# **Κατηγοριοποίηση tweets και ανάλυση πρόβλεψης**

# Abstract

Τα κοινωνικά μέσα δικτύωσης παράγουν τεράστια ποσά δεδομένων κάθε λεπτό, γεγονός το οποίο οφείλεται στην υιοθέτησή τους κατά τα τελευταία χρόνια. Οι καινοτομίες στον κλάδο επέτρεψαν νέους τρόπους επικοινωνίας μεταξύ των ανθρώπων και δημιούργησαν πολλές επιχειρηματικές ευκαιρίες. Μεγάλα δεδομένα στα κοινωνικά μέσα απαιτούν αποτελεσματικές και προηγμένες τεχνολογίες επεξεργασίας. Σκοπός των αναλύσεων εξόρυξης δεδομένων είναι να βρεθούν αξιόπιστα πρότυπα και πληροφορίες από τα δεδομένα του Twitter.

Σκοπός αυτής της εργασίας είναι να αναπτύξει μια τεχνική με την οποία με βάση τη σημασιολογική ανάλυση που κάνουμε, αλλά και με τη χρήση κατάλληλων αλγορίθμων, να ταξινομούμε κατάλληλα τα tweets. Οι τεχνικές εξόρυξης δεδομένων, μαζί με τη φυσική επεξεργασία των γλωσσών, υλοποιούνται προκειμένου να βρεθεί το πλαίσιο «πίσω» από τα δεδομένα και να μπορούμε να βγάλουμε χρήσιμα συμπεράσματα.

# Περιεχόμενα

1. Εισαγωγή………………………………………………………………………………………………………3
   1. Κίνητρα και υπόβαθρο……………………………………………………………………………….4
   2. Twitter……………………………………………………………………………………………………….5
   3. Χαρακτηριστικά των Big Data…………………………………………………………………….7
2. Διατύπωση και προσέγγιση του προβλήματος……………………………………………..8
   1. Διατύπωση του προβλήματος…………………………………………………………………….8
   2. Περιγραφή αλγόριθμου……………………………………………………………………………..9
   3. Ταξινόμηση των tweets – Αλγόριθμοι………………………………………………………10
      1. Naïve Bayes(NB)………………………………………………………………………………10
      2. SupportVector Machines(SVM)………………………………………………………..13
   4. Μετρικές αξιολόγησης……………………………………………………………………………..17
3. Μοντέλο κατηγοριοποίησης……………………………………………………………………….19
   1. Σύνολο δεδομένων…………………………………………………………………………………..19
   2. Περιγραφή μοντέλου πρόβλεψης…………………………………………………………….20
      1. Εισαγωγή των δεδομένων……………………………………………………………….21
      2. Επεξεργασία των δεδομένων…………………………………………………………..21
         1. Bag of Words………………………………………………………………………………..21
         2. Stemming and Lemmatization………………………………………………………22
         3. Μοντέλο Διανυσματικού χώρου…………………………………………………..23
         4. Αναπαράσταση δεδομένων – TF-IDF…………………………………………….23
      3. Επιλογή αλγόριθμου……………………………………………………………………….24
      4. Αξιολόγηση του μοντέλου……………………………………………………………….25
4. Πειραματικά αποτελέσματα……………………………………………………………………….26
   1. Κατηγοριοποίηση με Naïve Bayes…………………………………………………………….26
   2. Κατηγοριοποίηση με SVM………………………………………………………………………..32
   3. Σύγκριση αποτελεσμάτων………………………………………………………………………..34

# Εισαγωγή

Κατά τα τελευταία έτη, τα ποσά των δεδομένων που παράγονται από τις υπηρεσίες Διαδικτύου έχουν αυξηθεί σημαντικά. Οι καινοτομίες στον τομέα της Πληροφορικής και των Τηλεπικοινωνιών επέτρεψαν νέες επιχειρηματικές ευκαιρίες για τη δημιουργία υπηρεσιών ικανών να χειρίζονται μεγάλους όγκους δεδομένων. Η τεχνολογία έφτασε στο επίπεδο όπου οι άνθρωποι μπορούν και συνδέονται καθημερινά με τα κοινωνικά μέσα και μοιράζονται τη ζωή τους μέσω των κοινωνικών δικτύων.

Η κοινωνική δικτύωση μέσω του Διαδικτύου έχει γίνει δημοφιλής τα τελευταία χρόνια, γεγονός που δικαιολογείται και από τους αυξημένους όγκους δεδομένων. Παρουσιάστηκαν νέες προκλήσεις σε σχέση με τις αρχιτεκτονικές αποθήκευσης δεδομένων με χαρακτηριστικά κλιμάκωσης και αποτελεσματικούς αλγόριθμους επεξεργασίας. Η ανάλυση εξόρυξης δεδομένων έχει μεγάλες δυνατότητες για την εξεύρεση ουσιαστικών στοιχείων μέσα στα δεδομένα των κοινωνικών δικτύων. Το κοινωνικό δίκτυο του Twitter είναι μια υπηρεσία που αναπτύσσεται προκειμένου να καταστεί δυνατή η επικοινωνία μεταξύ των ανθρώπων στέλνοντας σύντομα μηνύματα.

Σύμφωνα με έρευνα, το Twitter ως η δεύτερη μεγαλύτερη πλατφόρμα κοινωνικών μέσων, ακριβώς πίσω από το Facebook, δημιουργεί περίπου 350.000 tweets κάθε λεπτό ή 21 εκατομμύρια ανά ώρα. Αυτοί οι όγκοι δεδομένων παρουσιάζουν προκλήσεις για τους μηχανικούς να αναπτύξουν καινοτόμες λύσεις για αποτελεσματική αρχιτεκτονική δεδομένων και δυνατότητες επεξεργασίας για την εφαρμογή της εξόρυξης δεδομένων. Η σημασία της υλοποίησης μεγάλων δεδομένων σε επιχειρήσεις σε διάφορους τομείς, όπως η βιομηχανία της υγείας, το λιανικό εμπόριο, τα τηλεπικοινωνιακά δίκτυα ή τα κοινωνικά δίκτυα, διαδραματίζει κρίσιμο σενάριο για τη βελτιστοποίηση των επιχειρηματικών διαδικασιών και τη δημιουργία νέων προτάσεων αξίας για τις ροές εσόδων.

Το Twitter έχει μεγάλη επίδραση στην εξόρυξη δεδομένων, καθώς οι χρήστες παράγουν Big Data που μπορεί να επεξεργαστεί. Επιπλέον, υπάρχουν απαιτήσεις αρχιτεκτονική ανάπτυξης που μπορεί να εξελίξει για συνεχή νέα-ροή tweets και επίσης δυνατότητα να ενσωματώσει με προχωρημένους αλγόριθμους μηχανικής μάθησης. Γνωρίζοντας τι σκέφτονται οι χρήστες ή πώς αισθάνονται για προϊόντα είναι πολύτιμη πρόταση για τις εταιρείες.

Η σημασιολογική ανάλυση είναι μέρος της εξόρυξης δεδομένων, η οποία παρακολουθεί τις αντιλήψεις του κοινού σχετικά με διάφορα θέματα. Αυτό μπορεί να αναλύσει τι σκέφτονται οι άνθρωποι για τα επιχειρηματικά προϊόντα και την ποιότητά τους, τα επώνυμα προϊόντα, τις στρατηγικές τιμολόγησης ή παγκόσμιες τάσεις. Επιπλέον, μπορεί να προσδιορίσει τις επιχειρηματικές ευκαιρίες και έτσι να γίνει ένας αποτελεσματικός παράγοντας για τις εταιρείες ώστε να καινοτομούν τις υπηρεσίες τους.

Το Twitter ως πλατφόρμα που υποστηρίζεται από τους ενεργούς χρήστες δημιουργεί ευκαιρίες για εξόρυξη δεδομένων και πιο συγκεκριμένες σημασιολογικές αναλύσεις με βάση τα tweets. Οι χρήστες Twitter συχνά εκφράζουν τις απόψεις τους σχετικά με διάφορα θέματα μέσα στα δημοσιευμένα tweets τους. Και συνεπώς, εφαρμόζοντας τεχνική επεξεργασία κειμένων, η τεχνική εξόρυξης δεδομένων μπορεί να εξυπηρετήσει τις εταιρείες μέσω ανατροφοδότηση για την καλύτερη διαχείριση της επωνυμίας της εταιρίας.

Από την άλλη πλευρά, δεδομένου ότι το Twitter παράγει τεράστιους όγκους δεδομένων κάθε μέρα, οι σημασιολογικές αναλύσεις μπορούν να βοηθήσουν στις εκστρατείες που σχετίζονται με το μάρκετινγκ, για να διερευνήσουν τις απόψεις του κοινού σχετικά με το νεοεμφανιζόμενο προϊόν. Για παράδειγμα για μια κινηματογραφική ταινία να αναλύσουν το συναίσθημα για την ικανοποίηση των χρηστών.

# Κίνητρα και υπόβαθρο

Το Twitter είναι ένα μέσο κοινωνικής δικτύωσης, όπου οι χρήστες μπορούν να μοιράζονται σύντομα μηνύματα, που ονομάζονται tweets. Η υπηρεσία ξεκίνησε το 2006 με έδρα στην Καλιφόρνια των ΗΠΑ. Κατά τα χρόνια της ανάπτυξής του κέρδισε παγκόσμια δημοτικότητα με 320 εκατομμύρια ενεργούς χρήστες. Οι χρήστες που έχουν συνδεθεί με την υπηρεσία μπορούν να μοιραστούν τις στιγμές τους με φίλους, να συμπεριλάβουν συνδέσμους σε φωτογραφίες ή βίντεο σε tweets, να κάνουν σχόλια και να επαναδημοσιεύσουν άλλα tweets. Το Twitter υποστηρίζει επίσης τη λειτουργία παρακολούθησης άλλων χρηστών.

