AI 2 Project 1

Βορλόου Φώτιος

1115201900026

Νοέμβριος 2023

Αρχικά φορτώνουμε το csv αρχείο σε μια μεταβλητή η οποία θα κρατάει το data frame.

Preprocessing

Για την αφαίρεση άχρηστων πληροφοριών από τη στήλη των tweets, εφάρμοσα τις συναρτήσεις lemmatization, text_preprocessing, remove_stopwords. Έτσι γίνεται λημματοποίηση με ελληνικούς χαρακτήρες, αφαιρούνται urls tags και λοιποί χαρακτήρες που δυσχεραίνουν το μοντέλο ενώ αφαιρούνται και κοινές λέξεις που απαντώνται πολύ συνά και είναι ουδέτερες, όπως και,, το,, η, κλπ.

,

Vectorizer-Training

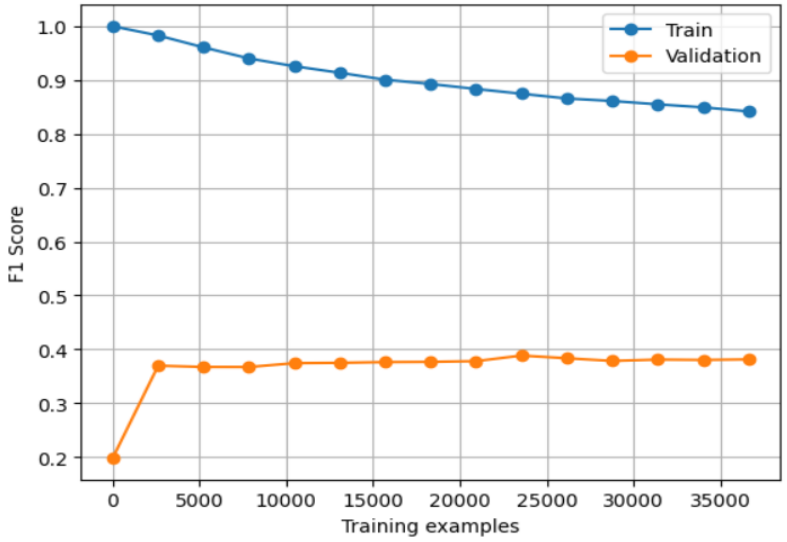
Για το vectorization, χρησιμοποίησα τον TfidfVectorizer για τη μετατροπή των tweets σε bag-of-words αναπαράσταση. Στη συνέχεια, εκπαίδευσα το μοντέλο μου χρησιμοποιώντας την LogisticRegression με κατάλληλα επιλεγμένες παραμέτρους.

Learning Curve

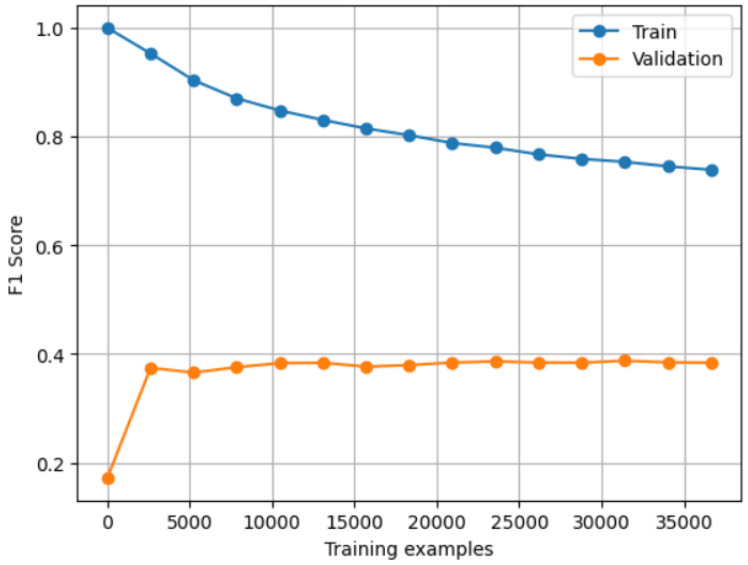
Για να σχεδιάσουμε το learning curve εκπαιδεύουμε το μοντέλο μας πολλές φορές με διαφορετικά ¨τμήματα' του training set . Παρατηρώ ότι το μοντέλο κάνει overfit καθώς τα πηγαίνει καλύτερα για τα στοιχεία του training data από ότι του validation data .

