονική ακολουθία των γεγονότων λαμβάνεται υπόψη. (P. Fournier-Viger et al., 2017).

Η παλινδρόμηση (Regression) είναι μια στατιστική μεθοδολογία που χρησιμοποιείται κυρίως στην ανάλυση αριθμητικών προβλέψεων (J. Han, 2006). Αποτελεί την πλέον δημοφιλή μέθοδο πρόβλεψης της ικανοποίησης των πελατών και σε σύγκριση με τις υπόλοιπες τεχνικές είναι η πιο κατάλληλη για την εξόρυξη συνόλων με λιγότερα από 1000 στοιχεία (S. Bhanap & S. Kawthekar, 2015).

2.3 Εξόρυξη δεδομένων από κοινωνικά δίκτυα

Τα κοινωνικά δίκτυα, αποτελούν μια σημαντική πηγή εξόρυξης δεδομένων, με έντονο ερευνητικό ενδιαφέρον. Σύμφωνα με τους Gürsoy et al. (2017), η εξόρυξη των κοινωνικών δικτύων ορίζεται ως η διαδικασία αναπαράστασης, ανάλυσης και εξαγωγής ενεργών προτύπων από δεδομένα μέσων κοινωνικής δικτύωσης η οποία εισάγει βασικές έννοιες και αλγορίθμους κατάλληλους για τη διερεύνηση μεγάλων σε όγκο δεδομένων. Χρησιμοποιεί μια σειρά από βασικές θεωρίες και μεθοδολογίες της επιστήμης των υπολογιστών, της εξόρυξης δεδομένων, της μηχανικής μάθησης, της ανάλυσης των κοινωνικών δικτύων, της επιστήμης δικτύων, της κοινωνιολογίας και της στατιστικής. Η εξόρυξη των κοινωνικών δικτύων, περιλαμβάνει τα εργαλεία για την τυπική αναπαράσταση, μέτρηση, μοντελοποίηση και την εξόρυξη σημαντικών προτύπων από δεδομένα μεγάλης κλίμακας που προέρχονται από τα μέσα κοινωνικής δικτύωσης (Zafarani et al., 2014).

2.3.1 Κίνητρα εξόρυξης δεδομένων από κοινωνικά δίκτυα

Τα ψηφιακά μέσα μπορούν να ξεπεράσουν τα όρια του φυσικού κόσμου για να μελετήσουν τις ανθρώπινες σχέσεις (H. Lauw et al., 2010) και να συμβάλουν στον προσδιορισμό του συναισθήματος που αποδίδεται στους περιφερειακούς πληθυσμούς χωρίς ρητές έρευνες.

Τα μέσα κοινωνικής δικτύωσης καταγράφουν αποτελεσματικά τις ιογενείς τάσεις του μάρκετινγκ και αποτελούν την πλέον κατάλληλη πηγή μελέτης για την καλύτερη κατανόηση και αξιοποίηση των μηχανισμών επιρροής (P. Domingos & M. Richardson,

2011). Επιχειρήσεις με έντονη παρουσία στα μέσα κοινωνικής δικτύωσης, έχουν στη διάθεσή τους μεγάλα σύνολα δεδομένων η ανάλυση των οποίων τους επιτρέπει να διαχειριστούν αποτελεσματικά τις προτιμήσεις των πελατών, να βελτιώσουν τις επιδόσεις τους και να παραμείνουν μπροστά στον ανταγωνισμό (S. Balan, J. Rege, 2017). Επιπλέον, η μελέτη των κοινωνικών δικτύων μπορεί να ενισχύσει τη χρήση των μέσων κοινωνικής δικτύωσης, να βελτιώσει την εμπορική νοημοσύνη και να συμβάλει στην παροχή βελτιωμένων υπηρεσιών (S. Pippal et al., 2014). Τα στελέχη επιχειρήσεων αναζητούν τρόπους για την αξιοποίηση της εξαγόμενης πληροφορίας και την εφαρμογή αυτής στις πωλήσεις, το μάρκετινγκ και τη διαφήμιση.

Ωστόσο, η απόκτηση χρήσιμης πληροφορίας από δεδομένα κοινωνικών δικτύων αποτελεί μια εξαιρετικά δύσκολη διαδικασία χωρίς την εφαρμογή τεχνικών εξόρυξης λόγω τριών βασικών προκλήσεων. Πρώτον ο μεγάλος όγκος που τα χαρακτηρίζει. Η απουσία αυτοματοποιημένων τεχνικών για την ανάλυση των πληροφοριών που προέρχονται από τα κοινωνικά δίκτυα, θα καθιστούσε αδύνατη την επεξεργασία των δεδομένων μέσα σε ένα εύλογο χρονικό διάστημα. Δεύτερον, τα δεδομένα κοινωνικών δικτύων είναι εγγενώς θορυβώδη καθώς διαθέτουν σημαντικά προβλήματα τόσο σε λεκτικό όσο και συντακτικό επίπεδο γεγονός που μπορεί να επηρεάσει αρνητικά τα αποτελέσματα της ανάλυσης. Τρίτον, ο δυναμικός τους χαρακτήρας. Οι συχνές αλλαγές και ενημερώσεις σε σύντομο χρονικό διάστημα αποτελούν μια σημαντική διάσταση που θα πρέπει να λαμβάνεται υπόψη κατά την επεξεργασία των δεδομένων που προέρχονται από τα μέσα κοινωνικής δικτύωσης (G. Barbier & H. Liu, 2011).

Μια άλλη πτυχή των δεδομένων που παράγονται στα μέσα κοινωνικής δικτύωσης, αποτελεί το γεγονός ότι δεν διαθέτουν προκαθορισμένη μορφή καθώς και ο σχεσιακός τους χαρακτήρας που μπορεί να περιπλέξει την ανάλυση. Τα χαρακτηριστικά αυτά, δεν αποτελούν πρόβλημα κατά την εξόρυξη γνώσης καθώς υπάρχουν διαθέσιμες αρκετές τεχνικές που έχουν σχεδιαστεί ειδικά για τον εντοπισμό μοτίβων και κανόνων βάσει των σχεσιακών χαρακτηριστικών (J. Han, 2006). Επιπλέον, οι εφαρμογές εξόρυξης γνώσης αξιοποιούν τεχνικές οι οποίες αναπτύχθηκαν για την αντιμετώπιση προβλημάτων σε άλλους τομείς και καθιστούν εφικτή την εφαρμογή αυτών και σε δεδομένα κοινωνικών δικτύων.

Συνεπώς, η εξόρυξη γνώσης από μέσα κοινωνικής δικτύωσης, μπορεί να βοηθήσει επιχειρήσεις και οργανισμούς να ξεπεράσουν προκλήσεις που σχετίζονται με δεδομένα κοινωνικών δικτύων. Επιπλέον, επιτρέπει στις επιχειρήσεις την εφαρμογή εξειδικευμένου

και προσωποποιημένου μάρκετινγκ, συμβάλει στην μελέτη της συμπεριφοράς των χρηστών, παρέχει ενδείξεις σχετικά με την κοινωνική δομή, εξατομικεύει τις υπηρεσίες ιστού για τους καταναλωτές και καθιστά εφικτή την ανίχνευση και πρόσληψη ανεπιθύμητων γεγονότων (I. King et al., 2009).

2.4 Ανάλυση συναισθήματος και εφαρμογές

Τα κοινωνικά δίκτυα έχουν αποτελέσει το μέσο για τη δημιουργία εμπορικών εφαρμογών που ενσωματώνουν τεχνικές εξόρυξης με σκοπό την καλύτερη κατανόηση των δεδομένων για ερευνητικούς και επιχειρησιακούς σκοπούς (G. Barbier & H. Liu, 2011). Η ανίχνευση κοινοτήτων ή ομάδων, η ανάλυση συναισθήματος και η εξόρυξη γνώσης, η δημιουργία συστάσεων, η μοντελοποίηση της επιρροής και η διάχυση της πληροφορίας αποτελούν ορισμένα αντιπροσωπευτικά παραδείγματα εφαρμογών εξόρυξης δεδομένων που σχετίζονται με τα μέσα κοινωνικής δικτύωσης (P. Gundecha & H. Liu, 2012). Δεδομένου του ερευνητικού ερωτήματος, η παρούσα εργασία θα επικεντρωθεί στην ανάλυσης συναισθήματος.

2.4.1 Τεχνικές ανάλυσης συναισθήματος

Η ανάλυση συναισθήματος η οποία συχνά αναφέρεται και ως εξόρυξη γνώμης, μπορεί να οριστεί ως η αυτόματη ανακάλυψη θετικών ή αρνητικών εκφράσεων σε περιεχόμενο που δημιουργείται από τους χρήστες των κοινωνικών δικτύων (M. Adedoyin-Olowe et al., 2011) και καθιστά εφικτή την περιγραφή του τρόπου διάδοσης των συναισθημάτων και των επιπτώσεών τους. Για την αυτοματοποιημένη εξαγωγή ή ταξινόμηση των συναισθημάτων, η ανάλυση συναισθήματος, χρησιμοποιεί την επεξεργασία φυσικής γλώσσας (Natural Language Processing – NLP), την ανάλυση κειμένου και διάφορες υπολογιστικές τεχνικές (A. Basant et al. 2015).

Οι απόψεις που καθημερινά εκφράζουν οι χρήστες των κοινωνικών δικτύων μπορούν να χρησιμοποιηθούν ως βάση για τη λήψη αποφάσεων. Η ανάλυση συναισθήματος, μπορεί να συμβάλει στην ανάπτυξη στρατηγικών μάρκετινγκ και την αποτελεσματική διαχείριση του εμπορικού σήματος καθώς επιτρέπει σε επιχειρήσεις και οργανισμούς να κατανοούν τις απόψεις των πελατών σχετικά με τα προϊόντα και τις παρεχόμενες υπηρεσίες, να αντιλαμβάνονται τον τρόπο με τον οποίο το κοινό τους αντιδρά στις διαφημιστικές εκστρατείες και σε νέα προϊόντα καθώς και τον λόγο για τον οποίο δεν αγοράζουν ένα προϊόν (D. Kurka et al., 2016; A. D'Andrea et al., 2015).

Κύριος στόχος κατά την ανάλυση συναισθήματος αποτελεί η κατανόηση των κριτικών/απόψεων που εκφράζουν οι χρήστες του δικτύου και η εξέταση της βαθμολογίας των συναισθημάτων (D. Hussein, 2015) με την εφαρμογή διάφορων προσεγγίσεων και εργαλείων ταξινόμησης. Η ταξινόμηση συναισθήματος, μπορεί να οριστεί ως η διαδικασία κατηγοριοποίησης μιας συγκεκριμένης μονάδας ενός εγγράφου σε θετική – ευνοϊκή ή αρνητική – δυσμενής κλάση (A. D'Andrea et al., 2015). Η κλάση αυτή, συχνά αναφέρεται και ως πολικότητα.

Σύμφωνα με τους W. Medhat et al. (2014) για την ανάλυση συναισθήματος υπάρχουν τρία κύρια επίπεδα ταξινόμησης: το επίπεδο εγγράφου, το επίπεδο φράσης και το επίπεδο πτυχής. Το επίπεδο εγγράφου (A. Yesenalina et al. 2010), ταξινομεί ένα έγγραφο βάσει της θετικής ή αρνητικής γνώμης που εκφράζεται σε αυτό και θεωρεί το σύνολο του εγγράφου ως τη βασική μονάδα πληροφόρησης που πραγματεύεται ένα συγκεκριμένο θέμα. Το επίπεδο φράσης (N. Farra et al., 2010), ταξινομεί το συναίσθημα που εκφράζεται σε κάθε πρόταση ενός εγγράφου. Στην περίπτωση αυτή, εάν η φράση είναι υποκειμενική, η ταξινόμηση γίνεται με βάσει τη θετική ή αρνητική άποψη που εκφράζει. Το επίπεδο πτυχής (H. Zhou & F. Song, 2015) ταξινομεί το συναίσθημα λαμβάνοντας υπόψη συγκεκριμένες πτυχές των εγγράφων και βασίζεται στο γεγονός ότι οι χρήστες εκφράζουν διαφορετικές απόψεις για διαφορετικές πτυχές ενός εγγράφου.

Οι Jebaseeli & Kirubakaran (2012) σε έργο τους παρουσίασαν διάφορες τεχνικές ανάλυσης συναισθήματος όπως η μηχανική μάθηση, ο σημασιολογικός προσανατολισμός και η ολιστική προσέγγιση με λέξεις που είναι διαθέσιμες στη διεθνή βιβλιογραφία και σχετίζονται με τις κριτικές προϊόντων. Η έρευνα υποστηρίζει ότι η ανάλυση συναισθήματος και η εξόρυξη γνώμης είναι ζωτικής σημασίας για την υποστήριξη της διαδικασίας λήψης αποφάσεων που σχετίζονται με τα προϊόντα και τις υπηρεσίες των επιχειρήσεων.

Μια άλλη έρευνα σχετικά με τις προσεγγίσεις που χρησιμοποιούνται κατά την ανάλυση συναισθήματος παρέχεται από τους Kaur & Gupta (2013). Οι συγγραφείς στο έργο τους παρουσιάζουν τρείς προσεγγίσεις για την εξαγωγή συναισθήματος: την υποκειμενική προσέγγιση λεξικού η οποία αποτελεί έναν κατάλογο λέξεων σε κάθε μια από τις οποίες έχει αποδοθεί μια βαθμολογία που υποδηλώνει τη φύση της ως θετική, αρνητική ή ουδέτερη, την προσέγγιση με βάση το μοντέλο n-gram η οποία χρησιμοποιεί uni-gram, bi-gram, tri-gram ή συνδυασμό αυτών για την ταξινόμηση των συναισθημάτων και την προσέγγιση μηχανικής μάθησης η οποία εκτελεί την ημι-επιβλεπόμενη ή και επιβλεπόμενη

μάθηση για την εξαγωγή χαρακτηριστικών από το κείμενο και την εκπαίδευση του μοντέλο.

Σε αντίστοιχες έρευνες (Maynard & Funk, 2011; P. Kumar et al. 2016) οι διάφορες προσεγγίσεις ανάλυσης συναισθήματος κατατάσσονται σε τρείς επιμέρους κατηγορίες: την προσέγγιση μηχανικής μάθησης, την προσέγγιση βάσει λεξικών και την υβριδική προσέγγιση.

Η προσέγγιση μηχανικής μάθησης, χρησιμοποιείται για την πρόβλεψη της πολικότητας των συναισθημάτων βασιζόμενη τόσο στα εκπαιδευμένα όσο και στα δεδομένα δοκιμής. Για την ταξινόμηση των συναισθημάτων, η τεχνική αυτή, εφαρμόζει αλγόριθμους μηχανικής μάθησης και χρησιμοποιεί γλωσσικά χαρακτηριστικά. Κύριο πλεονέκτημα της μεθόδου αποτελεί η ικανότητα προσαρμογής και η δημιουργία εκπαιδευμένων μοντέλων για συγκεκριμένους σκοπούς και πλαίσια. Ένα βασικό μειονέκτημα της μεθόδου αποτελεί η μειωμένη δυνατότητα εφαρμογής της σε νέα δεδομένα καθώς η ύπαρξη ετικετών, που σε ορισμένες περιπτώσεις μπορεί να είναι δαπανηρή ή ακόμα και απαγορευτική, είναι απαραίτητη. Η προσέγγιση είναι σε θέση να χρησιμοποιεί μεθόδους επιβλεπόμενης και μη επιβλεπόμενης μηχανικής μάθησης. Οι πιο διαδεδομένες και ευρέως χρησιμοποιούμενες προσεγγίσεις μηχανικής μάθησης είναι τα δίκτυα Bayesian, η ταξινόμηση Naive Bayes, η μέθοδος μέγιστης εντροπίας, τα νευρωνικά δίκτυα και η μηχανή διανυσματικής υποστήριξης.

Η προσέγγιση που βασίζεται σε λεξικά δεν απαιτεί εκπαίδευση για την ανάλυση συναισθήματος. Η μέθοδος, χρησιμοποιεί μια προκαθορισμένη λίστα λέξεων κάθε μια από τις οποίες συνδέεται με ένα συγκεκριμένο συναίσθημα και βασίζεται στην καταμέτρηση των θετικών και αρνητικών λέξεων. Οι προσεγγίσεις βάσει λεξικών, ποικίλουν ανάλογα με το πλαίσιο για το οποίο έχουν δημιουργηθεί και δεν απαιτούν την ύπαρξη ετικετών στα δεδομένα. Ωστόσο, η δημιουργία και χρήση ενός μόνο λεξικού για διαφορετικά πλαίσια είναι δύσκολη. Ορισμένες από τις πιο συνηθισμένες προσεγγίσεις που βασίζονται σε λεξικά είναι, η μέθοδος βάσει λεξικών η οποία μεταφράζει το σύνολο των λέξεων χωρίς να συσχετίζει την μεταξύ τους έννοια, η νέα προσέγγιση μηχανικής μάθησης η οποία ενσωματώνει σημαντικά γλωσσικά χαρακτηριστικά στην αυτόματη μάθηση, η προσέγγιση βάσει σώματος κατά την οποία ένας παράγοντας συναισθήματος που εξαρτάται από την συχνότητα εμφάνισης της κάθε λέξης ανατίθεται σε αυτή και ο συνδυασμός προσεγγίσεων για την ταξινόμηση των συναισθημάτων όπου ο συνδυασμός ειδικά εκπαιδευμένων στοιχισμένων λέξεων αυξάνει την ακρίβεια της ταξινόμησης.

Τέλος, η υβριδική προσέγγιση, αποτελεί ένα συνδυασμό τεχνικών μηχανικής μάθησης και τεχνικών που βασίζονται σε λεξικά. Αρκετές έρευνες έχουν δείξει ότι ο συνδυασμός τόσο της μηχανικής μάθησης όσο και των προσεγγίσεων που βασίζονται σε λεξικά, βελτιώνουν την ταξινόμηση των συναισθημάτων (S.M. Vohra & J.B. Teraiya, 2013).

2.4.2 Αξιολόγηση τεχνικών ανάλυσης συναισθήματος

Σύμφωνα με τους Kharde & Sonawane (2016), η απόδοση των τεχνικών που χρησιμοποιούνται για την ανάλυση συναισθήματος μπορούν να αξιολογηθούν με τη χρήση τεσσάρων δεικτών: Accuracy, Precision, Recall και F-score. Ο πλέον κοινός τρόπος για τον υπολογισμό αυτών των δεικτών βασίζεται στον πίνακα σύγχυσης (Confusion Matrix) (Padmaja & Fatima, 2013) όπως αυτός παρουσιάζεται στη συνέχεια.

Πίνακας 1: Πίνακας Σύγχυσης

	Προβλεπόμενα Θετικά Στοιχεία	Προβλεπόμενα Αρνητικά Στοιχεία
Πραγματικά Θετικά Στοιχεία	Αριθμός Πραγματικά Θετικών Στοιχείων (True Positive – TP)	Αριθμός Ψευδώς Αρνητικών Στοιχείων (False Negative – FN)
Πραγματικά Αρνητικά Στοιχεία	Αριθμός Ψευδώς Θετικών Στοιχείων (False Positive – FP)	Αριθμός Πραγματικά Αρνητικών Στοιχείων (True Negative – TN)

Ο δείκτης Accuracy, είναι ο λόγος του αριθμού των σωστών προβλέψεων προς τον συνολικό αριθμό των στοιχείων της ανάλυσης. Μια τιμή της τάξης του 100% υποδηλώνει ότι τα προβλεπόμενα αποτελέσματα ταυτίζονται πλήρως με τα πραγματικά.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{TP + TN + FP + FN}$$

Ο δείκτης Precision είναι ο λόγος των πραγματικά θετικών προβλεπόμενων στοιχείων προς το σύνολο των θετικά προβλεπόμενων στοιχείων.

$$\text{Precision} = \frac{\text{TP}}{TP + FP}$$

Ο δείκτης Recall είναι ο λόγος των πραγματικά θετικών προβλεπόμενων στοιχείων προς το σύνολο των πραγματικών θετικών στοιχείων.

$$\text{Recall} = \frac{\text{TP}}{TP + FN}$$

Τέλος, ο δείκτης F1-score, είναι ένας αρμονικός μέσος Precision και Recall.

$$\mathrm{F-measure} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precisio} + \mathrm{Recall}}$$

Σε γενικές γραμμές, για την αξιολόγηση ενός συστήματος ανάλυσης συναισθήματος, ουσιαστικά ελέγχεται το κατά πόσο συμφωνούν τα αποτελέσματά του με την ανθρώπινη κρίση. Ωστόσο, το γεγονός ότι οι άνθρωποι τείνουν να συμφωνούν σε ποσοστό περίπου 80% όσον αφορά την κατηγοριοποίηση κειμένου σε θετικό ή αρνητικό έχει ως αποτέλεσμα να είναι αδύνατο για ένα σύστημα να επιτύχει ακρίβεια πάνω από το ποσοστό αυτό (M. Ogneva, 2010). Ως συνέπεια, ένα σύστημα το οποίο καταφέρνει να συμφωνήσει κατά 70% με την ανθρώπινη κρίση, θεωρείται αποτελεσματικό (K. Roebuck, 2012).

