
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΚΑΤΑΝΕΜΗΜΕΝΩΝ ΣΥΣΤΗΜΑΤΩΝ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗΣ ΓΝΩΣΗΣ


# Μελέτη Αποδοτικότητας Αλγορίθμων Μηχανικής Μάθησης Για Παραμετροποιημένες Εισόδους

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της


Μαρίνας Α. Μαγκαφά

Επιβλέπων: Θεοδώρα Βαρβαρίγου,
Καθηγήτρια Ε.Μ.Π



Αθήνα, Σεπτέμβριος 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΚΑΤΑΝΕΜΗΜΕΝΩΝ ΣΥΣΤΗΜΑΤΩΝ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗΣ ΓΝΩΣΗΣ

# Μελέτη Αποδοτικότητας Αλγορίθμων Μηχανικής Μάθησης Για Παραμετροποιημένες Εισόδους

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Μαρίνας Α. Μαγκαφά

Επιβλέπων: Θεοδώρα Βαρβαρίγου,
Καθηγήτρια Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 28η Σεπτεμβρίου 2018.

| | | |
|---|---|---|
| ..... | ..... | ..... |
| Θεοδώρα Βαρβαρίγου | Εμμανουήλ Βαρβαρίγος | Βασίλειος Λούμος |
| Καθηγήτρια Ε.Μ.Π. | Καθηγήτρια Ε.Μ.Π. | Καθηγήτρια Ε.Μ.Π. |

Αθήνα, Σεπτέμβριος 2018

**ΜΑΡΙΝΑ Α. ΜΑΓΚΑΦΑ**
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

# Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαιτέρως την καθηγήτρια Θεοδώρα Βαρβαρίγου για την ευκαιρία που μου έδωσε να εργαστώ στον τομέα της ανάλυσης συναισθημάτων, στο Εργαστήριο Κατανεμημένων Συστημάτων και Διαχείρισης Γνώσης.

Επίσης θα ήθελα να εκφράσω τις ευχαριστίες μου στον Μεταδιδακτορικό Ερευνητή Βρεττό Μούλο για την καθοδήγηση, τις πολύτιμες συμβουλές του και τον χρόνο που αφιέρωσε καθ' όλη τη διάρκεια εκπόνησης αυτής της διπλωματικής εργασίας.

Τέλος ευχαριστώ την οικογένεια μου και τους ανθρώπους που ήταν δίπλα μου και με στήριξαν κατά τη διάρκεια των σπουδών μου.

# Περίληψη

Ο τομέας της ανάλυσης συναισθημάτων αναπτύχθηκε τα τελευταία χρόνια από την ανάγκη για εξαγωγή γνώμης και συναισθημάτων με αυτοματοποιημένο τρόπο από μεγάλο όγκο δεδομένων παραγόμενων από χρήστες, που διατίθεται ελεύθερα στο διαδίκτυο. Για τον σκοπό αυτό γίνεται χρήση μεθόδων που στηρίζονται σε τεχνικές μηχανικής μάθησης. Με την εκπαίδευση των αλγορίθμων μηχανικής μάθησης με χρήση χαρακτηριστικών κειμένων που γνωρίζουμε την πολικότητα τους, κατασκευάζονται μοντέλα για την πρόβλεψη της πολικότητας νέων κειμένων.

Στο πλαίσιο αυτό, η παρούσα εργασία μελετά την ανάλυση συναισθήματος σε δεδομένα κριτικής ταινιών και επιχειρήσεων, με τη χρήση τριών αλγορίθμων μηχανικής μάθησης, με σκοπό την ανίχνευση της πολικότητας τους. Πιο συγκεκριμένα, εξετάζουμε την μεταβολή της απόδοσης κάθε αλγορίθμου, σε κάθε βάση δεδομένων, για διαφορετικές παραμέτρους του συνόλου εκπαίδευσης. Οι βασικές παράμετροι των μοντέλων, που μελετάμε την επίδρασή τους, είναι το πλήθος των δεδομένων, το είδος της βάσης δεδομένων και η χρήση κειμένων στο σύνολο εκπαίδευσης με πιο έντονη ή με περισσότερο αμφιλεγόμενη έκφραση της πολικότητας τους.

Λέξεις Κλειδιά: ανάλυση συναισθήματος, επιβλεπόμενη μηχανική μάθηση, ταξινόμηση πολικότητας κειμένου, n-gram γράφοι, bag of words αναπαράσταση, απλοϊκό μοντέλο Bayes

# Abstract

The field of sentiment analysis has been developed in recent years by the need for opinion and sentiment extraction in an automated way from a large volume of user-generated data, freely available on the Internet. For this purpose, methods based on machine learning techniques are used. By training the machine learning algorithms, using attributes from texts which we know their polarity, models are constructed to predict the polarity of new texts.

In this context, this thesis studies the sentiment analysis in movie and business reviews data, using three machine learning algorithms, in order to detect their polarity. More specifically, we examine how the performance of each algorithm changes, in each database for different parameters of the training set. The basic parameters of the models in the study are the number of data, the type of database and the use of texts in the training set with a more pronounced or more controversial expression of their polarity.

Keywords: sentiment analysis, supervised machine learning, polarity text classification, n-gram graphs, bag of words model, naïve Bayes model

# Πίνακας Περιεχομένων

# Πίνακας Πινάκων

# Πίνακας σχημάτων

# Κεφάλαιο 1ο : ΕΙΣΑΓΩΓΗ

## 1.1 Ανάλυση Συναισθήματος

### 1.1.1 Ορισμός

Η ανάλυση συναισθημάτων είναι η διαδικασία που στόχος της είναι να καθορίσει τη στάση ενός ατόμου σχετικά με κάποιο θέμα ή τη συνολική πολικότητα - συναισθηματική αντίδραση του σε ένα έγγραφο, αλληλεπίδραση ή γεγονός, ταξινομώντας τη σε θετική, αρνητική και ουδέτερη. Η στάση αυτή μπορεί να είναι κρίση, αξιολόγηση ή συναισθηματική κατάσταση. Σε ένα κείμενο όμως, τις περισσότερες φορές, περιλαμβάνονται περισσότερες από μία απόψεις, πάνω σε κάποια θέματα και για αυτό το λόγο συνηθίζεται να προηγείται η εξαγωγή απόψεων, πριν από την κατηγοριοποίηση της πολικότητας του κειμένου. Συνεπώς είναι δυνατή η εφαρμογή της ανάλυσης συναισθήματος είτε σε ολόκληρο έγγραφο, είτε σε επιμέρους παραγράφους, προτάσεις ή και σε χαρακτηριστικά του κειμένου. Η διαδικασία αυτή της εξαγωγής απόψεων από ένα κείμενο ονομάζεται εξόρυξη γνώμης. Οι δύο όροι, της ανάλυσης συναισθήματος και της εξόρυξης γνώμης, έχουν παρόμοια έννοια, εμπίπτουν και οι δύο στο πεδίο της ανάλυσης υποκειμενικότητας και χρησιμοποιούνται για να αναφέρουν τον ίδιο σκοπό της ανάλυσης συναισθήματος. [1]

Η ανάλυση συναισθημάτων, πιο συγκεκριμένα αναφέρεται στην επεξεργασία της φυσικής γλώσσας, της ανάλυσης κειμένου, της υπολογιστικής γλωσσολογίας και της βιομετρίας για τον συστηματικό εντοπισμό, εξαγωγή, ποσοτικοποίηση και μελέτη των συναισθηματικών καταστάσεων και των υποκειμενικών πληροφοριών.

### 1.1.2 Εφαρμογές

Η ανάλυση συναισθήματος αποτελεί ένα σημαντικό πεδίο έρευνας, ειδικά τα τελευταία χρόνια, που τείνει να αποκτήσει ακόμη μεγαλύτερες διαστάσεις, λόγω της εισαγωγής της έννοιας των Big Data στον κλάδο της επιστήμης των υπολογιστών. Εφαρμόζεται, ήδη, ευρέως στον επιχειρηματικό τομέα, καθώς οι υπεύθυνοι επιχειρήσεων μπορούν να λάβουν αποφάσεις βασιζόμενοι στις γνώμες των χρηστών, σχετικά με τα προϊόντα τους.

Η εφαρμογή της, όμως δεν περιορίζεται μόνο στον κλάδο των επιχειρήσεων, καθώς την συναντάμε και στον τομέα του χρηματιστηρίου, σε άρθρα ενημέρωσης και στον κλάδο της πολιτικής. Ειδικότερα στον τομέα της πολιτικής, με την εφαρμογή της ανάλυσης συναισθημάτων στις γνώμες των πολίτων από δημοσιεύσεις στα κοινωνικά δίκτυα, υπάρχει η δυνατότητα πρόβλεψης των αποτελεσμάτων πολιτικών εκλογών. Τα κοινωνικά δίκτυα

θεωρούνται μία καλή πηγή πληροφοριών για την εξαγωγή γνώμης του κοινού, διότι οι χρήστες εκφράζουν και κοινοποιούν ελεύθερα τις απόψεις τους σε αυτά.

### 1.1.3 Προκλήσεις στη διαδικασία της ανάλυσης συναισθήματος

Η διαδικασία της ανάλυσης συναισθήματος σε ένα κείμενο αποτελεί δύσκολο εγχείρημα, λόγω των ιδιαιτεροτήτων που παρουσιάζουν σε σχέση με άλλα δεδομένα. Κάποιες από τις κυριότερες δυσκολίες που αντιμετωπίζονται είναι οι ακόλουθες: [2]

- Λεξιλογική ή σημασιολογική αμφισημία. Πολλές φορές σε ένα κείμενο, συναντάμε λέξεις με πολλαπλό νόημα, είτε φράσεις που εκφράζουν διαφορετικές σημασίες, ανάλογα με το περιεχόμενο του κειμένου στο οποίο εμφανίζονται.

- Σχετική εξάρτηση. Μία έννοια σε μία γλώσσα εκφράζεται από έναν συνδυασμό λέξεων και φράσεων.

- Θόρυβος στα δεδομένα. Ορθογραφικά λάθη, συντακτικά ή και γραμματικά λάθη και συντομογραφίες στο γραπτό λόγο. Συχνό σε περιπτώσεις ανάλυσης δεδομένων που περιέχουν δημοσιεύσεις χρηστών σε κοινωνικά δίκτυα.

- Σαρκασμός και ειρωνεία. Χρήση επιθέτων από τον συγγραφέα, με σκοπό να εκφράσουν την ακριβώς αντίθετη σημασία τους.

- Λεξιλόγιο. Στα περισσότερα κείμενα που εφαρμόζεται ανάλυση συναισθημάτων στο διαδίκτυο, χρησιμοποιείται ανεπίσημη γλώσσα, που περιλαμβάνει αργκό, νεολογισμούς, επιμήκυνση φθόγγων, χρήση κεφαλαίων και σημείων στίξης για έμφαση και χρήση συντομογραφιών. Αυτή η ιδιαίτερη μορφή λεξιλογίου δεν ευνοεί τη χρήση λεκτικών αναλυτών.

## 1.2 Σκοπός της εργασίας

Σε αυτή την εργασία εφαρμόζουμε ανάλυση συναισθημάτων, με σκοπό την ανίχνευση πολικότητας κριτικών χρηστών σε δύο ευρέως γνωστές βάσεις δεδομένων, με χρήση λεξικού και αλγορίθμων επιβλεπόμενης μηχανικής μάθησης. Πιο συγκεκριμένα, η ταξινόμηση αφορά κριτικές χρηστών στις βάσεις δεδομένων του IMDB και του YELP κάνοντας χρήση του αλγορίθμου n-grams, χρησιμοποιώντας διαφορετικές παραμέτρους εισόδου σε κάθε εκτέλεση.

Σύμφωνα με τις παραμέτρους εισόδου που χρησιμοποιούνται η απόδοση του αλγορίθμου μεταβάλλεται και σκοπός της εργασίας είναι να εντοπίσουμε την εξάρτηση της κάθε παραμέτρου εισόδου με την απόδοση και πως επιτυγχάνουμε μέγιστη απόδοση και αξιοπιστία, χωρίς υπάρξει τροποποίηση στον κώδικα του αλγορίθμου.

## 1.3 Οργάνωση κειμένου εργασίας

Η εργασία αυτή αποτελείται από τέσσερα κεφάλαια συνολικά. Αυτό είναι το πρώτο κεφάλαιο το οποίο έχει σκοπό την εισαγωγή στο γενικό θέμα της εργασίας, της ανάλυσης συναισθημάτων, καθώς και την ανάλυση του σκοπού και της διάρθρωσης της εργασίας.

Στο δεύτερο κεφάλαιο αναλύονται οι βασικές έννοιες της εργασίας, όπως η έννοια της μηχανικής μάθησης, οι αλγόριθμοι που χρησιμοποιήθηκαν και τα μοντέλα ταξινομητών. Επιπλέον, επεξηγούνται τα στάδια υλοποίησης του προγράμματος.

Στο τρίτο κεφάλαιο παρουσιάζονται οι βάσεις δεδομένων στις οποίες στηρίχτηκαν τα πειράματα αυτής της εργασίας και χαρακτηριστικά αυτών. Το δεύτερο μέρος του κεφαλαίου αναφέρεται στην οργάνωση των πειραμάτων και τις παραμέτρους που χρησιμοποιήθηκαν.

Στο τέταρτο και τελευταίο κεφάλαιο παρουσιάζονται τα πειραματικά μας αποτελέσματα και τα συμπεράσματα που προκύπτουν από αυτά, καθώς και προτεινόμενες μελλοντικές επεκτάσεις της έρευνας.

# Κεφάλαιο 2$^{ο}$ : Βασικές έννοιες

## 2.1 Μηχανική Μάθηση

### 2.1.1 Ορισμός

Η μηχανική μάθηση αποτελεί υποπεδίο της τεχνητής νοημοσύνης στον τομέα της επιστήμης των υπολογιστών. Αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα, μέσω στατιστικής ανάλυσης τους, και να κάνουν προβλέψεις σχετικά με αυτά. [3] Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Ο κυριότερος στόχος ενός μαθητευόμενου είναι η γενίκευση της εμπειρία του. [4] Γενίκευση θεωρείται η ικανότητα μιας μηχανής μάθησης να αποδίδει με ακρίβεια σε καινούριες, πρωτόγνωρες εργασίες, εφόσον αρχικά έχει εκπαιδευτεί σε ένα σύνολο δεδομένων εκπαίδευσης. Τα δεδομένα που χρησιμοποιούνται προς εκπαίδευση θεωρούνται αντιπροσωπευτικά του χώρου καταστάσεων και η μηχανή είναι ικανή για την κατασκευή ενός γενικού μοντέλου που θα έχει τη δυνατότητα να κάνει προβλέψεις για νέες καταστάσεις με επαρκή ακρίβεια.

### 2.1.2 Είδη μηχανικής μάθησης

Ο τομέας της μηχανικής μάθησης διακρίνεται σε τρεις περιπτώσεις μάθησης, με κριτήριο το βαθμό παρέμβασης του ανθρώπου στη διαδικασία της μάθησης:

- Μάθηση με επίβλεψη (supervised learning) ή μάθηση με παραδείγματα (learning from examples): το σύστημα καλείται να "μάθει" μια έννοια ή συνάρτηση από ένα σύνολο δεδομένων, η οποία αποτελεί περιγραφή ενός μοντέλου.
- Μάθηση χωρίς επίβλεψη (unsupervised learning) ή μάθηση από παρατήρηση (learning from observation): το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.
- ενισχυτική μάθηση (reinforcement learning): αποτελείται από μια οικογένεια τεχνικών στις οποίες το σύστημα μάθησης προσπαθεί να μάθει μέσα από την άμεση αλληλεπίδραση του με το περιβάλλον.

### 2.1.3 Ταξινόμηση

Όπως αναφέρθηκε, στη διαδικασία της επιβλεπόμενης μάθησης, δίνεται στη μηχανή ένα αντικείμενο ή μια κατάσταση ως είσοδο και η αναμενόμενη έξοδος που επιθυμείται να επιστρέφει η μηχανή. Το αντικείμενο ή η κατάσταση περιγράφεται ως ένα σύνολο χαρακτηριστικών, το οποίο εκφράζεται με ένα διάνυσμα αριθμών (διάνυσμα χαρακτηριστικών). [4] Οι τιμές εισόδου και εξόδου των διανυσμάτων μπορεί να είναι είτε συνεχείς, είτε διακριτές. Όταν βρισκόμαστε σε περίπτωση μάθησης με διακριτές τιμές εξόδου, η διαδικασία που ακολουθούμε ονομάζεται μάθηση ταξινόμησης και οι διακριτές τιμές εξόδου ονομάζονται κλάσεις. Κατά την εκπαίδευση του αλγορίθμου, λοιπόν, στο στάδιο της εκπαίδευσης έχουμε ένα επιπλέον χαρακτηριστικό στο διάνυσμα, το διάνυσμα κλάσης, που εκφράζει την αναμενόμενη έξοδο.

Εφόσον ολοκληρωθεί η εκπαίδευση του αλγορίθμου, το μοντέλο ελέγχεται με νέα δεδομένα, που καλείται να προβλέψει την έξοδο. Κατά τη διαδικασία ελέγχου το χαρακτηριστικό κλάσης δεν χρησιμοποιείται για την εξαγωγή του αποτελέσματος, συνεπώς η μηχανή δεν γνωρίζει την κλάση στην οποία ανήκει το ενδεχόμενο και επιστρέφει μία πρόβλεψη με χρήση της συνάρτησης που κατασκεύασε κατά την εκπαίδευση της. Το χαρακτηριστικό κλάσης χρησιμοποιείται τελικά, για τη σύγκριση με την έξοδο που πρόβλεψε το μοντέλο.

### 2.1.4 Μετρικές αξιολόγησης μοντέλων

Μία συχνά χρησιμοποιούμενη μετρική αξιολόγησης είναι η ακρίβεια ταξινόμησης των μοντέλων (accuracy) και υπολογίζεται από την παρακάτω σχέση:

$$\text{Accuracy} = \frac{\text{number of correclty classified test instances}}{\text{number of test instances}}$$

Στη συγκεκριμένη εργασία, βρισκόμαστε στην περίπτωση επιβλεπόμενης μάθησης κατά την οποία το σύστημα «μαθαίνει» μία συνάρτηση, με δεδομένα ένα σύνολο παραδειγμάτων εισόδου και εξόδου το οποίο ονομάζεται σύνολο εκπαίδευσης (training set). Στην περίπτωσή μας, χρησιμοποιούμε τις κριτικές ταινιών στο IMDB και τις κριτικές του YELP, για τις οποίες γνωρίζουμε την πολικότητα στην οποία ανήκουν, για να κατασκευάσουμε το σύνολο εκπαίδευσης της μηχανής. Ως είσοδο στον αλγόριθμο για εκπαίδευση, δίνουμε τα χαρακτηριστικά των κειμένων αυτών και την κλάση στην οποία ανήκουν. Έτσι η μηχανή, ύστερα από το στάδιο της εκπαίδευσης, μπορεί να προβλέπει για κάθε νέο κείμενο με νέα χαρακτηριστικά σε ποια κλάση ανήκει. Επομένως, αυτή η διαδικασία ανάγεται σε ένα πρόβλημα ταξινόμησης.

## 2.2 Αλγόριθμος Bag of words

### 2.2.1 Ορισμός

Το μοντέλο Bag of words ορίζεται ως μία απλοποιημένη αναπαράσταση, που χρησιμοποιείται στην επεξεργασία φυσικής γλώσσας. Σύμφωνα με αυτό, ένα κείμενο αναπαρίσταται με ένα σύνολο λέξεων, που συλλέγετε από το περιεχόμενο του κειμένου, αγνοώντας πληροφορίες όπως η σειρά εμφάνισής τους ή η γραμματική. Η μοναδική πληροφορία του αρχικού κειμένου, που έχει σημασία σε αυτό το μοντέλο, είναι η συχνότητα εμφάνισης των λέξεων. [5][6]

### 2.2.2 Εφαρμογές

Ο αλγόριθμος αυτός είναι πολύ εύκολος στην κατανόηση και την υλοποίηση και για αυτό το λόγο εφαρμόζεται συχνά σε μεθόδους επεξεργασίας φυσικής γλώσσας, όπως αναγνώριση γλώσσας, φιλτράρισμα ανεπιθύμητων μηνυμάτων, κατηγοριοποίηση κειμένου, ανάλυση συναισθήματος κ.α., ακόμη και σε εφαρμογές στον τομέα της όρασης υπολογιστών.

### 2.2.3 Αναπαράσταση κειμένων με χρήση Bag of words

Η αναπαράσταση κάθε κειμένου, με τη μέθοδο αυτή, επιτυγχάνεται με την χρήση διανύσματος. Κάθε θέση του διανύσματος αντιστοιχεί σε μία λέξη του λεξιλογίου (vocabulary set) που έχουμε δημιουργήσει, και είτε λαμβάνει τιμές της συχνότητας εμφάνισης κάθε λέξης, είτε απλώς αναπαριστά την ένδειξη εμφάνισής της. Οι συχνότητες εμφάνισης παίρνουν ακέραιες τιμές και δηλώνουν το πλήθος φορών που συναντάμε μία λέξη στο κείμενο, ενώ η ένδειξη εμφάνισης είναι δυαδικό ψηφίο που παίρνει την τιμή αληθές όταν η αντίστοιχη λέξη του λεξιλογίου εμφανίζεται στο κείμενο και ψευδές στην αντίθετη περίπτωση. Το λεξιλόγιο που χρησιμοποιούμε σε αυτή τη διαδικασία, δημιουργείται αποθηκεύοντας λέξεις των κειμένων του συνόλου εκπαίδευσης της μηχανής και με τη χρήση των εργαλείων της βιβλιοθήκης Weka[1]. Η μέθοδος αυτή χρησιμοποιεί ως διάνυσμα χαρακτηριστικών, το διάνυσμα λέξεων της προσέγγισης bag of words με την προσθήκη του χαρακτηριστικού κλάσης, για την εκπαίδευση και τον έλεγχο του ταξινομητή.

[1]http://www.cs.waikato.ac.nz/_ml/weka/

## 2.3 Αλγόριθμος N-gram

### 2.3.1 Ορισμός n-gram

Στους τομείς της υπολογιστικής γλωσσολογίας και της πιθανότητας, ένα n-gram αποτελεί μια συνεχή ακολουθία n στοιχείων από μια ακολουθία κειμένου ή ομιλίας. Τα στοιχεία εξαγωγής, αναλόγως με τη μορφή της ακολουθίας και το σκοπό τις εφαρμογής, μπορούν να είναι φωνήματα, συλλαβές, γράμματα ή λέξεις. Στην περίπτωση που έχουμε n χαρακτήρες ονομάζονται n-grams χαρακτήρων (character n-grams) και στην περίπτωση που έχουμε ν λέξεις ονομάζονται n-grams λέξεων (word n-grams).

Τα n-grams που προκύπτουν από ένα κείμενο είναι επικαλυπτόμενα, όπως παρατηρούμε στο επόμενο παράδειγμα εξαγωγής n-grams χαρακτήρων (για n=1,2,3):

| Παράδειγμα | 1-gram | 2-gram | 3-gram |
|---|---|---|---|
| ...sample_text... | ..., s, a, m, p, l, e, _, t, e, x, t, ... | ..., sa, am, mp, pl, le, e_, _t, te, ex, xt, ... | ..., sam, amp, mpl, ple, le_, e_t, _te, tex, ext, ... |

Το βασικό σημείο των n-grams είναι ότι με τη χρήση τους μελετάμε τη γλωσσική δομή από στατιστική άποψη. Όσο μεγαλύτερο το n-gram, τόσο περισσότεροι οι πόροι και η υπολογιστική ισχύς που χρειάζεται για να εκπαιδευτεί το μοντέλο.

