Attitude, sentiment and emotion detection

Βροχίδης Αλέξανδρος

Λιαπίκος Θεόδωρος

Η Ανάλυση Συναισθημάτων και η Χρησιμότητά της

Ορισμός

Ως ανάλυση συναισθημάτων ορίζεται η διαδικασία υπολογιστικής μελέτης των απόψεων και των συναισθημάτων των ανθρώπων ως προς μια οντότητα.

Εστίαση κυρίως σε απόψεις, που εκφράζονται σε ένα κομμάτι κειμένου, από κάποιο συγγραφέα, αναφορικά με μια οντότητα.

Χρησιμότητα της ανάλυσης συναισθημάτων

Πρακτική εφαρμογή σε:

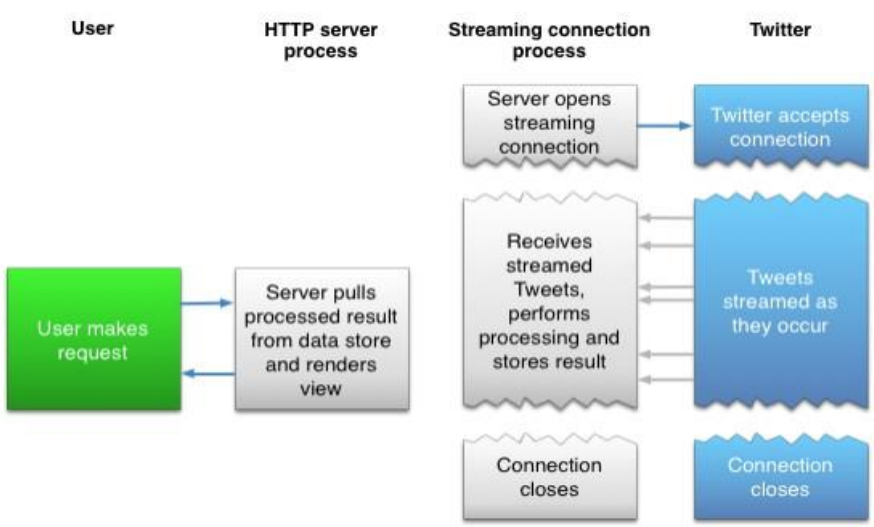
- Επιχειρήσεις
  - Στάση πελατών απέναντι στα προϊόντα μιας εταιρίας
  - Πρόβλεψη κρίσεων
- Ατομικά

  Άποψη ψηφοφόρων για τις θέσεις ενός πολιτικού
- Χρηματοοικονομικά

  Πρόβλεψη πορείας μιας μετοχής

Συλλογή των Δεδομένων

- Εταιρικές πηγές (fora, chats)
- Μέσα κοινωνικής δικτύωσης (Twitter, FB κλπ.)
- Χρήση εξειδικευμένων APIs
  - Προγραμματιστική πρόσβαση στα Δεδομένα και διαχείριση της ροής της πληροφορίας
- Twitter APIs
  - Twitter REST API (στατικά δεδομένα)
  - Twitter Streaming API (ροή δεδομένων)



Προεπεξεργασία των Δεδομένων

Χαρακτηριστικά των Δεδομένων

- Τεράστιος όγκος
- Συνεχής ροή σε πραγματικό χρόνο
- Είναι αδόμητα
- Μεταφορά θορύβου
  - Ορθογραφικά λάθη
  - Χρήση ιδιωματισμών
  - Υπερβολική χρήση σημείων στίξης
  - Χρήση συντομογραφιών
  - Χρήση εικονιδίων
  - Συνδυασμός διαφόρων γλωσσών

Αποτελέσματα προεπεξεργασίας Δεδομένων

- Βελτίωση “ποιότητας” Δεδομένων
- Μείωση όγκου Δεδομένων
- Μείωση υπολογιστικού όγκου και χρόνου
- Βελτίωση απόδοσης των αποτελεσμάτων