2.4.3 Ανασκόπηση εφαρμογών ανάλυσης συναισθήματος

Το πεδίο της ανάλυσης συναισθήματος, παρουσιάζει έντονο ερευνητικό ενδιαφέρον και ως εκ τούτου υπάρχει διαθέσιμος ένας μεγάλος αριθμός ερευνών.

Σε μελέτη τους οι Pournarakis et al. (2017), εισήγαγαν ένα υπολογιστικό μοντέλο που συνδυάζει την ταξινόμηση θεμάτων και συναισθημάτων για τον εντοπισμό θεμάτων με μεγάλη επιρροή βάσει των απόψεων που οι καταναλωτές εκφράζουν στα μέσα κοινωνικής δικτύωσης. Το προτεινόμενο μοντέλο, δημιουργεί ένα νέο γενετικό αλγόριθμο για τη βελτίωση της ομαδοποίησης των tweets σε σημασιολογικά συνεκτικές ομάδες οι οποίες δρουν ως βασική προϋπόθεση για τον εντοπισμό επικρατούντων θεμάτων και συναισθημάτων σε μεγάλες ομάδες δεδομένων. Για την επικύρωση του μοντέλου, οι συγγραφείς εφάρμοσαν τον αλγόριθμο σε σύνολο δεδομένων που συλλέχθηκε από την πλατφόρμα του Twitter σχετικά με το δίκτυο μεταφοράς Uber. Τα αποτελέσματα, παρουσιάζουν τις αντιλήψεις των καταναλωτών και παρέχουν ενδείξεις σχετικά με την αναγνωρισιμότητα και την σημασία του εταιρικού σήματος.

Οι Kang et al. (2017) σε έρευνά τους, μελέτησαν τις απόψεις του κοινού σχετικά με μια νέα πολιτική σχολικών γευμάτων για την πρόληψη της παιδικής παχυσαρκίας φέρνοντας στην επιφάνεια πτυχές που σχετίζονται με αυτές τις απόψεις καθώς και τυχόν διαφορές μεταξύ φύλων και περιφερειών στις ΗΠΑ. Η έρευνα βασίστηκε σε συνολικά 14.317 tweets προερχόμενα από 11.715 χρήστες τα οποία δημοσιεύτηκαν από την ημερομηνία εφαρμογής της πολιτικής στις 9/2/2010 έως τις 31/12/2015. Οι συγγραφείς, εφάρμοσαν τεχνικές εξόρυξης γνώσης για την ταξινόμηση των tweets σε θετικά, αρνητικά και ουδέτερα και διεξήγαγαν ανάλυση περιεχομένου για την εις βάθος κατανόηση όσον αφορά

το στόχο, τον χρήστη, την πηγή και τη λειτουργία. Τα αποτελέσματα της έρευνας αποκάλυψαν τις απόψεις του κοινού σχετικά με την βελτίωση της πολιτικής, συνέβαλαν στην τεκμηρίωση της θετικής επίδρασης που παρουσιάζουν τα οφέλη για την υγεία στην προώθηση της πολιτικής και της κοινοτικής συνεργασίας και αποκάλυψαν ενδιαφέρουσες διαφορές μεταξύ φύλων και απόψεων ανά περιφέρεια.

Η εξαγωγή γνώσης από δεδομένα κοινωνικών δικτύων αποτελεί σύνθετη διαδικασία και μπορεί να επιτευχθεί με διάφορους τρόπους. Σε μελέτη του ο Younis (2016), παρουσιάζει μια προσέγγιση ανοιχτού κώδικα στο πλαίσιο της οποίας συλλέγονται, επεξεργάζονται, αναλύονται και οπτικοποιούνται δεδομένα Twitter. Στόχος της έρευνας είναι η εξόρυξη κειμένου και η ανάλυση συναισθήματος των online κριτικών που σχετίζονται με δυο μεγάλες αλυσίδες λιανικής πώλησης στο Ηνωμένο Βασίλειο κατά τη διάρκεια των Χριστουγέννων του έτους 2014. Η μελέτη καταλήγει στο συμπέρασμα ότι η ανάλυση και απόδοση συναισθήματος στις απόψεις των πελατών, διευκολύνει τις επιχειρήσεις να κατανοήσουν την ανταγωνιστική τους αξία σε μια μεταβαλλόμενη αγορά καθώς και τη γνώμη των πελατών σχετικά με τα προϊόντα και τις υπηρεσίες οι οποίες παράλληλα παρέχουν ενδείξεις για τις μελλοντικές στρατηγικές και πολιτικές λήψεις αποφάσεων.

Σύμφωνα με τους Gürsoy et al. (2017), οι επιχειρήσεις προκειμένου να αυξήσουν το ανταγωνιστικό τους πλεονέκτημα και την αποτελεσματική αξιολόγηση του ανταγωνιστικού τους περιβάλλοντος οφείλουν να παρακολουθούν και να αναλύουν εκτός από το περιεχόμενο των εταιρικών σελίδων και τις πληροφορίες κειμένου στους ιστοτόπους κοινωνικής δικτύωσης των κύριων ανταγωνιστών τους. Σε μια προσπάθεια να βοηθήσουν τις επιχειρήσεις, οι He et al. (2013) σε έργο τους παρουσιάζουν μια εις βάθος ανάλυση αδόμητων δεδομένων κειμένου που προερχόμενα από το Facebook και το Twitter και σχετίζονται με τις τρεις μεγαλύτερες αλυσίδες πίτσας. Η μελέτη επικεντρώνεται στην ανάλυση του ανταγωνισμού και την μετατροπή των διαθέσιμων δεδομένων σε χρήσιμη γνώση για τους υπεύθυνους λήψης αποφάσεων και ηλεκτρονικού εμπορίου. Τα αποτελέσματα, επιβεβαιώνουν τη σημασία της ανάλυσης ανταγωνισμού στα μέσα κοινωνικής δικτύωσης και τη δύναμη της εξόρυξης γνώσης από δεδομένα κειμένου ως εργαλείο για την εκτίμηση της επιχειρησιακής αξίας από μεγάλα σε όγκο δεδομένα κοινωνικών δικτύων.

Σε άρθρο του ο Mostafa (2013) προκειμένου να εξάγει το συναίσθημα των καταναλωτών χρησιμοποίησε ένα τυχαίο δείγμα 3.516 tweets τα οποία σχετίζονται με γνωστά εμπορικά

σήματα. Στην εν λόγω μελέτη, η απόδοση του συναισθήματος έγινε με τη βοήθεια λεξικού με τα αποτελέσματα της ανάλυσης να υποδεικνύουν τη θετική στάση των καταναλωτών ως προς τα εξεταζόμενα εμπορικά σήματα. Χρησιμοποιώντας μια ποιοτική και ποσοτική μεθοδολογία για την ανάλυση των tweets, η μελέτη προσθέτει εύρος και βάθος στη συζήτηση σχετικά με τη στάση των χρηστών ως προς τις δημοφιλείς μάρκες.

Η υπάρχουσα βιβλιογραφία σχετικά με την ανάλυση συναισθήματος στο Twitter, χρησιμοποιεί μια ποικιλία χαρακτηριστικών και μεθόδων πολλές από τις οποίες αποτελούν προσαρμογή των παραδοσιακών τεχνικών ταξινόμησης κειμένου. Οι Ghiassi et al. (2013), σε έρευνά τους παρουσίασαν μια επιβλεπόμενη προσέγγιση μείωσης χαρακτηριστικών με τη χρήση n-grams και στατιστικής ανάλυσης για την ανάπτυξη ενός λεξικού κατάλληλου για την ανάλυση συναισθήματος στο Twitter. Για την ενίσχυση του μειωμένου λεξικού με όρους που σχετίζονται με συγκεκριμένα εμπορικά σήματα, οι συγγραφείς εισήγαγαν όρους από tweets τα οποία και αναφέρονται σε αυτά. Τα αποτελέσματα έδειξαν ότι το μειωμένο σύνολο λέξεων, μειώνει την πολυπλοκότητα του μοντέλου, διατηρεί υψηλό βαθμό κάλυψης και αποδίδει υψηλότερη ακρίβεια στα αποτελέσματα ανάλυσης συναισθήματος.

Η εξόρυξη γνώσης και η ανάλυση συναισθήματος προσφέρουν την δυνατότητα κατανόησης του περιεχομένου που παράγεται από τους χρήστες και του τρόπου με τον οποίο ένα συγκεκριμένο προϊόν ή μάρκα γίνεται αντιληπτό. Οι Aravindan et al. (2014) σε έργο τους, επικεντρώθηκαν στη σημασία των διαδικτυακών εφαρμογών καθώς ενδείκνυνται για ηλεκτρονικές αγορές και παρέχουν τη δυνατότητα δημιουργίας κριτικών προκειμένου να αυξήσουν την παραγωγικότητα και την απόδοση των υπηρεσιών. Οι συγγραφείς, πρότειναν ένα σύστημα που εξάγει αυτόματα τις κριτικές των προϊόντων και ανάλογα με το συναίσθημα που εκφράζουν τις κατατάσσει σε θετικές ή αρνητικές. Ο προτεινόμενος αλγόριθμος περιλαμβάνει δυο στάδια: την εξαγωγή χαρακτηριστικών και την ταξινόμηση βάσει πολικότητας. Κατά το πρώτο στάδιο, οι συγγραφείς χρησιμοποίησαν κανόνες συσχέτισης για την εξαγωγή των ιδιαίτερων χαρακτηριστικών ενός προϊόντος. Στη συνέχεια ανέπτυξαν έναν επιβλεπόμενο ταξινομητή πολικότητας που βασίζεται σε αλγόριθμο μηχανικής μάθησης και καθορίζει το συναίσθημα των κριτικών σε σχέση με τα ιδιαίτερα χαρακτηριστικά του προϊόντος. Η αξιολόγηση του αλγορίθμου έδειξε ότι ο ταξινομητής είναι ιδιαίτερα αποδοτικός και επιτυγχάνει ακρίβεια 79,67%.

3 ΜΕΘΟΔΟΛΟΓΙΑ

3.1 Πλαίσιο έρευνας

Όπως αναφέρθηκε, το ερώτημα που καλείται να απαντήσει η παρούσα εργασία είναι το κατά πόσο είναι εφικτό να ανιχνευτεί συναίσθημα σε δημοσιεύσεις χρηστών σε ένα κοινωνικό δίκτυο με την αξιοποίηση τεχνικών εξόρυξης και ανάλυσης δεδομένων. Το ερώτημα αυτό, πηγάζει από την ανάγκη των επιχειρήσεων να κατανοούν τη γνώμη των καταναλωτών, που αποτελεί σημαντική παράμετρος για την αποτελεσματικότητα και την γενική τους επιτυχία. Ωστόσο, ο μεγάλος όγκος δεδομένων που παράγονται από τους χρήστες των κοινωνικών δικτύων, καθιστά αδύνατη τη χειροκίνητη ανάλυσή τους και συνεπώς αδύνατη τη χειροκίνητη ανίχνευση συναισθήματος. Λύση στο πρόβλημα αυτό, επιχειρούν να δώσουν οι τεχνικές εξόρυξης και ανάλυσης δεδομένων, για τις οποίες ωστόσο είναι ανάγκη να εξεταστεί πόσο αποτελεσματικές είναι.

Για την απάντηση του ερευνητικού ερωτήματος, θα αναπτυχθεί ένα σύστημα το οποίο θα συλλέγει, θα επεξεργάζεται και θα αναλύει δεδομένα από κοινωνικά δίκτυα, προσδίδοντάς τους συναίσθημα. Στη συνέχεια, θα επιλεγεί ένα τυχαίο δείγμα από τα δεδομένα που επεξεργάστηκε το σύστημα στα οποία με αντικειμενικά κριτήρια θα αποδοθεί χειροκίνητα συναίσθημα. Τέλος, θα συγκριθεί το συναίσθημα που αποδόθηκε χειροκίνητα με το συναίσθημα που απέδωσε το σύστημα και έτσι θα εκτιμηθεί η αποτελεσματικότητά του.

3.2 Χαρακτηριστικά πηγής δεδομένων

Δεδομένου του στόχου της μελέτης, πηγή των δεδομένων θα είναι ένα κοινωνικό δίκτυο το οποίο πρέπει να έχει τα ακόλουθα χαρακτηριστικά:

- Να είναι ευρείας χρήσης και υψηλού ρυθμού παραγωγής δεδομένων
- Να λειτουργεί ως μέσο για την έκφραση απόψεων
- Να διαθέτει μηχανισμούς που καθιστούν εφικτή την πρόσβαση και εξόρυξη δεδομένων

3.3 Χαρακτηριστικά συστήματος

Τα βασικά χαρακτηριστικά του συστήματος που θα αναπτυχθεί είναι τα παρακάτω:

- Δυνατότητα σύνδεσης με ένα κοινωνικό δίκτυο για την άντληση δεδομένων.
- Υλοποίηση σε πλατφόρμα η οποία να καθιστά εφικτή τη διαχείριση δεδομένων μεγάλου όγκου.

- Δυνατότητα να χρησιμοποιεί μια συγκεκριμένη λέξη κλειδί ώστε να συλλέγει δεδομένα σχετικά με αυτή για την επίτευξη στοχευμένης αναζήτησης.
- Δυνατότητα απόδοσης συναισθήματος στα προς ανάλυση δεδομένα.

3.4 Μεθοδολογία ανάλυσης συναισθήματος

Εφόσον πηγή δεδομένων θα αποτελέσει ένα κοινωνικό δίκτυο, το σύστημα επιχειρεί την ανίχνευση συναισθήματος σε αναρτήσεις χρηστών.

Συγκεκριμένα, το σύστημα, θα αντλεί δημοσιεύσεις χρηστών που θα περιέχουν μια λέξη κλειδί, η οποία θα επιλεγεί με τέτοιο τρόπο ώστε να είναι διαθέσιμος ένας μεγάλος όγκος δεδομένων.

Δεδομένου ότι, για την ανίχνευση του συναισθήματος, σε κάθε επιμέρους ανάρτηση θα αποδοθεί πολικότητα (θετική, αρνητική ή ουδέτερη), θα πρέπει να χρησιμοποιηθεί ένας αλγόριθμος κατηγοριοποίησης συνδυαστικά με ένα λεξικό αναφοράς. Η παραπάνω διαδικασία θα πρέπει να περιλαμβάνει επιμέρους τεχνικές για τον καθαρισμό και την προετοιμασία των δεδομένων.

Η επιλογή της μεθοδολογίας ανάλυσης συναισθήματος, θα βασιστεί στο θεωρητικό υπόβαθρο που παρουσιάστηκε στο δεύτερο κεφάλαιο της εργασίας.

3.5 Αξιολόγηση αποτελεσμάτων

Για την αξιολόγηση των αποτελεσμάτων και συνεπώς για την απάντηση του ερευνητικού ερωτήματος, θα επιλεγεί ένα τυχαίο δείγμα 100 αναρτήσεων στα οποία με βάση αντικειμενικά κριτήρια θα αποδοθεί χειροκίνητα συναίσθημα - πολικότητα (θετικό, αρνητικό ή ουδέτερο). Στη συνέχεια, θα γίνει σύγκριση με το συναίσθημα που αποδόθηκε σε κάθε ανάρτηση από το σύστημα και θα υπολογιστεί το ποσοστό των σωστών προβλέψεων. Με βάση τη βιβλιογραφική ανασκόπηση, θα θεωρήσουμε ότι το ερευνητικό ερώτημα απαντήθηκε θετικά, εάν το ποσοστό επιτυχίας του συστήματος είναι μεγαλύτερο από 70%.

4 ΑΠΟΤΕΛΕΣΜΑΤΑ

Στο κεφάλαιο αυτό, παρουσιάζονται τα αποτελέσματα της έρευνας. Αρχικά, γίνεται αναφορά στις τεχνολογίες που χρησιμοποιήθηκαν και στον τρόπο επιλογής τους. Στη συνέχεια, παρατίθεται η σχεδίαση και η αρχιτεκτονική του συστήματος και ακολουθεί η υλοποίηση και η εκτέλεσή του. Τέλος, παρουσιάζονται τα αποτελέσματα του συστήματος, πραγματοποιείται η αξιολόγησή τους και απαντάται το ερευνητικό ερώτημα.

4.1 Επιλογή τεχνολογιών

Έπειτα από διαδικτυακή έρευνα και λαμβάνοντας υπόψη τα χαρακτηριστικά που περιεγράφηκαν στη μεθοδολογία, το κοινωνικό δίκτυο που επιλέχθηκε ως πηγή δεδομένων είναι το Twitter. Σύμφωνα με τον Sheela (2016), οι λόγοι που καθιστούν το Twitter μια σημαντική πηγή εξόρυξης και ανάλυσης δεδομένων απορρέουν από την ευρεία χρήση της πλατφόρμας και τον υψηλό ρυθμό παραγωγής δεδομένων. Το Twitter, χρησιμοποιείται από διαφορετικούς ανθρώπους για την έκφραση γνώμης πάνω σε ποικίλα θέματα ενώ παράλληλα περιέχει έναν μεγάλο αριθμό θέσεων οι οποίες αυξάνονται καθημερνά. Επιπλέον, το κοινό του Twitter, απαρτίζεται από χρήστες διάφορων χωρών και ποικίλει από απλούς χρήστες έως προσωπικότητες, εταιρικούς αντιπροσώπους, πολιτικούς ακόμα και προέδρους χωρών. Ως εκ τούτου, καθιστά δυνατή τη συλλογή χρήσιμων δεδομένων κειμένου από διαφορετικές κοινωνικές ομάδες και ομάδες συμφερόντων.

Πέραν της ευρείας αποδοχής και χρήσης, το Twitter παρουσιάζει ορισμένα εγγενή χαρακτηριστικά και μηχανισμούς τα οποία διευκολύνουν την εξόρυξη δεδομένων. Η επαρκώς τεκμηριωμένη διεπαφή προγραμματισμού εφαρμογών (API) του Twitter, είναι ανοιχτή και πλήρως προσβάσιμη, γεγονός που αυξάνει τη διαθεσιμότητα λογισμικού και ηλεκτρονικών εργαλείων για την εξόρυξη δεδομένων. Επιπλέον, το Twitter διευκολύνει την εύρεση και παρακολούθηση συνομιλιών καθώς διαθέτει λειτουργία αναζήτησης που επιτρέπει τον εύκολο εντοπισμό tweets ενώ παράλληλα η χρήση hashtags για τον προσδιορισμό του θέματος ενός tweet , διευκολύνει τον εντοπισμό, την ταξινόμηση και την επέκταση των αναζητήσεων κατά τη συλλογή δεδομένων (Ahmed, 2015).

Όσον αφορά την ανάπτυξη του συστήματος ανάλυσης των δεδομένων, η πλατφόρμα που επιλέχθηκε είναι το Apache Hadoop. Το Hadoop, αποτελεί ένα εργαλείο ανοιχτού κώδικα το οποίο ενδείκνυται για την επεξεργασία δεδομένων Twitter καθώς επιτρέπει την

αποθήκευση και επεξεργασία μεγάλων σε όγκο κατανεμημένων δεδομένων, δεδομένων ροής, δεδομένων με χρονική σφραγίδα, δεδομένων κειμένου και πολλά άλλα.

Ένα σημαντικό πλεονέκτημα του εργαλείου αποτελεί η ικανότητα κλιμάκωσης και η υψηλή του απόδοση. Η κατανεμημένη επεξεργασία δεδομένων σε μεμονωμένους κόμβους του συμπλέγματος επιτρέπει στο Hadoop να αποθηκεύει, να διαχειρίζεται, να επεξεργάζεται και να αναλύει δεδομένα σε όρους petabyte.