### 2.3.2 Εφαρμογές

Η εξαγωγή n- gram χρησιμοποιείται ευρέως στην στατιστική ανάλυση της φυσικής γλώσσας. Χρήσιμη εφαρμογή έχει στην αναγνώριση ομιλίας, όπου τα φωνήματα και οι ακολουθίες φωνημάτων μοντελοποιούνται με βάση την λογική των n-gram. [6]

Τα n-grams μπορούν επίσης να χρησιμοποιηθούν για ακολουθίες σχεδόν οποιουδήποτε τύπου δεδομένων. Για παράδειγμα, έχουν χρησιμοποιηθεί για την εξαγωγή χαρακτηριστικών από μεγάλα σύνολα δορυφορικών εικόνων, για τον προσδιορισμό των τμημάτων της Γης που απεικονίζει μια συγκεκριμένη εικόνα. Επίσης, στον τομέα τις βιοιατρικής, έχουν μεγάλη επιτυχία ως πρώτη διερεύνηση της αναζήτησης γενετικής ακολουθίας και στην ταυτοποίηση του είδους από το οποίο προέρχονται σύντομες αλληλουχίες DNA.

Τέλος, με την χρήση n-gram γράφων μπορούμε να πετύχουμε εξαγωγή περίληψης κειμένου (text summarization),[7][8] γεγονός που κάνει τον αλγόριθμο των n-grams να

εφαρμόζεται αποτελεσματικά σε περιεχόμενο μέσων κοινωνικής δικτύωσης, για την ανάλυση συναισθήματος.[9]

### 2.3.3 Εφαρμογή του n-gram αλγορίθμου στην Ανάλυση Συναισθήματος:

Στην δική μας περίπτωση χρησιμοποιείται το μοντέλο κειμένου ως αναπαράσταση ενός συνόλου κειμένων κριτικών ταινιών και επιχειρήσεων που έχουν κοινή γνώμη (θετική ή αρνητική κριτική).

#### 2.3.3.1 Αναπαράσταση κειμένου με n-gram γράφο χαρακτήρων:

Για την δημιουργία του μοντέλου κειμένου αναπαριστούμε κάθε κείμενο με έναν n-gram γράφο ως εξής:

Έστω ένα τυχαίο κείμενο κριτικής: “sample text”

Επιλέγουμε n ίσο με 3 και εξάγουμε τα 3-grams χαρακτήρων από το κείμενο :

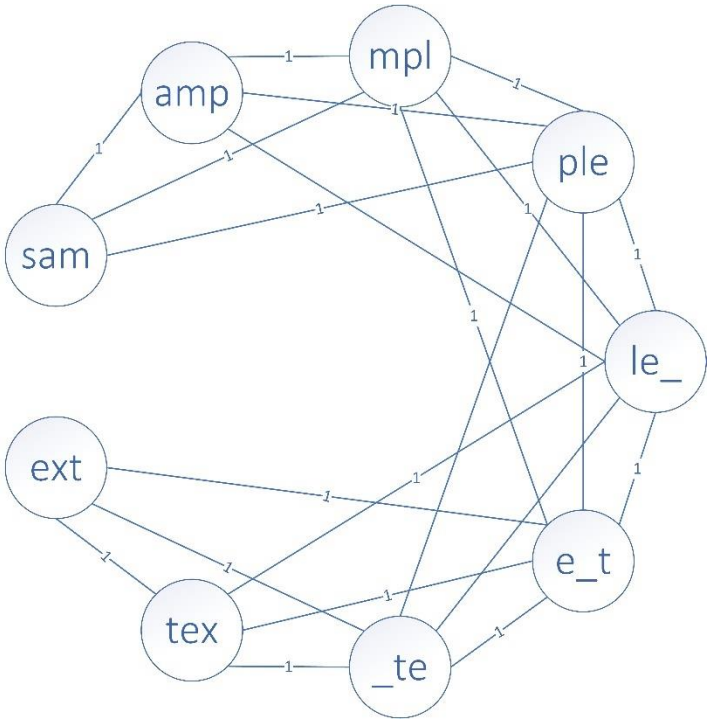
“sample text ”

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **sam** | **amp** | **mpl** | **ple** | **le_** | **e_t** | **_te** | **tex** | **ext** |

Στη συνέχεια, ξεκινάει η διαδικασία δημιουργίας του 3-gram γράφου. Αρχικά, δημιουργούμε για κάθε τρίγραμμα έναν κόμβο στο γράφο. Ο γράφος αυτός θέλουμε τελικά να απεικονίζει την γειτνίαση των n-grams, ώστε να μας δίνει πληροφορίες για την σειρά εμφάνισης των χαρακτήρων στο κείμενο.

Για τον σχεδιασμό του γράφου χρησιμοποιούμε τη μεταβλητή “μήκος παραθύρου” (Dwin), στην οποία δίνουμε μια τιμή σύμφωνα με το πλήθος των n-grams, που επιθυμούμε να θεωρηθούν γειτονικά με το υπό μελέτη n-gram. Σύμφωνα με την τιμή του παραθύρου αν ένα n-gram βρίσκεται μέσα στο διάστημα του παραθύρου τότε τοποθετούμε μια ακμή μεταξύ των δύο κόμβων. Το βάρος της ακμής ορίζεται από το πλήθος εμφάνισης του ζεύγους των n-grams εντός του παραθύρου.

Ο τελικός γράφος του κειμένου του παραδείγματος, για τιμή παραθύρου ίση με 3, φαίνεται στο σχήμα 2.1:




*Σχήμα 1 Αναπαράσταση κειμένου με χρήση n-gram χαρακτήρων*

### 2.3.3.2 Αναπαράσταση κειμένου με n-gram γράφο λέξεων (word graph):

Με σκοπό την ακριβέστερη αναπαράσταση ενός κειμένου κριτικής μέσω γράφων χρησιμοποιήθηκε και μία νέα μέθοδος που προκύπτει από το συνδυασμό της μεθόδου bag of words και των n-gram γράφων χαρακτήρων. Αυτή η νέα μέθοδος στην εργασία αυτή ονομάζεται μέθοδος word graphs και υλοποιείται με την ίδια λογική του αλγορίθμου εξαγωγής n-gram γράφων χαρακτήρων, με τη διαφορά ότι αντί για ακολουθίες χαρακτήρων έχουμε ακολουθίες λέξεων και αντίστοιχα εξαγωγή γράφου λέξεων.

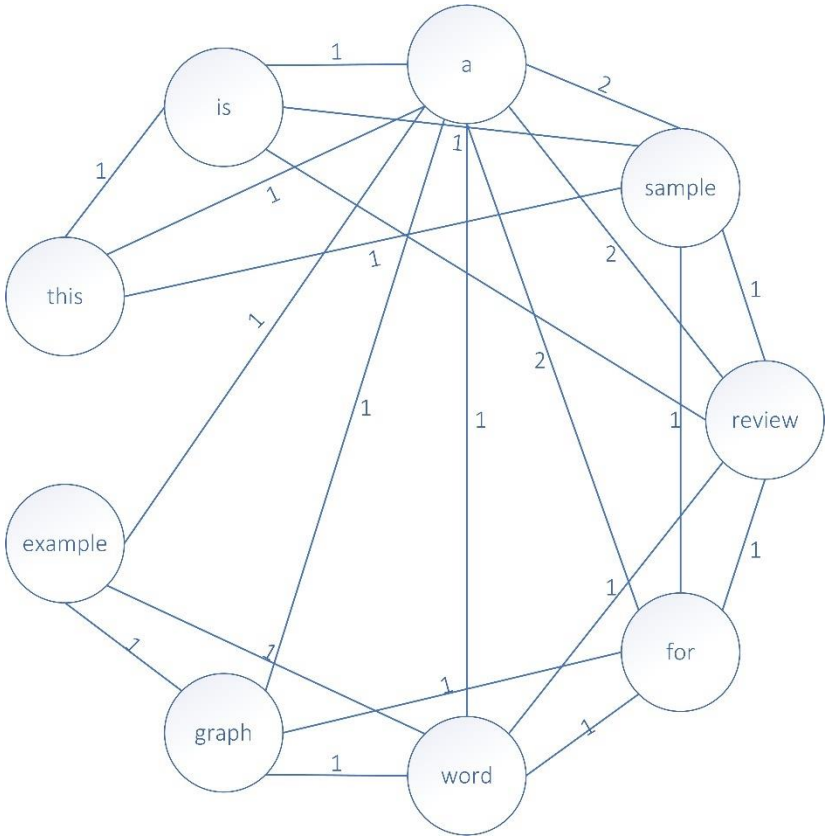
Πιο συγκεκριμένα, για την εξαγωγή ενός γράφου λέξεων ενός κειμένου ακολουθούμε την εξής διαδικασία:

Έστω το κείμενο κριτικής :

"this is a sample review for a word graph example"

Οι κόμβοι του γράφου που δημιουργούνται για μέγεθος παραθύρου = 1 λέξη είναι :

| this | is | a | sample | review | for | word | graph | example |
|---|---|---|---|---|---|---|---|---|




*Σχήμα 2 Αναπαράσταση κειμένου με χρήση n-gram λέξεων*

Ο γράφος που προκύπτει για μέγεθος n = 3 είναι παρουσιάζεται στο σχήμα 2.2. Αξίζει να προσέξουμε ότι ο κόμβος «a» δημιουργείται μία μοναδική φορά και προσαρμόζουμε την τιμή του βάρους των ακμών, που εμφανίζονται πάνω από μία φορά στον γράφο.

### 2.3.3.2 Συγχώνευση των γράφων και δημιουργία μοντέλου:

Η εξαγωγή του μοντέλου γράφου γίνεται ύστερα από συγχώνευση των επιμέρους γράφων λέξεων κάθε κειμένου κριτικής του συνόλου εκπαίδευσης με κοινά χαρακτηριστικά (στην περίπτωσή μας θετική ή αρνητική βαθμολογία). Η διαδικασία συγχώνευσης των γράφων ξεκινάει με την αρχικοποίηση του γράφου μοντέλου με τον πρώτο γράφο. Στη συνέχεια προσθέτουμε στον γράφο τα νέα ζεύγη n-grams που προκύπτουν από τους επόμενους γράφους, προσθέτοντας τους νέους κόμβους και ακμές και ανανεώνοντας τα βάρη των ήδη υπαρχουσών ακμών, έτσι ώστε να αποτελούν τον μέσο όρο των βαρών. Αξίζει να δώσουμε προσοχή στη περίπτωση που μία ακμή εμφανίζεται στον γράφο μοντέλου και σε ν-1 γράφους κειμένων, τότε το βάρος της ακμής πρέπει να ανανεωθεί κατάλληλα ώστε να αποτελεί το μέσο όρο των ν βαρών. Ο αλγόριθμος, όμως, επειδή δεν αποθηκεύει τα προηγούμενα βάρη για να υπολογίζει εκ νέου

τον μέσο όρο, γνωρίζοντας το πλήθος των υπό συγχώνευση ακμών, το νέο βάρος μπορεί να υπολογιστεί από τον τύπο:

$$\text{new average} = \text{old average} + \frac{1}{n} \times (\text{new weight} - \text{old average})$$

Όπου $$\text{old average} = \frac{w_1 + w_2 + \cdots + w_{n-1}}{n-1}$$

### 2.3.3.3 Εξαγωγή χαρακτηριστικών με χρήση των γράφων:

Στην εργασία αυτή χρησιμοποιήθηκε δυϊκή ταξινόμηση (σε δύο κλάσεις) κατά την οποία ταξινομούμε κάθε ενδεχόμενο κατάστασης ως αληθές ή ψευδές (1 και 0 αντίστοιχα) και στην περίπτωση μας η ταξινόμηση αφορά την κλάση θετικής πολικότητας και την κλάση αρνητικής πολικότητας.

Από κάθε κείμενο- instance προς ταξινόμηση εξάγουμε ένα σύνολο χαρακτηριστικών το οποίο είναι αντιπροσωπευτικό για την πολικότητα του. Για τον σκοπό αυτό χρησιμοποιούνται τα μοντέλα. Αρχικά δημιουργούμε τον n-gram γράφο του κάθε κειμένου-ενδεχόμενου και στη συνέχεια τον συγκρίνουμε με κάθε γράφο πολικότητας, ώστε να εξάγουμε κάποιες συγκεκριμένες τιμές ομοιότητας για κάθε κείμενο, οι οποίες αποτελούν τα χαρακτηριστικά για την ταξινόμηση του κειμένου στην κατάλληλη πολικότητα, μέσα από την μηχανη.

Οι δείκτες ομοιότητας που χρησιμοποιήθηκαν χρησιμοποιούνται για σύγκριση γράφων στην αξιολόγηση περιλήψεων [7] και είναι οι εξής:

- Coocurrence ή Containment Similarity (CS): δηλώνει το ποσοστό των ακμών εκείνου του γράφου με τις λιγότερες ακμές που εμφανίζονται και στον μεγαλύτερο γράφο (δεν λαμβάνουμε υπόψη το βάρος των ακμών).
- Size Similarity (SS): δηλώνει την αναλογία μεγέθους των δύο γράφων.
- Value Similarity (VS): δηλώνει τον αριθμό των ακμών του μικρότερου γράφου που εμφανίζονται και στον μεγαλύτερο ακριβώς όπως και στην περίπτωση του Containment Similarity μόνο που αυτή τη φορά λαμβάνουμε υπόψη τα βάρη των ακμών.
- Normalized Value Similarity (NVS): δηλώνει την αναλογία : $\frac{VS}{SS}$

Στην πράξη σε αυτή την εργασία εξάγουμε για κάθε κείμενο κριτικής τρεις τιμές ομοιότητας με τον κάθε γράφο πολικότητας δηλαδή συνολικά έξι τιμές στο διάστημα {0,1} για κάθε κείμενο. Συμπεριλαμβάνοντας και την τιμή C της κλάσης του κειμένου δημιουργείται ένα διάνυσμα εφτά τιμών για κάθε ενδεχόμενο με τις οποίες γίνεται η εκπαίδευση της μηχανής. Αντίστοιχα λειτουργούμε για το σύνολο κειμένων ελέγχου. Στην περίπτωση μας δεν θα ασχοληθούμε με την κλάση ουδέτερης πολικότητας συνεπώς το πρόβλημα μας ανάγεται στην κατηγορία Binary Polarity Problems.

## 2.4 Μοντέλα Ταξινομητών

Το πιο δημοφιλές μοντέλο ταξινόμησης κειμένων είναι το μοντέλο Naive Bayes [10][11] το οποίο και εφαρμόζουμε στη συγκεκριμένη εργασία.

### 2.4.1 Απλοικό μοντέλο Bayes (naive Bayes)

Ο όρος naïve Bayes αναφέρεται για να δηλώσει τα μοντέλα μηχανικής μάθησης που εφαρμόζουν τον κανόνα του Bayes κατά την ταξινόμηση σε κλάσεις των ενδεχομένων κατάστασης, που εκφράζονται με ένα διάνυσμα χαρακτηριστικών. Η λέξη naïve που χαρακτηρίζει το μοντέλο, εκφράζει μία αφελή υπόθεση ανεξαρτησίας των μεταβλητών των χαρακτηριστικών, που λαμβάνουμε. Επιπλέον, στο μοντέλο αυτό η κλάση των ενδεχομένων λαμβάνει διακριτές τιμές σε ένα πεπερασμένο σύνολο, συνεπώς το μοντέλο αποτελεί μοντέλο ταξινόμησης.

Πιο συγκεκριμένα, σύμφωνα με τον κανόνα του Bayes ισχύει:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)},$$

όπου A και B είναι γεγονότα.

- P(A) και P(B) είναι οι πιθανότητες των A και B που είναι ανεξάρτητα μεταξύ τους.
- P(A | B), η υπό συνθήκη πιθανότητα, είναι η πιθανότητα του A δεδομένου του B να είναι αληθής.
- P(B | A), είναι η πιθανότητα του B δεδομένου του A να είναι αληθής.

Επομένως για την ταξινόμηση ενός ενδεχόμενου σε κάποια κλάση στην περίπτωση μας χρησιμοποιούμε τα μοντέλα Bayes. Για κάθε ενδεχόμενο υπολογίζουμε την πιθανότητα να ανήκει αυτό σε κάθε μία από τις κλάσης, σύμφωνα με τα χαρακτηριστικά του, χρησιμοποιώντας το θεώρημα του Bayes και στη συνέχεια αυτό ταξινομείται στην αντίστοιχη κλάση.

# Κεφάλαιο 3ο Ικανότητα γενίκευσης μοντέλου μηχανικής μάθησης

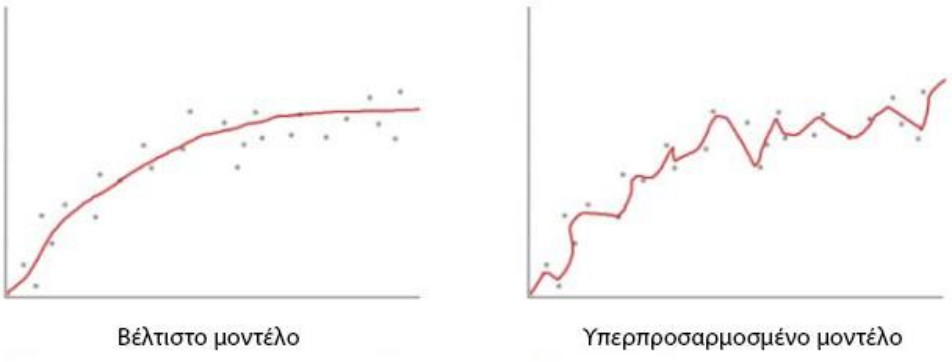
Οι αλγόριθμοι της επιβλεπόμενης μηχανικής μάθησης, όπως αναφέρεται στο προηγούμενο κεφάλαιο, επιδιώκουν την δημιουργία μοντέλων τα οποία εκπαιδεύονται από ένα σύνολο δεδομένων εκπαίδευσης με την εξαγωγή χαρακτηριστικών που προκύπτουν κατά την ανάλυση των δεδομένων πηγής. Σκοπός ενός μοντέλου μπορεί είναι να ταξινομεί νέες περιπτώσεις βάση ενός συνόλου κατηγοριών. Για να είναι εφικτή αυτή η λειτουργία θα πρέπει το μοντέλο που θα δημιουργηθεί να μπορεί να γενικεύσει. Σε αυτό το κεφάλαιο επικεντρωνόμαστε στα ζητήματα που προκύπτουν κατά τη διαδικασία κατασκευής ενός κατάλληλου μοντέλου σε ένα σύστημα μηχανικής μάθησης.

## 3.1 Υπερπροσαρμογή - Υπερμοντελοποίηση

Ένα μοντέλο που δεν μας οδηγεί στη γενίκευση της εμπειρίας του συστήματος σημαίνει ότι δεν μπορεί να κάνει ακριβείς προβλέψεις για νέα δεδομένα που συναντά. Στην περίπτωση λοιπόν, που το μοντέλο είναι προσαρμοσμένο σε ένα συγκεκριμένο σύνολο δεδομένων σε τέτοιο βαθμό ώστε να μην μπορεί να προσαρμοστεί σε νέα δεδομένα ή να προβλέψει νέες παρατηρήσεις αξιόπιστα, παρατηρείται το φαινόμενο της υπερπροσαρμογής-υπερμοντελοποίησης (Overfitting).

Το φαινόμενο αυτό προκαλείται συνήθως, όταν το στατιστικό μοντέλο του συστήματος περιέχει περισσότερες παραμέτρους από αυτές που είναι αναγκαίες και εκφράζουν τα χαρακτηριστικά των δεδομένων που μας ενδιαφέρουν, προκειμένου να επεξηγήσει όσο το δυνατόν πιο πιστά την όποια συμπεριφορά εμφανίζει το περιορισμένο πλήθος δεδομένων εκπαίδευσης. Στην ουσία σε αυτή τη περίπτωση εξάγονται εν αγνοία μας στοιχεία των δεδομένων που εμπεριέχουν κάποιο βαθμό σφάλματος ή αποτελούν τυχαίο θόρυβο και αυτά στη συνέχεια αντιπροσωπεύουν την υποκείμενη δομή του μοντέλου και αυξάνουν τον κίνδυνο μειωμένης προβλεπτικής ικανότητας του συστήματος.

Με άλλα λόγια, η υπερμοντελοποίηση είναι η χρήση μοντέλων ή διαδικασιών που παραβιάζουν την αρχή της οικονομίας (το ξυράφι του Occam), για παράδειγμα περιλαμβάνοντας περισσότερες ρυθμιζόμενες παραμέτρους από τις αναγκαίες ή χρησιμοποιώντας μια πιο περίπλοκη προσέγγιση από την βέλτιστη. Το ξυράφι του Occam υποδηλώνει ότι οποιαδήποτε δεδομένη πολύπλοκη λειτουργία δεν είναι απαραίτητα καλύτερη από οποιαδήποτε δεδομένη απλή λειτουργία. Αυτό το φαινόμενο διακρίνεται καθαρά στο σχήμα 3, στο οποίο παρατηρούμε πως το υπερπροσαρμοσμένο μοντέλο καθορίζεται αυστηρά από τα δεδομένα εκπαίδευσης, χάνοντας έτσι την ικανότητα του για γενίκευση όταν θα συναντήσει νέα δεδομένα.




*Σχήμα 3 Σύγκριση βέλτιστου μοντέλου - υπερπροσαρμοσμένου μοντέλου*

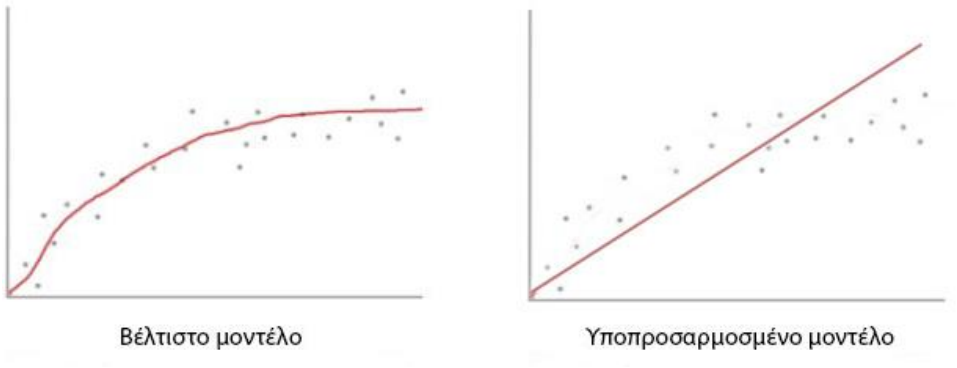
Πιο συγκεκριμένα, σε ένα χώρο υποθέσεων H , μια υπόθεση h ∈ H υπερμοντελοποιεί τα δεδομένα αν υπάρχει μια άλλη υπόθεση h' ∈ H με μεγαλύτερο σφάλμα από την h στα δεδομένα εκπαίδευσης, αλλά μικρότερο σε όλο το σύνολο των περιπτώσεων.

Η υπερμοντελοποίηση είναι ιδιαίτερα πιθανή στις περιπτώσεις όπου το μοντέλο που χρησιμοποιείται είναι αρκετά σύνθετο ή όταν το σύνολο δεδομένων εκπαίδευσης είναι μικρό, προκαλώντας το μοντέλο να προσαρμοστεί σε πολύ συγκεκριμένα τυχαία χαρακτηριστικά των δεδομένων εκπαίδευσης, τα οποία φέρουν ανεπιθύμητα αποτελέσματα και δεν επιτρέπουν στο μοντέλο να γενικεύσει με επιτυχία. Έτσι όταν το μοντέλο που χρησιμοποιείται είναι υπερπροσαρμοσμένο, η ακρίβεια στα παραδείγματα εκπαίδευσης είναι υψηλή, ενώ η απόδοση σε νέα, άγνωστα δεδομένα αρχίζει να μειώνεται.

