ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Αναγνώριση Συναισθηματικής Κατάστασης στα Κοινωνικά Δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

των

Σταυριανού Κωνσταντίνου
Αδάμ Ιωάννη

Επιβλέπων : Σταματάτος Ευστάθιος

Μέλη εξεταστικής επιτροπής: Εργίνα Καβαλλιεράτου, Ανδρέας Παπασαλούρος

Σάμος, Φεβρουάριος 2020

Πρόλογος και ευχαριστίες

Θα θέλαμε να ευχαριστήσουμε τον καθηγητή κ. Σταματάτο Ευστάθιο που είχε την επίβλεψη της συγκεκριμένης διπλωματικής εργασίας για την ευκαιρία που μας έδωσε να ασχοληθούμε με ένα τόσο ενδιαφέρον και σύγχρονο θέμα.

Σταυριανού Κων/νου και Αδάμ Ιωάννη

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Πίνακας περιεχομένων

1 Εισαγωγή 1

1.1 Αναγνώριση συναισθηματικής κατάστασης (Affect Detection) 2

1.2 Μηχανική Μάθηση (Machine Learning) 3

1.2.1 Ιστορία μηχανικής μάθησης 4

1.2.2 Εφαρμογές Μηχανικής Μάθησης 6

1.3 Εξόρυξη Κειμένου (Text Mining) 7

1.3.1 Σημασία της εξαγωγής πληροφοριών 9

1.3.2 Αποτελέσματα και επιδόσεις της εξόρυξης κειμένου 10

1.4 Δυσκολίες και Προκλήσεις 11

1.5 Στόχοι της εργασίας 13

2 Επιβλεπόμενη μηχανική μάθηση 14

2.1 Εισαγωγή στην επιβλεπόμενη μηχανική μάθηση 14

2.1.1 Classification 15

2.1.2 Regression 16

2.2 Αλγόριθμοι 18

2.2.1 Random Forest 18

Πλεονεκτήματα και Μειονεκτήματα 19

Πλεονεκτήματα 19

Μειονεκτήματα 19

2.2.2 Logistic Regression 20

3 Αναγνώριση συναισθηματικής κατάστασης στο SemEval-2018 22

3.1 Παρουσίαση διαγωνισμού SemEval-2018 22

3.1.1 Δημιουργία δεδομένων διαγωνισμού (Dataset) 24

3.1.2 Αξιολόγηση για αυτόματες προβλέψεις 28

3.2 Μέτρηση απόδοσης αλγορίθμων 31

3.2.1 Συντελεστής Συσχέτισης Pearson (Pearson Correlation Coefficient) 31

3.2.2 Συντελεστής Συσχέτισης – Βασικά 32

3.2.3 Συντελεστής Συσχέτισης - Προβλήματα Διερμηνείας 34

3.2.4 Πίνακας Συσχέτισης 34

3.2.5 Jaccard Index 36

3.3 Αποτελέσματα διαγωνισμού 37

3.3.1 Προσεγγίσεις Ομάδων υψηλής βαθμολογίας 37

4 Η προσέγγιση μας 42

4.1 Τεχνική που ακολουθήσαμε 42

4.1.1 Πρώτο στάδιο (Προ-επεξεργασία Δεδομένων) 43

4.1.2 Χαρακτηριστικά (features) – Διανυσματοποίηση 47

4.1.3 Ταξινομητές 56

4.2 Παρεμφερείς προσεγγίσεις με συμμετέχοντες διαγωνισμού 57

5 Πειραματικά αποτελέσματα 60

5.1 Υπολογισμός Ακρίβειας για αποτελέσματα 60

5.2 Λειτουργία και ανάλυση υπολογισμού μέτρου απώλειας μοντέλου παλινδρόμησης 60

5.3 Λειτουργία και ανάλυση υπολογισμού μέτρου απώλειας μοντέλου κατηγοριοποίησης 62

5.4 Ανάλυση συστήματος και αποτελεσμάτων 64

6 Συμπεράσματα - Προοπτικές 71

6.1.1 Εφαρμογές 71

6.1.2 Ανακεφαλαίωση 72

6.1.3 Συμπεράσματα γενικής θεματολογίας 72

6.1.4 Συμπεράσματα διαγωνισμού 74

6.1.5 Άνθρωποι ή μηχανές 74

6.1.6 Που μπορούν να βοηθήσουν οι μηχανές 75

6.1.7 Επόμενα βήματα 76

Βιβλιογραφία 77

Περίληψη

Τα τελευταία χρόνια, με την αυξανόμενη χρήση Web 2.0 εφαρμογών έχουν δημιουργηθεί καινούργιοι τρόποι επικοινωνίας. Τα μέσα κοινωνικής δικτύωσης περιέχουν ένα θησαυρό πληροφοριών, μεταξύ των οποίων γνώμες και συναισθήματα των χρηστών τους και ο σκοπός της Αναγνώρισης Συναισθηματικής Κατάστασης είναι να βγάλει χρήσιμα συμπεράσματα από αυτά. Ειδικότερα, οι επιχειρήσεις μπορούν να αξιοποιήσουν αυτές τις πληροφορίες ώστε να γίνουν πιο ανταγωνιστικές και να βελτιώσουν το επίπεδο των παρεχόμενων προϊόντων και υπηρεσιών τους. Πιο συγκεκριμένα, η παρούσα διπλωματική εργασία ασχολείται με τις τεχνικές και τα εργαλεία της αναγνώρισης συναισθηματικής κατάστασης. Παρουσιάζονται κατηγορίες στις οποίες μπορεί να ενταχθεί η αναγνώριση συναισθηματικής κατάστασης, ως προς τις χρήσεις της, και παραδείγματα αυτών. Έπειτα, σε τεχνικό επίπεδο, παρατίθενται οι κατηγορίες σε σχέση με την προσέγγιση κειμένου, και σε σχέση με την προσέγγιση της τεχνικής που ακολουθούμε κατά την διαδικασία. Ύστερα, εξετάζονται εκτεταμένα οι κατηγορίες των τεχνικών οι οποίες είναι οι τεχνικές με λεξικά, τεχνικές με επιβλεπόμενη μηχανική μάθηση, τεχνικές μη-επιβλεπόμενη μηχανική μάθηση, υβριδικές τεχνικές. Παρουσιάζονται παραδείγματα από μελέτες πάνω στην κάθε κατηγορία και συγκεντρωτικοί πίνακες που βοηθούν στην επισκόπηση της τεχνολογίας αιχμής και εξάγονται χρήσιμα συμπεράσματα. Έπειτα, με γνώμονα τη χρησιμότητα της αναγνώρισης συναισθηματικής κατάστασης γενικά, αλλά και ειδικότερα στις επιχειρήσεις, γίνεται επισκόπηση των εφαρμογών και δίνονται παραδείγματα αυτών. Τέλος, συμπεραίνονται κάποιες χρήσιμες παρατηρήσεις για την παρούσα κατάσταση που βρίσκεται η αναγνώριση συναισθηματικής κατάστασης, αλλά και για το μέλλον αυτής.

Λέξεις Κλειδιά: *ανάλυση συναισθηματικής κατάστασης, (μη) επιβλεπόμενη μάθηση, λεξικά, εξόρυξη κειμένων, εκπαίδευση συστήματος*

Abstract

In recent years, with the increasing use of Web 2.0 applications they have create new ways of communicating. Social media contain spectacular information, including the opinions and feelings of their users and the purpose of Sensitivity Analysis is to draw useful conclusions from them. In particular, businesses can leverage this information to make them more competitive and improve the level of their products and services. In particular, this thesis deals with the techniques and tools of affect detection. There are categories that can be incorporated into the analysis of emotion, its uses, and examples. Then, at the technical level, the categories are listed in relation to the text approach, and in relation to the technique we follow in the process. Then, the categories of techniques which are lexical techniques, supervised machine learning techniques, non-supervised machine learning techniques, and hybrid techniques are examined extensively. Examples of studies on each category and aggregate tables are presented to help review the state of the art and draw useful conclusions. Then, based on the usefulness of affect detection in general, but also in business in particular, applications are reviewed and examples are given. Finally, there are some useful observations about the present state of affect detection and the future of it.

Keywords: *affect detection, dataset, supervised learning, text mining, machine learning*

1

Εισαγωγή

Τη σημερινή εποχή, το διαδίκτυο έχει διευρυνθεί και έχει αλλάξει σε ένα μεγάλο βαθμό την καθημερινότητα των ανθρώπων καθώς τους προσφέρει νέους τρόπους να επικοινωνούν και να ενημερώνονται. Οι χρήστες του διαδικτύου δεν είναι πλέον παθητικοί αποδέκτες πληροφοριών αλλά μπορούν να συμμετέχουν σε κοινωνικά δίκτυα όπως το Twitter και να συζητούν, να μοιράζονται τη γνώμη τους και να ανταλλάζουν απόψεις και ιδέες. Έτσι παρατηρούμε ότι οι παλαιότεροι τρόποι επικοινωνίας όπως για παράδειγμα το e-mail έχουν σχεδόν σταματήσει να υπάρχουν. Οι πλατφόρμες κοινωνικής δικτύωσης έχουν αντικαταστήσει το παλαιό τρόπο επικοινωνίας, τα e-mails καθώς οι χρήστες μπορούν να εκφράσουν τη γνώμη τους, να κάνουν μια κριτική σχετικά με ένα προϊόν που αγόρασαν ή ακόμη και μια υπηρεσία που χρησιμοποιούν. Τα δεδομένα αυτά είναι πολύ σημαντικά για πολλούς παράγοντες και επειδή βρίσκονται σε ηλεκτρονική μορφή η αποθήκευση τους είναι αρκετά εύκολη. Η ανάλυση συναισθήματος (Sentiment Analysis) και ανάλυση συναισθηματικής κατάστασης (Affect Detection) έκανε την εμφάνιση της διότι υπήρχε η ανάγκη ανάλυσης και αξιοποίησης των πληροφοριών αυτών με αυτοματοποιημένο τρόπο. Έτσι η αναγνώριση συναισθηματικής κατάστασης στα κοινωνικά δίκτυα έχει αναπτυχθεί στον επιστημονικό χώρο καθώς και στον ακαδημαϊκό χώρο.

1.1 Αναγνώριση συναισθηματικής κατάστασης (Affect Detection)

Η ανίχνευση συναισθημάτων από το κείμενο είναι ένα πρόσφατο πεδίο έρευνας που σχετίζεται στενά με την Ανάλυση Συναισθημάτων. Η Ανάλυση Συναισθημάτων στοχεύει στον εντοπισμό θετικών, ουδέτερων ή αρνητικών συναισθημάτων από το κείμενο, ενώ η Αναγνώριση συναισθηματικής κατάστασης αποσκοπεί στην ανίχνευση και αναγνώριση των τύπων συναισθημάτων μέσω της έκφρασης κειμένων, όπως ο θυμός, η αηδία, ο φόβος, η ευτυχία, η θλίψη, η έκπληξη και άλλα.

Η ανίχνευση της συναισθηματικής κατάστασης ενός ατόμου με την ανάλυση κειμένου που γράφτηκε από οποιονδήποτε φαίνεται δύσκολη. Η αναγνώριση του συναισθήματος του κειμένου παίζει βασικό ρόλο στην αλληλεπίδραση ανθρώπου-υπολογιστή. Τα συναισθήματα μπορούν να εκφράζονται από την ομιλία ενός ατόμου, την έκφραση προσώπου ή ένα γραπτό κείμενο. Έχει γίνει αρκετή δουλειά όσον αφορά την αναγνώριση συναισθημάτων λόγου και προσώπου, αλλά το σύστημα αναγνώρισης συναισθημάτων με βάση το κείμενο χρειάζεται ακόμα έλξη ερευνητών. Στην υπολογιστική γλωσσολογία, η ανίχνευση των ανθρώπινων συναισθημάτων στο κείμενο γίνονται ολοένα και πιο σημαντικά από πρακτική άποψη. Η ανίχνευση του συναισθήματος του συγγραφέα από το κείμενο στοχεύει να εντοπίσει τα συναισθήματα που κρύβονται με τη μελέτη της εισόδου, δηλαδή κείμενα γραμμένα από συγγραφέα. Αυτό βασίζεται στην αρχή ότι εάν ένα άτομο είναι ευτυχισμένο, αυτό τους επηρεάζει να χρησιμοποιούν θετικές λέξεις. Ομοίως, εάν ένα άτομο είναι λυπημένο, απογοητευμένο ή θυμωμένο, το είδος των λέξεων που χρησιμοποιούν μπορεί να μας επιτρέψει να συμπεράνουμε το αρνητικό συναίσθημα.

Σήμερα οι περισσότεροι άνθρωποι αλληλεπιδράσουν μέσω του διαδικτύου με τη μορφή κειμένου, γράφοντας κείμενα σε blogs, ηλεκτρονικό ταχυδρομείο, γράφοντας κριτικές και σχόλια σε μέσα κοινωνικής δικτύωσης. Η Αναγνώριση Συναισθηματικής Κατάστασης αναλύει τα ηλεκτρονικά κείμενα και αναφέρει τα συναισθήματα των συγγραφέων ή τις απόψεις τους προς ένα συγκεκριμένο αντικείμενο ή οποιαδήποτε από τα υποσυνείδητα χαρακτηριστικά του ή πτυχές. Για να αξιολογήσουμε τα συναισθήματα στην ανάλυση του συναισθήματος, πρέπει να εντοπίσουμε όρους οι οποίοι αποτελούν τη βάση για την ανάλυση αισθήσεων για να εκτελέσει την αξιολόγηση.

1.2 Μηχανική Μάθηση (Machine Learning)

Μηχανική μάθηση είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Το 1959, ο Άρθουρ Σάμουελ ορίζει τη μηχανική μάθηση ως "Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί". Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Η μηχανική μάθηση είναι στενά συνδεδεμένη και συχνά συγχέεται με υπολογιστική στατιστική, ένας κλάδος, που επίσης επικεντρώνεται στην πρόβλεψη μέσω της χρήσης των υπολογιστών. Έχει ισχυρούς δεσμούς με την μαθηματική βελτιστοποίηση, η οποία παρέχει μεθόδους, τη θεωρία και τομείς εφαρμογής. Η Μηχανική μάθηση εφαρμόζεται σε μια σειρά από υπολογιστικές εργασίες, όπου τόσο ο σχεδιασμός όσο και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος. Παραδείγματα εφαρμογών αποτελούν τα φίλτρα spam (spam filtering), η οπτική αναγνώριση χαρακτήρων (OCR), οι μηχανές αναζήτησης και η υπολογιστική όραση. Η Μηχανική μάθηση μερικές φορές συγχέεται με την εξόρυξη δεδομένων, όπου η τελευταία επικεντρώνεται περισσότερο στην εξερευνητική ανάλυση των δεδομένων, γνωστή και ως μη επιτηρούμενη μάθηση.

Στο πεδίο της ανάλυσης δεδομένων, η μηχανική μάθηση είναι μια μέθοδος που χρησιμοποιείται για την επινόηση πολύπλοκων μοντέλων και αλγορίθμων που οδηγούν στην πρόβλεψη. Τα αναλυτικά μοντέλα επιτρέπουν στους ερευνητές, τους επιστήμονες δεδομένων, τους μηχανικούς και τους αναλυτές να παράγουν αξιόπιστες αποφάσεις και αποτελέσματα και να αναδείξουν αλληλοσυσχετίσεις μέσω της μάθησης από ιστορικές σχέσεις και τάσεις στα δεδομένα.

Στη μηχανική μάθηση το κύριο αντικείμενό είναι η μελέτη αλγορίθμων οι οποίοι επιτρέπουν στους υπολογιστές να μαθαίνουν και να εξελίσσονται, χωρίς να είναι προγραμματισμένοι. Λειτουργούν αυτοματοποιημένα για να εξάγουν την πληροφορία και χρησιμοποιούνται για να προβλέπουν καινούργια δεδομένα. Συχνά χρησιμοποιείται ένα σύνολο δεδομένων εκπαίδευσης (dataset) από το οποίο ο

αλγόριθμος είναι σε θέση να αντιληφθεί την δομή τους και να αναζητήσει ένα μοντέλο που θα χειρίζεται άγνωστα νέα δεδομένα.

Οι κατηγορίες μηχανικής μάθησης είναι :

Α) Η επιβλεπόμενη μάθηση όπου στο σύστημα κάνουν την εμφάνισή τους κάποια δεδομένα εισόδου και είναι σε θέση να μάθει πώς οι είσοδοι αντιστοιχίζονται στην κάθε επιθυμητή έξοδο. Με αυτό το τρόπο μπορεί να προβλέπει μελλοντικά εισόδους άγνωστης εξόδου. Αλγόριθμοι που ανήκουν σ'αυτή την κατηγορία είναι ο Naïve Bayes, ο SVM και ο Maximum Entropy.

Β) Μη επιβλεπόμενη μάθηση ή μάθηση χωρίς επίβλεψη: ο αλγόριθμός δημιουργεί ένα μοντέλο για μια σειρά εισόδων, χωρίς να είναι γνωστές οι έξοδοι. Οι αλγοριθμοι μη επιβλεπόμενης μάθησης είναι αλγόριθμοι ομαδοποίησης, όπως k-means, k-methoids, hierarchical clustering.

Γ) Ημι-επιβλεπόμενη μάθηση: Η εκπαίδευση των ταξινομητών επιτυγχάνεται με γνωστές και άγνωστες εξόδους.

1.2.1 Ιστορία μηχανικής μάθησης

Ως επιστημονικό εγχείρημα, η μηχανική μάθηση αναπτύχθηκε από την αναζήτηση για την τεχνητή νοημοσύνη. Ήδη από την πρώιμη περίοδο της έρευνας στον τομέα της τεχνητής νοημοσύνης σε ακαδημαϊκό επίπεδο, το ζήτημα της κατασκευής μηχανών που θα μάθαιναν από δεδομένα απασχόλησε τους ερευνητές. Προσπάθησαν να προσεγγίσουν το πρόβλημα με διάφορες συμβολικές μεθόδους, καθώς και με τα λεγόμενα νευρωνικά δίκτυα. Αυτά ήταν ως επί το πλείστον perceptrons[1] και μοντέλα, που όπως διαπιστώθηκε αργότερα ήταν επανεφευρέσεις των γενικευμένων γραμμικών μοντέλων της στατιστικής. Επίσης χρησιμοποιήθηκε η πιθανοθεωρητική λογική, ιδιαίτερα στην αυτοματοποιημένη ιατρική διάγνωση.

Ωστόσο, μια αυξανόμενη έμφαση σε προσεγγίσεις που βασίζονται στην λογική γνώση προκάλεσε ένα ρήγμα μεταξύ Τεχνητής Νοημοσύνης και μηχανικής μάθησης. Τα πιθανοθεωρητικά συστήματα μαστίζονταν από θεωρητικά και πρακτικά προβλήματα απόκτησης δεδομένων και αναπαράστασής τους. Από το 1980,έμπειρα συστήματα επικράτησαν στο πεδίο της Τεχνητής Νοημοσύνης, και ο

[1] https://el.wikipedia.org/wiki/Perceptron

ρόλος της στατιστικής υποχώρησε. Η εργασία σε συμβολική/βασισμένη σε γνώση εκμάθηση συνεχίστηκε εντός της ΤΝ, οδηγώντας στον επαγωγικό λογικό προγραμματισμό, αλλά οι κατευθυντήριες γραμμές της στατιστικής ήταν τώρα έξω από το χώρο της τεχνητής νοημοσύνης, στην αναγνώριση προτύπων και στην ανάκτηση πληροφοριών. Η έρευνα για νευρωνικά δίκτυα εγκαταλείφτηκε από την ΤΝ και την Επιστήμη Υπολογιστών τον ίδιο περίπου καιρό. Η ίδια επίσης κατεύθυνση ακολουθήθηκε πέρα από την ΤΝ και την πληροφορική, από ερευνητές άλλων ειδικοτήτων, συμπεριλαμβανομένων των Hopfield, Rumelhart και Hinton. Η επιτυχία ήρθε στα μέσα της δεκαετίας του 1980 με την επανεφεύρεση της μεθόδου ανάστροφης μετάδοσης (backpropagation).

Η Μηχανική μάθηση, αναδιοργανώθηκε ως ένα ξεχωριστό πεδίο, που άρχισε να ακμάζει κατά τη δεκαετία του 1990. Η προσοχή μετατοπίστηκε από τις συμβολικές προσεγγίσεις που κληρονόμησε από την Τεχνητή Νοημοσύνη, που στόχο είχαν την αντιμετώπιση επιλήσμων προβλημάτων πρακτικής φύσης, και δόθηκε έμφαση σε μεθόδους και μοντέλα της στατιστικής και της θεωρίας πιθανοτήτων. Επίσης επωφελήθηκε από την διαθεσιμότητα ψηφιοποιημένων πληροφοριών και της δυνατότητας να διανεμηθούν μέσω του Διαδικτύου.

Η Μηχανική μάθηση και η εξόρυξη δεδομένων συχνά χρησιμοποιούν τις ίδιες μεθόδους και επικαλύπτονται σημαντικά. Μπορούν να διακριθούν ως εξής :

- Η μηχανική μάθηση εστιάζει στην πρόβλεψη, που βασίζεται σε γνωστές ιδιότητες που απορρέουν από το σύνολο εκπαίδευσης.
- Η εξόρυξη δεδομένων εστιάζει στην ανακάλυψη ιδιοτήτων μη γνωστών εκ των προτέρων. Αυτό είναι το βήμα ανάλυσης στην ανακάλυψη γνώσης από βάσεις δεδομένων

Οι δύο τομείς επικαλύπτονται με πολλούς τρόπους. Η εξόρυξη δεδομένων χρησιμοποιεί πολλές μεθόδους μηχανικής μάθησης, αλλά συχνά με διαφορετικούς στόχους. Από την άλλη πλευρά και η μηχανική μάθηση χρησιμοποιεί μεθόδους εξόρυξης δεδομένων, όπως η μη επιτηρούμενη μάθηση, ή στο στάδιο προεπεξεργασίας για να βελτιώνει την ακρίβεια της μάθησης. Ένα μεγάλο μέρος της σύγχυσης μεταξύ των δύο ερευνητικών τομέων (που συχνά έχουν ξεχωριστά συνέδρια και περιοδικά, με το ECML PKDD να αποτελεί σημαντική εξαίρεση) προκύπτει από τις βασικές υποθέσεις πάνω στις οποίες και οι δύο δουλεύουν. Όμως, στην μηχανική μάθηση η απόδοση συνήθως αξιολογείται ως προς την ικανότητα αναπαραγωγής γνώσης, την οποία ήδη κατέχουμε, ενώ στην ανακάλυψη γνώσης και την εξόρυξη δεδομένων το κλειδί είναι η ανακάλυψη γνώσης που δεν προκατέχουμε. Στην πρώτη περίπτωση μια μέθοδος επιτηρούμενης μάθησης μπορεί να έχει καλύτερα αποτελέσματα, ενώ σε μία τυπική διεργασία Ανακάλυψης Γνώσης και Εξόρυξης δεδομένων οι επιτηρούμενες μέθοδοι μάθησης δεν λειτουργούν εξαιτίας της μη διαθεσιμότητας συνόλου εκπαίδευσης

Η μηχανική μάθηση συνδέεται επίσης με την βελτιστοποίηση: πολλά προβλήματα μάθησης διατυπώνονται ως η ελαχιστοποίηση της συνάρτησης απώλειας από ένα σύνολο δεδομένων εκπαίδευσης. Η συνάρτηση απώλειας εκφράζει τη διαφορά μεταξύ των προβλέψεων του εκπαιδευμένου μοντέλου και των πραγματικών καταστάσεων του προβλήματος. Η διαφορά των δύο τομέων απορρέει από τον στόχο της γενίκευσης: ενώ οι αλγόριθμοι βελτιστοποίησης μπορούν να ελαχιστοποιήσουν την απώλεια ενός συνόλου εκπαίδευσης, η μηχανική μάθηση εστιάζει στην ελαχιστοποίηση της απώλειας σε άγνωστες καταστάσεις.