Σε αντίθεση με τα παραδοσιακά συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων, η χρήση του Hadoop δεν απαιτεί τη δημιουργία δομημένων σχημάτων πριν την αποθήκευση των δεδομένων. Το γεγονός αυτό, παρέχει μεγάλη ευελιξία ως προς τη διαχείριση των δεδομένων καθώς επιτρέπει την αποθήκευση δεδομένων οποιασδήποτε μορφής, συμπεριλαμβανομένων των ημι-δομημένων και μη δομημένων δεδομένων όπως αυτά του Twitter και στην συνέχεια την ανάλυση και εφαρμογή σχήματος στα ήδη αναγνωσμένα δεδομένα.

4.1.1 Twitter API

Το Twitter API, μπορεί να κατηγοριοποιηθεί σε δυο επιμέρους τύπους με βάση τον σχεδιασμό και τον τρόπο ανάκτησης δεδομένων από την πλατφόρμα του Twitter, το STREAMING API και το REST API. Στην παρούσα εργασία, για τη λήψη δεδομένων σε πραγματικό χρόνο, έγινε χρήση του πρώτου.

Το Streaming API παρέχει μια συνεχή ροή δεδομένων πραγματικού χρόνου και απαιτεί την ανάπτυξη και διατήρηση μιας μακράς σύνδεσης HTTP. Λαμβάνει ενημερώσεις σχετικά με τα πλέον πρόσφατα tweets καθώς και πληροφορίες σχετικά με το συγγραφέα του εκάστοτε tweet, τα οποία ταιριάζουν με το ερώτημα αναζήτησης και παρέχει στους προγραμματιστές πρόσβαση στην παγκόσμια ροή των στοιχείων tweets με πολύ μικρή καθυστέρηση.

Για την αποστολή ασφαλών εξουσιοδοτημένων αιτημάτων προς το Twitter API, το Twitter χρησιμοποιεί το πρωτόκολλο εξουσιοδότησης OAuth. Με την χρήση του πρωτοκόλλου, οι χρήστες δεν χρειάζεται να μοιράζονται τα διαπιστευτήρια των λογαριασμών τους με εφαρμογές τρίτων αυξάνοντας έτσι την ασφάλεια του λογαριασμού τους. Επιπλέον, το OAuth, είναι συμβατό με έναν μεγάλο αριθμό βιβλιοθηκών γεγονός που το καθιστά εύκολο στη χρήση (Twitter Developer). Σε γενικές γραμμές, ο έλεγχος ταυτότητας μπορεί να γίνει με δυο τρόπους: τον έλεγχο ταυτότητας χρήστη (user authentication) και τον έλεγχο ταυτότητας εφαρμογής (application only authentication).

Η αυθεντικοποίηση χρήστη, αποτελεί την πιο κοινή μορφή επαλήθευσης πόρων στο Twitter OAuth. Στην περίπτωση αυτή, ένα πιστοποιημένο αίτημα, προσδιορίζει τόσο την ταυτότητα της εφαρμογής όσο και την ταυτότητα που συνοδεύει τα παραχωρηθέντα δικαιώματα του χρήστη ο οποίος εκτελεί κλήσεις προς το API, εκπροσωπούμενο από το αναγνωριστικό πρόσβασης του χρήστη (user's access token).

Η αυθεντικοποίηση εφαρμογής, είναι μια μορφή ταυτοποίησης που επιτρέπει σε μια εφαρμογή να εκτελεί αιτήματα προς τον API για λογαριασμό της χωρίς το περιβάλλον χρήστη και απευθύνεται σε προγραμματιστές που επιθυμούν πρόσβαση σε δημόσια δεδομένα. Οι κλήσεις API, εξακολουθούν να τελούν υπό περιορισμούς οι οποίοι όμως υπόκεινται σε ολόκληρη την εφαρμογή και όχι ανά χρήστη. Οι μέθοδοι API που υποστηρίζουν αυτή τη μορφή ελέγχου ταυτότητας, υφίστανται δυο όρια ταχύτητας στην τεκμηρίωσή τους, ένα ανά χρήστη (application-user authentication) και ένα ανά εφαρμογή (application-only authentication).

4.1.2 Apache Hadoop

Ο πυρήνας του Apache Hadoop αποτελείται από έναν κατανεμημένο σύστημα αποθήκευσης αρχείων γνωστό ως Hadoop Distributed File System - HDFS και ένα σύστημα επεξεργασίας το οποίο βασίζεται στο μοντέλο προγραμματισμού MapReduce. Το Hadoop, χωρίζει τα αρχεία σε μεγάλα μπλοκ τα οποία κατανέμει στους κόμβους του συμπλέγματος και στη συνέχεια μεταφέρει συσκευασμένο κώδικα για την παράλληλη επεξεργασία αυτών. Αυτή η προσέγγιση, εκμεταλλεύεται την τοπική αποθήκευση των δεδομένων (ibm.org) και επιτρέπει στους κόμβους να χειρίζονται δεδομένα στα οποία έχουν πρόσβαση.

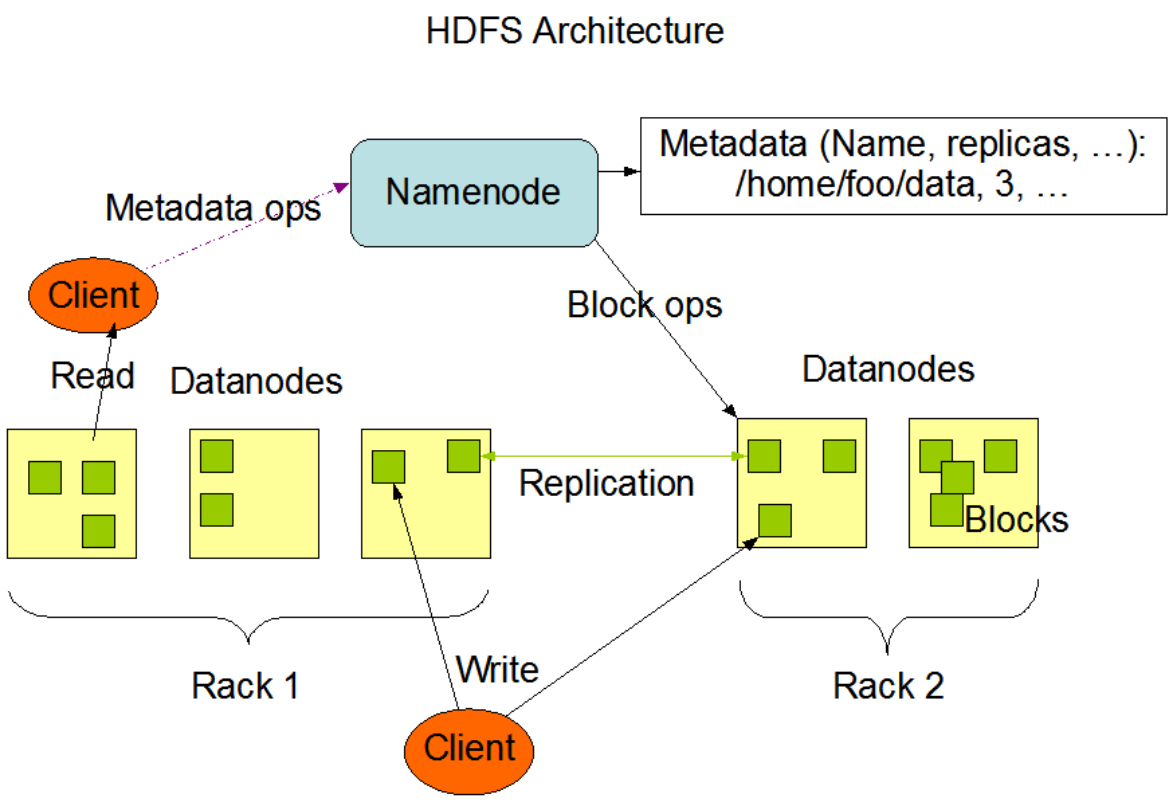
Το πλαίσιο του Apache Hadoop, δομείται από τις ακόλουθες τέσσερεις (4) βασικές ενότητες: το Hadoop Common το οποίο περιέχει το σύνολο των βιβλιοθηκών και των βοηθητικών προγραμμάτων που απαιτούνται για τη λειτουργία των λοιπών ενοτήτων του Hadoop, το κατανεμημένο σύστημα αποθήκευσης αρχείων (Hadoop Distributed File System – HDFS) το οποίο αποθηκεύει τα σύνολα δεδομένων σε βασικό υλικό και παρέχει πολύ υψηλό εύρος ζώνης σε ολόκληρο το σύμπλεγμα, την πλατφόρμα Hadoop YARN είναι υπεύθυνη για τη διαχείριση των πόρων πληροφορικής και την χρήση αυτών για την οργάνωση των εφαρμογών και το Hadoop MapReduce που βασίζεται στο μοντέλο προγραμματισμού MapReduce και χρησιμοποιείται για την επεξεργασία δεδομένων μεγάλης κλίμακας (Apache Hadoop).

4.1.2.1 Hadoop Distributed File System – HDFS

Το Hadoop Distributed File System (HDFS), είναι ένα κατανεμημένο σύστημα αρχείων το οποίο αποτελείται από ένα σύμπλεγμα μηχανών που τρέχουν σε εμπορικό υλικό με στόχο την αποθήκευση πολύ μεγάλων αρχείων και αποτελεί το κύριο σύστημα αποθήκευσης δεδομένων που χρησιμοποιείται από τις εφαρμογές Hadoop. Η αρχιτεκτονική του βασίζεται σε δυο τύπους κόμβων το NameNode και DataNode για την υλοποίηση ενός κατανεμημένου συστήματος αρχείων που παρέχει πρόσβαση υψηλής απόδοσης σε δεδομένα μέσω εξαιρετικά κλιμακούμενων συμπλεγμάτων. Αποτελεί βασικό κομμάτι πολλών τεχνολογιών του οικοσυστήματος Hadoop καθώς παρέχει ένα αξιόπιστο μέσο για τη διαχείριση μεγάλων συνόλων δεδομένων και την υποστήριξη εφαρμογών ανάλυσης ενώ παράλληλα υποστηρίζει την ταχεία μεταφορά δεδομένων μεταξύ των κόμβων του συμπλέγματος (Search Data Management).

Το HDFS, αφού λάβει τα δεδομένα, διασπά τις πληροφορίες σε ξεχωριστά μπλοκ τα οποία διανέμει στους κόμβους του συμπλέγματος επιτρέποντας έτσι την παράλληλη επεξεργασίας υψηλής απόδοσης. Επιπλέον, ο σχεδιασμός του το καθιστά ιδιαίτερα ανεκτικό σε σφάλματα και βλάβες υλικού, δεδομένου ότι το σύστημα αρχείων επαναλαμβάνει κάθε μεμονωμένο μπλοκ σε περισσότερους από έναν κόμβους του συμπλέγματος. Το γεγονός αυτό, επιτρέπει την ανάκτηση των δεδομένων από διάφορους κόμβους και εξασφαλίζει την απρόσκοπτη λειτουργία τους συστήματος (White, 2012).

Η αρχιτεκτονική του HDFS βασίζεται στο μοντέλο Master/Slave. Αποτελείται από έναν master διακομιστή τον NameNode ο οποίος διαχειρίζεται το χώρο ονομάτων του συστήματος αρχείων (namespace) και ρυθμίζει την πρόσβαση των κόμβων σε αυτά και πολλούς DataNodes (slaves), συνήθως έναν ανά κόμβο του συμπλέγματος, οι οποίοι διαχειρίζονται το χώρο αποθήκευσης που αναλογεί στους κόμβους που τρέχουν. Το HDFS διαθέτει ένα χώρο ονομάτων και επιτρέπει την αποθήκευση δεδομένων χρήστη σε μορφή αρχείων. Τα αρχεία, χωρίζονται σε ένα ή περισσότερα τεμάχια και στη συνέχεια αποθηκεύονται σε διάφορους DataNodes τους συστήματος. Ο NameNode, εκτελεί λειτουργίες του χώρου ονομάτων όπως το άνοιγμα, το κλείσιμο και η μετονομασία αρχείων και καταλόγων ενώ παράλληλα καθορίζει τον τρόπο διανομής των τεμαχίων στους DataNodes. Τα DataNodes, είναι υπεύθυνα για την εξυπηρέτηση των αιτημάτων ανάγνωσης και εγγραφής αρχείων από τους πελάτες (clients) του συστήματος. Επιπλέον, εκτελούν λειτουργίες δημιουργίας, διαγραφής και αναπαραγωγής τεμαχίων έπειτα από εντολή του NameNode.



Εικόνα 4: HDFS Architecture

Τα NameNode και DataNode, είναι κομμάτια λογισμικού σχεδιασμένα να λειτουργούν σε μηχανές βασικών προϊόντων και χρησιμοποιούν συνήθως λειτουργικό σύστημα GNU/Linux. Η ανάπτυξη του HDFS, έχει γίνει με τη γλώσσα προγραμματισμού Java και ως εκ τούτου, κάθε μηχανή που υποστηρίζει τη γλώσσα αυτή μπορεί να εκτελεί το λογισμικό NameNode και DataNode. Μια τυπική ανάπτυξη του συστήματος αποτελείται από ένα σύμπλεγμα μηχανών εκ των οποίων η μια εκτελεί αποκλειστικά λειτουργίες NameNode ενώ οι υπόλοιπες εκτελούν ένα μέρος του λογισμικού DataNode. Η αρχιτεκτονική του HDFS, δεν αποκλείει την εκτέλεση πολλαπλών DataNodes στην ίδια μηχανή, αλλά σε μια πραγματική εφαρμογή συναντάται σπάνια. Η ύπαρξη ενός μοναδικού NameNode σε ένα σύμπλεγμα μηχανών, απλοποιεί σε μεγάλο βαθμό την αρχιτεκτονική του συστήματος. Το NameNode διαχειρίζεται το σύνολο των δεδομένων και αποτελεί το αποθετήριο για το σύνολο των μεταδεδομένων του HDFS. Το σύστημα έχει σχεδιαστεί με τέτοιο τρόπο ώστε τα δεδομένα χρήστη να μην ρέουν ποτέ μέσω του NameNode (HDFS Architecture Guide).

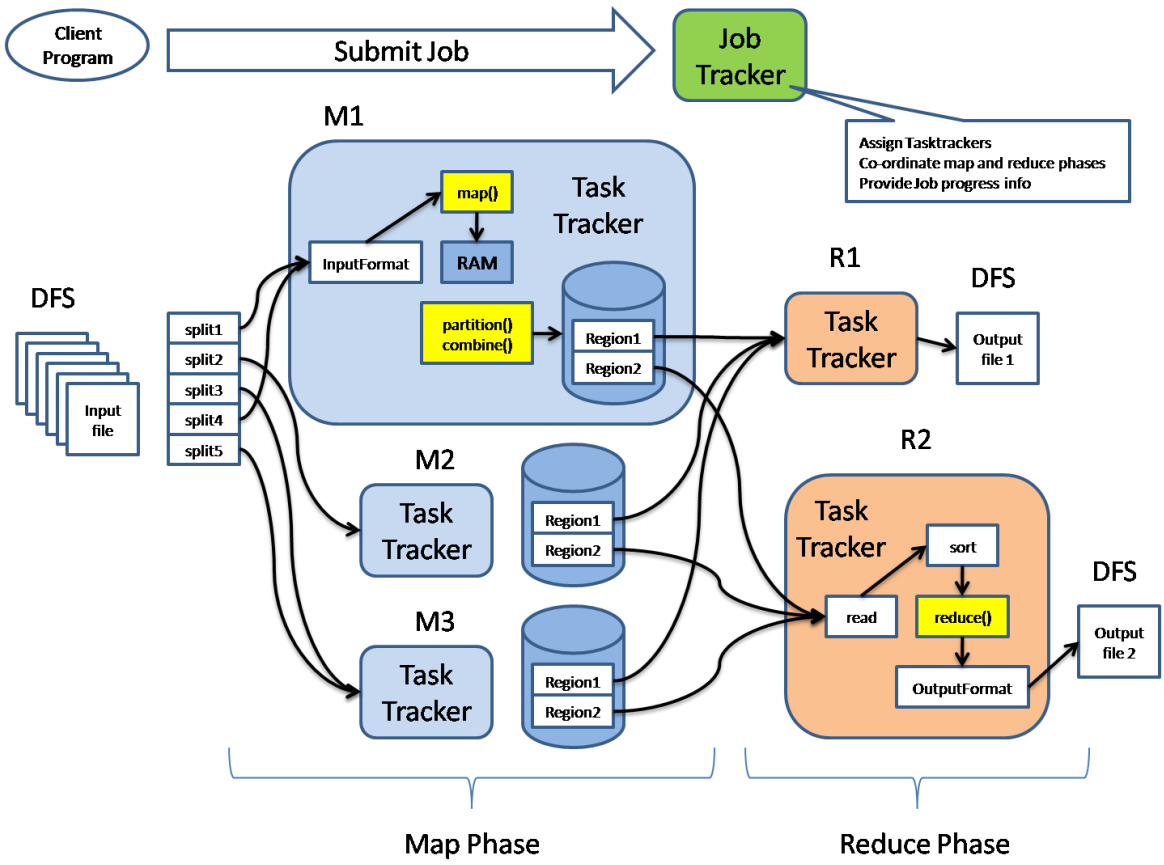
4.1.2.2 Hadoop MapReduce

Το Hadoop MapReduce, είναι ένα πλαίσιο λογισμικού για την εύκολη ανάπτυξη εφαρμογών που επιτρέπει την παράλληλη επεξεργασία μεγάλων ποσοτήτων δεδομένων σε συστάδες βασικού υλικού με αξιόπιστο τρόπο. Αποτελεί ένα προγραμματιστικό μοντέλο για κατανεμημένη υπολογιστική που βασίζεται σε Java.

Για την επεξεργασία των δεδομένων, ο αλγόριθμος MapReduce εφαρμόζει δυο σημαντικές λειτουργίες το Map και το Reduce. Μια εργασία MapReduce συνήθως χωρίζει το σύνολο των δεδομένων σε ανεξάρτητα κομμάτια τα οποία επεξεργάζονται από τη λειτουργία Map με έναν εντελώς παράλληλο τρόπο. Το πλαίσιο, ταξινομεί τα αποτελέσματα του Map τα οποία στη συνέχεια εισάγονται στη λειτουργία Reduce. Τόσο η είσοδος, όσο και η έξοδος του πλαισίου αποθηκεύονται σε ένα σύστημα αρχείων. Το MapReduce, φροντίζει παράλληλα για τις εργασίες προγραμματισμού, παρακολούθησης και της εκ νέου εκτέλεσης τυχόν αποτυχημένων λειτουργιών.

Κύριο πλεονέκτημα του MapReduce, αποτελεί η δυνατότητα κλιμάκωσης της επεξεργασίας δεδομένων σε πολλούς κόμβους υπολογιστών. Στο μοντέλο MapReduce, τα πρωτογενή δεδομένα που υπόκεινται σε επεξεργασία ονομάζονται Mappers και Reducers. Η αποσύνθεση μιας εφαρμογής επεξεργασίας δεδομένων σε Mappers και Reducers σε αρκετές περιπτώσεις απαιτεί σημαντική υπολογιστική δύναμη. Ωστόσο, η χρήση του MapReduce επιτρέπει την κλιμάκωση εφαρμογών ώστε να εκτελούνται σε εκατοντάδες, χιλιάδες ή ακόμα και δεκάδες χιλιάδες μηχανές μέσω της κατάλληλης διαμόρφωσης του συμπλέγματος.

Τυπικά, οι υπολογιστικοί κόμβοι και οι κόμβοι αποθήκευσης του MapReduce, ταυτίζονται σε αυτούς του HDFS. Αυτή η διαμόρφωση, επιτρέπει στο πλαίσιο να προγραμματίζει αποτελεσματικά τις εργασίες στους κόμβους στους οποίους υπάρχουν δεδομένα προσδίδοντας πολύ υψηλό εύρος ζώνης σε ολόκληρο το σύμπλεγμα.



