EESTech Challenge 2022

Machine Forgetting

Antonis Parlapanis
Iasonas Pavlakis
Konstantinos Tsakonas

2 Απριλίου 2022

Περιεχόμενα

Εισαγωγή

Στον συγκεκριμένο διαγωνισμό ασχοληθήκαμε με Ταξινόμηση Κειμένου (Text Classifiation) πάνω στο σύνολο δεδομένων 20 Newsgroups. Παρακάτω θα αναλυθούν:

1. οι τεχνικές προεπεξεργασίας των δεδομένων που ακολουθήσαμε
2. οι αλγόριθμοι μάθησης που χρησιμοποιήσαμε
3. τα συμπεράσματα που βγάλαμε

Σημείωση: Ο κώδικας που γράψαμε για όλες τις παραπάνω διαδικασίες υπάρχει στο αρχείο ... που υπάρχει στο συμπιεσμένο αρχείο που στείλαμε.

1 Προεπεξεργασία Δεδομένων

Για την προεπεξεργασία δεδομένων ακολουθήσαμε την εξής διαδικασία. Αρχικά, ξεκινήσαμε με το να καθαρίσουμε όλα τα κείμενα από άχρηστες πληροφορίες, όπως το email στην αρχή του κάθε άρθρου, τα σημεία στίξης, χαρακτήρες newline και whitespaces και κανονικοποιήσαμε όλα τα γράμματα σε μικρά.

Στην συνέχεια, κάναμε tokenize όλες τις υπόλοιπες λέξεις, κάναμε lematize πάνω στο αποτέλεσμα (μιάς και στην συγκεκριμένη εφαρμογή έχει μεγάλη σημασία το context στο οποίο βρίσκονται οι λέξεις). Στην παραπάνω διαδικασία δοκιμάσαμε να βγάλουμε τις πιο συχνά και τις λιγότερο συχνά εμφανιζόμενες λέξεις ώστε να πετύχουμε καλύτερο αποτέλεσμα στα βάρη που προκύπτουν κατά την διαδικασία του TF-IDF.

Τέλος δοκιμάσαμε την τεχνική Data Augmentation, η οποία αναλύεται παρακάτω.

1.1 Data Augmentation

Το Data Augmentation είναι μία τεχνική η οποία μας επιτρέπει να δίνουμε στους classifiers μεγαλύτερη ποικιλία δεδομένων με στόχο να αυξήσουμε την αποτελεσματικότητα τους. Στην περίπτωσή μας αλλάζαμε με τυχαίο τρόπο λέξεις με συνώυμα, δημιουργώντας έτσι καινούρια δείγματα. Για να το πετύχουμε αυτό, χρησιμποποιήσαμε την συνάρτηση wordnet.synsets.

2 Αλγόριθμοι Μάθησης

Οι αλγόριθμοι μάθησης που χρησιμοποιήσαμε είναι οι εξής:

1. Logistic Regression
2. Stohastic Gradient Decent
3. Multinomial Naive Bayes

Με βάση την παραπάνω προεπεξεργασία, τα δεδομένα αξιολόγησης του κάθε αλγορίθμου έχουν ως εξής:

Πίνακας 1: Δεδομένα Αξιολόγησης

Αλγόριθμος	Accuracy	F1-Score
Logistic Regression	84%	84%
Stohastic Gradient Decent	86%	85%
Multinomial Naive Bayes	84%	84%

Τα Confusion Matrix που παρήγαγε ο κάθε αλγόριθμος δίνεται στο αρχείο του κώδικα.

3 Συμπεράσματα

Αποφασίσαμε να κρατήσουμε τους αριθμούς μέσα στο κείμενο, διότι οι αριθμοί εμφάνιζαν μεγάλη συχνότητα σε συγκεκριμένες κατηγορίες, κάτι που έδινε μεγαλύτερη έμφαση στο context κατά την διαδικασία του preprocessing. Επίσης δεν αφαιρέσαμε από τα κείμενα το θέμα, καθώς παρατηρήσαμε ότι είναι πολύ σημαντικό για την κατηγοριοποίηση του κειμένου.