## 3.2 Υποπροσαρμογή - Ατελής μάθηση

Ένα μοντέλο πρόβλεψης, το οποίο δεν είναι αρκετά σύνθετο, μπορεί να αποτύχει να μοντελοποιήσει επιτυχώς τα δεδομένα εκπαίδευσης και να μη μπορεί να γενικεύσει σε νέα δεδομένα, οδηγώντας σε ατελή μάθηση. Η ατελής μάθηση, ή αλλιώς υποπροσαρμογή (Underfitting), αποτελεί ακριβώς το αντίθετο φαινόμενο από την υπερπροσαρμογή και ένα τέτοιο μοντέλο μηχανικής μάθησης αναγνωρίζεται εύκολα λόγω των χαμηλών τιμών απόδοσης, τόσο στα δεδομένα εκπαίδευσης, όσο και στα δεδομένα ελέγχου.

Συχνά, το φαινόμενο αυτό συμβαίνει όταν έχουμε λίγα δεδομένα για να κατασκευάσουμε ένα ακριβές μοντέλο και επίσης όταν προσπαθούμε να χτίσουμε ένα γραμμικό μοντέλο με μη γραμμικά δεδομένα. Σε τέτοιες περιπτώσεις, οι κανόνες του μοντέλου μηχανικής μάθησης είναι πολύ απλοί, λαμβάνοντας υπόψη ανεπαρκή χαρακτηριστικά και αποτυγχάνουν να αποδώσουν τη συμπεριφορά του συνόλου των δεδομένων εκπαίδευσης και ως εκ τούτου το μοντέλο θα κάνει πιθανώς πολλές λανθασμένες προβλέψεις.




*Σχήμα 4 Σύγκριση βέλτιστο μοντέλο - Υποπροσαρμοσμένο μοντέλο*

Όπως παρατηρείται στο σχήμα 4, ένα υπόπροσαρμοσμένο μοντέλο εξαρτάται ελάχιστα από τα δεδομένα εκπαίδευσης και κάνει μία ισχυρή παραδοχή για τα δεδομένα. Στο συγκεκριμένο διάγραμμα ένα υποπροσαρμοσμένο μοντέλο, όπως ένα πολυώνυμο $1^{ου}$ βαθμού δεν δίνει ιδιαίτερη προσοχή στα σημεία, εφόσον η υπόθεση είναι ότι τα δεδομένα είναι γραμμικά, κάτι που προφανώς δεν συμπίπτει με την πραγματικότητα

.

## 3.3 Επιλογή κατάλληλου μοντέλου

Στην ιδανική περίπτωση, θέλουμε να επιλέξουμε ένα μοντέλο το οποίο ούτε θα είναι αρκετά απλοικό ώστε να είμαστε στην περίπτωση της ατελούς μάθησης, αλλά ούτε και αρκετά σύνθετο και να παρατηρείται υπερμοντελοποίηση. Αυτός είναι ο στόχος, αλλά είναι πολύ δύσκολο να επιτευχθεί στην πράξη.

Για να κατανοήσουμε αυτόν τον στόχο, μπορούμε να παρατηρήσουμε την απόδοση ενός αλγορίθμου μηχανικής μάθησης με την πάροδο του χρόνου, καθώς εκπαιδεύεται. Μπορούμε να σχεδιάσουμε τόσο την απόδοση στα δεδομένα εκπαίδευσης όσο και την απόδοση σε ένα σύνολο δεδομένων που έχουμε αποκλείσει από τη διαδικασία εκπαίδευσης.

Με την πάροδο του χρόνου, όπως ο αλγόριθμος μαθαίνει, το σφάλμα για το μοντέλο στα δεδομένα εκπαίδευσης μειώνεται και το ίδιο συμβαίνει και με το σφάλμα στο σύνολο δεδομένων ελέγχου. Αν εκπαιδεύσουμε για πολύ μεγάλο χρονικό διάστημα, το σφάλμα στο σύνολο δεδομένων εκπαίδευσης μπορεί να συνεχίσει να μειώνεται επειδή αρχίζει να λαμβάνει χώρα το φαινόμενο της υπερπροσαρμογής και το σύστημα εκπαιδεύεται με άσχετες λεπτομέρειες και στοιχεία θορύβου στο σύνολο δεδομένων εκπαίδευσης και προσαρμόζεται σε αυτά. Ταυτόχρονα, το σφάλμα για το σύνολο ελέγχου αρχίζει να αυξάνεται και πάλι καθώς μειώνεται η ικανότητα του μοντέλου να γενικεύει.

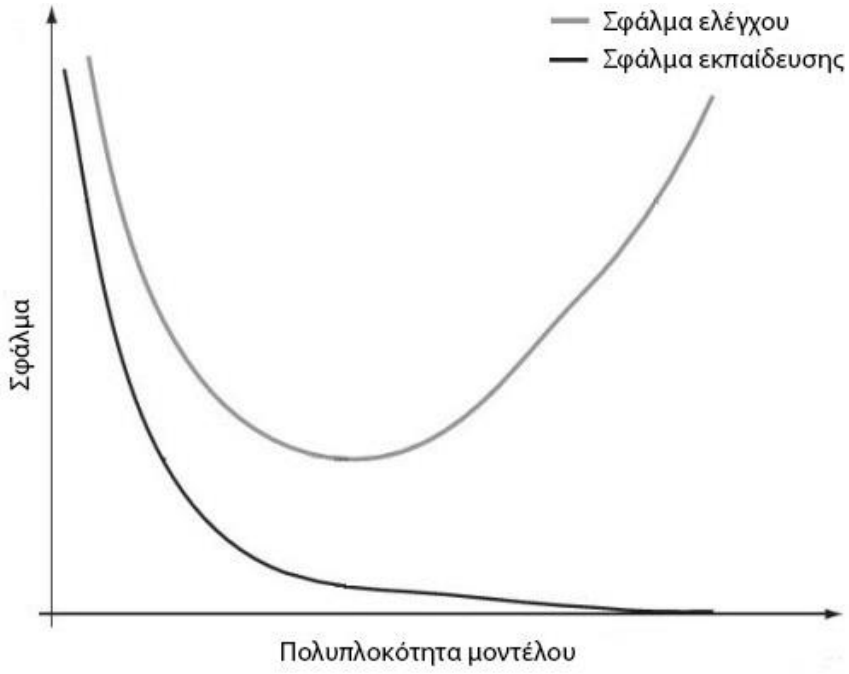
Το σημείο που μας ενδιαφέρει για την επιλογή ενός κατάλληλου μοντέλου για το σύστημα μας, είναι το σημείο ακριβώς πριν αρχίσει να αυξάνεται το σφάλμα στο σύνολο

δεδομένων ελέγχου, όπου το μοντέλο έχει καλές δεξιότητες τόσο στο σύνολο δεδομένων εκπαίδευσης όσο και στο σύνολο νέων αγνώστων δεδομένων.

Πιο αναλυτικά, στο φαινόμενο της υπερπροσαρμογής αναφερόμαστε λέγοντας ότι το μοντέλο έχει υψηλή διακύμανση (high variance). Αντίστοιχα στην περίπτωση της υπόπροσαρμογής λέμε ότι το μοντέλο χαρακτηρίζεται από υψηλή μεροληψία (high bias). Η μεροληψία είναι ένα σφάλμα που αποτελεί αποτέλεσμα εσφαλμένων υποθέσεων στον αλγόριθμο μάθησης. Η υψηλή μεροληψία μπορεί να προκαλέσει έναν αλγόριθμο να χάσει τις σχετικές σχέσεις μεταξύ των χαρακτηριστικών και των στόχων εξόδου (υποπροσαρμογή). Η διακύμανση είναι ένα σφάλμα που προκαλείται από την ευαισθησία στις μικρές διακυμάνσεις στο σύνολο εκπαίδευσης. Η μεγάλη διακύμανση μπορεί να προκαλέσει έναν αλγόριθμο να εντάξει στο μοντέλο τον τυχαίο θόρυβο στα δεδομένα εκπαίδευσης, και να οδηγήσει σε λανθασμένες υποθέσεις (υπερπροσαρμογή). Το δίλημμα μεροληψίας-διακύμανσης αποτελεί τη σύγκρουση στην προσπάθεια ταυτόχρονης ελαχιστοποίησης αυτών των δύο πηγών σφάλματος που εμποδίζουν τους αλγόριθμους επιβλεπόμενης μάθησης να γενικεύουν πέρα από το σύνολο δεδομένων εκπαίδευσης.

### 3.3.1 Πολυπλοκότητα μοντέλου και σφάλμα

Στο σχήμα 5 απεικονίζεται η σχέση της πολυπλοκότητας του μοντέλου με το σφάλμα που παρατηρείται. Είναι φανερό ότι το μοντέλο που ελαχιστοποιεί το σφάλμα εκπαίδευσης δεν ταυτίζεται με το μοντέλο που ελαχιστοποιεί το σφάλμα ελέγχου. Ενώ το σφάλμα εκπαίδευσης μειώνεται όσο αυξάνεται η πολυπλοκότητα του μοντέλου, το σφάλμα ελέγχου εμφανίζει πολύ διαφορετική συμπεριφορά. Για μοντέλα με μικρή πολυπλοκότητα, στα αριστερά του διαγράμματος (περιοχή με υψηλή μεροληψία), το σφάλμα ελέγχου σταδιακά μειώνεται όσο η πολυπλοκότητα αυξάνεται. Στην πορεία, όμως, μετά από ένα συγκεκριμένο σημείο αρχίζει και πάλι να αυξάνεται με την αύξηση της πολυπλοκότητας (περιοχή με υψηλή διακύμανση). Το σημείο αυτό που παρουσιάζει ελάχιστο το σφάλμα ελέγχου αντιστοιχεί στο βέλτιστο μοντέλο για το σύστημα μας και σηματοδοτεί το σημείο έναρξης του φαινομένου της υπερπροσαρμογής των μοντέλων με υψηλότερη πολυπλοκότητα. Η χαμηλή απόδοση των μοντέλων με αυξημένη πολυπλοκότητα οφείλεται στο γεγονός ότι αποδίδουν μία δομή προσαρμοσμένη στα δεδομένα εκπαίδευσης, η οποία προφανώς δεν υφίσταται στα δεδομένα ελέγχου.

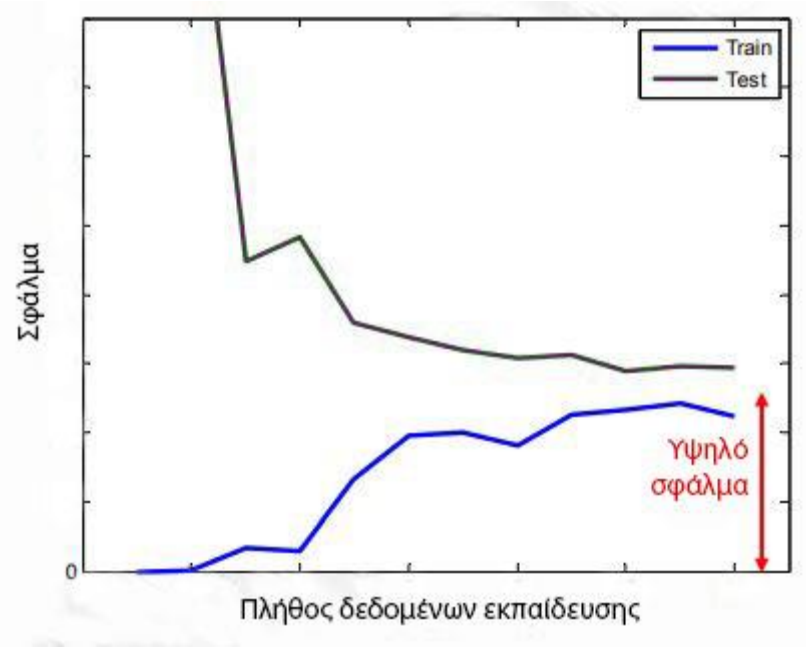



*Σχήμα 5 Καμπύλες σφάλματος - πολυπλοκότητας του μοντέλου*

### 3.3.2 Μέγεθος συνόλου δεδομένων εκπαίδευσης και σφάλμα

Άλλη μία μεταβλητή που μας ενδιαφέρει να εξετάσουμε, εκτός από τη πολυπλοκότητα του μοντέλου, είναι το πλήθος των δεδομένων εκπαίδευσης. Σε αυτή τη περίπτωση για να διαπιστώσουμε αν είμαστε αντιμέτωποι με το πρόβλημα υψηλής μεροληψίας ή με το πρόβλημα υψηλής διακύμανσης αποτυπώνουμε τις καμπύλες σφάλματος εκπαίδευσης και ελέγχου ως συναρτήσεις του πλήθους δεδομένων του συνόλου εκπαίδευσης.

Στο σχήμα 6 παρουσιάζεται μία περίπτωση υποπροσαρμογής ή αλλιώς υψηλής μεροληψίας. Από το διάγραμμα είναι φανερό ότι όσο μικρότερο είναι το πλήθος του συνόλου εκπαίδευσης τόσο πιο ακριβής είναι η προσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης και αντίστοιχα τόσο μικρότερο το σφάλμα εκπαίδευσης. Αντίθετα όσο αυξάνεται το πλήθος των δεδομένων εκπαίδευσης το μοντέλο παρουσιάζει μικρότερη ικανότητα προσαρμογής και το σφάλμα εκπαίδευσης αυξάνεται.

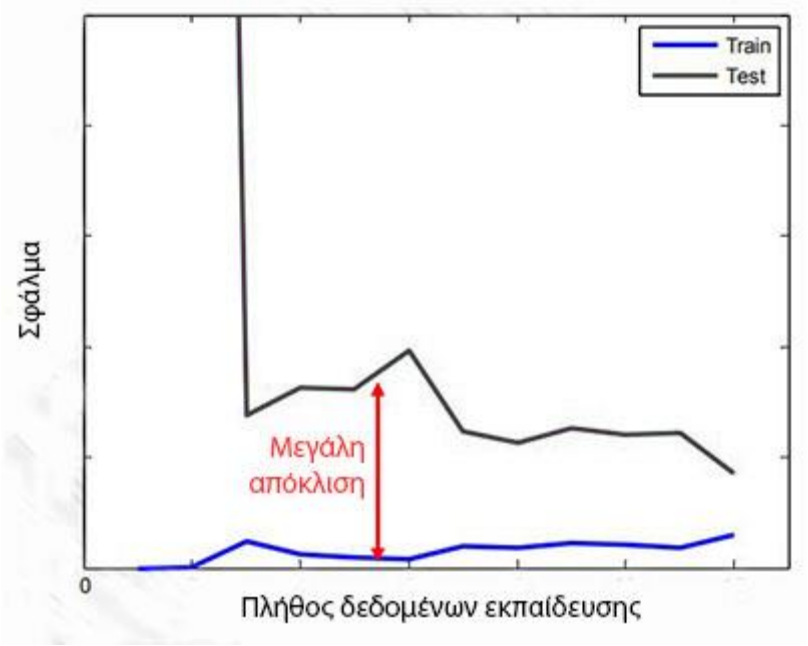



*Σχήμα 6 Καμπύλες σφάλματος - πλήθος δεδομένων εκπαίδευσης : Περίπτωση υψηλής μεροληψίας*

Προφανώς, το αντίθετο φαινόμενο παρατηρείται όσον αφορά το σύνολο ελέγχου. Για μικρότερες τιμές του πλήθους του συνόλου εκπαίδευσης, το μοντέλο παρουσιάζει μειωμένη ικανότητα γενίκευσης και συνεπώς μεγαλύτερο σφάλμα ελέγχου. Όσο αυξάνεται το πλήθος των δεδομένων εκπαίδευσης παρατηρείται μείωση του σφάλματος ελέγχου, εφόσον όσο περισσότερα είναι τα δεδομένα με τα οποία εκπαιδεύτηκε το μοντέλο, τόσο μεγαλύτερη ικανότητα γενίκευσης παρουσιάζει κατά την εφαρμογή του σε νέα δεδομένα.

Όταν, λοιπόν, αντιμετωπίζουμε την περίπτωση υψηλής μεροληψίας, το σφάλμα ελέγχου μειώνεται με τη σταδιακή αύξηση του πλήθους των δεδομένων εκπαίδευσης, αλλά ύστερα από ένα σημείο η βελτίωση που παρουσιάζει το υποπροσαρμοσμένο μοντέλο θα είναι μηδαμινή. Το σφάλμα εκπαίδευσης αντίστοιχα λαμβάνει χαμηλές τιμές για μικρά σύνολα εκπαίδευσης και καταλήγει σε τιμές παραπλήσιες του σφάλματος ελέγχου για μεγαλύτερα σύνολα εκπαίδευσης.

Χαρακτηριστικό στοιχείο του φαινομένου της υποπροσαρμογής ενός μοντέλου είναι η υψηλή τιμή που καταλήγουν να έχουν το σφάλμα ελέγχου και το σφάλμα εκπαίδευσης. Συνεπώς στην περίπτωση που η χαμηλή απόδοση του μοντέλου οφείλεται σε υψηλή μεροληψία, η αύξηση του πλήθους των δεδομένων εκπαίδευσης από μόνη της δεν προσφέρει σημαντική βελτίωση μετά από ένα σημείο και χρειάζεται να ληφθούν επιπλέον παράγοντες υπόψιν, όπως η προσθήκη μεγαλύτερου αριθμού χαρακτηριστικών.




*Σχήμα 7 Καμπύλες σφάλματος - πλήθος δεδομένων εκπαίδευσης : Περίπτωση υψηλής διακύμανσης*

Αντίθετα, στην περίπτωση που ένα μοντέλο χαρακτηρίζεται από υψηλή διακύμανση, το σφάλμα εκπαίδευσης θα ξεκινήσει από πολύ χαμηλές τιμές για μικρά σύνολα εκπαίδευσης και θα συνεχίσει να έχει αρκετά καλή απόδοση για μεγαλύτερα σύνολα εκπαίδευσης με μικρή αύξηση του σφάλματος εκπαίδευσης.

Από την άλλη, το σφάλμα ελέγχου ενός υπερπροσαρμοσμένου μοντέλου, ξεκινάει με πολύ υψηλές τιμές και μειώνεται όσο αυξάνονται τα δεδομένα εκπαίδευσης και στη συνέχεια παρά την μικρή πτωτική τάση που παρουσιάζει, διατηρεί αρκετά υψηλές τιμές ακόμη και για αρκετά μεγάλα σύνολα εκπαίδευσης.

Χαρακτηριστικό ενδεικτικό στοιχείο αυτής της περίπτωσης υπερπροσαρμοσμένου μοντέλου είναι η παρατήρηση μεγάλης απόκλισης ανάμεσα στο σφάλμα εκπαίδευσης και στο σφάλμα ελέγχου. Για μεγαλύτερες τιμές του πλήθους των δεδομένων εκπαίδευσης οι δύο καμπύλες σφαλμάτων τείνουν να συγκλίνουν, συνεπώς χρησιμοποιώντας περισσότερα δεδομένα στην διαδικασία της εκπαίδευσης μπορεί πράγματι να βελτιωθεί η απόδοση του μοντέλου.

## 3.4 Μελέτη της απόδοσης γνωστών μεθόδων

Σε αυτή την εργασία, αναλύσαμε μερικούς αλγόριθμους για τους οποίους θα ήταν χρήσιμο να γνωρίζουμε σε ποιες περιπτώσεις οδηγούμαστε σε φαινόμενα υπερπροσαρμογής ή υποπροσαρμογής με τη χρήση τους, καθώς και την επίδραση του ταξινομητή που έχουμε επιλέξει στην απόδοση του συστήματος.

- Η μέθοδος bag of words

Η πλειοψηφία των μεθόδων μηχανικής μάθησης ταξινόμησης κειμένων χρησιμοποιεί το μοντέλο διανυσματικού χώρου (Vector Space Model - VSM). Μία από αυτές τις μεθόδους είναι η μέθοδος bag of words, η οποία χρησιμοποιεί τις συχνότητες εμφάνισης μεμονωμένων λέξεων ως χαρακτηριστικά των διανυσμάτων. Ωστόσο, αυτή η προσέγγιση χρησιμοποιεί πολλά πλεονάζοντα χαρακτηριστικά και έχει μια αραιή μήτρα μεγάλου μεγέθους, η οποία είναι πιθανό να οδηγήσει σε υπερμοντελοποίηση στην εκπαίδευση και χαμηλή ακρίβεια στα δεδομένα ελέγχου [14]. Συνεπώς, η χρήση αυτής της μεθόδου είναι αποδίδει ικανοποιητικά όταν τα δεδομένα που έχουμε στη διάθεσή μας είναι επαρκή.

- Η μέθοδος n-gram γράφων

Ένας ακόμη αλγόριθμος που μελετάμε σε αυτή την εργασία, είναι ο αλγόριθμος των n-gram γράφων. Καθώς το μέγεθος των n-gram αυξάνεται, η πιθανότητα να συναντήσουμε ίδια n-grams στα δεδομένα μειώνεται. Για το λόγο αυτό, παρά το γεγονός, ότι ένα n-gram, με μεγάλο n, θεωρητικά, περιέχει περισσότερες πληροφορίες σχετικά με το περιεχόμενο μιας λέξης, δεν μπορεί εύκολα να γενικεύσει σε άλλα νέα σύνολα δεδομένων (υπερπροσαρμογή του μοντέλου). Από την άλλη πλευρά, ένα μοντέλο που κατασκευάστηκε για πολύ μικρή τιμή του n, δεν έχει αρκετές πληροφορίες σχετικές με τα συμφραζόμενα και μπορεί να οδηγήσει σε υποπροσαρμογή του συστήματος.

Συνεπώς, εάν το λεξιλόγιο των δεδομένων κειμένων που έχουμε στη διάθεσή μας είναι πολύ πλούσιο, αλλά οι λέξεις αυτές έχουν πολύ χαμηλή συχνότητα εμφάνισης, είναι πιθανό να έχουμε καλύτερα αποτελέσματα για μικρότερες τιμές του n. Ομοίως, εάν το σύνολο των δεδομένων εκπαίδευσης είναι πολύ μικρό, καλό θα ήταν να γίνει χρήση μικρής τιμής του n. Ωστόσο, υποθέτοντας ότι έχουμε αρκετά δεδομένα για να αποφύγετε το φαινόμενο της υπερπροσαρμογής, τότε επιτυγχάνουμε καλύτερες επιδόσεις για υψηλότερες τιμές του n.

- Ο ταξινομητής Naive Bayes

Όσον αφορά τον ταξινομητή, ο Naive Bayes είναι ένας ταξινομητής υψηλής μεροληψίας, με χαμηλή διακύμανση και μπορεί να δημιουργήσει ένα καλό μοντέλο, ακόμη και όταν το σύνολο δεδομένων είναι μικρό. Είναι απλός στη χρήση και έχει μικρό υπολογιστικό κόστος. Η ακρίβεια του Naive Bayes θεωρείται γενικά καλή σε αποφάσεις ταξινόμησης, και είναι σε θέση να αντέξει το θόρυβο που εισάγεται κατά τη διάρκεια της εκπαίδευσης αλλά έχει υπερβολική εμπιστοσύνη στις αποφάσεις του. Συνεπώς, ένας ταξινομητής Naive Bayes έχει ένα σταθερό δομικό σχήμα ανεξάρτητα από τα δεδομένα εκπαίδευσης. Αυτός ο περιορισμός οδηγεί συχνά σε φαινόμενα υποπροσαρμογής, ωστόσο εξακολουθεί να αποδίδει αρκετά καλά.

Πιο αναλυτικά , με τον ταξινομητή Naive Bayes, κάνουμε μία αφελή υπόθεση ανεξαρτησίας των μεταβλητών, που σημαίνει ότι οι αλληλεπιδράσεις μεταξύ μεταβλητών μπορούν να αγνοηθούν. Αυτό έχει ως αποτέλεσμα:

i) να έχει μια απλούστερη και γρηγορότερη λειτουργία υποθέσεων (σε σύγκριση με ταξινομητές, π.χ. λογιστική παλινδρόμηση)

ii) δεδομένου ότι οι αλληλεπιδράσεις δεν λαμβάνονται υπόψη, ορισμένες από τις πληροφορίες στα δεδομένα αγνοούνται. Αυτό τον καθιστά μοντέλο υψηλής μεροληψίας. Παρουσιάζει υψηλό σφάλμα, αλλά είναι αρκετά ανθεκτικό στον θόρυβο των δεδομένων.

iii) τέλος, εφόσον αγνοούνται οι εξαρτήσεις, απαιτούνται λιγότερα δεδομένα εκπαίδευσης.