Εικόνα 5: MapReduce Architecture

Το σχεδιαστικό μοντέλο του προγραμματιστικού πλαισίου MapReduce, αποτελείται από έναν Master JodTracker και έναν Slave TaskTracker ανά κόμβο του συμπλέγματος. Όπως προκύπτει και από την παραπάνω εικόνα, ο Master είναι υπεύθυνος για την ανάθεση εργασιών στους Slaves, τον προγραμματισμό κάθε επιμέρους εργασίας τους, την παρακολούθησή τους και την εκ νέου εκτέλεση των αποτυχημένων εργασιών ενώ οι Slaves, εκτελούν τις εργασίες σύμφωνα με τις οδηγίες του Master. (MapReduce Tutorial, Tutorials Point)

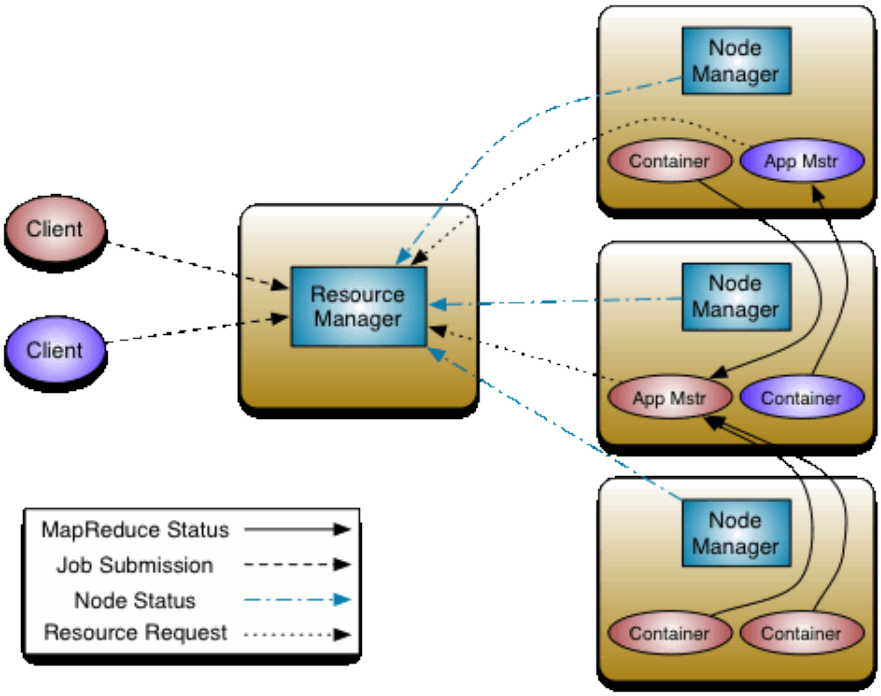
4.1.2.3 Hadoop YARN

Το Hadoop YARN «Yet Another Resource Negotiator», ένα από τα βασικά συστατικά του Apache Hadoop, αποτελεί το επίπεδο διαχείρισης πόρων και προγραμματισμού εργασιών του Hadoop. Είναι υπεύθυνο για την κατανομή των πόρων του συστήματος στις διάφορες εφαρμογές που εκτελούνται σε ένα σύμπλεγμα Hadoop καθώς επίσης και τον προγραμματισμό των εργασιών που εκτελούνται στους διάφορους κόμβους του συμπλέγματος.

Η τεχνολογία εισήχθη στο πλαίσιο του Hadoop το 2012 και επέκτεινε σημαντικά τις πιθανές χρήσεις και τη δύναμη του Hadoop καθώς επιτρέπει την αξιοποίηση των πλεονεκτημάτων του HDFS και του συμπλέγματος κόμβων. Εισήγαγε μια νέα προσέγγιση που αποσυνδέει τους κόμβους διαχείρισης πόρων και τις λειτουργίες προγραμματισμού από τα στοιχεία επεξεργασίας δεδομένων του MapReduce επιτρέποντας στο Hadoop να υποστηρίζει ποικίλους τρόπους επεξεργασίας δεδομένων και ένα ευρύτερο φάσμα εφαρμογών.

Σε μια αρχιτεκτονική συμπλέγματος, το Hadoop YARN βρίσκεται μεταξύ του HDFS και των μηχανών επεξεργασίας που χρησιμοποιούνται για την εκτέλεση των εφαρμογών. Για την παρακολούθηση των διαδικασιών επεξεργασίας στους μεμονωμένους κόμβους του συμπλέγματος, συνδυάζει έναν κεντρικό διαχειριστή (resource manager), συντονιστές εφαρμογών και πράκτορες σε επίπεδο κόμβων. Με τον τρόπο αυτό, είναι σε θέση να διαθέτει δυναμικά πόρους σε εφαρμογές ανάλογα με τις ανάγκες τους, μια δυνατότητα που σχεδιάστηκε για τη βελτίωση της αξιοπιστίας των πόρων και της απόδοσης των εφαρμογών σε σύγκριση με την πιο στατική κατανομή του MapReduce.

Το Hadoop YARN αποκεντρώνει την εκτέλεση και την παρακολούθηση των εργασιών επεξεργασίας, διαχωρίζοντας τις λειτουργίες JodTracker και TaskTracker στις ακόλουθες οντότητες: έναν κοινό διαχειριστή πόρων (global ResourceManager), ένα ApplicationMaster ανά εφαρμογή, έναν κόμβο διαχείρισης (NodeManager slave) ανά κόμβο του συμπλέγματος και ένα κοντέινερ πόρων ανά εφαρμογή (per-application Container) που ελέγχονται από τους NodeManager



Εικόνα 6: YARN Architecture

Το ResourceManager και το NodeManager δημιούργησαν το νέο γενικό σύστημα για τη διαχείριση εφαρμογών με κατανεμημένο τρόπο. Το ResourceManager δέχεται εντολές εργασιών από τους χρήστες, προγραμματίζει τις εργασίες και κατανέμει τους πόρους στο σύνολο των εφαρμογών του συστήματος. Το ApplicationMaster είναι η οντότητα που διαπραγματεύεται πόρους και συνεργάζεται με το NodeManager για την εκτέλεση και παρακολούθηση των εργασιών

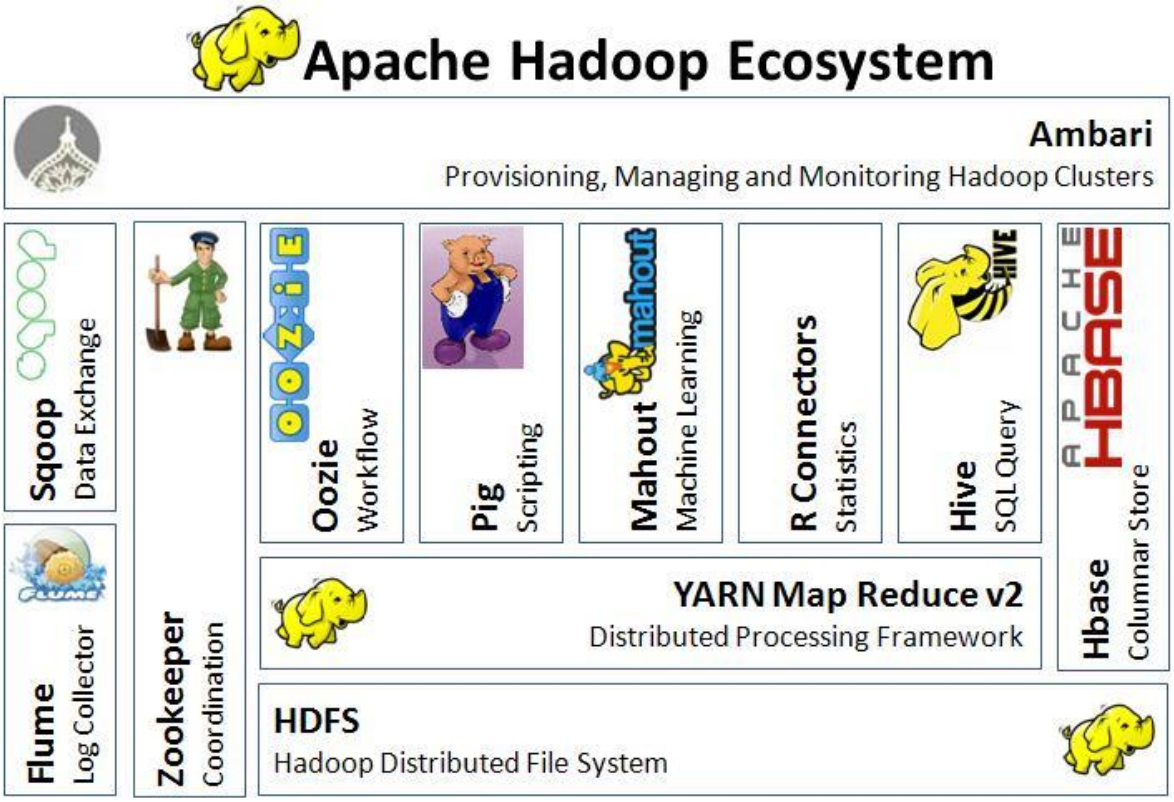
Το ResourceManager διαθέτει έναν προγραμματιστή, ο οποίος είναι υπεύθυνος για την κατανομή πόρων στις διάφορες εφαρμογές που εκτελούνται στο σύμπλεγμα, σύμφωνα με περιορισμούς όπως τα όρια των χρηστών καθώς επίσης και τις απαιτήσεις πόρων κάθε εφαρμογής. Κάθε ApplicationMaster είναι υπεύθυνο για την διαπραγμάτευση πόρων με τον προγραμματιστή του ResourceManager, την παρακολούθηση της κατάστασής τους και της προόδου τους.

Ο NodeManager κάθε κόμβου είναι υπεύθυνος για την εκκίνηση των εφαρμογών, την παρακολούθηση χρήσης των πόρων του συστήματος (CPU, μνήμη, δίσκος, δίκτυο) και την αναφορά τους στον ResourceManager (Hadoop YARN, Hortonworks Apache Hadoop YARN).

4.1.2.4 Οικοσύστημα Hadoop

Το Hadoop, εκτός από ένα πλαίσιο για τη διαχείριση και ανάλυση μεγάλων δεδομένων προσφέρει μια πληθώρα πρόσθετων πακέτων λογισμικού που μπορούν να εγκατασταθούν στην κορυφή του ή παράλληλα με αυτό επιτρέποντας την επεξεργασία διαφόρων τύπων δεδομένων. Ο συνδυασμός αυτός αποτελεί το οικοσύστημα του Hadoop (Hadoop Tutorials).

Ορισμένες από τις τεχνολογίες που απαρτίζουν το οικοσύστημα είναι το Apache Hive, το Apache Pig, το Apache Flume και το Apache Oozie. Η δημιουργία και χρήση του οικοσυστήματος, προϋποθέτει αρχικά την εγκατάσταση του Hadoop και εν συνεχεία την ενσωμάτωση των πρόσθετων εργαλείων παράλληλα με την υλοποίηση των απαιτούμενων εργασιών διαμόρφωσης.



Εικόνα 7: Apache Hadoop Ecosystem

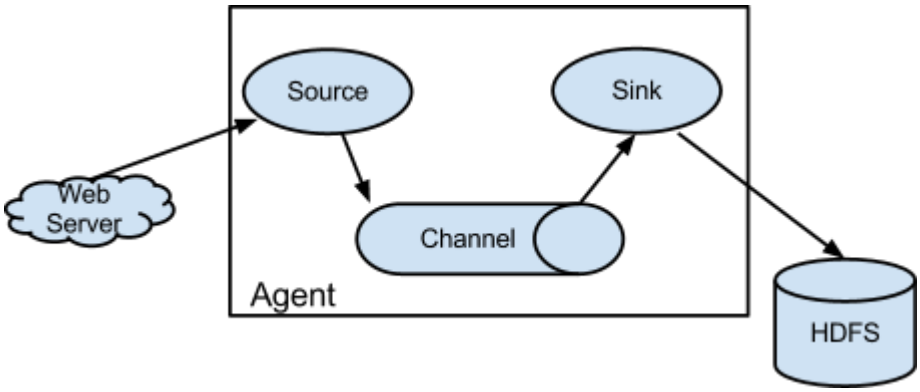
4.1.2.5 Apache Flume

Η Apache Flume είναι μια κατανεμημένη, αξιόπιστη και διαθέσιμη υπηρεσία για την αποτελεσματική συλλογή, συνάθροιση και μεταφορά μεγάλων ποσοτήτων δεδομένων καταγραφής σε ένα αποθετήριο HDFS. Διαθέτει μια απλή και ευέλικτη αρχιτεκτονική που

βασίζεται σε δεδομένα συνεχούς ροής και είναι ανθεκτική σε σφάλματα χάρη στην χρήση αξιόπιστων συγχρονισμένων μηχανισμών ανακατεύθυνσης και ανάκτησης.

Η αρχιτεκτονική υψηλού επιπέδου της Flume βασίζεται σε έναν απλοποιημένο κώδικα ο οποίος είναι επεκτάσιμος και εύκολος στη χρήση. Η υπηρεσία, χαρακτηρίζεται από υψηλή αξιοπιστία καθώς δεν ενέχει τον κίνδυνο απώλειας δεδομένων. Υποστηρίζει επίσης τη δυναμική διαμόρφωση γεγονός που μειώνει το χρόνο εκτέλεσης των εργασιών της.

Τα στοιχεία που συνθέτουν την Apache Flume είναι: τα Events, οι Sources, οι Sinks, τα Channels, οι Agents και ο Client. Το Event – γεγονός, είναι κάθε ξεχωριστή μονάδα δεδομένων που μεταφέρεται από τη Flume. Η Source – πηγή, αποτελεί την οντότητα μέσω της οποίας τα δεδομένα εισέρχονται στη Flume. Οι πηγές, είτε διερευνούν ενεργά είτε περιμένουν παθητικά τα δεδομένα που πρέπει να παραδοθούν σε αυτές. Ο Sink – νεροχύτης αποτελεί το στοιχείο που παραδίδει τα δεδομένα στον τελικό προορισμό. Για τη ροή δεδομένων στον εκάστοτε προορισμό υπάρχουν διαθέσιμοι αρκετοί sinks. Ένα παράδειγμα αποτελεί ο HDFS sink ο οποίος καταγράφει συμβάντα στον αποθηκευτικό χώρο HDFS. Το Channel – κανάλι, είναι ο αγωγός μεταξύ της πηγής και του sink. Οι πηγές, διοχετεύουν τα γεγονότα στο κανάλι τα οποία στη συνέχεια αντλούνται από τους sinks. Ο Agent – πράκτορας αναφέρεται σε κάθε φυσική εικονική μηχανή Java στην οποία τρέχει η Flume. Ουσιαστικά πρόκειται για μια συλλογή από sources, sinks και channels. Τέλος, ο Client – πελάτης είναι η οντότητα που παράγει και μεταδίδει το συμβάν στην πηγή που λειτουργεί εντός του πράκτορα.



Εικόνα 8: Flume Agent

Όπως προκύπτει και από την αρχιτεκτονική της Flume, η ροή δεδομένων ξεκινάει από τον client ο οποίος μεταδίδει τα events σε μια source που λειτουργεί εντός του agent. Η source, αφού λάβει το event το μεταδίδει σε ένα ή περισσότερα channels από όπου και αντλείται από έναν ή περισσότερους sinks οι οποίοι λειτουργούν εντός του ίδιου agent. Τα channels,

για την αποσύνδεση του ρυθμού έκχυσης από τον ρυθμό άντλησης κατά την ανταλλαγή των γεγονότων, χρησιμοποιούν το μοντέλο παραγωγής-κατανάλωσης. Όταν η αυξημένη δραστηριότητα από πλευράς client προκαλεί την ταχύτερη παραγωγή δεδομένων από αυτή που μπορεί να χειριστεί ο προβλεπόμενος προορισμός, το μέγεθος του channel αυξάνεται γεγονός που επιτρέπει στις sources να συνεχίζουν να λειτουργούν κανονικά ακόμα και κατά τη διάρκεια αιχμής. Επιπλέον, ο sink ενός agent μπορεί να συνδεθεί με την πηγή άλλου agent επιτρέποντας τη δημιουργία σύνθετων τοπολογιών ροής δεδομένων. Τέλος, καθώς η κατανεμημένη αρχιτεκτονική του Apache Flume δεν απαιτεί την ύπαρξη κεντρικού σημείου συντονισμού κάθε agent λειτουργεί ανεξάρτητα από τους υπόλοιπους χωρίς κανένα εγγενές σημείο αποτυχίας γεγονός που διευκολύνει την οριζόντια κλιμάκωση (Apache Fume Project Site, Hortonworks Apache Flume).

4.1.2.6 Apache Hive

Η Apache Hive, είναι μια συλλογή λογισμικού που διευκολύνει την ανάγνωση, εγγραφή και διαχείριση μεγάλων συνόλων δεδομένων που βρίσκονται αποθηκευμένα σε κατανεμημένους αποθηκευτικούς χώρους με τη χρήση μιας γλώσσας τύπου SQL που ονομάζεται HiveQL (HQL).

Η ενσωμάτωσή της στην κορυφή ενός συστήματος Hadoop παρέχει εργαλεία που διευκολύνουν την πρόσβαση σε ήδη αποθηκευμένα δεδομένα επιτρέποντας έτσι την εκτέλεση εργασιών όπως η εξαγωγή, ο μετασχηματισμός, η φόρτωση, η αναφορά και η ανάλυση δεδομένων. Επιπλέον, διαθέτει ένα μηχανισμό για την απόδοση σχήματος σε μια ποικιλία διάφορων τύπων δεδομένων. Η Hive, παρέχει επίσης πρόσβαση σε δεδομένα που αποθηκεύονται απευθείας είτε στον αποθηκευτικό χώρο HDFS είτε σε άλλα συστήματα αποθήκευσης δεδομένων όπως το HBase και επιτρέπει την εκτέλεση ερωτημάτων με την χρήση των πλαισίων Apache Hadoop MapReduce, Apache Tez και Apache Spark.

Η Apache Hive, διαθέτει τυπικές λειτουργίες SQL συμπεριλαμβανομένων πολλών λειτουργιών ανάλυσης όπως λειτουργίες OLAP και κοινές εκφράσεις πινάκων. Μπορεί επίσης να επεκταθεί μέσω της ανάπτυξης λειτουργιών (User Defined Functions – UDFs), συνόλων (User Defined Aggregates – UDAFs) και πινάκων (User Defined Tables – UDTFs) που ορίζονται από το χρήστη.

Κατά την εκτέλεση λειτουργιών μεγάλης διάρκειας (Live Long and Process) η Hive χρησιμοποιεί διακομιστές για την εκτέλεση συχνών ερωτημάτων οι οποίοι διαχειρίζονται αποτελεσματικά την προσωρινή αποθήκευση δεδομένων. Με τον τρόπο αυτό, επιταχύνει

τη διαδικασία προσανατολισμού των παρτίδων του Hadoop και επιτυγχάνει την όσο το δυνατόν ταχύτερη ανταπόκριση των ερωτημάτων.

Οι πίνακες της Hive, διαθέτουν παρόμοια χαρακτηριστικά με αυτούς των σχεσιακών βάσεων δεδομένων και οι μονάδες δεδομένων, οργανώνονται με μια φθίνουσα ταξινόμηση. Οι βάσεις δεδομένων απαρτίζονται από πίνακες κάθε ένας από τους οποίους αποτελείται από ξεχωριστά ανεξάρτητα μέρη – partitions. Τα δεδομένα των πινάκων μπορούν να ανακτηθούν και να αντικατασταθούν μέσω απλών ερωτημάτων HQL ενώ παράλληλα εφικτή είναι και η προσθήκη δεδομένων σε αυτούς.