Αλλά

- Επιπλέον φάση επεξεργασίας (χρόνος, κόστος)
- Απαιτεί πειραματισμό, για την επιλογή των κατάλληλων συνδυασμών τεχνικών

Τεχνικές Προεπεξεργασίας των Δεδομένων

- Tokenization

  Διαχωρισμός του κειμένου σε:

  - Προτάσεις
  - Λέξεις

- Αφαίρεση Stop Words

  Συχνές λέξεις που δεν μεταφέρουν κάποια ιδιαίτερη πληροφορία

- Αφαίρεση σημείων στίξης

- Stemming

  Αφαίρεση κατάληξης ή προθέματος λέξης, διατήρηση της ρίζας (όχι απαραίτητα έγκυρη λέξη)

  (studies -> studi, studying -> study)

- Lemmatization

  Επιστροφή της λεξικογραφικής μορφής (λήμμα) της λέξης (έγκυρη λέξη)

  (studies -> study, studying -> study)

- Εύρεση μέρους του λόγου (Part of Speech tagging)

  Γραμματική κατηγοριοποίηση λέξεων (π.χ. ρήμα, επίθετο)

- Εντοπισμός Οντοτήτων (named Entities)

  Ταυτοποίηση βασικών Οντοτήτων του κειμένου και κατηγοριοποίησή τους

  - π.χ. Άτομα, Τοποθεσίες, Εταιρίες, Οργανισμοί

- Εντοπισμός Συσχετίσεων Οντοτήτων

  “Ο Χ είναι πρόεδρος της εταιρίας Υ”

Εξαγωγή Χαρακτηριστικών (Feature Extraction) των Δεδομένων

- Σημασιολογικά χαρακτηριστικά (semantic features)

  Φορτισμένα με θετικό ή αρνητικό συναίσθημα
  - emoticons
- Συντακτικά χαρακτηριστικά (syntactic features)
  - Μέρη του Λόγου (POS)
  - N-grams

    unigram (…w1 w2 w3 w4 w5…)

    bigram (…w1 w2 w3 w4 w5…)

    trigram (…w1 w2 w3 w4 w5…)
  - Οντότητες (named Entities)

- Στυλιστικά χαρακτηριστικά (stylistic features)

  Ιδιομορφίες που χαρακτηρίζουν το στυλ του συγγραφέα του κειμένου
  - emoticons
  - Χρήση ιδιωματισμών
- Ειδικά χαρακτηριστικά (specific features)

  Χαρακτηρίζουν το μέσο προέλευσης του κειμένου, π.χ. για τα tweets:
  - Hashtags
  - Retweets
  - Friends
  - Followers
  - Geo Tags

Πρακτικές προσεγγίσεις

- Χρήση Λεξικών Συναισθήματος
- Τεχνικές Μηχανικής Μάθησης
- Υβριδική προσέγγιση

Λεξικά Συναισθήματος (Sentiment Lexicons)

- Ένα λεξικό συναισθήματος είναι μια λίστα λέξεων και φράσεων που συνήθως χρησιμοποιούνται για να εκφράσουν ένα θετικό ή αρνητικό συναίσθημα
- Πρέπει να φτιάχνονται με προσοχή, καθώς εκεί στηρίζεται σε μεγάλο ποσοστό η επιτυχία της ανάλυσης

Κατασκευή των λεξικών

- Δίνουν βάρη σε συγκεκριμένες λέξεις ανάλογα με το συναίσθημα που αυτές αποδίδουν (θετικές, αρνητικές)
- Λέξεις που δεν υπάρχουν στο λεξικό θεωρούνται ουδέτερες
- Επέκταση λεξικών:
  - Εντοπισμός ομώνυμων και αντώνυμων λέξεων και απόδοση αντίστοιχων βαρών
  - Χρήση συντακτικών χαρακτηριστικών όπως “και”, “είτε”

