Ανάπτυξη Αναδρομικού Νευρωνικού Δικτύου για Πρόβλεψη Συναισθημάτων στο Twitter

Ονοματεπώνυμο : Ευθύμιος Πατέλης

Αριθμός Μητρώου : 1115201300141

Προ-επεξεργασία Δεδομένων

Για τον καθαρισμό των δεδομένων πριν την εκπαίδευση του μοντέλου υλοποιήθηκαν οι παρακάτω τεχνικές :

1.) Tokenization
2.) Αφαίρεση αναφορών χρηστών (user mentions)
3.) Αφαίρεση συνδέσμων (URLs)
4.) Αφαίρεση τόνων
5.) Αφαίρεση ειδικών χαρακτήρων (τελειών, κομμάτων, κ.λπ.)
6.) Αφαίρεση προ-επιλεγμένων stopwords
7.) Μετατροπή των λέξεων σε uppercase
8.) **Custom** Lemmatization
9.) Αφαίρεση spaces εντός ή στα άκρα ενός token
10.) Αφαίρεση λέξεων μικρού μήκους

Το data cleaning έγινε με σκοπό την αφαίρεση στοιχείων που πιθανώς να μην φέρουν ιδιαίτερο συναισθηματικό νόημα, καθώς και για τη μεγαλύτερη δυνατή μείωση διάστασης του dataset ως προς το μέγεθος του vocabulary. Για τα stopwords χρησιμοποιήθηκε ένα custom set, το οποίο προέκυψε από την εύρεση κοινά χρησιμοποιουμένων λέξεων εντός του training set. Στη φάση του lemmatization, αφαιρέθηκαν τα γράμματα ‘Σ’ και ‘Ν’ από τις λέξεις με τέτοια κατάληξη, για περεταίρω pruning του vocabulary. Το παραπάνω data cleaning είναι όμοιο με των δυο πρώτων ασκήσεων, με σκοπό την όσο το δυνατόν **δικαιότερη** σύγκριση των αποτελεσμάτων του τρέχοντος και των προηγουμένων μοντέλων ως προς τις μετρικές απόδοσης (precision, recall και f-1).

Σχεδιασμός Μοντέλου

Το νευρωνικό δίκτυο αποτελείται από δυο κύρια μέρη. Το πρώτο μέρος, το οποίο αναλαμβάνει τα input data, είναι ένα αναδρομικό νευρωνικό δίκτυο (**RNN**) του οποίου η δομή καθορίζεται από πολλαπλές υπερπαραμέτρους, οι οποίες είναι ο συνολικός αριθμός των layers, το dimensionality του hidden state, ο τύπος του δικτύου (**LSTM** ή **GRU**), το aggregation method (**MEAN**, **LAST** ή **MAX**), καθώς και ένα dropout probability, το οποίο εισάχθηκε **μόνο** στο RNN. Το δεύτερο μέρος είναι ένα γραμμικό μοντέλο το οποίο δέχεται ως input το output του RNN και είναι υπεύθυνο για την υλοποίηση των προβλέψεων μεταξύ των τριών κλάσεων.

Αποτελέσματα Πειραμάτων

Παρακάτω ακολουθούν τα scores του μοντέλου ως προς τις μετρικές Precision, Recall, F-1, και Accuracy. Το καλύτερο μοντέλο προέκυψε ως συνδυασμός των μετρικών αυτών **και** δεδομένων των learning curves (οι οποίες δεν παρουσιάζονται για κάθε πείραμα, για εξοικονόμηση χώρου) έτσι ώστε το μοντέλο να εμφανίζει όσο το λιγότερο δυνατό underfit ή overfit.