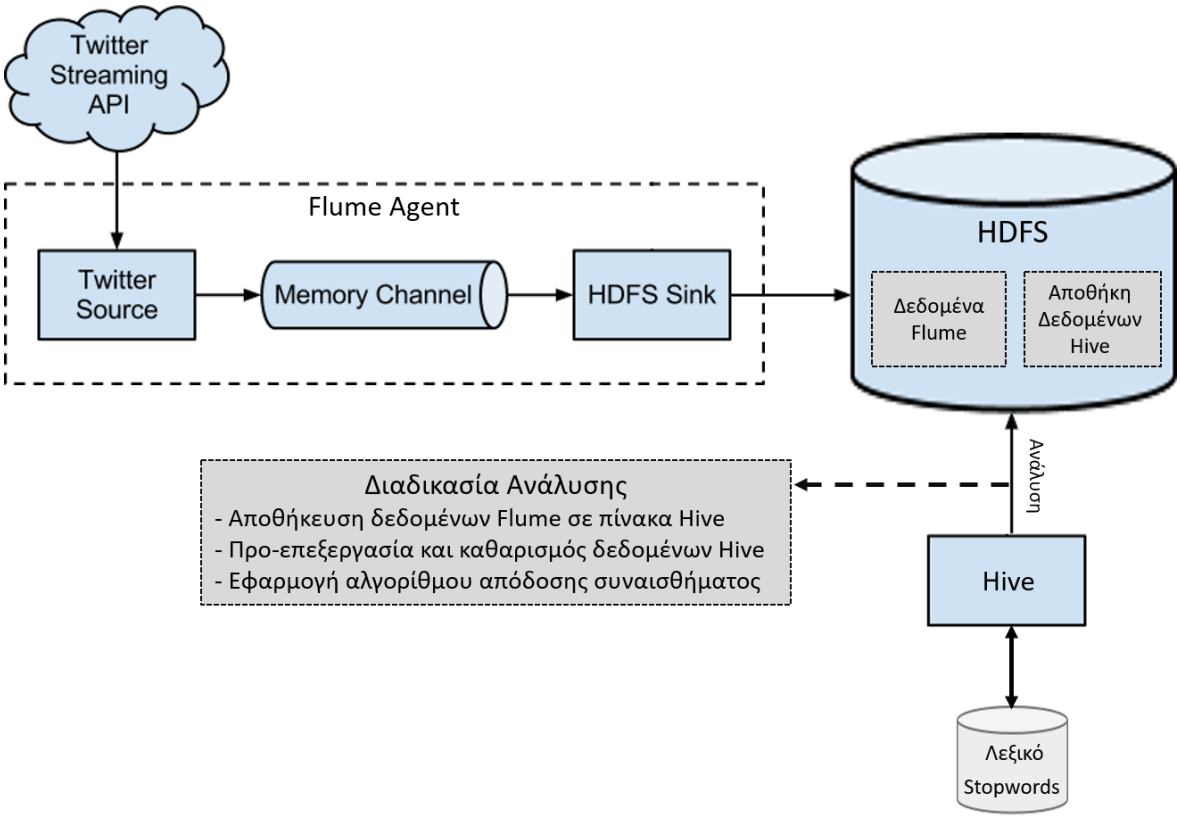
Σε μια συγκεκριμένη βάση δεδομένων, τα δεδομένα πινάκων αποθηκεύονται με τέτοιο τρόπο ώστε να διατηρούν πληροφορίες σχετικά με τη δομή τους (serialized data) και σε κάθε πίνακα αντιστοιχεί ένα ευρετήριο HDFS. Επιπλέον, κάθε πίνακας μπορεί να υποδιαιρεθεί σε partitions που καθορίζουν τον τρόπο κατανομής των δεδομένων σε δευτερεύοντα ευρετήρια ενώ τα δεδομένα ενός partition μπορούν να αναλυθούν περαιτέρω σε κάδους (buckets).

Η Hive υποστηρίζει το σύνολο των βασικών τύπων δεδομένων όπως BIGINT, BINARY, BOOLEAN, CHAR, DECIMAL, DOUBLE, FLOAT, INT, SMALLINT, STRING, TIMESTAMP και TINYINT. Επιπλέον, επιτρέπει τον συνδυασμό βασικών τύπων για τη δημιουργία σύνθετων τύπων δεδομένων, όπως δομές, χάρτες και συστοιχίες (Apache Hive, Apache Hive GitHub Project, Hortonworks Apache Hive).

4.2 Σχεδίαση συστήματος

Το προτεινόμενο σύστημα ανάλυσης επικεντρώνεται στην συλλογή, αποθήκευση και περεταίρω επεξεργασία περιεχομένου πραγματικού χρόνου από την πλατφόρμα του Twitter με σκοπό την απόδοση συναισθήματος σε αναρτήσεις χρηστών που εμπεριέχουν μια συγκεκριμένη λέξη κλειδί. Συγκεκριμένα, η λέξη κλειδί που επιλέχθηκε είναι «tesla», η οποία συνδέεται με την πολυεθνική εταιρεία Tesla Inc., που δραστηριοποιείται στο χώρο των ηλεκτρικών αυτοκινήτων και των μπαταριών παρέχοντας έτσι σημαντικό όγκο δεδομένων για το σύστημα.

Στην εικόνα που ακολουθεί, παρουσιάζονται όλα τα βήματα που περιλαμβάνει το προτεινόμενο σύστημα για την ανάλυση συναισθήματος από τη συλλογή των δεδομένων μέχρι και το τελικό αποτέλεσμα.



Εικόνα 9: Αρχιτεκτονική Συστήματος

Αρχικά, δημιουργείται ένα Twitter Application για να είναι εφικτή η χρήση του Twitter Streaming API και η εξαγωγή tweets πραγματικού χρόνου από το Twitter το οποίο αποτελεί και την κύρια πηγή πληροφόρησης.

Στη συνέχεια, για τη διοχέτευση των δεδομένων στον αποθηκευτικό χώρο της Hadoop το HDFS γίνεται χρήση της Flume. Το Twitter API χρησιμοποιείται κατά την διαμόρφωση της Flume μέσω της οποίας όλα τα tweets μεταφέρονται και αποθηκεύονται σε μη δομημένη μορφή (raw data) από την πλατφόρμα του Twitter στη θέση «/user/flume/twitter_data» του HDFS.

Έπειτα από την αποθήκευση των δεδομένων, ακολουθεί η επεξεργασία αυτών με τη χρήση εντολών Hive. Σε αυτό το στάδιο τα μη δομημένα και σύνθετα δεδομένα φιλτράρονται για την απαλοιφή της περιττής πληροφορίας και μεταφέρονται στον πίνακα twitter_data ο οποίος βρίσκεται στη θέση «/user/hive/warehouse/» του HDFS όπου και πραγματοποιείται η ανάλυση. Στη συνέχεια με τη χρήση ορισμένων UDF (User Defined Functions) εντολών και τη βοήθεια του πίνακα αναφοράς stopwords, τα προς ανάλυση δεδομένα μετασχηματίζονται ώστε να λάβουν την κατάλληλη μορφή και καθαρίζονται για την αφαίρεση του θορύβου. Πιο αναλυτικά, το εν λόγω στάδιο περιλαμβάνει τα εξής βήματα:

1. Διατήρηση των δεδομένων που είναι χρήσιμα για την ανάλυση συναισθήματος. Οι πληροφορίες που απαιτούνται είναι ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet και το περιεχόμενό του.
2. Μετατροπή των tweets σε πεζά γράμματα
3. Απομάκρυνση μη αγγλικών tweets
4. Διαχωρισμό κάθε tweet σε μια συστοιχία λέξεων
5. Διαχωρισμός κάθε λέξης της συστοιχίας λέξεων των tweets σε νέα γραμμή
6. Φιλτράρισμα των επιμέρους λέξεων κάθε tweet για την απομάκρυνση των URLs, ειδικών χαρακτήρων, κενών πεδίων και stopwords.

Για την απόδοση συναισθήματος, εφαρμόζεται ένας αλγόριθμος κατηγοριοποίησης που συνίσταται από τα ακόλουθα δυο (2) βήματα:

1. Απόδοση πολικότητας σε κάθε λέξη του tweet
2. Απόδοση συνολικού συναισθήματος σε κάθε tweet

Το πρώτο βήμα πραγματοποιείται με τη χρήση ενός λεξικού αναφοράς το οποίο μέσω της Hive εισάγεται στον πίνακα dictionary. Αρχικά, για κάθε λέξη του tweet για την οποία υπάρχει ταύτιση με μια λέξη του λεξικού, αποδίδεται η αντίστοιχη πολικότητα (θετική, αρνητική ή ουδέτερη) εκφρασμένη αριθμητικά ως εξής: θετική = 5, ουδέτερη = 2,5 και αρνητική = 0.

Στο δεύτερο βήμα, υπολογίζεται ο μέσος όρος της πολικότητας των λέξεων που συνιστούν το κάθε tweet και προσδιορίζεται το γενικό συναίσθημά του ως θετικό εάν μέση πολικότητα είναι μεγαλύτερη από 2,5, αρνητικό εάν είναι μικρότερη από 2,5 και ουδέτερο σε κάθε άλλη περίπτωση.

Όσον αφορά το λεξικό αναφοράς, θα χρησιμοποιηθεί ένα απλό λεξικό το οποίο συνίσταται από 8.221 λέξεις και ανακτήθηκε από την πλατφόρμα GitHub και εξυπηρετεί τον αλγόριθμο κατηγοριοποίησης καθώς η χρήση ενός πιο πλούσιου λεξικού είναι εμπορική και απαιτεί ειδική άδεια.

Οι ενδιάμεσοι πίνακες που συνθέτουν το μοντέλο ανάλυσης, δημιουργούνται στον HDFS και είναι οι ακόλουθοι: tweet_simple, split_words, explode_word, clean_word1, clean_word2, clean_word3, clean_word4, και word_join.

4.2.1 Δομές δεδομένων

Το σύνολο των πινάκων που απαρτίζουν το σύστημα ανάλυσης, δημιουργήθηκαν στον αποθηκευτικό χώρο HDFS με τη χρήση εντολών Hive έτσι ώστε να είναι αναγνωρίσιμοι από το Hadoop. Πιο αναλυτικά, οι πίνακες που δημιουργήθηκαν είναι οι ακόλουθοι:

- twitter_data ο οποίος περιλαμβάνει πληροφορίες για κάθε ένα από τα tweets που συλλέχθηκαν και έχει τις ακόλουθες στήλες:
    - id BIGINT: ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet
    - created_at STRING: η ημερομηνία και ώρα που δημιουργήθηκε το tweet
    - source STRING: η πηγή από την οποία ο χρήστης δημιούργησε το tweet
    - retweet_count INT: ο αριθμός των retweet του tweet
    - favorite_count INT: ο αριθμός των like του tweet
    - text STRING: το περιεχόμενο του tweet
    - lang STRING: η γλώσσα στην οποία είναι γραμμένο το tweet
    - hashtags ARRAY<STRING>: το σύνολο των hashtag που περιέχονται στο tweet
    - friends_count INT: ο αριθμός των χρηστών που ακολουθεί ο χρήστης
    - favourites_count INT: ο αριθμός των στοιχείων τα οποία ο χρήστης έχει δηλώσει ότι του αρέσουν
    - screen_name STRING: το screen name του χρήστη
    - statuses_count INT: ο αριθμός των δημοσιεύσεων του χρήστη
    - followers_count INT: ο αριθμός ακολούθων του χρήστη

- tweet_simple ο οποίος αποτελεί μια απλούστευση του πίνακα twitter_data και περιλαμβάνει τα σύνολο των δεδομένων που απαιτούνται για την ανάλυση με στήλες:
    - id BIGINT: ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet
    - text STRING: το περιεχόμενο του tweet σε πεζά γράμματα

- split_words ο οποίος περιλαμβάνει μια συστοιχία με το σύνολο των λέξεων κάθε tweet διαιρεμένες με κόμμα, με στήλες:
    - id BIGINT: ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet
    - words ARRAY<STRING>: το περιεχόμενο του tweet διαιρεμένο σε λέξεις

- explode_word ο οποίος περιλαμβάνει κάθε στοιχείο (λέξη) ενός tweet σε μια σειρά, με στήλες:
    - id BIGINT: ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet
    - word STRING: κάθε λέξη ενός tweet ανά γραμμή

- clean_word1: ο οποίος περιλαμβάνει το σύνολο των στοιχείων του πίνακα explode_word πλην των URLs, με στήλες:
    - id BIGINT: ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet
    - word STRING: κάθε λέξη ενός tweet ανά γραμμή

- clean_word2: ο οποίος περιλαμβάνει το σύνολο των στοιχείων του πίνακα clean_word1 από τα οποία έχουν αφαιρεθεί οι ειδικοί χαρακτήρες, με στήλες:
    - id BIGINT: ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet
    - word STRING: κάθε λέξη ενός tweet ανά γραμμή

- clean_word3: ο οποίος περιλαμβάνει το σύνολο των στοιχείων του πίνακα clean_word2 πλην των κενών πεδίων και λέξεων με λιγότερους από τρείς χαρακτήρες, με στήλες:
    - id BIGINT: ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet
    - word STRING: κάθε λέξη ενός tweet ανά γραμμή

- stopwords: ο οποίος περιλαμβάνει ορισμένες από πιο συνηθισμένες λέξεις της αγγλικής γλώσσας που δεν εκφράζουν συναίσθημα, με στήλη:
    - word STRING: μια λέξη που περιλαμβάνεται στη λίστα των stopword λέξεων

- clean_word4: ο οποίος περιλαμβάνει το σύνολο των στοιχείων του πίνακα clean_word3 πλην των στοιχείων για τα οποία υπάρχει ταύτιση με τα στοιχεία του πίνακα stopwords, με στήλες:
    - id BIGINT: ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet
    - word STRING: κάθε λέξη ενός tweet ανά γραμμή

- dictionary ο οποίος περιλαμβάνει την πολικότητα ενός συνόλου αγγλικών λέξεων, με στήλες:
    - word STRING: μια λέξη που περιλαμβάνεται στη λίστα λέξεων του λεξικού

- polarity STRING: η πολικότητα κάθε λέξης εκφρασμένη ως θετική, αρνητική ή ουδέτερη

- word_join ο οποίος περιλαμβάνει πληροφορίες σχετικά με την πολικότητα κάθε λέξης του πίνακα clean_word4, και έχει τις ακόλουθες στήλες:
  - id BIGINT: ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet
  - word STRING: κάθε λέξη ενός tweet ανά γραμμή
  - polarity INT: το συναίσθημα κάθε λέξης του tweet εκφρασμένο αριθμητικά

- tweet_sentiment ο οποίος περιλαμβάνει πληροφορίες με το συνολικό συναίσθημα κάθε tweet και έχει τις ακόλουθες στήλες:
  - id BIGINT: ο μοναδικός αριθμός που αποδίδεται σε κάθε tweet
  - polarity STRING: το συναίσθημα που χαρακτηρίζει το περιεχόμενο ολόκληρου του tweet εκφρασμένο ως θετικό, αρνητικό ή ουδέτερο

4.2.2 Υλικό – Λογισμικό

Το υλικό που χρησιμοποιήθηκε για την ανάπτυξη του συστήματος αποτελείται από έναν φορητό υπολογιστή HP με τα εξής χαρακτηριστικά:

- CPU Intel i3-2330M
- RAM 8GB DDR3
- Operating System Ubuntu 18.04

Σε αυτή τη μηχανή, εγκαταστάθηκαν οι εξής εκδόσεις των εργαλείων που χρησιμοποιήθηκαν:

- Apache Hadoop 2.8.4
- Apache Flume 1.8.0
- Apache Hive 2.3.3

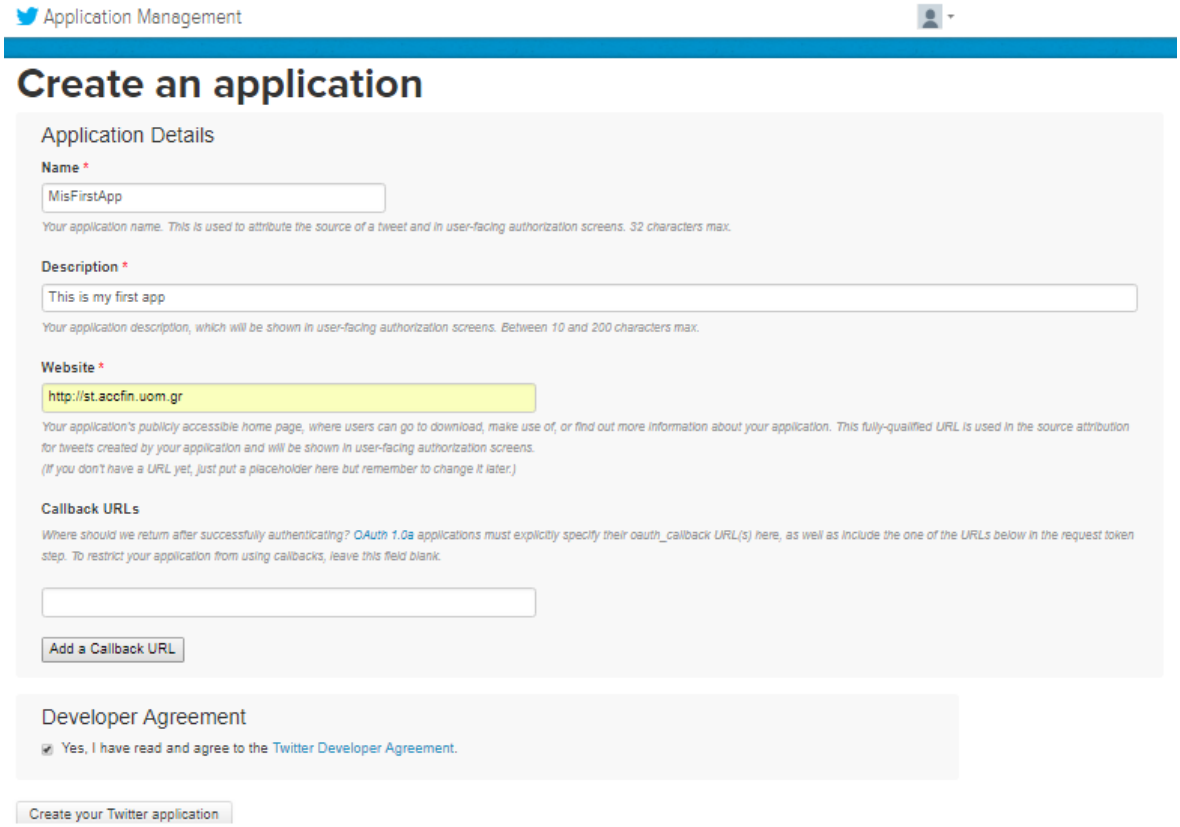
4.3 Υλοποίηση συστήματος

4.3.1 Δημιουργία Twitter App

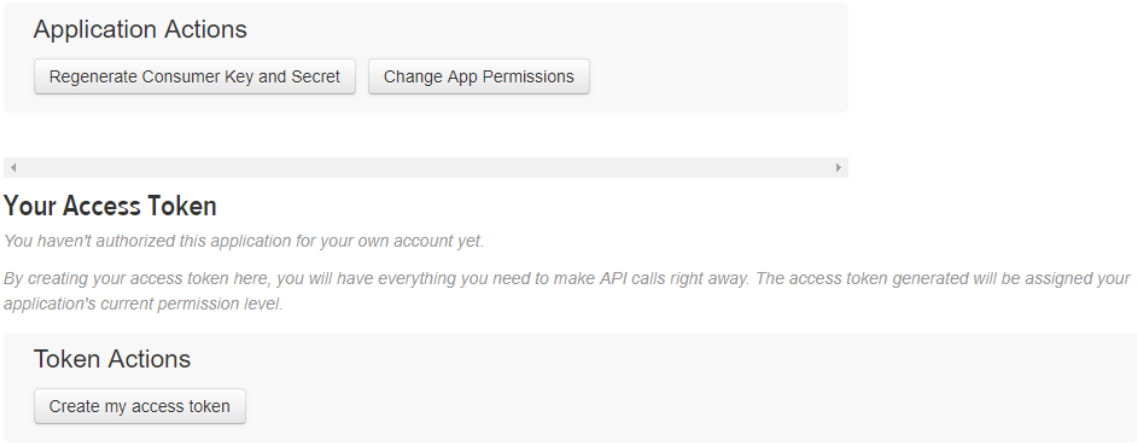
Αρχικά, για να είναι εφικτή η απόκτηση μιας συνεχούς ροής δεδομένων πραγματικού χρόνου από την πλατφόρμα του Twitter, έγινε χρήση του Twitter Streaming API. Για την απόκτηση πρόσβασης στο Streaming API του Twitter δημιουργήθηκε ένα twitter

application το οποίο παρέχει τα απαιτούμενα διαπιστευτήρια (Consumer key, Consumer secret, Access token και Access Token Secret) για τη σύνδεση και την λήψη δεδομένων από την πλατφόρμα του Twitter.