Χρήση Λεξικών Συναισθήματος

- Κάθε λέξη του εξεταζόμενου κειμένου ελέγχεται στο λεξικό και αποδίδεται σε αυτή ένα σκορ, ανάλογα με το θετικό, αρνητικό ή ουδέτερο συναίσθημα, με το οποίο συνδέεται
- Το σκορ δείχνει την ένταση της έκφρασης των συναισθημάτων της κάθε λέξη
- Με ειδικό τύπο λαμβάνεται υπόψη η επίδραση ειδικών λέξεων, όπως επιρρήματα, που ενισχύουν ή αποδυναμώνουν την έννοια της λέξης στην οποία αναφέρονται
- Γίνεται κανονικοποίηση του σκορ [0,1]
- Το συνολικό σκορ του κειμένου προκύπτει από την άθροιση των επιμέρους σκορ των λέξεων που την αποτελούν. Το συνολικό σκορ καθορίζει και την κατηγοριοποίηση του κειμένου ως θετικό, αρνητικό ή ουδέτερο
- Η απόδοση της Ανάλυσης εξαρτάται άμεσα από την ποιότητα του χρησιμοποιούμενου Λεξικού

Πρακτικές προσεγγίσεις

Μηχανική μάθηση με επίβλεψη

- Στην επιβλεπόμενη μάθηση θεωρείται ότι υπάρχει ένα πεπερασμένο σύνολο κλάσεων στις οποίες το έγγραφο μπορεί να ενταχθεί. Στην απλούστερη έκδοση υπάρχουν μόνο δύο κλάσεις, η θετική και η αρνητική.

Μηχανική μάθηση χωρίς επίβλεψη

- Οι προσεγγίσεις μη επιβλεπόμενης μηχανικής μάθησης για την ανάλυση συναισθημάτων των εγγράφων, βασίζονται στον εντοπισμό και προσδιορισμό του σημασιολογικού προσανατολισμού (semantic orientation) συγκεκριμένων φράσεων μέσα στο έγγραφο.

Support Vector Machines

Πως λειτουργεί

- Supervised Αλγόριθμος
- Το βασικό χαρακτηριστικό της μεθόδου είναι η αναπαράσταση των παραδειγμάτων ως σημεία χαρτογραφημένα έτσι ώστε τα παραδείγματα των διαφορετικών κατηγοριών να χωρίζονται από ένα σαφές κενό, το οποίο πρέπει να είναι όσο το δυνατόν ευρύτερο. Στη συνέχεια τα νέα παραδείγματα χαρτογραφούνται επίσης στο χώρο και προβλέπεται σε ποια πλευρά ανήκουν.

Πλεονεκτήματα και μειονεκτήματα

- ✓ Μεταχειρίζονται πολύπλοκα μη γραμμικά συστήματα, με καλή απόδοση στα κείμενα
- ✓ Χρησιμοποιούν απλούς γραμμικούς αλγορίθμους
- ✓ Ανθεκτικοί στην υπερμοντελοποίηση με χαμηλό κόστος
- Μεγάλες απαιτήσεις μνήμης
- Δεν είναι ερμηνεύσιμα μοντέλα

Naïve Bayes

Πως λειτουργεί

- Βασίστηκε στο θεώρημα του Bayes
- Μελετήθηκε τη δεκαετία του 1960
- Βασίζεται σε πιθανότητες
- Χρειάζεται κατάλληλη προεπεξεργασία για να είναι ανταγωνιστικός

Πλεονεκτήματα και μειονεκτήματα

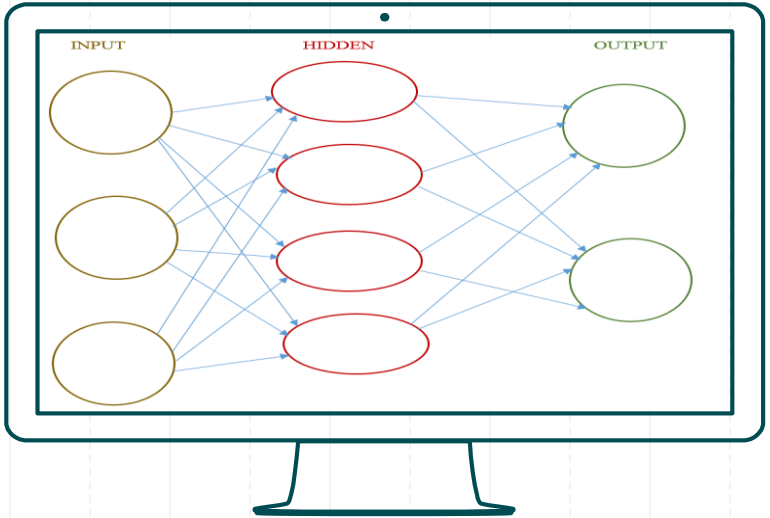
✓ Πολύ καλές αποδόσεις σχετικά με την ανάλυση συναισθημάτων σε κείμενα

