Τεχνητή Νοημοσύνη 2

Εργασία 3

Δημητρακόπουλος Κωνσταντίνος ΑΜ: 1115201500034

Περιεχόμενα

Προεπεξεργασία των Δεδομένων

Τα δεδομένα δέχονται **προεπεξεργασία** για την βέλτιστη αντιστοίχιση κατά τη διανυσματοποίηση τους.

Δοκιμάστηκαν οι παρακάτω μέθοδοι:

1. Αφαίρεση συνδέσμων.
2. Αφαίρεση hashtags.
3. Αφαίρεση mentions.
4. Αντικατάσταση συνδέσμων σε <link>.
5. Αντικατάσταση hashtags σε <hashtag>.
6. Αντικατάσταση mentions σε <mention>.
7. Αντικατάσταση αριθμών σε <number>.
8. Αντικατάσταση κεφαλαίων χαρακτήρων σε <upper>.
9. Διατήρηση μόνο των αλφαριθμητικών χαρακτήρων.
10. Μετατροπή σε πεζά γράμματα.
11. Lemmatizing.
12. Αφαίρεση stop words.

Για την **αντιστοίχιση tokens-embeddings** πραγματοποιήθηκαν πειράματα με τα pre-trained embeddings **glove.twitter.26B** και **glove.6B**. Οι περιπτώσεις 4 έως 8 συνεισφέρουν στη διανυσματοποίηση των οντοτήτων όπως περιγράφονται εντός των glove twitter pretrained embeddings.

Δοκιμές με το σετ GloVe 6B

Με χρήση της βοηθητικής συνάρτησης **token_statistics** έγινε εξαγωγή των χαρακτηριστικών που παρουσιάζει ο παρακάτω πίνακας για τους διαφορετικούς **συνδυασμούς προεπεξεργασίας** των δεδομένων. Συγκεκριμένα αναγράφονται το συνολικό πλήθος των tokens που παρήγαγε ο tokenizer, το συνολικό πλήθος των tokens που βρέθηκαν στο λεξικό των pretrained embeddings, αντίστοιχα το συνολικό πλήθος των tokens που δεν εντοπίστηκαν, το ποσοστό κάλυψης των συνολικών tokens από το λεξικό και σχόλια σχετικά με τη κάθε δοκιμή.

No	Modules	Total	Found	Not Found	Percentage Coverage	Comments
Π1	• remove_non_alpha • to_lowercase • lemmatize • remove_stop_words	210226	186094	24132	0.88	• Αγνοεί ενωμένα, ανορθόγραφα και δυσνόητα tokens. • “timehealth”, ”mncqnpdqgc”, “aaaaaa”
Π2	• to_lowercase • lemmatize • remove_stop_words	275535	245948	29587	0.89	• Αγνοεί emoji, links, ενωμένα tokens και αριθμούς.
Π3	• remove_links • remove_hashtags • remove_mentions • to_lowercase • lemmatize • remove_stop_words	214250	204721	9529	0.95	• Αγνοεί tokens που χρησιμοποιούν emoji, παύλες ή είναι ανορθόγραφα.
Π4	• replace_links • replace_hashtags • replace_mentions • replace_numbers • replace_upper_words • to_lowercase • lemmatize • remove_stop_words	250272	198291	51981	0.79	• Αγνοεί το μετασχηματισμό των ειδικών tokens όπως αναμενόταν.