1.2.2 Εφαρμογές Μηχανικής Μάθησης

- Αναγνώριση ομιλίας και γραφικού χαρακτήρα
- Ανάκτηση πληροφορίας
- Βελτιστοποίηση
- Βιοπληροφορική
- Διαδικτυακή Διαφήμιση
- Εντοπισμός Διαδικτυακής απάτης
- Εντοπισμός απάτης πιστωτικής κάρτας
- Επεξεργασία φυσικής γλώσσας
- Ηλεκτρονικά παιχνίδια
- Ιατρική Διάγνωση
- Κατηγοριοποίηση ακολουθιών DNA
- Λογισμικά
- Μαρκετινγκ
- Μετακίνηση Ρομπότ
- Μηχανές αναζήτησης
- Μηχανική αντίληψη
- Οικονομία
- Συναισθηματική υπολογιστική
- Συστήματα σύστασης
- Υπολογιστική ανατομία
- Υπολογιστική όραση- συμπεριλαμβανομένης της αναγνώρισης αντικειμένου
- Χημειοπληροφορική
- Χρηματιστηριακή ανάλυση

1.3 Εξόρυξη Κειμένου (Text Mining)

Η εξόρυξη δεδομένων είναι η διαδικασία της ανεύρεσης σχεδίων σε μεγάλα σύνολα δεδομένων που περιλαμβάνουν μεθόδους στη διασταύρωση της μηχανικής μάθησης , των στατιστικών στοιχείων και των συστημάτων βάσεων δεδομένων. Η εξόρυξη δεδομένων είναι ένας διεπιστημονικός υποτομέας της επιστήμης των υπολογιστών και των στατιστικών, με γενικό στόχο την εξαγωγή πληροφοριών (με έξυπνες μεθόδους) από ένα σύνολο δεδομένων και τη μετατροπή των πληροφοριών σε κατανοητή δομή για περαιτέρω χρήση. Η εξόρυξη δεδομένων είναι το βήμα ανάλυσης της διαδικασίας "αποκάλυψης γνώσεων σε βάσεις δεδομένων" ή Knowledge Discovery in Databases (KDD). Εκτός από το πρώτο βήμα ανάλυσης, περιλαμβάνει επίσης θέματα διαχείρισης βάσεων δεδομένων και δεδομένων , προεπεξεργασία δεδομένων , εκτιμήσεις μοντέλων και συμπερασμάτων , μετρήσεις ενδιαφερόντων, εκτιμήσεις πολυπλοκότητας , μετα-επεξεργασία ανακαλυφθέντων δομών, οπτικοποίηση και ενημέρωση μέσω διαδικτύου. Σύμφωνα με τους Hotho μπορούμε να διακρίνουμε τρεις διαφορετικές προοπτικές εξόρυξης κειμένου, δηλαδή την εξόρυξη κειμένου ως εξόρυξη πληροφοριών, την εξόρυξη κειμένου ως εξόρυξη δεδομένων κειμένου και την εξόρυξη κειμένου ως διαδικασία Knowledge Discovery in Databases (KDD).

Ο όρος "εξόρυξη δεδομένων" είναι μια εσφαλμένη ονομασία , επειδή ο στόχος είναι η εξαγωγή μοτίβων και γνώσεων από μεγάλες ποσότητες δεδομένων, όχι η εξαγωγή (εξόρυξη) των ίδιων των δεδομένων. Είναι επίσης μια λέξη - κλειδί και εφαρμόζεται συχνά σε οποιαδήποτε μορφή επεξεργασίας δεδομένων ή πληροφοριών μεγάλης κλίμακας (συλλογή , εξαγωγή , αποθήκευση , ανάλυση και στατιστικά στοιχεία) καθώς και οποιαδήποτε εφαρμογή συστήματος υποστήριξης αποφάσεων υπολογιστών , την τεχνητή νοημοσύνη (π.χ. μηχανική μάθηση) και την επιχειρηματική ευφυΐα .

Η εξόρυξης δεδομένων είναι η ημιαυτόματη ή αυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένων για την εξαγωγή προηγουμένως άγνωστων ενδιαφερόντων προτύπων όπως ομάδων αρχείων δεδομένων (ανάλυση συμπλέγματος), ασυνήθιστων αρχείων (ανίχνευση ανωμαλιών) και εξαρτήσεων (εξόρυξη κανόνα σύνδεσης , διαδοχική εξόρυξη προτύπων). Αυτό συνήθως συνεπάγεται τη χρήση

τεχνικών βάσης δεδομένων, όπως χωρικών δεικτών . Αυτά τα πρότυπα μπορούν στη συνέχεια να θεωρηθούν ως ένα είδος περίληψης των δεδομένων εισόδου και μπορούν να χρησιμοποιηθούν σε περαιτέρω ανάλυση ή για παράδειγμα στη μηχανική μάθηση και στις προβλέψεις ανάλυσης. Για παράδειγμα, η εξόρυξη δεδομένων μπορεί να εντοπίσει πολλαπλές ομάδες στα δεδομένα, τα οποία στη συνέχεια μπορούν να χρησιμοποιηθούν για την επίτευξη ακριβέστερων αποτελεσμάτων πρόβλεψης από ένα σύστημα υποστήριξης αποφάσεων . Ούτε η συλλογή δεδομένων, ούτε η προετοιμασία δεδομένων, ούτε η ερμηνεία και η αναφορά αποτελεσμάτων αποτελούν μέρος της εξόρυξης δεδομένων, αλλά ανήκουν στη συνολική διαδικασία Knowledge Discovery in Databases (KDD) ως πρόσθετα βήματα.

Η ανάλυση κειμένων περιλαμβάνει την ανάκτηση πληροφοριών , τη λεξική ανάλυση για τη μελέτη των κατανομών συχνότητας λέξεων, την αναγνώριση προτύπων , την επισήμανση ,τον σχολιασμό ,την εξαγωγή πληροφοριών ,τις τεχνικές εξόρυξης δεδομένων, της απεικόνισης και των προγνωστικών αναλύσεων. Ο πρωταρχικός στόχος είναι, ουσιαστικά, η μετατροπή του κειμένου σε δεδομένα για ανάλυση, μέσω εφαρμογής επεξεργασίας φυσικής γλώσσας (NLP), διαφορετικών τύπων αλγορίθμων και αναλυτικών μεθόδων. Μια σημαντική φάση αυτής της διαδικασίας είναι η ερμηνεία των πληροφοριών που συλλέγονται.

Μια τυπική εφαρμογή είναι η σάρωση ενός συνόλου εγγράφων γραμμένων σε μια φυσική γλώσσα και είτε η μοντελοποίηση του εγγράφου που έχει οριστεί για σκοπούς ταξινόμησης πρόβλεψης ή η συμπλήρωση μιας βάσης δεδομένων ή ενός ευρετηρίου αναζήτησης με τις πληροφορίες που εξάγονται. Το έγγραφο είναι το βασικό στοιχείο κατά την εκκίνηση με την εξόρυξη κειμένου. Ορίζουμε ένα έγγραφο ως μια μονάδα κειμένων δεδομένων, η οποία συνήθως υπάρχει σε πολλούς τύπους συλλογών.

Το Data Mining συμπεριλαμβάνει τις ακόλουθες κατηγορίες διαδικασιών

- **Anomaly detection (Outlier/change/deviation detection)**

Η αναγνώριση μη συνηθισμένων δεδομένων, τα οποία θα μπορούσαν είτε να αποτελέσουν ενδιαφέρουσα πηγή γνώσης και ένδειξη πληροφορίας, είτε λάθη που επιβάλουν περαιτέρω ανάλυση.

- **Association rule learning (Dependency modelling)**

Η αναζήτηση συσχετίσεων μεταξύ μεταβλητών, π.χ. των προϊόντων ενός εμπορικού καταστήματος και των τιμών τους, τα προϊόντα που αγοράζονται συχνά μαζί, οι αγορές σε σχέση με το χρονικό διάστημα στο οποίο πραγματοποιούνται κ.α.

- **Clustering**

Η διαδικασία ανακάλυψης ομάδων και δομών σε δεδομένα που είναι κατά κάποιο τρόπο όμοια, χωρίς να υπάρχει κάποια γνώση για τις δομές τους και τις κατηγορίες στις οποίες εμπίπτουν.

- **Classification**

Η διαδικασία γενίκευσης μιας γνωστής δομής και η εφαρμογή της σε νέα δεδομένα.

- **Regression**

Η προσπάθεια ανακάλυψης μιας εξίσωσης που μοντελοποιεί τα δεδομένα με το ελάχιστο λάθος.

- **Summarization**

Η προσπάθεια απόδοσης του νοήματος ή της παρουσίασης μιας πιο συμπυκνωμένης μορφής ενός συνόλου δεδομένων, συμπεριλαμβανομένου της οπτικοποίησης και της δημιουργίας αναφορών

1.3.1 Σημασία της εξαγωγής πληροφοριών

Ο ρόλος της Εξαγωγής πληροφοριών, στα πλαίσια της Ανάκτησης πληροφοριών και Διαχείρισης γνώσης, είναι η αναγνώριση εξειδικευμένης πληροφορίας και η εξαγωγή γνώσης από μη δομημένα δεδομένα με μηχανικό (αυτόματο) τρόπο. Αντίθετα με την κλασσική Ανάκτηση πληροφοριών, σύμφωνα με την οποία η αναζήτηση γίνεται με βάση συγκεκριμένες λέξεις-κλειδιά και το αποτέλεσμα περιλαμβάνει μόνο κείμενα στα οποία βρίσκεται (ενδεχομένως) η χρήσιμη πληροφορία, η Εξόρυξη πληροφοριών στοχεύει ακριβώς στην αναγνώριση

της χρήσιμης μόνο πληροφορίας και το περιβάλλον (context) στο οποίο αυτή εμφανίζεται.

Δεδομένου του μεγάλου όγκου πληροφοριών που παράγονται και διακινούνται σήμερα (κύριο χαρακτηριστικό του διαδικτύου) το ζητούμενο στις μέρες μας είναι όχι απλώς η κατοχή της πληροφορίας – ο οποιοσδήποτε σήμερα μπορεί να έχει πρόσβαση σε σχετικά οποιαδήποτε πληροφορία – αλλά η διαχείριση της πληροφορίας και ο εντοπισμός της «σχετικής» πληροφορίας. Έτσι, ενώ με μια κλασσική μηχανή αναζήτησης ο ενδιαφερόμενος θα λάβει ως απάντηση ένα σύνολο κειμένων που ενδεχομένως περιέχουν την απάντηση που περιμένει, η Εξόρυξη πληροφοριών στοχεύει στην απάντηση και μόνο σε αυτή.

1.3.2 Αποτελέσματα και επιδόσεις της εξόρυξης κειμένου

Οι επιδόσεις αναλόγων συστημάτων που έχουν παρουσιαστεί κατά καιρούς σε συνέδρια και ερευνητικές εργασίες αναφέρονται σε έρευνα πάνω σε συγκεκριμένους τομείς (κείμενα γραμμένα για συγκεκριμένους θεματικούς χώρους) και διαφέρουν ανάλογα με το είδος της πληροφορίας που στοχεύουν να αναγνωρίσουν. Έτσι, αναφορικά με της πέντε προαναφερθείσες κατηγορίες της Εξαγωγής πληροφοριών, έχουν παρατηρηθεί τα κάτωθι:

Όσο προχωράμε από την αναγνώριση κυρίων ονομάτων προς την αναγνώριση σχέσεων μεταξύ γεγονότων η δυσκολία αυξάνεται εκθετικά, αφού εξάλλου τα μετέπειτα στάδια ανάλυσης εξαρτώνται σε μεγάλο βαθμό στα προγενέστερα. Έχοντας ως δεδομένο το γεγονός αυτό, η αναγνώριση κυρίων ονομάτων επιτυγχάνεται κατά 93% περίπου και λαμβάνοντας υπ’ όψιν το γεγονός ότι ένας άνθρωπος δε μπορεί να επιτύχει το 100%, μπορούμε να ισχυριστούμε πως τα συστήματα αναγνώρισης ονομάτων τα καταφέρνουν το ίδιο ή σχεδόν το ίδιο καλά με τον άνθρωπο.

Όμως, τα επόμενα στάδια ανάλυσης δεν εφαρμόζονται με την ίδια επιτυχία. Τα αποτελέσματα που δημοσιεύονται είναι ενδεικτικά περισσότερες πληροφορίες εδώ. Τελικά, η αναγνώριση γεγονότων (που είναι ο κύριος στόχος της εξαγωγής πληροφοριών) δεν εφαρμόζεται με μεγάλη επιτυχία ούτε αυτή: τα καλύτερα συστήματα αναφέρουν επιδόσεις της τάξεως του 60%, σύμφωνα με τα πρακτικά των συνεδρίων MUC. Τα νούμερα που αναφέρονται εδώ δηλώνουν κατά προσέγγιση την επίδοση των συστημάτων και είναι το F-measure των αποτελεσμάτων.

Στο σημείο αυτό πρέπει να σημειώσουμε δύο πράγματα. Πρώτον, τα αναφερόμενα μεγέθη αναφέρονται σε πειράματα πάνω σε συγκεκριμένους τομείς (domains) και συγκεκριμένες γλώσσες (Αγγλικά συνήθως). Το γεγονός αυτό είναι ιδιαίτερα σημαντικό, αφού κάθε τομέας και κάθε γλώσσα έχουν ιδιαιτερότητες οι οποίες πρέπει να ενσωματωθούν στα συστήματα εξόρυξης πληροφοριών. Από τη φύση της λοιπόν η Εξαγωγή πληροφοριών είναι άρρηκτα συνδεδεμένη με τη γλώσσα και το είδος των δεδομένων (κειμένων) που πρόκειται να αναλυθούν.

Δεύτερον, η σχετική έρευνα έχει γίνει πιο εντατική τα τελευταία χρόνια. Συνεχώς προτείνονται νέες μέθοδοι κατασκευής αναλόγων αρχιτεκτονικών και συστημάτων οι οποίες στοχεύουν σε καλύτερες επιδόσεις και στην απλοποίηση της επέκτασης της λειτουργίας υπαρχόντων συστημάτων σε νέους τομείς και νέες γλώσσες. Έμφαση δίνεται κυρίως στην ανάπτυξη ειδικών συστημάτων μέσω αυτόματης εκμάθησης της γλωσσικής πληροφορίας και των ιδιαιτεροτήτων του χώρου από το ίδιο το σύστημα. Στο μέλλον, λοιπόν, είναι βέβαιο ότι θα έχουμε καλύτερα συστήματα και σαφώς καλύτερες επιδόσεις.

1.4 Δυσκολίες και Προκλήσεις

Πολλές είναι οι δυσκολίες οι οποίες συμβαίνουν κατά τη διάρκεια της διαδικασίας εξόρυξης κειμένου και της επίδρασης στην αποτελεσματικότητα της λήψης αποφάσεων. Καθορίζονται διάφοροι κανόνες και κανονισμοί στην επεξεργασία του κειμένου έτσι ώστε να καθιστά την διαδικασία της εξόρυξης κειμένου αποτελεσματική. Πριν από την εφαρμογή της ανάλυσης μοτίβων στο έγγραφο υπάρχει μια ανάγκη μετατροπής μη δομημένων δεδομένων σε ενδιάμεση μορφή αλλά σε αυτό το στάδιο η διαδικασία εξόρυξης έχει τις δικές της επιπλοκές. Ένα σημαντικό ζήτημα είναι μια πολύγλωσση εξάρτηση από το κείμενο που δημιουργεί προβλήματα. Υπάρχουν μόνο λίγα εργαλεία που υποστηρίζουν πολλαπλές γλώσσες. Χρησιμοποιούνται διάφοροι αλγόριθμοι και τεχνικές ανεξάρτητα για την υποστήριξη πολυγλωσσικού κειμένου. Πολλά σημαντικά έγγραφα παραμένουν εκτός της διαδικασίας εξόρυξης κειμένου γιατί τα διάφορα

εργαλεία δεν τα υποστηρίζουν. Αυτά τα θέματα δημιουργούν πολλά προβλήματα στην ανακάλυψη της γνώσης και στη λήψη αποφάσεων επεξεργασίας.

Σύμφωνα με τις απαιτήσεις του κάθε τομέα, χρειάζονται εμπειρογνώμονες για να συνεργαστούν έτσι ώστε να εξάγουν πιο αποτελεσματικά και ακριβή αποτελέσματα. Η χρήση συνωνύμων και αντωνύμων στα έγγραφα δημιουργούν προβλήματα στα εργαλεία για την εξόρυξη κειμένου που λαμβάνουν και τα δύο στο ίδιο πλαίσιο. Είναι δύσκολο να κατηγοριοποιηθούν τα έγγραφα όταν η συλλογή εγγράφων είναι μεγάλη και δημιουργούνται από διάφορες περιοχές που έχουν παρόμοιο περιεχόμενο. Οι συντομογραφίες δίνουν διαφορετική σημασία σε διαφορετικές περιπτώσεις και είναι επίσης ένα μεγάλο ζήτημα. Οι γραμματικοί κανόνες σύμφωνα παραμένουν ένα ανοικτό ζήτημα στον τομέα της εξόρυξης κειμένων.

Συνεχίζοντας αναγνωρίζοντας ένα συγκεκριμένο σύνολο λέξεων-κλειδιών (keywords) θα μπορούσαμε να προσδιορίσουμε τη συνολική πολικότητα της άποψης που εκφράζεται στο κείμενο, καθώς η πολικότητα ενός κειμένου προκύπτει από την πολικότητα των μεμονωμένων λέξεων από τις οποίες απαρτίζεται. Η προσέγγιση μέσω λέξεων κλειδιών στην ανίχνευση συναισθήματος δεν εμφανίζει υψηλά ποσοστά ακρίβειας και έχει αποδειχθεί ελλιπής σε ορισμένες περιπτώσεις. Επιπλέον, εκτός από τον προσδιορισμό της πολικότητας, όταν απουσιάζουν συναισθηματικά φορτισμένες λέξεις, ιδιαίτερα απαιτητικός είναι και ο διαχωρισμός των υποκειμενικών και αντικειμενικών λέξεων και φράσεων ενός κειμένου. Η γενικότερη αντίληψη της θετικής ή αρνητικής άποψης δεν εξαρτάται άμεσα από το εκάστοτε θέμα συζήτησης. Ωστόσο, το συναίσθημα και η υποκειμενικότητα ενός κειμένου εξαρτώνται από το σημασιολογικό πλαίσιο στο οποίο τοποθετούνται . Άλλος ένας παράγοντας που επηρεάζει την πολικότητα είναι η σειρά των λέξεων και φράσεων στο κείμενο. Οι ίδιες λέξεις με διαφορετική σειρά μπορεί να οδηγήσουν σε τελείως διαφορετική συνολική πολικότητα. Τέλος, στις δυσκολίες που συναντά η Αναγνώριση Συναισθηματικής Κατάστασης πρέπει να συμπεριληφθούν και οι προκλήσεις της ευρύτερης περιοχής της Επεξεργασίας Φυσικής Γλώσσας όπως ο χειρισμός της άρνησης, η ειρωνεία και ο σαρκασμός.

1.5 Στόχοι της εργασίας

Το κύριο στάδιο της εργασίας είναι η ανάλυση και εξεύρεση της συναισθηματικής κατάστασης ενός ενεργού χρήστη των μέσων κοινωνικής δικτύωσης. Σύμφωνα με το διαγωνισμό για την αναγνώριση της συναισθηματικής κατάστασης των ανθρώπων οι οποίοι είναι χρήστες του twitter είναι η δημιουργία μοντέλου μηχανικής μάθησης χρησιμοποιώντας τα κείμενα και τα άρθρα που δημοσιεύουν.

2

Επιβλεπόμενη μηχανική μάθηση

2.1 Εισαγωγή στην επιβλεπόμενη μηχανική μάθηση

Η εποπτευόμενη μάθηση είναι το πιο συνηθισμένο υποτμήμα της μηχανικής μάθησης σήμερα. Συνήθως, οι νέοι επαγγελματίες της εκμάθησης μηχανών θα ξεκινήσουν το ταξίδι τους με αλγόριθμους εποπτευόμενης μάθησης. Επομένως, η πρώτη από αυτές τις τρεις σειρές θα αφορά την επιβλεπόμενη μάθηση.

Οι επιβλεπόμενοι αλγόριθμοι μηχανικής μάθησης σχεδιάζονται για να μάθουν με παράδειγμα. Το όνομα " επιβλεπόμενη" μάθηση προέρχεται από την ιδέα ότι η κατάρτιση αυτού του τύπου του αλγορίθμου είναι σαν να έχεις έναν εκπαιδευτικό να εποπτεύει όλη τη διαδικασία.

Κατά την κατάρτιση ενός επιβλεπόμενου αλγορίθμου μάθησης, τα δεδομένα εκπαίδευσης θα αποτελούνται από εισόδους που αντιστοιχίζονται με τις σωστές εξόδους. Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος θα αναζητήσει πρότυπα στα δεδομένα που σχετίζονται με τις επιθυμητές εξόδους. Μετά την εκπαίδευση, ένας αλγόριθμος επιβλεπόμενης μάθησης θα λάβει νέες αόρατες εισροές και θα

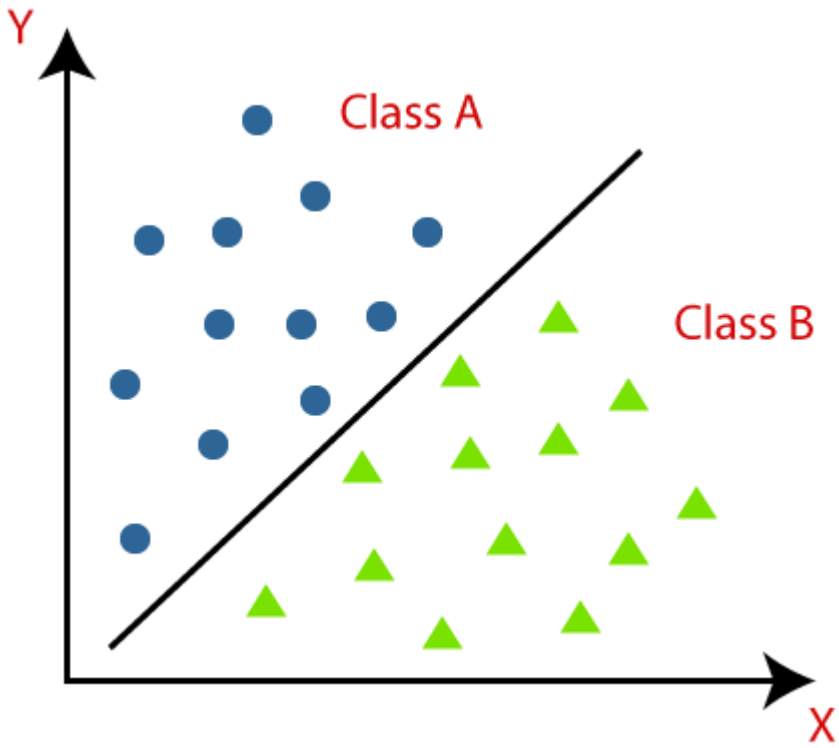
καθορίσει ποια ετικέτα θα ταξινομηθούν οι νέες εισροές με βάση προηγούμενα δεδομένα εκπαίδευσης. Ο στόχος ενός μοντέλου επιβλεπόμενης μάθησης είναι η πρόβλεψη της σωστής ετικέτας για τα δεδομένα εισαγωγής που παρουσιάστηκαν πρόσφατα. Στην πιο βασική του μορφή, ένας αλγόριθμος επιβλεπόμενης μάθησης μπορεί να γραφτεί απλά ως εξής:

$$Y = f(x)$$

Όπου Y είναι η προβλεπόμενη έξοδος που καθορίζεται από μια συνάρτηση χαρτογράφησης που εκχωρεί μια κλάση σε μια τιμή εισόδου x. Η λειτουργία που χρησιμοποιείται για τη σύνδεση των χαρακτηριστικών εισόδου με την προβλεπόμενη έξοδο δημιουργείται από το μοντέλο μηχανικής μάθησης κατά τη διάρκεια της εκπαίδευσης.

Η επιβλεπόμενη μάθηση μπορεί να χωριστεί σε δύο υποκατηγορίες: Ταξινόμηση (Classification) και παλινδρόμηση (Regression).

2.1.1 Classification



Εικόνα 1 κατηγοριοποίηση δύο μεταβλητών

[2] **https://www.javatpoint.com/classification-algorithm-in-machine-learning**

Κατά τη διάρκεια της κατάρτισης, στους αλγόριθμους ταξινόμησης δίνονται στοιχεία δεδομένων με μια καθορισμένη κατηγορία. Η δουλειά ενός αλγορίθμου ταξινόμησης είναι να πάρει μια τιμή εισόδου και να την αναθέσει σε μια τάξη ή κατηγορία, την οποία χωρίζει με βάση τα δεδομένα εκπαίδευσης που παρέχονται.

Το πιο συνηθισμένο παράδειγμα ταξινόμησης είναι ο προσδιορισμός εάν ένα μήνυμα ηλεκτρονικού ταχυδρομείου είναι spam ή όχι. Με δύο κατηγορίες για να διαλέξετε (spam, ή όχι spam), το πρόβλημα αυτό ονομάζεται πρόβλημα δυαδικής ταξινόμησης. Στον αλγόριθμο θα δοθούν δεδομένα κατάρτισης με μηνύματα ηλεκτρονικού ταχυδρομείου τα οποία είναι και τα spam και όχι spam. Το μοντέλο θα βρει τις δυνατότητες μέσα στα δεδομένα που σχετίζονται με κάθε κατηγορία και θα δημιουργήσει τη συνάρτηση χαρτογράφησης που αναφέρθηκε νωρίτερα: Y = f (x). Στη συνέχεια, όταν παρέχεται με ένα αόρατο μήνυμα ηλεκτρονικού ταχυδρομείου, το μοντέλο θα χρησιμοποιήσει αυτή τη λειτουργία για να καθορίσει εάν το μήνυμα ηλεκτρονικού ταχυδρομείου είναι spam ή όχι.