# Κεφάλαιο 4ο Υλοποίηση και οργάνωση πειραμάτων

Στο κεφάλαιο αυτό, παρουσιάζεται η αρχιτεκτονική υλοποίησης στην οποία στηρίχθηκαν τα πειράματα. Για τους σκοπούς αυτής της εργασίας, χρησιμοποιήθηκαν για την εξαγωγή των πειραματικών αποτελεσμάτων, δύο βάσεις δεδομένων , οι οποίες παρουσιάζονται στη συνέχεια, σε αυτήν την ενότητα και έπειτα γίνεται η ανάλυση των παραμέτρων εκτέλεσης, κάθε πειράματος που υλοποιήθηκε.

## 4.1 Υλοποίηση προγράμματος

Ο αλγόριθμος, που χρησιμοποιήθηκε σε αυτή την εργασία, χωρίζεται σε τρία επιμέρους στάδια επεξεργασίας:

1. Δημιουργία και αποθήκευση των γραφών.
2. Δημιουργία και αποθήκευση αρχείων εκπαίδευσης και ελέγχου (training kai test set).
3. Δημιουργία ταξινομητή και αξιολόγηση του.

### 4.1.1 Πρώτο στάδιο

Σε αυτό το στάδιο δημιουργούνται οι n-gram και οι word-gram γράφοι οι οποίοι αποτελούν τα μοντέλα, ένα για κάθε πολικότητα. Για τους δύο αλγόριθμους εξαγωγής γράφων, γίνεται χρήση των ιδίων κειμένων κριτικής του συνόλου εκπαίδευσης.

Στο στάδιο αυτό έχουμε τη δυνατότητα να επιλέξουμε τα κείμενα κριτικής που θα χρησιμοποιηθούν με βάση τις βαθμολογίες κριτικών που επιθυμούμε, το πλήθος τους, καθώς και να ορίσουμε τις παραμέτρους που χρησιμοποιούμε για κάθε αλγόριθμο όπως το μέγεθος παραθύρου και το μέγεθος του n.

Το στάδιο αυτό δεν αποτελεί μέρος της μηχανικής μάθησης, αλλά θεωρείται στάδιο προεπεξεργασίας και προετοιμασίας των δεδομένων. Εν τέλει, οι γράφοι που θα προκύψουν από το ορισμένο στάδιο, αποθηκεύονται για χρήση τους σε διαφορετικά πειράματα.

Για την κατασκευή των γράφων της μεθόδου n-gram γράφων και word- gram γράφων, έγινε χρήση της βιβλιοθήκης JInsect. Συγκεκριμένα, χρησιμοποιήθηκαν οι μεθόδοι της κλάσης DocumentNGramGraph και η επέκτασή της για την υλοποίηση της δεύτερης μεθόδου. Επιπλέον, η βιβλιοθήκη JInsect κάνει χρήση της βιβλιοθήκης OpenJGraph της Java για τον χειρισμό των γράφων.

### 4.1.2 Δεύτερο στάδιο

Στο δεύτερο στάδιο δημιουργούνται τα αρχεία εκπαίδευσης και ελέγχου. Συγκεκριμένα γίνεται σύγκριση των γράφων πολικότητας που δημιουργήσαμε στο πρώτο στάδιο με τους γράφους των υποψήφιων για εκπαίδευση ή ελέγχου κειμένων κριτικής. Στο τέλος του δεύτερου σταδίου δημιουργούνται τα αρχεία των διανυσμάτων χαρακτηριστικών των κειμένων και αποθηκεύονται στη μορφή Attribute Relation File Format (ARFF). Σε αυτό το στάδιο γίνεται και η δημιουργία των αρχείων εκπαίδευσης και ελέγχου της μεθόδου bag of words.

Το μοντέλο bag of words υλοποιήθηκε με τη βοήθεια της βιβλιοθήκης Weka. Η βιβλιοθήκη αυτή προσφέρει ένα φίλτρο, που επιτρέπει τη μετατροπή ενός κειμένου, που είναι αποθηκευμένο σε ένα string, σε ένα διάνυσμα λέξεων. Η διαδικασία επιλογής των λέξεων του διανύσματος γίνεται από τα δεδομένα εκπαίδευσης και το πλήθος των λέξεων που επιλέγονται καθορίζεται από μία παράμετρο, με προκαθορισμένη αρχική τιμή τις 1000 λέξεις.

### 4.1.3 Τρίτο στάδιο

Στο τρίτο στάδιο δημιουργείται και αξιολογείται ο ταξινομητής που καθορίζεται από τον χρήστη με τις παραμέτρους των αρχείων που δίνονται ως είσοδο σε κάθε εκτέλεση και δημιουργήθηκαν στο δεύτερο στάδιο. Και σε αυτό το στάδιο, λοιπόν, χρησιμοποιείται η βιβλιοθήκη Weka, η οποία ενσωματώνει υλοποιήσεις μεγάλης ποικιλίας ταξινομητών, καθώς και μεθόδους αξιολόγησής τους. Στην δικιά μας περίπτωση, για την εκπόνηση των πειραμάτων, ο ταξινομητής που χρησιμοποιήθηκε είναι ο naïve bayes.

## 4.2 Βάσεις δεδομένων

### 4.2.1 Βάση δεδομένων κριτικών ταινιών του IMDB

Το IMDB αποτελεί μία από τις μεγαλύτερες διαδικτυακές βάσεις δεδομένων με πληροφορίες για ηθοποιούς, ταινίες, τηλεοπτικά προγράμματα, παρουσιαστές της τηλεόρασης, βιντεοπαιχνίδια και συντελεστές παραγωγής ταινιών ή προγραμμάτων. Η βάση δεδομένων κριτικών του IMDB που επιλέξαμε στην συγκεκριμένη εργασία είναι διαθέσιμη στην ιστοσελίδα του πανεπιστημίου του Stanford[1].

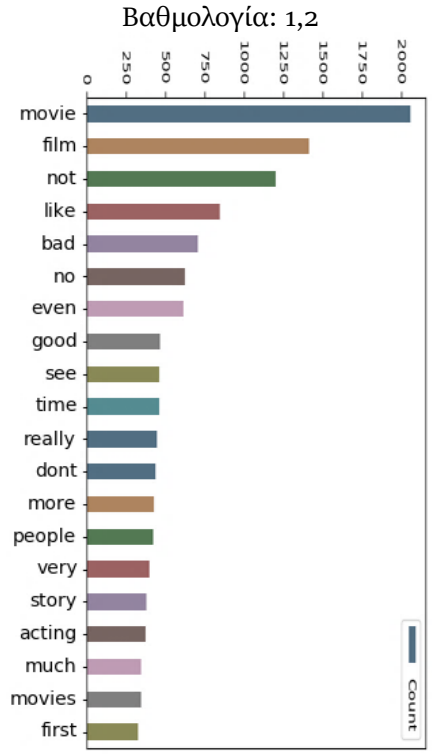
Τα δεδομένα αυτά συγκεντρώθηκαν από το πανεπιστήμιο του Stanford και προσφέρονται ένα σύνολο 25.000 αρχείων κειμένων κριτικών με σκοπό την εκπαίδευση και ένα σύνολο 25.000 αρχείων κειμένων κριτικών για την διαδικασία ελέγχου-αξιολόγησης. Κάθε ένα από αυτά τα σύνολα είναι χωρισμένα σε θετικές και αρνητικές κριτικές.

Η βάση αυτή αποτελείται από αρχεία κειμένου .txt τα οποία περιέχουν το περιεχόμενο κειμένου των κριτικών και το όνομα των αρχείων έχει τη μορφή number_rating.txt όπου number ο αύξοντας αριθμός του αρχείου και rating η βαθμολογία της κριτικής στο διάστημα {1,10}. Οι θετικές βαθμολογίες ορίζονται στο διάστημα {1,4} και αντίστοιχα οι αρνητικές στο διάστημα {6-10}.

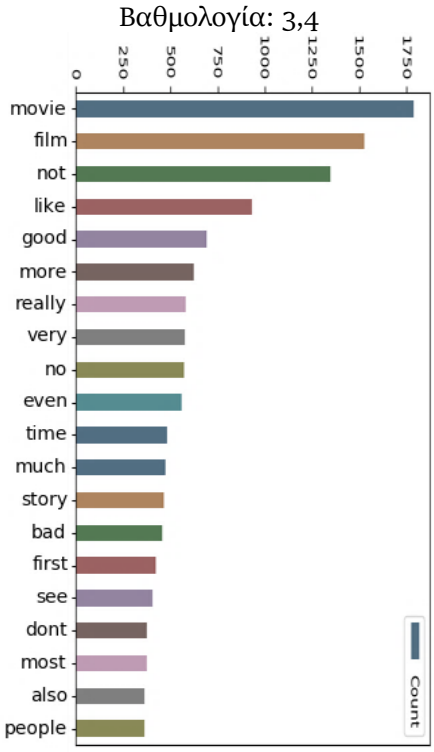
#### 4.2.1.1 Περιεχόμενο Κριτικών

Για την καλύτερη κατανόηση του περιεχομένου των κριτικών της βάσης δεδομένων, με μία σύντομη στατιστική ανάλυση, παρουσιάζονται στη συνέχεια οι λέξεις με τη μεγαλύτερη συχνότητα εμφάνισης στα κείμενα. Η ανάλυση αυτή πραγματοποιήθηκε στα αρχεία κριτικών που χρησιμοποιήθηκαν στη συνέχεια για την εκπαίδευση του αλγορίθμου. Ο κώδικας που χρησιμοποιήθηκε για την ανάλυση βρίσκεται στο παράρτημα στο τέλος της εργασίας.

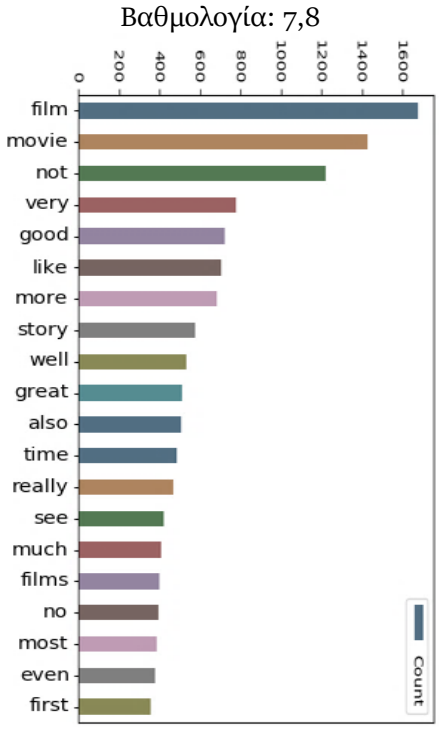
[1] http://ai.stanford.edu/~amaas/data/sentiment



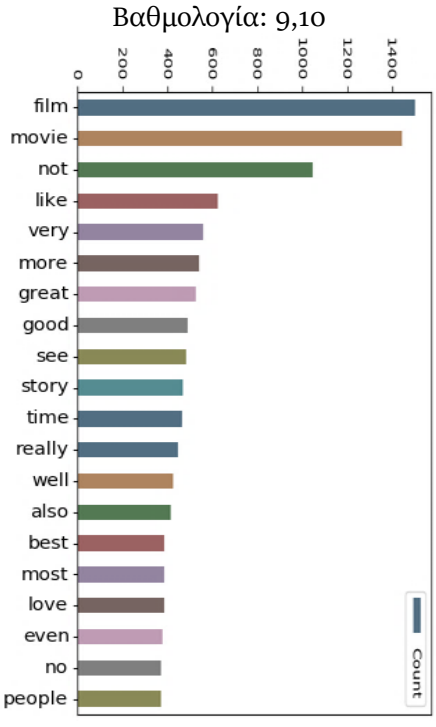

*Σχήμα 8 IMDB - Top 20 words - Rating:1,2*




*Σχήμα 9 IMDB - Top 20 words - Rating:3,4*




*Σχήμα 10 IMDB - Top 20 words - Rating:7,8*

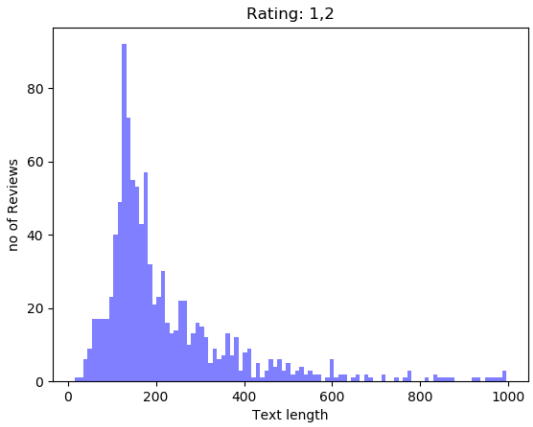



*Σχήμα 11 IMDB - Top 20 words - Rating:9,10*

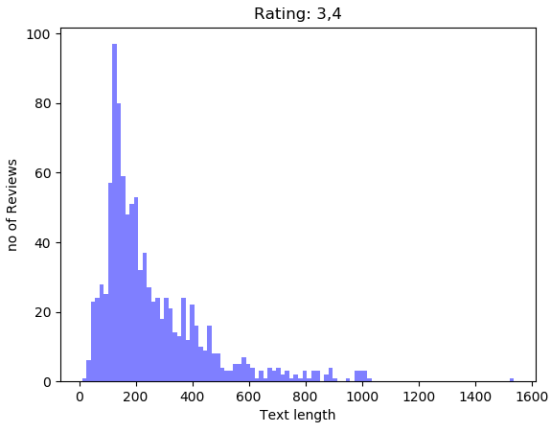
#### 4.2.1.2 Δομή κειμένων κριτικών

Για την κατανόηση της δομής των κειμένων της βάσης, μας ενδιαφέρει η κατανομή των μηκών κειμένων, ανά ζεύγη βαθμολογιών (εφόσον εργαζόμαστε ανά ζεύγη και στη συνέχεια). Ο κώδικας που χρησιμοποιήθηκε για την κατασκευή των ιστογραμμάτων βρίσκεται στο παράρτημα στο τέλος της εργασίας.

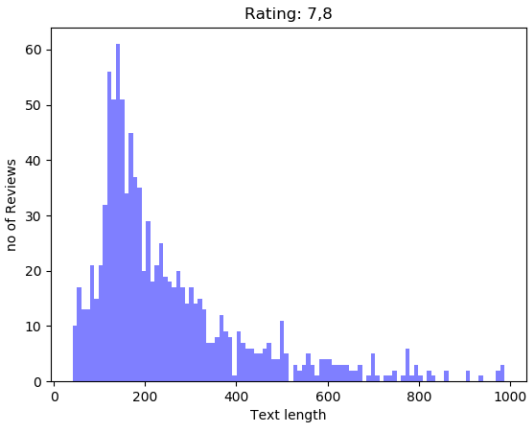
Παρατηρείται, στα σχήματα 3.5-8 ότι οι κατανομές έχουν αρκετές ομοιότητες και στις τέσσερις περιπτώσεις και μας ενδιαφέρει ιδιαίτερα ότι σε όλες τις βαθμολογίες παρουσιάζεται μέγιστο στο διάστημα 100-200 λέξεις. Συνεπώς, το συγκεκριμένο σύνολο αρχείων εκπαίδευσης είναι κατάλληλο για χρήση στα πειράματα της εργασίας.



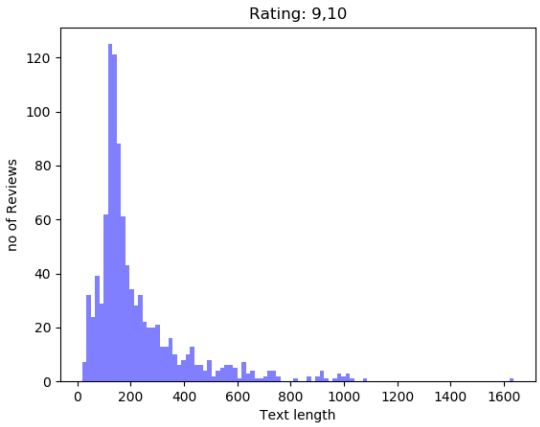

*Σχήμα 12 IMDB - Text Length Distribution - Rating: 1,2*




*Σχήμα 13 IMDB - Text Length Distribution - Rating: 3,4*




*Σχήμα 14 IMDB - Text Length Distribution - Rating: 7,8*




*Σχήμα 15 IMDB - Text Length Distribution - Rating: 9,10*

### 4.2.2 Βάση δεδομένων κριτικών επιχειρήσεων του YELP

Το YELP είναι πολυεθνικός οργανισμός που δημοσιεύει βαθμολογίες και κριτικές χρηστών για τοπικές επιχειρήσεις. Οι κύριες κατηγορίες επιχειρήσεων που είναι εγγεγραμμένες στη βάση του YELP είναι εστιατόρια, πολυκαταστήματα, κέντρα νυχτερινής διασκέδασης, τοπικές υπηρεσίες, καφετέριες και αυτοκινητοβιομηχανίες. Τα δεδομένα αυτά είναι διαθέσιμα στον ιστότοπο του YELP[1] , για χρήση από φοιτητές για εκπαιδευτικούς σκοπούς.

Η βάση δεδομένων αποτελείται από 5.200.000 κριτικές επιχειρήσεων και είναι οργανωμένη με διαφορετική δομή, οπότε χρειάστηκε να την μορφοποιήσουμε αντίστοιχα με αυτή του IMDB, ώστε να μην γίνουν τροποποιήσεις στον κώδικα του κυρίως προγράμματος. Οι κριτικές στην βάση στην αρχική της μορφή βρίσκονται όλες σε ένα αρχείο .json, μεγάλου όγκου, αποθηκευμένες σειριακά. Για να εργαστούμε σε αυτή τη βάση δεδομένων, την μετατρέψαμε στη μορφή που περιγράψαμε προηγουμένως. Ο κώδικας για την μετατροπή βρίσκεται στο παράρτημα στο τέλος της εργασίας.

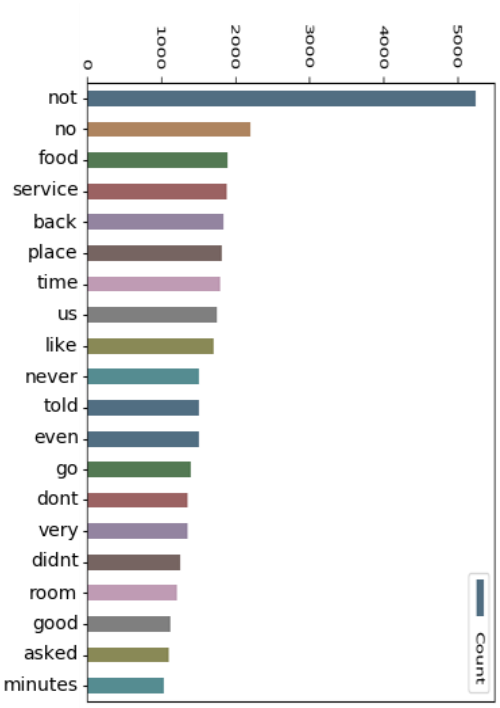
Στη συγκεκριμένη βάση δεδομένων, οι κριτικές είναι σύντομα κείμενα που αποτελούνται από λίγες γραμμές με περίπου εκατό λέξεις. Συνήθως, μια κριτική κάνει μία πολύπλευρη περιγραφή σχετικά με τις υπηρεσίες μιας επιχείρησης και της εμπειρίας του χρήστη. Επιπλέον, όσον αφορά το σύστημα βαθμολόγησης. οι θετικές βαθμολογίες ορίζονται στο διάστημα {4,5}, ενώ οι αρνητικές στο διάστημα {1,2}. Με βάση αυτό το σύστημα βαθμολόγησης, επιλέχθηκαν ορισμένα αρχεία κριτικών για εκπαίδευση του αλγορίθμου, τα οποία τα διακρίναμε σε θετικές και αρνητικές κριτικές και αντίστοιχα ορισμένα αρχεία για έλεγχο.

#### 4.2.2.1 Περιεχόμενο κριτικών

Ομοίως με την προηγούμενη περίπτωση, αναλύσαμε τη βάση αυτή ως προς την συχνότητα εμφάνισης λέξεων στα κείμενα και παρουσιάζουμε στη συνέχεια τις είκοσι πιο συχνές λέξεις, ανά βαθμολογία κειμένων. Η ανάλυση αυτή μας βοηθάει να έχουμε μία οπτική εικόνα του περιεχομένου των κριτικών κάθε βαθμολογίας.

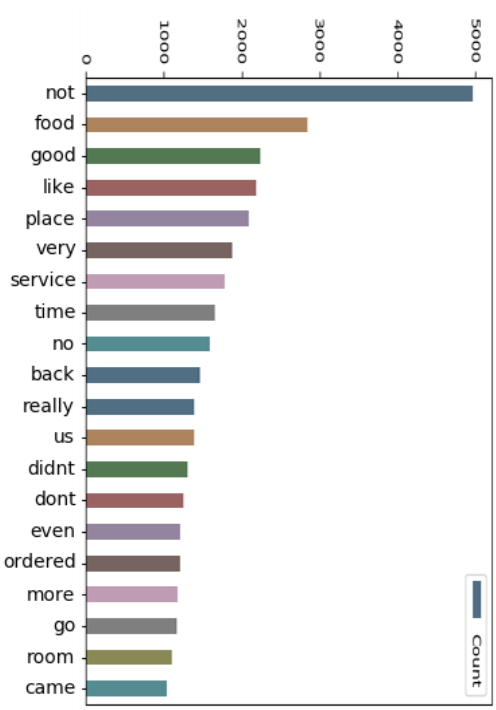
[1] https://www.yelp.com/dataset/challenge

Βαθμολογία: 1



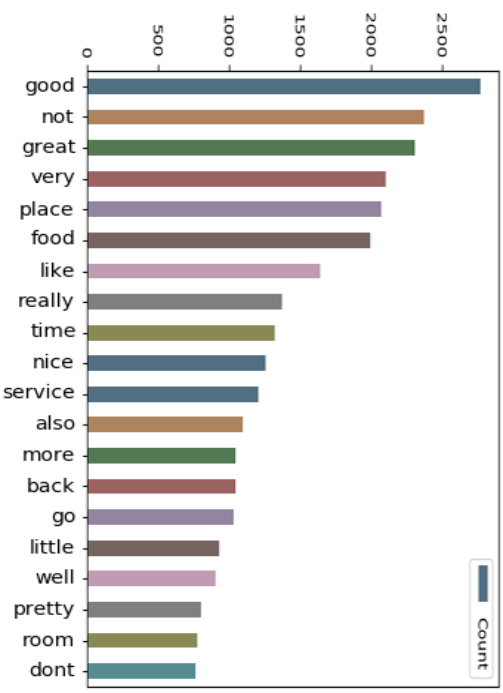

*Σχήμα 16 YELP - Top 20 words - Rating:1*

Βαθμολογία: 2



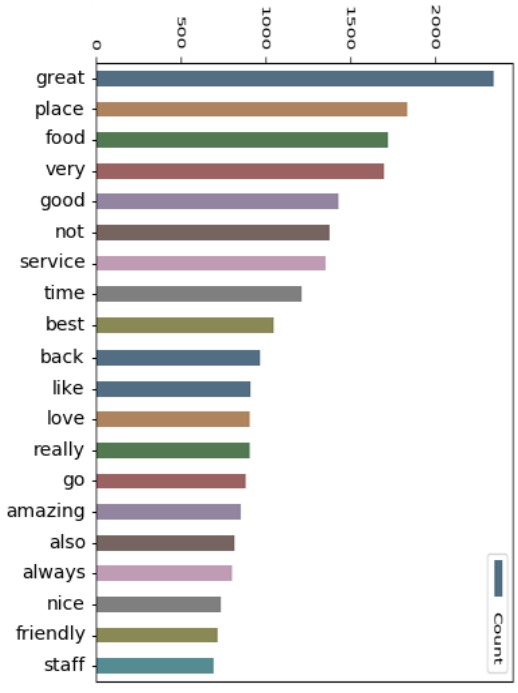

*Σχήμα 17 YELP - Top 20 words - Rating:2*

Βαθμολογία: 4




*Σχήμα 18 YELP - Top 20 words - Rating:4*

Βαθμολογία: 5

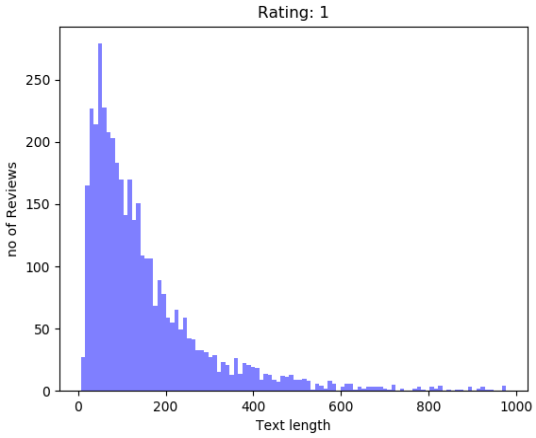



*Σχήμα 19 YELP - Top 20 words - Rating:5*

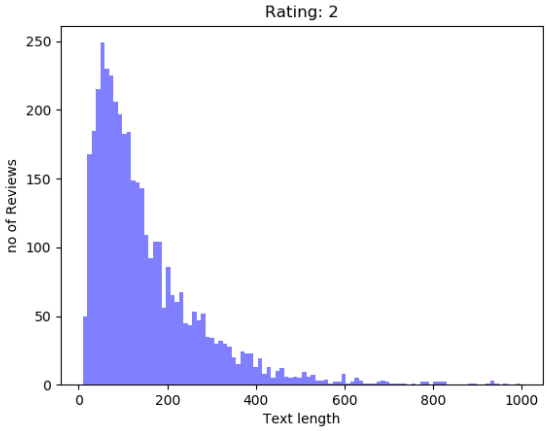
#### 4.2.2.2 Δομή κειμένων κριτικών

Επιπλέον πληροφορίες για αυτή τη βάση δεδομένων δίνουν τα σχήματα 15-18, σχετικά με την κατανομή του μήκους του κειμένου, για κάθε μία βαθμολογία ξεχωριστά.