Δοκιμές

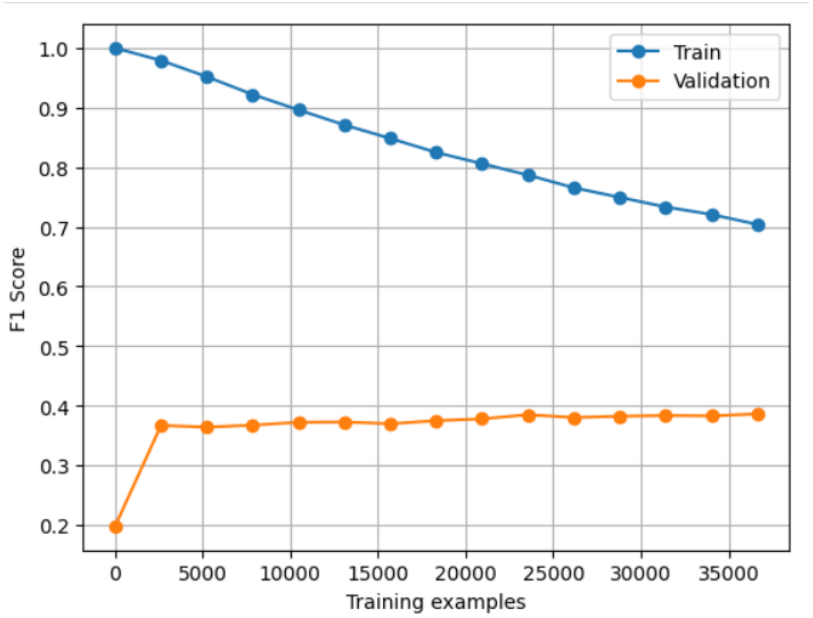
Δοκίμασα 4 μοντέλα μέχρι να καταλήξω σε αυτό με τα καλύτερα αποτελέσματα.
Αρχικά χρησιμοποίησα CountVectorizer χωρίς όρια και το αποτέλεσμα ήταν το παρακάτω:



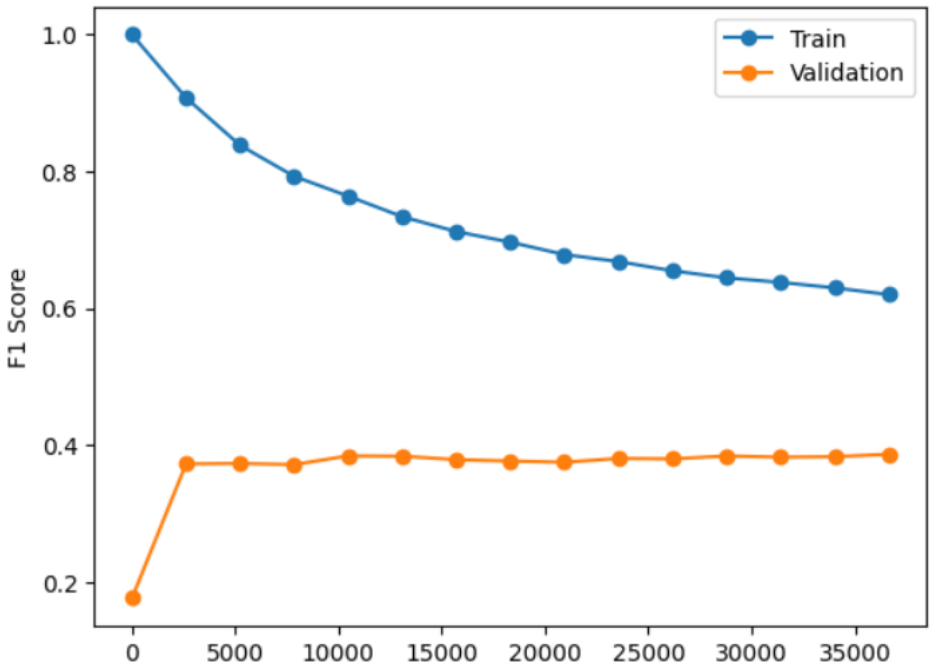
Ξανά θα μείωσα τις διαστάσεις των vector tweet λαμβάνοντας υπόψη μόνο λέξεις που εμφανίζονται σε τουλάχιστον 4 έγγραφα και το πολύ στο 30% όλων των εγγράφων και είχα:



Σε αυτή τη φάση κατάλαβα ότι το μοντέλο κάνει overfiting, διότι το train score ήταν πολύ πιο πάνω από το validation. Οπότε, δοκίμασα τον TfidfVectorizer, αφού ο συγκεκριμένος δίνει μεγαλύτερη σημασία στις σπανιότερες λέξεις, άρα τις πιο σημαντικές και χωρίς όρια είχα το παρακάτω αποτέλεσμα:



Τέλος, εφάρμοσα τα όρια που δοκίμασα στο δεύτερο μοντέλο και διαμόρφωσα το τελικό αποτέλεσμα:



```
F1 Score Train: 1.0
F1 Score Validation: 0.17183081477199122
F1 Score Train: 0.9130195613758707
F1 Score Validation: 0.3722925993151301
F1 Score Train: 0.8387688529183678
F1 Score Validation: 0.37031791580189855
F1 Score Train: 0.7966368306845252
F1 Score Validation: 0.3712753778603954
F1 Score Train: 0.7657265579554764
F1 Score Validation: 0.3814937513172472
F1 Score Train: 0.740740270344369
F1 Score Validation: 0.38007813415629427
F1 Score Train: 0.7136102800114045
F1 Score Validation: 0.3812901795610729
F1 Score Train: 0.7016940738464472
F1 Score Validation: 0.37801437691753303
F1 Score Train: 0.6841160821428687
F1 Score Validation: 0.381773910527861
F1 Score Train: 0.6737782430427149
F1 Score Validation: 0.3873585328629997
F1 Score Train: 0.6573927309995076
F1 Score Validation: 0.3890629673198606
F1 Score Train: 0.646570188952105
F1 Score Validation: 0.3885295169844977
F1 Score Train: 0.6386034777075486
F1 Score Validation: 0.38264173761993486
F1 Score Train: 0.6295373735417881
F1 Score Validation: 0.3847474747474748
F1 Score Train: 0.6235766862371235
F1 Score Validation: 0.3858836149811941
```