Τα προβλήματα ταξινόμησης μπορούν να λυθούν με πολυάριθμους αλγόριθμους. Όποιος αλγόριθμος επιλέγετε να χρησιμοποιήσετε εξαρτάται από τα δεδομένα και την κατάσταση. Ακολουθούν ορισμένοι δημοφιλείς αλγόριθμοι ταξινόμησης:

- Linear Classifiers
- Support Vector Machines
- Decision Trees
- K-Nearest Neighbor
- Random Forest

2.1.2 Regression

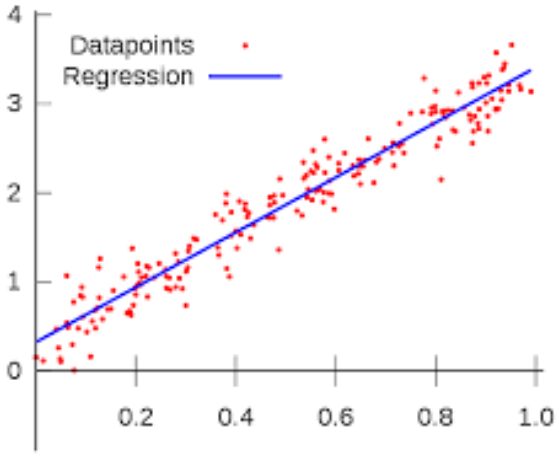
Η παλινδρόμηση (Regression) είναι μια προγνωστική στατιστική διαδικασία όπου το μοντέλο επιχειρεί να βρει τη σημαντική σχέση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Ο στόχος ενός αλγόριθμου παλινδρόμησης είναι να προβλέψουμε έναν συνεχή αριθμό όπως οι πωλήσεις, το εισόδημα και τα αποτελέσματα δοκιμών. Η εξίσωση για τη βασική γραμμική παλινδρόμηση μπορεί να γραφεί ως εξής:

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + \cdots + w[i] * x[i] + b$$

Όπου x [i] είναι το χαρακτηριστικό για τα δεδομένα και όπου w [i] και b είναι παράμετροι που αναπτύσσονται κατά τη διάρκεια της εκπαίδευσης. Για απλά μοντέλα γραμμικής παλινδρόμησης με μόνο ένα χαρακτηριστικό στα δεδομένα, ο τύπος μοιάζει με αυτόν:

$$\hat{y} = wx + b$$

Όπου w είναι η κλίση, το x είναι το μοναδικό χαρακτηριστικό και το b είναι το y-intercept. Για απλά προβλήματα παλινδρόμησης όπως αυτό, οι προβλέψεις μοντέλων αντιπροσωπεύονται από τη γραμμή της καλύτερης προσαρμογής.



Εικόνα 2 αλγοριθμος παλινδρόμησης

Υπάρχει μια σαφής θετική συσχέτιση μεταξύ των ωρών που μελετήθηκαν (ανεξάρτητη μεταβλητή) και του τελικού σκορ του μαθητή (εξαρτημένη μεταβλητή). Μια γραμμή καλύτερης προσαρμογής μπορεί να εξαχθεί μέσω των σημείων δεδομένων για να δείξει τις προβλέψεις των μοντέλων όταν δοθεί μια νέα είσοδος. Ας υποθέσουμε ότι θέλαμε να μάθουμε πόσο καλά θα περάσει ένας σπουδαστής με πέντε ώρες σπουδών. Μπορούμε να χρησιμοποιήσουμε τη γραμμή της καλύτερης προσαρμογής για να προβλέψουμε το αποτέλεσμα της δοκιμής με βάση τις επιδόσεις άλλων σπουδαστών.

[3] **https://medium.com/simple-ai/linear-regression-intro-to-machine-learning-6-6e320dbdaf06**

Υπάρχουν πολλοί διαφορετικοί τύποι αλγόριθμων παλινδρόμησης. Οι τρεις πιο συνηθισμένες παρατίθενται παρακάτω:

- Γραμμική παλινδρόμηση
- Λογιστική παλινδρόμηση
- Πολυωνυμική παλινδρόμηση

2.2 Αλγόριθμοι

2.2.1 Random Forest

Ο Random Forest είναι ένας αλγόριθμος εποπτευόμενης μάθησης που χρησιμοποιείται τόσο για ταξινόμηση όσο και για παλινδρόμηση. Ωστόσο, χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης. Όπως γνωρίζουμε ότι ένα δάσος αποτελείται από δέντρα και περισσότερα δέντρα σημαίνει πιο ισχυρό δάσος. Ομοίως, ο τυχαίος δασικός αλγόριθμος δημιουργεί δέντρα αποφάσεων σε δείγματα δεδομένων και στη συνέχεια παίρνει την πρόβλεψη από καθένα από αυτά και τελικά επιλέγει την καλύτερη λύση μέσω ψηφοφορίας. Είναι μια μέθοδος σύνολο που είναι καλύτερο από ένα ενιαίο δέντρο απόφασης επειδή μειώνει την υπερβολική τοποθέτηση με τον μέσο όρο του αποτελέσματος.

Μπορούμε να κατανοήσουμε τη λειτουργία του αλγόριθμου Random Forest με τη βοήθεια των παρακάτω βημάτων

- **Βήμα 1** - Αρχικά, ξεκινήστε με την επιλογή τυχαίων δειγμάτων από ένα δεδομένο σύνολο δεδομένων.
- **Βήμα 2** - Στη συνέχεια, αυτός ο αλγόριθμος θα κατασκευάσει ένα δέντρο απόφασης για κάθε δείγμα. Τότε θα πάρει το αποτέλεσμα πρόβλεψης από κάθε δέντρο αποφάσεων.
- **Βήμα 3** - Σε αυτό το βήμα, η ψηφοφορία θα πραγματοποιηθεί για κάθε προβλεπόμενο αποτέλεσμα.

- **Βήμα 4** - Επιτέλους, επιλέξτε το πιο ψηλό αποτέλεσμα πρόβλεψης ως το τελικό αποτέλεσμα πρόβλεψης.

Πλεονεκτήματα και Μειονεκτήματα

Πλεονεκτήματα

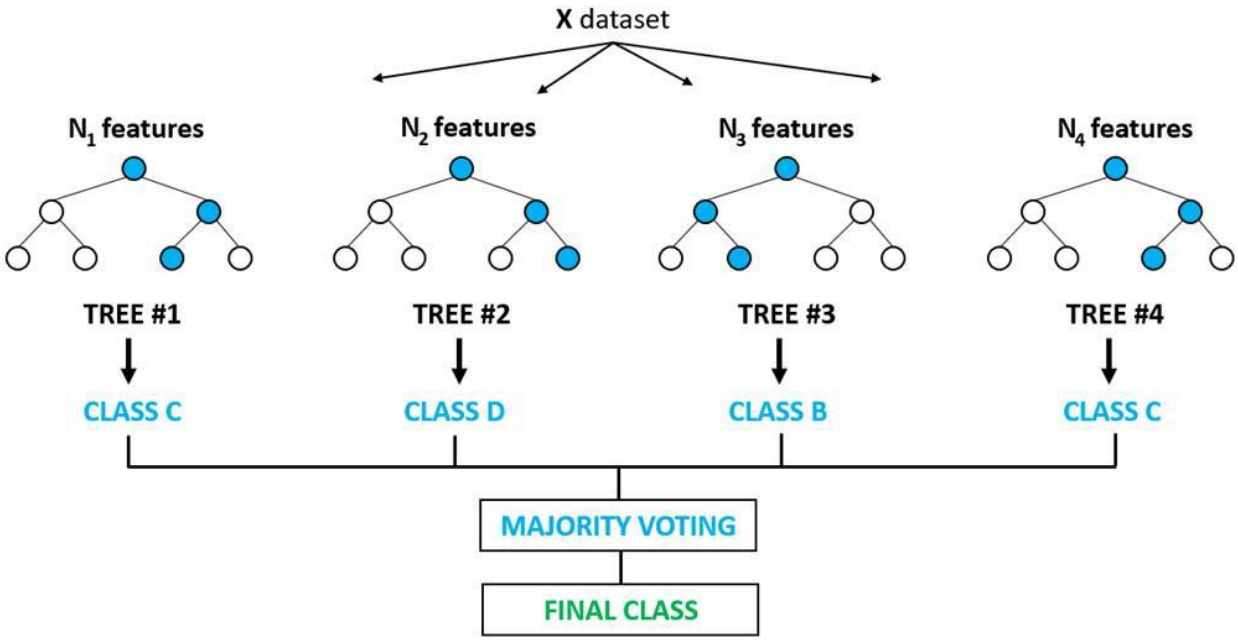
Τα παρακάτω είναι τα πλεονεκτήματα του αλγόριθμου Random Forest -

- Αντιμετωπίζει το πρόβλημα της υπερφόρτωσης με τον μέσο όρο ή το συνδυασμό των αποτελεσμάτων των διαφορετικών δέντρων αποφάσεων.
- Τυχαία δάση λειτουργούν καλά για ένα ευρύ φάσμα στοιχείων δεδομένων από ένα ενιαίο δέντρο αποφάσεων κάνει.
- Το τυχαίο δάσος έχει μικρότερη διακύμανση, στη συνέχεια δέντρο απόφασης.
- Τα τυχαία δάση είναι πολύ ευέλικτα και έχουν πολύ υψηλή ακρίβεια.
- Η κλιμάκωση των δεδομένων δεν απαιτεί τον τυχαίο αλγόριθμο δασών. Διατηρεί καλή ακρίβεια ακόμη και μετά την παροχή δεδομένων χωρίς κλιμάκωση.
- Οι τυχαίοι δασικοί αλγόριθμοι διατηρούν καλή ακρίβεια, ακόμη και ένα μεγάλο μέρος των δεδομένων λείπει.

Μειονεκτήματα

Τα παρακάτω είναι τα μειονεκτήματα του αλγόριθμου Random Forest -

- Η πολυπλοκότητα είναι το κύριο μειονέκτημα των τυχαίων δασικών αλγορίθμων.
- Η κατασκευή τυχαίων δασών είναι πολύ πιο δύσκολη και χρονοβόρα από τα δέντρα αποφάσεων.
- Απαιτούνται περισσότεροι υπολογιστικοί πόροι για την εφαρμογή του αλγόριθμου Random Forest.
- Είναι λιγότερο διαισθητικό σε περίπτωση που έχουμε μια μεγάλη συλλογή από δέντρα αποφάσεων.
- Η διαδικασία πρόβλεψης χρησιμοποιώντας τυχαία δάση είναι πολύ χρονοβόρα σε σύγκριση με άλλους αλγόριθμους.



Εικόνα 3 αλγόριθμος Random Forest [4]

2.2.2 *Logistic Regression*

Τύποι λογικής παλινδρόμησης

1.Binary Logistic Regression Δυο πιθανά αποτελέσματα. Παράδειγμα: θυμωμένος ή όχι

2.Multinomial Logistic Regression (Emotion Classification) Τρεις ή περισσότερες κατηγορίες χωρίς παραγγελία.

3.Ordinal Logistic Regression Τρεις ή περισσότερες κατηγορίες.

Θετικά

[4] https://www.globalsoftwaresupport.com/random-forest-classifier/

- Βολικές βαθμολογίες πιθανότητας για παρατηρήσεις
- Αποτελεσματικές υλοποιήσεις διαθέσιμες σε όλα τα εργαλεία
- Το Multicollinearity[5] δεν είναι πραγματικά ένα ζήτημα και μπορεί να αντιμετωπιστεί με τη ρύθμιση L2 σε κάποιο βαθμό
- Ευρεία ευελιξία στη βιομηχανία για λογικές λύσεις παλινδρόμησης

Αρνητικά

- Δεν λειτουργεί καλά όταν το μέγεθος του χώρου είναι πολύ μεγάλο
- Δεν χειρίζεται καλά μεγάλο αριθμό χαρακτηριστικών κατηγοριών / μεταβλητών
- Αντιστοιχεί σε μετασχηματισμούς για μη γραμμικά χαρακτηριστικά
- Αναφέρεται σε ολόκληρα δεδομένα

[5] https://www.tandfonline.com/doi/abs/10.1080/09720502.2010.10700699?journalCode=tjim20

3

Αναγνώριση συναισθηματικής κατάστασης στο SemEval-2018

3.1 Παρουσίαση διαγωνισμού SemEval-2018

Παρουσιάζουμε το διαγωνισμό SemEval-2018 Affects in Tweets, το οποίο περιλαμβάνει μια σειρά από subtasks από την επηρεαστική κατάσταση ενός ατόμου από το κείμενο, άρθρο (tweet) τους. Για κάθε εργασία δημιουργήσαμε ετικέτα δεδομένα από αγγλικά, αραβικά και ισπανικά tweets.

Τα μεμονωμένα καθήκοντα είναι:

1. η ένταση των συναισθημάτων, παλινδρόμηση

2. κανονική ταξινόμηση έντασης συναισθημάτων

3. υποχώρηση σθένους (αίσθησης)

4. η συσχέτιση κατάταξης σθένους

5. η κατηγοριοποίηση συναισθήματος

Εβδομήντα πέντε ομάδες (περίπου 200 μέλη της ομάδας) συμμετείχαν στην κοινή εργασία.

Σύμφωνα με το διαγωνισμό τα συναισθήματα είναι βασικά για τη γλώσσα και τη σκέψη. Είναι εξοικειωμένοι και συνηθισμένοι, όμως είναι πολύπλοκη και απαλή.

Οι άνθρωποι είναι γνωστό ότι αντιλαμβάνονται εκατοντάδες διαφορετικά συναισθήματα. Σύμφωνα με στο βασικό μοντέλο συναισθημάτων (γνωστό και ως το κατηγορηματικό μοντέλο), μερικά συναισθήματα, όπως η χαρά, η θλίψη και ο φόβος είναι τα πιο βασικά. Κάθε ένα από αυτά τα συναισθήματα μπορεί να γίνει αισθητό σε διαφορετικές εντάσεις. Για παράδειγμα, οι δηλώσεις μας μπορεί να μεταφέρουν ότι είμαστε πολύ θυμωμένοι, ελαφρώς λυπημένοι κλπ. Εδώ, ένταση αναφέρεται στο βαθμό ή την ποσότητα ενός συναισθήματος όπως ο θυμός ή η θλίψη. Τα συναισθήματα είναι σημεία σε ένα τρισδιάστατο χώρο του σθένους (θετικότητα-αρνητικότητα), διέγερση (ενεργητική-παθητική) και κυριαρχία (κυρίαρχος-υποτακτικός). Χρησιμοποιούμε τον όρο επηρεάζουν για να αναφερθούμε σε διάφορες κατηγορίες που σχετίζονται με τις συγκινήσεις όπως η χαρά, ο φόβος, το σθένος και η διέγερση.

Σύμφωνα με το διαγωνισμό οι εφαρμογές φυσικής γλώσσας στο εμπόριο, τη δημόσια υγεία, τη διαχείριση καταστροφών και το κοινό πολιτική μπορούν να επωφεληθούν από τη γνώση του επηρεασμού καταστάσεις ανθρώπων - και οι δύο κατηγορίες και το τις εντάσεις των συναισθημάτων που αισθάνονται. Ο διαγωνισμός περιλαμβάνει μια σειρά από subtasks όπου τα αυτόματα συστήματα πρέπει να συμπεράνουν την επηρεασμένη κατάσταση ενός ατόμου από το tweet τους.

Οι υποκατηγορίες με τις οποίες ασχολήθηκε ο διαγωνισμός είναι οι εξής:

1. ***Emotion Intensity Regression*** (EI-reg): Δίνεται ένα άρθρο (tweet) και ένα συναίσθημα, έτσι καθορίζεται η ένταση του συναισθήματος που αντιπροσωπεύει καλύτερα την ψυχική κατάσταση του συγγραφέα ενός tweet. Με άλλα λόγια οι τιμές που βρίσκονται στα δεδομένα μας κυμαίνονται με ένα πραγματικό σκορ μεταξύ 0 και 1.
2. ***Emotion Intensity Ordinal Classification*** (EI-oc): Με δεδομένο ένα tweet και ένα συναίσθημα, ταξινομείται το άρθρο σε μία από τις τέσσερις σειρές τάξεων της έντασης του συναισθήματος που αντιπροσωπεύει καλύτερα την ψυχική την κατάσταση του άρθρου.
3. ***Valence (Sentiment) Regression*** (V-reg): Δίνεται ένα άρθρο, το οποίο καθορίζει την ένταση του συναισθήματος ή το σθένος (V) που αντιπροσωπεύει καλύτερα την ψυχική κατάσταση του συγγραφέα του άρθρου. Στα δεδομένα οι πραγματική βαθμολογία είναι μεταξύ 0 (πιο αρνητικό) και 1 (πιο θετικό).
4. ***Valence Ordinal Classification*** (V-oc): Δίνεται ένα άρθρο, το οποίο ταξινομείται σε μια από επτά κανονικές τάξεις, που αντιστοιχούν σε διάφορα επίπεδα θετικής και αρνητικής έντασης συναισθήματος, ότι

αντιπροσωπεύει καλύτερα την ψυχική κατάσταση του συγγραφέα του άρθρου.

5. ***Emotion Classification*** (E-c): Με δεδομένο ένα άρθρο, ταξινομείται ως «ουδέτερο ή χωρίς συναισθήματα» ή ως ένα, ή περισσότερα, από ένδεκα συναισθήματα που αντιπροσωπεύουν την ψυχική κατάσταση του συγγραφέα του άρθρου.

3.1.1 Δημιουργία δεδομένων διαγωνισμού (Dataset)

Σε αυτό το υποκεφάλαιο θα αναφερθούμε στο πως δημιουργήθηκαν τα δεδομένα που χρησιμοποιήθηκαν στο διαγωνισμό αλλά και από την μεριά μας.

Τα σύνολα δεδομένων EI-reg αντιστοιχούν σε θυμό, φόβο, χαρά, και θλίψη. Τα σύνολα δεδομένων EI-oc περιλαμβάνουν τα ίδια tweets όπως στο EI-reg, δηλαδή το σύνολο δεδομένων θυμού, το EI-oc έχει τα ίδια tweets με το σύνολο δεδομένων θυμού EI-reg, το σύνολο δεδομένων φόβου EI-oc έχει τα ίδια tweets με αυτά του φόβου του EI-reg, και ούτω καθεξής. Ωστόσο, οι ετικέτες για τα tweets EI-oc είναι κανονικές τάξεις αντί για realvalued βαθμού έντασης. Το σύνολο δεδομένων V-reg περιλαμβάνει ένα υποσύνολο tweets από κάθε ένα από τα τέσσερα EI-reg συναισθηματικά σύνολα δεδομένων. Το σύνολο δεδομένων V-oc έχει το ίδιο tweets όπως στο σύνολο δεδομένων V-reg. Το σύνολο δεδομένων E-c περιλαμβάνει όλα τα tweets από τα τέσσερα σύνολα δεδομένων EI-reg. Ο συνολικός αριθμός περιπτώσεων στις περιπτώσεις E-c, EI-reg.

Για τη δημιουργία ενός συνόλου δεδομένων των tweets πλούσιων σε ένα συγκεκριμένο συναισθήματα, χρησιμοποιήθηκε η ακόλουθη μεθοδολογία. Για κάθε συναίσθημα X, επιλέξαμε 50 έως 100 όρους που συνδέονται με αυτό το συναίσθημα σε διαφορετικά επίπεδα έντασης. Για παράδειγμα, για το «θυμό» σύνολο δεδομένων, χρησιμοποιήσαμε τους όρους: θυμωμένος, τρελός, απογοητευμένος, ενοχλημένος, γεμάτος, ερεθισμένος, μιμημένος, μανία, ανταγωνισμός, και ούτω καθεξής. Περιλαμβάνονται οι όροι αναζήτησης που επιλέξαμε λέξεις συναίσθημα που αναφέρονται στον θησαυρό του Ρότζετ, πλησιέστερους γείτονες αυτών των λέξεων συναισθημάτων σε ένα χώρο που ενσωματώνει λέξη, καθώς συνήθως χρησιμοποιούνται emoji και emoticons.

Παρακολουθώντας το API του Twitter, σε διάστημα δύο μηνών (Ιούνιος και Ιούλιος, 2017), για tweets που περιλαμβάνονται τους όρους των ερωτημάτων. Επιλέχτηκαν τυχαία 1.400 tweets από τη χαρά που έχουν οριστεί για σχολιασμό της έντασης της χαράς. Για τα τρία αρνητικά συναισθήματα, αρχικά επιλέχθηκαν τυχαία

200 tweets από το καθένα τις αντίστοιχες συλλογές tweet. Αυτά τα 600 tweets σχολιάστηκαν και για τα τρία αρνητικά συναισθήματα έτσι ώστε να μπορούμε να μελετήσουμε τις σχέσεις μεταξύ ο φόβος και ο θυμός, ανάμεσα στον θυμό και στη θλίψη, και ανάμεσα στη θλίψη και τον φόβο. Για καθένα από τα αρνητικά συναισθήματα, επιλέχθηκαν επίσης 800 επιπλέον tweets, από τα αντίστοιχα σετ tweet και σχολιάστηκαν μόνο για το αντίστοιχο συναίσθημα. Έτσι, ο αριθμός των tweets για καθένα από τα αρνητικά συναισθήματα ήταν επίσης 1.400 (τα 600 περιλαμβάνονται και στα τρία αρνητικά συναισθήματα + 800 μοναδικά στο συναίσθημα εστίασης). Για κάθε συγκίνηση, 100 tweets που είχαν ένα hashtag συναισθημάτων, emoticon, ή emoji επιλέχθηκαν τυχαία. Καταργήθηκαν οι τελευταίοι όροι των ερωτημάτων από αυτά τα tweets. Σαν αποτέλεσμα, το σύνολο δεδομένων περιελάμβανε επίσης μερικά tweets χωρίς να υπάρχουν σαφείς συναισθηματικοί-ενδεικτικοί όροι.

Έτσι, το σύνολο δεδομένων EI-reg περιλάμβανε 1.400 νέα tweets για κάθε ένα από τα τέσσερα συναισθήματα. Να σημειωθεί ότι το σύνολο δεδομένων EmoInt περιελάμβανε ήδη 1.500 έως 2.300 tweets ανά συναισθηματικό σχολιασμό για ένταση. Εκείνα τα tweets δεν επαναπροσδιορίστηκαν. Τα EmoInt EI-reg tweets καθώς και τα νέα tweets του EI-reg ήταν και τα δύο για συνηθισμένες τάξεις συναισθημάτων (EI-oc) . Τα νέα tweets της EI-reg σχημάτισαν την ανάπτυξη του EI-reg (dev).

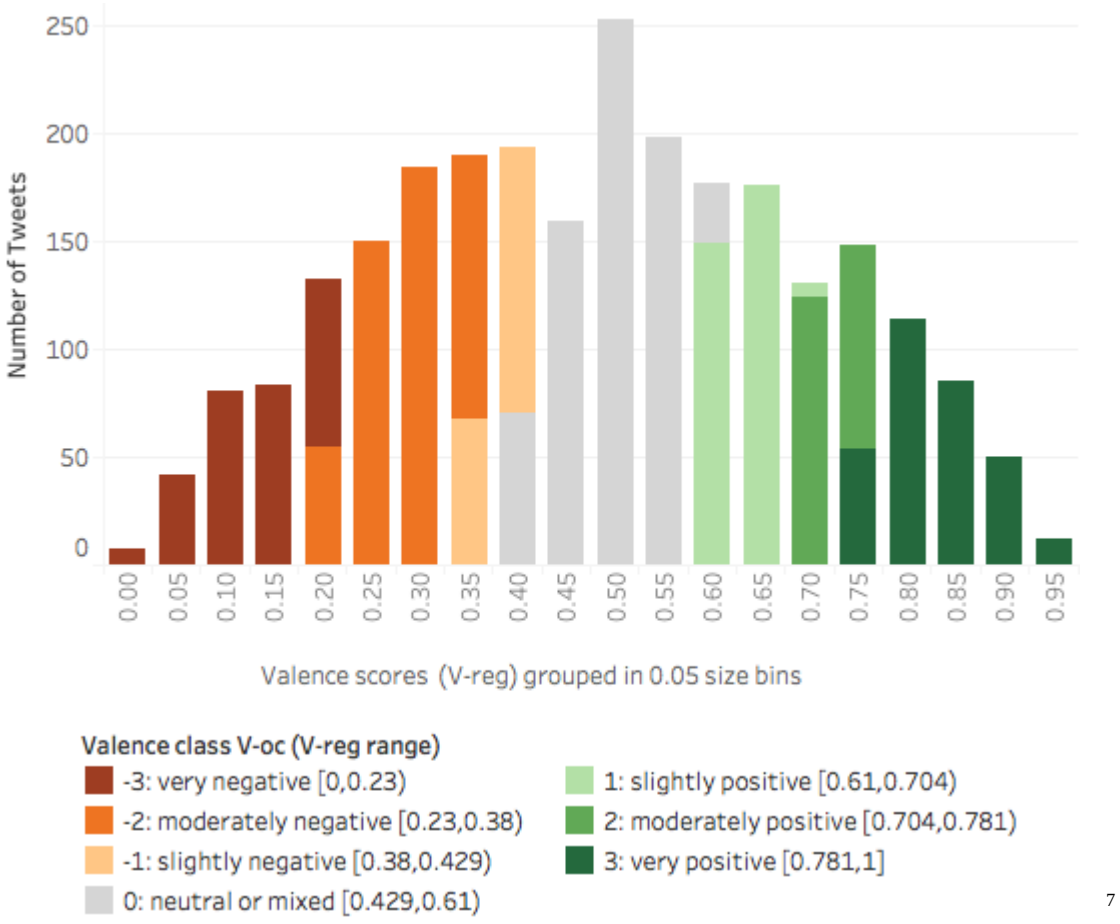
Το σύνολο δεδομένων σθένους (El-reg) περιελάμβανε tweets από το νέο το EI-reg set και το σύνολο EmoInt. Τα tweets που συμπεριλήφθηκαν ήταν και τα 600 κοινά για τα τρία αρνητικά σύνολα των συναισθημάτων και 600 τυχαία επιλεγμένα tweets χαράς. Τα EmoInt tweets που περιλαμβάνονται ήταν 600 τυχαία επιλεγμένα tweets χαράς και 200 από το καθένα, τυχαία επιλεγμένα, για θυμό, φόβο, και θλίψη. Για να μελετήθεί το σθένος σε σαρκαστικά tweets, συμπεριελήφθηκαν επίσης 200 tweets που είχαν hashtags #sarcastic, #sarcasm, #irony ή #ironic Έτσι, το V-reg αποτελείται από 2.600 tweets συνολικά. Το σύνολο V-oc αποτελείται από τα ίδια tweets όπως στο σύνολο V-reg.