Hidden Dimension	Total Layers	RNN Type	Aggregation Method	Skip Connections	Batch Size	Gradient Clip	Dropout Probability	Precision	Recall	F-1
2	1	GRU	LAST	TRUE	8	1	0	0.340	0.340	0.340
4	1	GRU	LAST	TRUE	16	5	0	0.340	0.340	0.399
8	2	GRU	LAST	FALSE	32	10	0	0.351	0.351	0.350
16	2	GRU	MEAN	TRUE	64	1	0.2	0.349	0.349	0.348
2	4	GRU	MEAN	TRUE	128	5	0.2	0.345	0.345	0.345
4	4	GRU	MEAN	FALSE	128	10	0.2	0.378	0.374	0.372
16	1	GRU	MAX	FALSE	128	1	0	0.385	0.386	0.385
16	**2**	**GRU**	**MAX**	**FALSE**	**128**	**1**	**0**	**0.403**	**0.398**	**0.387**
2	1	LSTM	LAST	TRUE	8	1	0	0.356	0.356	0.356
4	1	LSTM	LAST	TRUE	16	5	0	0.370	0.373	0.368
8	2	LSTM	LAST	FALSE	32	10	0	0.389	0.386	0.385
16	2	LSTM	MEAN	TRUE	64	1	0.2	0.370	0.371	0.370
32	4	LSTM	MEAN	FALSE	128	5	0.2	0.381	0.383	0.380
64	4	LSTM	MAX	TRUE	256	1	0.2	0.378	0.377	0.376
128	8	LSTM	MAX	FALSE	512	5	0.4	0.387	0.387	0.386

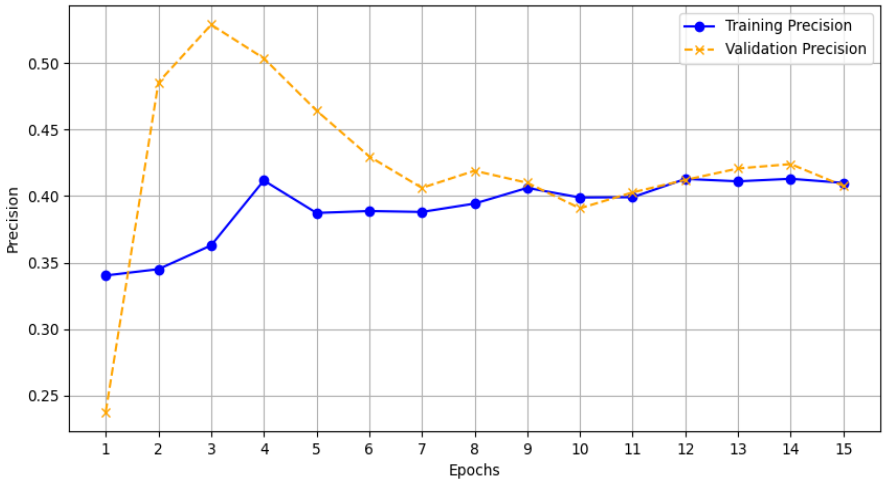
Fine Tuning

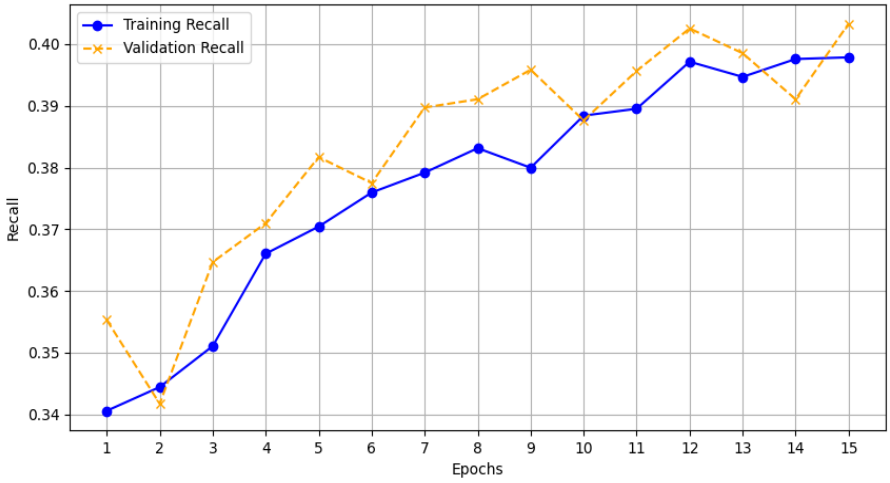
Για τη βελτιστοποίηση του καλύτερου, **βασικού** μοντέλου (το οποίο φαίνεται με bold στον παραπάνω πίνακα), χρησιμοποιήθηκε ο scheduler, με τη λογική ότι η προοδευτική μείωση του learning rate βοηθά το μοντέλο να εντοπίσει το τοπικό ελάχιστο της συνάρτησης απώλειας, αποτρέποντας το από το να "αναπηδά" στις υψηλότερες τιμές γύρω από αυτό το ελάχιστο, όπως και με όμοιο σκεπτικό για τις μετρικές απόδοσης. Παρακάτω, παρουσιάζονται τα αποτελέσματα :