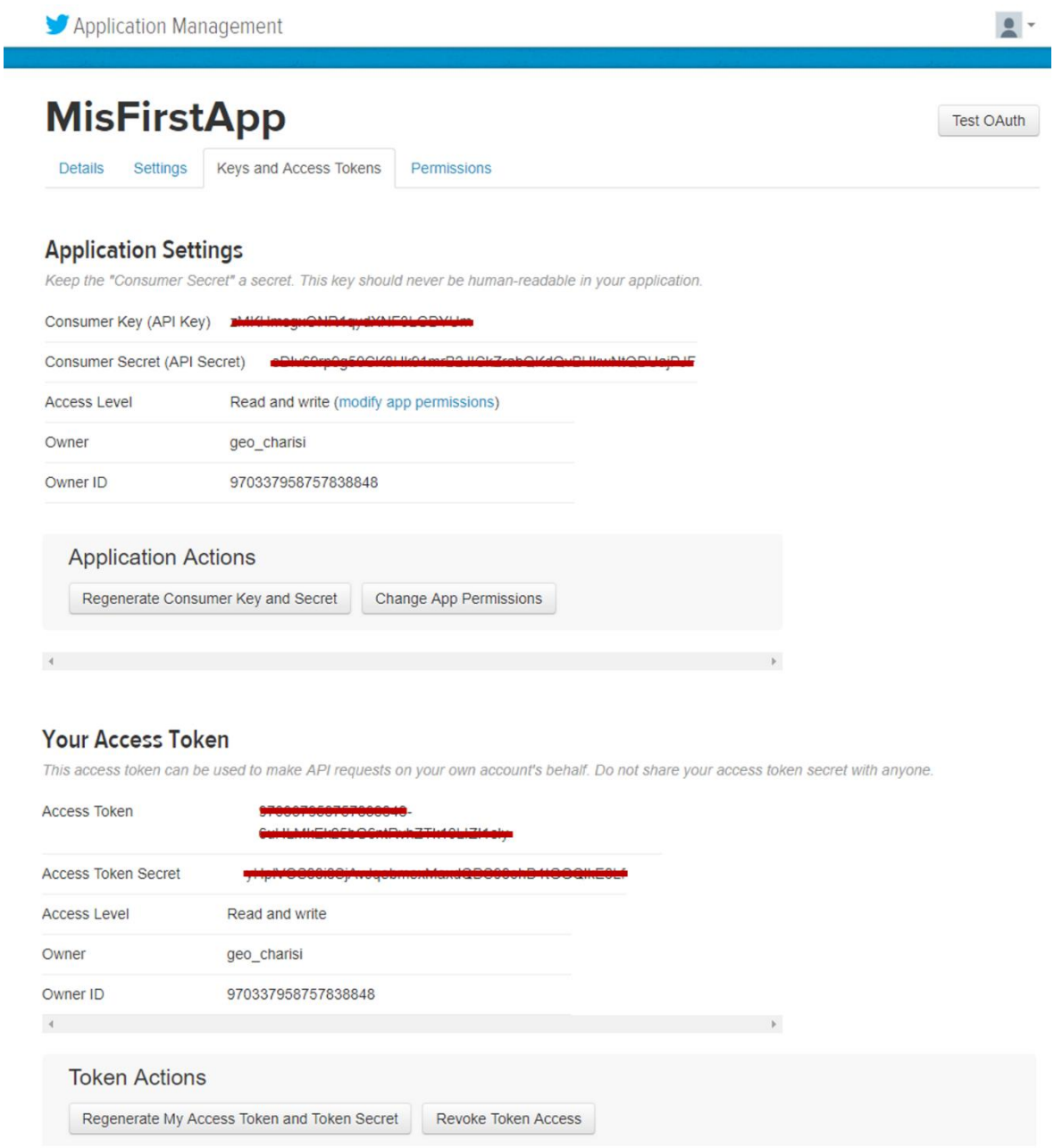
Στις εικόνες που ακολουθούν, παρουσιάζονται τα βήματα που ακολουθήθηκαν.



Εικόνα 10: Δημιουργία Twitter Application



Εικόνα 11: Δημιουργία Access Token



Εικόνα 12: Twitter Application

4.3.2 Διαμόρφωση Flume για εξόρυξη δεδομένων

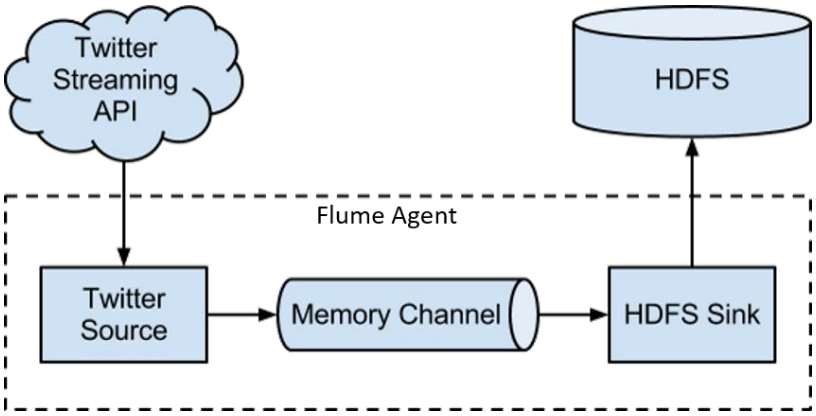
Η συλλογή των δεδομένων από την πλατφόρμα του Twitter και η αποθήκευσή τους στον αποθηκευτικό χώρο HDFS γίνεται μέσω της Apache Flume.

Η χρήση της Flume, προϋποθέτει την ύπαρξη ενός Hadoop cluster η εγκατάσταση του οποίου έγινε σε μια μηχανή Ubuntu (single node set up) λόγω περιορισμένης διαθεσιμότητας υλικού. Στη συνέχεια ακολούθησε η εγκατάσταση και η διαμόρφωση της Flume όπου σε συνδυασμό με τα διαπιστευτήρια που παρήχθησαν κατά τη δημιουργία του

twitter application επιτρέπουν την εξόρυξη δεδομένων από το Twitter και την αποθήκευση αυτών στον αποθηκευτικό χώρο HDFS.

Όπως αναφέρθηκε και σε προηγούμενη ενότητα, η Flume αποτελεί ένα σύστημα λήψης δεδομένων που διαμορφώνεται με τον προσδιορισμό τελικών σημείων σε μια ροή δεδομένων τα οποία ονομάζονται source και sink. Στην Flume, κάθε μεμονωμένο κομμάτι δεδομένων (στη συγκεκριμένη περίπτωση κάθε tweet) ορίζεται ως ένα event (συμβάν). Οι πηγές (sources) παράγουν συμβάντα (μεμονωμένα δεδομένα) τα οποία μέσω ενός καναλιού αποστέλλονται στο sink και στη συνέχεια, το sink καταγράφει τα γεγονότα σε μια προκαθορισμένη τοποθεσία. Κάθε ανεξάρτητη διαδικασία της Apache Flume, όπως αυτή περιεγράφηκε παραπάνω, ορίζεται ως ένας Flume-agent.

Για την υλοποίηση του παρόντος συστήματος, σχεδιάστηκε μια ανεξάρτητη διαδικασία (flume agent) κατά την οποία μια προσαρμοσμένη πηγή αποκτά πρόσβαση στο Streaming API του Twitter και μέσω του καναλιού Memory Channel αποστέλλει τα δεδομένα στο HDFS Sink. Στη συνέχεια, ο HDFS Sink διοχετεύει τα αρχεία στην προκαθορισμένη τοποθεσία /user/flume/twitter_data του αποθηκευτικού χώρου HDFS σε μια πρωτογενή μορφή JSON (raw JSON format). Επιπλέον, κατά τον σχεδιασμό του Flume agent, ορίστηκε η λέξη κλειδί «tesla», για το φιλτράρισμα των δεδομένων και τον εντοπισμό σχετικών με το θέμα της ανάλυσης tweets.



Εικόνα 13: Flume Agent

Η δημιουργία της προκαθορισμένης τοποθεσίας στον HDFS για την αποθήκευση των δεδομένων έγινε με τη χρήση της εντολής -mkdir στο περιβάλλον της Hadoop.

```
hduser@pcgeo: /usr/local/hadoop/bin
File  Edit  View  Search  Terminal  Help
geo@pcgeo:~$ su hduser
Password:
hduser@pcgeo:/home/geo$ cd ~
hduser@pcgeo:~$ cd $HADOOP_HOME/sbin
hduser@pcgeo:/usr/local/hadoop/sbin$ start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-namenode-pcgeo.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-datanode-pcgeo.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hduser-secondarynamenode-pcgeo.out
hduser@pcgeo:/usr/local/hadoop/sbin$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resourcemanager-pcgeo.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduser-nodemanager-pcgeo.out
hduser@pcgeo:/usr/local/hadoop/sbin$ jps
8355 Jps
7300 NameNode
7913 ResourceManager
7740 SecondaryNameNode
7500 DataNode
8079 NodeManager
hduser@pcgeo:/usr/local/hadoop/sbin$ jps
8432 Jps
7300 NameNode
7913 ResourceManager
7740 SecondaryNameNode
7500 DataNode
8079 NodeManager
hduser@pcgeo:/usr/local/hadoop/sbin$ cd $HADOOP_HOME/bin
hduser@pcgeo:/usr/local/hadoop/bin$ hdfs dfs -mkdir hdfs://localhost:54310/user/flume/twitter_data
hduser@pcgeo:/usr/local/hadoop/bin$
```

Εικόνα 14: Δημιουργία path στον HDFS

Στην εικόνα που ακολουθεί, παρουσιάζεται το αρχείο διαμόρφωσης του Flume Agent για τη συλλογή δεδομένων από το Twitter και την αποθήκευση αυτών στον HDFS.

```
hduser@pcgeo: /usr/local/flume/conf
File  Edit  View  Search  Terminal  Help
  GNU nano 2.9.3                    twitter.conf                    Modified

#Naming the components on the current agent.
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey=[redacted]
TwitterAgent.sources.Twitter.consumerSecret=[redacted]
TwitterAgent.sources.Twitter.accessToken=[redacted]
TwitterAgent.sources.Twitter.accessTokenSecret=[redacted]
TwitterAgent.sources.Twitter.keywords = tesla

# Describing/Configuring the sink
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:54310/user/flume/twitter_data/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
```

Εικόνα 15: Αρχείο Διαμόρφωσης Flume Agent

Όπως προκύπτει και από το αρχείο διαμόρφωσης η πηγή που χρησιμοποιήθηκε είναι η Cloudera Twitter Source ώστε να είναι εφικτή η αποθήκευση των δεδομένων σε μορφή

JSON. Στο πλαίσιο αυτό, πριν την εκτέλεση του Flume agent, με τη χρήση του Apache Maven, ένα εργαλείο ανάπτυξης και διαχείρισης έργων λογισμικού του βασίζονται σε Java, δημιουργήθηκε αρχείο jar (flume-sources-1.0-SNAPSHOT.jar) με το σύνολο των αρχείων εκτέλεσης της πηγής το οποίο τοποθετήθηκε στη βιβλιοθήκη της Flume.

4.3.3 Δημιουργία πινάκων για προ-επεξεργασία δεδομένων

Η προ-επεξεργασία, ο καθαρισμός και η ανάλυση των δεδομένων υλοποιούνται με τη χρήση της Apache Hive η οποίο εγκαταστάθηκε στην κορυφή του Hadoop cluster.

Κατά το στάδιο της προ-επεξεργασίας, τα δεδομένα που αποθηκεύονται σε μορφή αρχείου JSON μέσω της Flume στη θέση /user/flume/twitter_data του HDFS φορτώνονται εν συνεχεία στον πίνακα twitter_data ο οποίος δημιουργήθηκε στη θέση /user/hive/warehouse του HDFS όπου και υλοποιείται η επεξεργασία μέσω εντολών Hive.

Για να διασφαλιστεί η ορθή ερμηνεία των δεδομένων του πίνακα twitter_data, χρησιμοποιείται η βιβλιοθήκη Hive HCatalog Core[1] η οποία μέσω της ενσωματωμένης διεπαφής JSON SerDe[2] που διαθέτει, υποστηρίζει την ανάγνωση και εγγραφή αρχείων που προέρχονται σε μορφή JSON. Το ακρωνύμιο SerDe προέρχεται από τις λέξεις Serializer και Deserializer και αποτελεί την διεπαφή που επιτρέπει την μετάφραση δεδομένων οποιασδήποτε μορφής, σε μορφή που η Hive μπορεί να επεξεργαστεί. Στο πλαίσιο αυτό, για να είναι εφικτή η χρήση της διεπαφής το αρχείο jar το οποίο περιλαμβάνει το σύνολο των αρχείων εκτέλεσης της Hive HCatalog Core φορτώθηκε στην Hive.

[1] https://cwiki.apache.org/confluence/display/Hive/HCatalog

[2] https://cwiki.apache.org/confluence/display/Hive/SerDe

```
1   ADD jar /usr/local/hive/lib/hive-hcatalog-core-2.3.3.jar;
2
3   CREATE EXTERNAL TABLE twitter_data (
4       id BIGINT,
5       created_at STRING,
6       source STRING,
7       retweet_count INT,
8       favorite_count INT,
9       text STRING,
10      lang STRING,
11      entities STRUCT<
12          urls:ARRAY<STRUCT<expanded_url:STRING>>,
13          user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
14          hashtags:ARRAY<STRUCT<text:STRING>>>,
15      `user` STRUCT<
16          friends_count:INT,
17          favourites_count:INT,
18          screen_name:STRING,
19          statuses_count:INT,
20          followers_count:INT>
21  )
22  ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe'
23  LOCATION 'hdfs://localhost:54310/user/flume/twitter_data/';
```

Εικόνα 16: Δημιουργία πίνακα twitter_data

4.3.4 Δημιουργία πινάκων για προετοιμασία και καθαρισμό δεδομένων

Για την προετοιμασία και τον καθαρισμό των προς ανάλυση δεδομένων, αρχικά δημιουργείται το view tweet_simple, που εξάγει το μοναδικό αριθμό (id) και το περιεχόμενο (text) εκφρασμένο σε πεζά γράμματα ενώ παράλληλα απομακρύνει όσα tweet δεν είναι γραμμένα στην αγγλική γλώσσα.

```
1   CREATE VIEW tweet_simple AS SELECT
2       id,
3       LOWER(text) AS text
4   FROM twitter_data
5   WHERE lang="en";
```

Εικόνα 17: Δημιουργία view tweet_simple

Στη συνέχεια για την προετοιμασία των προς ανάλυση δεδομένων και τον περεταίρω καθαρισμό αυτών, γίνεται χρήση ορισμένων έτοιμων συναρτήσεων της Hive οι οποίες καθορίζονται από το χρήστη (User Defined Functions UDF's).

Καθώς το προτεινόμενο σύστημα βασίζεται σε έναν αλγόριθμο κατηγοριοποίησης λέξεων για την απόδοση συναισθήματος, κατά την προετοιμασία των δεδομένων το περιεχόμενο κάθε tweet διαχωρίζεται σε επιμέρους λέξεις. Στο πλαίσιο αυτό, αρχικά γίνεται χρήση της UDF συνάρτησης split() η οποία διατίθεται από τη Hive και επιστρέφει ως αποτέλεσμα μια σειρά (array) τιμών. Δημιουργήθηκε έτσι το view split_words στο οποίο αποθηκεύεται ο μοναδικός αριθμός κάθε tweet και η λίστα με το σύνολο των λέξεων που το απαρτίζουν.

```
1    CREATE VIEW split_words AS SELECT
2      id AS id,
3      split(text,' ') AS words
4    FROM tweet_simple;
```

Εικόνα 18: Δημιουργία view split_words

Στη συνέχεια, για την εξαγωγή κάθε λέξης από τη συστοιχία λέξεων και την καταχώρηση αυτής σε νέα γραμμή χρησιμοποιείται η UDF συνάρτηση δημιουργίας πίνακα explode(). Μέσω της συνάρτησης αυτής, δημιουργήθηκε το view explode_word το οποίο περιέχει το σύνολο των λέξεων του κάθε tweet διατηρώντας παράλληλα το id του tweet.

```
1    CREATE VIEW explode_word AS SELECT
2        id,
3        word
4    FROM split_words
5    LATERAL VIEW explode(words) DUMMY AS word;
```

Εικόνα 19: Δημιουργία view explode_word

Τέλος, για την απομάκρυνση του θορύβου, κάθε επιμέρους λέξη ενός tweet ελέγχεται προκειμένου να απομακρυνθούν URLs, ειδικοί χαρακτήρες, κενές εγγραφές και λέξεις οι οποίες δεν φέρουν συναίσθημα όπως άρθρα και σύνδεσμοι με τη βοήθεια του πίνακα αναφοράς stopwords ο οποίος εισήχθη στο σύστημα μέσω εντολών Hive.

```
1    CREATE EXTERNAL TABLE stopwords (
2        word string
3    )
4    ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
5    STORED AS TEXTFILE;
6
7    LOAD DATA INPATH '/user/sample/stopwords/'
8    INTO TABLE stopwords;
```

Εικόνα 20: Δημιουργία πίνακα stopwords

Για τον καθαρισμό των δεδομένων, δημιουργούνται συνολικά τέσσερα (4) view, τα clean_word1, clean_word2, clean_word3 και clean_word4, ο προγραμματιστικός κώδικας των οποίων παρατίθεται στη συνέχεια.

```
1    CREATE VIEW clean_word1 AS SELECT
2        ew.id,
3        ew.word
4    FROM explode_word ew
5    WHERE ew.word NOT LIKE "https%";
```

Εικόνα 21: Δημιουργία view clean_word1

```
1   CREATE VIEW clean_word2 AS SELECT
2       id AS id,
3       REGEXP_REPLACE(word, '[^a-zA-Z0-9\u00E0-\u00FC ]+',
        '') AS word
4   FROM clean_word1;
```

Εικόνα 22: Δημιουργία view clean_word2

```
1   CREATE VIEW clean_word3 AS SELECT
2       cw.id,
3       cw.word
4   FROM clean_word2 cw
5   WHERE cw.word != '' AND length(word) > 3;
```

Εικόνα 23: Δημιουργία view clean_word3

```
1   CREATE VIEW clean_word4 AS SELECT
2       cw.id,
3       cw.word
4   FROM clean_word3 cw LEFT OUTER JOIN stopwords sw
5   ON cw.word = sw.word
6   WHERE sw.word IS NULL;
```

Εικόνα 24: Δημιουργία view clean_word4

4.3.5 Κατασκευή αλγορίθμου για απόδοση συναισθήματος

Η ανάλυση συναισθήματος γίνεται με την χρήση ενός λεξικού αναφοράς[3], το οποίο φορτώθηκε στον πίνακα dictionary μέσω εντολών Hive.

```
1   CREATE EXTERNAL TABLE dictionary (
2       word string,
3       polarity string
4   )
5   ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
6   STORED AS TEXTFILE;
7
8   LOAD DATA INPATH '/user/sample/dictionary/' INTO TABLE
    dictionary;
```

Εικόνα 25: Δημιουργία πίνακα dictionary

[3] https://github.com/AshwanthRamji/Depression-Sentiment-Analysis-with-Twitter-Data

Έπειτα από τη φόρτωση του λεξικού, σε κάθε λέξη του πίνακα clean_word4 για την οποία υπάρχει ταύτιση με μια λέξη του πίνακα dictionary αποδίδεται η πολικότητα κάθε λέξης εκφρασμένη αριθμητικά και τα αποτελέσματα καταχωρούνται στο view word_join διατηρώντας παράλληλα το id κάθε λέξης.

```
1    CREATE VIEW word_join AS SELECT
2        id,
3        cw.word,
4        case d.polarity
5            when  'negative' then 0
6            when 'positive' then 5
7            else 2.5 end as polarity
8    FROM clean_word4 cw
9    LEFT OUTER JOIN dictionary d ON cw.word = d.word;
```

Εικόνα 26: Δημιουργία view word_join

Για την απόδοση συνολικού συναισθήματος σε κάθε tweet, υπολογίστηκε η μέση πολικότητα των λέξεων που φέρουν κοινό αναγνωριστικό (id). Έτσι, κάθε tweet χαρακτηρίζεται ως θετικό αν η μέση πολικότητα είναι μεγαλύτερη 2.5, αρνητικό αν είναι μικρότερη από 2,5 και ουδέτερο σε κάθε άλλη περίπτωση.

```
1    CREATE TABLE tweet_sentiment AS SELECT
2        id,
3        case
4            when avg( polarity ) > 2.5 then 'positive'
5            when avg( polarity ) < 2.5 then 'negative'
6            else 'neutral' end as polarity
7    FROM word_join GROUP BY id;
```