Παρακάτω παρατηρούμαι το σύνολο των δεδομένων καθώς και κάποια ποσοστά του διαγωνισμού.

Dataset	Train	Dev	Test	Total
English				
E-c	6,838	886	3,259	10,983
EI-reg, EI-oc				
anger	1,701	388	1,002	3,091
fear	2,252	389	986	3,627
joy	1,616	290	1,105	3,011
sadness	1,533	397	975	2,905
V-reg, V-oc	1,181	449	937	2,567
Arabic				
E-c	2,278	585	1,518	4,381
EI-reg, EI-oc				
anger	877	150	373	1,400
fear	882	146	372	1,400
joy	728	224	448	1,400
sadness	889	141	370	1,400
V-reg, V-oc	932	138	730	1,800
Spanish				
E-c	3,561	679	2,854	7,094
EI-reg, EI-oc				
anger	1,166	193	627	1,986
fear	1,166	202	618	1,986
joy	1,058	202	730	1,990
sadness	1,154	196	641	1,991
V-reg, V-oc	1,566	229	648	2,443

[6]

Εικόνα 4 Πληροφορίες των Συνόλων δεδομένων



[7]

Εικόνα 5 διακριτές κλάσεις υποενοτήτων αναγνώρισης σθένους

[6] http://saifmohammad.com/WebDocs/semeval2018-task1.pdf

[7] http://saifmohammad.com/WebDocs/semeval2018-task1.pdf

Dataset	Scheme	Location	Item	#Items	#Annotators	MAI	#Q/Item	#Annotat.
English								
E-c	categorical	World	tweet	11,090	303	7	2	174,356
EI-reg								
anger	BWS	USA	4-tuple of tweets	2,780	168	4	2	27,046
fear	BWS	USA	4-tuple of tweets	2,750	220	4	2	26,908
joy	BWS	USA	4-tuple of tweets	2,790	132	4	2	26,676
sadness	BWS	USA	4-tuple of tweets	2,744	118	4	2	26,260
V-reg	BWS	USA	4-tuple of tweets	5,134	175	4	2	49,856
Total								**331,102**
Arabic								
E-c	categorical	World	tweet	4,400	175	7	1	36,274
EI-reg								
anger	BWS	World	4-tuple of tweets	2,800	221	4	2	25,960
fear	BWS	World	4-tuple of tweets	2,800	197	4	2	25,872
joy	BWS	World	4-tuple of tweets	2,800	133	4	2	24,690
sadness	BWS	World	4-tuple of tweets	2,800	177	4	2	25,834
V-reg	BWS	World	4-tuple of tweets	3,600	239	4	2	36,824
Total								**175,454**
Spanish								
E-c	categorical	World	tweet	7,182	160	7	1	56,274
EI-reg								
anger	BWS	World	4-tuple of tweets	3,972	157	3	2	27,456
fear	BWS	World	4-tuple of tweets	3,972	388	3	2	29,530
joy	BWS	World	4-tuple of tweets	3,980	323	3	2	28,300
sadness	BWS	World	4-tuple of tweets	3,982	443	3	2	28,462
V-reg	BWS	World	4-tuple of tweets	4,886	220	3	2	38,680
Total								**208,702**

Εικόνα 4 λεπτομερής ανάλυση συνόλου δεδομένων και μέθοδος εξόρυξης [8]

Ο παραπάνω του παραρτήματος συνοψίζει τις βασικές λεπτομέρειες του τρέχοντος συνόλου σχολιασμών που έγιναν για το SemEval-2018. Επηρεάζει το σετ δεδομένων Tweets (AIT). Το AIT χωρίστηκε σε εκπαίδευση, ανάπτυξη, και σύνολα δοκιμών για πειράματα μηχανικής μάθησης. Όλα τα αγγλικά προήλθαν από τα tweets-2016 ήταν μέρος της προπόνησης. Όλα τα αγγλικά tweets που προήλθαν από τα tweets-2017 χωρίστηκαν σε ανάπτυξη και δοκιμή σύνολα.

Ο Πίνακας 2 παρουσιάζει τη συμφωνία μεταξύ των διαμεσολαβητών και την συμφωνία Fleiss για τους σχολιασμούς συναισθημάτων πολλαπλών ετικετών. Η συμφωνία μεταξύ διαχειριστών (IRA) υπολογίζεται ως τοις εκατό, που κάθε ζεύγος σχολιαστών συμφωνεί. Για λόγους σύγκρισης, δείχνουμε επίσης τις βαθμολογίες που λαμβάνονται τυχαία επιλέγοντας εάν ιδιαίτερο συναίσθημα ισχύει ή όχι.

[8] http://saifmohammad.com/WebDocs/semeval2018-task1.pdf

Παρατηρούμαι ότι τα αποτελέσματα που προκύπτουν από τους πραγματικούς σχολιασμούς είναι σημαντικά υψηλότερες από τις βαθμολογίες που λαμβάνονται τυχαία.

3.1.2 Αξιολόγηση για αυτόματες προβλέψεις

3.1.2.1 Για EI-reg, EI-oc, V-reg και V-oc

Στον επίσημο διαγωνισμό για τα EI-reg, EI-oc, V-reg, και V-oc ένα σημαντικό κομμάτι ήταν η συσχέτιση Pearson Correlation με τις χρυσές αξιολογήσεις / ετικέτες (gold tags). Για το EI-reg και το EI-oc, τα αποτελέσματα συσχετίζονται μεταξύ τους καθώς και τα τέσσερα συναισθήματα ήταν κατά μέσο όρο για να καθορισθεί η μέτρηση του ανταγωνισμού. Εκτός από τον επίσημο διαγωνισμό μετρητών που περιγράφηκαν παραπάνω, μερικές επιπλέον μετρήσεις έγιναν για κάθε υποβολή. Αυτές είχαν σκοπό να παρέχουν μια διαφορετική προοπτική για τα αποτελέσματα. Η δευτερεύουσα μέτρηση που χρησιμοποιείται για το τα καθήκοντα παλινδρόμησης ήταν:

I. Pearson Correlation για ένα υποσύνολο του σετ δοκιμών που περιλαμβάνει μόνο τα tweets με ένταση βαθμολογία μεγαλύτερη ή ίση με 0,5.

Οι δευτερεύουσες μετρήσεις που χρησιμοποιούνται για τη συστηματική ταξινόμηση είναι:

II. Pearson Correlation για ένα υποσύνολο του σετ δοκιμών που περιλαμβάνει μόνο τα tweets με ένταση με τάξεις χαμηλές X, μέτρια X ή υψηλή X (όπου X είναι ένα συναίσθημα).

III. Σταθμισμένο τετραγωνικό kappa στο πλήρες σύνολο δοκιμών.

IV. Ζυγισμένο τετραγωνικό kappa σε ένα υποσύνολο συναισθημάτων του συνόλου δοκιμών.

3.1.2.2 Για το E-c

Η επίσημη μέτρηση που χρησιμοποιήθηκε για το E-c ήταν η ακρίβεια πολλών ετικετών (ή δείκτης Jaccard). Δηλαδή είναι ένα έργο ταξινόμησης πολλαπλών ετικετών που κάθε άρθρο μπορεί να έχει μία ή περισσότερες ετικέτες χιούμορ, και μία ή περισσότερες προβλεπόμενες ετικέτες συναισθημάτων. Πολλαπλή ετικέτα η ακρίβεια ορίζεται ως το μέγεθος της τομής των προβλεπόμενων και χρυσών ετικετών που διαιρούνται από το μέγεθος της ένωσης τους. Το μέτρο αυτό υπολογίζεται για κάθε άρθρο t, και στη συνέχεια υπολογίζεται κατά μέσο όρο για όλα τα tweets T στο σύνολο δεδομένων:

$$Accuracy = \frac{1}{|T|} * \sum_{t \in T} Gt \cap Pi / Gt \cup Pi$$

όπου Gt είναι το σύνολο των χρυσών ετικετών για ένα tweet t, Pt είναι το σύνολο των προβλεπόμενων ετικετών για τα tweet t και T είναι το σύνολο των tweets.

	#Teams				
ML algorithm	EI-reg	EI-oc	V-reg	V-oc	E-c
AdaBoost	1	1	3	1	0
Bi-LSTM	10	8	10	6	6
CNN	10	8	7	6	3
Gradient Boosting	8	3	5	4	1
Linear Regression	11	2	7	2	1
Logistic Regression	9	7	8	6	6
LSTM	13	9	10	5	4
Random Forest	8	7	5	6	6
RNN	0	0	0	0	1
SVM or SVR	15	9	8	6	6
Other	14	16	13	12	7

Εικόνα 5 αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήθηκαν [9]

[9] http://saifmohammad.com/WebDocs/semeval2018-task1.pdf

Features/Resources	EI-reg	EI-oc	V-reg	V-oc	E-c
affect-specific word embeddings	10	8	9	9	5
affect/sentiment lexicons	24	16	16	15	12
character ngrams	6	4	3	4	2
dependency/parse features	2	3	3	3	2
distant-supervision corpora	10	8	7	5	4
manually labeled corpora (other)	6	4	4	5	3
AIT-2018 train-dev (other task)	6	5	5	5	3
sentence embeddings	10	8	7	8	6
unlabeled corpora	6	3	5	3	0
word embeddings	32	21	25	21	20
word ngrams	19	14	12	10	9
Other	5	5	5	5	5

Εικόνα 6 μέθοδοι δημιουργίας και εξαγωγής χαρακτηριστών [10]

[10] http://saifmohammad.com/WebDocs/semeval2018-task1.pdf

3.2 Μέτρηση απόδοσης αλγορίθμων

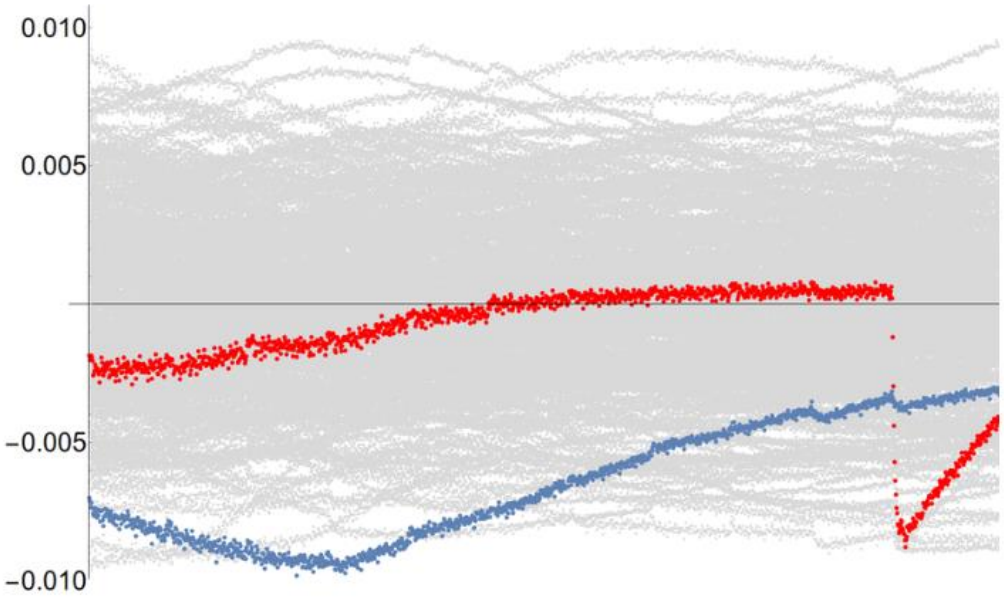
3.2.1 Συντελεστής Συσχέτισης Pearson (Pearson Correlation Coefficient)

Μια συσχέτιση Pearson είναι ένας αριθμός μεταξύ -1 και 1 που δείχνει την έκταση στην οποία δύο μεταβλητές είναι γραμμικές.

Ο συσχετισμός Pearson είναι επίσης γνωστός ως "συντελεστής συσχέτισης ροπής προϊόντος" (PMCC) ή απλά " συσχέτιση ".
Οι συσχετίσεις Pearson είναι κατάλληλες μόνο για μετρικές μεταβλητές (οι οποίες περιλαμβάνουν διχοτομικές μεταβλητές).

- Για τις κανονικές μεταβλητές , χρησιμοποιήστε τη συσχέτιση Spearman ή το tau και το Kendall's

Για τις ονομαστικές μεταβλητές , χρησιμοποιήστε το V Cramér .



Εικόνα 7 συσχέτιση μεταξύ τιμών δύο μεταβλητών [11]

3.2.2 Συντελεστής Συσχέτισης – Βασικά

Μερικά βασικά σημεία σχετικά με τους συντελεστές συσχέτισης παρουσιάζονται ωραία στο παρακάτω σχήμα.

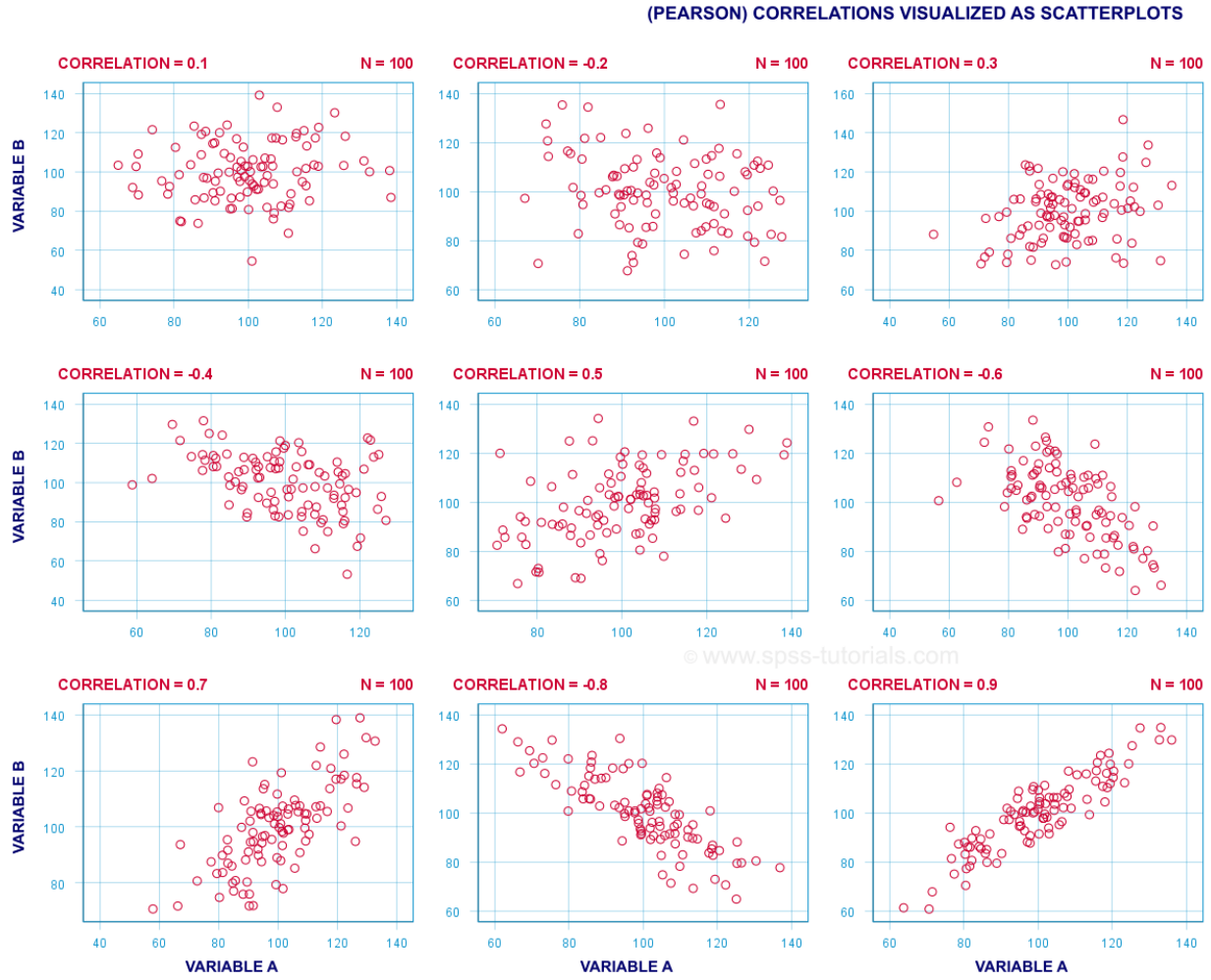
1. Οι συσχετίσεις δεν είναι ποτέ χαμηλότερες από -1. Μια συσχέτιση του -1 δείχνει ότι τα σημεία δεδομένων σε μια γραφική παράσταση σκέδασης βρίσκονται ακριβώς σε ευθεία κατηφορική γραμμή. οι δύο μεταβλητές είναι απολύτως αρνητικά γραμμικές.

2. Μια συσχέτιση του 0 σημαίνει ότι δύο μεταβλητές δεν έχουν καμιά γραμμική σχέση. Ωστόσο, μπορεί να υπάρχει κάποια μη γραμμική σχέση μεταξύ των δύο μεταβλητών.

3. Οι συντελεστές συσχέτισης δεν είναι ποτέ υψηλότεροι από 1. Ο συντελεστής συσχέτισης 1 σημαίνει ότι δύο μεταβλητές είναι απολύτως θετικά γραμμικές.

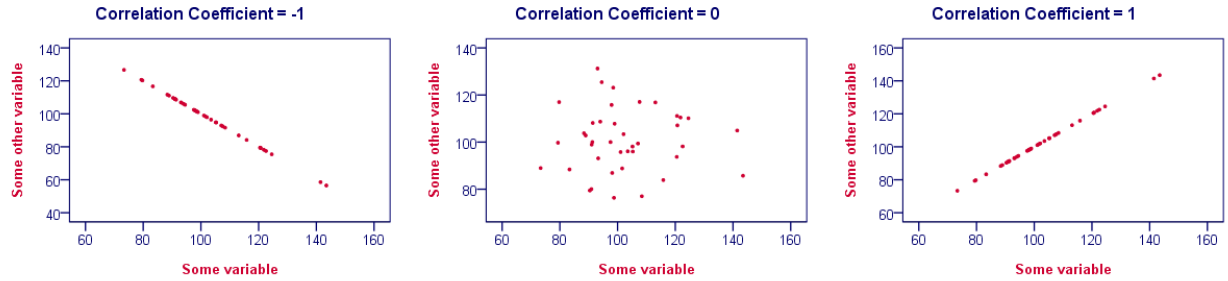
[11] https://www.researchgate.net/figure/Correlation-traces-a-time-series-of-a-Pearson-correlation-coefficient-during-the_fig1_328310203

οι κουκίδες σε μια διάσπαση σκέδασης βρίσκονται ακριβώς σε μια ευθεία αύξουσα γραμμή.



Εικόνα 8 Συσχετίσεις μεταξύ δύο μεταβλητών [12]

[12] https://www.spss-tutorials.com/pearson-correlation-coefficient/



Εικόνα 9 Παραδείγματα μεταβλητών με βάση τις συσχετίσεις [13]

3.2.3 *Συντελεστής Συσχέτισης - Προβλήματα Διερμηνείας*

Κατά την ερμηνεία των συσχετισμών, πρέπει να έχετε κατά νου ορισμένα πράγματα. Μια περίπλοκη συζήτηση αξίζει ένα ξεχωριστό φροντιστήριο, αλλά θα αναφερθούμε εν συντομία σε δύο βασικά σημεία.

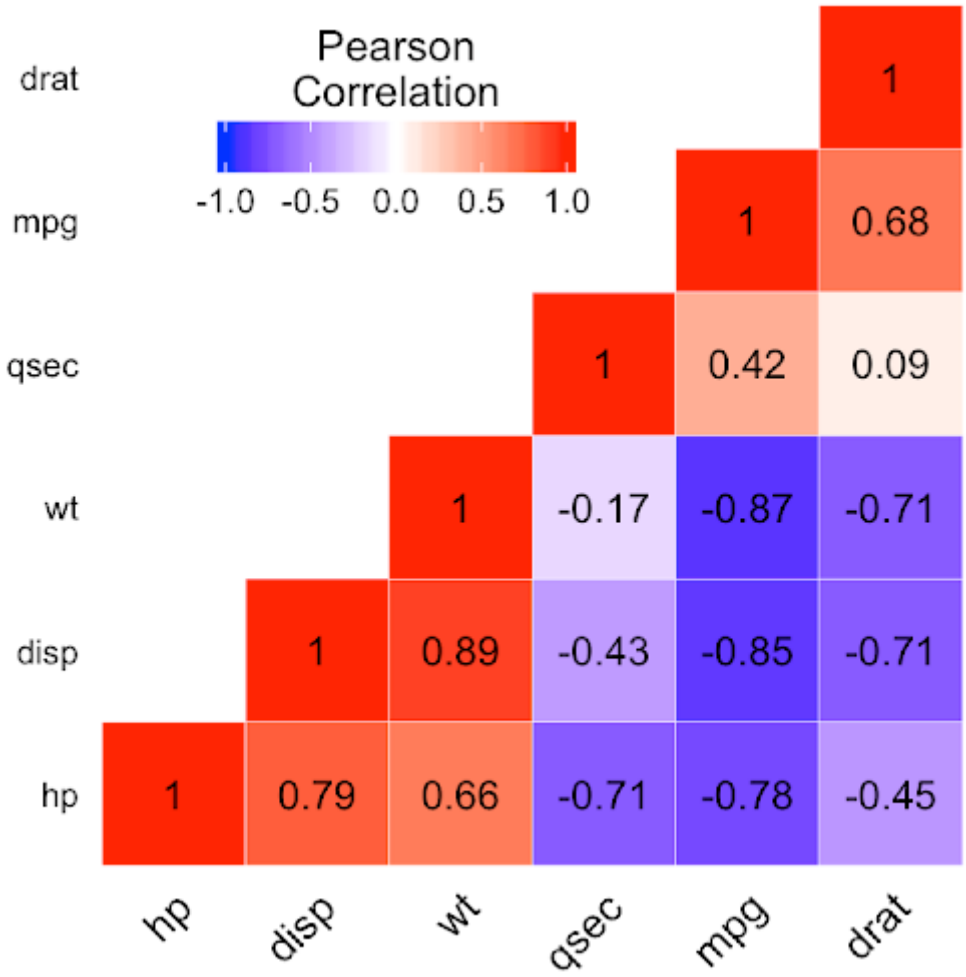
Οι συσχετισμοί μπορεί να δείχνουν ή να μην υποδηλώνουν αιτιώδεις σχέσεις . Αντιστρόφως, οι αιτιώδεις σχέσεις από κάποια μεταβλητή σε άλλη μεταβλητή ενδέχεται να οδηγήσουν ή όχι σε συσχέτιση μεταξύ των δύο μεταβλητών.

Οι συσχετισμοί είναι πολύ ευαίσθητοι στις αποκλίσεις . μια μοναδική ασυνήθιστη παρατήρηση μπορεί να έχει τεράστιο αντίκτυπο σε μια συσχέτιση. Αυτά τα αποθέματα ανιχνεύονται εύκολα από μια γρήγορη επιθεώρηση ενός scatterplot.

3.2.4 *Πίνακας Συσχέτισης*

Λάμβάνοντας υπόψη ότι οι συσχετισμοί ισχύουν για ζεύγη μεταβλητών. Αυτές οι συσχετίσεις παρουσιάζονται συνήθως σε έναν τετραγωνικό πίνακα γνωστό ως μήτρα συσχέτισης .Ένα παράδειγμα φαίνεται παρακάτω.