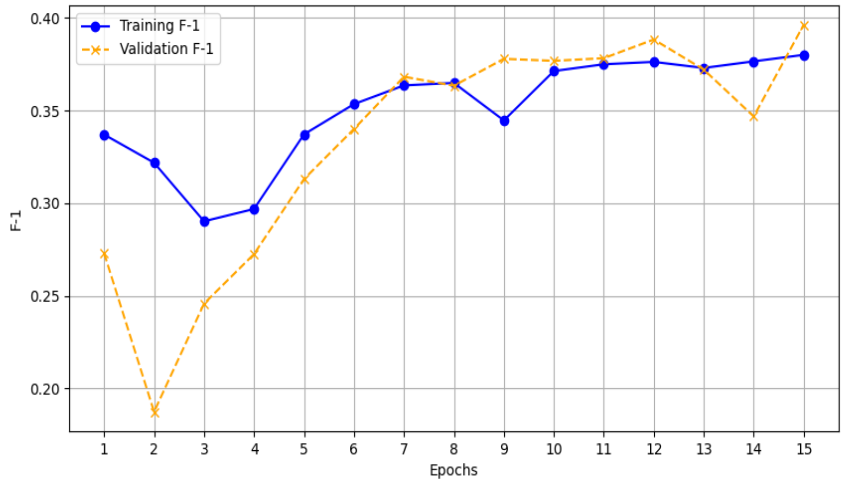
Scheduler	Gamma	Step	Precision	Recall	F-1
ExponentialLR	0.8	-	0.403	0.398	0.387
StepLR	0.4	5	0.406	0.402	0.394
StepLR	0.4	10	0.405	0.401	0.393
StepLR	**0.8**	**5**	**0.407**	**0.403**	**0.396**
StepLR	0.8	10	0.401	0.398	0.392

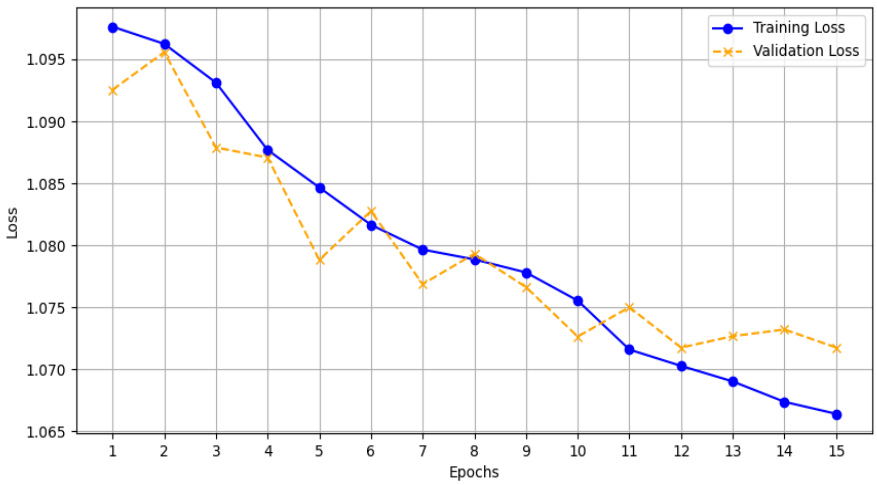
Learning Curves

Παρακάτω παρατίθενται τα learning curves του καλύτερου μοντέλου που προέκυψε από το fine – tuning, στις μετρικές precision, recall, f-1 και loss, καθώς και οι τιμές του καλύτερου instance του καλύτερου μοντέλου ως προς αυτές:


Training Precision
Validation Precision
Precision
Epochs
0.25
0.30
0.35
0.40
0.45
0.50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