Εκτός από τη δημιουργία προσωπικών λογαριασμών, πολλές επιχειρήσεις δημιουργούν τους δικούς τους λογαριασμούς για να συνεργαστούν με τους πελάτες και να λάβουν ανατροφοδότηση για τα προϊόντα, να προωθήσουν δραστηριότητες της εταιρείας, εκπτώσεις ή ενημερωτικά δελτία. Ως εκ τούτου, το Twitter έγινε ενδιαφέρουσα πλατφόρμα για το branding και το μάρκετινγκ.

Το Twitter αποτελεί τον αποτελεσματικό τρόπο επικοινωνίας μεταξύ φίλων, χρηστών ή πελατών. Ως εκ τούτου, η εξόρυξη δεδομένων πάνω από το Twitter μπορεί να οδηγήσει σε ενδιαφέρουσες πληροφορίες για τους χρήστες και να δημιουργήσει αξία για τις εταιρείες.

# Twitter

Το Twitter επιτρέπει στους χρήστες να επικοινωνούν με σύντομα μηνύματα που ονομάζονται tweets. Κάθε tweet μπορεί να περιέχει 140 χαρακτήρες το πολύ. Σύμφωνα με την ιστοσελίδα του Twitter [1], 140 χαρακτήρες παρουσιάζουν το τέλειο μήκος για την αποστολή ενημερώσεων κατάστασης μέσω μηνυμάτων κειμένου. Επιπλέον, 20 άλλοι χαρακτήρες προορίζονται για τα ονόματα των ανθρώπων. Μόλις οι χρήστες συνδεθούν σε υπηρεσία και εγγραφούν για δωρεάν λογαριασμό, τα μέλη μπορούν να στείλουν tweets ή να ακολουθήσουν άλλα μέλη για να ενημερωθούν για τα τελευταία νέα. Αυτά τα σύντομα μηνύματα δημοσιεύονται στο προφίλ των χρηστών. Επιπλέον, μπορούν να σταλούν στους οπαδούς και να μπορούν να αναζητηθούν στο Twitter. [1]

Η υπηρεσία κοινωνικής δικτύωσης δεν περιορίζεται μόνο στην πρόσβαση στον ιστότοπο, αλλά οι χρήστες μπορούν να εμπλέκονται και ν’ αλληλεπιδρούν μέσω εφαρμογών που έχουν αναπτυχθεί για έξυπνες συσκευές(smartphones, tablets). Στην πραγματικότητα, σύμφωνα με τα στοιχεία χρήσης του Twitter [5], υπάρχει το 80% των ενεργών χρηστών στο κινητό. Η δημοτικότητα σε σχέση με τα micro blogging κέρδισε επιτυχία σε ολόκληρο τον κόσμο, γεγονός που υποστηρίζεται από το ότι το Twitter έχει περίπου 320 εκατομμύρια ενεργούς χρήστες που ασχολούνται με την υπηρεσία σε μηνιαία βάση.

Το περιεχόμενο που δημιουργείται από το χρήστη μπορεί να δημιουργήσει πολλές ευκαιρίες για μάρκετινγκ και διαφήμιση, στις οποίες χρησιμοποιούνται οι τεχνικές εξόρυξης δεδομένων. Το Twitter είναι χρήσιμο για την ανάγνωση και εύρεση ενδιαφερόντων θεμάτων που προσελκύουν την προσοχή του χρήστη. Οι άνθρωποι μπορούν να ανακαλύψουν νέα σε πραγματικό χρόνο σχετικά με το τι συμβαίνει στον κόσμο ή να μένουν σε επαφή με φίλους.

Από την άλλη πλευρά, πολλές εταιρείες χρησιμοποιούν το Twitter για να ενημερώσουν τους πελάτες σχετικά με τις προσφορές τους. Το περιεχόμενο των tweets σχετίζεται με δύο επιπλέον μεταδεδομένα που διακρίνονται σε οντότητες και μέρη. Οι οντότητες Tweet είναι αναφορές χρηστών, οι οποίες αντιπροσωπεύουν τον τρόπο αναφοράς των άλλων χρηστών στα δικά τους tweets, συμπεριλαμβάνοντας το @ σύμβολο, ακολουθούμενο από το όνομα χρήστη τους. Επιπλέον, οι οντότητες tweet ενδέχεται να περιέχουν επίσης hashtags και διευθύνσεις URL. Αντίθετα, οι θέσεις tweet αντιπροσωπεύουν τοποθεσίες πραγματικού κόσμου που μπορούν να ενσωματωθούν σε ένα tweet. [9]

Η ορολογία είναι ένα σημαντικό μέρος του Twitter, διότι διδάσκει στους χρήστες τη λειτουργικότητα και τις λειτουργίες της υπηρεσίας. Επιπλέον, ορίζει τις πτυχές του Twitter και διάφορες δυνατότητες για τη χρήση του. Για να κατανοήσουμε την ορολογία του Twitter, παρουσιάζεται μια σύντομη επισκόπηση. Στο Twitter χρησιμοποιείται το (@) για να καλέσετε κάποιον χρήστη στο tweet ή για να στείλετε ένα μήνυμα στο χρήστη. Επιπλέον, αυτό το σύμβολο χρησιμοποιείται όποτε ο χρήστης θέλει να δημιουργήσει σύνδεση με άλλους χρήστες και να συνδεθεί με το Twitter προφίλ του. Το όνομα χρήστη προσδιορίζει με μοναδικό τρόπο κάθε χρήστη και γενικά χρησιμοποιείται με το σύμβολο @, για παράδειγμα, ο Andy Murray είναι @andy\_murray. [11]

Ένα άλλο δημοφιλές σύμβολο που χρησιμοποιείται στο Twitter ονομάζεται hashtag (#). Βοηθά τους χρήστες να κατηγοριοποιούν τα μηνύματα. Στην πραγματικότητα, η δομή του έρχεται με το (#) σύμβολο ακολουθούμενο από τη σχετική λέξη-κλειδί σε σχέση με το μήνυμα tweet. Ουσιαστικά, βοηθά στην κατηγοριοποίηση των tweets με βάση το περιεχόμενό τους και επιτρέπει καλύτερα αποτελέσματα αναζήτησης από το Twitter Search. [10] Hashtags μπορεί να βρίσκεται οπουδήποτε μέσα στο tweet. Όταν οι χρήστες κάνουν κλικ σε αυτό, θα οδηγηθούν σε κατηγορία που ομαδοποιεί όλα τα tweets από τους χρήστες Twitter στο ίδιο θέμα.

Το twitter επίσης υποστηρίζει αποστολή άμεσων προσωπικών μηνυμάτων μεταξύ των χρηστών, ως εκ τούτου αναπτύσσουν βασικές λειτουργίες ανταλλαγής μηνυμάτων, αλλά με προστιθέμενη αξία της micro blogging υπηρεσία(εμπεριέχονται video, posts, links). Το γεγονός αυτό είναι το κύριο πλεονέκτημα σε σύγκριση με τις απλές υπηρεσίες ανταλλαγής, όπως το WhatsApp ή το Viber.

Η δυνατότητα να εγγραφούν οι χρήστες σε ένα διαφορετικό λογαριασμό Twitter είναι γνωστό υπό όρο «following». Όταν ο χρήστης αποφασίσει να ακολουθήσει άλλο λογαριασμό, το Twitter θα ενημερώσει το προφίλ του ακόλουθου με τα πιο πρόσφατα tweets. Ο χρήστης γίνεται στους όρους του Twitter «follower». Αριθμός των followers εμφανίζεται στην επισκόπηση προφίλ χρήστη.

To Geotag χρησιμοποιείται για να ενημερώσει τους χρήστες σχετικά με την τοποθεσία όπου βρισκόταν ο δημιουργός, όταν καταχώρησε το tweet.

Κουμπί «Like» δείχνει την θετική αντίδραση προς το περιεχόμενο ενός tweet. Το προφίλ χρήστη επιτρέπει στους χρήστες να δημιουργήσουν ειδικές λίστες που θα μπορεί να ομαδοποιεί όλα τα θέματα σε κατηγορίες. Έτσι θα μπορεί να εξασφαλίζει καλύτερη πλοήγηση και συνοχή μεταξύ των θεμάτων ενδιαφέροντος του.

Προτεινόμενοι αλγόριθμοι μπορούν να προτείνουν στους χρήστες «Ποιούς να ακολουθούν». Η πληροφορία βρίσκεται στη σελίδα του προφίλ και συνήθως αποτελείται από προωθούμενους Twitter λογαριασμούς, τάσεις και tweets από διαφήμιση.

Η επιλογή της απάντησης έρχεται με κάθε καταχωρημένο tweet. Δίνει τη δυνατότητα σε κάποιο χρήστη να αντιδρά και να στείλει απάντηση. Επιπλέον, η επιλογή retweet είναι μέρος κάθε μηνύματος. Έχει επίσης τη λειτουργικότητα να μοιράζονται tweets με δικούς του «followers». Αν το tweet αποτελείται από περισσότερα από ένα σύμβολα @, λέγεται «αναφορά». [8]

# Χαρακτηριστικά των Big Data

Τα Big Data χαρακτηρίζονται ως δεδομένα που ξεπερνούν τις δυνατότητες της παραδοσιακής επεξεργασίας σε επίπεδο επιδόσεων και απαιτήσεων, οι οποίες δυνατότητες μπορούν να ταξινομηθούν σε 5 ομάδες. Ως εκ τούτου, τα μεγάλα δεδομένα [9] καθορίζεται από 5 «V»:

**Volume** 🡪 όγκος

**Velocity** 🡪 Ταχύτητα

**Variety** 🡪 Ποικιλία

**Veracity** 🡪 Αξιοπιστία

**Value** 🡪 Αξία

**Volume**: Χαρακτηρίζει το συνολικό ποσό των δεδομένων που δημιουργούνται από τις υπηρεσίες. Σχετικά με τα social media, το Twitter οδηγεί σε αυξημένες απαιτήσεις για συστήματα προηγμένης βάσης δεδομένων.