[13] https://www.spss-tutorials.com/pearson-correlation-coefficient/



Εικόνα 10 Συσχετίσεις ζευγών μεταβλητών [14]

[14] http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization

3.2.5 Jaccard Index

Ο δείκτης Jaccard, γνωστός και ως συντελεστής ομοιότητας Jaccard, είναι ένα στατιστικό στοιχείο που χρησιμοποιείται για την κατανόηση των ομοιότητας μεταξύ των συνόλων δειγμάτων. Η μέτρηση δίνει έμφαση στην ομοιότητα μεταξύ συνόλων πεπερασμένων δειγμάτων και ορίζεται τυπικά ως το μέγεθος της διατομής διαιρούμενο με το μέγεθος της ένωσης των συνόλων δειγμάτων. Η μαθηματική αναπαράσταση του δείκτη είναι γραμμένη ως εξής:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Παρόμοια με τον δείκτη Jaccard, που είναι μια μέτρηση της ομοιότητας, η απόσταση Jaccard μετρά την ανομοιογένεια μεταξύ των συνόλων δειγμάτων. Η απόσταση Jaccard υπολογίζεται με την εύρεση του δείκτη Jaccard και την αφαίρεση του από το 1, ή εναλλακτικά διαιρώντας τις διαφορές ny τη διασταύρωση των δύο συνόλων. Ο τύπος για την απόσταση Jaccard αντιπροσωπεύεται ως εξής:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

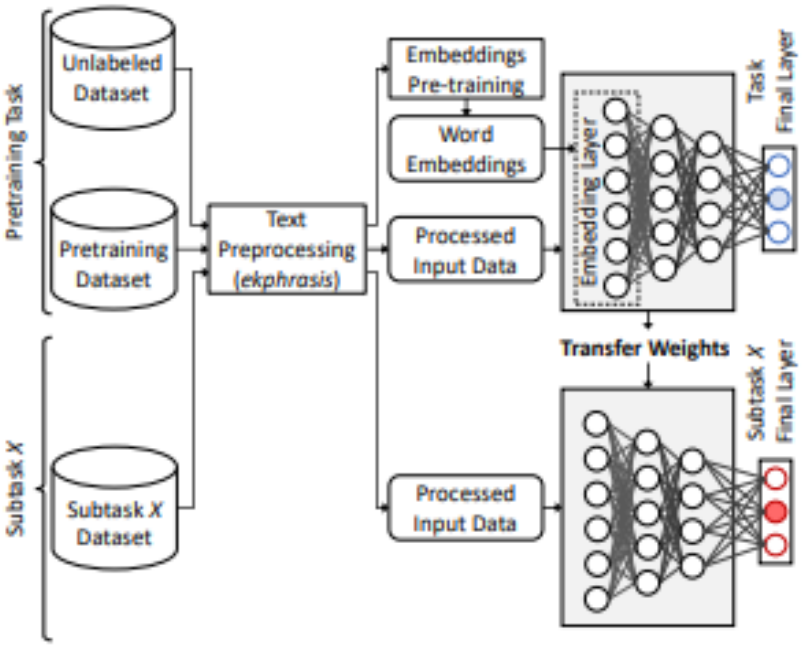
Αν σπάσει ο τύπος, ο δείκτης Jaccard είναι ουσιαστικά ο αριθμός και στις δύο ομάδες, διαιρούμενος με τον αριθμό σε κάθε ομάδα, πολλαπλασιασμένο επί 100. Αυτό θα προκαλέσει μια ποσοστιαία μέτρηση της ομοιότητας μεταξύ των δύο συνόλων δειγμάτων. Συνεπώς, για να βρείτε την απόσταση Jaccard, απλώς αφαιρέστε την ποσοστιαία τιμή από 1. Για παράδειγμα, αν η μέτρηση ομοιότητας είναι 35%, τότε η απόσταση Jaccard (1 - .35) είναι .65 ή 65%.

3.3 Αποτελέσματα διαγωνισμού

3.3.1 Προσεγγίσεις Ομάδων υψηλής βαθμολογίας

Σε αυτή την ενότητα θα γίνει η λεπτομερής ανάλυση των προσεγγίσεων των ομάδων οι οποίες πέτυχαν αποτελέσματα που τις κατατάσσουν στη κορυφή της κατάταξης ακρίβειας των μοντέλων. Αρχικά θα αναφερθούμε στη ομάδα SeerNet η οποία κατέχει την πρώτη θέση σχεδόν σε όλες τις υποενότητες, Θα ξεκινήσουμε με αυτή όχι μόνο γιατί κατέχει την περίοπτη θέση στην κατάταξη αλλά επειδή η προσέγγιση που ακολουθεί έχει αρκετές ομοιότητες με την προσεγγίση που υλοποιήσαμε στο δικό μας ερευνητικό κομμάτι. Όσο αφορά τη προεπεξεργασία τροποποιεί τα ακατέργαστα tweets ώστε να καταλήξουν στην εξαγωγή χαρακτηριστικών. Οι όροι προεπεξεργάζονται με τη χρήση του tweettokeniz. Οι συγκεκριμένες λέξεις-κλειδιά του Twitter αντικαθίστανται τμηματικά, δηλαδή, από USERNAME, PHONENUMBER, διευθύνσεις URL, timestamps. Όλοι οι χαρακτήρες μετατρέπονται σε πεζά γράμματα. Έπειτα μια ακολουθία από emojis χωρίζονται ανά μοναδιαία τμήματα. Και τέλος αντικαθιστούνται από την περιγραφή τους από ένα λεξικό. Συμμετείχαν σε 4 υποενότητες, EI-oc, EI-reg, V-oc, V-reg. Δύο από τις υποενότητες είναι κανονική ταξινόμηση και τα υπόλοιπα δύο είναι παλινδρομήσεις. Περιγράφετε η προσέγγιση για την κατασκευή μοντέλων μηχανικής μάθησης αμφότερες για τις παραπάνω υποενότητες. Συμμετείχαν στη συστηματική ταξινόμηση της έντασης των συναισθημάτων, που σκοπός ήταν η πρόβλεψη της ένταση των συναισθημάτων από τις κατηγορίες θυμό, φόβο, χαρά και θλίψη. Ξεχωριστά σύνολα δεδομένων προωθήθηκαν για κάθε τάξη συναισθημάτων. Ο στόχος της κατηγορικής ταξινόμησης του σθένους ήταν να αποδώσει στο tweet μία από τις 7 τάξεις [3, 3]. Πειραματίστηκαν με τον Classifier XG Boost, τον Random Forest Classifier του sklearn. Ακόμη για τις υποενότητες της παλινδρόμησης (E-reg, V-reg), σκοπός τουε ήταν να προβλεφθεί η ένταση σε κλίμακα 0-1. Πειραματίστηκαν με XG Boost Regressor, Random Forest Regressor της sklearn βιβλιοθήκης. Οι υπέρ-παραμέτροι κάθε μοντέλου ορίστηκαν για διαφορετικές περιπτώσεις. Τα δύο καλύτερα μοντέλα που αντιστοιχούν σε κάθε τύπο υποενότητας επιλέχθηκαν μετά την έβδομη πραγματοποίηση του cross-validation. Τα αποτελέσματα από όλους τους αλγορίθμους κατηγοριοποίησης και παλινδρόμησης για τα δεδομένα, προκύπτουν με την τεχνική επιλογής δέντρων αποφάσεων συγκεντρωτικών μοντέλων ούτως ώστε να συνδυάσουμε τα αποτελέσματα. Σε αυτή την περίπτωση, περάσαμε τα αποτελέσματα από τα μοντέλα σε αλγορίθμους επόμενου σταδίου ως είσοδο. Η έξοδος αυτού του σύνθετου μοντέλου αντιμετωπίζεται ως η τελική έξοδος του συστήματος. Παρατήρησαν ότι η χρήση κανονικών αλγορίθμων παλινδρόμησης έδωσε μεγαλύτερη ακρίβεια από τη χρήση αλγορίθμων κατηγοριοποίησης που επεξεργάζονται κάθε τιμή εξόδου ως κατηγορική τιμή.

Η επόμενη ακριβώς ομάδα που θα αναλύσουμε τον τρόπο με τον οποίο προσέγγισε τον διαγωνισμό ονομάζεται NTUA-SLP. Θα παρουσιάσουμε αρχικά το μοντέλο τους όπως το έχει σχεδιάσει η ίδια η ομάδα στη παρακάτω φωτογραφία.



Εικόνα 11 αρχιτεκτονική μοντέλου ομάδας NTUA-SLP[15]

Το πρώτο επίπεδο αφορά την προεκπαίδευση των του επιπέδων που περιέχουν τις ακολουθίες λέξεων(embedding layer) όπου το σύνολο δεδομένων εκπαιδεύεται με τις ακολουθίες λέξεων, το δεύτερο επίπεδο όπου δημιουργείται το νευρωνικό δίκτυο και τέλος η εφαρμογή αυτού του μοντέλου με τις κατάλληλες παραμέτρους στη κάθε υποενότητα. Στο κομμάτι της προεπεξεργασίας των όρων χρησιμοποίησαν τη βιβλιοθήκη ekphrasis. Συγκεκρίμενα η συγκεκριμένη βιβλιοθήκη τμηματοποιεί τους όρους, διορθώνει τα ορθογραφικά λάθη, κανονικοποιεί τις λέξεις και κατακερματίζει τις λέξεις από τα hashtags. Έπειτα για τη δημιουργία του νευρωνικού δικτύου αρχικοποίησαν τα βάρη του αρχικού επιπέδου ακολουθιών λέξεων με το προεπεξεργασμένο word2vec Twitter embeddings για να εκπαιδευτεί ένα αμφίδρομο νευρωνικό δίκτυο με ένα μηχανισμό εσωτερικής ανατροφοδότησης της εκπαίδευσης.

[15] https://arxiv.org/pdf/1804.06658.pdf

Τέλος η ομάδα PlusEmo2Vec χρησιμοποίησε τρία χαρακτηριστικά ως είσοδο. Αρχικά προχώρησε στη δημιουργία δύο μοντέλων με το πρώτο να είναι προεκπαιδευμένο μοντέλο DeepMoji, το οποίο εκπαιδεύτηκε σε 1,2 δισεκατομμυρίων σύνολο δεδομένων με προβλέψεις σε 64 ετικέτες emoji. Έπειτα χρησιμοποιείται διπλής κατεύθυνσης νευρωνικό δίκτυο με μνήμη μακράς διάρκειας και στο τελευταίο επίπεδο με τη συνάρτηση softmax για την εξαγωγή χαρακτηριστικών από το δεδομένα του διαγωνισμού. Κάθε δείγμα μετασχηματιζεται σε ενα διάνυσμα 2304 διαστάσεων. Μια παρόμοια μεθοδολογία για εκπαίδευση των λέξεων στο επόμενο βήμα και όχι των emoji ακολοθούθησαν σειριακά. Ακολούθησε η προεπεξεργασία των συνόλων δεδομένων τμηματοποιώντας τους όρους ολοκληρώνοντας έτσι το δέσιμο με τα τρία προηγούμενα επίπεδα.

Αποτελέσματα για El-reg

			Pearson r (all instances)					Pearson r (gold in 0.5-1)				
Test Set	Rank	Team Name	avg.	anger	fear	joy	sadness	avg.	anger	fear	joy	sadness
English												
	1	SeerNet	79.9	82.7	77.9	79.2	79.8	63.8	70.8	60.8	56.8	66.6
	2	NTUA-SLP	77.6	78.2	75.8	77.1	79.2	61.0	63.6	59.5	55.4	65.4
	3	PlusEmo2Vec	76.6	81.1	72.8	77.3	75.3	57.9	66.3	49.7	54.2	61.3
	23	Median Team	65.3	65.4	67.2	64.8	63.5	49.0	52.6	49.7	42.0	51.7
	37	SVM-Unigrams	52.0	52.6	52.5	57.5	45.3	39.6	45.5	30.2	47.6	35.0
	46	Random Baseline	-0.8	-1.8	2.4	-5.8	2.0	-4.8	-8.8	-1.1	-3.2	-5.9
Arabic												
	1	AffecThor	68.5	64.7	64.2	75.6	69.4	53.7	46.9	54.1	57.0	56.9
	2	EiTAKA	66.7	62.7	62.7	73.8	67.5	53.3	47.9	60.4	49.0	56.0
	3	EMA	64.3	61.5	59.3	70.9	65.6	49.0	44.4	45.7	49.7	56.2
	6	Median Team	54.2	50.1	50.1	62.8	53.7	44.6	39.1	43.0	45.4	51.0
	7	SVM-Unigrams	45.5	40.6	43.5	53.0	45.0	35.3	34.4	36.6	33.2	36.7
	13	Random Baseline	1.3	-0.6	1.6	-1.0	5.2	-0.7	0.2	0.7	1.1	-4.8
Spanish												
	1	AffecThor	73.8	67.6	77.6	75.3	74.6	58.7	54.9	60.4	59.1	60.4
	2	UG18	67.7	59.5	68.9	71.2	71.2	51.6	42.2	52.1	54.0	58.1
	3	ELiRF-UPV	64.8	59.1	63.2	66.3	70.5	44.0	41.0	37.5	45.6	51.7
	6	SVM-Unigrams	54.3	45.7	61.9	53.6	56.0	46.2	42.9	47.4	47.9	46.4
	8	Median Team	44.1	34.8	53.3	41.4	47.1	38.2	24.6	42.5	44.8	41.0
	15	Random Baseline	-1.2	-5.6	0.4	1.8	-1.4	-0.5	0.1	-4.6	1.8	0.8

Εικόνα 12 **http://saifmohammad.com/WebDocs/semeval2018-task1.pdf**

Αποτελέσματα για El-oc

			Pearson r (all classes)					Pearson r (some-emotion)				
Test Set	Rank	Team Name	avg	anger	fear	joy	sadness	avg	anger	fear	joy	sadness
English												
	1	SeerNet	69.5	70.6	63.7	72.0	71.7	54.7	55.9	45.8	61.0	56.0
	2	PlusEmo2Vec	65.9	70.4	52.8	72.0	68.3	50.1	54.8	32.0	60.4	53.3
	3	psyML	65.3	67.0	58.8	68.6	66.7	50.5	51.7	46.8	57.0	46.3
	17	Median Team	53.0	53.0	47.0	55.2	56.7	41.5	40.8	31.0	49.4	44.8
	26	SVM-Unigrams	39.4	38.2	35.5	46.9	37.0	29.6	31.5	18.3	39.6	28.9
	37	Random Baseline	-1.6	-6.2	4.7	1.4	-6.1	-1.1	-3.8	-0.7	-0.2	0.1
Arabic												
	1	AffecThor	58.7	55.1	55.1	63.1	61.8	43.7	42.6	47.2	44.6	40.4
	2	EiTAKA	57.4	57.2	52.9	63.4	56.3	46.0	48.8	47.6	50.9	36.6
	3	UNCC	51.7	45.9	48.3	53.8	58.7	36.3	34.1	33.1	38.3	39.8
	6	SVM-Unigrams	31.5	28.1	28.1	39.6	30.2	23.6	25.1	25.2	24.1	20.1
	7	Median Team	30.5	30.1	24.2	36.0	31.5	24.8	24.2	17.2	28.3	29.4
	11	Random Baseline	0.6	-5.7	-1.9	0.8	9.2	1.2	0.2	-2.0	2.9	3.7
Spanish												
	1	AffecThor	66.4	60.6	70.6	66.7	67.7	54.2	47.4	58.8	53.5	57.2
	2	UG18	59.9	49.9	60.6	66.5	62.5	48.5	38.0	49.3	53.1	53.4
	3	INGEOTEC	59.6	46.8	63.4	65.5	62.8	46.3	33.0	49.8	53.3	49.2
	6	SVM-Unigrams	48.1	44.4	54.6	45.1	48.3	40.8	37.1	46.1	37.1	42.7
	8	Median Team	36.0	26.3	28.3	51.3	38.0	33.1	24.0	26.1	50.5	31.6
	15	Random Baseline	-2.2	1.1	-6.9	-0.5	-2.7	1.6	0.2	-1.8	4.4	3.6

Εικόνα 13 **http://saifmohammad.com/WebDocs/semeval2018-task1.pdf**

Αποτελέσματα για El-oc

			Pearson r (all classes)					Pearson r (some-emotion)				
Test Set	Rank	Team Name	avg	anger	fear	joy	sadness	avg	anger	fear	joy	sadness
English												
	1	SeerNet	69.5	70.6	63.7	72.0	71.7	54.7	55.9	45.8	61.0	56.0
	2	PlusEmo2Vec	65.9	70.4	52.8	72.0	68.3	50.1	54.8	32.0	60.4	53.3
	3	psyML	65.3	67.0	58.8	68.6	66.7	50.5	51.7	46.8	57.0	46.3
	17	Median Team	53.0	53.0	47.0	55.2	56.7	41.5	40.8	31.0	49.4	44.8
	26	SVM-Unigrams	39.4	38.2	35.5	46.9	37.0	29.6	31.5	18.3	39.6	28.9
	37	Random Baseline	-1.6	-6.2	4.7	1.4	-6.1	-1.1	-3.8	-0.7	-0.2	0.1
Arabic												
	1	AffecThor	58.7	55.1	55.1	63.1	61.8	43.7	42.6	47.2	44.6	40.4
	2	EiTAKA	57.4	57.2	52.9	63.4	56.3	46.0	48.8	47.6	50.9	36.6
	3	UNCC	51.7	45.9	48.3	53.8	58.7	36.3	34.1	33.1	38.3	39.8
	6	SVM-Unigrams	31.5	28.1	28.1	39.6	30.2	23.6	25.1	25.2	24.1	20.1
	7	Median Team	30.5	30.1	24.2	36.0	31.5	24.8	24.2	17.2	28.3	29.4
	11	Random Baseline	0.6	-5.7	-1.9	0.8	9.2	1.2	0.2	-2.0	2.9	3.7
Spanish												
	1	AffecThor	66.4	60.6	70.6	66.7	67.7	54.2	47.4	58.8	53.5	57.2
	2	UG18	59.9	49.9	60.6	66.5	62.5	48.5	38.0	49.3	53.1	53.4
	3	INGEOTEC	59.6	46.8	63.4	65.5	62.8	46.3	33.0	49.8	53.3	49.2
	6	SVM-Unigrams	48.1	44.4	54.6	45.1	48.3	40.8	37.1	46.1	37.1	42.7
	8	Median Team	36.0	26.3	28.3	51.3	38.0	33.1	24.0	26.1	50.5	31.6
	15	Random Baseline	-2.2	1.1	-6.9	-0.5	-2.7	1.6	0.2	-1.8	4.4	3.6

Εικόνα 14 **http://saifmohammad.com/WebDocs/semeval2018-task1.pdf**

Αποτελέσματα για V-oc

Rank	Team Name	r (all)	r (some emo)
English			
1	SeerNet	83.6	88.4
2	PlusEmo2Vec	83.3	87.8
3	Amobee	81.3	86.5
18	Median Team	68.2	75.4
24	SVM-Unigrams	50.9	56.0
36	Random Baseline	-1.0	-1.2
Arabic			
1	EiTAKA	80.9	84.7
2	AffecThor	75.2	79.2
3	INGEOTEC	74.9	78.9
7	Median Team	55.2	59.6
8	SVM-Unigrams	47.1	50.5
14	Random Baseline	1.1	0.9
Spanish			
1	Amobee	76.5	80.4
2	AffecThor	75.6	79.2
3	ELiRF-UPV	72.9	76.5
6	Median Team	55.6	59.1
8	SVM-Unigrams	41.8	46.1
13	Random Baseline	-4.2	-4.3

Εικόνα 15 **http://saifmohammad.com/WebDocs/semeval2018-task1.pdf**

Αποτελέσματα για E-c

Rank	Team Name	acc.	micro F1	macro F1
English				
1	NTUA-SLP	58.8	70.1	52.8
2	TCS Research	58.2	69.3	53.0
3	PlusEmo2Vec	57.6	69.2	49.7
17	Median Team	47.1	59.9	46.4
21	SVM-Unigrams	44.2	57.0	44.3
28	Random Baseline	18.5	30.7	28.5
Arabic				
1	EMA	48.9	61.8	46.1
2	PARTNA	48.4	60.8	47.5
3	Tw-StAR	46.5	59.7	44.6
6	SVM-Unigrams	38.0	51.6	38.4
7	Median Team	25.4	37.9	25.0
9	Random Baseline	17.7	29.4	27.5
Spanish				
1	MILAB_SNU	46.9	55.8	40.7
2	ELiRF-UPV	45.8	53.5	44.0
3	Tw-StAR	43.8	52.0	39.2
4	SVM-Unigrams	39.3	47.8	38.2
7	Median Team	16.7	27.5	18.7
8	Random Baseline	13.4	22.8	21.3

Εικόνα 16 **http://saifmohammad.com/WebDocs/semeval2018-task1.pdf**

4

Η προσέγγιση μας

4.1 Τεχνική που ακολουθήσαμε

Η τεχνική που ακολουθήσαμε είναι αυτή της επιβλεπόμενης μηχανικής μάθησης. Έχοντας ένα προκαθορισμένο σύνολο από κλάσεις, σκοπός μας είναι να τοποθετήσουμε τα αντικείμενα που εξετάζουμε σε κάθε μια από τις προκαθορισμένες κλάσεις η οποία θα περιγράφει τη συναισθηματική κατάσταση εξάγοντας από το εκάστοτε κείμενο.

Στη συγκεκριμένη τεχνική και αλγοριθμοποίηση της επιβλεπόμενης μηχανικής μάθησης αποτελείται από τρία στάδια.

1. Προ-επεξεργασία δεδομένων
2. Χαρακτηριστικά (features) – Διανυσματοποίηση
3. Ταξινομητές

4.1.1 Πρώτο στάδιο (Προ-επεξεργασία Δεδομένων)

1. Οι διαδικασίες που ακολουθήσαμε στη προεπεξεργασία δεδομένων κειμένου αναλύονται παρακάτω:

2. Αφαίρεση των σημείων στίξης

3. Αφαίρεση Stopwords δηλαδή των λέξεων που η σπανιότητα τους σε ένα κείμενο είναι μικρή

4. Stemming, μετατροπή των λέξεων

5. Μετατροπή των λέξεων σε πεζά γράμματα

4.1.1.1 Αφαίρεση σημείων στίξης

Αρχικός στόχος της προ-επεξεργασίας των εγγράφων αποτελεί η αφαίρεση όλων των περιττών συμβόλων και των σημείων στίξης καθώς αυτά δεν προσφέρουν καμία πληροφορία και δεν έχουν καμία σχέση με το εννοιολογικό περιεχόμενο του κειμένου. Η διαδικασία αφαίρεσης όλων των περιττών συμβόλων και σημείων στίξης, όπως για παράδειγμα: _ ( ) + = @ # % - . , / ! &, πραγματοποιήθηκε μέσω ελέγχου των συνόλου δεδομένων και αφαίρεση τους στα σημεία που βρίσκονταν. Τα σημεία στίξης συνήθως αφαιρούνται από το κείμενο πριν από την προεπεξεργασία και δεν περιλαμβάνονται στο παραγόμενο κείμενο. Στην εσωτερική όψη του κειμένου μας δίνεται η εντύπωση οπτικά ότι εννοιολογικά δεν είναι σωστό, καθώς τα σημεία στίξης είναι αναπόσπαστο κομμάτι της γραπτής γλώσσας. Είναι τρομερά δύσκολο να παράγεται κομμάτι κειμένου μεγάλης σημαντικότητας χωρίς σημεία στίξης, ή ακόμα να κατανοεί οποιοδήποτε ένα τέτοιο σώμα κειμένου.

Ωστόσο, αυτό έχει γίνει στον τομέα της επεξεργασίας φυσικής γλώσσας. Η στίξη θεωρείται εδώ και καιρό ότι είναι στενά συνδεδεμένη με τον τονισμό, δηλαδή ότι τα διαφορετικά σημεία στίξης δίνουν απλά τις ενδείξεις του αναγνώστη ως προς τα πιθανά χαρακτηριστικά έμφασης και

διακοπής του κειμένου. Ωστόσο, ακόμη και αν αναγνωρίσουμε ότι η στίξη πληρεί τον δικό της γλωσσικό ρόλο, δεν είναι καθόλου σαφές πώς ορίζεται αυτός ο ρόλος.

4.1.1.2 Αφαίρεση Stopwords

Τα stopwords είναι γενικά λέξεις που εμφανίζονται πολύ συχνά σε ένα κείμενο, χωρίς όμως να φέρουν ιδιαίτερη πληροφορία σχετικά με το περιεχόμενο του κειμένου. Η ύπαρξη τους καθορίζεται από συντακτικούς και εννοιολογικούς κανόνες. Παράδειγμα stopwords αποτελούν οι σύνδεσμοι των προτάσεων. Επισημαίνεται, πως ο αριθμός των όρων μειώθηκε σε μεγάλο βαθμό.

Η μέθοδος αυτή βασίζεται στην κατάργηση των stopwords λέξεων που έχουν ληφθεί από προ-καταρτισμένες λίστες. Για το σκοπό αυτής της εργασίας επιλέξαμε την κλασική λίστα της NLTK Corpus.

Ο αλγόριθμος υλοποιείται σύμφωνα με τα παρακάτω βήματα.

Βήμα 1: Το κείμενο του εγγράφου προορισμού είναι tokenized και μεμονωμένες λέξεις αποθηκεύονται σε πίνακα.

Βήμα 2: Λαμβάνεται μια ενιαία λέξη stop από τη λίστα stopword.

Βήμα 3: Η λέξη stop είναι σε σύγκριση με το κείμενο που ελέγχουμε σε μορφή πίνακα χρησιμοποιώντας τεχνική διαδοχικής αναζήτησης.

Βήμα 4: Εάν ταιριάζει, η λέξη στον πίνακα αφαιρείται και η σύγκριση συνεχίζεται μέχρι το μήκος του πίνακα.

Βήμα 5: Εμφανίζεται κείμενο που δεν περιέχει προειδοποιητικές λέξεις.