Παρατηρήσεις:

- Φαίνεται πως η δοκιμή Π3 προσφέρει την καλύτερη κάλυψη από το λεξικό και ίσως αποτελεί την καλύτερη αναπαράσταση του κειμένου, όμως η δοκιμή Π2 προσφέρει περισσότερες αντιστοιχήσεις σε tokens συνεπώς μία μεγαλύτερη αναπαράσταση από την Π3.
- Η δοκιμή Π4 όπως αναμενόταν αγνοεί τα ειδικά tokens που σχηματίστηκαν για τα twitter pretrained embeddings, συνεπώς είναι πιθανό να μην αποτελεί καλή αναπαράσταση.
- Όλες οι δοκιμές αδυνατούν να αναγνωρίσουν ενωμένα, ανορθόγραφα και δυσνόητα tokens

Δοκιμές στο σετ GloVe Twitter 26B

Αντίστοιχα, εφαρμόζονται οι ίδιες δοκιμές για το σετ glove.twitter.26B. Παρουσιάζεται ο πίνακας των αποτελεσμάτων.

No	Modules	Total	Found	Not Found	Percentage Coverage	Comments
Π5	• remove_non_alpha • to_lowercase • lemmatize • remove_stop_words	210226	188028	22198	0.89	• Αγνοεί ενωμένα, ανορθόγραφα και δυσνόητα tokens.
Π6	• to_lowercase • lemmatize • remove_stop_words	275535	241178	34357	0.87	• Αγνοεί emoji, links, ενωμένα tokens και αριθμούς.
Π7	• remove_links • remove_hashtags • remove_mentions • to_lowercase • lemmatize • remove_stop_words	214250	198881	15369	0.92	• Αγνοεί tokens που χρησιμοποιούν emoji, παύλες ή είναι ανορθόγραφα.
Π8	• replace_links • replace_hashtags • replace_mentions • replace_numbers • replace_upper_words • to_lowercase • lemmatize • remove_stop_words	250272	239805	10467	0.95	• Αγνοεί tokens που χρησιμοποιούν emoji, παύλες ή είναι ανορθόγραφα.

Παρατηρήσεις:

- Ο συνολικός αριθμός των tokens είναι κοινός με των προηγούμενων δοκιμών, ωστόσο υπάρχουν διαφοροποιήσεις ως προς την κάλυψη από το λεξικό.
- Όπως στις προηγούμενες δοκιμές, τα πειράματα αδυνατούν να αναγνωρίσουν ενωμένα, ανορθόγραφα και δυσνόητα tokens.
- Η δοκιμή Π8 αναγνωρίζει τα ειδικά tokens προσφέροντας καλή κάλυψη από το λεξικό.

Δίκτυα LSTM Πολλαπλών Επιπέδων

Για τη συγκεκριμένη ενότητα πραγματοποιήθηκαν δοκιμές σε δίκτυα **bidirectional LSTM** πολλαπλών επιπέδων. Ο παρακάτω πίνακας συνοψίζει τις δοκιμές αυτές και παρουσιάζει το **σχήμα του δικτύου (scheme)**, την τιμή της **gradient clipping norm (CG)** αν αυτή εφαρμόστηκε, την τιμή της **πιθανότητας απόρριψης Dropout (DO)** πριν κάθε layer αν αυτό εφαρμόστηκε, το **σφάλμα**, το **f1 score**, την **εποχή** καθώς και **σχόλια που αφορούν την εκπαίδευση**.

Περαιτέρω Επεξήγηση Σχήματος:

Στο σχήμα περιγράφονται ο αριθμός των layers, το μέγεθος των output χαρακτηριστικών από κάθε layer και η bidirectional ιδιότητα κάθε layer.

Για παράδειγμα, η **δοκιμή 7** παρουσιάζει ένα δίκτυο όπου το πρώτο layer είναι bidirectional lstm layer με μέγεθος χαρακτηριστικών εξόδου 80 και το δεύτερο layer είναι lstm layer με μέγεθος χαρακτηριστικών εξόδου 3.

Η **δοκιμή 3** είναι παραλλαγή της δοκιμής 2 όπου προστέθηκε batch normalization.