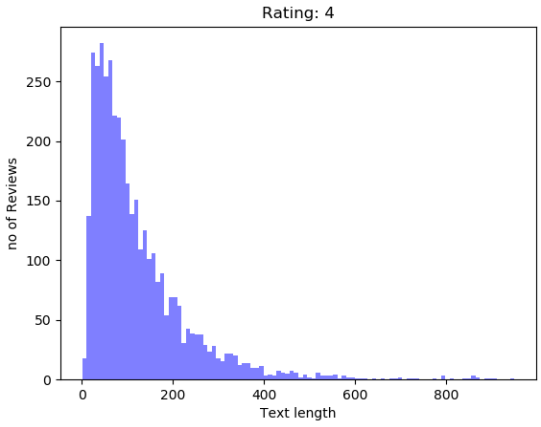
Παρατηρούμε ότι η κατανομή του μήκους κειμένου είναι πανομοιότυπη και για τις τέσσερις βαθμολογίες που εξετάζουμε και η πλειοψηφία των κειμένων κυμαίνεται στις 50 – 100 λέξεις, συνεπώς το συγκεκριμένο σύνολο αρχείων κριτικών κρίνεται κατάλληλο για την πειραματική μελέτη.



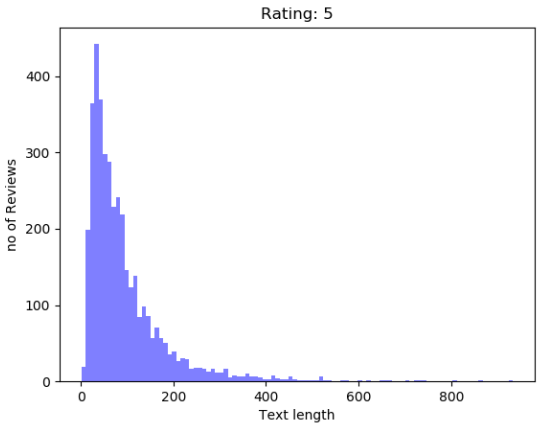

*Σχήμα 20 YELP - Text Length Distribution - Rating: 1*




*Σχήμα 21 YELP - Text Length Distribution - Rating: 2*




*Σχήμα 22 YELP - Text Length Distribution - Rating: 4*




*Σχήμα 23 YELP - Text Length Distribution - Rating: 5*

## 4.3 Παράμετροι εκτέλεσης

Οι αλγόριθμοι που χρησιμοποιήθηκαν για το πείραμα είναι ο word-graphs , ο n-gram graphs, με το μέγεθος n ίσο με 4 και ίδιο μήκος παραθύρου, όπως συνηθίζεται στη σχετική βιβλιογραφία [12] και ο bag-of-words, με χρήση συχνότητας εμφάνισης λέξεων και με πλήθος λέξεων που επιλέχθηκαν για την δημιουργία του λεξιλογίου ορισμένο στις 100.000 λέξεις. Περιγραφή του κάθε αλγορίθμου γίνεται αναλυτικά στο δεύτερο κεφάλαιο.

### 4.3.1 Πείραμα 1$^{ο}$ : Κριτικές ταινιών του IMDB

→ Παράμετροι εκτέλεσης σχετικές με τα αρχεία εκπαίδευσης:

Διακρίνουμε τρεις γενικές περιπτώσεις με βάση τη βαθμολογία του συνόλου εκπαίδευσης για την δημιουργία των μοντέλων γράφων πολικότητας:

- Negative Rating: {1,4}, Positive Rating: {6,10} (Όλες οι βαθμολογίες)
- Negative Rating: {1,2}, Positive Rating: {9,10} (Ακραίες βαθμολογίες)
- Negative Rating: {3,4}, Positive Rating: {7,8} (Ενδιάμεσες βαθμολογίες)

Άλλη μία παράμετρο που εξετάζουμε στην εργασία αυτή είναι το μέγεθος του συνόλου εκπαίδευσης. Διακρίνουμε τις εξής περιπτώσεις:

- Training set : 1000 reviews
- Training set : 2000 reviews

Για κάθε μία από τις προηγούμενες περιπτώσεις, συγκεκριμένα για τον αλγόριθμο word-graphs, διακρίνουμε δύο επιμέρους περιπτώσεις με βάση το μέγεθος παραθύρου:

- Window = 1
- Window = 4

Συνολικά, εξετάζουμε δώδεκα συνδυασμούς παραμέτρων εισόδου στην εκτέλεση του πρώτου σταδίου του προγράμματος για τον αλγόριθμο word-graphs, συνεπώς δημιουργούνται δώδεκα μοντέλα γράφοι κατά το στάδιο εκπαίδευσης του αλγορίθμου και έξι συνδυασμοί παραμέτρων εισόδου για τους αλγόριθμους n-gram και bag of words.

→ Παράμετροι εκτέλεσης των αρχείων ελέγχου:

Εξετάζουμε τρείς γενικές περιπτώσεις, με βάση το μέγεθος του συνόλου των αρχείων ελέγχου και της βαθμολογίας ως εξής :

- Testing set : 2000 reviews, Positive Rating: {5}, Negative Rating: {1}
- Testing set : 2000 reviews, Positive Rating: {4}, Negative Rating: {2}
- Testing set : 10000 reviews, Positive Rating: {4,5}, Negative Rating: {1,2}

### 4.3.2 Πείραμα 2ο : Κριτικές επιχειρήσεων του YELP:

→ Παράμετροι εκτέλεσης με βάση αρχεία εκπαίδευσης :

Ομοίως εργαζόμαστε με τη βάση δεδομένων του YELP , διακρίνοντας τρεις γενικές περιπτώσεις, με βάση τη βαθμολογία του συνόλου εκπαίδευσης:

- Negative Rating: {1,2}, Positive Rating: {4,5} (Όλες οι βαθμολογίες)
- Negative Rating: {1}, Positive Rating: {5} (Ακραίες βαθμολογίες)
- Negative Rating: {2}, Positive Rating: {4} (Ενδιάμεσες βαθμολογίες)

Επιπλέον, διακρίνουμε τις εξής περιπτώσεις με βάση το μέγεθος του συνόλου εκπαίδευσης.

- Training set : 2000 reviews
- Training set : 8000 reviews

Με την ίδια λογική, για κάθε μία από τις γενικές περιπτώσεις διακρίνουμε δύο επιμέρους περιπτώσεις με βάση το μέγεθος παραθύρου του αλγορίθμου word-graphs :

- Window = 1
- Window = 4

Συνολικά, όπως ακριβώς και προηγουμένως, εξετάζουμε δώδεκα συνδυασμούς παραμέτρων εισόδου στην εκτέλεση του πρώτου σταδίου του προγράμματος του αλγορίθμου word-graphs και έξι συνδυασμούς παραμέτρων στο στάδιο εκπαίδευσης των αλγορίθμων n-gram και bag of words.

→ Παράμετροι εκτέλεσης των αρχείων ελέγχου:

Το σύνολο των αρχείων ελέγχου διακρίνεται με ακριβώς τον ίδιο τρόπο, όπως στην βάση δεδομένων του IMDB, με τις τρείς γενικές περιπτώσεις (με βάση το μέγεθος του συνόλου των αρχείων ελέγχου και της βαθμολογίας) να είναι :

- Testing set : 2000 reviews, positive ranking: {5}, negative ranking: {1}
- Testing set : 2000 reviews, positive ranking: {4}, negative ranking: {2}
- Testing set : 10000 reviews, positive ranking: {4,5}, negative ranking: {1,2}

Συνεπώς, συνολικά εξετάζουμε 36 δυνατούς συνδυασμούς παραμέτρων στην περίπτωση του αλγορίθμου word-graphs και 18 συνδυασμούς παραμέτρων στην περίπτωση των αλγορίθμων n-gram και bag of words, για κάθε μία από τις βάσεις δεδομένων, οι οποίοι καλύπτουν ένα ευρύ φάσμα για τη διεξοδική μελέτη της συμπεριφοράς των αλγορίθμων.

# Κεφάλαιο 5ο Αποτελέσματα Πειραμάτων και Παρατηρήσεις

Στην ενότητα αυτή, γίνεται παρουσίαση και σύγκριση των αποτελεσμάτων των πειραμάτων που πραγματοποιήθηκαν.

## 5.1 Παρουσίαση αποτελεσμάτων και σύγκριση μεθόδων

### 5.1.1 Πείραμα 1ο : Κριτικές ταινιών IMDB - Αποτελέσματα

→ Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1,2 - 9,10 (Ακραίες Βαθμολογίες):

- Αποτελέσματα **WordGraphs**
  - Για window size = 1 :

| Model graphs: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 75.15% | 70.40% | 75.25% | 79% | 70.30% | 76.05% |
| Incorrectly Classified Instances | 24.85% | 29.60% | 24.75% | 21% | 29.70% | 23.95% |
| Kappa statistic | 0.503 | 0.408 | 0.505 | 0.58 | 0.406 | 0.521 |
| Mean absolute error | 0.2749 | 0.3455 | 0.2935 | 0.2448 | 0.3565 | 0.2757 |
| Root mean squared error | 0.426 | 0.4533 | 0.4149 | 0.399 | 0.4589 | 0.4094 |
| Relative absolute error | 54.99% | 69.10% | 58.71% | 48.96% | 71.31% | 55.14% |
| Root relative squared error | 85.20% | 90.65% | 82.98% | 79.81% | 91.78% | 81.89% |
| Coverage of cases (0.95 level) | 94.85% | 96.35% | 98% | 94.60% | 95.70% | 96.10% |
| Mean rel. region size (0.95 level) | 77.10% | 87.75% | 84.80% | 74.55% | 87.90% | 80.08% |

*Πίνακας 1 IMDB-WordGraphs Results (window size = 1) - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1,2 - 9,10*

Για window size = 4 :

| Model graphs: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 83.20% | 74.50% | 82.40% | 85.15% | 72.65% | 80.95% |
| Incorrectly Classified Instances | 16.80% | 25.50% | 17.60% | 14.85% | 27.35% | 19.05% |
| Kappa statistic | 0.664 | 0.49 | 0.648 | 0.703 | 0.453 | 0.619 |
| Mean absolute error | 0.2057 | 0.3078 | 0.2299 | 0.1811 | 0.3208 | 0.2267 |
| Root mean squared error | 0.359 | 0.4142 | 0.3693 | 0.3335 | 0.4259 | 0.3713 |
| Relative absolute error | 41.13% | 61.57% | 45.98% | 36.22% | 64.16% | 45.34% |
| Root relative squared error | 71.80% | 82.85% | 73.85% | 66.70% | 85.19% | 74.27% |
| Coverage of cases (0.95 level) | 95.60% | 98.10% | 97.10% | 96.75% | 97.65% | 96.45% |
| Mean rel. region size (0.95 level) | 72.30% | 87.85% | 78.18% | 70% | 86.55% | 75.90% |

*Πίνακας 2 IMDB-WordGraphs Results (window size = 4) - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1,2 - 9,10*

- Αποτελέσματα **NGramGraphs**

| Model graphs: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results NGramGraphs** | | | | | | |
| Correctly Classified Instances | 81.80% | 73.85% | 79.75% | 81.50% | 69.75% | 80.50% |
| Incorrectly Classified Instances | 18.20% | 26.15% | 20.25% | 18.50% | 30.25% | 19.50% |
| Kappa statistic | 0.636 | 0.477 | 0.595 | 0.63 | 0.395 | 0.61 |
| Mean absolute error | 0.2003 | 0.303 | 0.2409 | 0.2071 | 0.3408 | 0.2434 |
| Root mean squared error | 0.3775 | 0.4299 | 0.3872 | 0.3756 | 0.4406 | 0.3788 |
| Relative absolute error | 40.06% | 60.61% | 48.17% | 41.43% | 68.17% | 48.68% |
| Root relative squared error | 75.50% | 85.98% | 77.45% | 75.12% | 88.11% | 75.76% |
| Coverage of cases (0.95 level) | 93.60% | 96.30% | 95.55% | 94.95% | 97.05% | 96.45% |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mean rel. region size (0.95 level) | 67.05% | 84.55% | 76.55% | 69.73% | 89.03% | 79.03% |

*Πίνακας 3 IMDB-NGramGraphs Results - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1,2 - 9,10*

- Αποτελέσματα **Bag of Words**

| Training Set: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results Bag of Words** | | | | | | |
| Correctly Classified Instances | 77.10% | 72.75% | 75.75% | 78.30% | 75.75% | 78.30% |
| Incorrectly Classified Instances | 22.90% | 27.25% | 24.25% | 21.70% | 24.25% | 21.70% |
| Kappa statistic | 54.20% | 45.50% | 51.50% | 56.60% | 51.50% | 56.60% |
| Mean absolute error | 22.85% | 27.45% | 24.69% | 22.00% | 24.92% | 21.96% |
| Root mean squared error | 46.29% | 50.55% | 48.19% | 45.47% | 47.64% | 45.19% |
| Relative absolute error | 45.70% | 54.90% | 49.39% | 43.99% | 49.83% | 43.91% |
| Root relative squared error | 92.58% | 101.10% | 96.37% | 90.93% | 95.27% | 90.38% |
| Coverage of cases (0.95 level) | 81.30% | 77.95% | 79.20% | 81.40% | 80.65% | 81.85% |
| Mean rel. region size (0.95 level) | 54.25% | 55.75% | 54.40% | 53.95% | 56.73% | 54.30% |

*Πίνακας 4 IMDB- Bag of Word Results - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1,2 - 9,10*

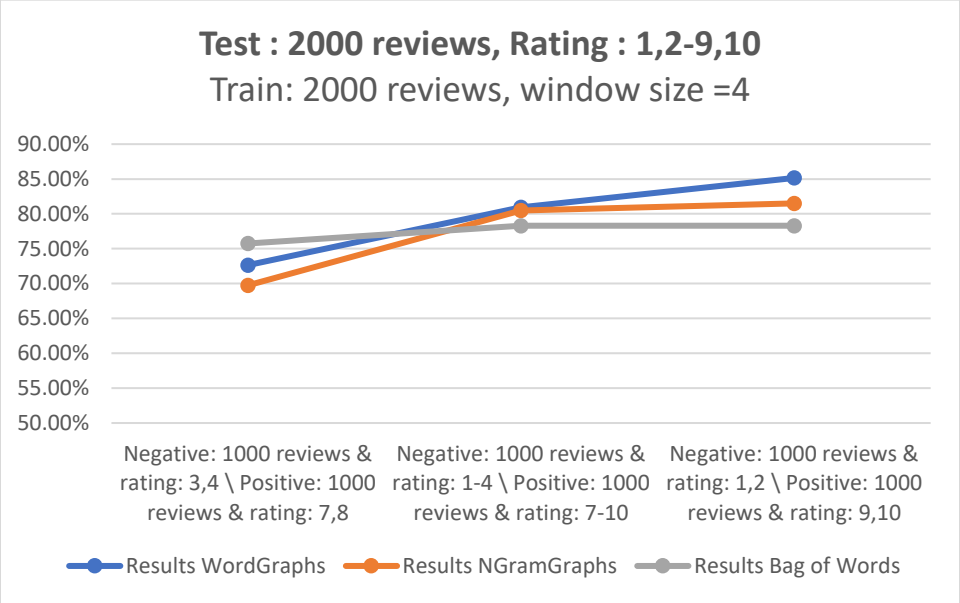
Παρατηρούμε λοιπόν, ότι στην περίπτωση που το σύνολο ελέγχου, που εξετάζουμε, είναι **2000 reviews και Rating : 1,2-9,10** , τότε για σταθερό πλήθος κριτικών στο σύνολο εκπαίδευσης, ίσο με 2000 (1000 κριτικές για κάθε πολικότητα) και window size ίσο με 4 στη περίπτωση του word graphs, πετυχαίνουμε τη μέγιστη ακρίβεια για κάθε αλγόριθμο. Στο σχήμα 19 δίνεται η γραφική αναπαράσταση των αποτελεσμάτων, για κάθε αλγόριθμο ξεχωριστά. Στον οριζόντιο άξονα έχουμε στα αριστερά τα μοντέλα γράφους που δημιουργήσαμε με τις ενδιάμεσες - αμφιλεγόμενες βαθμολογίες, στην μέση τους γράφους με όλες τις βαθμολογίες και στα δεξιά τους γράφους με την μεγαλύτερη πόλωση (μόνο ακραίες βαθμολογίες).

Σύμφωνα με το σχήμα 19, παρατηρούμε ότι η μεταβολή της πόλωσης στο σύνολο εκπαίδευσης δείχνει να επηρεάζει την απόδοση των αλγορίθμων n-gram graphs και word graphs, και συγκεκριμένα για μεγαλύτερη πόλωση έχουμε βελτίωση του ποσοστού ακρίβειας, ενώ η μέθοδος bag of words είναι ανθεκτική στην μεταβολή της πόλωσης του συνόλου εκπαίδευσής της, εφόσον οι αυξομειώσεις της ακρίβειας στις τρεις περιπτώσεις για αυτή τη μέθοδο είναι αμελητέες.

Η μέθοδος bag of words όταν το σύνολο εκπαίδευσης αποτελείται μόνο από ενδιάμεσες βαθμολογίες επιτυγχάνει καλύτερη απόδοση από τους άλλους δύο αλγορίθμους, ενώ όταν η πόλωση γίνεται πιο έντονη, εφόσον αυτή δεν μεταβάλλεται, υψηλότερα ποσοστά ακρίβειας επιτυγχάνουν οι αλγόριθμοι word graphs και n-gram graphs.

Συγκεκριμένα, στη περίπτωση του αλγορίθμου word graphs, έχουμε σημαντική αύξηση της απόδοσης όσο το training set γίνεται περισσότερο πολωμένο, πετυχαίνοντας τη μέγιστη απόδοση με τιμή 85.15% για μεγάλη πόλωση.

Ο αλγόριθμος n-gram graphs παρουσιάζει επίσης βελτίωση της απόδοσής του, όσο η πόλωση του συνόλου εκπαίδευσης γίνεται πιο έντονη και επιτυγχάνει μέγιστη απόδοση για μεγάλη πόλωση με τιμή 81.5%.