- Υψηλή πολυπλοκότητα

Νευρωνικά Δίκτυα

- Μαθηματικά μοντέλα, εμπνευσμένα από τα βιολογικά νευρικά μοντέλα, τα οποία χρησιμοποιούν νευρώνες
- ✓ Εύκολη και αποδοτική μάθηση μέσω παραδειγμάτων
- ✓ Μεγάλη ανοχή σε σφάλματα
- ✓ Εξαιρετικά στην αναγνώριση προτύπων

- Πολυπλοκότητα και αυξημένος χρόνος εκτέλεσης
- Ερμηνευσιμότητα αποτελεσμάτων



Word2Vec

Πως λειτουργεί

- Νευρωνικά δίκτυα δύο επιπέδων
- Παράγουν διανύσματα στο χώρο για κάθε λέξη
- Μετράνε την εγγύτητα των διανυσμάτων για να βρουν ομοιότητα λέξεων
- Νοηματικά παρόμοιες λέξεις βρίσκονται σε κοντινές αποστάσεις στο χώρο διανυσμάτων

Πλεονεκτήματα και μειονεκτήματα

- ✓ Αποδίδει πολύ καλά σε μεταβλητές που είναι κατηγορικές
- ✓ Εύκολος υπολογισμός γραμμικών συνδυασμών των μεταβλητών
- \- Δυσκολία παραμετροποίησης

Υβριδικές Προσεγγίσεις

Πως λειτουργεί

- Γίνεται προεπεξεργασία των δεδομένων με βάση της τεχνικές που προαναφέρθηκαν
- Πραγματοποιείται η σχεδίαση και η αξιοποίηση λεξικών συναισθημάτων με προσεγμένες και ισοκατανεμημένες κλάσεις
- Ελέγχονται κείμενα από χρήστες που είναι θετικοί απέναντι στην εταιρεία αλλά και αρνητικοί, ώστε να υπάρχει ποικιλία συναισθημάτων
- Βρίσκονται και εφαρμόζονται τα πιο αποδοτικά μοντέλα μηχανικής μάθησης που αναφέρθηκαν
- Αξιολογούνται βάση συγκεκριμένων μετρικών απόδοσης

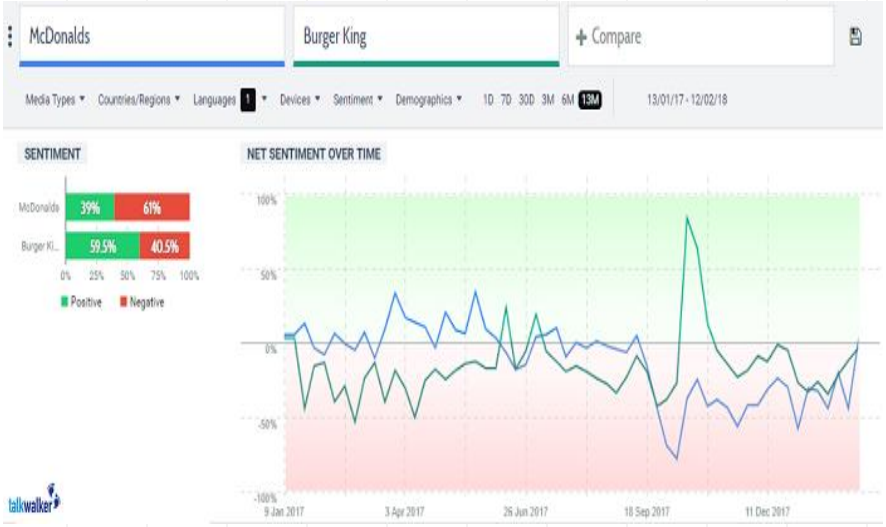
Προβλήματα Κατά την Ανάλυση Συναισθημάτων