**Velocity:** Περιγράφει την ταχύτητα με την οποία τα δεδομένα δημιουργούνται από τους χρήστες twitter. Τέτοιες απαιτήσεις σχετικά με την αρχιτεκτονική δεδομένων εντός των επιχειρήσεων απαιτούν ισχυρή και επεκτάσιμη επεξεργασία.

**Variety**: Έχει να κάνει με τη μορφή των δεδομένων, όπως βίντεο ή ήχος, αλλά επίσης υπάρχει διαφορά μεταξύ αδόμητης και δομημένης μορφής πληροφορίας.

**Veracity**: Ασχολείται με τη σχετικότητα των δεδομένων, που είναι απαραίτητο να φτάσει πριν από κάποιες αναλύσεις δεδομένων. Περιττές πληροφορίες και θόρυβος πρέπει να αφαιρούνται από τα δεδομένα μέσω διαδικασιών κανονικοποίησης.

**Value:** Αφορά δεδομένα σχετικά με την εύρεση πολύτιμων πληροφορίων, π.χ. με τεχνικές εξόρυξης δεδομένων. Η αξία στα δεδομένα είναι προσβάσιμη από αναλύσεις συναισθήματος στο Twitter βρίσκοντας τα μοτίβα των δεδομένων και την προετοιμασία αλγόριθμων μηχανικής μάθησης. [56] Περίληψη των χαρακτηριστικών Big Data απεικονίζονται στο παρακάτω σχήμα 1

# Διατύπωση και Προσέγγιση του Προβλήματος

# 2.1 Διατύπωση του προβλήματος

Ο σκοπός αυτής της εργασίας είναι να επικεντρώσει στην εξόρυξη πληροφορίας από κείμενο(text mining) με συγκεκριμένους αλγόριθμους προκειμένου να ανακαλύψουμε πληροφορία «πίσω» από το περιεχόμενο των tweets και να βγάλουμε κάποια χρήσιμα συμπεράσματα.

Έχουμε επικεντρώσει στις αμερικανικές εκλογές και στην επίδραση των tweets στους χρήστες σχετικά με τους δύο υποψηφίους, Donald Trump και τη Hilary Clinton.

Η διαδικασία που ακολουθούμε έχει ως εξής:

* Φορτώνουμε και επεξεργαζόμαστε το σύνολο δεδομένων(dataset)
* Παρατηρούμε δημοφιλείς λέξεις για κάθε υποψήφιο και βλέπουμε την επίδραση αυτών των λέξεων στα tweets(πόσες φορές εμφανίζονται πχ).
* Στη συνέχεια, δημιουργούμε ένα μηχανισμό πρόβλεψης με τη χρήση συγκεκριμένου αλγόριθμου. Δηλαδή, εκπαιδεύοντας το συγκεκριμένο dataset, μπορούμε να προβλέψουμε ποσοστιαία για ένα νέο tweets, αν ταιριάζει περισσότερο(και πόσο) στο @realDonaldTrump ή @HillaryClinton. Συνεπώς μπορούμε να δείξουμε, ότι η χρήση συγκεκριμένων λέξεων κλειδιών, που δείξαμε ότι είναι δημοφιλείς, αν περιέχονται μέσα σ’ ένα νέο tweet μπορούν να χαρακτηρίζουν περισσότερο έναν υποψήφιο από τον άλλον.
* Μετράμε την απόδοση του αλγόριθμου για το συγκεκριμένο σύνολο εκπαίδευσης που έχουμε.
* Θα χρησιμοποιήσουμε το εργαλείο Orange3 της Python με τη βοήθεια του οποίου θα μπορέσουμε να αναλύσουμε και να επεξεργαστούμε τα δεδομένα μας, αλλά και να κάνουμε μια αντίστοιχη αξιολόγηση του δείγματος εκπαίδευσης, όπως προηγουμένως. Με αυτόν τον τρόπο θα μπορέσουμε να αξιολογήσουμε τις δυνατότητες, αλλά και την ευχρηστία του συστήματος.

# 2.2 Περιγραφή Αλγόριθμου

* Εισαγωγή και φόρτωση του dataset
* Προεπεξεργασία δεδομένων: Φιλτράρουμε τα tweets προκειμένου να μετρήσουμε τη συχνότητα εμφάνισης των λέξεων και να υπολογίσουμε το «βάρος» κάθε λέξης, αλλά και να μετατρέψουμε τις λέξεις σε λήμματα.
* Δημιουργούμε ένα training/testing μοντέλο
* Εντοπισμός λέξεων: Εντοπίζουμε δημοφιλείς λέξεις μέσα στα tweets και τις καταγράφουμε.
* Υπολογισμός συνολικού score και ταξινόμηση: Κάθε νέο tweet, που δημιουργείται ταξινομείται σε έναν από τους δύο υποψηφίους με ένα συγκεκριμένο ποσοστό επιτυχίας.
* Σύγκριση αλγόριθμων

# Ταξινόμηση των tweets – Αλγόριθμοι

Η ταξινόμηση κειμένου έγινε ένα πεδίο σημαντικής μελέτης επέκτασης των μέσων ενημέρωσης, όπως τα κοινωνικά δίκτυα. Προκλήσεις ως προς την ταξινόμηση συνδέονται με ταξινομητές και τις δυνατότητες απόδοσής τους, που εξαρτώνται από το συγκεκριμένο μοντέλο. Η ταξινόμηση απαιτείται στις μέρες μας σε διάφορες περιπτώσεις χρήσης, όπως η ανίχνευση spam, η ανάλυση συναισθήματος, προβλέψεις σε σχεδόν πραγματικό χρόνο.

# 2.3.1 Naïve Bayes

Ο Naïve Bayes αλγόριθμος βασίζεται στο θεώρημα πιθανοτήτων Bayes, που στόχο έχει να προβλέψει αποτελέσματα από μη-επισημασμένα δεδομένα. Στο θεώρημα Bayes, η μέθοδος ταξινόμησης προϋποθέτει την ανεξαρτησία μεταξύ των χαρακτηριστικών(features) και της κλάσης κατηγοριοποίησης(data).

Τα Naïve Bayes μοντέλα χωρίζονται σε διάφορους τύπους ανάλογα με το χειρισμό των χαρακτηριστικών. Στο Bernoulli μοντέλο, οι τιμές των χαρακτηριστικών πρέπει να είναι δυαδικές(0-1, T-F κ.α). Επιπλέον, σε αυτό το μοντέλο δε μας ενδιαφέρει η συχνότητα εμφάνισης κάθε λέξης μες στο κείμενο. Από τη άλλη, στο πολυωνυμικό(multinomial) μοντέλο μας ενδιαφέρει η συχνότητα των λέξεων μες στο κείμενο.

Ο Naïve Bayes αλγόριθμος ταξινόμησης είναι χρήσιμος για να χαρακτηρίσει ακόμα και σύνολα δεδομένων με υψηλό όγκο πληροφοριών, καθώς εκτελείται αποτελεσματικά και είναι εύκολο να εφαρμοστεί.

P(c|X) = [P(X|c) \* P(c)] / P(X)

* P(c|x) = εκ των υστέρων πιθανότητα
* P(x|c) = Δεσμευμένη πιθανότητα
* P(c) = εκ των προτέρων πιθανότητα κλάσης
* P(X) = εκ των προτέρων πιθανότητα ταξινομητή

P (c | X) = P (x1 | c) x P (x2 | c) x … x P (xn | c) x P (c)

Η εκ των υστέρων πιθανότητα υπολογίζει την πιθανότητα του αποτελέσματος που προκύπτει από μια νέα πληροφορία. Στο P(c|x), το c αναπαριστά την κλάση που ταξινομούνται τα δεδομένα και το x τον ταξινομητή. Η δεσμευμένη πιθανότητα είναι η πιθανότητα να βρίσκεται ο ταξινομητής(χαρακτηριστικό) μέσα στην κλάση.

Βάσει λοιπόν των πιο πάνω, μπορούμε να υπολογίσουμε τη μέγιστη a Posteriori πιθανότητα P(X) των δεδομένων.



=  P(X) σταθερό => Μπορεί να παραλειφθεί

= 

Υποθέτοντας ότι όλες οι πιθανότητες c είναι το ίδιο πιθανές, δηλαδή , μπορούμε να παραλείψουμε το και το P(c). Άρα:



Η πιο πάνω υπόθεση ονομάζεται **Maximum Likelihood Hypothesis**.

# Multinomial Naïve Bayes

Τα έγγραφα του training set περιγράφονται από συνδέσεις των τιμών των χαρακτηριστικών τους, δηλαδή των λέξεων που τα αποτελούν, δηλαδή παίρνουν διακριτές τιμές, με πιθανότητες εμφάνισης στην κατηγορία C, , τότε υποθέτουμε πολυωνυμική κατανομή. Συνεπώς, υπάρχει ένα πολυώνυμο πιθανοτήτων για κάθε κατηγορία, που εκφράζει την εμφάνιση των γεγονότων {1,2,…,n} σε αυτή την κατηγορία, με συχνότητες . Με άλλα λόγια . Θέλουμε να το αναθέσουμε σε μια κλάση cj  C.



= 

= 

Λόγω του ότι είναι δύσκολο να υπολογιστεί το  θα κάνουμε ακόμη μια υπόθεση ότι τα χαρακτηριστικά είναι στατιστικώς ανεξάρτητα, δηλαδή αν γνωρίζουμε την τιμή ενός χαρακτηριστικού, δεν μπορούμε να πούμε τίποτα για την τιμή κάποιου άλλου χαρακτηριστικού. Αυτό λέγεται **Conditional Independence Assumption**. Έτσι:



Το αποτέλεσμα είναι ο Naïve Bayes Classifier:



Για να αποφύγουμε να έχουμε αποτέλεσμα 0 σε περιπτώσεις όπου κάποια κλάση δεν περιέχει κάποιο χαρακτηριστικό το οποίο περιέχεται στο έγγραφο που προσπαθούμε να ταξινομήσουμε, δηλαδή , θα ορίσουμε το  ως



όπου:

k: αριθμός των τιμών που μπορεί να πάρει το xi.