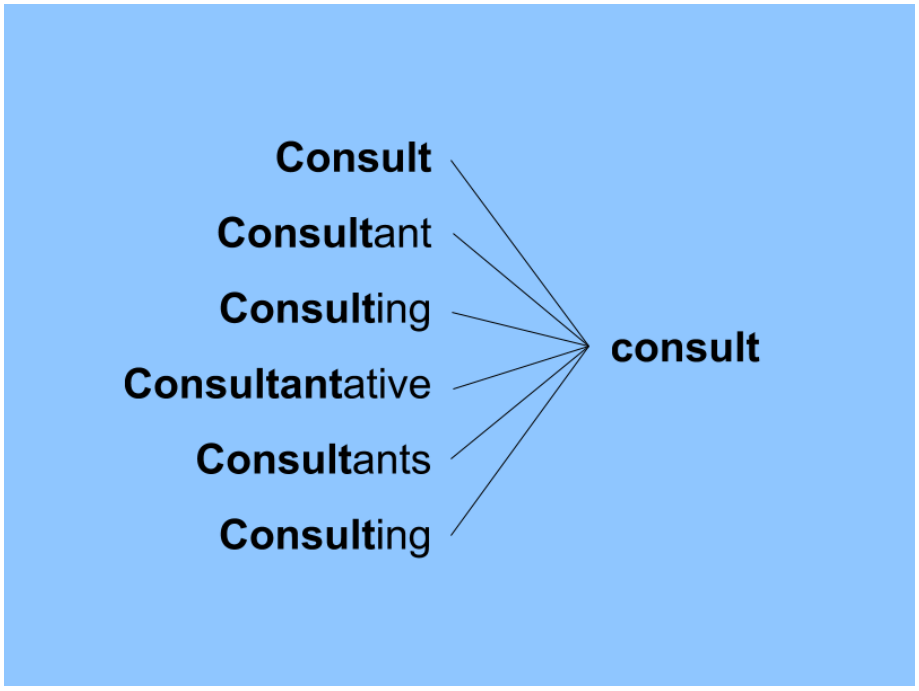
Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Εικόνα 17 Παράδειγμα αφαίρεσης stop words [16]

4.1.1.3 Stemming

Το stemming των λέξεων αποσκοπεί στη μείωση του αριθμού των όρων-λέξεων, μέσω των οποίων γίνεται η αναπαράσταση των κειμένων. Η συγκεκριμένη διαδικασία είναι ιδιαίτερα σημαντική κατά τη διαδικασία της ομαδοποίησης, αφού καθιστά τη διαδικασία λιγότερο εξαρτώμενη από τις ιδιαίτερες μορφές των λέξεων. Ουσιαστικά, κατά τη διαδικασία του stemming, πραγματοποιείται η αναγωγή του συνόλου των λέξεων ενός κειμένου στη ρίζα τους, δημιουργώντας ταυτόχρονα ομάδες ομόριζων λέξεων. Για παράδειγμα οι λέξεις «αγαπάω», «αγάπη», «αγάπησε», «αγαπητικός», «αγαπημένος» μπορούν να αναχθούν στην κοινή τους ρίζα «αγαπ». Μια μορφή μετασχηματισμού λέξεων, stemming, είναι η διαδικασία της μετατροπής του πληθυντικού αριθμού των λέξεων στον ενικό και η μετατροπή των παρελθοντικών χρόνων των ρημάτων στον ενεστώτα. Οι αλγόριθμοι που εκτελούν τη μετατροπή μιας λέξης στη γλωσσολογικά ορθή ρίζα της καλούνται «Lemmatizers» και η διαδικασία «Lemmatization». Υπάρχουν αρκετοί αλγόριθμοι για stemming. Ένας από αυτούς είναι ο αλγόριθμος Lovins Stemmer . Ο πλέον όμως, δημοφιλής αλγόριθμος για stemming είναι και ο αλγόριθμος του Porter, τον οποίο χρησιμοποιεί και το λογισμικό ανάλυσης περιεχομένου WordStat. Αξίζει να σημειωθεί όμως, πως οι δύο παραπάνω αλγόριθμοι βρίσκουν εφαρμογή σε λέξεις της αγγλικής γλώσσας, ενώ δεν υποστηρίζουν το ελληνικό αλφάβητο.

[16] https://www.geeksforgeeks.org/removing-stop-words-nltk-python/



Εικόνα 18 Παράδειγμα εφαρμογής stemming [17]

4.1.1.4 Μετατροπή των λέξεων σε πεζά γράμματα

Αφού ολοκληρώσαμε όλες τις παραπάνω διαδικασίες και πριν καταχωρήσουμε τα δεδομένα σε πίνακες για την εύκολη διαχείριση τους μετατρέψαμε όλα τα γράμματα σε πεζά έτσι ώστε να έχουν όλα την ίδια δομή.

[17] https://devopedia.org/stemming

4.1.2 Χαρακτηριστικά (features) – Διανυσματοποίηση

Επιλογή μεθόδου διανυσματοποίησης για εξαγωγή χαρακτηριστικών

- Term frequency * Inverse document frequency (TF-IDF)
- Count Vectorizer
- Υπολογισμός παραμέτρων υπολογισμού διανύσματος

*4.1.2.1 Term frequency * Inverse document frequency (TF-IDF)*

Ορισμός

1. Το tf-idf είναι προϊόν δύο στατιστικών στοιχείων, *συχνότητας όρων* και *αντίστροφης συχνότητας εγγράφων* . Υπάρχουν διάφοροι τρόποι για τον προσδιορισμό των ακριβών τιμών και των δύο στατιστικών στοιχείων.

2. Μια φόρμουλα που έχει ως στόχο να καθορίσει τη σημασία μιας λέξης-κλειδιού ή μιας φράσης μέσα σε ένα έγγραφο ή μια ιστοσελίδα.

Συχνότητα όρου

Στην περίπτωση του **όρου συχνότητα** tf (t , d) , η απλούστερη επιλογή είναι να χρησιμοποιήσετε τον *ακατέργαστο αριθμό* ενός όρου σε ένα έγγραφο, δηλαδή τον αριθμό των φορών που εμφανίζεται ο όρος t στο έγγραφο d . Αν συμβολίζουμε την ακατέργαστη μέτρηση από $f_{t,d}$, τότε η απλούστερη καθεστώς tf είναι tf (t , d) = $f_{t,d}$. Άλλες δυνατότητες περιλαμβάνουν :

- Boolean "συχνότητες": tf (t , d) = 1 αν t συμβαίνει στο d και 0 διαφορετικά?
- η συχνότητα που έχει οριστεί για το μήκος του εγγράφου: $f_{t,d}$ ÷ (αριθμός λέξεων στο d)
- Λογαριθμική κλίμακα : tf (t , d) = log (1 + $f_{t,d}$) .

- αυξημένη συχνότητα, για να αποφευχθεί η μεροληψία προς μεγαλύτερα έγγραφα, π.χ. ακατέργαστη συχνότητα διαιρούμενη με την ακατέργαστη συχνότητα του πιο εμφανιζόμενου όρου στο έγγραφο:

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

Παραλλαγές του χρόνου συχνότητας (tf)

βάρους	tf βάρος
δυάδικος	$0, 1$
ακατέργαστη μέτρηση	$f_{t,d}$
μακροπρόθεσμη συχνότητα	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
κανονικοποίηση λογαρίθμου	$\log(1 + f_{t,d})$
διπλή κανονικοποίηση 0,5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
διπλή κανονικοποίηση K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

[18]

Αντίστροφη συχνότητα όρου

Η **αντίστροφη συχνότητα εγγράφων** είναι ένα μέτρο της ποσότητας πληροφοριών που παρέχει η λέξη, δηλαδή εάν είναι συνηθισμένη ή σπάνια σε όλα τα έγγραφα. Είναι το λογαριθμικά κλιμακωτό αντίστροφο τμήμα των εγγράφων που περιέχουν τη λέξη (που λαμβάνεται διαιρώντας τον συνολικό αριθμό των εγγράφων με τον αριθμό των εγγράφων που περιέχουν τον όρο και στη συνέχεια λαμβάνοντας τον λογάριθμο αυτού του πηκτικού):

[18] https://en.wikipedia.org/wiki/Tf%E2%80%93idf#Definition

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

με

- N : ο συνολικός αριθμός των εγγράφων στο σώμα

- $|\{d \in D : t \in d\}|$: αριθμός εγγράφων όπου ο όρος t εμφανίζεται (δηλαδή, tf(t,d)≠0). Εάν ο όρος δεν είναι στο σώμα, αυτό θα οδηγήσει σε διαίρεση-με-μηδέν. Συνεπώς, είναι κοινό να προσαρμόζεται ο παρονομαστής σε

$$1 + |\{d \in D : t \in d\}|$$

Παραλλαγές βάρους ανάστροφης συχνότητας εγγράφων (idf)

βάρους	βάρος IDF ($n_t = |\{d \in D : t \in d\}|$)
unary	1
αντίστροφη συχνότητα εγγράφων	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
αντίστροφη συχνότητα εγγράφου ομαλή	$\log\left(\frac{N}{1+n_t}\right) + 1$
αντίστροφη συχνότητα εγγράφου max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
πιθανή αντίστροφη συχνότητα εγγράφων	$\log \frac{N-n_t}{n_t}$

19

[19] https://en.wikipedia.org/wiki/Tf%E2%80%93idf#Definition

Θεματική συχνότητα-Αντίστροφη συχνότητα εγγράφου

Στη συνέχεια, το tf-idf υπολογίζεται ως

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

Ένα υψηλό βάρος στο tf-idf επιτυγχάνεται με μια υψηλή συχνότητα (στο δεδομένο έγγραφο) και μια χαμηλή συχνότητα εγγράφων του όρου σε ολόκληρη τη συλλογή των εγγράφων. Τα βάρη επομένως τείνουν να φιλτράρουν τους κοινούς όρους. Δεδομένου ότι ο λόγος μέσα στη λειτουργία καταγραφής του idf είναι πάντα μεγαλύτερος ή ίσος με 1, η τιμή του idf (και tf-idf) είναι μεγαλύτερη ή ίση με 0. Καθώς ένας όρος εμφανίζεται σε περισσότερα έγγραφα, ο λόγος μέσα στον λογάριθμο προσεγγίζει 1 , φέρνοντας το idf και tf-idf πιο κοντά στο 0.

Συνιστώμενα σχήματα βαρύτητας tf-idf

βάρους	το βάρος του εγγράφου	ερώτημα βάρος βάρος
1	$f_{t,d} \cdot \log \frac{N}{n_t}$	$\left(0.5 + 0.5 \frac{f_{t,q}}{\max_t f_{t,q}}\right) \cdot \log \frac{N}{n_t}$
2	$1 + \log f_{t,d}$	$\log\left(1 + \frac{N}{n_t}\right)$
3	$(1 + \log f_{t,d}) \cdot \log \frac{N}{n_t}$	$(1 + \log f_{t,q}) \cdot \log \frac{N}{n_t}$

20

[20] https://en.wikipedia.org/wiki/Tf%E2%80%93idf#Definition

4.1.2.2 Count Vectorizer (Bag-of-Words)

Το μοντέλο bag-of-words είναι μια απλουστευμένη αναπαράσταση που χρησιμοποιείται στην επεξεργασία της φυσικής γλώσσας και στην ανάκτηση πληροφοριών (IR). Σε αυτό το μοντέλο, ένα κείμενο (όπως μια πρόταση ή ένα έγγραφο) αναπαρίσταται ως η τσάντα (multiset) των λέξεων της, αγνοώντας τη γραμματική και ακόμη και την τάξη των λέξεων αλλά διατηρώντας την πολλαπλότητα . Το μοντέλο bag-of-words έχει επίσης χρησιμοποιηθεί για την όραση του υπολογιστή .

Το μοντέλο bag-of-words χρησιμοποιείται συνήθως σε μεθόδους ταξινόμησης εγγράφων όπου η συχνότητα εμφάνισης κάθε λέξης χρησιμοποιείται ως χαρακτηριστικό γνώρισμα για την κατάρτιση ενός ταξινομητή.

Παράδειγμα

```
(1) John likes to watch movies. Mary likes movies too.
(2) Mary also likes to watch football games.
```

```
"John","likes","to","watch","movies","Mary","likes","movies","too"

"Mary","also","likes","to","watch","football","games"
```

Εφαρμογή

Στην πράξη, το μοντέλο Bag-of-words χρησιμοποιείται κυρίως ως εργαλείο δημιουργίας χαρακτηριστικών. Αφού μετατρέψουμε το κείμενο σε λίστα λέξεων, μπορούμε να υπολογίσουμε διάφορα μέτρα για να χαρακτηρίσουμε το κείμενο. Ο συνηθέστερος τύπος χαρακτηριστικών ή χαρακτηριστικών που υπολογίζονται από το μοντέλο Bag-of-words είναι η μακροχρόνια συχνότητα, δηλαδή η συχνότητα που εμφανίζεται ένας όρος στο κείμενο. Για το παραπάνω παράδειγμα, μπορούμε να κατασκευάσουμε για τους ακόλουθους δύο καταλόγους ώστε να καταγράψουμε τις συχνότητες όρων όλων των διακριτών λέξεων:

```
(1) [1, 2, 1, 1, 2, 1, 1, 0, 0, 0]
(2) [0, 1, 1, 1, 0, 1, 0, 1, 1, 1]
```

Κάθε καταχώρηση των λιστών αναφέρεται στην καταμέτρηση της αντίστοιχης καταχώρησης στη λίστα (αυτή είναι και η παράσταση ιστόγραμμα). Για παράδειγμα, στην πρώτη λίστα (η οποία αντιπροσωπεύει το έγγραφο 1), οι δύο πρώτες καταχωρήσεις είναι "1,2":

- Η πρώτη καταχώρηση αντιστοιχεί στη λέξη "John", η οποία είναι η πρώτη λέξη στη λίστα και η τιμή της είναι "1" επειδή εμφανίζεται το "John" στο πρώτο έγγραφο 1 φορά.
- Η δεύτερη καταχώρηση αντιστοιχεί στη λέξη "likes", η οποία είναι η δεύτερη λέξη στη λίστα και η τιμή της είναι "2" επειδή το "likes" εμφανίζεται στο πρώτο έγγραφο 2 φορές.

Αυτή η αναπαράσταση λίστας (ή φορέα) δεν διατηρεί τη σειρά των λέξεων στις αρχικές προτάσεις. Αυτό είναι μόνο το κύριο χαρακτηριστικό του μοντέλου Bag-of-words. Αυτό το είδος αναπαράστασης έχει αρκετές επιτυχημένες εφαρμογές, όπως το φιλτράρισμα ηλεκτρονικού ταχυδρομείου.

Ωστόσο, οι συχνότητες όρων δεν είναι απαραιτήτως η καλύτερη αντιπροσώπευση για το κείμενο. Κοινές λέξεις όπως "the", "a", "to" είναι σχεδόν πάντα οι όροι με την υψηλότερη συχνότητα στο κείμενο. Επομένως, η κατοχή ενός υψηλού ακαθάριστου αριθμού δεν σημαίνει απαραίτητα ότι η αντίστοιχη λέξη είναι πιο σημαντική. Για να αντιμετωπιστεί αυτό το πρόβλημα, ένας από τους πιο δημοφιλείς τρόπους για την "ομαλοποίηση" των όρων συχνότητας είναι να βρεθεί ένας όρος από το αντίστροφο της συχνότητας εγγράφου ή tf-idf . Επιπλέον, για τον συγκεκριμένο σκοπό της ταξινόμησης, έχουν αναπτυχθεί ελεγχόμενες εναλλακτικές λύσεις για να λαμβάνουν υπόψη την ετικέτα κλάσης ενός εγγράφου. Τέλος, χρησιμοποιείται δυαδική στάθμιση (παρουσία / απουσία ή 1/0) στη θέση των συχνοτήτων για ορισμένα προβλήματα (π.χ., σύστημα λογισμικού εκμάθησης μηχανών).

4.1.2.3 N-gramms

Στους τομείς της υπολογιστικής γλωσσολογίας και της πιθανότητας , ένα n -gram είναι μια συνεχόμενη ακολουθία n στοιχείων από ένα δεδομένο δείγμα κειμένου ή ομιλίας. Τα στοιχεία μπορούν να είναι συλλαβές , γράμματα , λέξεις ή

ζεύγη σύμφωνα με την εφαρμογή. Τα n -grams τυπικά συλλέγονται από ένα σώμα κειμένου ή ομιλίας.

Χρησιμοποιώντας λατινικά αριθμητικά προθέματα , ένα n -gram μεγέθους 1 αναφέρεται ως "unigram". το μέγεθος 2 είναι ένα " bigram " (ή, λιγότερο συχνά, ένα "digram") το μέγεθος 3 είναι ένα " τρίγραμμα ". Χρησιμοποιούνται μερικές φορές αγγλικές κάρτες , π.χ. "τετράγωνα", "πέντε γραμμάρια". Ένα πολυμερές ή ολιγομερές γνωστού μεγέθους ονομάζεται k-mer αντί ενός n-gram, με συγκεκριμένα ονόματα χρησιμοποιώντας ελληνικά αριθμητικά προθέματα όπως "μονομερές", "διμερές", "τριμερές", "τετραμερές" "πενταμερές", κ.λπ.

Εφαρμογές

Ένα μοντέλο n -gram είναι ένας τύπος πιθανολογικού μοντέλου γλωσσομάθειας για την πρόβλεψη του επόμενου στοιχείου σε μια τέτοια ακολουθία με τη μορφή ενός (n - 1) -order Markov μοντέλου[21] . Τα μοντέλα n -gram χρησιμοποιούνται σήμερα ευρέως στην πιθανότητα , στη θεωρία της επικοινωνίας , στην υπολογιστική γλωσσολογία (για παράδειγμα στη στατιστική επεξεργασία της φυσικής γλώσσας), στην υπολογιστική βιολογία (για παράδειγμα στην ανάλυση βιολογικών ακολουθιών) και στη συμπίεση δεδομένων . Δύο οφέλη από n(και αλγόριθμοι που τις χρησιμοποιούν) είναι η απλότητα και η δυνατότητα κλιμάκωσης - με μεγαλύτερο n , ένα μοντέλο μπορεί να αποθηκεύσει περισσότερα περιβάλλοντα με ένα καλά κατανοητό εμπόριο χωροχρόνου , επιτρέποντας στα μικρά πειράματα να κλιμακώνονται αποτελεσματικά.

Τα μοντέλα n -gram χρησιμοποιούνται ευρέως στη στατιστική επεξεργασία φυσικής γλώσσας . Στην αναγνώριση ομιλίας , τα φωνήματα και οι ακολουθίες των φωνημάτων μοντελοποιούνται χρησιμοποιώντας μια κατανομή n -gram. Για την ανάλυση, οι λέξεις μοντελοποιούνται έτσι ώστε κάθε n -gram να αποτελείται από n λέξεις. Για την αναγνώριση γλώσσας , ακολουθίες χαρακτήρων / γραφημάτων (π.χ. , τα γράμματα της αλφαβήτου) είναι το πρότυπο για διαφορετικές γλώσσες. [4] Για τις ακολουθίες χαρακτήρων, τα 3-γραμμάρια (που μερικές φορές αναφέρονται ως "trigrams") που μπορούν να δημιουργηθούν από το "καλό καλοκαίρι" είναι "goo", "ood", "od", "dm", "mo" "και ούτω καθεξής, μετρώντας τον χαρακτήρα χώρου ως γραμμάριο (μερικές φορές η αρχή και το τέλος ενός κειμένου διαμορφώνονται ρητά, προσθέτοντας" _ □ _g "," _go "," ng_ "και" g_ □ _ "). Για τις ακολουθίες λέξεων, τα τριγράμματα που μπορεί να δημιουργηθούν από το "σκυλί που μύριζε σαν skunk" είναι "το σκυλί", "το σκυλί μύριζε", "το σκυλί μύριζε σαν", "μύριζε σαν", " ένα skunk "και" ένα skunk # ".

21 https://el.wikipedia.org/wiki/%CE%91%CE%BB%CF%85%CF%83%CE%AF%CE%B4%CE%B1_%CE%9C%CE%AC%CF%81%CE%BA%CE%BF%CF%86

Οι ασκούμενοι που ενδιαφέρονται περισσότερο για όρους πολλαπλών λέξεων ενδέχεται να προεπεξεργαστούν χορδές για να αφαιρέσουν κενά. Πολλοί απλά καταργούν το κενό σε έναν ενιαίο χώρο διατηρώντας ταυτόχρονα σημάδια παραγράφων, διότι ο κενός χώρος είναι συχνά είτε στοιχείο στυλ γραφής είτε εισάγει διάταξη ή παρουσίαση που δεν απαιτείται από τη μεθοδολογία πρόβλεψης και έκπτωσης. Η στίξη είναι επίσης συνήθως μειωμένη ή αφαιρεθεί με προεπεξεργασία και χρησιμοποιείται συχνά για να ενεργοποιήσει τη λειτουργικότητα.

Τα n -grams μπορούν επίσης να χρησιμοποιηθούν για ακολουθίες λέξεων ή σχεδόν οποιουδήποτε τύπου δεδομένων. Για παράδειγμα, έχουν χρησιμοποιηθεί για την εξαγωγή χαρακτηριστικών για τη συσσώρευση μεγάλων συνόλων δορυφορικών εικόνων και για τον προσδιορισμό σε ποιο τμήμα της γης προέκυψε μια συγκεκριμένη εικόνα. Έχουν επίσης μεγάλη επιτυχία ως πρώτη διέλευση στην αναζήτηση γενετικής αλληλουχίας και στην ταυτοποίηση του είδους από το οποίο προέκυψαν σύντομες ακολουθίες DNA.

Τα μοντέλα n -gram επικρίνονται συχνά επειδή δεν έχουν καμία ρητή αναπαράσταση της εξάρτησης από μεγάλη απόσταση. Αυτό οφείλεται στο γεγονός ότι η μόνη ρητή περιοχή εξάρτησης είναι (n - 1) μάρκες για ένα μοντέλο n -gram, και δεδομένου ότι οι φυσικές γλώσσες ενσωματώνουν πολλές περιπτώσεις απεριόριστων εξαρτήσεων (όπως wh-movement), αυτό σημαίνει ότι ένα n -gram μοντέλο δεν μπορεί η αρχή διακρίνει τις απεριόριστες εξαρτήσεις από τον θόρυβο (επειδή οι συσχετίσεις μεγάλης απόστασης μειώνονται εκθετικά με την απόσταση για οποιοδήποτε μοντέλο Markov). Για το λόγο αυτό, τα μοντέλα n -gram δεν έχουν επηρεάσει σημαντικά τη γλωσσική θεωρία, όπου μέρος του ρητού στόχου είναι να μοντελοποιήσουμε αυτές τις εξαρτήσεις.

Μια άλλη κριτική που έχει γίνει είναι ότι τα μοντέλα της Μάρκοφ της γλώσσας, συμπεριλαμβανομένων των μοντέλων n -gram, δεν καταγράφουν ρητά τη διάκριση απόδοσης / ικανότητας. Αυτό συμβαίνει επειδή τα μοντέλα n-gram δεν έχουν σχεδιαστεί για να μοντελοποιούν τη γλωσσική γνώση καθαυτή και να μην ισχυρίζονται ότι είναι (ακόμη και δυνητικά) πλήρη μοντέλα γλωσσικής γνώσης. Αντίθετα, χρησιμοποιούνται σε πρακτικές εφαρμογές.

Στην πράξη, τα μοντέλα n-gram έχουν αποδειχθεί εξαιρετικά αποτελεσματικά στη μοντελοποίηση γλωσσικών δεδομένων, τα οποία αποτελούν βασικό στοιχείο στις σύγχρονες εφαρμογές στατιστικής γλώσσας .

Οι περισσότερες σύγχρονες εφαρμογές που βασίζονται σε μοντέλα n -gram, όπως εφαρμογές μηχανικής μετάφρασης , δεν βασίζονται αποκλειστικά σε τέτοια μοντέλα. Αντίθετα, τυπικά ενσωματώνουν και Bayesian συμπεράσματα . Τα σύγχρονα στατιστικά μοντέλα αποτελούνται συνήθως από δύο μέρη, μια

προηγούμενη κατανομή που περιγράφει την εγγενή πιθανότητα ενός πιθανού αποτελέσματος και μιας συνάρτησης πιθανότητας που χρησιμοποιείται για να εκτιμηθεί η συμβατότητα ενός πιθανού αποτελέσματος με τα παρατηρούμενα δεδομένα. Όταν χρησιμοποιείται ένα γλωσσικό μοντέλο, χρησιμοποιείται ως μέρος της προηγούμενης διανομής (π.χ. για να μετρήσει την έμφυτη "καλοσύνη" μιας πιθανής μετάφρασης), και ακόμα και τότε δεν είναι συχνά το μόνο συστατικό της διανομής.

Χρησιμοποιούνται επίσης χειροποίητα χαρακτηριστικά διαφόρων ειδών, για παράδειγμα μεταβλητές που αντιπροσωπεύουν τη θέση μιας λέξης σε μια πρόταση ή το γενικό θέμα του λόγου. Επιπλέον, συχνά χρησιμοποιούνται χαρακτηριστικά που βασίζονται στη δομή του δυνητικού αποτελέσματος, όπως είναι οι συντακτικές εκτιμήσεις. Τέτοια χαρακτηριστικά χρησιμοποιούνται επίσης ως μέρος της συνάρτησης πιθανότητας, η οποία χρησιμοποιεί τα παρατηρούμενα δεδομένα. Συμβατικά γλωσσική θεωρία μπορεί να ενσωματωθεί σε αυτά τα χαρακτηριστικά (αν και στην πράξη, είναι σπάνιο που διαθέτει ειδικά για παραγωγική ή άλλες ιδιαίτερες θεωρίες της γραμματικής ενσωματώνονται, όπως στην υπολογιστική γλωσσολογία τείνουν να είναι «αγνωστικιστής» προς επιμέρους θεωρίες της γραμματικής.

4.1.2.4 ***Διαχείριση emoticon***

:)	:-\
:-)	:\
:-D	:-(
:))	:-((
:-))	:(
(:	:o
:D	:O
:)	D:
;)	=/
;-)	=(
XD	:'-(
=]	:\\
;D	:/
:]	:S
:o)	

Τα αρχικά emoticons που χρησιμοποιήθηκαν είναι τα παραπάνω αλλά θεωρήσαμε ανάγκη να προσθέσουμε πολλά περισσότερα έτσι ώστε να μην έχουμε αλλοίωση των αποτελεσμάτων.

