Ονοματεπώνυμο: Ανδρέας Γουλέτας (3170031), Λούντζης Λάμπρος (3170095)
Εργασία: Τεχνητή Νοημοσύνη
Ημερομηνία: 20 Ιανουαρίου 2021

1 Εισαγωγή

1.1 Περιγραφή Εργασίας

Δοθέντος του συνόλου δεδομένων **IMDb Dataset**, καλούμαστε να εκπαιδεύσουμε δύο μοντέλα μηχανικής μάθησης με σκοπό την **κατάταξη (classification)** κριτικών ταινιών σε δυο κατηγορίες, θετικές και ανρητικές.

Για την λύση του προβλήματος της κατάταξης, επιλέξαμε τους αλγορίθμους επιβλεπόμενης μάθησης **Multinomial Naive Bayes** και **Random Forest** με χρήση **ID3** δέντρων. Οι αλγόριθμοι εκπαιδεύονται στο 90% των δεδομένων του συνόλου εκπαίδευσης του IMDb Dataset, οι υπερπαράμετροί τους επιλέγονται με την την χρήση του υπόλοιπου 10% του συνόλου εκπαίδευσης, ως δεδομένα επικύρωσης, και τέλος αξιολογούνται στα δεδομένα του συνόλου αξιολόγησης του IMDb Dataset.

1.2 Δομή παραδοτέου

Αρχικά, στην Ενότητα 2 περιγράφεται η επεξεργασία που έχουν υποστεί τα δεδομένα. Στην Ενότητα 3 παρουσιάζεται η μέθοδος της επιλογής χαρακτηριστικών, τα προβλήματα που λύνουμε κάνοντας χρήση της καθώς επίσης και το μοντέλο του κλερδους πληροφορίας. Στην Ενότητα 4 παρουσιάζεται ο ταξινομητής Multinomial Naive Bayes, καθώς και τα αποτελέσματα που δίνει μέσω καμπυλών αξιολόγησης. Τέλος, στην Ενότητα 5 παρουσιάζεται ο ταξινομητής Random Forest με δέντα αποφάσεων ID3 μαζί και τα αποτελέσματά του που έχουμε εξάγει σε καμπύλες αξιολόγησης.

2 Επεξεργασία Δεδομένων

Για να χρησιμοποιήσουμε τη συλλογή IMDb Dataset, θα πρέπει τα κείμενα να μετασχηματιστούν με τέτοιο τρόπο ώστε να μπορούν να χρησιμοποιηθούν απο τα μοντέλα μας.
Αρχικά, τα κείμενα σπάνε σε όρους (tokenization), οπου κάθε όρος είναι μια λέξη του κείμενου. Έπειτα, γίνεται αφαίρεση σημείων στίξης και άλλων χαρακτήρων που δεν σχηματίζουν ακολουθίες από γράμματα και αριθμούς. Τέλος, κάθε όρος κανονικοποιείται χρησιμοποιώντας case-folding, δηλαδή μετατρέποντάς τον σε πεζά γράμματα.

2.1 Αφαίρεση stop words

Παρότι την παραπάνω επεξεργασία, παρατηρούμε ότι υπάρχουν κάποιοι όροι που εμφανίζονται συχνά στα κείμενα οι οποίοι παρέχουν μικρή αξία στην κατάταξη των κειμένων. Χαρακτηριστικά παραδείγματα, είναι οι όροι the, a, an, this κλπ. Αυτοί οι όροι ονομάζονται **stop words**, και επειδή δεν βοηθούν στην κατάταξη των κειμένων τους αφαιρούμε.

2.2 Stemming

Μια ακόμα μορφή κανονικοποίησης που χρησιμοποιούμε, για να μειώσουμε το λεξιλόγιό μας, είναι η αποκοπή καταλήξεων με **stemming**. Σκοπός της τεχνικής αυτής είναι να περικόψει τις κλιτικές καταλήξεις και να ανάγει τις παράγωγες λέξεις σε μια κοινη μορφή, το **stem**. Την παραπάνω υλοποίηση την επιτυγχάνουμε με την χρήση του **Porter stemmer**, που είναι ένας αλγοριθμικός stemmer.