Στη συνέχεια της εργασίας, για την κατηγοριοποίηση συναισθήματος, θα χρησιμοποιηθεί το Multinomial Naive Bayes μοντέλο.

# Gaussian Naïve Bayes

Για τον υπολογισμό των υπό συνθήκη πιθανοτήτων θα πρέπει να γνωρίζουμε την κατανομή των πιθανοτήτων των Xi. Αν τα χαρακτηριστικά των στιγμιοτύπων παίρνουν συνεχείς τιμές, τότε υποθέτουμε κανονική κατανομή για τις πιθανότητες των Xi δεδομένης της κατηγορίας C, . Κατά την εκπαίδευση, αρχικά επιλέγουμε για την κατηγορία C τα αντίστοιχα στιγμιότυπα που έχουν ταξινομηθεί χειροκίνητα σε αυτή. Εν συνεχεία, από το σύνολο αυτό, για κάθε χαρακτηριστικό, υπολογίζουμε το μέσο όρο των τιμών του, μC και τη διασπορά και βάσει αυτών προσδιορίζουμε την κανονική κατανομή του χαρακτηριστικού για την κατηγορία C. Στο τέλος, για κάθε χαρακτηριστικό και για κάθε κατηγορία, έχουμε μία κανονική κατανομή, η οποία καθορίζει την κατανομή του χαρακτηριστικού για την κατηγορία αυτή. Η πιθανοφάνεια των χαρακτηριστικών κατά τον έλεγχο υπολογίζεται ως εξής:



# 2.3.2 Support Vector Machines (SVM)

Πρόκειται για μια δυαδική μηχανή μάθησης. Η κεντρική ιδέα είναι ότι για ένα δείγμα εκπαίδευσης, η svm κατασκευάζει ενα υπερεπίπεδο ως επιφάνεια απόφασης με τρόπο ώστε το περιθώριο διαχωρισμού μεταξύ θετικών και αρνητικών παραδειγμάτων να μεγιστοποιείται. Η ιδέα αυτή βασίζεται στη θεωρία του πυρήνα εσωτερικού γινομένου μεταξύ ενός διανύσματος υποστήριξης Xi και ενός διανύσματος X το οποίο αντλείται από το χώρο δεδομένων εισόδου. Στη μέθοδο πυρήνα ο αλγόριθμος εξάγει ένα μικρό υποσύνολο σημείων δεδομένων από το ίδιο το δείγμα εκπαίδευσης. Η μέθοδος πυρήνα είναι βέλτιστη.

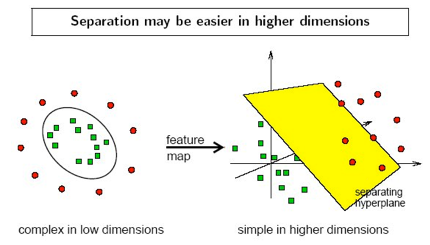
Η SVM **χρησιμοποιείται** για:

* *Ταξινόμηση προτύπων*
* *Προβλήματα μη-γραμμικής παλινδρόμησης*

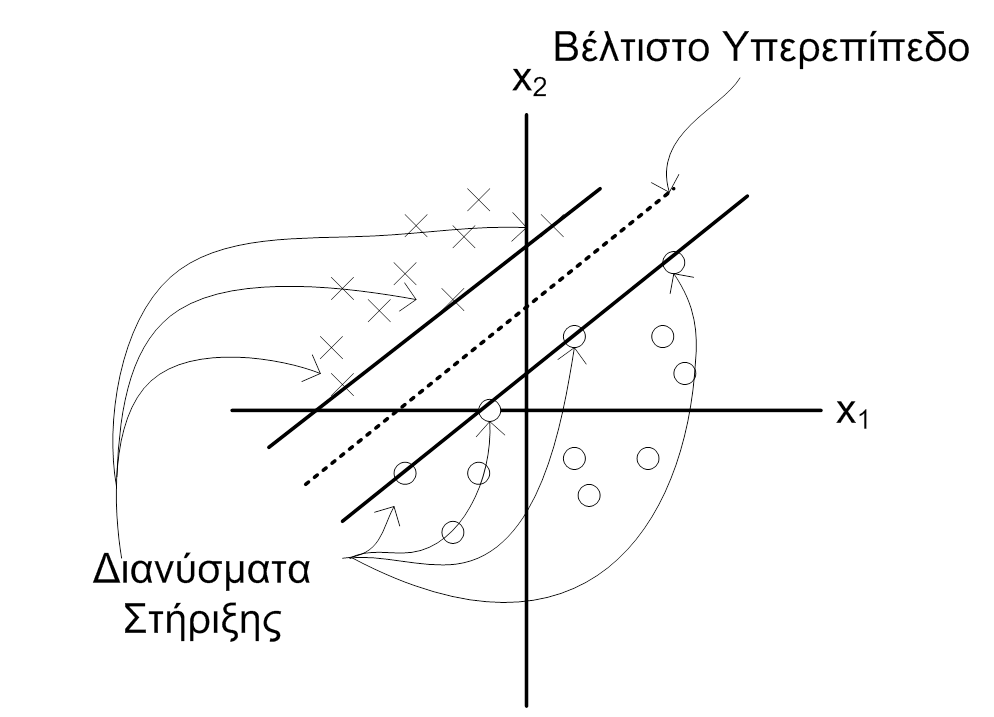
Οι ταξινομητές SVMs μπορεί να χρησιμοποιηθούν τόσο σε **γραμμικά** όσο και σε **μη γραμμικά** δεδομένα.

Μια μηχανή υποστήριξης διανυσµάτων χρησιμοποιεί μια μη γραμμική αντιστοίχιση για να μετασχηματίσει τα αρχικά δεδομένα εκπαίδευσης σε δεδομένα υψηλότερης διάστασης. Με μια κατάλληλη μη γραμμική αντιστοίχιση σε μια επαρκώς **υψηλότερη διάσταση**, τα δεδομένα πλέον από τις δύο κλάσεις μπορούν να διαχωριστούν από ένα **υπερεπίπεδο**. Με βάση λοιπόν τη νέα διάσταση, ο ταξινομητής ψάχνει για γραμμικώς διαχωριζόμενα υπερεπίπεδα (όρια απόφασης).

Η μέθοδος SVM βρίσκει αυτά τα υπερεπίπεδα χρησιμοποιώντας τα διανύσματα υποστήριξης (support vectors), που αποτελούν τα σημαντικότερα στιγμιότυπα εκπαίδευσης και τα περιθώρια (margins), που καθορίζονται από τα διανύσματα υποστήριξης. Διαισθητικά, ένας **καλός διαχωρισμός** επιτυγχάνεται από το υπερεπίπεδο το οποίο έχει τη μεγαλύτερη απόσταση από τα κοντινότερα σημεία εκπαίδευσης των κλάσεων ταξινόμησης (αποκαλούμενο περιθώριο συνάρτησης). Γενικά **όσο μεγαλύτερο το περιθώριο συνάρτησης τόσο χαμηλότερο είναι το παραγόμενο λάθος από τη χρήση του ταξινομητή**



Το βέλτιστο υπερεπίπεδο για γραμμικά διαχωρίσιμα πρότυπα

****

# Γραμμική κατηγοριοποίηση

Η **εξίσωση** που περιγράφει το υπερεπίπεδο είναι:

wTxi + b ≥ 0 για di = +1

wTxi + b < 0 για di = -1

x είναι το **διάνυσμα εισόδου,**w το **διάνυσμα βαρών** και b η **πόλωση**

**Βέλτιστο υπερεπίπεδο:** w0Tx + b0 = 0

Μπορούμε κάθε φορά να επανακλιμακώσουμε τα w0

και τα b0 έτσι ώστε να ισχύει:

w0Tx + b0 ≥ 0 για di = +1

w0Tx + b0 < 0 για di = -1

Για δεδομένο διάνυσμα βαρών **w** και πόλωσης **b**, η απόσταση του πιο κοντινού σημείου στην γραμμή η οποία ορίζεται από το βέλτιστο υπερεπίπεδο ονομάζεται περιθώριο διαχωρισμού και συμβολίζεται ως *ρ*. Ο στόχος των SV μηχανών είναι να βρουν το υπερεπίπεδο όπου **το *ρ* μεγιστοποιείται**

H διακρίνουσα συνάρτηση g(x) = W0Tx + b0, η οποία δίνει ένα αλγεβρικό μέτρο για το πόσο μακριά είναι το X από την γραμμή του βέλτιστου υπερεπιπέδου. Το Χ μπορεί να γραφεί ως **** ,όπου Xρ  είναι η προβολή του Χ στη γραμμή του βέλτιστου υπερεπιπέδου και r είναι η επιθυμητή απόσταση και ισοδύναμα 

Η απόσταση της W0Tx + b0 = 0 από την αρχή των αξόνων δίνεται από τον τύπο Εάν

* **b0 > 0** η αρχή των αξόνων βρίσκεται στην πλευρά της κλάσης +1
* **b0 < 0** η αρχή των αξόνων βρίσκεται στην πλευρά της κλάσης -1
* **b0 = 0** το **w0Tx + b0** περνά από την αρχή των αξόνων

Υποθέτουμε πως η *ρ* υποδηλώνει την βέλτιστη τιμή του περιθωρίου διαχωρισμού ανάμεσα στις δύο κλάσεις, οι οποίες αποτελούν το σύνολο εκπαίδευσης *Τ*, τότε ρ = 2r. Από το παραπάνω καταλαβαίνουμε πως για να μεγιστοποιήσουμε το περιθώριο διαχωρισμού μεταξύ των κλάσεων, μπορούμε ισοδύναμα να ελαχιστοποιήσουμε την ευκλείδεια νόρμα του διανύσματος βαρών W. Για την εύρεση του βέλτιστου υπερεπιπέδου χρησιμοποιούμε την τετραγωνική βελτιστοποίηση.