*Σχήμα 24 IMDB - Σύγκριση αλγορίθμων - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1,2 - 9,10*

→ Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 3,4 - 7,8 (Ενδιάμεσες βαθμολογίες):

- Αποτελέσματα **WordGraphs**

Για window size = 1:

| Model graphs: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 68% | 64.50% | 65.95% | 70.30% | 66.70% | 67.15% |
| Incorrectly Classified Instances | 32% | 35.50% | 34.05% | 29.70% | 33.30% | 32.85% |
| Kappa statistic | 0.36 | 0.29 | 0.319 | 0.406 | 0.334 | 0.343 |
| Mean absolute error | 0.3549 | 0.3908 | 0.3745 | 0.3303 | 0.3787 | 0.3597 |
| Root mean squared error | 0.4948 | 0.4883 | 0.4867 | 0.472 | 0.4697 | 0.4854 |
| Relative absolute error | 70.98% | 78.16% | 74.90% | 66.07% | 75.73% | 71.94% |
| Root relative squared error | 98.95% | 97.67% | 97.34% | 94.40% | 93.94% | 97.09% |
| Coverage of cases (0.95 level) | 90.85% | 96.05% | 95.50% | 92.85% | 96.75% | 93.95% |
| Mean rel. region size (0.95 level) | 81.30% | 90.68% | 88.58% | 81.20% | 90.10% | 84.55% |

*Πίνακας 5 IMDB-WordGraphs Results (window size = 1) - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 3,4 - 7,8*

Για window size = 4 :

| Model graphs: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 71.30% | 69.65% | 71.25% | 73.65% | 68.40% | 71.95% |

| | | | | | | |
|---|---|---|---|---|---|---|
| Incorrectly Classified Instances | 28.70% | 30.35% | 28.75% | 26.35% | 31.60% | 28.05% |
| Kappa statistic | 0.426 | 0.393 | 0.425 | 0.473 | 0.368 | 0.439 |
| Mean absolute error | 0.3093 | 0.3596 | 0.3241 | 0.2888 | 0.3642 | 0.3154 |
| Root mean squared error | 0.4595 | 0.4551 | 0.4565 | 0.441 | 0.4598 | 0.452 |
| Relative absolute error | 61.86% | 71.93% | 64.81% | 57.75% | 72.84% | 63.08% |
| Root relative squared error | 91.90% | 91.03% | 91.30% | 88.19% | 91.96% | 90.41% |
| Coverage of cases (0.95 level) | 92.75% | 97.60% | 94.70% | 93.35% | 96.90% | 94.95% |
| Mean rel. region size (0.95 level) | 78.53% | 91.45% | 83.13% | 76.73% | 88.98% | 81.65% |

*Πίνακας 6 IMDB-WordGraphs Results (window size = 4) - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 3,4 - 7,8*

- Αποτελέσματα **NGramGraphs**

| Model graphs: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results NGramGraphs** | | | | | | |
| Correctly Classified Instances | 70.95% | 68.50% | 71.65% | 70.70% | 64.85% | 70.80% |
| Incorrectly Classified Instances | 29.05% | 31.50% | 28.35% | 29.30% | 35.15% | 29.20% |
| Kappa statistic | 0.419 | 0.37 | 0.433 | 0.414 | 0.297 | 0.416 |
| Mean absolute error | 0.3064 | 0.3513 | 0.316 | 0.3134 | 0.3881 | 0.3306 |
| Root mean squared error | 0.4822 | 0.4664 | 0.46 | 0.4788 | 0.4802 | 0.4621 |
| Relative absolute error | 61.27% | 70.27% | 63.20% | 62.69% | 77.62% | 66.11% |
| Root relative squared error | 96.43% | 93.27% | 92.00% | 95.76% | 96.04% | 92.41% |
| Coverage of cases (0.95 level) | 88.65% | 96.10% | 93.70% | 90.00% | 96.45% | 94.45% |
| Mean rel. region size (0.95 level) | 71.93% | 88.68% | 80.65% | 74.55% | 91.40% | 83.58% |

*Πίνακας 7 IMDB-NGramGraphs Results - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 3,4 - 7,8*

- Αποτελέσματα **Bag of Words**

| Training Set: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & |
|---|---|---|---|---|---|---|

| | & rating: 9,10 | & rating: 7,8 | & rating: 7,8,9,10 | & rating: 9,10 | & rating: 7,8 | rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results Bag of Words** | | | | | | |
| Correctly Classified Instances | 65.15% | 66.10% | 65.90% | 67.65% | 67.95% | 67.05% |
| Incorrectly Classified Instances | 34.85% | 33.90% | 34.10% | 32.35% | 32.05% | 32.95% |
| Kappa statistic | 30.30% | 32.20% | 31.80% | 35.30% | 35.90% | 34.10% |
| Mean absolute error | 35.01% | 34.07% | 34.28% | 32.74% | 32.46% | 33.15% |
| Root mean squared error | 57.80% | 56.60% | 57.21% | 55.77% | 54.86% | 56.00% |
| Relative absolute error | 70.02% | 68.15% | 68.56% | 65.49% | 64.93% | 66.31% |
| Root relative squared error | 115.60% | 113.20% | 114.42% | 111.54% | 109.72% | 112.00% |
| Coverage of cases (0.95 level) | 69.60% | 72.00% | 70.25% | 71.80% | 73.75% | 71.95% |
| Mean rel. region size (0.95 level) | 54.25% | 55.75% | 54.40% | 53.95% | 56.73% | 54.30% |

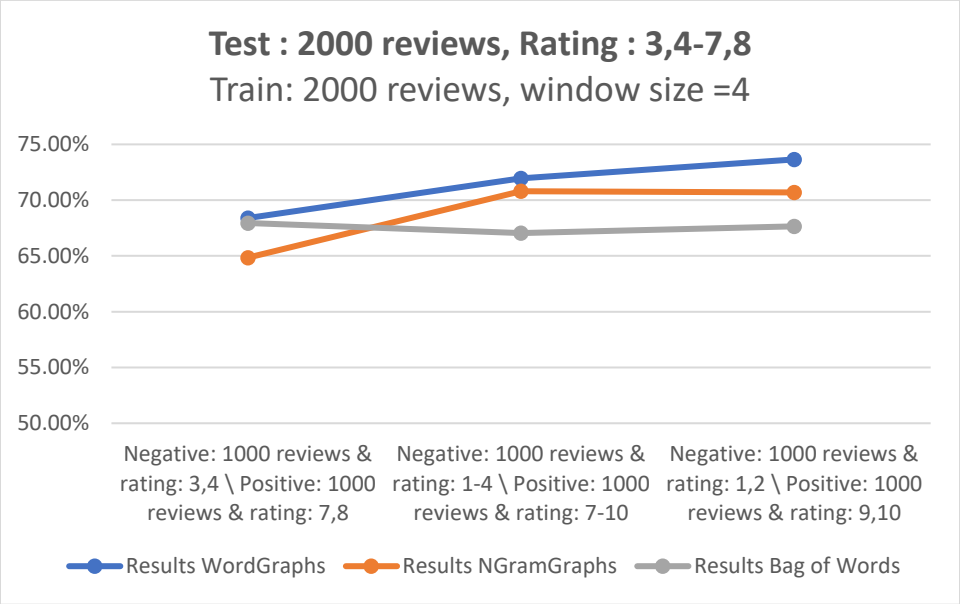
*Πίνακας 8 IMDB-Bag of Words Results - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 3,4 - 7,8*

Σε αυτή τη περίπτωση που το test set, που εξετάζουμε, είναι **2000 reviews και Rating: 3,4-7,8** , τότε για σταθερό πλήθος κριτικών στο σύνολο εκπαίδευσης, ίσο με 2000 (1000 κριτικές για κάθε πολικότητα) και window size ίσο με 4 (για τον αλγόριθμο word graphs), πετυχαίνουμε τιμές μέγιστης ακρίβειας για κάθε αλγόριθμο. Στο σχήμα 20 γίνεται γραφικά αναπαράσταση των αποτελεσμάτων με σκοπό τη σύγκριση της απόδοσης των τριών αλγορίθμων.

Ομοίως με την προηγούμενη περίπτωση, σύμφωνα με το σχήμα 20, επαληθεύεται ότι η μέθοδος bag of words μένει ανεπηρέαστη στη μεταβολή της πόλωσης του συνόλου εκπαίδευσης, ενώ οι δύο αλγόριθμοι γράφων παρουσιάζουν διαφορές.

Ο αλγόριθμος word graphs παρατηρείται ότι βελτιώνει την ακρίβεια του, όσο μεγιστοποιούμε την πόλωση στην εκπαίδευση και πετυχαίνει τη μέγιστη τιμή και από τους τρεις αλγορίθμους στην περίπτωση με τη μεγαλύτερη πόλωση. Αξιοσημείωτο είναι ότι, παρόλο που το σύνολο ελέγχου αφορά μόνο τις ενδιάμεσες βαθμολογίες, έχουμε μεγαλύτερη ακρίβεια όταν το σύνολο εκπαίδευσης περιέχει ακραίες βαθμολογίες. Συνεπώς, ανεξάρτητα ποιες βαθμολογίες περιέχει το σύνολο ελέγχου, για να έχουμε την καλύτερη ακρίβεια που μπορούμε να επιτύχουμε, σε αυτή τη βάση δεδομένων, χρησιμοποιούμε μόνο ακραίες βαθμολογίες για την εκπαίδευση της μηχανής.

Ο αλγόριθμος n-gram graphs, για σύνολο εκπαίδευσης με ενδιάμεσες μόνο βαθμολογίες, παρουσιάζει τα χειρότερα αποτελέσματα συγκριτικά με τους άλλους δύο αλγορίθμους, για όλες τις βαθμολογίες ανταγωνίζεται τον word graphs, ενώ εάν απομονώσουμε μόνο τις ακραίες βαθμολογίες στην εκπαίδευση δείχνει να μη μεταβάλλεται.




*Σχήμα 25 IMDB - Σύγκριση αλγορίθμων - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 3,4 - 7,8*

→ Σύνολο ελέγχου με 10000 κριτικές και βαθμολογίες 1,2,3,4 - 7,8,9,10 (Όλες οι βαθμολογίες):

- Αποτελέσματα **WordGraphs**

Για window size = 1:

| Model graphs: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 72.37% | 67.85% | 71.02% | 74.43% | 68.20% | 72.44% |
| Incorrectly Classified Instances | 27.63% | 32.15% | 28.98% | 25.57% | 31.80% | 27.56% |
| Kappa statistic | 0.4474 | 0.357 | 0.4204 | 0.4886 | 0.364 | 0.4488 |
| Mean absolute error | 0.3071 | 0.3644 | 0.3318 | 0.2866 | 0.3677 | 0.3133 |
| Root mean squared error | 0.4552 | 0.4683 | 0.4514 | 0.4373 | 0.4644 | 0.4454 |
| Relative absolute error | 61.42% | 72.88% | 66.36% | 57.32% | 73.55% | 62.66% |
| Root relative squared error | 91.03% | 93.66% | 90.28% | 87.45% | 92.88% | 89.08% |
| Coverage of cases (0.95 level) | 93.23% | 96.40% | 96.23% | 93.89% | 96.45% | 95.02% |
| Mean rel. region size (0.95 level) | 78.94% | 89.36% | 86.11% | 77.55% | 89.10% | 82.12% |

*Πίνακας 9 IMDB-WordGraphs Results (window size = 1) - Σύνολο ελέγχου με 10.000 κριτικές και βαθμολογίες 1,2,3,4 - 7,8,9,10*

Για window size = 4 :

| Model graphs: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 78.28% | 72.25% | 76.59% | 79.91% | 70.90% | 77.38% |
| Incorrectly Classified Instances | 21.72% | 27.75% | 23.41% | 20.09% | 29.10% | 22.62% |
| Kappa statistic | 0.5656 | 0.445 | 0.5318 | 0.5982 | 0.418 | 0.5476 |
| Mean absolute error | 0.2473 | 0.3319 | 0.2735 | 0.2294 | 0.3407 | 0.2644 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Root mean squared error | 0.4029 | 0.4345 | 0.4132 | 0.3874 | 0.4404 | 0.4067 |
| Relative absolute error | 49.46% | 66.39% | 54.70% | 45.87% | 68.14% | 52.88% |
| Root relative squared error | 80.57% | 86.91% | 82.64% | 77.48% | 88.08% | 81.33% |
| Coverage of cases (0.95 level) | 94.89% | 97.50% | 95.96% | 95.31% | 97.23% | 95.99% |
| Mean rel. region size (0.95 level) | 74.96% | 89.28% | 80.09% | 72.83% | 87.78% | 78.39% |

*Πίνακας 10 IMDB-WordGraphs Results (window size = 4) - Σύνολο ελέγχου με 10.000 κριτικές και βαθμολογίες 1,2,3,4 - 7,8,9,10*

- Αποτελέσματα **NGramGraphs**

| Model graphs: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results NGramGraphs** | | | | | | |
| Correctly Classified Instances | 77% | 71.43% | 75.75% | 77.38% | 68.11% | 76.23% |
| Incorrectly Classified Instances | 23% | 28.57% | 24.25% | 22.62% | 31.89% | 23.77% |
| Kappa statistic | 0.54 | 0.4286 | 0.515 | 0.5476 | 0.3622 | 0.5246 |
| Mean absolute error | 0.2479 | 0.3267 | 0.2751 | 0.2501 | 0.3612 | 0.283 |
| Root mean squared error | 0.4285 | 0.449 | 0.4241 | 0.421 | 0.4586 | 0.4177 |
| Relative absolute error | 49.58% | 65.33% | 55.01% | 50.02% | 72.23% | 56.59% |
| Root relative squared error | 85.70% | 89.79% | 84.83% | 84.20% | 91.72% | 83.55% |
| Coverage of cases (0.95 level) | 91.14% | 95.93% | 94.85% | 92.82% | 96.87% | 95.75% |
| Mean rel. region size (0.95 level) | 68.93% | 86.12% | 77.97% | 71.60% | 89.98% | 81.12% |

*Πίνακας 11 IMDB-NGramGraphs Results - Σύνολο ελέγχου με 10.000 κριτικές και βαθμολογίες 1,2,3,4 - 7,8,9,10*

- Αποτελέσματα **Bag of Words**

| Training Set: | Negative: 500 reviews & rating: 1,2 \ Positive: 500 reviews & rating: 9,10 | Negative: 500 reviews & rating: 3,4 \ Positive: 500 reviews & rating: 7,8 | Negative: 500 reviews & rating: 1,2,3,4 \ Positive: 500 reviews & rating: 7,8,9,10 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 9,10 | Negative: 1000 reviews & rating: 3,4 \ Positive: 1000 reviews & rating: 7,8 | Negative: 1000 reviews & rating: 1,2,3,4 \ Positive: 1000 reviews & rating: 7,8,9,10 |
|---|---|---|---|---|---|---|
| **Results Bag of Words** | | | | | | |
| Correctly Classified Instances | 71.75% | 70.07% | 71.10% | 74.04% | 71.72% | 73.30% |

| | | | | | | |
|---|---|---|---|---|---|---|
| Incorrectly Classified Instances | 28.25% | 29.93% | 28.90% | 25.96% | 28.28% | 26.70% |
| Kappa statistic | 0.435 | 0.4014 | 0.422 | 0.4808 | 0.4344 | 0.466 |
| Mean absolute error | 0.2837 | 0.3025 | 0.2916 | 0.2614 | 0.2876 | 0.2683 |
| Root mean squared error | 0.5191 | 0.5323 | 0.5254 | 0.4969 | 0.5147 | 0.5017 |
| Relative absolute error | 56.74% | 60.50% | 58.31% | 52.27% | 57.52% | 53.66% |
| Root relative squared error | 103.82% | 106.46% | 105.08% | 99.38% | 102.94% | 100.33% |
| Coverage of cases (0.95 level) | 75.65% | 75.17% | 75.10% | 77.83% | 77.15% | 77.76% |
| Mean rel. region size (0.95 level) | 54.19% | 55.78% | 54.64% | 54.28% | 56.76% | 54.93% |

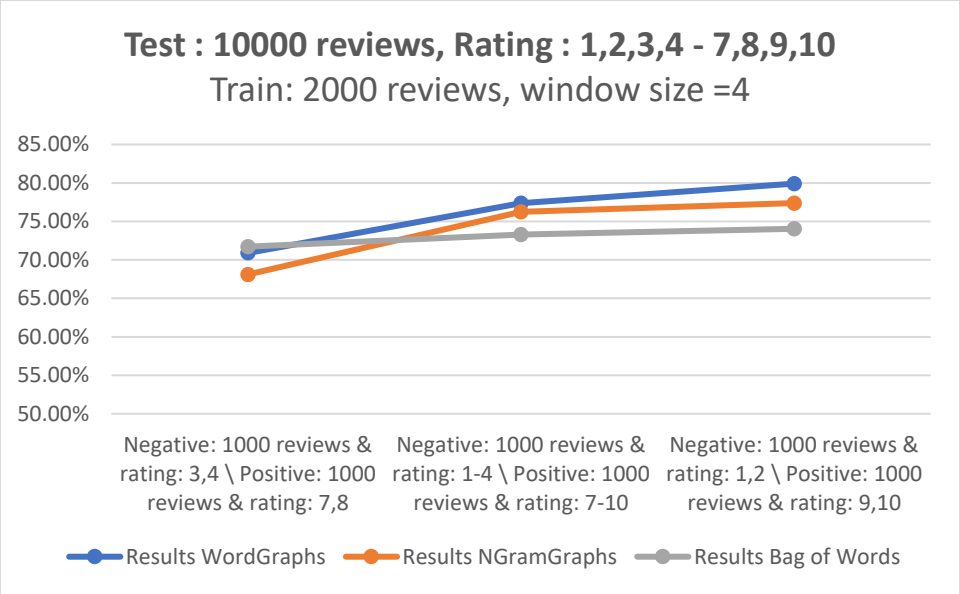
*Πίνακας 12 IMDB-Bag of Words Results - Σύνολο ελέγχου με 10.000 κριτικές και βαθμολογίες 1,2,3,4 - 7,8,9,10*

Στην περίπτωση, λοιπόν, που **το σύνολο ελέγχου περιέχει όλες τις βαθμολογίες** συγκεντρώνουμε τα αποτελέσματα για 2000 κριτικές στο σύνολο εκπαίδευσης και window size = 4 (για την περίπτωση του αλγορίθμου word graphs), που όπως ήταν αναμενόμενο επιτυγχάνονται οι υψηλότερες αποδόσεις για κάθε αλγόριθμο και τα αποτελέσματα παρουσιάζονται στο διάγραμμα στο σχήμα 21.

Όπως παρατηρήσαμε και στα δύο σύνολα ελέγχου που παρουσιάσαμε προηγουμένως, η μέθοδος bag of words δείχνει να μη μεταβάλλεται σημαντικά για διαφορετικές πολώσεις στο σύνολο εκπαίδευσης. Η απόδοση της είναι συγκρίσιμη με τις αποδόσεις των άλλων δύο αλγορίθμων μόνο στη περίπτωση που το σύνολο εκπαίδευσης αποτελείται από ενδιάμεσες βαθμολογίες μόνο.

Αντίθετα με τη μέθοδο bag of words, όμως, ο αλγόριθμος word graphs και ο αλγόριθμος n-gram graphs βελτιώνουν την απόδοσή τους όσο το σύνολο εκπαίδευσης γίνεται περισσότερο πολωμένο. Ο αλγόριθμος word graphs όταν στο σύνολο εκπαίδευσης εκτός από τις ενδιάμεσες βαθμολογίες προστεθούν και ακραίες επιτυγχάνει μεγαλύτερη απόδοση, με τιμές 77,38% για σύνολο εκπαίδευσης που περιέχει όλες τις βαθμολογίες και 79.91% για σύνολο εκπαίδευσης που αποτελείται μόνο από ακραίες βαθμολογίες. Συνεπώς, στην περίπτωση που γίνεται χρήση αυτού του αλγορίθμου η εκπαίδευση με σύνολο μεγάλης έντασης πόλωσης βελτιώνει αρκετά την απόδοση, για οποιοδήποτε σύνολο ελέγχου.

Τέλος, ο αλγόριθμος n-gram παρατηρούμε ότι, ενώ παρουσιάζει μεγάλη βελτίωση της απόδοσής του (8%) όταν στο σύνολο εκπαίδευσης εκτός από τις ενδιάμεσες βαθμολογίες προστεθούν και οι ακραίες , όταν αφαιρεθούν οι ενδιάμεσες βαθμολογίες από το σύνολο εκπαίδευσης παρατηρείται επίσης αύξηση της ακρίβειας αλλά μικρότερη της τάξης του 1%.




*Σχήμα 26 IMDB - Σύγκριση αλγορίθμων - Σύνολο ελέγχου με 10.000 κριτικές και βαθμολογίες 1,2,3,4 - 7,8,9,10*

### 5.1.2 Πείραμα 2ο : Κριτικές επιχειρήσεων YELP - Αποτελέσματα

- Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1 - 5 (Ακραίες Βαθμολογίες):

  - Αποτελέσματα **WordGraphs**

Για window size = 1:

| Model graphs: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 82.40% | 76.40% | 82.55% | 86.40% | 77.95% | 82.85% |
| Incorrectly Classified Instances | 17.60% | 23.60% | 17.45% | 13.60% | 22.05% | 17.15% |
| Kappa statistic | 0.648 | 0.528 | 0.651 | 0.728 | 0.559 | 0.657 |
| Mean absolute error | 0.1866 | 0.2674 | 0.2058 | 0.1546 | 0.2511 | 0.1928 |
| Root mean squared error | 0.3695 | 0.4093 | 0.3519 | 0.3276 | 0.4064 | 0.3559 |
| Relative absolute error | 37.31% | 53.48% | 41.17% | 30.92% | 50.22% | 38.57% |
| Root relative squared error | 71.21% | 81.86% | 70.37% | 65.52% | 81.27% | 71.17% |
| Coverage of cases (0.95 level) | 95.90% | 96% | 96.80% | 95.65% | 94.90% | 96.30% |
| Mean rel. region size (0.95 level) | 67.73% | 79.03% | 72.50% | 64.08% | 75.13% | 69.95% |

*Πίνακας 13 YELP-WordGraphs Results (window size = 1) - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1 - 5*

Για window size = 4 :

| Model graphs: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|

| **Results WordGraphs** | | | | | | |
|---|---|---|---|---|---|---|
| Correctly Classified Instances | 84.30% | 77.45% | 82.60% | 83.35% | 79.80% | 87.80% |
| Incorrectly Classified Instances | 15.70% | 22.55% | 17.40% | 16.65% | 20.20% | 12.20% |
| Kappa statistic | 0.686 | 0.549 | 0.652 | 0.667 | 0.596 | 0.756 |
| Mean absolute error | 0.1737 | 0.2742 | 0.1898 | 0.1711 | 0.2134 | 0.1444 |
| Root mean squared error | 0.3413 | 0.3905 | 0.3451 | 0.371 | 0.3906 | 0.2977 |
| Relative absolute error | 34.74% | 54.83% | 37.96% | 34.22% | 42.68% | 28.89% |
| Root relative squared error | 68.26% | 78.10% | 69.03% | 74.20% | 78.12% | 59.54% |
| Coverage of cases (0.95 level) | 95.45% | 97.35% | 97.35% | 91.75% | 93.80% | 97.35% |
| Mean rel. region size (0.95 level) | 65.98% | 81.98% | 70.28% | 60% | 67.60% | 65.93% |

*Πίνακας 14 YELP-WordGraphs Results (window size = 4) - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1 - 5*

- Αποτελέσματα **NGramGraphs**

| Model graphs: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results NGramGraphs** | | | | | | |
| Correctly Classified Instances | 87.80% | 77% | 88.75% | 83.85% | 76.30% | 88.90% |
| Incorrectly Classified Instances | 12.20% | 23% | 11.25% | 16.15% | 23.70% | 11.10% |
| Kappa statistic | 0.756 | 0.54 | 0.775 | 0.677 | 0.526 | 0.778 |
| Mean absolute error | 0.1482 | 0.2628 | 0.1486 | 0.1774 | 0.251 | 0.1559 |
| Root mean squared error | 0.3078 | 0.3896 | 0.2894 | 0.3591 | 0.4253 | 0.2932 |
| Relative absolute error | 29.63% | 52.56% | 29.72% | 35.49% | 50.20% | 31.18% |
| Root relative squared error | 61.55% | 77.91% | 57.88% | 71.82% | 85.05% | 58.64% |
| Coverage of cases (0.95 level) | 96.25% | 97.65% | 98.20% | 93.75% | 91.95% | 97.80% |
| Mean rel. region size (0.95 level) | 65.20% | 80.68% | 69.53% | 64.38% | 69.73% | 70.70% |

*Πίνακας 15 YELP-NGramGraphs Results - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1 - 5*

- Αποτελέσματα **Bag of Words**

| Training Set: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & |
|---|---|---|---|---|---|---|

| | & rating: 5 | & rating: 4 | & rating: 4,5 | & rating: 5 | & rating: 4 | rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results Bag of Words** | | | | | | |
| Correctly Classified Instances | 69.45% | 69.35% | 71.50% | 70.15% | 70.20% | 70.05% |
| Incorrectly Classified Instances | 30.55% | 30.65% | 28.50% | 29.85% | 29.80% | 29.95% |
| Kappa statistic | 38.90% | 38.70% | 43.00% | 40.30% | 40.40% | 40.10% |
| Mean absolute error | 30.60% | 30.75% | 28.48% | 29.80% | 29.79% | 29.91% |
| Root mean squared error | 55.16% | 54.99% | 52.99% | 54.37% | 54.19% | 54.42% |
| Relative absolute error | 61.21% | 61.50% | 56.97% | 59.61% | 59.58% | 59.83% |
| Root relative squared error | 110.32% | 109.98% | 105.98% | 108.75% | 108.39% | 108.83% |
| Coverage of cases (0.95 level) | 69.95% | 70.65% | 72.80% | 71.00% | 71.55% | 70.90% |
| Mean rel. region size (0.95 level) | 50.55% | 51.45% | 51.13% | 50.73% | 51.33% | 50.85% |

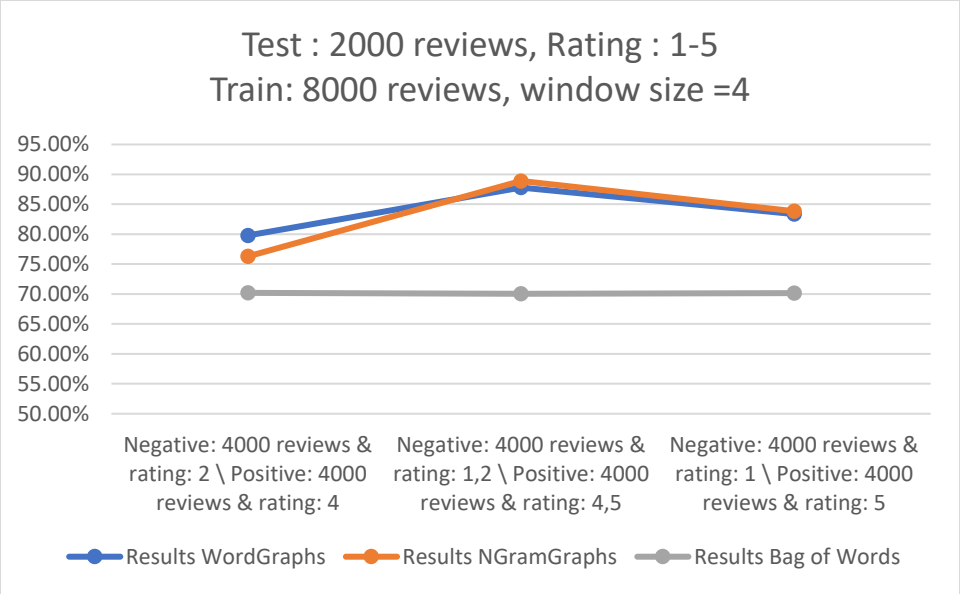
*Πίνακας 16 YELP-Bag of Words Results - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1 - 5*

Στην περίπτωση που το σύνολο ελέγχου, που εξετάζουμε, είναι **2000 κριτικές και ακραίες βαθμολογίες 1 - 5,** τότε για σταθερό πλήθος κριτικών στο σύνολο εκπαίδευσης, ίσο με 8000 (4000 κριτικές για κάθε πολικότητα) και window size ίσο με 4 στη περίπτωση του word graphs, πετυχαίνουμε τη μέγιστη απόδοση για κάθε αλγόριθμο. Στο σχήμα 22 δίνεται η γραφική αναπαράσταση των αποτελεσμάτων, για κάθε αλγόριθμο ξεχωριστά. Στον οριζόντιο άξονα έχουμε στα αριστερά τα μοντέλα γράφους που δημιουργήσαμε με τις ενδιάμεσες - αμφιλεγόμενες βαθμολογίες, στην μέση τους γράφους με όλες τις βαθμολογίες και στα δεξιά τους γράφους με την μεγαλύτερη πόλωση (μόνο ακραίες βαθμολογίες).