παράδειγμα: biology, biological, biologically → biolog

2.3 Vectorization

Είδικα στο μοντέλο Random Forest, τα κείμενα παίρνουν την μορφή **διανυσμάτων (vector)**. Τα διανύσματα έχουν μέγεθος ανάλογο του συνόλου των χαρακτηριστικών που χρησιμοποιούνται και κάθε θέση του διανύσματος έχει τιμή 0, άν το χαρακτηριστικό δεν υπάρχει στο κείμενο, ή 1, αν το χαρακτηριστικό υπάρχει.

παράδειγμα: Έστω το κείμενο d_1 σε διανυσματική μορφή γίνεται: d_1: $< X_1, X_2, X_3..., X_n > = < 1, 0, 0, ..., 1 >$ οπου X_i είναι ένα χαρακτηριστικο.

3 Επιλογή Χαρακτηριστικών

Οι αλγόριθμοι μηχανικής μάθησης που υλοποιούμε, προκειμένου να κατατάξουν τα κείμενα στην σωστή κατηγορία χρειάζονται ένα σύνολο **χαρακτηρισιτικών**. Αυτό το σύνολο χαρακτηρισιτκών ταυτίζεται με τις μοναδικές λέξεις που μπορούν να προκείψουν από τα δεδομένα εκπαίδευσης, γνωστό και ως **λεξιλόγιο (vocabulary)**.
Είναι εύλογο να αποφανθούμε οτι λόγω της μεγάλης έκτασης των χαρακτηριστικών, η απόδοση των ταξινομητών μας κατά την εκπαίδευση και την εφαρμογή τους δεν είναι η καλύτερη δυνατή. Εκτός αυτού, υπάρχουν **χαρακτηριστικά θορύβου (noise features)** τα οποία αυξάνουν το σφάλμα κατάταξης (classification error) σε νέα δεδομένα.
Για την επίλυση των παραπάνω προβλημάτων επιλέγουμε ένα υποσύνολο των διαθέσιμων χαρακτηριστικών και συγκεκριμένα αυτών που παρέχουν αρκετή ποσότητα πληροφορίας, ώστε ο αλγόριθμός μας να πάρει την σωστή απόφαση κατάταξης. Η διαδικασία αυτή ονομάζεται **επιλογή χαρακτηριστικών (feature selection)**. Η η υλοποίησή της βασίζεται στο **κέρδος πληροφορίας (information gain)**, που δίνεται από την εξίσωση:

$$IG(X;C) = H(C) - \sum_{x} P(X = x) * H(C|X = x)$$

όπου $H(C)$ είναι η εντροπία της κλάσσης C και $H(C|X = x)$ είναι η εντροπία της κλάσσης C αν γνωρίζουμε ότι η τιμή του χαρακτηριστικού X είναι x.

4 Multinomial Naive Bayes

Το μοντέλο **multinomial Naive Bayes** αποτελεί μια πιθανοτική μέθοδο επιβλεπόμενης μάθησης. Σκοπός του μοντέλου είναι να βρεί την πιθανότερη κλάση (θετική ή αρνητική κριτική) για κάθε κείμενο:

$$c_{map} = \arg\max_{c \in \mathbb{C}} P(c|d) = \arg\max_{c \in \mathbb{C}} P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

όπου $P(t_k|c)$ είναι η δεσμευμένη πιθανότητα ενός όρου t_k να βρίσκεται σε ένα κείμενο της κλάσης c και $P(c)$ είναι η a priori πιθανότητα ενός κειμένου να ανήκει στην κλάση c.