Η δημιουργία ενός υπερεπιπέδου διαχωρισμού είναι σχεδόν αδύνατο να γίνει χωρίς την **ύπαρξη σφαλμάτων ταξινόμησης**. Στόχος είναι η **ελαχιστοποίηση** της πιθανότητας σφάλματος ταξινόμησης. Εισάγουμε λοιπόν ένα σύνολο μη-αρνητικών βαθμωτών μεταβλητών ξ, οι οποίες αποκαλούνται μεταβλητές χαλάρωσης και η σχέση που περιγράφει το βέλτιστο υπερεπίπεδο είναι:



Μετρούν την απόκλιση ενός σημείου δεδομένων από την ιδανική συνθήκη διαχωρισιμότητας προτύπων.

* Για 0<ξ<1, το σημείο εμπίπτει στη σωστή περιοχή διαχωρισμού.
* Για ξ>1 εμπίπτει στη λάθος.

Το βέλτιστο υπερεπίπεδο στην περίπτωση αυτή θα δίνεται από την λύση του ακόλουθου προβλήματος βελτιστοποίησης:



Η ελαχιστοποίηση του πρώτου όρου σχετίζεται με την ελαχιστοποίηση της VC διάστασης του SVM δικτύου και άρα με τον έλεγχο της μάθησης. Όσον αφορά την ελαχιστοποίηση του δεύτερου όρου, αυτή έχει να κάνει με την ελαχιστοποίηση του αριθμού των λαθών. Η παράμετρος C προκαθορίζεται από τον χρήστη και αποτελεί έναν συμβιβασμό μεταξύ πολυπλοκότητας της μηχανής και του αριθμού των μη διαχωρίσιμων σημείων.

# Μη - Γραμμική κατηγοριοποίηση

Το πρόβλημα βελτιστοποίησης για μη-διαχωρίσιμα πρότυπα περιλαμβάνει το πρόβλημα της βελτιστοποίησης για γραμμικά διαχωρίσιμα πρότυπα ως ειδική περίπτωση. Συγκεκριμένα θέτοντας ξi = 0 για όλα τα i.

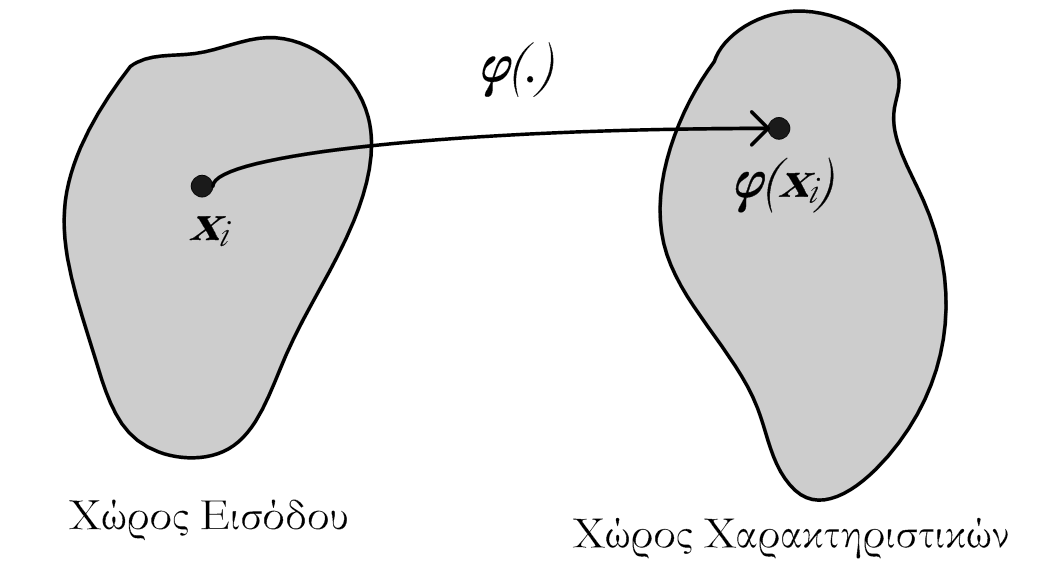
Η περίπτωση των μη-διαχωρίσιμων προτύπων διαφέρει από την περίπτωση των διαχωρίσιμων στο ότι ο περιορισμός αi>0 αντικαθίσταται από τον πιο αυστηρό περιορισμό 0<αi<C.

Η παράμετρος C είναι μια καθοριζόμενη από το χρήστη θετική παράμετρος, η οποία ελέγχει το συμβιβασμό μεταξύ της πολυπλοκότητας της μηχανής και του αριθμού των μη-διαχωρίσιμων σημείων

Μεγάλο C σημαίνει ότι υπάρχει μεγάλη εμπιστοσύνη στην ποιότητα του δείγματος εκπαίδευσης

Μικρό C σημαίνει ότι το δείγμα εκπαίδευσης είναι θορυβώδες.

Εν τούτοις, αν ο βέλτιστος τρόπος να διακριθούν τα δεδομένα είναι μη γραμμικά υπερεπίδεδα, τότε κρίνεται καταλληλότερη η χρήση μίας μη γραμμικής διανυσματικής συνάρτησης . Για το σκοπό αυτό, ο χώρος του προβλήματος μετασχηματίζεται, σε έναν άλλο χώρο μεγαλύτερης ή και άπειρης διάστασης μέσω της απεικόνισης

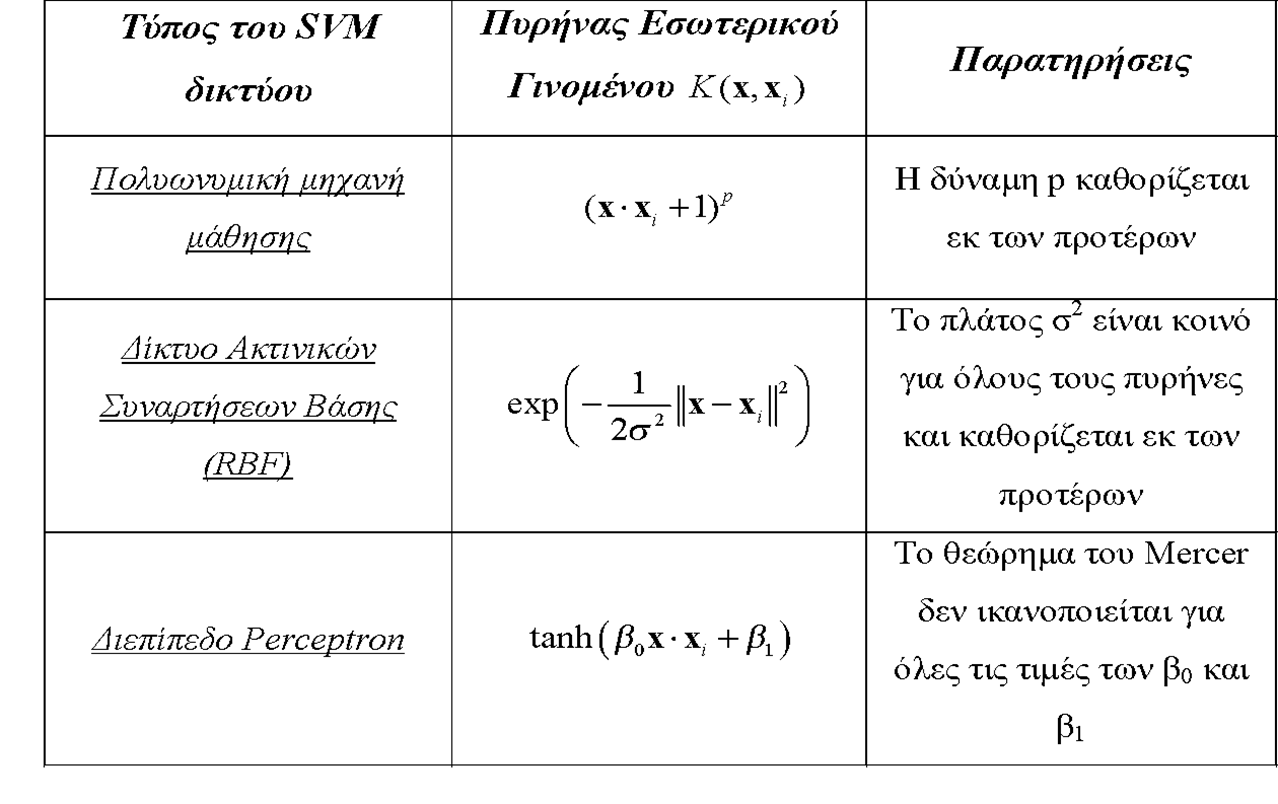


Ο πυρήνας K(x,xi) είναι μια συνάρτηση που υπολογίζει το εσωτερικό γινόμενο των εικόνων που παράγονται από το χώρο χαρακτηριστικών βάσει του Φ δύο σημείων δεδομένων στο χώρο εισόδου



Το βέλτιστο υπερεπίπεδο περιγράφεται από την παρακάτω σχέση:





# 2.4 Μετρικές Αξιολόγησης

Προκειμένου να αξιολογηθεί η επίδοση ενός ταξινομητή, έχει προταθεί πληθώρα μετρικών αξιολόγησης. Στη συνέχεια του κεφαλαίου, παρουσιάζονται οι δημοφιλέστερες μετρικές αξιολόγησης αλγορίθμων μηχανικής μάθησης.