Παρατηρούμε, σύμφωνα με το διάγραμμα του σχήματος 22, ότι η βάση του YELP παρουσιάζει διαφορετική συμπεριφορά από αυτή της βάσης του IMDB. Σε αυτή τη βάση δεδομένων, η μέθοδος bag of words δίνει πολύ χαμηλότερη απόδοση, συγκριτικά με τους δύο άλλους αλγόριθμους, που η απόδοση τους σχεδόν ταυτίζεται σε αυτή τη περίπτωση.

Μία ακόμη παρατήρηση, είναι ότι η πόλωση του συνόλου εκπαίδευσης, δεν επηρεάζει καθόλου τα αποτελέσματα για τη μέθοδο bag of words. Στο συγκεκριμένο πείραμα, παρατηρούμε μεγάλα ποσοστά ακρίβειας για τους αλγορίθμους word graphs και n-gram graphs και συγκεκριμένα παρουσιάζεται μέγιστο για σύνολο αρχείων εκπαίδευσης που περιλαμβάνει όλες τις βαθμολογίες, με τιμές 88.9% για τον αλγόριθμο n-gram graphs και 87,8% για τον word graphs.

Αξίζει να δώσουμε προσοχή, στο ότι, ακόμη και αν το σύνολο ελέγχου μας αποτελείται μόνο από ακραίες βαθμολογίες, μέγιστη ακρίβεια έχουμε όταν το σύνολο εκπαίδευσης αποτελείται από όλες τις βαθμολογίες.




*Σχήμα 27 YELP - Σύγκριση αλγορίθμων - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 1 - 5*

- Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 2 - 4 (Ενδιάμεσες βαθμολογίες):

  - Αποτελέσματα **WordGraphs**

Για window size = 1:

| Model graphs: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 82.40% | 76.40% | 82.55% | 86.40% | 77.95% | 82.85% |
| Incorrectly Classified Instances | 17.60% | 23.60% | 17.45% | 13.60% | 22.05% | 17.15% |
| Kappa statistic | 0.648 | 0.528 | 0.651 | 0.728 | 0.559 | 0.657 |
| Mean absolute error | 0.1866 | 0.2674 | 0.2058 | 0.1546 | 0.2511 | 0.1928 |
| Root mean squared error | 0.3695 | 0.4093 | 0.3519 | 0.3276 | 0.4064 | 0.3559 |
| Relative absolute error | 37.31% | 53.48% | 41.17% | 30.92% | 50.22% | 38.57% |
| Root relative squared error | 71.21% | 81.86% | 70.37% | 65.52% | 81.27% | 71.17% |
| Coverage of cases (0.95 level) | 95.90% | 96% | 96.80% | 95.65% | 94.90% | 96.30% |
| Mean rel. region size (0.95 level) | 67.73% | 79.03% | 72.50% | 64.08% | 75.13% | 69.95% |

*Πίνακας 17 YELP-WordGraphs Results (window size = 1) - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 2 - 4*

Για window size = 4 :

| Model graphs: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 84.30% | 77.45% | 82.60% | 83.35% | 79.80% | 87.80% |
| Incorrectly Classified Instances | 15.70% | 22.55% | 17.40% | 16.65% | 20.20% | 12.20% |
| Kappa statistic | 0.686 | 0.549 | 0.652 | 0.667 | 0.596 | 0.756 |
| Mean absolute error | 0.1737 | 0.2742 | 0.1898 | 0.1711 | 0.2134 | 0.1444 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Root mean squared error | 0.3413 | 0.3905 | 0.3451 | 0.371 | 0.3906 | 0.2977 |
| Relative absolute error | 34.74% | 54.83% | 37.96% | 34.22% | 42.68% | 28.89% |
| Root relative squared error | 68.26% | 78.10% | 69.03% | 74.20% | 78.12% | 59.54% |
| Coverage of cases (0.95 level) | 95.45% | 97.35% | 97.35% | 91.75% | 93.80% | 97.35% |
| Mean rel. region size (0.95 level) | 65.98% | 81.98% | 70.28% | 60% | 67.60% | 65.93% |

*Πίνακας 18 YELP-WordGraphs Results (window size = 4) - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 2 - 4*

- Αποτελέσματα **NGramGraphs**

| Model graphs: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results NGramGraphs** | | | | | | |
| Correctly Classified Instances | 69.75% | 73.95% | 78.05% | 65.15% | 70.65% | 75% |
| Incorrectly Classified Instances | 30.25% | 26.05% | 21.95% | 34.85% | 29.35% | 25% |
| Kappa statistic | 0.395 | 0.479 | 0.561 | 0.303 | 0.413 | 0.5 |
| Mean absolute error | 0.3106 | 0.3114 | 0.2614 | 0.3457 | 0.3075 | 0.2758 |
| Root mean squared error | 0.4973 | 0.4158 | 0.4059 | 0.5468 | 0.4608 | 0.4234 |
| Relative absolute error | 62.12% | 62.29% | 52.27% | 69.15% | 61.50% | 55.16% |
| Root relative squared error | 99.46% | 83.17% | 81.17% | 109.36% | 92.16% | 84.68% |
| Coverage of cases (0.95 level) | 86.45% | 97.55% | 95.05% | 79.90% | 92.95% | 94.15% |
| Mean rel. region size (0.95 level) | 68.53% | 86.85% | 77.83% | 63.50% | 76.63% | 76.85% |

*Πίνακας 19 YELP-NGramGraphs Results - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 2 - 4*

- Αποτελέσματα **Bag of Words**

| Training Set: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results Bag of Words** | | | | | | |
| Correctly Classified Instances | 63.10% | 65.00% | 64.30% | 63.80% | 64.50% | 63.50% |
| Incorrectly Classified Instances | 36.90% | 35.00% | 35.70% | 36.20% | 35.50% | 36.50% |

| | | | | | | |
|---|---|---|---|---|---|---|
| Kappa statistic | 26.20% | 30.00% | 28.60% | 27.60% | 29.00% | 27.00% |
| Mean absolute error | 36.98% | 35.05% | 35.50% | 36.23% | 35.38% | 36.46% |
| Root mean squared error | 60.65% | 58.67% | 59.19% | 60.00% | 58.94% | 60.13% |
| Relative absolute error | 73.96% | 70.11% | 71.00% | 72.46% | 70.75% | 72.92% |
| Root relative squared error | 121.30% | 117.33% | 118.38% | 119.99% | 117.88% | 120.27% |
| Coverage of cases (0.95 level) | 63.60% | 66.65% | 65.85% | 64.40% | 66.75% | 64.60% |
| Mean rel. region size (0.95 level) | 50.63% | 51.70% | 51.30% | 50.63% | 52.00% | 51.00% |

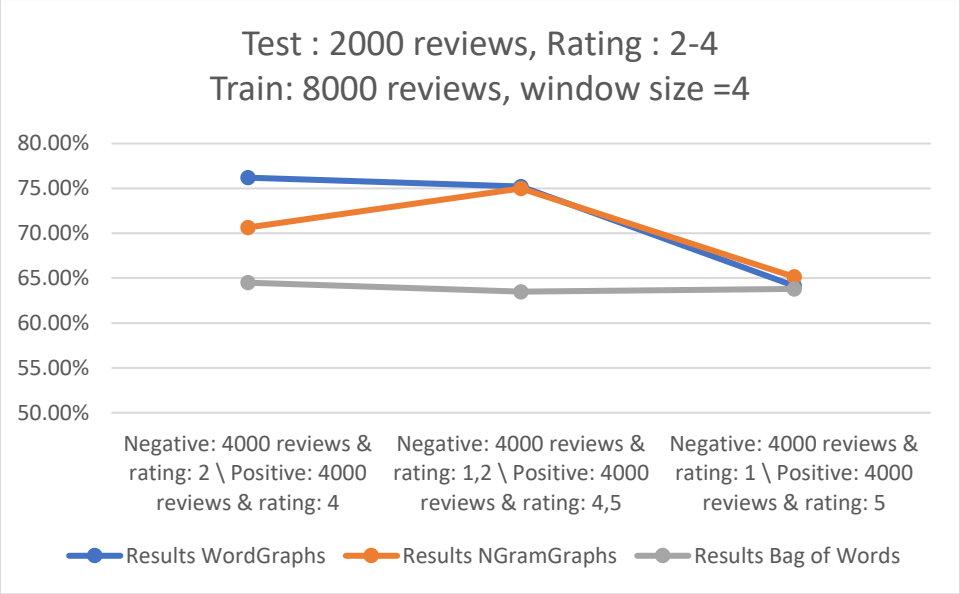
*Πίνακας 20 YELP-Bag of Words Results - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 2 – 4*

Σε αυτή τη περίπτωση, που το **σύνολο ελέγχου μας αποτελείται από τις ενδιάμεσες βαθμολογίες (2,4),** όπως φαίνεται και στο σχήμα 23, κάθε αλγόριθμος έχει τη δική του διαφορετική συμπεριφορά.

Ο αλγόριθμος word graphs δίνει την υψηλότερη τιμή ακρίβειας για σύνολο εκπαίδευσης με βαθμολογίες ίδιες με το σύνολο ελέγχου (ενδιάμεσες). Για γράφους μοντέλα με όλες τις βαθμολογίες δίνει σχεδόν την ίδια απόδοση (με μία αμελητέα πτώση της τιμής) και στη συνέχεια για πολωμένους γράφους έχουμε αρκετά χαμηλότερη ακρίβεια.

Από την άλλη πλευρά, ο n-gram graphs για γράφους με ενδιάμεσες βαθμολογίες έχει χαμηλότερη απόδοση από τον word graphs, ενώ τα υπόλοιπα αποτελέσματα ταυτίζονται με μέγιστη τιμή για σύνολα εκπαίδευσης με όλες τις βάσεις δεδομένων.

Ο αλγόριθμος bag of words εξακολουθεί να παρουσιάζει πολύ χαμηλές τιμές ακρίβειας και να είναι ανθεκτικός στη μεταβολή της πολικότητας του συνόλου εκπαίδευσης, με αμελητέες αυξομειώσεις στην απόδοση.




*Σχήμα 28 YELP - Σύγκριση αλγορίθμων - Σύνολο ελέγχου με 2000 κριτικές και βαθμολογίες 2 - 4*

- Σύνολο ελέγχου με 10000 κριτικές και βαθμολογίες 1,2 – 4,5 (Όλες οι βαθμολογίες):

  - Αποτελέσματα **WordGraphs**

Για window size = 1:

| Model graphs: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 76.27% | 74.04% | 76.85% | 79.24% | 75.05% | 76.94% |
| Incorrectly Classified Instances | 23.73% | 25.96% | 23.15% | 20.76% | 24.95% | 23.06% |
| Kappa statistic | 0.5254 | 0.4808 | 0.537 | 0.5848 | 0.501 | 0.5388 |
| Mean absolute error | 0.2474 | 0.2938 | 0.2587 | 0.2182 | 0.2825 | 0.2484 |
| Root mean squared error | 0.4244 | 0.435 | 0.4073 | 0.4048 | 0.4332 | 0.4202 |
| Relative absolute error | 49.48% | 58.77% | 51.74% | 43.63% | 56.50% | 49.68% |
| Root relative squared error | 84.87% | 87.01% | 81.45% | 80.96% | 86.65% | 84.04% |
| Coverage of cases (0.95 level) | 92.76% | 95.17% | 95.07% | 92.29% | 93.89% | 93.67% |
| Mean rel. region size (0.95 level) | 70.00% | 80.33% | 75.74% | 66.34% | 77.36% | 72.10% |

*Πίνακας 21 YELP-WordGraphs Results (window size = 1) - Σύνολο ελέγχου με 10.000 κριτικές και βαθμολογίες 1,2 - 4,5*

Για window size = 4 :

| Model graphs: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results WordGraphs** | | | | | | |
| Correctly Classified Instances | 74.32% | 75.02% | 76.76% | 72.99% | 77.10% | 80.80% |
| Incorrectly Classified Instances | 25.68% | 24.98% | 23.24% | 27.01% | 22.90% | 19.20% |
| Kappa statistic | 0.4864 | 0.5004 | 0.5352 | 0.4598 | 0.542 | 0.616 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mean absolute error | 0.2626 | 0.3 | 0.249 | 0.2706 | 0.242 | 0.2093 |
| Root mean squared error | 0.4497 | 0.4175 | 0.4117 | 0.4858 | 0.4156 | 0.3795 |
| Relative absolute error | 52.52% | 60.00% | 49.80% | 54.13% | 48.40% | 41.86% |
| Root relative squared error | 89.94% | 83.49% | 82.34% | 97.17% | 83.13% | 75.90% |
| Coverage of cases (0.95 level) | 89.79% | 96.48% | 94.92% | 82.98% | 93.39% | 94.55% |
| Mean rel. region size (0.95 level) | 67.41% | 83.08% | 73.16% | 60.33% | 70.34% | 68.59% |

*Πίνακας 22 YELP-WordGraphs Results (window size = 4) - Σύνολο ελέγχου με 10.000 κριτικές και βαθμολογίες 1,2 - 4,5*

- Αποτελέσματα **NGramGraphs**

| Model graphs: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results NGramGraphs** | | | | | | |
| Correctly Classified Instances | 77.99% | 74.69% | 81.61% | 74.13% | 72.01% | 81.12% |
| Incorrectly Classified Instances | 22.01% | 25.31% | 18.39% | 25.87% | 27.99% | 18.88% |
| Kappa statistic | 0.5598 | 0.4938 | 0.6322 | 0.4826 | 0.4402 | 0.6224 |
| Mean absolute error | 0.2382 | 0.2952 | 0.2172 | 0.2686 | 0.2904 | 0.2261 |
| Root mean squared error | 0.4211 | 0.4106 | 0.3667 | 0.4695 | 0.4553 | 0.3752 |
| Relative absolute error | 47.64% | 59.03% | 43.44% | 53.72% | 58.07% | 45.23% |
| Root relative squared error | 84.23% | 82.12% | 73.34% | 93.89% | 91.06% | 75.04% |
| Coverage of cases (0.95 level) | 90.97% | 97.53% | 96.30% | 85.80% | 91.91% | 95.51% |
| Mean rel. region size (0.95 level) | 67.47% | 84.44% | 74.11% | 63.99% | 73.57% | 74.19% |

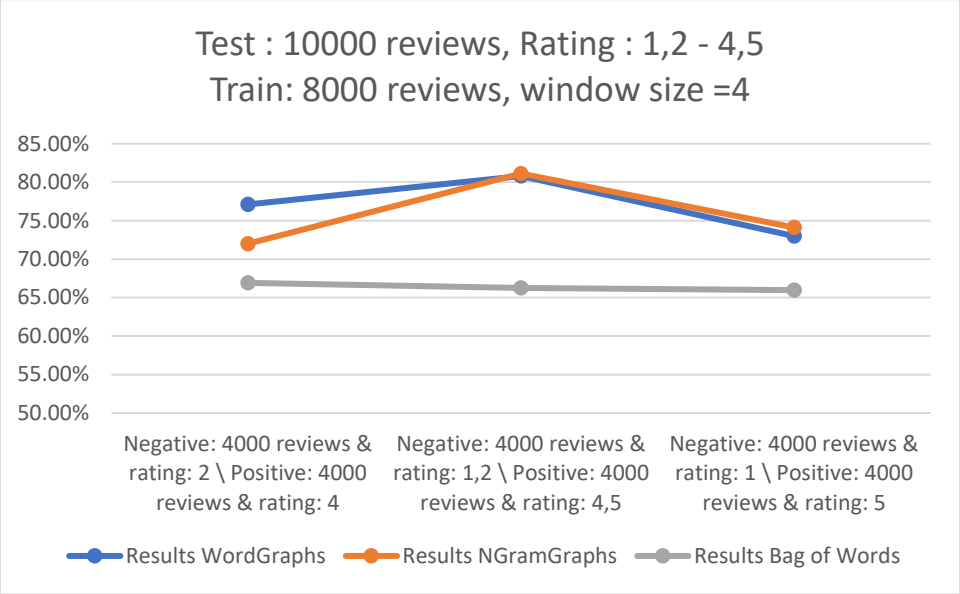
*Πίνακας 23 YELP-NGramGraphs Results - Σύνολο ελέγχου με 10.000 κριτικές και βαθμολογίες 1,2 - 4,5*

- Αποτελέσματα **Bag of Words**

| Training Set: | Negative: 1000 reviews & rating: 1 \ Positive: 1000 reviews & rating: 5 | Negative: 1000 reviews & rating: 2 \ Positive: 1000 reviews & rating: 4 | Negative: 1000 reviews & rating: 1,2 \ Positive: 1000 reviews & rating: 4,5 | Negative: 4000 reviews & rating: 1 \ Positive: 4000 reviews & rating: 5 | Negative: 4000 reviews & rating: 2 \ Positive: 4000 reviews & rating: 4 | Negative: 4000 reviews & rating: 1,2 \ Positive: 4000 reviews & rating: 4,5 |
|---|---|---|---|---|---|---|
| **Results Bag of Words** | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Correctly Classified Instances | 65.82% | 66.49% | 67.08% | 65.97% | 66.91% | 66.26% |
| Incorrectly Classified Instances | 34.18% | 33.51% | 32.92% | 34.03% | 33.09% | 33.74% |
| Kappa statistic | 0.3164 | 0.3298 | 0.3416 | 0.3194 | 0.3382 | 0.3252 |
| Mean absolute error | 0.3417 | 0.3348 | 0.3284 | 0.34 | 0.3305 | 0.3373 |
| Root mean squared error | 0.5826 | 0.5733 | 0.5694 | 0.5809 | 0.5704 | 0.5782 |
| Relative absolute error | 68.34% | 66.95% | 65.68% | 68.00% | 66.10% | 67.47% |
| Root relative squared error | 116.51% | 114.65% | 113.87% | 116.18% | 114.08% | 115.65% |
| Coverage of cases (0.95 level) | 66.54% | 68.33% | 68.42% | 66.74% | 68.45% | 67.15% |
| Mean rel. region size (0.95 level) | 50.70% | 51.77% | 51.20% | 50.72% | 51.53% | 50.87% |

*Πίνακας 24 YELP-Bag of Words Results - Σύνολο ελέγχου με 10.000 κριτικές και βαθμολογίες 1,2 - 4,5*




*Σχήμα 29 YELP - Σύγκριση αλγορίθμων - Σύνολο ελέγχου με 10.000 κριτικές και βαθμολογίες 1,2 - 4,5*

## 5.2 Σύγκριση αποτελεσμάτων κάθε αλγορίθμου για διαφορετικά σύνολα εκπαίδευσης και ελέγχου

Στη συνέχεια συγκεντρώνουμε τα αποτελέσματα και από τα τρία διαφορετικά σύνολα ελέγχου και παρουσιάζονται αναλυτικά οι παρατηρήσεις για κάθε αλγόριθμο ξεχωριστά.

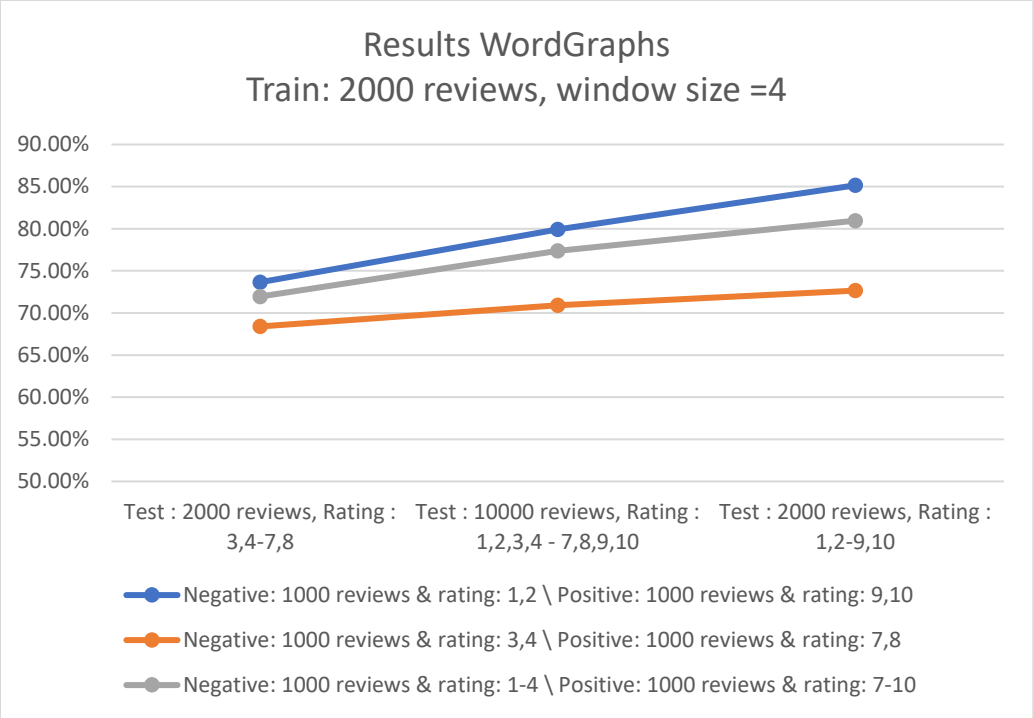
### 5.2.1 Βάση δεδομένων IMDB – Σύγκριση αποτελεσμάτων

Σε αυτή την ενότητα συγκεντρώσαμε τα αποτελέσματα κάθε αλγορίθμου, για κάθε σύνολο ελέγχου που εξετάσαμε με παραμέτρους του συνόλου εκπαίδευσης που επιτυγχάνουμε μέγιστη απόδοση (πλήθος κριτικών ίσο με 2000 και μέγεθος παραθύρου ίσο με 4, για την περίπτωση του αλγορίθμου word graphs). Κάθε γραμμή ανήκει σε ένα σύνολο εκπαίδευσης.