Στην παραπάνω εξίσωση παρατηρούμε ότι γίνεται πολλαπλασιασμός πολλών πιθανοτήτων, με κίνδυνο υποχείλισης. Γι' αυτό, επιλέγουμε να λογαριθμίσουμε την παραπάνω εξίσωση ώστε να έχουμε άθροίσματα πιθανοτήτων:

$$c_{map} = \arg\max_{c \in \mathbb{C}} (\log P(c) + \sum_{1 \leq k \leq n_d} \log P(t_k|c))$$

Η a priori πιθανότητα δίνεται από την εξίσωση:

$$P(c) = \frac{N_c}{N}$$

όπου N_c είναι ο αριθμός των κειμένων στην κλάση c και N είναι ο συνολικός αριθμός των κείμένων. Αντίστοιχα, η δεσμευμένη πιθανότητα είναι:

$$P(t_k|c) = \frac{T_{tc} + 1}{\sum_{t' \in V} (T_{t'c} + 1)}$$

όπου T_{tc} είναι η **συχνότητα (term frequency)** του όρου t στα δεδομένα εκπαίδευσης της κλάσης c.

Η συχνότητα ενός όρου στα κείμενα εκπαίδευσης της κλάσης c έχει ένα μειονέκτημα: όλοι οι όροι θεωρούνται εξίσου σημαντικοί όσο αναφορά την κατάταξη του κειμένου. Γι' αυτό, επιλέγουμε να χρησιμοποιήσουμε την **συχνότητα κειμένου (document frequency)**, που ορίζεται ως το πλήθος των κειμένων της συλλογής που περιέχουν τον όρο t. Η παραπνάνω εξίσωση μετατρέπεται ως εξης:

$$P(t_k|c) = \frac{df_{tc} + 1}{\sum_{t' \in V} (df_{t'c} + 1)}$$

4.1 Επιλογή Χαρακτηριστικών

Όπως αναφέρθηκε στην Ενότητα 3, προκειμένου ο ταξινομητής Naive Bayes να είναι αποδοτικός κατά την εκπαίδευση και την εφαρμογή του, καθώς επίσης ανθεκτικός προς τα χαρακτηριστικά θορύβου, επιλέξαμε ένα υποσύνολο χαρκτηριστικών με μέγεθος 0.007% του αρχικού συνόλου χαρακτηριστικών. Η επιλογή αυτή, όπως παρουσιάζεται στην Εικόνα 1, δίνει τα καλύτερα ποσοστά ακρίβειας (accuracy) κατά την εφαρμογή του ταξινομητή στα δεδομένα επικύρωσης.

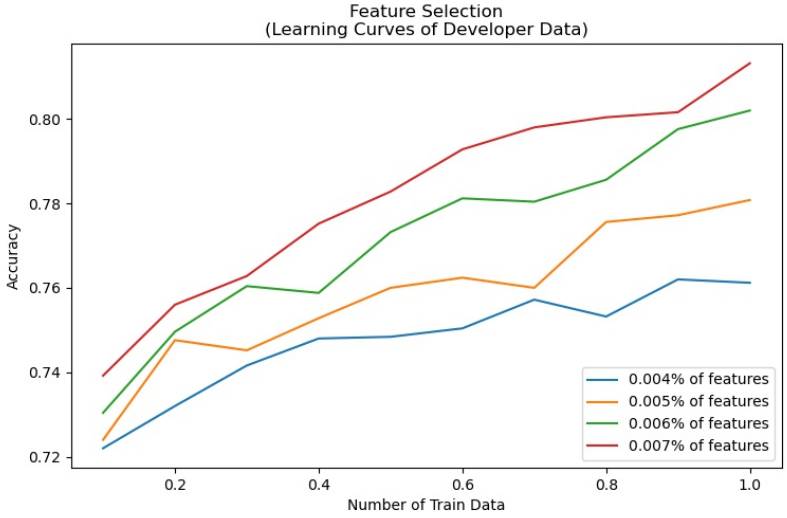

Figure 1: Feature selection in multinomial Naive Bayes

4.2 Καμπύλη Μάθησης

Για την αποτίμηση του μοντέλου multinomial Naive Bayes, χρησιμοποιούμε ως μετρική την **ακρίβεια (accuracy)**. Στην Εικόνα 2. παρουσιάζονται οι **καμπύλες μάθησης (learning curves)** από τα δεδομένα εκπαίδευσης και αξιολόγησης. Παρατηρούμε οτι το μοντέλο μας επιτυγχάνει ποσοστά ακρίβειας κοντά στο 81%.