Όπου,

* a (ή TP) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 1 και

ταξινομήθηκαν στην 1

* b (ή FN) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 1, αλλά

ταξινομήθηκαν στην 2

* c (ή FP) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 2, αλλά

ταξινομήθηκαν στην 1

* d (ή TN) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 2 και

ταξινομήθηκαν στην 2

Τέλος, η μετρική F-Measure παρέχει μία συνολική εκτίμηση των μοντέλων, καθώς συνδυάζει δύο άλλες μετρικές, την ανάκληση και την ακρίβεια. Η μετρική F-Measure στην ουσία είναι ο αρμονικός μέσος όρος (harmonic mean) της ανάκλησης και της ακρίβειας, και υπολογίζεται ως εξής:



Ουσιαστικά για κάθε κλάση βλέπουμε το πρόβλημα ως δυαδικό οπού η πρώτη έξοδος είναι η ίδια η κλάση και δεύτερη έξοδος όλες οι υπόλοιπες.

# Μοντέλο Κατηγοριοποίησης

# 3.1 Σύνολο Δεδομένων

Έχουμε ένα σύνολο από περίπου 6500 tweets και μεταδεδομένα από τους δύο βασικούς υποψηφίους των Αμερικανικών εκλογών, τη Hillary Clinton και τον Donald Trump.

|  |  |
| --- | --- |
| **Name** | **Type** |
| Id | Numeric |
| handle | String |
| text | String |
| is\_retweet | Boolean |
| original\_author | String |
| time | DateTime |
| lang | String |
| retweet\_count | Numeric |
| favorite\_count | Numeric |
| longitude | String |
| latitude | String |
| place\_full\_name | String |
| source\_url | String |

***Πίνακας 1: Χαρακτηριστικά συνόλου δεδομένων***

Στη συνέχεια θα δούμε κι ένα παράδειγμα από tweet έτσι όπως εμφανίζονται μες στο σύνολο δεδομένων.

|  |  |
| --- | --- |
| **Name** | **Type** |
| Id | 780925634159796224 |
| handle | HillaryClinton |
| text | The question in this election: Who can put the plans into action that will make your life better? https://t.co/XreEY9OicG |
| is\_retweet | 0 |
| original\_author | - |
| time | 2016-09-28 00:22:34 |
| lang | en |
| retweet\_count | 218 |
| favorite\_count | 651 |
| longitude | - |
| latitude | - |
| place\_full\_name | - |
| source\_url | https://studio.twitter.com |

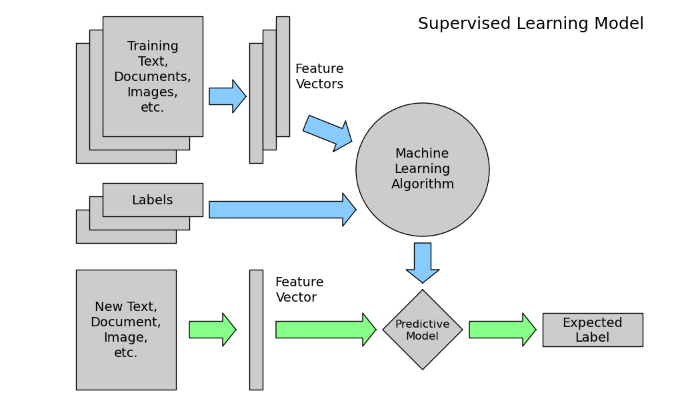
***Πίνακας 2: Ένα παράδειγμα εγγραφής στο σύνολο δεδομένων***

# 3.2 Περιγραφή μοντέλου πρόβλεψης

Παρατηρώντας το .csv αρχείο παρατηρούμε πως η στήλη handle είναι η ετικέτα που μας λέει εάν το συγκεκριμένο μήνυμα αναφέρεται στον Donald Trump ή στη Hillary Clinton.

Το σώμα του κειμένου είναι το επιβλεπόμενο σύνολο εκπαίδευσης. Επιβλεπόμενο γιατί υπάρχουν συγκεκριμένες ετικέτες/κλάσεις. Χρησιμοποιώντας λοιπόν αυτά τα δεδομένα θα εκπαιδεύσουμε ένα μοντέλο εκπαίδευσης, προκειμένου να μπορεί να διαχωρίζει τα tweets ανάμεσα στους δύο βασικούς υποψηφίους αυτόματα.

Συνεπώς, με αυτό το μοντέλο, θα είμαστε σε θέση να κατηγοριοποιούμε αυτόματα μη-κατηγοριοποιημένα(unlabeled) tweets σε έναν από τους δύο υποψηφίους.



# 3.2.1 Εισαγωγή των δεδομένων

Χρησιμοποιούμε τη βιβλιοθήκη Panda της Python για την εισαγωγή των δεδομένων μου.

tweets = pd.read\_csv('US\_elections.csv', encoding="utf-8")

Μπορούμε συνολικά να δούμε το σύνολο των tweets, αλλά και να αποτυπώσουμε σ έναν πίνακα κάποια βασικά στατιστικά στοιχεία.

|  |  |  |
| --- | --- | --- |
| handle |  | text |
| **HillaryClinton** | Count | 3226 |
|  | Unique | 3224 |
|  | top | A man who talks about our veterans and militar... |
|  | freq | 2 |
| **DonaldTrump** | Count | 3218 |
|  | Unique | 3210 |
|  | top | MAKE AMERICA GREAT AGAIN! |
|  | freq | 8 |

***Πίνακας 3: Βασικά στοιχεία για τους 2 υποψηφίους***

# 3.2.2 Επεξεργασία δεδομένων

# 3.2.2.1 Bag of Words

Το μοντέλο bag-of-words είναι μια απλοποιημένη αναπαράσταση που χρησιμοποιούνται στην επεξεργασία φυσικής γλώσσας και ανάκτηση πληροφοριών (IR). Σε αυτό το μοντέλο, ένα κείμενο (όπως μια φράση ή ένα έγγραφο) αναπαρίσταται ως μία «τσάντα», δηλαδή μια πληθώρα από λέξεις, παραβλέποντας γραμματική, ακόμα και τη σειρά των λέξεων, αλλά διατηρώντας την πολλαπλότητα.

Το μοντέλο Bag-of-Words χρησιμοποιείται συνήθως σε μεθόδους ταξινόμησης κειμένων όπου η εμφάνιση (συχνότητα) της κάθε λέξης χρησιμοποιείται ως ένα χαρακτηριστικό γνώρισμα για την εκπαίδευση του ταξινομητή.

def split\_into\_tokens(message):

message = message # convert bytes into proper unicode

return TextBlob(message).words

messages.message.head()

messages.message.head().apply(split\_into\_tokens)

0 [the, question, in, this, election, who, can, ...

1 [if, we, stand, together, there, 's, nothing, ...

2 [both, candidates, were, asked, about, how, th...

3 [join, me, for, a, 3pm, rally, tomorrow, at, t...

4 [this, election, is, too, important, to, sit, ...

# 3.2.2.2 Stemming and Lemmatization

Stemming είναι μια μέθοδος κανονικοποίησης δεδομένων, όπου η κατάληξη της λέξης αφαιρείται από τη βασική δομή ρίζας της, η οποία ονομάζεται stem. Για παράδειγμα, το παιχνίδι λέξεων έχει πολλές μορφές από τη γλωσσική άποψη: playing, played, plays κλπ. Οι αναλύσεις κειμένων δεν απαιτούν διαφοροποίηση μεταξύ των διαφόρων χρονικών περιόδων της λέξης.

Lemmatization είναι παρόμοια με την τεχνική που χρησιμοποιείται για την κανονικοποίηση του κειμένου. Η κύρια διαφορά είναι ότι ενώ το stemming μπορεί να μετατρέψει τη ρίζα της λέξης σε μη υπάρχουσα λέξη, η Lemmatization βασίζεται σε μέρη των κανόνων ομιλίας και μετατρέπει τη ρίζα της λέξης, που λέγεται λήμμα, σε πραγματική υπάρχουσα λέξη που αναζητείται στο λεξικό NLTK. Ουσιαστικά πρόκειται για μια πιο σύνθετη διαδικασία κανονικοποίησης του κειμένου.

def split\_into\_lemmas(message):

message = message.lower()

words = TextBlob(message).words

# for each word, take its "base form" = lemma

return [word.lemma for word in words]

# 3.2.2.3 Μοντέλο διανυσματικού χώρου

Στη συνέχεια πρέπει να μετατρέψουμε κάθε μήνυμα κειμένου, που πιο πάνω το αναπαραστήσαμε διαχωρισμένο σε λήμματα, σε διάνυσμα προκειμένου να γίνει κατανοητό από το μοντέλο μηχανικής μάθησης.

Για να γίνει αυτό, απαιτούνται 3 βήματα, στο μοντέλο bag-of-words:

1. Τη μέτρηση της συχνότητας εμφάνισης μιας λέξης μέσα σ ένα κείμενο.
2. Τον υπολογισμό του «βάρους» αυτής της μέτρησης. Τα λήμματα που εμφανίζονται συχνότερα θα έχουν μεγαλύτερο βάρος.
3. Κανονικοποίηση των διανυσμάτων σε μονάδα μήκους(L2 νόρμα)

Ουσιαστικά με αυτή τη διαδικασία δημιουργείται ένα μεγάλο αραιό μητρώο με συγκεκριμένες γραμμές, στήλες και μηδενικά.

messages\_bow = bow\_transformer.transform(messages['message'])

print('sparse matrix shape:', messages\_bow.shape)

print('number of non-zeros:', messages\_bow.nnz)

print('sparsity: %.2f%%' % (100.0 \* messages\_bow.nnz / (messages\_bow.shape[0] \* messages\_bow.shape[1])))

Out: sparse matrix shape: (5722, 9016)

number of non-zeros: 91236

sparsity: 0.18%

# Αναπαράσταση δεδομένων – TF-IDF

Κατά την αναπαράσταση των κειμένων, ως επί το πλείστον, το βάρος κάθε όρου, ισούται µε τη συχνότητα εμφάνισης του όρου, στο αντίστοιχο κείμενο. Η επιλογή της συχνότητας, ως στάθμιση, έχει ως αποτέλεσμα, οι όροι με τη μεγαλύτερη συχνότητα, να θεωρούνται ως οι περισσότερο αντιπροσωπευτικοί όροι του κειμένου, λόγω της βαρύτητάς τους. Μία λύση σε αυτό το πρόβλημα αποτελεί η στάθμιση TF-IDF, όπου TF (Term Frequency) η συχνότητα του όρου, ενώ IDF (Inverse Document Frequency) είναι ένα βάρος που δηλώνει τη σημαντικότητα ενός όρου του κείμενου, σε σχέση με το σύνολο των κειμένων. Η στάθμιση υπολογίζεται από τον πολλαπλασιασμό των TF και IDF.