No	Scheme	CG	DO	Loss (Train/Val)	Score (Train/Val)	Ep	Comments
1	LSTM (3)	-	-	0.795867/ 0.844803	0.636654/ 0.585740	20	• Καλό μοντέλο • Εμφάνιση overfitting
2	LSTM (100) LSTM (50) LSTM (3)	-	0.2	0.972222/ 0.97556	0.324968/ 0.33421	35	• Εμφάνιση Vanishing gradients
3	2 -> + Batch Norm	-	0.2	0.965144/ 0.96573	0.468938/ 0.47350	8	• Χαμηλό f1 • Τείνει σε σύγκλιση • Περισσότερες εποχές
4	LSTM (75) LSTM (3)	-	0.2	0.941710/ 0.94435	0.509177/ 0.51350	15	• Καλή σύγκλιση σφάλματος • Απότομες εναλλαγές f1
5	LSTM (3)	-	0.2	0.824040/ 0.83875	0.611877/ 0.60518	49	• Καλό μοντέλο
6	LSTM (100, bid) LSTM (50, bid) LSTM (3)	-	0.2	0.895485/ 0.89606	0.558384/ 0.54903	8	• Μόνιμη αύξηση σφάλματος • Μόνιμη μείωση f1
7	LSTM (80, bid) LSTM (3)	-	0.2	0.843929/ 0.85413	0.611622/ 0.60835	42	• Καλή σύγκλιση • Καλό σφάλμα και f1 • Περισσότερες εποχές
8	7 -> + epochs	-	0.2	0.843089/ 0.85619	0.617186/ 0.60303	19	• Αμετάβλητο • Τιμές των Gradients έως 6-7
9	7 -> + Gradient Clip	2	0.2	0.80189/ 0.8362	0.65215/ 0.6229	48	• Μέση Βελτίωση f1 Κατά 1-2%

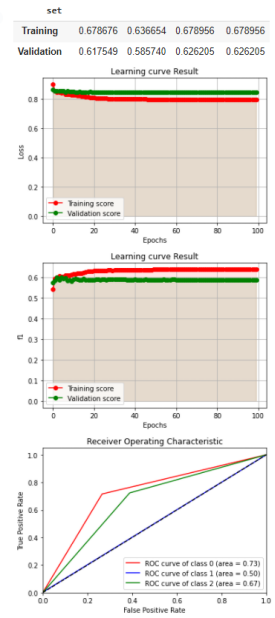
Κατά τη **δοκιμή 1** τα αποτελέσματα είναι ικανοποιητικά εμφανίζοντας ελάχιστο **overfitting**. Γίνεται εισαγωγή **Dropout** με τιμή απόρριψης 0.2 (ακολουθήθηκε η προτεινόμενη τιμή 0.2~0.3 για αναδρομικά δίκτυα σύμφωνα με τη θεωρία) για την αντιμετώπιση του φαινομένου και αύξηση του τελικού f1 σκορ.

Κατά τη **δοκιμή 2** εμφανίστηκε το φαινόμενο **vanishing gradients** καθώς το μοντέλο δεν εκπαιδευόταν με την πάροδο των εποχών και αντιμετωπίστηκε με την εισαγωγή **batch normalization** πριν τα κατάλληλα layers μειώνοντας όμως την απόδοση των μετρικών.

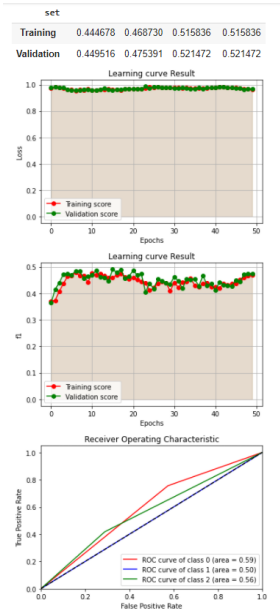
Ικανοποιητικά μοντέλα φαίνονται αυτά των **δοκιμών 5 και 7**, ενώ παρατηρούνται έντονες τιμές gradients στη δοκιμή 8. Με την εισαγωγή **gradient clipping** παρουσιάζεται το βέλτιστο μέχρι στιγμής **μοντέλο 9**.