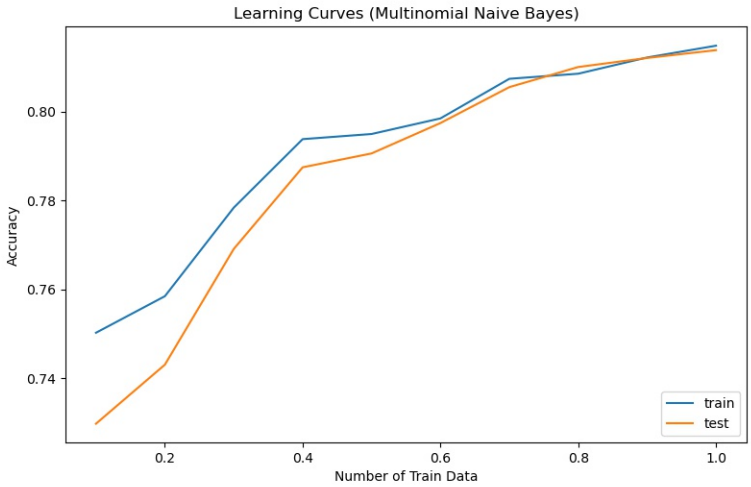

Figure 2: Multinomial Naive Bayes learning curves

4.3 Καμπύλη Precision-Recall

Για τον υπολογισμό των μετρικών **ακρίβεια (precision)** και **ανάκληση (recall)** έχουμε επιλέξει την χρήση 10 thresholds. Στην Εικόνα 3 παρατηρούμε ότι καθώς μειώνεται το precision, αυξάνεται το recall με αποτέλεσμα να έχουμε λιγότερα false negatives κατά την κατάταξη. Αντίστοιχα, καθώς μειώνεται το recall αυξάνεται το precision επιτυγχάνοντας λιγότερα false positives.

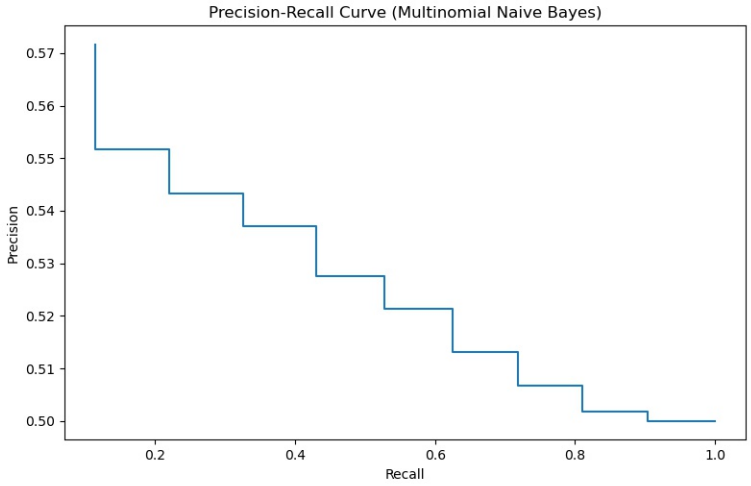

Figure 3: Multinomial Naive Bayes precision-recall curve

4.4 Καμπύλη F1

Στην Εικόνα 4 παρουσιάζεται η τιμή της μετρικής **f1** για κάθε ένα από τα 10 thresholds που είχαμε επιλέξει για τον υπολογισμό της καμπύλης precision-recall.