Η στάθμιση TF-IDT, δίνει αρκετά καλά αποτελέσματα, καθώς το βάρος IDF παίρνει μεγάλες τιμές, όταν ένας όρος, υπάρχει σε λίγα κείμενα, ενώ, όταν ο όρος συναντάται σε πολλά από τα κείμενα, τότε το βάρος IDF παίρνει μικρές. τιμές. Με αυτή τη στάθμιση, οι σπάνιοι όροι έχουν υψηλό IDF, και όροι με μεγάλη συχνότητα βαρύνονται με χαμηλότερο IDF. Αυτή η προσέγγιση, έχει ως αποτέλεσμα, τα stopwords να παίρνουν σχετικά μικρό βάρος και να µην αποτελούν πλέον τους πιο αντιπροσωπευτικές όρους στα κείμενα.

tfidf\_transformer = TfidfTransformer().fit(messages\_bow)

messages\_tfidf = tfidf\_transformer.transform(messages\_bow)

print(messages\_tfidf)

print(messages\_tfidf.shape)

(0, 8755) 0.217922181616

(0, 8593) 0.161469587626

(0, 8572) 0.195824391455

(0, 7779) 0.183979385543

(0, 7722) 0.196876463794

(0, 7720) 0.162516355764

(0, 6287) 0.321250380345

(5722, 9016)

# Επιλογή Αλγόριθμου

Οι αλγόριθμοι που έχουν επιλεγεί, για την υλοποίηση των μοντέλων, προέρχονται από την επιβλεπόμενη μηχανική μάθηση. Ο Multinomial Naive Bayes και ο αλγόριθμός SVΜ, με γραμμική συνάρτηση πυρήνα, θα αποτελέσουν τη βάση των μοντέλων κατηγοριοποίησης που μελετηθούν στη συνέχεια της εργασίας. Μετά την ολοκλήρωση των μοντέλων, θα πραγματοποιηθεί συγκριτική ανάλυση των αποδόσεών τους, στα τρία σύνολα δεδομένων. Η εισαγωγή των αλγορίθμων γίνεται με τη χρήση της βιβλιοθήκης sklearn

# Αξιολόγηση του μοντέλου

Για την αξιολόγηση των μοντέλων κατηγοριοποίησης, έχουν επιλεγεί οι παρακάτω μετρικές, τις οποίες και αναλύσαμε προηγουμένως:

* Ορθότητα(Accuracy)
* Ανάκληση(Recall)
* Ακρίβεια(Precision)
* F-Measure

Επιπρόσθετα, θα εξάγουμε και τους Πίνακες Σύγχυσης (Confusion Matrices) των ταξινομητών, για την περαιτέρω ερμηνεία των αποτελεσμάτων.

# Πειραματικά Αποτελέσματα

Στη συνέχεια, παρουσιάζονται τα πειραματικά αποτελέσματα των ταξινομητών που υλοποιήθηκαν, για το σύνολο δεδομένων. Για κάθε σύνολο, ο καλύτερος MultinomialNB ταξινομητής, ο καλύτερος SVM ταξινομητής και εν τέλει, η τελική επιλογή ταξινομητή.

# 4.1 Κατηγοριοποίηση με Naïve Bayes

Αρχικά θα αξιολογήσουμε το σύνολο εκπαίδευσης(training set)

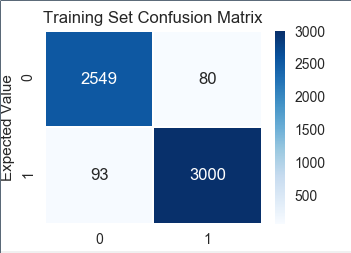
CV = 10: Χωρίζουμε το σύνολο δεδομένων τυχαία σε 10 μέρη, 9 για εκπαίδευση και 1 για τα αποτελέσματα

Test\_size = 0.2: Το μέγεθος του συνόλου ελέγχου αποτελεί το 20% του συνολικού δείγματος.

|  |  |  |
| --- | --- | --- |
| **Training\_size** | **Test\_size** | **Data\_size** |
| 4577 | 1145 | 5722 |

***Πίνακας 4: Διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο ελέγχου***

**Accuracy on training set**: 96.98%



***Πίνακας 5: Confusion Matrix για το σύνολο εκπαίδευσης***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **Hillary** | 0.96 | 0.97 | 0.97 | 2629 |
| **Trump** | 0.97 | 0.97 | 0.97 | 3093 |
| **avg / total** | 0.97 | 0.97 | 0.97 | 5722 |

***Πίνακας 6: Μετρικές αξιολόγησης για το σύνολο εκπαίδευσης***

Η συνήθης πρακτική που ακολουθείται είναι να χωρίσουμε το σύνολο εκπαίδευσης(Training\_set) σε μικρότερα υποσύνολα, εν προκειμένω σε 10. Συνεπώς εκπαιδεύουμε το μοντέλο σε 9 μέρη και υπολογίζουμε το Accuracy στο τελευταίο μέρος (validation set). Επαναλαμβάνουμε 10 φορές(παίρνοντας κάθε φορά διαφορετικό validation set). Με αυτόν τον τρόπο μπορούμε να έχουμε μια αίσθηση για την αξιοπιστία, τη σταθερότητα του μοντέλου, που δημιουργήσαμε. Εάν βγάλουμε πολύ διαφορετικά νούμερα από τη γενική αξιολόγηση του μοντέλου, η οποία είναι 96.98%, τότε σημαίνει ότι πρέπει να ξαναεπεξεργαστούμε τα δεδομένα μας, καθώς το μοντέλο μας δε δείχνει σταθερό.

Στην προκειμένη περίπτωση παρατηρούμε από τα παρακάτω αποτελέσματα, αλλά και από το μέσο όρο τους και τη διασπορά τους ότι έχω ένα αρκετά σταθερό μοντέλο, καθώς οι αποκλίσεις είναι πολύ μικρές.

|  |  |
| --- | --- |
| **#** | **Score\_Valid\_Set** |
| 1 | 0.93464052 |
| 2 | 0.91921397 |
| 3 | 0.93231441 |
| 4 | 0.93668122 |
| 5 | 0.93449782 |
| 6 | 0.93668122 |
| 7 | 0.92778993 |
| 8 | 0.91028446 |
| 9 | 0.94091904 |
| 10 | 0.93873085 |

***Πίνακας 7: Τα 10 Validatios Sets***

**Mean of 10 validation sets**: 0.9311753458

**Std of 10 validation sets:** 0.00909924550909

Ωστόσο υπάρχουν μια σειρά από παραμέτρους που πρέπει να συνυπολογιστούν στο μοντέλο. Μπορούμε λοιπόν να εξετάσουμε κατά πόσο η επεξεργασία και η λημματοποίηση βοηθούν στην αξιολόγηση του μοντέλου. Αυτό θα γίνει με τη χρήση του αλγόριθμου Naïve Bayes.

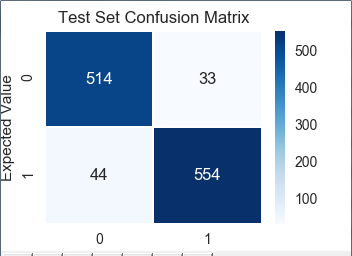
|  |  |  |
| --- | --- | --- |
| **bow\_\_analyzer** | **tfidf\_\_use\_idf** | **Mean/Std** |
| split\_into\_lemmas | True | 0.92965/0.00613 |
| split\_into\_lemmas | False | 0.92135/0.00424 |
| split\_into\_tokens | True | 0.93030/0.00585 |
| split\_into\_tokens | False | 0.92550/0.00396 |

***Πίνακας 8: Αξιολόγηση με τη χρήση των τεχνικών επεξεργασίας***

Παρατηρούμε πως τα καλύτερα αποτελέσματα(mean: 0.93030, std: 0.00585), τα έχουμε για:

* 'bow\_\_analyzer': <function split\_into\_tokens>
* 'tfidf\_\_use\_idf': True

Θα αξιολογήσουμε επίσης και το σύνολο ελέγχου(test set)