Τα παραγόμενα διαγράμματα

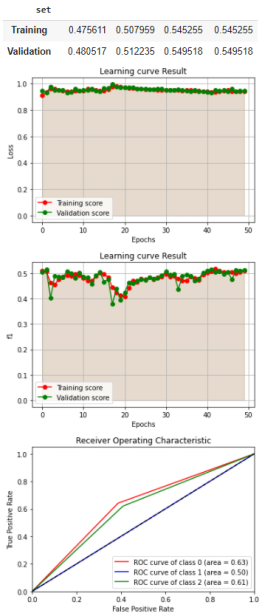
(Μετρικές από αριστερά προς τα δεξιά: Precision-f1-Recall-Accurasy) :



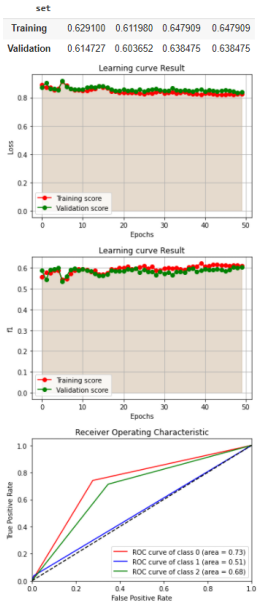
Δοκιμή 1



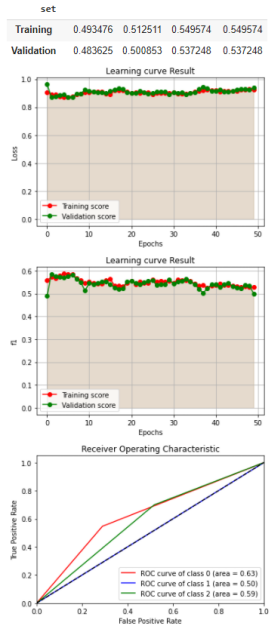
Δοκιμή 3



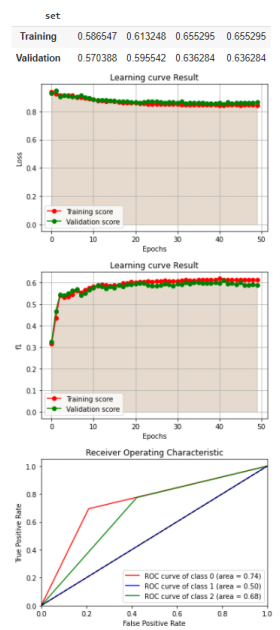
Δοκιμή 4



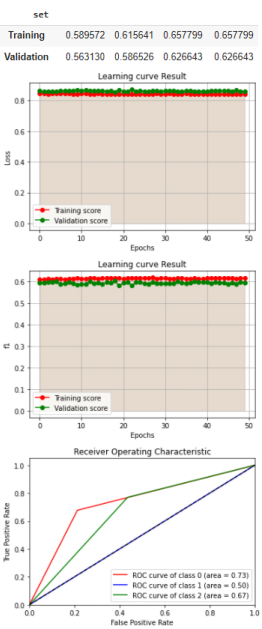
Δοκιμή 5



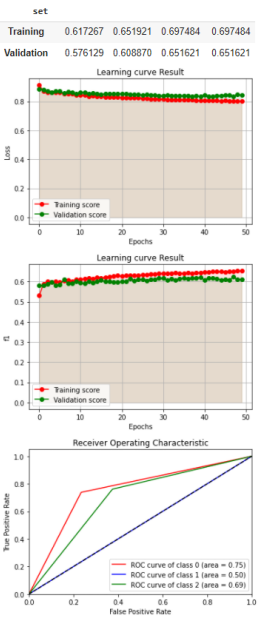
Δοκιμή 6



Δοκιμή 7



Δοκιμή 8



Δοκιμή 9

Δίκτυα GRU Πολλαπλών Επιπέδων

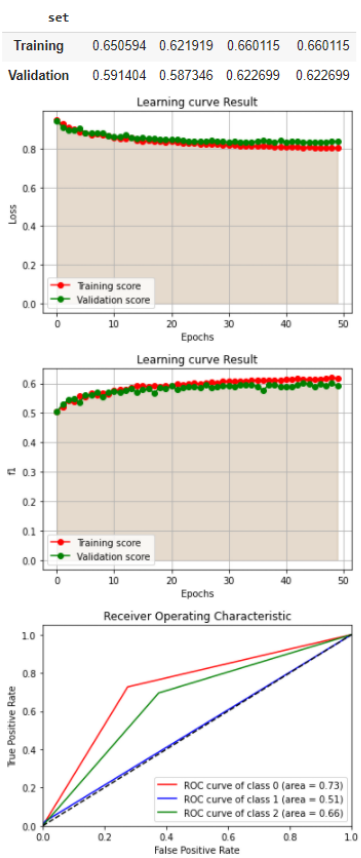
Ακολουθεί η ίδια διαδικασία για δίκτυα GRU πολλαπλών επιπέδων. Τα αποτελέσματα συνοψίζονται στον παρακάτω πίνακα.

No	Scheme	CG	DO	Loss (Train/Val)	Score (Train/Val)	Ep	Comments
10	• GRU (100, bid) • GRU (50, bid) • GRU (3)	2	0.2	0.98223/ 0.9848	0.30288/ 0.3036	8	• Πιθανό φαινόμενο vanishing gradients • Πιθανή λύση layer skipping