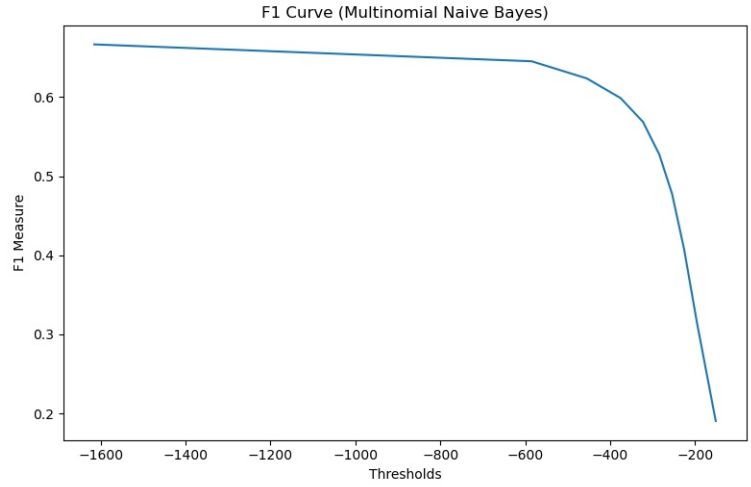

Figure 4: Multinomial Naive Bayes f1 curve

5 Random Forest με ID3

Το μοντέλο **Random Forest** αποτελεί ένα μοντέλο μάθησης συνόλου (ensemble learning). Συγκεκριμένα, δημιουργεί πολλά δένδρα αποφάσεων, όπου κάθε ένα τέτοιο δένδρο, δέχεται ως είσοδο ένα νέο σύνολο δεδομένων, ίδιου μεγέθους και δεδομένα με το αρχικό σύνολο. Χρησιμοποιεί τη τεχνική **bootstrapping**, δηλαδή για κάθε νέο σύνολο δεδομένων επιλέγει τυχαία κείμενα με επανατοποθέτηση. Μάλιστα, δέχεται ως είσοδο ένα σύνολο επιλεγμένων τυχαία ιδιοτήτων, από το αρχικό υποσύνολο καλύτερων ιδιοτήτων, χωρίς επανατοποθέτηση.

Για να προβλέψει την κλάση ενός κειμένου, τα δέντρα αποφάσεων τρέχουν παράλληλα. Κάθε ένα τέτοιο δέντρο παράγει μια απόφαση, δηλαδή την κλάση στόχο που θεωρεί οτι ανήκει το κείμενο. Αυτές οι αποφάσεις συλλέγονται και επιστρέγεται εκείνη που συμφωνεί με την πλειοψηφία.

Για την υλοποίηση των δέντρα αποφάσεων έχουμε επιλέξει τον ταξινομητή **ID3**. Η λειτουργία ενός δέντρου αποφάσεων είναι η εξής:

- Βρίσκει την ιδιότητα με το μεγαλύτερο κέρδος πληροφορίας (βλέπε Ενότητα 3). Με αυτό τον τρόπο ελπίζουμε να φτάσουμε στην σωστή ταξινόμηση με όσο το δυνατόν μικρό αριθμό ελέγχων χαρακτηριστικών.
- Χωρίζει τα δεδομένα που λαμβάνει ως είσοδο, σε δύο μέρη σύμφωνα με το αν υπάρχει η ιδιότητα ή όχι στα κείμενα.
- Επαναλαμβάνει την διαδικασία μεχρι να φτάσει σε τελικούς κόμβους. Γνωρίζουμε οτι βρισκόμαστε σε τελική κατάσταση όταν όλα τα δεδομένα, στον κόμβο προς εξέταη, ανήκουν σε μια από τις δύο κλάσεις, ή όταν δεν έχουμε άλλα χαρακτηριστικά για να ελέγξουμε ή όταν έχουμε φτάσει το μέγιστο βάθος.