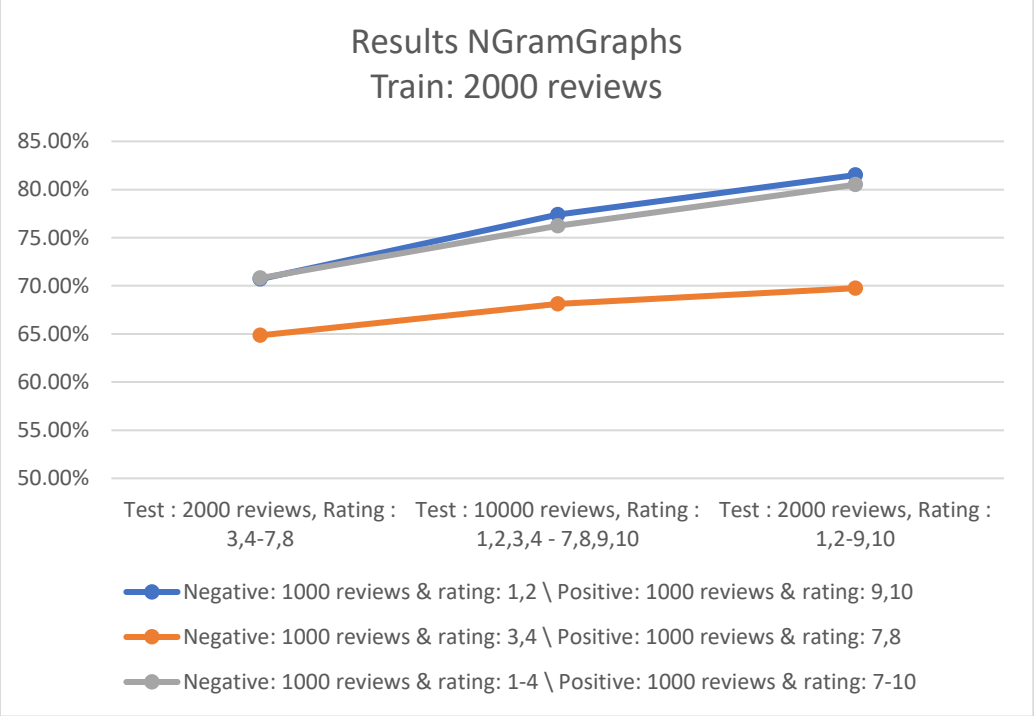
Όπως παρατηρούμε στο σχήμα 25, για τον αλγόριθμο word graphs, τις υψηλότερες τιμές απόδοσης λαμβάνει η μπλε γραμμή η οποία δηλώνει ότι χρησιμοποιήθηκε σύνολο εκπαίδευσης του αλγορίθμου το οποίο περιλαμβάνει μόνο τις ακραίες τιμές 1,2 – 9,10 (έντονη πόλωση). Στη συνέχεια ακολουθεί η περίπτωση που το σύνολο εκπαίδευσης περιέχει όλες τις βαθμολογίες και τέλος με πολύ χαμηλή απόδοση η περίπτωση με τις ενδιάμεσες βαθμολογίες.

Για τον αλγόριθμο n-gram graphs, σύμφωνα με το σχήμα 26, παρατηρούμε ότι, η περίπτωση με το σύνολο εκπαίδευσης να περιέχει τις ακραίες βαθμολογίες, και η περίπτωση εκείνου που τις περιέχει όλες σχεδόν ταυτίζουν τα αποτελέσματα τους. Ενώ παρατηρούμε ότι η περίπτωση που το σύνολο εκπαίδευσης περιέχει μόνο αμφιλεγόμενες βαθμολογίες, τότε η ακρίβεια μειώνεται σημαντικά.

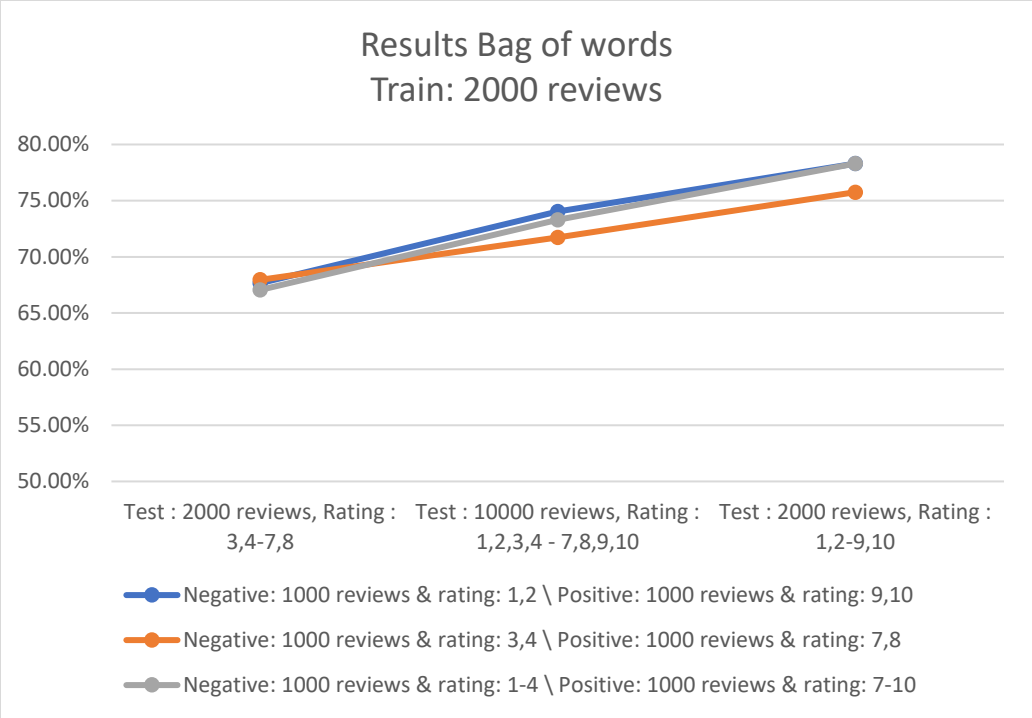
Για τα αποτελέσματα του αλγορίθμου bag of words, παρατηρούμε από το σχήμα 27, ότι δεν μεταβάλλονται ιδιαίτερα με την μεταβολή των βαθμολογιών, που περιέχει το σύνολο εκπαίδευσης, εφόσον οι γραμμές σχεδόν ταυτίζονται. Συνεπώς, σε αυτό το διάγραμμα, φαίνεται καθαρά, πως η μέθοδος bag of words δεν επηρεάζει την ακρίβεια της, όταν μεταβάλουμε την ένταση της πόλωσης στο σύνολο εκπαίδευσης του αλγορίθμου.




*Σχήμα 30 IMDB - Results WordGraphs (window size = 4) - Σύγκριση απόδοσης διαφορετικών συνόλων εκπαίδευσης*




*Σχήμα 31 IMDB - Results NGramGraphs - Σύγκριση απόδοσης διαφορετικών συνόλων εκπαίδευσης*




*Σχήμα 32 IMDB - Results Bag of Words - Σύγκριση απόδοσης διαφορετικών συνόλων εκπαίδευσης*

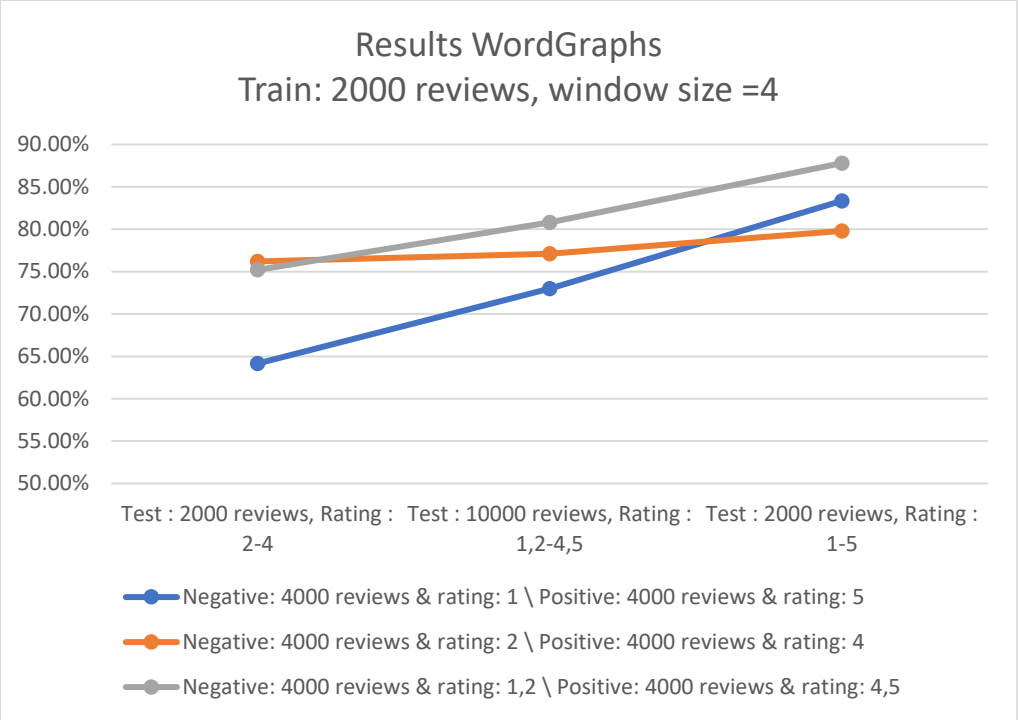
### 5.2.2 Βάση δεδομένων YELP – Σύγκριση αποτελεσμάτων

Στην περίπτωση του YELP, ακολουθήσαμε την ίδια ακριβώς διαδικασία και συγκεντρώσαμε τα αποτελέσματα από τα τρία σύνολα ελέγχου, για την περίπτωση που εκπαιδεύσαμε τους αλγορίθμους με σύνολο 2000 κριτικών και μέγεθος παραθύρου ίσο με 4 (για τον αλγόριθμο word graphs), σκοπό να μελετήσουμε τη συμπεριφορά κάθε αλγορίθμου ξεχωριστά για κάθε σύνολο εκπαίδευσης με διαφορετική ένταση πόλωσης.

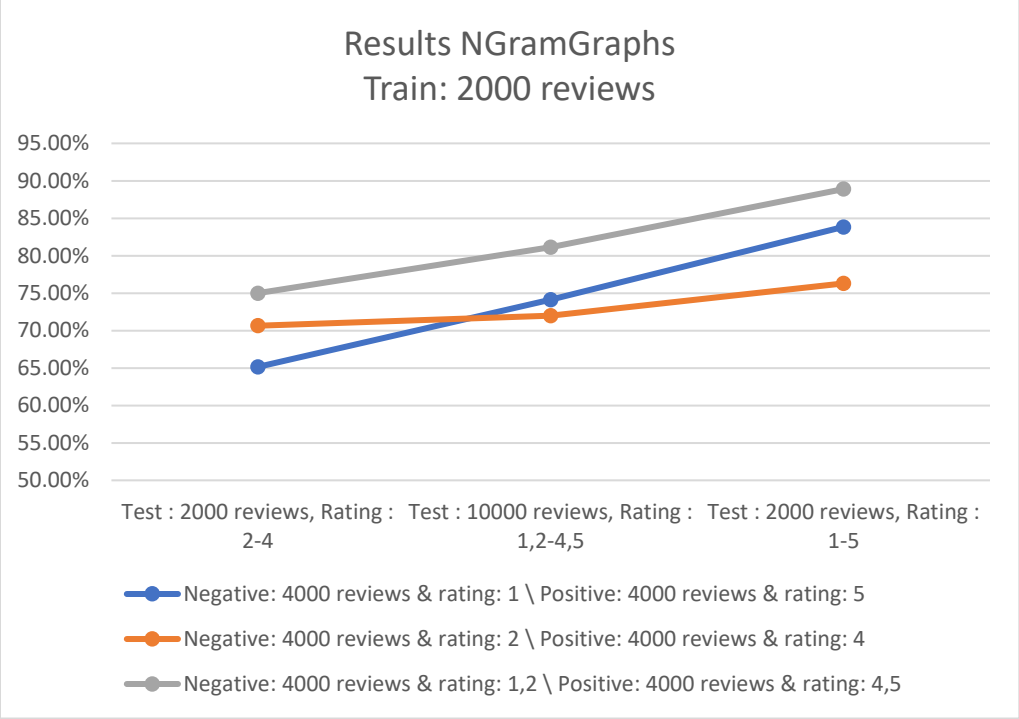
Όπως φαίνεται στο σχήμα 28, για τη μέθοδο **word graphs**, παρατηρείται να επιτυγχάνει καλύτερη συνολική απόδοση η γκρι γραμμή, που αντιπροσωπεύει τα μοντέλα γράφους που είναι εκπαιδευμένα με όλες τις βαθμολογίες. Επίσης, αξίζει να δώσουμε προσοχή στο γεγονός, ότι όταν το σύνολο ελέγχου αποτελείται από ενδιάμεσες βαθμολογίες είναι προτιμότερο να χρησιμοποιηθεί σύνολο εκπαίδευσης που να περιέχει ενδιάμεσες βαθμολογίες, καθώς η απουσία τους από το σύνολο εκπαίδευσης έχει σημαντικές επιπτώσεις στην ακρίβεια που επιτυγχάνουμε.

Για τον αλγόριθμο **n-gram graphs**, σύμφωνα με το σχήμα 29, παρατηρείται ότι ξεκάθαρα καλύτερα ποσοστά ακρίβειας επιτυγχάνουμε για σύνολο εκπαίδευσης που περιέχει όλες τις βαθμολογίες, ανεξάρτητα με το περιεχόμενο του συνόλου εκπαίδευσης. Αντίστοιχα, και για την περίπτωση αυτού του αλγορίθμου, όταν το σύνολο ελέγχου περιέχει ενδιάμεσες βαθμολογίες, καλύτερη απόδοση έχουμε για σύνολα εκπαίδευσης που περιέχουν ενδιάμεσες βαθμολογίες, καθώς εάν αυτές αφαιρεθούν το ποσοστό ακρίβειας μειώνεται σημαντικά.

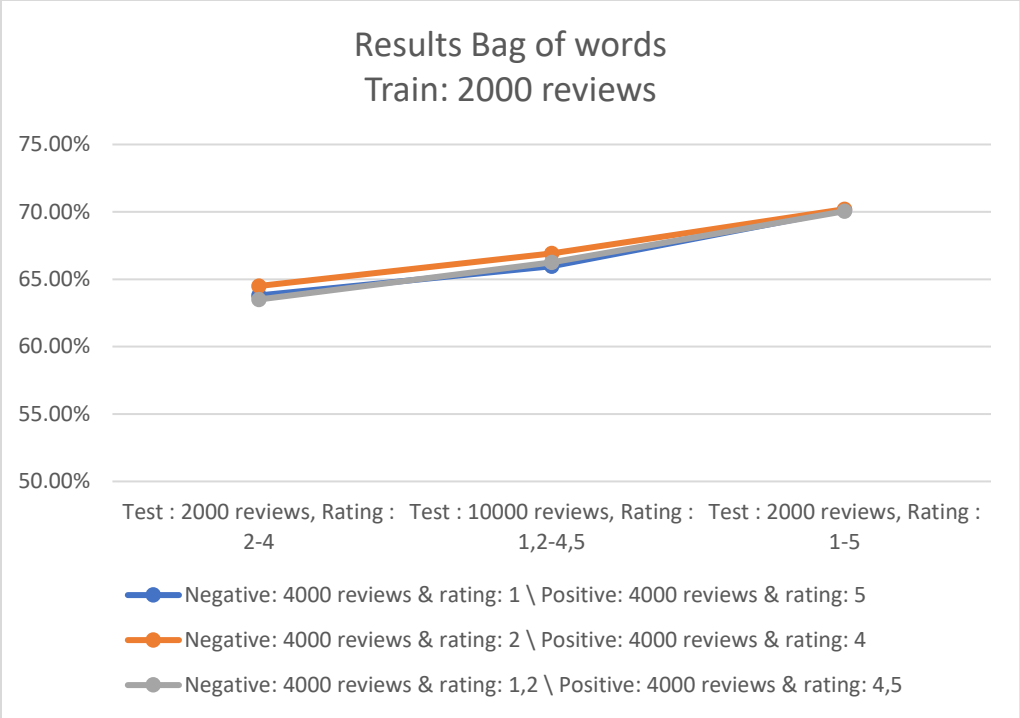
Στη μέθοδο **bag of words**, η μεταβολή της έντασης της πόλωσης στο σύνολο εκπαίδευσης δεν δείχνει να επηρεάζει τα αποτελέσματα της ακρίβειας της μεθόδου, όπως φαίνεται στο σχήμα 30, καθώς τα αποτελέσματα για κάθε σύνολο εκπαίδευσης που μελετήθηκε ταυτίζονται.




*Σχήμα 33 YELP - Results WordGraphs (window size = 4) - Σύγκριση απόδοσης διαφορετικών συνόλων εκπαίδευσης*




*Σχήμα 34 YELP - Results NGramGraphs - Σύγκριση απόδοσης διαφορετικών συνόλων εκπαίδευσης*




*Σχήμα 35 YELP - Results Bag of Words - Σύγκριση απόδοσης διαφορετικών συνόλων εκπαίδευσης*

## 5.3 Συμπεράσματα

Συνοψίζοντας, στα δύο πειράματα που εκτελέσαμε, παρατηρήσαμε ότι στη βάση δεδομένων κριτικών του IMDB η χρήση συνόλου εκπαίδευσης που χαρακτηρίζεται από εντονότερη έκφραση της έντασης της πόλωσης, βελτιώνει τα ποσοστά ακρίβειας του αλγορίθμου σημαντικά για τους αλγόριθμους word graphs και n-gram graphs, ενώ η επίδοση της μεθόδου bag of words δείχνει να μην εξαρτάται σε μεγάλο βαθμό από αυτή τη παράμετρο.

Όσον αφορά τη βάση δεδομένων κριτικών του YELP, βέλτιστες τιμές απόδοσης για τους αλγόριθμους word graphs και n-gram graphs επιτυγχάνουμε όταν το σύνολο εκπαίδευσης αποτελείται από κριτικές όλων των βαθμολογιών για οποιοδήποτε σύνολο ελέγχου. Ο αλγόριθμος bag of words μένει ανεξάρτητος από οποιαδήποτε μεταβολή της έντασης της πολικότητας του συνόλου εκπαίδευσης.

Τέλος, συνολικά τις καλύτερες τιμές ακρίβειας επιτυγχάνει ο αλγόριθμος word graphs, με τον αλγόριθμο n-gram graphs να ακολουθεί με μικρές διαφορές στην επίδοση, ενώ ο αλγόριθμος bag of words παρουσιάζει, κατά κύριο λόγο, τις χαμηλότερες τιμές ακρίβειας συγκριτικά με τους άλλους δύο αλγορίθμους.

## 5.4 Μελλοντικές εργασίες

Χρήσιμη έρευνα που θα μπορούσε να γίνει στο μέλλον είναι η εισαγωγή της ουδέτερης κλάσης πολικότητας στο πρόβλημα [13], που αποτελεί σημαντικό πεδίο μελέτης στον τομέα της ανάλυσης συναισθημάτων. Επιπλέον, εφόσον σε αυτή τη διπλωματική εργασία ερευνήσαμε τις περιπτώσεις των κριτικών ταινιών και επιχειρήσεων, χρήσιμη θα ήταν και η επέκταση της μελέτης και η αξιολόγηση της επίδοσης των αλγορίθμων σε βάση δεδομένων κριτικών προϊόντων.

# Βιβλιογραφία


[1] Opinion mining and sentiment analysis. Foundations and trends in information retrieval. Bo Pang and Lillian Lee. 2008.

[2] A state of the Art Opinion Mining And Its Application Domains. Haji Binali, Vidyasagar Potdar, Chen Wu. 2010

[3] Glossary of terms. Ron Kohavi Foster Provost. 1998

[4] Pattern Recognition and Machine Learning. Bishop, C. M. 2006

[5] Foundetions of statistical natural language processing. Christopher D Manning and Hinrich Schütze. 1999

[6] Speech and language processing. James H. Martin and Daniel Jurafsky. 2000

[7] N-gram graphs: Representing documents and document sets in summary system evaluation. George Giannakopoulos and Vangelis Karkaletsis. 2009

[8] Summarization system evaluation revisited: N-gram graphs. George Giannakopoulos, Vangelis Karkaletsis, George Vouros, and Panagiotis Stamatopoulos. 2008

[9] Sentiment analysis of social media content using n-gram graphs. Fotis Aisopos, George Papadakis, and Theodora Varvarigou. 2011

[10] Naive (bayes) at forty: The independence assumption in information retrieval. David D Lewis. 1998

[11] Introduction to information retrieval, volume 1. Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008

[12] Content vs. context for sentiment analysis: a comparative analysis over microblogs. Fotis Aisopos, George Papadakis, Konstantinos Tserpes, and Theodora Varvarigou. 2012

[13] The importance of neutral examples for learning sentiment. Moshe Koppel and Jonathan Schler. 2006

[14] Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. ChengXiang Zhai, Sean Massung, 2016

# Παράρτημα

- Μετατροπή μορφής βάσης δεδομένων του YELP από .json αρχείο σε αρχεία .txt

```
import json
data = []
with open('yelp_academic_dataset_review.json') as f:
    for index, line in enumerate(f):
        parsed_json = json.loads(line)
        stars = parsed_json['stars']
        with open('r/{}_%s.txt'.format(index) % stars, 'w') as outfile:
            json.dump(parsed_json['text'], outfile)
```

- Κατασκευή διαγραμμάτων κατανομής του μήκους κειμένων, ανά βαθμολογία

```
import numpy as np
import matplotlib.mlab as mlab
import matplotlib.pyplot as plt
import os

try:
      path = raw_input("Filenames path= ")
      data_path = raw_input("Dataset path= ")
      wordcount = []
      with open(path) as f:
            for line in f:
                  data =[]
                  data_path_temp = []
                  data_path_temp = data_path + '/' + line
                  data_path_temp = data_path_temp.rstrip()
                  with open(data_path_temp, 'r') as review:
                        data = review.read()
                        data    =    data.replace('.','    ').replace('!','
').replace('?',' ').replace(';',' ').lower()
                        wordcount.append(len(data.split()))
      num_bins = 100
      n,  bins,  patches  =  plt.hist(wordcount,  num_bins,  facecolor='blue',
alpha=0.5)
      plt.ylabel('no of Reviews');
      plt.xlabel('Text length');
      plt.show()
```

```
	os.system('pause')
except Exception as ex:
	print ex
	raw_input()
```

- Εύρεση των λέξεων με τη μεγαλύτερη συχνότητα εμφάνισης

```
import os
import collections
import pandas as pd
import matplotlib.pyplot as plt

try:
	path = raw_input("Filenames path= ")
	data_path = raw_input("Dataset path= ")
	wordcount = []

	# Stopwords
	stopwords = set(lines.strip() for lines in open('stopwords.txt'))

	with open(path) as f:
		for line in f:
			data =[]
			data_path_temp = []
			data_path_temp = data_path + '/' + line
			data_path_temp = data_path_temp.rstrip()
			with open(data_path_temp, 'r') as review:
				data = review.read()
				for word in data.lower().split():
					word = word.replace(".","")
					word = word.replace("-","")
					word = word.replace(",","")
					word = word.replace("?","")
					word = word.replace(":","")
					word = word.replace("<br /><br />","")
					word = word.replace("\"","")
					word = word.replace("/","")
					word = word.replace(">","")
					word = word.replace("!","")
					word = word.replace("â€œ","")
					word = word.replace("â€˜","")
					word = word.replace("*","")
					word = word.replace("'","")
```

```
                                if word not in stopwords:
                                        wordcount.append(word)

        # Print most common word
        n_print = int(input("How many most common words to print: "))
        print("\nOK. The {} most common words are as follows\n".format(n_print))
        word_counter = collections.Counter(wordcount)
        for word, count in word_counter.most_common(n_print):
                print(word, ": ", count)

        # Create a data frame of the most common words
        # Draw a bar chart
        lst = word_counter.most_common(n_print)
        df = pd.DataFrame(lst, columns = ['Word', 'Count'])
        df.plot.bar(x='Word',y='Count')
        plt.show()
        os.system('pause')
except Exception as ex:
        print ex
        raw_input()
```