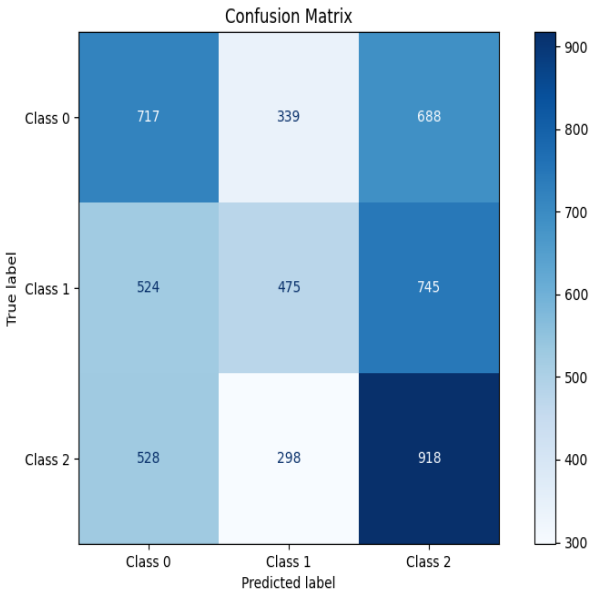

Training Recall
Validation Recall
Recall
Epochs
0.34
0.35
0.36
0.37
0.38
0.39
0.40
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15





Το μοντέλο φαίνεται να μαθαίνει από τα δεδομένα καθώς και να γενικεύει, αφού οι καμπύλες των μετρικών για το training και validation set αυξάνουν, ενώ το loss μειώνεται. Επιπλέον, το μοντέλο φαίνεται να μην κάνει overfit ή underfit, αφού οι τελικές τιμές των μετρικών αξιολόγησης αλλά και του σφάλματος κυμαίνονται σε περίπου ιδιά επίπεδα, και για τα δυο sets.

Confusion Matrix



Από το confusion matrix παρατηρούμε ότι :

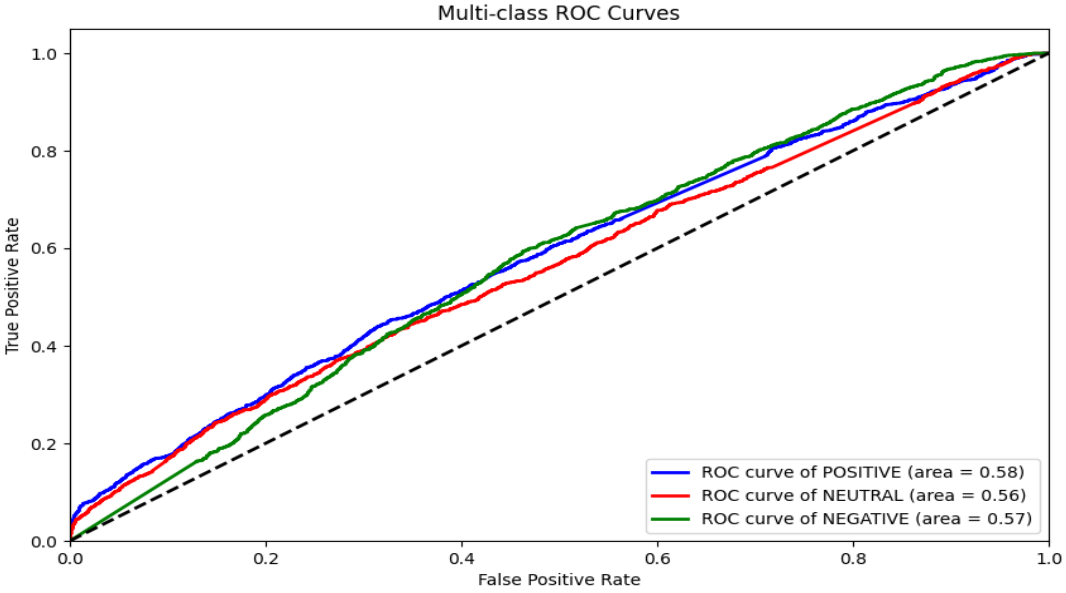
1.) Για το class 0 (Positive) , τα precision και recall είναι σχετικά ισορροπημένα (40.5% και 41.1% αντίστοιχα).

2.) Για το class 1 (Neutral), παρά το υψηλό precision (42.7%), το recall είναι χαμηλό (27.2%). Αυτό σημαίνει ότι αν και το μοντέλο είναι αρκετά ακριβές στο να **μην** κατατάσσει ένα non-neutral instance ως neutral, ταυτόχρονα κατατάσσει επίσης πολλά neutral instances ως non-neutral.