- Προβλήματα με εξόρυξη της χρονικής στιγμής της σύνταξης του κειμένου
- Λέξεις διαφορετικά εκφρασμένες που όμως έχουν το ίδιο νόημα (εικόνα, φωτογραφία)
- Λέξεις που δεν εκφράζουν πάντα το ίδιο συναίσθημα (μικρή οθόνη, μικρό usb)
- Αναφορά σε πολλαπλές οντότητες “Η Χ κάμερα είναι πολύ καλύτερη από την Υ κάμερα. Είναι φθηνότερη κιόλας”
- Cross lingual opinion mining για επιχειρήσεις με πελάτες σε όλο τον κόσμο.

Βασικά Εργαλεία και Εφαρμογές

Quick Search

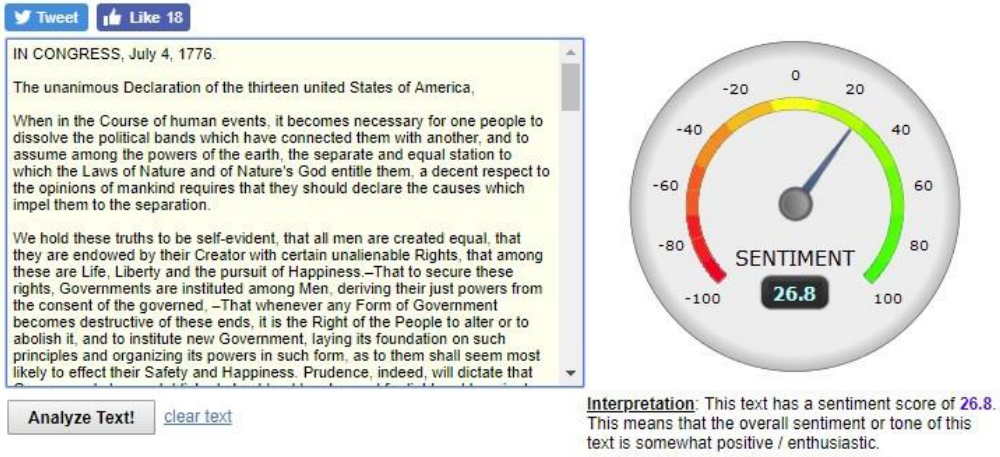
- Δωρεάν demo σε επιχειρήσεις
- Εισαγωγή δύο ή περισσοτέρων brands και απεικόνιση θετικών η αρνητικών συναισθημάτων από τον κόσμο που μιλάει γι' αυτά.
- Όμορφο γραφικό περιβάλλον με καλή ανάλυση



Βασικά Εργαλεία και Εφαρμογές

Sentiment Analyzer

- Δωρεάν χρήση
- Ικανοποιεί περιορισμένες απαιτήσεις
- Λειτουργεί καλύτερα στα αγγλικά
- 8.000 Δείγματα κειμένων
- Δεν μπορεί να δεχτεί ροή δεδομένων



Συμπεράσματα

- Η ανάλυση συναισθημάτων είναι ένας συνεχώς αναπτυσσόμενος τομέας με μεγάλο ερευνητικό ενδιαφέρον
- Τα λεξικά που αναπτύσσονται παρουσιάζουν ελλείψεις ως προς τις γλώσσες
- Υπάρχει δυσκολία στην αναγνώριση ειρωνικών, και σαρκαστικών σχολίων
- Οι υβριδικές προσεγγίσεις φαίνονται να παράγουν τα καλύτερα αποτελέσματα στις περισσότερες περιπτώσεις

THANK YOU!

Any questions?