5.1 Υπερπαράμετροι

Για την επιλογή των υπερπαραμέτρων στηριχθήκαμε στη καμπύλη μάθησης από τα δεδομένα επικύρωσης (validation data).
Πρώτα, με κάποιες τυχαίες δοκιμές αποφασίσαμε ότι ο καλύτερος αριθμός από **ιδιότητες (features)** είναι 100, εκ των οποίων οι 80 θα επιλέγονται τυχαία και θα δίνονται ως είσοδο στο δένδρο.
Στη συνέχεια, επιλέξαμε το **μέγιστο βάθος δέντρου (max depth)**, δοκιμάζοντας τα βάθη 5, 7, 10. Η καλύτερη τιμή accuracy επιτυγχάθηκε με βάθος 10 (την μέγιστη ακρίβεια θα μπορούσαμε να την έχουμε χρησιμοποιώντας το μέγιστο βάθος, πράγμα που απαιτεί αρκετή υπολογιστική ισχυς). Στα βάθη 5 και 7 παρατηρήθηκε το πρόβλημα του περιορισμένου χώρου αναζήτησης καθώς η τιμή της μετρικής accuracy από τα δεδομένα επικύρωσης ήταν χαμηλή.
Τέλος, για τον αριθμό των δένδρων που θα παράγονται από τον Random Forest, έγιναν δοκιμές για τις τιμές 10, 30, 50 και 100. Παρατηρήσαμε ότι όσο μεγαλύτερη είναι η τιμή των δένδρων που παράγονται τόσο καλύτερη είναι η τιμή της μετρικής accuracy.

5.2 Καμπύλη Μάθησης

Για την αποτίμηση του μοντέλου Random Forest, χρησιμοποιούμε ως μετρική την **ακρίβεια (accuracy)**. Στην Εικόνα 5. παρουσιάζονται οι **καμπύλες μάθησης (learning curves)** από τα δεδομένα εκπαίδευσης και αξιολόγησης. Παρατηρούμε οτι το μοντέλο μας επιτυγχάνει ποσοστά ακρίβειας κοντά στο 71%.

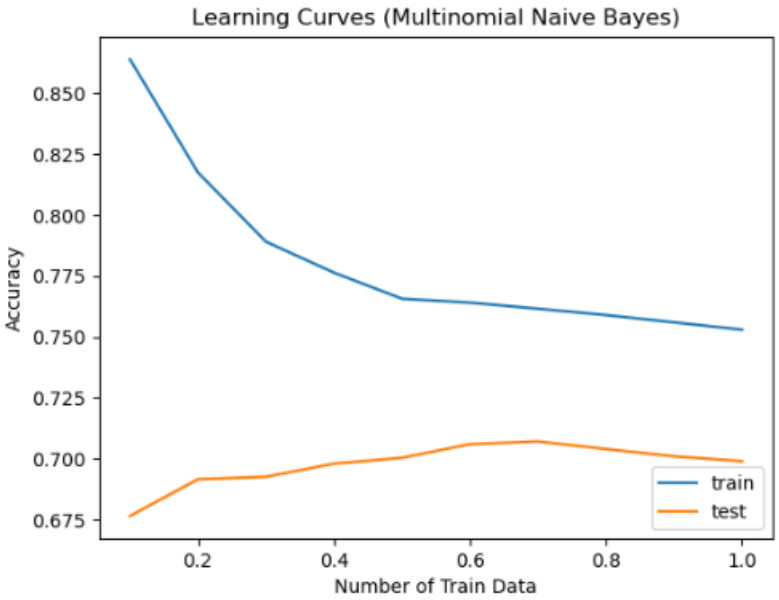

Figure 5: Random Forest learning curves

5.3 Καμπύλη Precision-Recall

Για τον υπολογισμό των μετρικών **ακρίβεια (precision)** και **ανάκληση (recall)** έχουμε επιλέξει την χρήση 10 thresholds. Στην Εικόνα 6 παρατηρούμε ότι μεταξύ των σημείων 0 και 0.2 του οριζόντιου άξονα υπάρχει μια πτώση την μετρικής precision και έπειτα αύξηση. Υποθέτουμε οτι η συμπεριφορά αυτή οφείλεται στο γεγονός ότι υπάρχουν περισσοτερα false positives από true positives. Αντίστοιχα, μεταξύ των τιμών 0.4 και 0.6 του οριζόντιου άξονα παρατηρούμε μια αύξηση της μετρικής precision καθώς η μετρική recall αυξάνεται. Αυτό μπορεί να οφείλεται στο γεγονός ότι τα true positives είναι πολυ περισσοτερα των false positives.