Εικόνα 27: Δημιουργία πίνακα tweet_sentiment

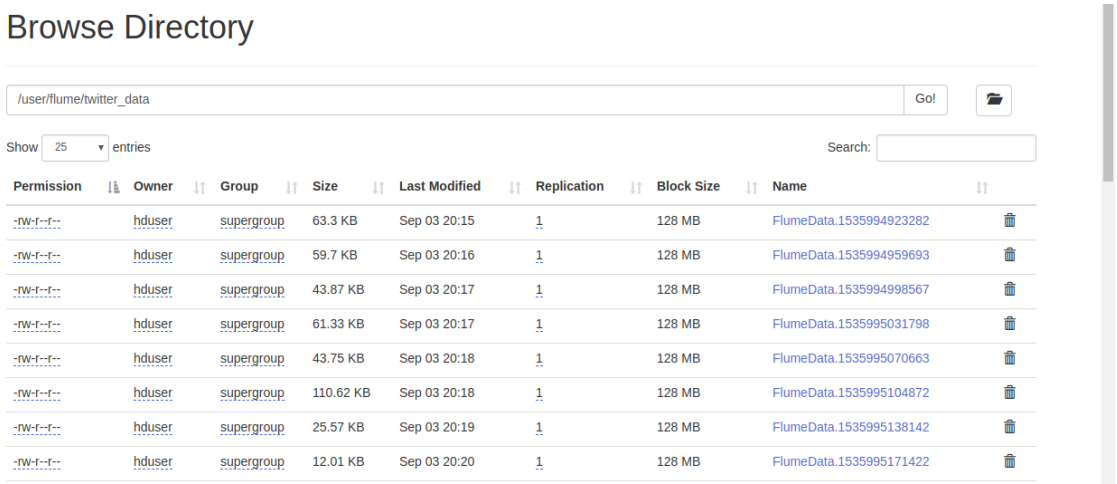
4.4 Εκτέλεση – Αποτελέσματα συστήματος

Αρχικά, για τη λήψη των δεδομένων, ο Flume agent τέθηκε σε λειτουργία μέσω εντολών από το terminal του Hadoop cluster όπως φαίνεται και στην εικόνα που ακολουθεί.

```
hduser@pcgeo: /usr/local/flume
File  Edit  View  Search  Terminal  Help
hduser@pcgeo:~$ cd $FLUME_HOME
hduser@pcgeo:/usr/local/flume$ bin/flume-ng agent --conf ./conf/ -f
conf/twitter.conf -Dflume.root.logger=DEBUG,console -n TwitterAgent
Info: Sourcing environment configuration script /usr/local/flume/con
f/flume-env.sh
Info: Including Hadoop libraries found via (/usr/local/hadoop/bin/ha
doop) for HDFS access
```

Εικόνα 28: Εκτέλεση Flume Agent

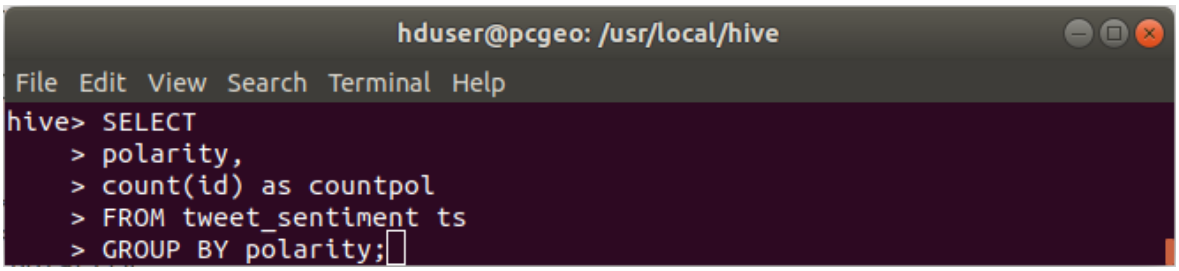
Έπειτα από την εκτέλεσή του, τα δεδομένα αποθηκεύτηκαν σε μορφή αρχείου JSON στην προκαθορισμένη τοποθεσία του HDFS.



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hduser	supergroup	63.3 KB	Sep 03 20:15	1	128 MB	FlumeData.1535994923282
-rw-r--r--	hduser	supergroup	59.7 KB	Sep 03 20:16	1	128 MB	FlumeData.1535994959693
-rw-r--r--	hduser	supergroup	43.87 KB	Sep 03 20:17	1	128 MB	FlumeData.1535994998567
-rw-r--r--	hduser	supergroup	61.33 KB	Sep 03 20:17	1	128 MB	FlumeData.1535995031798
-rw-r--r--	hduser	supergroup	43.75 KB	Sep 03 20:18	1	128 MB	FlumeData.1535995070663
-rw-r--r--	hduser	supergroup	110.62 KB	Sep 03 20:18	1	128 MB	FlumeData.1535995104872
-rw-r--r--	hduser	supergroup	25.57 KB	Sep 03 20:19	1	128 MB	FlumeData.1535995138142
-rw-r--r--	hduser	supergroup	12.01 KB	Sep 03 20:20	1	128 MB	FlumeData.1535995171422

Εικόνα 29: Hadoop Administration Web UI

Στη συνέχεια, οι πίνακες της Hive ενημερώθηκαν αυτόματα αποδίδοντας συναίσθημα στα προς ανάλυση δεδομένα. Τέλος, η εξαγωγή των αποτελεσμάτων πραγματοποιήθηκε μέσω του ακόλουθου ερωτήματος:



Εικόνα 30: Εξαγωγή αποτελεσμάτων

Συνολικά, συλλέχθηκαν πληροφορίες για 4.145 tweets τα οποία περιλαμβάνουν στο περιεχόμενό τους τη λέξη – κλειδί «tesla». Έπειτα από την προ-επεξεργασία και τον καθαρισμό, διατηρήθηκαν προς ανάλυση 3.255 δεδομένα (tweets), ενώ ο αλγόριθμος ανάλυσης συναισθήματος απέδωσε τα εξής αποτελέσματα:

- 929 χαρακτηρίστηκαν ως θετικά (28,54%)
- 1033 χαρακτηρίστηκαν ως αρνητικά (31,74%)
- 1293 χαρακτηρίστηκαν ως ουδέτερα (39,72%)

```
hduser@pcgeo: /usr/local/hive
File  Edit  View  Search  Terminal  Help
Total MapReduce CPU Time Spent: 5 seconds 990 msec
OK
negative        1033
neutral         1293
positive        929
Time taken: 24.145 seconds, Fetched: 3 row(s)
hive>
```

Εικόνα 31: Αποτελέσματα συστήματος

4.5 Αξιολόγηση αποτελεσμάτων συστήματος

Σύμφωνα με τα κριτήρια που τέθηκαν στη μεθοδολογία για την επικύρωση των αποτελεσμάτων, επιλέχθηκε ένα τυχαίο δείγμα 100 tweets στα οποία με βάση αντικειμενικά κριτήρια αποδόθηκε συναίσθημα. Στη συνέχεια, έγινε σύγκριση με το συναίσθημα που αποδόθηκε σε κάθε tweet από το σύστημα υπολογίστηκε ο δείκτης ακρίβειας του συστήματος.

$$Ακρίβεια\ Συστήματος = \frac{Σωστά\ προβλεπόμενα\ tweets}{Σύνολο\ tweets} = \frac{71}{100} = 71\%$$

Η τιμή του δείκτη ακρίβειας υποδηλώνει ότι τα προβλεπόμενα αποτελέσματα ταυτίζονται κατά 71% με τα πραγματικά, δηλαδή το σύστημα είναι σε θέση να προβλέψει σωστά το συναίσθημα σε 71% των tweet που ανέλυσε.

4.6 Αποτελέσματα έρευνας

Σύμφωνα με τα αποτελέσματα της παραπάνω αξιολόγησης, το σύστημα που αναπτύχθηκε είναι σε θέση να προβλέπει σωστά το συναίσθημα σε ποσοστό 71% των δεδομένων (tweets) που αναλύει. Με βάση τα κριτήρια που τέθηκαν στη μεθοδολογία για την απάντηση του ερευνητικού ερωτήματος, μπορούμε εκ του ασφαλούς να καταλήξουμε ότι η ανίχνευση συναισθήματος σε δεδομένα από κοινωνικά δίκτυα μέσω τεχνικών εξόρυξης και ανάλυσης είναι εφικτή.

5 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΤΑΣΕΙΣ

5.1 Συμπεράσματα

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, έγινε μια πρώτη προσπάθεια να εξακριβωθεί το κατά πόσο είναι δυνατό να ανιχνευτεί το συναίσθημα σε δεδομένα από κοινωνικά δίκτυα μέσω εξόρυξης και ανάλυσης.

Το σύστημα που αναπτύχθηκε, κατάφερε να προσδώσει σωστό συναίσθημα στο 71% των δεδομένων που ανέλυσε. Το ποσοστό αυτό, είναι ικανοποιητικό δεδομένου ότι ο αλγόριθμος που εφαρμόστηκε ήταν αρκετά απλός και βασίστηκε σε ένα ελεύθερο λεξικό αναφοράς. Ένα εμπορικό λεξικό αναφοράς το οποίο εξειδικεύεται στο θέμα της έρευνας, μπορεί να βελτιώσει περεταίρω τα αποτελέσματα της ανάλυσης.

Παράλληλα, οι δυνατότητες του Hadoop δεν έγιναν αντιληπτές ούτε αξιοποιήθηκαν στο έπακρον. Ιδανικά, η αποτελεσματικότητα που εγγυάται το Hadoop, προϋποθέτει την υλοποίηση του συστήματος σε ένα σύμπλεγμα τριών τουλάχιστον κόμβων.

Λαμβάνοντας υπόψη όλα τα παραπάνω, γίνεται εύκολα αντιληπτό ότι η ανίχνευση συναισθήματος σε δεδομένα κοινωνικών δικτύων μέσω εξόρυξης και ανάλυσης είναι εφικτή και υλοποιήσιμη.

Η έρευνα αυτή υποδεικνύει τις δυνατότητες των τεχνολογιών που υπάρχουν για εξόρυξη και ανίχνευση συναισθήματος σε δεδομένα από κοινωνικά δίκτυα και οφείλει να λειτουργήσει ως κίνητρο για τις επιχειρήσεις να επενδύσουν στις τεχνολογίες αυτές ώστε να καταφέρουν να αξιοποιήσουν τις πραγματικές τους δυνατότητες προς όφελός τους.

5.2 Προτάσεις – Μελλοντικές επεκτάσεις

Η εξόρυξη γνώσης από μέσα κοινωνικής δικτύωσης αποτελεί έναν ευρύ κλάδο έρευνας μέρος των πιο σημαντικών πτυχών της οποίας έχουν καλυφθεί στην παρούσα εργασία. Η ανάλυση συναισθήματος αποτελεί ένα εξελισσόμενο πεδίο του τομέα ανάλυσης δεδομένων με ιδιαίτερο ενδιαφέρον για τον κλάδο των επιχειρήσεων. Το προτεινόμενο σύστημα, εκτός από την ανάλυση του συναισθήματος των χρηστών, με την διατήρηση πρόσθετων πληροφοριών για το κάθε tweets, είναι σε θέση να υπολογίζει αποτελέσματα όπως τους χρήστες με τα περισσότερα tweets και τα πλέον χρησιμοποιούμενα hashtags. Ως εκ τούτου, μπορεί επίσης να χρησιμοποιηθεί για τον υπολογισμό πρόσθετων μέτρων

προκειμένου να προσδιοριστούν οι τρέχουσες τάσεις σε σχέση με συγκεκριμένα θέματα. Η επέκταση αυτή, μπορεί να είναι ιδιαίτερα χρήσιμη στον τομέα των επιχειρήσεων.

Επιπλέον, μελλοντική επέκταση της παρούσας εργασίας θα μπορούσε να αποτελέσει η υλοποίηση του συστήματος με τη χρήση του Apache Oozie για τη δημιουργία μιας ροής εργασιών και τον προσδιορισμό του χρόνου εκτέλεσης της εκάστοτε εργασίας του συστήματος. Επίσης, για την αυτοματοποίηση και αναπαράσταση των αποτελεσμάτων μπορεί να γίνει χρήση του Apache Zeppelin ένα web-based notebook ανοιχτού κώδικα, που επιτρέπει τη διαδραστική ανάλυση και οπτικοποίηση των αποτελεσμάτων.

Τέλος, επεκτάσεις θα μπορούσαν να γίνουν και στο κομμάτι της ανάλυσης. Η χρήση ενός προσαρμοσμένου λεξικού στο εκάστοτε πεδίο έρευνας καθώς και η απομάκρυνση ή διαφορετική διαχείριση μηνυμάτων με σαρκαστική ή ειρωνική διάθεση θα έδιναν καλύτερα αποτελέσματα και θα συνέβαλαν στη βελτίωση της ακρίβειας του συστήματος.

Βιβλιογραφία

1. About Twitter, Online Available: https://about.twitter.com/en_us/company.html (July 23 2018).
2. Adedoyin-Olowe M., Medhat Gaber M. & Stahl F. (2011), "A Survey of Data Mining Techniques for Social Network Analysis".
3. Aggarwal C. (2011), "An Introduction to Social Network Data Analytics", in Social Network Data Analytics, Springer.
4. Ahmed, W., "Using Twitter as a data source: An overview of current social media research tools", LSE Impact blog, July 7 2015, Available at: http://blogs.lse.ac.uk/impactofsocialsciences/2015/07/10/social-media-research-tools-overview/
5. Apache Flume, "Welcome to Apache Flume", Online Available: https://flume.apache.org/ (August 26 2018).
6. Apache Hadoop, "Hadoop YARN", Online available: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html (August 25 2018).
7. Apache Hadoop, "HDFS Architecture Guide", Online available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (August 24 2018).
8. Apache Hadoop, "MapReduce Tutorial", Online available: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html (August 24 2018).
9. Apache Hadoop, "Welcome to Apache™ Hadoop®!", Online available: http://hadoop.apache.org/, (August 23 2018).
10. Apache Hive, Online Available: http://hive.apache.org/
11. Aravindan S., Ekbal A. (2014), "Feature Extraction and Opinion Mining in Online Product Reviews,", IEEE International Conference on Information Technology.
12. Balan S. & Rege J. (2017), "Mining for Social Media: Usage Patterns of Small Businesses"
13. Barbier G. & Liu H. (2011), "Data mining in social media" In C. C. Aggarwal (Ed.) Social network data analytics (pp. 327–352). United States: Springer
14. Basant, A., Namita, M., Pooja, B., Sonal G. (2015), "Sentiment Analysis Using Common-Sense and Context Information", Hindawi Publishing Corporation Computational Intelligence and Neuroscience.

15. BD HD Team, "Hadoop – An Apache Hadoop Tutorials for Beginners", BD HD Team, Published September 23, 2017, Update October 17, 2017, Online available: https://techvidvan.com/tutorials/apache-hadoop-tutorials/, (August 23 2018).
16. Bernard Marr, Big Data: What Is Hadoop - An Easy Explanation For Absolutely Anyone, Bernard Marr, Online available: https://www.bernardmarr.com/default.asp?contentID=1080, (August 23 2018).
17. Cwiki Apache Org, "Hive HCatalog", Cwiki Apache Org, Available at: https://cwiki.apache.org/confluence/display/Hive/HCatalog, (September 3 2018)
18. Cwiki Apache Org, "Hive SerDe", Cwiki Apache Org, Available at: https://cwiki.apache.org/confluence/display/Hive/SerDe, (September 3 2018)
19. D'Andrea A., Ferri F., Grifoni P. & Guzzo T. (2015), "Approaches, Tools and Applications for Sentiment Analysis Implementation", International Journal of Computer Applications (0975 – 8887), Volume 125 – No.3.
20. David L., Passmore, A., Kosorukoff (2011), "Social Network Analysis Theory and Applications".
21. Deepesh Kumbhare, "Understanding Hadoop Ecosystem" Hadoop Tutorials, Online available: http://hadooptutorials.co.in/tutorials/hadoop/understanding-hadoop-ecosystem.html (August 23 2018).
22. Doaa Mohey El-Din Mohamed Hussein (2016), "A survey on sentiment analysis challenges", Journal of King Saud University - Engineering Sciences.
23. Domingos P. & Richardson M. (2001), "Mining the network value of customers", In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 57–66, New York, NY, USA.
24. Dr. Umadeni B. & Surya P. (2017), "A Review on Various Data Mining Techniques in Social Media", Vol. 5, Issue 4.
25. Eirinaki M. (2008), "Web Mining: A Road trip".
26. Farra N., Challita E., Assi R. A. & Hajj H. (2010), "Sentence-level and document-level sentiment mining for arabic texts", In: Proceeding IEEE International Conference on Data Mining Workshops.
27. Fournier-Viger P., Chun-Wei Lin J., Uday Kiran R., Koh Y. S. & Thomas R. (2017), "A Survey of Sequential Pattern Mining", Volume 1, Number 1.
28. Ghiassi, M., Skinner, J. and Zimbra D. (2013), "Twitter Brand Sentiment Analysis: A Hybrid System Using N-Gram Analysis and Dynamic Artificial Neural Network", Expert Systems with Applications, 40, Issue 16, 6266–6282.

29. GitHub, “Apache Hive Project”, Online Available: https://github.com/apache/hive
30. GitHub, “Reference Dictionary” GitHub, June 25 2017, Online Available: https://github.com/AshwanthRamji/Depression-Sentiment-Analysis-with-Twitter-Data, (September 3 2018)
31. Gundecha P. & Liu H. (2012), “Mining social media: A brief introduction”, Tutorials in Operations Research.
32. Gürsoy U. T., Bulu D., Yiğit C. (2017), “Social Media Mining and Sentiment Analysis for Brand Management”, Global Journal of Emerging Trends in e-Business, Marketing and Consumer Psychology, An Online International Research Journal (ISSN: 2311-3170) Vol: 3 Issue: 1
33. Han J. (2006), “Data Mining Concepts and Techniques”, Morgan Kaufmann, San Diego.
34. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar (2014), “A Review Paper on Big Data and Hadoop”, International Journal of Scientific and Research Publications, ISSN 2250-3153, Volume 4, Issue 10.
35. He W., Zha, S. and Li L. (2013), “Social Media Competitive Analysis and Text Mining: A Case Study in The Pizza Industry”, International Journal of Information Management, 33, Issue 3, 464–472.
36. Hortonworks, “Apache Flume”, Online Available: https://hortonworks.com/apache/flume/, (August 26 2018).
37. Hortonworks, “Apache Hadoop YARN”, Online Available: https://hortonworks.com/apache/yarn/, (August 25 2018).
38. Hortonworks, “Apache Hadoop, Online available: https://hortonworks.com/apache/hadoop/, (August 23 2018).
39. Hortonworks, “Apache Hive, Online available: https://hortonworks.com/apache/hive/, (August 26 2018).
40. IBM, "Hadoop Distributed File System (HDFS)?", Online available: https://www.ibm.com/support/knowledgecenter/en/SSPT3X_4.2.0/com.ibm.swg.im.infosphere.biginsights.product.doc/doc/c0057606.html, (August 23 2018).
41. Jaba Sheela L. (2016), “A Review of Sentiment Analysis in Twitter Data Using Hadoop”, International Journal of Database Theory and Application Vol.9, No.1, pp.77-86
42. Jebaseeli, A. N., & Kirubakaran, E. (2012), “A survey on sentiment analysis of (product) reviews”, International Journal of Computer Applications, 47(11).