4.1.3 Ταξινομητές

Στο τελευταίο κομμάτι
Ταξινομητές διακριτών τιμών:

- EI-oc:
    - Random Forest
    - Logistic Regression
    - LinearSVC
- V-oc
    - Random Forest
    - Logistic Regression
    - LinearSVC
- E-c
    - OneVsRestClassifier με τη χρήση του Logistic Regression

Ταξινομητές συνεχών τιμών:

- V-reg:
    - Linear Regression
    - Support Vector Regression
    - Linear SVR

- EI-reg: ταξινομητές που χρησιμοποιήθηκαν είναι:
  - Linear Regression
  - Support Vector Regression
  - Linear SVR

4.2 Παρεμφερείς προσεγγίσεις με συμμετέχοντες διαγωνισμού

Η αρχική μελέτη της προσέγγισης και υλοποίησης των τεχνικών δεν θα μπορούσε να μην αφορά την μελέτη των αλγορίθμων που ακολούθησαν άλλοι διαγωνιζόμενοι. Σαφώς αφού η γενική αρχιτεκτονική του συστήματος στηριζόταν στη επιβλεπόμενη μηχανική μάθηση, επόμενο βήμα θα ήταν η εκτενής ανάλυση άλλων συστημάτων που θα είχαμε ως υπόδειγμα, ώστε να οδηγηθούμε σε μια περπατημένη οδό και να έχουμε ένα ασφαλές αποτέλεσμα. Ακολούθως θα αναλυθούν οι προσεγγίσεις των ομάδων: SINAI , UWB και SeerNet συμμετέχουσες στο διαγωνισμό.

Αρχικά η SINAI ασχολήθηκε με τις ενότητες της έντασης των εκάστοτε κατηγοριών συναισθημάτων. Ακολούθησε τα εξής βήματα προεπεξεργασίας δεδομένων τμηματοποιώντας σε κατηγορίες χαρακτηριστικών τις λέξεις των κειμένων βάση της βιβλιοθήκης της επεξεργασίας φυσικής γλώσσας, έπειτα αφαιρώντας τις καταλήξεις αποσπώντας τη ρίζα των λέξεων, ακολούθως αφαιρέθηκαν τα stop words και τέλος όλα τα γράμματα μετατράπηκαν σε πεζά. Στο τελευταίο κομμάτι αυτού του επιπέδου για τα EI-oc και EI-reg χρησιμοποιήθηκαν τα λεξικά NRC Affect Intensity and WNA lexicons και για το E-c ομοίως τα λεξικά Emolex and WNA lexicon. Στη συνέχεια, περιγράφεται η μεθοδολογία που χρησιμοποιήθηκε:

EI-oc. Για να εκτελεστεί η ταξινόμηση, ελέγξαν την παρουσία των όρων στο λεξικό και στη συνέχεια πρόσθεσαν την ένταση αυτών των λέξεων ομαδοποιώντας τα με τη συναισθηματική κατηγορία (θυμός, φόβος, θλίψη και χαρά). Το αποτέλεσμα είναι ένα διάνυσμα τεσσάρων τιμών για κάθε λεξικό. Επιπλέον, κάθε όρος αντιπροσωπεύεται ως ένα διάνυσμα χρησιμοποιώντας τη τεχνική απόδοσης βάρους TF-IDF. Η ένωση των διανυσμάτων του λεξικού και η

αναπαράσταση των βαρέων του TF-IDF του όρου χρησιμοποιούνται ως χαρακτηριστικά για την ταξινόμηση εκπαιδεύοντας το σώμα των δεδομένων από τον αλγόριθμο SVM. Επιλέξαμε τον αλγόριθμο SVM, γνωστό ως C-SVC, η τιμή της παραμέτρου C ήταν 1.0 και ο πυρήνας που επιλέχθηκε ήταν γραμμικός.

EI-reg. Στη συγκεκριμένη υποενότητα, ανέλυσαν την παρουσία όρων από το λεξικό και στη συνέχεια προχώρησαν το άθροισμα, τον μέσο όρο και το μέγιστο της έντασης του όρου όπου ομαδοποιούνταν σε κάποια συναισθηματική κατηγορία (θυμός, φόβος, θλίψη και χαρά). Το αποτέλεσμα είναι ένας διάνυσμα δώδεκα τιμών για κάθε λεξικό. Η ένωση των διανυσμάτων του λεξικού και η αναπαράσταση του TF-IDF του κάθε όρου χρησιμοποιούνται ως χαρακτηριστικά για την ταξινόμηση χρησιμοποιώντας τον αλγόριθμο SVM με την ίδια διαμόρφωση όπως αυτή που χρησιμοποιείται στο υποενότητα EI-oc.

E-c. Ανέλυσαν την παρουσία όρων λεξικού στα tweets
και αποδόθηκε η τιμή 1 ως τιμή εμπιστοσύνης (CV). Στη συνέχεια, συνοψίσαν τις τιμές αυτές των λέξεων των οποίων το συναίσθημα είναι όμοιο με την τιμές διανύσματος των συναισθημάτων για κάθε λεξικό. Η ένωση αυτών των διανυσμάτων και η αναπαράσταση των βάρεων του TF-IDF των όρων χρησιμοποιούνται ως χαρακτηριστικά για την ταξινόμηση χρησιμοποιώντας τον αλγόριθμο Random Forest με παράμετρο την τιμή 25 ως αριθμό δέντρων.

Μία ομάδα ακόμα της οποίας η προσέγγιση είχε ομοιότητες με την τεχνική που δημιουργήσαμε και εφαρμόσαμε είναι η UWB. Πρώτον ομοίως το αρχικό στάδιο αφορούσε τη προεπεξεργασία δεδομένων ακολουθώντας τα παρακάτω βήματα:

1. Οι λέξεις τμηματοποιήθηκαν και μετατράπηκαν σε πεζά γράμματα
2. Οι σύνδεσμοι URL αντικαταστάθηκαν
3. Τα ονόματα των χρηστών Twitter (που αρχίζουν με @)
αντικαταστάθηκαν με το @user token
4. Τμήματα των όρων που περιέχουν αλληλουχίες γραμμάτων που συναντιόντουσαν περισσότερο από δύο φορές στη σειρά αντικαταστάθηκαν από δύο εμφανίσεις των γραμμάτων (π.χ. huuuungry μειώνεται σε huungry, looooove
σε loove)
5. Κοινές ακολουθίες λέξεων και emojis διαιρέθηκαν σε δύο τμήματα(π.χ. "
nice: D: D " χωρίστηκε σε δύο τμήματα "nice" και": D: D ")

Όσον αφορά τη δημιουργία χαρακτηριστικών το σύστημα που χρησιμοποίησαν αναλύεται όπως θα αναφέρουμε. Μία περίπτωση είναι λέξεις από n-grams από i έως n (για i = 1, n = 2, unigrams και bigrams χρησιμοποιήθηκαν), δεύτερη περίπτωση για χαρακτήρες ngrams από i σε n (για i = 2, n = 3 χαρακτήρες

όπως bigrams και trigrams χρησιμοποιήθηκαν).Έπειτα χρησιμοποιήθηκαν τα Word Embeddings, που αφορά ένα μέσο όρο λέξεων που αποτελούν ακολουθία λέξεων σε tweets. Και τέλος τα λεξικά συναισθημάτων όπου χρησιμοποιήθηκε η βιβλιοθήκη των Affective Tweets για να εξάγουν χαρακτηριστικά από λεξικά επί του θέματος. Ο αλγόριθμος που χρησιμοποιήθηκε στη επιβλεπόμενη μηχανική μάθηση από τη συγκεκριμένη ομάδα λογίζεται ως ο SVM με ένα L2 κανονικοποιημένο μοντέλο παλινδρόμησης και παράμετρο ρύθμισης C στο 1.

Η τελευταία προσέγγιση η οποία βαδίσαμε σε μεγάλο ποσοστό και η οποία διακατείχε υψηλό ποσοστό ακρίβειας στον διαγωνισμό αφορά την ομάδα SeerNet. Η ανάλυση του συγκεκριμένου μοντέλου πραγματοποιήθηκε στο επίπεδο 3.3.1 όπου αναλύονται λεπτομερώς οι υλοποιήσεις των μοντέλων των κορυφαίων ομάδων.

5

Πειραματικά αποτελέσματα

5.1 Υπολογισμός Ακρίβειας για αποτελέσματα

Από τη μελέτη και ανάλυση του προβλήματος βγαίνει το συμπέρασμα ότι το μοντέλο θέτει πολλαπλούς στόχους, απαιτούνται περισσότερες από μία λειτουργίες για τον υπολογισμό της ακρίβειας. Το μοντέλο προσπαθεί να επιλύσει δύο προβλήματα τα οποία δεν μπορούν να μοιραστούν τις λειτουργίες της ακρίβειας λόγω της εγγενής φύσης του προβλήματος, το ένα είναι πρόβλημα παλινδρόμησης και το άλλο ένα πρόβλημα κατηγοριοποίησης.

5.2 Λειτουργία και ανάλυση υπολογισμού μέτρου απώλειας μοντέλου παλινδρόμησης

Για την υπολογισμό της απώλειας του μοντέλου παλινδρόμησης, το μέσο τετραγωνικό σφάλμα(mean squared error) επιλέχθηκε ως ένας τρόπος βελτιστοποίησης του μοντέλου σε συσχέτιση με την βαθμολογία Pearson. Δεδομένου ότι το μέσο τετραγωνικό σφάλμα επιδιώκει να ελαχιστοποιήσει τη διαφορά μεταξύ της πρόβλεψης και της βαθμολογίας των πραγματικών τιμών

δεδομένων του δοκιμαστικού συνόλου, μία χαμηλή τιμή του μέσου τετραγωνικού σφάλματος θα φέρει το σκορ Pearson πιο κοντά στη τιμή 1 που είναι και η βέλτιστη τιμή. Η λειτουργία υπολογισμού απώλειας διασφαλίζει ότι τα άρθρα χωρίς ετικέτες ταξινόμησης δεν επηρεάζουν τον υπολογισμό των αποτελεσμάτων που εξέρχονται δια του μοντέλου δίνοντας τις προβλεπόμενες τιμές απώλεια μηδέν, άρα μελετάται για τις συνεχόμενες τιμές οι οποίες βρίσκονται στο διάστημα [0.5, 1].

Δεδομένου ότι υπάρχει σημαντική επικάλυψη στο εύρος και στον όγκο των άρθρων, ορισμένα άρθρα επαναχρησιμοποιούνται σε πολλαπλές συναισθηματικές καταστάσεις από το υποτμήμα ένα του διαγωνισμού, οι οποίες με τη σειρά τους επαναχρησιμοποιούνται μία φορά στο υποτμήμα πέντε του διαγωνισμού. Ο πραγματικός αριθμός μοναδικών και "διπλότυπων" άρθρων είναι δύσκολο να επιλυθούν και παρουσιάστηκαν μια πρόκληση τις πρώτες επαναλήψεις του μοντέλου. Οι χρυσές ετικέτες του δοκιμαστικού συνόλου και οι προβλεπόμενες βαθμολογίες παρουσιάζονται στους παρακάτω Πίνακες:

Anger score	Fear score	Joy score	Sadness score
Gold: 0.517	Gold: 0.800	Gold: 0.197	Gold: 0.707
Pred: 0.460	Pred: 0.957	Pred: 0.143	Pred: 0.748

Σύγκριση μεταξύ προβλεφθέτων και αληθών τιμών με χρήση της ακρίβειας

Η πρόβλεψη για τον υπολογισμό βαθμολογίας με στόχο παλινδρόμηση, κρίνεται ως θετική.
“we need to do something. something must bedone!!!!!’ your anxiety is amusing. nothing will be done. despair.”

Anger score	Fear score	Joy score	Sadness score
Gold: 0.953	Gold: 0.621	Gold: -------	Gold: 0.680
Pred: 0.632	Pred: 0.334	Pred: 0.450	Pred: 0.370

Σύγκριση μεταξύ προβλεφθέτων και αληθών τιμών με χρήση της ακρίβειας

Η πρόβλεψη για τον υπολογισμό βαθμολογίας με στόχο παλινδρόμηση, κρίνεται ως μη επιτυχημένη.

”Don’t fucking tag me in pictures as ’family first’ when you cut me out 5 years ago. You’re no one to me.”

Είναι αξιοσημείωτο από τη βαθμολογία που φαίνεται στους παραπάνω πίνακες ότι λέξεις-κλειδιά όπως η ‘family’, ‘anxiety’ και ‘fucking’, έχουν πολύ μεγάλη επίδραση στις τιμές που εξάγαμε από το μοντέλο. Αυτές οι συσχετίσεις λέξεων-κλειδιών ίσως είχαν αντιμετωπιστεί καλύτερα με τη χρήση εξωτερικών δεδομένων, όπως συναισθηματικά λεξικά.

5.3 Λειτουργία και ανάλυση υπολογισμού μέτρου απώλειας μοντέλου κατηγοριοποίησης

Η συνάρτηση υπολογισμού απώλειας για την κατηγοριοποίηση είναι λίγο πιο περίπλοκη αφού όλα τα άρθρα παλινδρόμησης λαμβάνουν ετικέτες συνεχόμενων τιμών, αλλά δεν έχουν όλα τα άρθρα παλινδρόμησης ετικέτες με διακριτές τιμές κατηγοριοποίησης. Εφόσον το μοντέλο έχει συγκεκριμένες διακριτές τιμές ανάλογα με το στόχο που θέτουμε των EI-oc, V-oc, E-c, υπάρχει μια λειτουργία υπολογισμού απώλειας για κάθε μία κατηγορία συναισθημάτων που αναλύουμε και μελετάμε. Η μέθοδος υπολογισμού απώλειας τροποποιήθηκε στο στόχο της κατηγοριοποίησης ώστε να υπολογίσει και να συσχετίσει την άνιση κατανομή άσσων και των μηδενικών. Ο σκοπός του μοντέλου είναι να προσδιορίσει τις τιμές των συναισθημάτων που υποδεικνύονται από ένα tweet ανάμεσα σε ένα συγκεκριμένο διάστημα με υψηλότερη τιμή τη μονάδα.

Για το τελευταίο υποσύστημα καθώς είναι εάν πρόβλημα πολλαπλών ετικετών. Ένα tweet λαμβάνει από καμία μέχρι και έντεκα ετικέτες του δοκιμαστικού συνόλου δεδομένων και παράλληλα οι τιμές που εξάγονται λαμβάνουν από το σύνολο ετικετών που όμοιο με του δοκιμαστικού συνόλου δηλαδή από μηδέν μέχρι και έντεκα. Τα μοντέλα αξιολογούνται με τον υπολογισμό της ακρίβειας πολλαπλών ετικετών, του δείκτη Jaccard, ο οποίος υπολογίζεται με τον ακόλουθο τύπο:

$$Accuracy = \frac{1}{|T|} \sum_{t \in T} \frac{|G_t \cap P_t|}{|G_t \cup P_t|}$$

όπου Gt το σύνολο των χρυσών ετικετών του δοκιμαστικού συνόλου δεδομένων για τον όρο t, Pt το σύνολο των προβλεπόμενων ετικετών για τον όρο t, και T είναι το σύνολο των άρθρων.

Λαμβάνοντας υπόψη ότι οι ετικέτες κατηγοριοποίησης αντιπροσωπεύουν τα συναισθήματα ο θυμός, η πρόβλεψη, η αηδία, ο φόβος, η χαρά, η αγάπη, η αισιοδοξία, η απαισιοδοξία, η θλίψη, η έκπληξη και η εμπιστοσύνη, οι υποδειγματικές προβλέψεις κατηγοριοποίησης παρουσιάζονται στους παρακάτω πίνακες.

Προβλέψεις	Ποσοστό ακρίβειας
Gold: 0 0 1 0 0 0 0 0 1 0 0	82,00%
Pred: 0 1 1 0 0 0 0 0 0 0 0	

υπολογισμός ακρίβειας (accuracy)

Βαθμολογία για το 5 υποτμήμα κατηγοριοποίησης της εργασίας: “Not sure tequila shots at my family birthday meal is up there with the best ideas I’ve ever had#grim”

Προβλέψεις	Ποσοστό ακρίβειας
Gold: 0 1 0 0 1 1 1 0 0 0 1	36,00%
Pred: 1 0 1 0 0 0 0 0 0 0 0	

υπολογισμός ακρίβειας (accuracy)

Μη αποδεκτή βαθμολογία για το 5 υποτμήμα κατηγοριοποίησης της εργασίας: ”SheenKL I assume the manga is #good?”

5.4 Ανάλυση συστήματος και αποτελεσμάτων

Σε αυτή την ενότητα περιγράφονται τα συστήματα που αναπτύχθηκαν για τα υποτμήματα EI-oc, EI-reg , V-oc, V-reg και E-c. Προεπεξεργαστήκαμε το σύνολο δεδομένων των άρθρων που συναντηθηκαν για κάθε υποτμήμα. Εφαρμόσαμε τα ακόλουθα βήματα προεπεξεργασίας: τα έγγραφα τμηματοποιήθηκαν λεκτικά με τη χρήση του NLTK TweetTokenizer, η απομάκρυνση καταλήξεων πραγματοποιήθηκε χρησιμοποιώντας το NLTK Snowball stemmer, οι λέξεις κλειδιά καταργήθηκαν και τέλος όλα τα γράμματα μετατράπηκαν σε πεζά.

Υποτμήμα EI-oc. Για να εκτελέσουμε την κατηγοριοποίηση, ελέγξαμε την παρουσία των όρων στο tweet και στη συνέχεια προσθέσαμε την ένταση αυτών των λέξεων που τις ομαδοποιούν με τη συναισθηματική κατηγορία (θυμός, φόβος, θλίψη και χαρά). Κάθε όρος αντιπροσωπεύεται με τις τεχνικές TF-IDF, Count Vectorizer την οποία παραμετροποιήσαμε με τη μεθοδο n-grams για n 1 και 3 λέξεων ως φορέας διανυσματοποίησης. Η ένωση και η αναπαράσταση TF-IDF των άρθρων χρησιμοποιούνται ως χαρακτηριστικά για την κατηγοριοποίηση χρησιμοποιώντας τους ακόλουθους αλγορίθμους. Επιλέξαμε τη συνταγή του SVM, γνωστή ως linearSVC, του αλγόριθμου Random forrest καθώς και του αλγόριθμου Logistic Regression. Η ομάδα η οποία έγινε η σύγκριση των αποτελεσμάτων ονομάζεται SeerNet και διακατείχε εκ των πρώτων θέσεων του διαγωνισμού υλοποιώντας τα μοντέλα της με αλγορίθμους νευρωνικών δικτύων και λεξικών λέξεων αλλά και συναισθημάτων. Η δικιά μας παραδοσιακή προσέγγιση παρατηρούμε ο κινείται στα ίδια επίπεδα απόδοσης διότι εμφανίσαμε ένα βελτιωμένα σύνολα δεδομένα εκπαίδευσης και έτσι καταφέραμε ανάλογα αποτελέσματα.

EI-OC φόβος	TF-IDF(n-grams=1,3)	Count Vectorizer(n-grams=1,3)
Random Forrest	0.56	0.53
Logistic Regression	5.34	0.47
LinearSVC	0.47	0.52

Αποτελέσματα για φόβο σε συσχέτιση με το Pearson

EI-OC θυμός	TF-IDF(n-grams=1,3)	Count Vectorizer(n-grams=1,3)
Random Forrest	0.54	0.57
Logistic Regression	0.41	0.50
LinearSVC	0.51	0.51

Αποτελέσματα για θυμό με Pearson

Count Vectorizer(n-grams=1,1)	TF-IDF(n-grams=1,1)	SeerNet
0.56	0.61	0.45
0.47	0.35	
0.54	0.58	

Αποτελέσματα για θυμό με Pearson

EI-OC χαρά	TF-IDF(n-grams=1,3)	Count Vectorizer(n-grams=1,3)
Random Forrest	0.61	0.63
Logistic Regression	0.52	0.51
LinearSVC	0.53	0.54

Αποτελέσματα για χαρά χρήση Pearson για συσχέτιση με Gold Labels

Count Vectorizer(n-grams=1,1)	TF-IDF(n-grams=1,1)	SeerNet
0.66	0.68	0.55
0.59	0.56	
0.56	0.57	

Αποτελέσματα για χαρά, σύγκριση με την ομάδα SeerNet, έκδοση ακρίβειας με Pearson

Count Vectorizer(n-grams=1,1)	TF-IDF(n-grams=1,1)	SeerNet
0.64	0.64	0.61
0.64	0.61	
0.59	0.59	

Αποτελέσματα για θυμό, σύγκριση με την ομάδα SeerNet, υπολογισμός ακρίβειας με το μοντέλο συσχετίσων Pearson

EI-OC λύπη	TF-IDF(n-grams=1,3)	Count Vectorizer(n-grams=1,3)
Random Forrest	0.53	0.55
Logistic Regression	0.43	0.47
LinearSVC	0.47	0.49

Αποτελέσματα για λύπη με Pearson

Count Vectorizer(n-grams=1,1)	TF-IDF(n-grams=1,1)	SeerNet
0.58	0.61	0.56
0.55	0.55	
0.52	0.53	

Αποτελέσματα για λύπη, σύγκριση με την ομάδα SeerNet εξαγωγή αποτελεσμάτων με μοντέλο Pearson

Υποτμήμα EI-reg. Σε αυτή την περίπτωση, ελέγξαμε την παρουσία λέξεων στο twitter και έπειτα υπολογίσαμε το άθροισμα, τον μέσο όρο και το μέγιστο της έντασης των λέξεων του άρθρου κατηγοριοποιώντας τους από τη συναισθηματική κατηγορία (θυμός, φόβος, θλίψη και χαρά).. Η ένωση των τεχνικών TF-IDF, Count Vectorizer με τη μεθοδολογία των n-grams υλοποιήθηκε για την αναπαράσταση του άρθρου χρησιμοποιήθηκαν ως χαρακτηριστικά για την κατηγοριοποίηση χρησιμοποιώντας τον αλγόριθμο SVM και Linear Regression με την ίδια διαμόρφωση με εκείνη που χρησιμοποιείται στο υποτμήμα της εργασίας EI-oc. Η προσέγγιση όπως αναφέραμε παραπάνω πέτυχε ανάλογα αποτελέσματα λόγω της επικείμενης πρόσθεσης δεδομένων στα δοκιμαστικά σύνολα δεδομένων με αποτέλεσμα την αύξηση των εκπαιδευτικών.

EI-reg φόβος	TF-IDF(n-grams=1,3)	Count Vectorizer(n-grams=1,3)
Linear Rergression	0.54	0.59
Support Vector Regression	0.52	0.44
Linear SVR	0.47	0.61

Αποτελέσματα για φόβο συσχετίζοντας με την ακρίβεια(mean squared error) Gold Labels μέσω Pearson

Count Vectorizer(n-grams=1,1)	TF-IDF(n-grams=1,1)	SeerNet
0.56	0.5	0.6
0.51	0.39	
0.54	0.49	

Αποτελέσματα για φόβο, σύγκριση με την ομάδα SeerNet συσχετίζοντας με την ακρίβεια Gold Labels μέσω Pearson

EI-reg θυμός	TF-IDF(n-grams=1,3)	Count Vectorizer(n-grams=1,3)
Linear Rergression	0.6	0.68
Support Vector Regression	0.43	0.58
Linear SVR	0.65	0.67

Αποτελέσματα για θυμό συσχετίζοντας με την ακρίβεια(mean squared error) Gold Labels μέσω Pearson

Count Vectorizer(n-grams=1,1)	TF-IDF(n-grams=1,1)	SeerNet
0.51	0.58	0.7
0.44	0.57	
0.68	0.65	

Αποτελέσματα για θυμό, σύγκριση με την ομάδα SeerNet συσχετίζοντας με την ακρίβεια(mean squared error) Gold Labels μέσω Pearson

EI-reg χαρά	TF-IDF(n-grams=1,3)	Count Vectorizer(n-grams=1,3)
Linear Rergression	0.53	0.54
Support Vector Regression	0.45	0.5
Linear SVR	0.61	0.64

Αποτελέσματα για χαρά συσχετίζοντας με την ακρίβεια(mean squared error) Gold Labels μέσω Pearson

Count Vectorizer(n-grams=1,1)	TF-IDF(n-grams=1,1)	SeerNet
0.58	0.46	0.56
0.51	0.48	
0.61	0.59	

Αποτελέσματα για χαρά, σύγκριση με την ομάδα SeerNet συσχετίζοντας με την ακρίβεια(mean squared error) Gold Labels μέσω Pearson

EI-reg λύπη	TF-IDF(n-grams=1,3)	Count Vectorizer(n-grams=1,3)
Linear Rergression	0.59	0.63
Support Vector Regression	0.54	0.47
Linear SVR	0.67	0.65

Αποτελέσματα για λύπη συσχετίζοντας με την ακρίβεια(mean squared error) Gold Labels μέσω Pearson

Count Vectorizer(n-grams=1,1)	TF-IDF(n-grams=1,1)	SeerNet
0.52	0.59	0.66
0.48	0.55	
0.65	0.66	

Αποτελέσματα για λύπη, σύγκριση με την ομάδα SeerNet συσχετίζοντας με την ακρίβεια(mean squared error) Gold Labels μέσω Pearson

Υποτμήμα E-c. Στο πέμπτο υποτμήμα του διαγωνισμού όπως μελετήσαμε, προχωρήσαμε στη διανυσματοποίηση με τις σχετικές τεχνικές της προεπεξεργασίας των άρθρων και ο υπολογισμός της ακρίβειας έγινε από τον δείκτη Jaccard όπως εξηγήθηκε και αναλύθηκε παραπάνω. Με τη χρησιμοποίηση της λογικής παλινδρόμησης και έπειτα με τον αλγόριθμο πολλαπλών ετικετών ένα προς υπόλοιπα εκτιμήθηκε με δυαδική κατηγοριοποίηση εάν το κάθε άρθρο ανήκει σε καμία ή πολλαπλές συναισθηματικές κατηγορίες. Η ομάδα η οποία έγινε σύγκριση αποτελεσμάτων ήταν η ομάδα NTUA-SLP και παρακάτω αντιπαραθέτουμε τα αποτελέσματα από το νέο βελτιωμένο σύνολο εκπαιδευτικών δεδομένων μας.