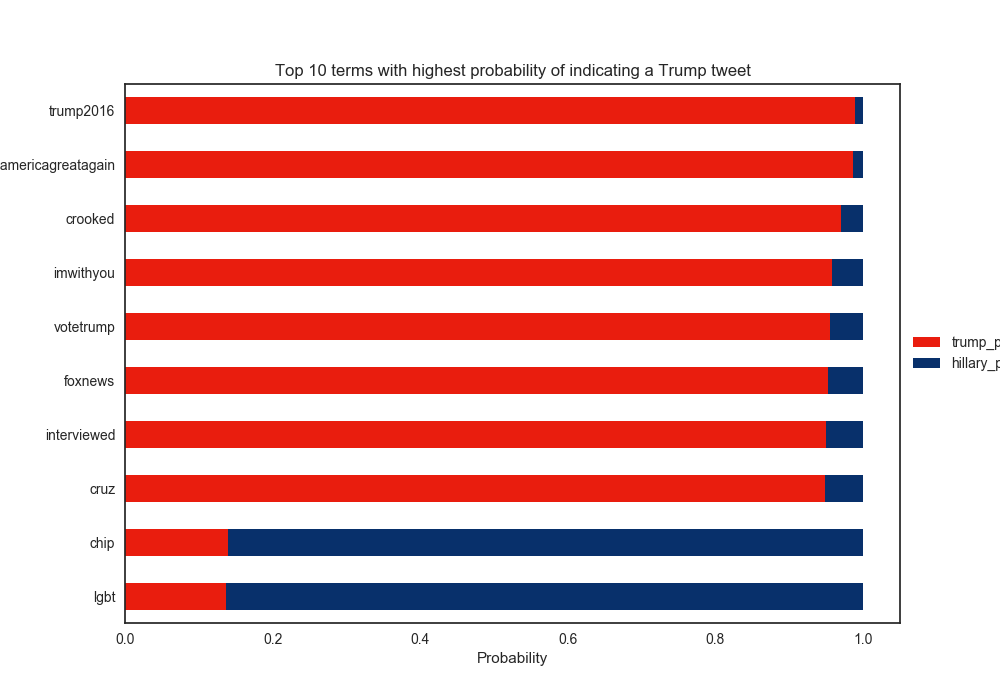
***Πίνακας 9: Confusion Matrix για το σύνολο ελέγχου***

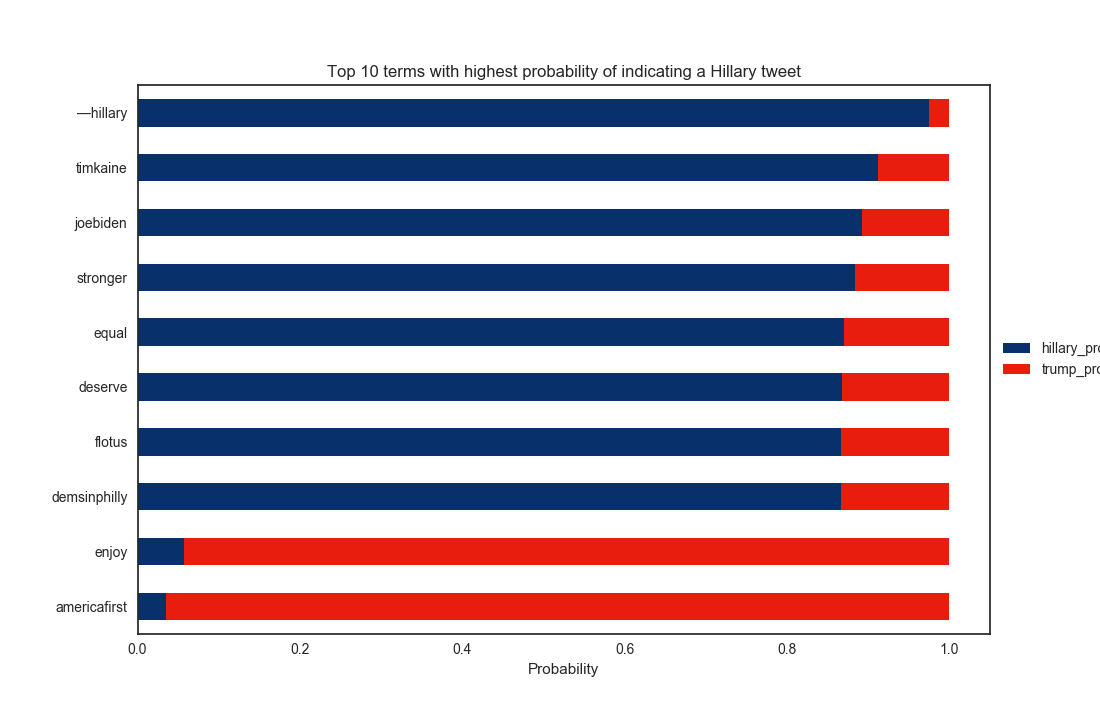
**Accuracy on test set**: 92.96%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **Hillary** | 0.92 | 0.94 | 0.93 | 547 |
| **Trump** | 0.94 | 0.93 | 0.94 | 598 |
| **avg / total** | 0.93 | 0.93 | 0.93 | 1145 |

***Πίνακας 10: Μετρικές αξιολόγησης για το σύνολο ελέγχου***

Στη συνέχεια θα χρησιμοποιήσουμε λέξεις με υψηλό TF-IDF score για να φτιάξουμε tweets, προκειμένου να εξετάσουμε την απόδοση και κατ’ επέκταση την αποτελεσματικότητα του αλγόριθμου που χρησιμοποιήσαμε.





Τέλος, βρίσκοντας κάποια δημοφιλή tweets και για τους δύο υποψηφίους στο Internet, κάνω κάποια ενδεικτικά tests.

#Popular tweets for Trump

print (nb\_detector.predict\_proba(["How long did it take your staff of 823 people to think that up--and where are your 33,000 emails that you deleted? "])[0])

print (nb\_detector.predict\_proba(["TODAY WE MAKE AMERICA GREAT AGAIN!"])[0])

print (nb\_detector.predict\_proba(["The media is spending more time doing a forensic analysis of Melania's speech than the FBI spent on Hillary's emails."])[0])

#Popular Tweets for Hillary

print (nb\_detector.predict\_proba(["Donald Trump called her Miss Piggy and Miss Housekeeping. Her name is Alicia Machado."])[0])

print (nb\_detector.predict\_proba(["Women have the power to stop."])[0])

print (nb\_detector.predict\_proba(["Delete your account."])[0])

#Top Trump tweets

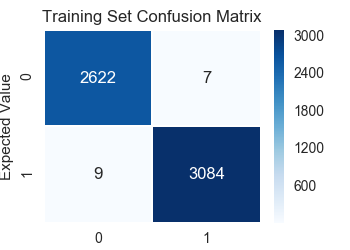
1. [ 0.42777501 0.57222499] 🡪 Trump
2. [ 0.13387538 0.86612462] 🡪 Trump
3. [ 0.10656411 0.89343589] 🡪 Trump

#Top Hillary tweets

1. [ 0.72442626 0.27557374] 🡪 Hillary
2. [ 0.65540288 0.34459712] 🡪 Hillary
3. [ 0.65449019 0.34550981] 🡪 Hillary

# Κατηγοριοποίηση με SVM

**Accuracy on training set:** 99.72%

****

***Πίνακας 11: Confusion Matrix για το σύνολο εκπαίδευσης***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **Hillary** | 1.00 | 1.00 | 1.00 | 2629 |
| **Trump** | 1.00 | 1.00 | 1.00 | 3093 |
| **avg / total** | 1.00 | 1.00 | 1.00 | 5722 |

***Πίνακας 12: Μετρικές αξιολόγησης για το σύνολο εκπαίδευσης***

Στη συνέχεια θα εκπαιδεύσουμε τις παραμέτρους του LinearSVC, προκειμένου να επιλέξουμε τον καλύτερο Classifier\_kernel και παράμετρο C.

|  |  |  |  |
| --- | --- | --- | --- |
| **classifier\_\_kernel** | **classifier\_\_C** | **'classifier\_\_gamma** | **Mean/Std** |
| Linear | 1 | - | 0.93424/0.00876 |
| Linear | 10 | - | 0.92899/0.00867 |
| Linear | 100 | - | 0.92353/0.01312 |
| Linear | 1000 | - | 0.91676/0.01231 |
| rbf | 1 | 0.001 | 0.54009/0.00025 |
| rbf | 1 | 0.0001 | 0.54009/0.00025 |
| rbf | 10 | 0.001 | 0.59537/0.01116 |
| rbf | 10 | 0.0001 | 0.54009/0.00025 |
| rbf | 100 | 0.001 | 0.92003/0.00846 |
| rbf | 100 | 0.0001 | 0.59559/0.01092 |
| rbf | 1000 | 0.001 | 0.93511/0.01002 |
| rbf | 1000 | 0.0001 | 0.91982/0.00809 |

***Πίνακας 13: Απόδοση του SVM σύμφωνα με τις παραμέτρους του***

For Linear Kernel and C = 1 and also for

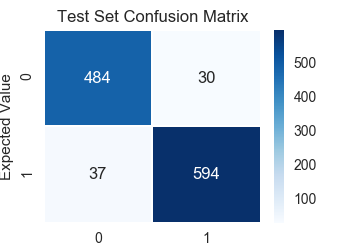
rbf Kernel, gamma = 0.001, C = 1000

Θα αξιολογήσουμε επίσης και το σύνολο ελέγχου(test set)

**Accuracy on test set:** 93.47%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **Hillary** | 0.93 | 0.94 | 0.94 | 514 |
| **Trump** | 0.95 | 0.94 | 0.95 | 631 |
| **avg / total** | 0.94 | 0.94 | 0.94 | 1145 |

***Πίνακας 14: Μετρικές αξιολόγησης για το σύνολο ελέγχου***



***Πίνακας 15: Confusion Matrix για το σύνολο ελέγχου***

#Top tweets for Trump

1. [ 0.09225056 0.90774944] 🡪 Trump
2. [ 0.05996241 0.94003759] 🡪 Trump
3. [ 0.01323039 0.98676961] 🡪 Trump

#Top tweets for Hillary

1. [ 0.96624967 0.03375033] 🡪 Hillary
2. [ 0.64675985 0.35324015] 🡪 Hillary
3. [ 0.98363469 0.01636531] 🡪 Hillary

# Σύγκριση αποτελεσμάτων

Παρατηρούμε πως με το **LinearSVC** παίρνουμε καλύτερα αποτελέσματα απ’ ότι με τον **Naïve Bayes**, όσον αφορά την αποδοτικότητα τις πρόβλεψης κατηγοριοποίησης τυχαίων tweets σε έναν από τους 2 υποψηφίους.