43. Kang, Y., Wang, Y., Zhang, D. and Zhou, L. (2017), "The Public's Opinions on A New School Meals Policy for Childhood Obesity Prevention in the U.S.: A Social Media Analytics Approach", International Journal of Medical Informatics. Vol. 103, 83-88.
44. Kaplan A. M., Haenlein M. (2010), "Users of the world, unite! The challenges and opportunities of Social Media".
45. Kaur, A., & Gupta, V. (2013), "A survey on sentiment analysis and opinion mining techniques", Journal of Emerging Technologies in Web Intelligence, 5(4), 367-371.
46. Kharde A. V. & Sonawane S. S. (2016), "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11.
47. King I., Li J., and Chan K. T. (2009), "A brief survey of computational approaches in social computing", In IJCNN'09: Proceedings of the 2009 international joint conference on Neural Networks, pages 2699–2706, Piscataway, NJ, USA, IEEE Press.
48. Kumar M. D. (2015), "Tweet Analysis: Twitter Data processing Using Apache Hadoop", International Journal Of Core Engineering & Management (IJCEM) Volume 1, Issue 11.
49. Kumar P. & Jaiswal U. C. (2016), "A Comparative Study on Sentiment Analysis and Opinion Mining", International Journal of Engineering and Technology (IJET), Vol 8.
50. Kurka D. B., Godoyy A., & Von Zubenz F. J. (2016), "Online Social Network Analysis: A Survey of Research Applications in Computer Science".
51. Lauw H., Shafer J. C., Agrawal R., and Ntoulas A. (2010), "Homophily in the digital world: A live journal case study", Internet Computing, IEEE, 14(2):15 –23.
52. Liu B. (2012), "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers.
53. Luna-Nevarez, C. (2016). "Exploring the use of Twitter opinion mining (TOM) in marketing courses." Marketing Management Association Fall 2016 Educators' Conference Proceedings, 110-111.
54. Margaret Rouse, "Hadoop Distributed File System (HDFS)", Search Data Management, Online Available: https://searchdatamanagement.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS, (August 24 2018).
55. Matthew A. Russell (2013), "Mining the Social Web", 2nd Edition.

56. Maynard, D., & Funk, A. (2011), “Automatic detection of political opinions in tweets” In: Proceedings of the 8th international conference on the semantic web, ESWC’11, p. 88-99.
57. Medhat, W., Hassan, A., Korashy, H. (2014), “Sentiment analysis algorithms and applications: A survey”, Ain Shams Eng.
58. Mostafa, M. M. (2013), “More Than Words: Social Networks’ Text Mining for Consumer Brand Sentiments”, Expert Systems with Applications, 40, Issue 10, 4241–4251
59. Mr. Vohra S. M. & Prof. Teraiya J. B. (2013), “A competitive Study of Sentiment Analysis Techniques”, Journal of Information, Knowledge and Research in Computer Engineering, Volume-02, Issue-02.
60. Mrs. Bhanap S., Dr. Kawthekar S. (2015), “Data Mining for Business Intelligence in Social Network: A survey”, International Advanced Research Journal in Science, Engineering and Technology, Vol. 2, Issue 12.
61. Ms. Patil P. S., Ms. Pranali B. S., Ms. Reshma J. F., Mr. Chougule P. A. (2016), “Sentiment Analysis on Twitter Data Using Apache Flume and Hive”, International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 02.
62. Nayak R. (2002), “Data Mining for Web-Enabled Electronic Business Applications”, School of Information Systems Queensland University of Technology Brisbane QLD 4001.
63. Ogneva, M. "How Companies Can Use Sentiment Analysis to Improve Their Business". Mashable, April 19 2010, Available at: https://mashable.com/2010/04/19/sentiment-analysis/?europe=true#wNPa2Yy1J5q6, (September 23 2018).
64. O’Reilly, T. and Milstein, S. (2011), The Twitter Book, O’Reilly Media
65. Padmaja S. and Prof. Fatima S. S. (2013), “Opinion Mining and Sentiment Analysis – An Assessment of Peoples’ Belief: A Survey”, International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.4, No.1.
66. Pippal S., Batra L., Krishna A., Gupta H., Arora K. (2014), “Data mining in social networking sites: A social media mining approach to generate effective business strategies”, International Journal of Innovations & Advancement in Computer Science, Volume 3, Issue 2.

67. Pournarakis, D.E., Sotiropoulos, D.N. and Giaglis, G.M. (2017), "A computational model for mining consumer perceptions in social media, Decision Support Systems", Vol.93, 98-110.
68. Purohit, H., Hampton, A., Shalin, V. L., Sheth, A. P., Flach, J., & Bhatt, S. (2013), "What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination", Computers in Human Behavior, 29(6), 2438-2447, doi:10.1016/j.chb.2013.05.007.
69. Roebuck, K. (2012). "Sentiment Analysis: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity", Vendors. ISBN 9781743049457
70. Rosen A., "Tweeting Made Easier", Twitter, November 7 2017, Available at: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html, (July 30 2018).
71. Russell, M.A. (2013), Mining the Social Web, O'reilly Media Inc., 5, 45.
72. Statista, https://www.statista.com/
73. Tan P. N., Steinbach M. & Kumar V. (2006), "Introduction to Data Mining", Chapter 6: Association Analysis: Basic Concepts and Algorithms, Available at: https://www-users.cs.umn.edu/~kumar001/dmbook/ch6.pdf, (November 19 2017).
74. Tom White (2012), "Hadoop: The Definitive Guide", O'Reilly, 3rd Edition.
75. Tutorials Point, "Hadoop – MapReduce", Online Available: https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm, (August 25 2018).
76. Twitter, "Authentication", Available at: https://developer.twitter.com/en/docs/basics/authentication/overview/oauth.html (July 30 2018).
77. Twitter, "Tweet updates", Available at: https://developer.twitter.com/en/docs/tweets/tweet-updates.html, (July 30 2018).
78. We are social, "Global Digital Report 2018", We are social, Available at: https://digitalreport.wearesocial.com/, (August 21 2018)
79. Wersm, "How much data is generated every minute?, Wersm, 2017, Available at: http://wersm.com/much-data-generated-every-minute/, (June 1 2018)
80. Yandong Wang, Robin Goldstone, Weikuan Yu & Teng Wang (2014), "Characterization and Optimization of Memory-Resident MapReduce on HPC Systems", IEEE 28th International Parallel & Distributed Processing Symposium.

81. Yesenalina A., Yue Y. & Cardie C., (2010), “Multi-level structured models for document-level sentiment classification”, In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. MIT, Massachusetts, Association for Computational Linguistics, USA, pp. 1046–1056.
82. Younis, E.M.G. (2015), “Sentiment Analysis and Text Mining for Social Media Microblogs Using Open Source Tools: An Empirical Study”, International Journal of Computer Applications, 112, 5.
83. Zafarani R., Abbasi, M. A. and Liu H. (2014), “Social Media Mining: An Introduction”, Cambridge University Press, 16.
84. Zhou H. & Song F, (2015), “Aspect-level sentiment analysis based on a generalized probabilistic topic and syntax model”, In: Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference. Association for the Advancement of Artificial Intelligence.

Παράρτημα

Σύγκριση αντικειμενικού συναισθήματος, με το συναίσθημα που αποδόθηκε σε κάθε tweet από το σύστημα.

Α/Α	Περιεχόμενο Tweet	Συναίσθημα Συστήματος	Αντικειμενικό Συναίσθημα
1	70-hour weeks and 'wtf' emails: 42 employees reveal the frenzy of working at #tesla under the 'cult' of #elonmusk… https://t.co/qrxlemyhmz	negative	negative
2	@dreamgiveaway: dream giveaway sweepstakes video: push a button. get ludicrous speed in this grand prize tesla- see how this car works	positive	positive
3	@street_insider: citron research's andrew left appears to sue tesla $tsla and elon musk https://t.co/84x44ws63f	neutral	neutral
4	fox news: utah woman sues tesla over ‘autopilot’ crash https://t.co/qbgvtvtsl4 https://t.co/yggyctcvuv	neutral	neutral
5	@tsm_myth: y'all make car buying sound like buying new shoes, living the best life gat damn.	negative	negative
6	news from @mercedesbenz- mercedes-benz eqc vs. jaguar i-pace vs. tesla model x: how they compare on paper https://t.co/pihlnmayy2	neutral	neutral
7	@david_kudla: $tsla falls to multi-week low below $295 in morning trading. #tesla $tslaq https://t.co/ut7cni1nme	neutral	neutral
8	@huby @tesla @elonmusk @bbcnews you do know that the ceo of audi is actually in jail for his involvement in the… https://t.co/bbktzu9yq9	neutral	neutral
9	@elonmusk: @mliebow @tesla much appreciated! great word-of-mouth is why model 3 is the best-selling electric car, despite no advertising	neutral	positive
10	@gerberkawasaki: no model 3 fires with 55k on the road. but ford recalls millions of vehicles. 17 have caught on fire already. be careful	neutral	neutral

11	#electriccar: mercedes announces first fully electric car to compete with tesla https://t.co/tqsuv4ogxz	neutral	positive
12	@coaching4today: musk's spacex could help fund take-private deal for tesla: report elon musk's rocket company https://t.co/21fp6wdcsa	positive	neutral
13	here are the key consumer demands via @automoblog. #5 is our pick, what's yours? https://t.co/7h20gytlio @tesla @amazon @hyundai @gm @ford	neutral	neutral
14	@susanwi77385995: @payoung28. who got the patent too the universe?? genesis: god said, "let there be light" or..."light, be"!	neutral	neutral
15	thanks @elonmusk and everyone who works hard @tesla. the electric car industry wouldn't be anywhere close to where… https://t.co/tbckgiogna	negative	positive
16	short seller andrew left files class action lawsuit against tesla and elon musk https://t.co/wp6x80z8m9 https://t.co/7opngdiyqb	negative	negative
17	@tesla @elonmusk please, pretty please, do not remove the ability to keep the rear camera in the top portion of the… https://t.co/y66k68hzzv	positive	neutral
18	i liked a @youtube video https://t.co/72fvz7qqsy first look inside the electric mercedes! - should tesla be worried?	negative	neutral
19	how long can tesla dominate the electric vehicle market? https://t.co/xl8kts3rw3	negative	neutral
20	"defendants artificially manipulated the price of tesla securities to damage the company’s short-sellers... by issues https://t.co/rtp5lpvlee	negative	neutral
21	@teslatunity: not a #tesla on the highway this morning.... https://t.co/lj5fdr3gho	neutral	neutral
22	@charlesmilander: the top 5 cars traded for tesla model 3 in 2018 are surprisingly affordable – roadshow	positive	neutral

	https://t.co/e6t7xtaypk		
23	woman who crashed her model s and broke her foot sues tesla - ars technica https://t.co/9572mdd3ih	negative	negative
24	hey fred - does @fredericlambert @electrekco cover the @tesla #model3 quality problems, such as panel gaps https://t.co/e9o5hlkki8	negative	negative
25	the mercedes-benz eqc electric vehicle is here (and tesla should be worried) | man of many https://t.co/plxkzy5g2k	negative	negative
26	@ups: .@illinoisupsers are charged up about taking the @tesla electric truck out for a test drive. we’re looking forward to introducing…	positive	positive
27	@sidlembirik: this is not looking good for @elonmusk! https://t.co/wugruejguk	positive	negative
28	@myteslabe: tesla says battery fire without crash in la was 'extraordinarily unusual occurrence', still investigating the cause	neutral	neutral
29	@oxygeno211: i call this german embrace of @tesla's mission. mark of a true leader is that he or she inspires the world. @tesla	positive	positive
30	goldman sachs says there's more trouble ahead for tesla - https://t.co/u6db9bmbap https://t.co/8ysiro41qd	negative	negative
31	short seller andrew left files class action lawsuit against #tesla and #elonmusk https://t.co/qjhm9orsxg	negative	negative
32	@elonmusk what's up with the model 3 pricing? i thought it was going to be an affordable electric vehicle that was… https://t.co/wuatnsz0be	positive	negative
33	tesla’s new bug bounty protects hackers — and your warranty https://t.co/1mxjelmasqneutral	neutral	neutral
34	@ianpavelko woohoo. i'm waiting for 5-8 months to turn into 4-7 on the configurator. may the tesla gods be with us	neutral	positive
35	@chadilac82: @elonmusk @yanquetino @tesla elon! do	neutral	negative

	you know what the hold up is in dallas? over 100 cars sitting in the lot, delivery		
36	was the @mercedesbenz #tesla killer behind the boring station-wagon? $tsla $tslaq https://t.co/qjq32ezhuw	negative	negative
37	@ndygrosso easier to believe when tesla model 3 can sell 8x more than toyota prius. and more than 14x nissan	positive	positive
38	@willfealey @elonmusk @tesla you can also add a fingerprint scanner in the future	neutral	neutral
39	tesla car owner fixed his own car after poor customer service https://t.co/wahvpm2nvm @mailonline	negative	negative
40	@joerogan can you ask elon musk if he thinks tesla shares would be worth more if he weren't on twitter?	positive	neutral
41	now if only i had a tesla to work on	neutral	neutral
42	me when a tesla almost ran me over the other day... i would've been honored	neutral	positive
43	i liked a @youtube video https://t.co/9fbmvy5spo will tesla stop for a person ? #model3	neutral	negative
44	how can people with sufficient education and capability to earn enough to afford #teslamotors cars be so incompeten… https://t.co/rgp61ku7t0	positive	positive
45	rt @reuters: mercedes' first all-electric car may present a major threat to tesla's domination https://t.co/c9gzr0rmno via @reuterstv	negative	negative
46	buying tesla stock or crypto coins is not investing, any more than buying an amazing gucci jacket is investing. https://t.co/faixhsocc6	positive	negative
47	@z3pt4r @zacksjerryrig @mercedesbenz probably not. the reason tesla doesn't lose range is that you charge it once https://t.co/6oj2hx8lic	neutral	positive
48	there is a lot more to tesla than the electric car. the real innovation will be in the energy storage systems https://t.co/fz9qqw5fn	positive	positive

49	@arlette_lisa but hey tesla isn’t going anywhere so you can let it do it’s thing 👌	neutral	neutral
50	@rocketl49 @kimbal @tesla are they confirming that is the case? or are they merely celebrating that a blog believes… https://t.co/rbzdpsrm9t	negative	negative
51	@radiopete22: elon musk: the volatile visionary at risk of steering tesla off the road https://t.co/ajbyluubfe	negative	negative
52	tesla to expand robotics opportunities for nevada schools. https://t.co/g4ex7hbsal	neutral	neutral
53	@elonmusk conquer your horizons #tesla https://t.co/lkumbszh2p	positive	positive
54	@businessinsider: audi has started production for its tesla rival — here's what we know about it so far https://t.co/2xgnx73mdq	negative	neutral
55	mercedes unveils electric car in direct german challenge to tesla https://t.co/snzyhgryfg https://t.co/uvj5sedvce	negative	neutral
56	70-hour weeks and 'wtf' emails: 42 employees reveal the frenzy of working at tesla under the 'cult' of elon musk	negative	neutral
57	@teslastats: welcome to https://t.co/xfov9xxd9o, your ultimate resource for any tesla stats and anything similar to that. see real-time…	positive	positive
58	everyone stop what you’re doing and look at me!! tesla $tsla $tslaq	neutral	positive
59	@lisaabramowicz1: tesla bonds due 2025 just hit a new low of a little more than 86 cents on the dollar, for an implied yield of 7.9%	negative	negative
60	@truththeorycom #einstein hands down. carl jung and tesla probably tie for second choice...	negative	negative
61	@teslamotorsclub: musk says access to v9 software is weeks away https://t.co/24zw5fz0cm #tesla https://t.co/7krddstlzo	neutral	neutral

62	tesla and musk sued for securities fraud https://t.co/x1yf6z9xl0 $tsla	negative	negative
63	@livingtesla: tesla model 3 tuxmat all-weather floor mats https://t.co/uv5bxr2is4 #win via @gleamapp @teslamotorsclub @model3owners #tesla	neutral	neutral
64	i liked a @youtube video https://t.co/y7uy46mrdb model rocket vs a motor home - tesla supercar? - new amazing mods in beamng drive	positive	positive
65	five new lawsuits on tesla and musk the fraudulent $tsla #tesla #saudiarabia https://t.co/ftkzhyaymx	negative	negative
66	#electriccar: mercedes rolls out new electric car to taker on tesla ... https://t.co/wk3i7c6nma	neutral	neutral
67	#cleanenergyke this is arguably the most important chart in energy: the falling price of batteries. elon musk https://t.co/knv1peuxin	positive	positive
68	@hvmtimeline: it's year 2020 ruins of ancient civilization,#tesla superchargers! https://t.co/ghsoeqxqcl	neutral	negative
69	@phoennix10: @cardudetopanga @wintoncapptnrs nice find. white interior, too. @benshooter another data point for special edition white…	positive	positive
70	@ophirgottlieb: $tsla software and disruption https://t.co/8wjkmxspbr "being first isn’t enough and having a great product isn’t enough…	negative	negative
71	@applehelix: unimpressed with the mercedes eqc. looks like a cross between a toyota highlander and a subaru outback. won't be a tesla ki…	negative	positive
72	@teslamotrsports: before you give up, think of the reason why you held on so long🔌🔋 #tesla https://t.co/3qcekp7waf	neutral	neutral
73	not just deli, sleepy's mattress among greenburgh stores unplugged by tesla https://t.co/ebfzcppdfs	neutral	neutral
74	'tesla killer' narrative comes back as german automakers	negative	positive

	launch ev offensive - don't fall for it… https://t.co/pohik4cnus		
75	@madmagazine: tesla employees frequently find elon musk sleeping under tables and desks. "when he gets like that we spray him with a mist	neutral	neutral
76	@teslarati tesla needs to make a cool looking reimagined wheel without causing more breakable parts in order to keep… https://t.co/7jwyhfxuul	neutral	negative
77	@valueanalyst1 if tesla wants to burn the shorts, all they need to do is build a better car that isn't in the shop… https://t.co/4onqdfttdv	negative	negative
78	@greatcrasho2018: hey @joerogan, here's 3 important questions for @elonmusk tonight: 1. what will he miss most about having been @tesla…	neutral	neutral
79	$tsla bounces can fail against 327.32 high in 3, 7 or 11 swing for a move lower towards 262.24-246.79 area next #elliottwave #trading #tesla	negative	negative
80	ok but collab with tesla	neutral	neutral
81	@elonmusk @yanquetino @tesla elon, i didn't know you were such a hero. thank you for speaking up for all the child… https://t.co/96mlmutnfp	positive	positive
82	opportunity cost of forgetting to charge or update failures is more significant than investors realize. the cars co… https://t.co/indgxdgykl	neutral	neutral
83	@olivierw4: wow! @tesla model 3 sales hit epic volume in august and nr 1 🏆in u.s. plug-in market in 2018 already confirmed	positive	positive
84	a @benedictevans post about tesla is, as they say, self-recommending. https://t.co/xpjqhjjwmm	neutral	neutral
85	“like tesla, mercedes is announcing eqc orders in norway even before its price. it has amassed more than 2k refund… https://t.co/thks6sokl2	positive	neutral

86	tesla bond hits record low, stock slips as investor worry depends https://t.co/dbu2omodwi https://t.co/aiuzouhh00	negative	negative
87	@gwestr: tesla needed $10b to build a scale electric automaker and battery facility.	neutral	neutral
88	@amitelazari: thanks @elonmusk @tesla for picking up this suggestion and adopting a legal safe harbor for researchers in your bug bounty…	positive	positive
89	amazon dethrones tesla as the stock short sellers hate the most (tsla, amzn) https://t.co/ixvsd3efot	negative	negative
90	@electrekco: tesla semi receives order of 30 more electric trucks from walmart https://t.co/3emgga6zk4	neutral	neutral
91	@chris_serro @tesla the 3 is growing on me a lot man. i've started looking at those more and more.	neutral	positive
92	if bitcoin could moon to 30k so i can forgive myself for buying this baby #tesla that would be real great.	positive	positive
93	@cazin678 be honest. elon made tesla a footnote in history. they were never going to be huge, but now they wont even live.	positive	negative
94	i liked a @youtube video https://t.co/ntel4mspwd we ran into the tesla semi today	neutral	neutral
95	somebody is in trouble #tesla	negative	negative
96	shortseller citron files lawsuit against tesla, musk https://t.co/ncgcj4sadj #news #reuters	negative	negative
97	@markleggett: due to the erratic, irresponsible actions of ceo elon musk, i hereby announce that i am boycotting all tesla products.	negative	negative
98	are tesla shares set to fall? goldman sachs thinks so. https://t.co/kkkwiynv73 via @mercnews	negative	negative
99	@tesla @elonmusk @roadshow and so pretty!	positive	positive
100	@rtehrani: don’t tell #uber or #tesla about this https://t.co/yfzm5ppnke	neutral	neutral