3.) Για το class 2 (Negative) , εμφανίζεται ο υψηλότερος δείκτης recall (52.6%), ωστόσο, παρατηρείται χαμηλό precision (39%). Αυτό σημαίνει ότι παρόλο που το μοντέλο εντοπίζει σχετικά αποτελεσματικά τις περιπτώσεις που ανήκουν πραγματικά στο class 2, ταυτόχρονα τείνει να κατατάσσει λανθασμένα και άλλες περιπτώσεις σε αυτό το class.

ROC curves

Αφού έχουμε τρία classes, χρησιμοποιήθηκαν τρία διαφορετικά ROC curves, με την τεχνική One vs Rest (OvR). Σε αυτήν, κάθε φορά ένα label ορίζεται ως το θετικό class και τα υπόλοιπα δύο συνθέτουν το αρνητικό class. Οι καμπύλες παρουσιάζονται παρακάτω:



Παρατηρείται ότι οι τιμές AUC του μοντέλου, οι οποίες είναι 0.58, 0.56 και 0.57 για τις κλάσεις 0, 1 και 2 αντιστοίχως, καθώς και η θέση των καμπυλών ROC πάνω από την κεντρική διαγώνιο, δείχνουν ότι το μοντέλο έχει καλύτερη απόδοση από ένα τυχαίο μοντέλο στην ταξινόμηση των κλάσεων. Σε περίπτωση που το μοντέλο είχε τυχαία απόδοση, τότε οι καμπύλες ROC θα συνέπιπταν με την κεντρική διαγώνιο, ανεξάρτητα από την τιμή threshold. Το μοντέλο διακρίνει με μεγαλύτερη ευκρίνεια τα δείγματα της κλάσης 2 (Negative), ακολουθούμενα από τις κλάσεις 0 (Positive) και 1 (Neutral), εμφανίζοντας μια σχετική ομοιομορφία στην απόδοση ανάμεσα στις κλάσεις.

Σχολιασμός Τελικού Μοντέλου

Συγκρίνοντας το τρέχον μοντέλο με αυτά των προηγουμένων δύο ασκήσεων, παρατηρούμε ότι :

1.) Στην πρώτη άσκηση, τα evaluation metrics μειώνονταν κατά τη την εκπαίδευση, δείχνοντας έλλειψη εκμάθησης. Αντιθέτως, στη δεύτερη αλλά και την τρέχουσα άσκηση, οι μετρικές βελτιώνονταν, υποδηλώνοντας αποτελεσματική (σχετικά) εκμάθηση και γενίκευση σε άγνωστα δεδομένα.

2.) Στο πρώτο μοντέλο, η σημαντική απόκλιση ανάμεσα στις μετρικές πάνω στο training και validation set υποδείκνυαν υπερπροσαρμογή, αφού οι τιμές του training set ήταν αρκετά μεγαλύτερες αυτών του validation set. Αντίθετα, στο δεύτερο αλλά και τρέχον μοντέλο, οι τιμές κυμαίνονται περίπου στο ίδιο επίπεδο.

3.) Στο αρχικό μοντέλο, τα ROC curves για τα positive και negative sentiments, ήταν υπό της κύριας διαγωνίου, το οποίο υποδείκνυε ότι το πρώτο μοντέλο απέδιδε χειρότερα από το αν τα sentiments επιλέγονταν τυχαία. Στη δεύτερη αλλά και τρίτη άσκηση, τα οι καμπύλες βρίσκονται από πάνω, υποδεικνύοντας καλύτερη διάκριση των κλάσεων σε σχέση με το αρχικό.

4.) Με τη χρήση, είτε ενός γραμμικού, είτε ενός αναδρομικού, Νευρωνικού δικτύου, η απόδοση του μοντέλου αυξήθηκε κατά περίπου 6%.

5.) Τέλος, η μετάβαση από ένα απλό γραμμικό σε ένα πιο σύνθετο, αναδρομικό, γραμμικό νευρωνικό δίκτυο δεν απέφερε ουσιώδη βελτίωση, αφού και των δυο η απόδοση κυμαίνεται περίπου στα ιδία επίπεδα (δηλαδή περίπου 40% ακρίβεια), το οποίο πολύ πιθανώς να σχετίζεται με το provided dataset.