TF-IDF(n-grams=1,1)	NTUA-SLP
0.54	0.588

Σύγκριση αποτελέσματος με την ομάδα NTUA-SLP με χρήση μοντέλου ακριβείας Jaccard

E-c one vs rest	TF-IDF(n-grams=1,3)	Count Vectorizer(n-gr	Count Vectorizer(n-grams=1,1)
Logistic Regression	0.63	0.65	0.61

Αποτελέσματα E-c με μοντέλο Jaccard

Υποτμήμα V-oc. Για την εκτέλεση της ταξινόμησης, όπως αναφέραμε και στο τμήμα EI-oc αναλύθηκε η παρουσία των όρων στο twitter και στη συνέχεια προσθέσαμε την ένταση αυτών των λέξεων που τις ομαδοποιούν ανάλογα το σθένος δηλαδή την ένταση του συναισθήματος σε 7 συναισθηματικές τάξεις. Κάθε όρος

αναλύεται με τις τεχνικές TF-IDF, Count Vectorizer την οποία παραμετροποιήσαμε με τη μέθοδο n-grams για 1 και 3 λέξεων ως φορέας διανυσματοποίησης. Η ένωση και η αναπαράσταση TF-IDF των άρθρων χρησιμοποιούνται ως χαρακτηριστικά για την κατηγοριοποίηση. Επιλέξαμε τους αλγορίθμους SVM, γνωστή ως linearSVC, του SVC, του αλγόριθμου Random forest καθώς και του αλγόριθμου Logistic Regression. Η ομάδα η οποία έγινε η σύγκριση των αποτελεσμάτων ονομάζεται SeerNet και διακατείχε εκ των πρώτων θέσεων του διαγωνισμού υλοποιώντας τα μοντέλα της με αλγορίθμους όπως είπαμε νευρωνικών δικτύων. Η δικιά μας προσέγγιση παρατηρούμε ότι παρότι έχουμε νέα βελτιωμένα σύνολα δεδομένα εκπαίδευσης τα αποτελέσματα το συγκεκριμένο τμήμα του διαγωνισμού δεν είναι ανάλογα με των κορυφαίων ομάδων στο διαγωνισμό.

V-oc	TF-IDF(n-grams=1,3)
Random Forrest	0.7
Logistic Regression	0.57
SVC	0.52
LinearSVC	0.59

Αποτελέσματα V-oc συσχετίζοντας με την ακρίβεια Gold Labels μέσω Pearson

Count Vectorizer(n-grams=1,3)	Count Vectorizer(n-grams=1,1)	TF-IDF(n-grams=1,1)	SeerNet
0.66	0.76	0.78	0.83
0.67	0.71	0.65	
0.66	0.76	0.68	
0.65	0.74	0.68	

Σύγκριση αποτελεσμάτων με την ομάδα SeerNet συσχετίζοντας με την ακρίβεια Gold Labels μέσω Pearson

Υποτμήμα V-reg. Η συγκεκριμένη περίπτωση, ελέγχει την σπανιότητα τους καθώς και τη συχνοτήτων λέξεων στο κάθε tweet και έπειτα υπολογίσαμε το άθροισμα, τον μέσο όρο και το μέγιστο της έντασης των λέξεων του άρθρου αποδίδοντας τιμές από ένα σύνολο συνεχών τιμών. Η ένωση των τεχνικών TF-IDF, Count Vectorizer με τη μεθοδολογία των n-grams υλοποιήθηκε για την αναπαράσταση του άρθρου χρησιμοποιήθηκαν ως χαρακτηριστικά για την κατηγοριοποίηση χρησιμοποιώντας τον αλγόριθμο SVM και Linear Regression με

την ίδια διαμόρφωση με εκείνη που χρησιμοποιείται στο υποτμήμα της εργασίας V-oc. Αν αναλογιστούμε την επικείμενο μετασχηματισμό των εκπαιδευτικών και δοκιμαστικών δεδομένων παρατηρούμε ότι οι τεχνικές που εφαρμόσαμε

V-reg	TF-IDF(n-grams=1,3)
Linear Regression	0.58
SVR	0.58
LinearSVR	0.62

Αποτελέσματα για V-reg συσχετίζοντας με την ακρίβεια(mean squared error) Gold Labels μέσω Pearson

παραμένουν αδύναμες σε συσχέτιση με τις κυρίαρχες μεθοδολογίες που

Count Vectorizer(n-grams=1,3)	Count Vectorizer(n-grams=1,1)	TF-IDF(n-grams=1,1)	SeerNet
0.63	0.5	0.58	0.55
0.56	0.57	0.61	
0.62	0.57	0.6	

Σύγκριση αποτελεσμάτων με την ομάδα SeerNet συσχετίζοντας με την ακρίβεια(mean squared error) Gold Labels μέσω Pearson

αναπτύσσονται από άλλες ομάδες.

6

Συμπεράσματα - Προοπτικές

6.1.1 Εφαρμογές

Στα παραπάνω παραρτήματα έγινε η αναφορά και αναλύθηκε η προσέγγιση μας στον διαγωνισμό. Από την αρχή υπήρχε η ανάγκη δημιουργίας πραγματικών εφαρμογών που θα υλοποιούσαν την αναγνώριση συναισθηματικής κατάστασης στην πράξη. Τη σημερινή εποχή υπάρχουν πάρα πολλές εφαρμογές που μπορεί κάποιος να της βρει ελεύθερα στο ίντερνετ ή να της αγοράσει. Και πάρα πολύς κόσμος χρησιμοποιεί εφαρμογές αναγνώρισης συναισθηματικής κατάστασης που συνήθως συνοδεύονται με άλλες εφαρμογές επεξεργασίας φυσικής γλώσσας NLP. Κάθε μια χρησιμοποιεί τις δικές της μεθόδους, αλγορίθμους και χαρακτηριστικά για να πετύχει το καλύτερο δυνατό αποτέλεσμα. Συνήθως η διαδικασία που ακολουθούν είναι κάτι που δεν φανερώνεται με λεπτομέρειες, αποτελεί τη τεχνογνωσία του κατασκευαστή. Οι εφαρμογές της αναγνώρισης συναισθηματικής κατάστασης που συνήθως συναντάμε αφορούν κυρίως δύο τομείς. Τις ελεύθερες εφαρμογές και τις επιχειρηματικές εφαρμογές.

6.1.2 Ανακεφαλαίωση

Η παρούσα εργασία ασχολήθηκε με τη δημιουργία μοντέλων και αλγορίθμων ανίχνευσης συναισθημάτων σε ένα από τα πιο διαδεδομένα μέσα κοινωνικής δικτύωσης όπως είναι το twitter. Αρχικά έγινε η ανάλυση της εξόρυξης πληροφορίας από κείμενο καθώς και μία εισαγωγή στο διαγωνισμό τον οποίο συμμετείχαμε SemEval-2018, τον δρόμο τον οποίο θα βαδίσουμε. Στο επόμενο παράρτημα έγινε μια λεπτομερής ανάλυση όλων των μεθόδων και τεχνικών που έχουν δημιουργηθεί μέχρι την παρούσα χρονική περίοδο με τη μεθοδολογία της επιβλεπόμενης μηχανικής μάθησης σε πρώτο σκέλος και σε δεύτερο έγινε η λεπτομερής αναφορά στο αντικείμενο του διαγωνισμού καθώς αναλύθηκε εξονυχιστικά το κάθε κομμάτι του. Στο τρίτο παράρτημα μπαίνουμε στο κύριο μέλος της εργασίας που αναφέρεται στη δικιά μας προσέγγιση, δηλαδή το μοντέλο που δημιουργήσαμε καθώς αναφέρονται όλοι οι κανόνες και τεχνικές που χρησιμοποιήθηκαν. Ακολούθως πραγματοποιήθηκε το πειραματικό σκέλος της προσέγγισης που υλοποιήσαμε σε συνδυασμό με την μελέτη των στόχων και των δεδομένων τους οποίους ακολουθήσαμε. Τα αποτελέσματα είναι ενθαρρυντικά από τα αποτελέσματα τα οποία αναρτήθηκαν συγκριτικά με τις καλύτερες ομάδες που συμμετείχαν στο διαγωνισμό. Τέλος να αναφερθούν κάποια ακόμα συμπεράσματα καθώς και κάποιες άλλες μεθοδολογίες που μπορούν στο μέλλον να οδηγήσουν καλύτερα αποτελέσματα.

6.1.3 Συμπεράσματα γενικής θεματολογίας

Η αναγνώριση συναισθηματικής κατάστασης είναι ένα πάρα πολύ σημαντικό εργαλείο για την ανίχνευση απόψεων. Αυτό γίνεται εμφανές από την πληθώρα των χρήσεων, των τεχνικών, των μεθόδων και των εφαρμογών που υπάρχουν τη σημερινή εποχή. Υπάρχουν προβλήματα που δεν μας επιτρέπουν να κάνουμε σωστά αναγνώριση συναισθηματικής κατάστασης. Αυτά είναι προβλήματα που απορρέουν από τη φύση των κειμένων όπως, το μήκος κειμένου, το πολυγλωσσικό περιεχόμενο, το θόρυβο κτλ. Επιπλέον, το συναίσθημα μπορεί πολλές φορές να εκφρασθεί με πιο λεπτό/έμμεσο τρόπο χωρίς τη χρήση συναισθηματικά φορτισμένων λέξεων. Επιπρόσθετα, ο εκφραστής της άποψης που διατυπώνεται στο κείμενο μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα. Ένα άλλο ζήτημα είναι ότι το συναίσθημα και η υποκειμενικότητα ενός κειμένου εξαρτώνται από το σημασιολογικό πλαίσιο στο οποίο τοποθετούνται. Σημαντικό επίσης πρόβλημα είναι ότι μπορεί να προκύψει διαφορετικό συναίσθημα από τη διαφορετική σειρά των λέξεων και των φράσεων στο κείμενο. Τέλος, υπάρχουν δυσκολίες που προέρχονται από την γενικότερη περιοχή της Επεξεργασίας Φυσικής Γλώσσας,

όπως, η αμφισημία, ο χειρισμός της άρνησης, η ειρωνεία και ο σαρκασμός. Υπάρχουν όμως και προβλήματα που προέρχονται αποκλειστικά από την τεχνική που εφαρμόζουμε για την υλοποίηση της ανάλυσης, τα οποία διακρίνονται παρακάτω.

Προσεγγίσεις με λεξικά :
• Το πλήθος των λέξεων είναι πεπερασμένο, σε αντίθεση με τις ανάγκες για ανάλυση σε ένα δυναμικό μέσο.
• Λέξεις που μπορούν να έχουν διαφορετικό συναίσθημα σε διαφορετικά περιεχόμενα.
• Η συναισθηματική πολικότητα μπορεί να είναι άμεσα σχετιζόμενη με τη θεματολογία.

Προσεγγίσεις με επιβλεπόμενη μηχανική μάθηση :
• Το κόστος.
• Η αδυναμία να εκπαιδευτούν οι ταξινομητές σε περισσότερα από ένα πεδία.
• Ο πολύ μεγάλος χρόνος και κόπος που απαιτείται για την κατασκευή του συνόλου εκπαίδευσης του ταξινομητή. Η δυσκολία στην εύρεση αντιπροσωπευτικού δείγματος για την κατάλληλη εκπαίδευση του ταξινομητή. Παρά τα παραπάνω προβλήματα έχουν βρεθεί τρόποι να ξεπεραστούν αποδοτικά με τις τεχνικές να ανταγωνίζονται μεταξύ τους, παρατηρούμε ότι κάθε μια να έχει τα δικά της πλεονεκτήματα σε σχέση με την άλλη. Οι επιβλεπόμενες προσεγγίσεις μηχανικής μάθησης έχουν πολύ καλά αποτελέσματα καθώς επιτυγχάνουν πολύ καλά ποσοστά ακρίβειας που υπερτερούν σε πολλές περιπτώσεις από τις
τεχνικές μη-επιβλεπόμενης μάθησης. Αυτές οι μέθοδοι μπορούν να πετύχουν 80%-84% ακρίβεια (accuracy). Πράγματι από τα στοιχεία που συγκεντρώσαμε επαληθεύεται ότι για επιβλεπόμενη μηχανική μάθηση η ακρίβεια είναι περίπου 79,97%. Οι τεχνικές βασισμένες σε λεξικά (lexicon-based) έχουν επίσης κάποια πολύ σημαντικά πλεονεκτήματα, γιατί δεν χρειάζονται εκπαίδευση, και είναι πιο κατάλληλες για μεγάλο εύρος περιεχομένων, μιας και στηρίζονται σε προκατασκευασμένα λεξικά από λέξεις με συγγενική συναισθηματική κατεύθυνση. Οι υβριδικές προσεγγίσεις εκμεταλλεύονται τα πλεονεκτήματα και προσπαθούν να αποσοβήσουν τα μειονεκτήματα των μεθόδων που εμπεριέχουν, διαπιστώσαμε οτι πετυχαίνουν πολύ καλά αποτελέσματα σε ποσοστό περίπου 77,18%. Στη μη-επιβλεπόμενη μηχανική μάθηση, σε αντίθεση με τις λεξικολογικές μεθόδους και τις μεθόδους επιβλεπόμενης μηχανικής μάθησης, το σύστημα τροφοδοτείται μόνο από εισόδους, και καλείται να κάνει την αναγνώριση συναισθηματικής κατάστασης χωρίς ανάγκη από σύνολα ήδη
χαρακτηρισμένα. Έτσι, επιτυγχάνουν και αυτές ικανοποιητικά αποτελέσματα, με καλά ποσοστά ακρίβειας όταν εφαρμόζονται σε γνωστά θεματικά πεδία, όπου το λεξιλόγιο των κειμένων τους καλύπτεται από τα λεξικά συναισθήματος. Τα

αποτελέσματα για τη μη επιβλεπόμενη μάθηση είδαμε ότι μπορούν να φτάσουν την ακρίβεια σε ποσοστό περίπου 65,66%. Το συμπέρασμα που βγαίνει μελετώντας και πολλά παραδείγματα εφαρμογών είναι ότι δεν υπάρχει μια μοναδική κατάλληλη συνταγή για αναγνώριση συναισθηματικής κατάστασης, αλλά εξαρτάται από την εκάστοτε περίπτωση και τα διάφορα δεδομένα που καλούμαστε να αναλύσουμε. Τα καλύτερα αποτελέσματα συνήθως έρχονται με τον κατάλληλο συνδυασμό των τεχνικών και των μεθόδων.

6.1.4 Συμπεράσματα διαγωνισμού

Το πρώτο και πιο σημαντικό κομμάτι το οποίο θα έπρεπε να αναλυθεί είναι ο παραμελισμός των παραδοσιακών αλγορίθμων και μεθόδων της επιβλεπόμενης μηχανικής μάθησης και η χρήση των νευρωνικών δικτύων για τον σχεδιασμό των μοντέλων από τις περισσότερες ομάδες του διαγωνισμού. Ο σκοπός που θέσαμε σαν ομάδα ήταν η πλήρης κατανόηση και υλοποίηση της μεθοδολογιών της κλασσικής επιβλεπόμενης μηχανικής μάθησης. Ακόμη και με την ανάπτυξη των εκπαιδευτικών συνόλων, δεδομένων παρατηρήσαμε ότι σε κάποια τμήματα του διαγωνισμού δεν καταφέραμε να προσεγγίσουμε τις κορυφαίες βαθμολογίες, πράγμα που φανερώνει την ανάγκη σαν ερευνητικό σκέλος, το επόμενο βήμα να είναι η ανάπτυξη της αναγνώρισης συναισθηματικής κατάστασης με τους νέους αλγόριθμους των νευρωνικών δικτύων.

6.1.5 Άνθρωποι ή μηχανές

Η απάντηση στο ερώτημα που θέτει η αναγνώριση συναισθηματικής κατάστασης είναι κάτι που δυσκολεύει τις μηχανές, αλλά εξίσου και τους ανθρώπους. Το να αποφανθεί ο άνθρωπος για το θετικό, αρνητικό ή ουδέτερο κρύβει δυσκολίες και παγίδες. Οι άνθρωποι συχνά δεν μπορούν να αποφασίσουν το συναίσθημα ενός κειμένου επειδή, όπως και οι μηχανές, δεν έχουν την αντίστοιχη γνώση για να το κάνουν. Για παράδειγμα δύο άνθρωποι με διαφορετικό γνωστικό υπόβαθρο, διαφορετικές εμπειρίες και διαφορετικά σημεία αναφοράς μπορεί να κάνουν μια διαφορετική ανάλυση για το ίδιο θέμα. Που υστερεί η αναγνώριση συναισθηματικής κατάστασης από ανθρώπους. Όπως είπαμε οι άνθρωποι μπορεί να βρεθούν σε δύσκολη θέση για να αποφανθούν για κάποιο συναίσθημα. Αλλά το πρόβλημα μπορεί να είναι πολύ μεγαλύτερο όταν πρέπει να αποφανθούν για

τεράστιο όγκο δεδομένων σε πολύ μικρό χρονικό διάστημα. Ο άνθρωπος είναι ικανός για συνεπείς και ακριβείς εκτιμήσεις κειμένων πάνω σε περιορισμένο αριθμό κατηγοριών και σε γνωστό πεδίο, αλλά όχι για πολύ μεγάλο όγκο δεδομένων.

6.1.6 Που μπορούν να βοηθήσουν οι μηχανές

Οι μηχανές επειδή μπορούν να μειώσουν τόσο το χρόνο όσο και να διαχειριστούν καλύτερα τον όγκο της εργασίας που έχουν να κάνουν, αποτελούν το πιο χρήσιμο εργαλείο της αναγνώρισης συναισθηματικής κατάστασης. Απαιτούν όμως ακρίβεια στα αποτελέσματα. Το μυστικό της αποτελεσματικότητας της αυτόματης αναγνώρισης συναισθηματικής κατάστασης είναι να καταλάβουμε τις επικίνδυνες περιοχές αυτής, οι οποίες είναι, η εξάρτηση περιεχομένου και η εξάρτηση χρόνου. Εξάρτηση περιεχομένου, σημαίνει ότι ο ταξινομητής που είναι σχεδιασμένος να ταξινομεί ένα θέμα δεν μπορεί να ταξινομήσει επίσης καλά κάποιο άλλο. Για παράδειγμα όταν ταξινομεί κριτικές για ποτήρια δεν θα τα πάει καλά σε ένα πολιτικό ντιμπέιτ. Η χρονική εξάρτηση αναφέρεται στο πότε ένας ταξινομητής γίνεται μη αποδοτικός μετά από πέρασμα κάποιας χρονικής περιόδου. Το λεξιλόγιο του θέματος μπορεί να έχει αλλάξει τόσο που ο ταξινομητής δεν μπορεί πια να "καταλάβει" τα δεδομένα όπως παλιά. Πως να αντιμετωπίσουμε τα προβλήματα των συστημάτων ανάλυσης. Για να είναι ακριβής ένας ταξινομητής συναισθήματος πρέπει να είναι συνεπής στην εξάρτηση θεματολογίας, και στην εξάρτηση χρόνου όντας συγκεκριμένος και σύγχρονος. Δεν πρέπει να στηριζόμαστε σε ένα και μόνο ταξινομητή που κάνει για όλα, αλλά πρέπει να χρησιμοποιούμε τον καταλληλότερο κάθε φορά και τον πιο ανανεωμένο. Υπάρχουν δύο τρόποι που μπορούν τα συστήματα αναγνώρισης συναισθηματικής κατάστασης να χτιστούν και να συντηρηθούν. Ο ένας είναι βασισμένος στη γνώση/γλωσσικές πηγές και ο άλλος στη μηχανική μάθηση. Όσον αφορά στη μηχανική μάθηση που είναι η πιο συνηθισμένη μέθοδος πρέπει να είμαστε σίγουροι ότι οι ταξινομητές είναι σωστά εκπαιδευμένοι. Η δυσκολία έγκειται στο κάθε πότε πρέπει να ενημερώνουμε τους ταξινομητές. Αυτό δεν είναι μια εύκολη απόφαση, διότι χρειάζεται πολύς χρόνος και χρήματα για αυτή την συντήρηση και τις αλλαγές στον ταξινομητή, και δεν υπάρχει κανένας κανόνας που μπορεί κάποιος να ακολουθήσει για να αντιμετωπίσει αυτό το πρόβλημα. Η αυτοματοποιημένη ανάλυση συναισθήματος είναι ένα εργαλείο και σαν όλα τα εργαλεία έχει κάποια όρια και κάποιες βέλτιστες μεθόδους και απαιτεί συντήρηση για να εξασφαλίσει το καλύτερο αποτέλεσμα. Φυσικά δεν μπορεί να αντικαταστήσει την ανθρώπινη ανάλυση συναισθήματος, και δεν γίνεται να πούμε ότι η αυτόματη ανάλυση συναισθήματος μπορεί να σταθεί από μόνη της και ότι ο ανθρώπινος παράγοντας δεν είναι απαραίτητος. Η αυτοματοποιημένη ανάλυση συναισθήματος πρέπει να θεωρείται σαν ένα πολύ καλό εργαλείο που λειτουργεί

συμπληρωματικά στην ανθρώπινη ανάλυση, διότι και οι δύο έχουν τα πλεονεκτήματα και τα μειονεκτήματά τους.

6.1.7 Επόμενα βήματα

Ένας νέος τομέας με ποιο προχωρημένες λύσεις στον τομέα της αυτόματης αναγνώρισης συναισθηματικής κατάστασης προχωρούν ένα βήμα παραπέρα προσπαθώντας να εντοπίσουν συναίσθημα, όχι μόνο στο κείμενο, αλλά και σε φωτογραφίες και βίντεο. Στο μεθοδολογικό κομμάτι, μερικά συστήματα συνδέουν το συναίσθημα με εγγραφές, όπως τις πωλήσεις, τις έρευνες, τις πληρωμές κτλ, συμπεριλαμβανομένου του συσχετισμού θέσης που μας οδηγεί προς έναν κόσμο ολοκληρωμένων αναλύσεων.

Βιβλιογραφία

Hotho, A., Nurnberger, A. And Paaß, G. (2005). "Μια σύντομη έρευνα για εξόρυξη κειμένου". Στο Ldv Forum, Vol. 20 (1), σελ. 19-62

Lovins, J.B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics. 11, 22-31.

The Porter Stemming Algorithm, ανακτήθηκε 13 Μαρτίου 2016. https://tartarus.org/martin/PorterStemmer/

Mohammad, S. M., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). Semeval-2018 Task 1: Affect in tweets. In Proceedings of International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA,USA

Saif M. Mohammad and Svetlana Kiritchenko. 2018.Understanding emotions: A dataset of tweets tostudy interactions between affect categories.InProceedings of the 11th Edition of the LanguageResources and Evaluation Conference, Miyazaki,Japan.

Saif M Mohammad, Svetlana Kiritchenko, and Xiao-dan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets.arXivpreprint arXiv:1308.6242

Saif M Mohammad. 2017.Word affect intensities.arXiv preprint arXiv:1704.08798.Saif M Mohammad and Felipe Bravo-Marquez. 2017.Wassa-2017 shared task on emotion intensity.arXivpreprint arXiv:1708.03700

Vipin Kumar and Sonajharia Minz. 2016. Multi-view ensemble learning: an optimal feature set partition-ing for high-dimensional data classification. Knowledge and Information Systems, 49(1):1–59.

SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in TweetsVenkatesh Duppada, Royal Jain and Sushant Hiray

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 128–132New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational LinguisticsSINAI at SemEval-2018 Task 1: Emotion Recognition in TweetsFlor Miriam Plaza-del-Arco, Salud Marıa Jimenez-Zafra,M. Teresa Mart´ın-Valdivia, L. Alfonso Ure˜na-L´opezDepartment of Computer Science, Advanced Studies Center in ICT (CEATIC)Universidad de Ja en, Campus Las Lagunillas, 23071, Jaen, Spain

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 133–140New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational LinguisticsUWB at SemEval-2018 Task 1: Emotion Intensity Detection in TweetsPavel P˘rib´a˘n1,2,Tom´a˘s Hercig1,2, andLadislav Lenc11NTIS – New Technologies for the Information Society,Faculty of Applied Sciences, University of West Bohemia, Czech Republic2Department of Computer Science and Engineering,Faculty of Applied Sciences, University of West Bohemia, Czech Republic

Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. 2012.Emoticon smoothed language models for twittersentiment analysis. InAaai.

Albert Mehrabian. 1980. Basic dimensions for a gen-eral psychological theory implications for personal-ity, social, environmental, and developmental stud-ies.

Bravo-Marquez et al. (2013) Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. 2013. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, page2. ACM.

Isabelle Augenstein and Anders Søgaard. 2018. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In Proceedings of NAACL, to appear