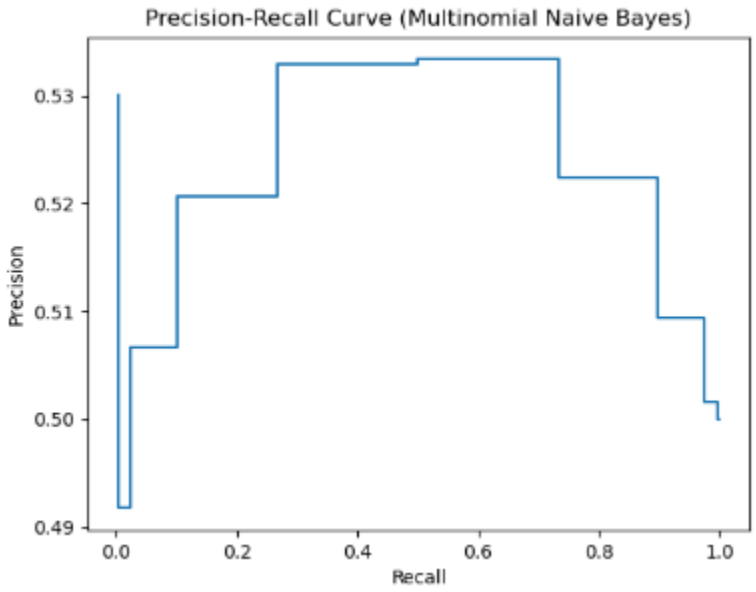

Figure 6: Random Forest precision-recall curve

5.4 Καμπύλη F1

Στην Εικόνα 7 παρουσιάζεται η τιμή της μετρικής **f1** για κάθε ένα από τα 10 thresholds που είχαμε επιλέξει για τον υπολογισμό της καμπύλης precision-recall.

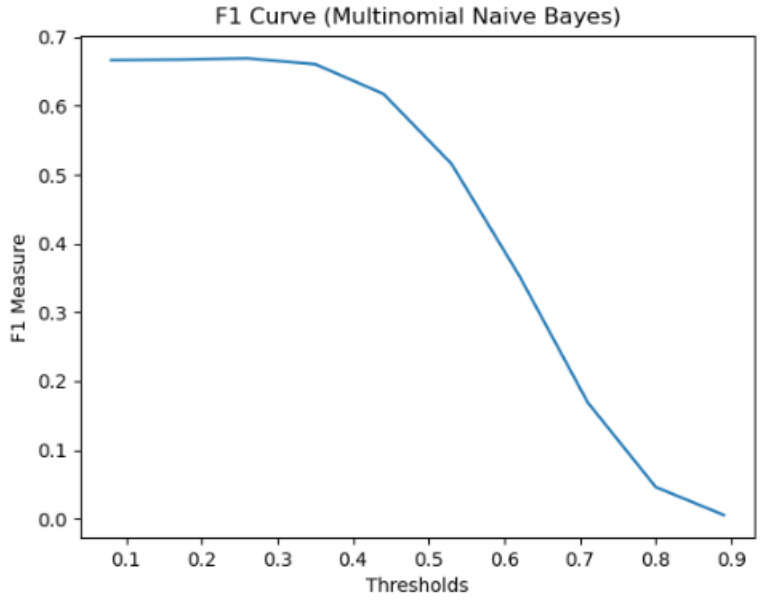

Figure 7: Random Forest f1 curve

Βιβλιογραφία

1. C. D. Manning, P. Raghavan, H. Schutze "An Introduction to Information Retrieval"
2. S. Buttcher, C. L. A. Clarke, G. V.Cormack "Information Retrieval: Implementing and Evaluating Search Engines"
3. S. Russel, P. Norvig "Artificial Intelligence: A Modern Approach"
4. T. M. Mitchell "Machine Learning"