11	10 -> + Batch Norm	2	0.2	0.99197894/0.991144	0.3029489/ 0.306185	21	• Όμοια συμπεριφορά
12	• GRU (80, bid) • GRU (3)	2	0.2	0.803736/ 0.838224	0.61974/ 0.60035	49	• Καλό μοντέλο • Χρειάζεται περισσότερες εποχές.
13	12 -> + epochs	2	0.2	0.7990404/ 0.832291	0.6213335/ 0.601295	48	• Αμετάβλητο • Καλό μοντέλο
14	• GRU (3)	2	0.2	0.7756615/ 0.8241831	0.6455074/ 0.6252712	50	• Καλό μοντέλο • Απότομες Εναλλαγές σε σφάλμα και f1 στις εποχές 3-7

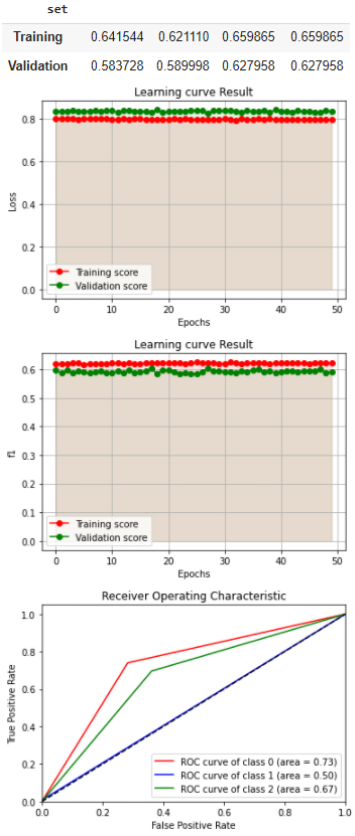
Κατά τη δοκιμή 10 εμφανίζεται το φαινόμενο vanishing gradients το οποίο δεν αντιμετωπίζεται με επιτυχία κατά την δοκιμή 11. Τα μοντέλα 12 και 14 παρουσιάζουν καλή απόδοση σύμφωνα με τις μετρικές και είναι ικανοποιητικά. Τελικά, τα μοντέλα που χρησιμοποιούν GRU layers είναι περισσότερο αποδοτικά σε χρόνο.

Τα παραγόμενα διαγράμματα

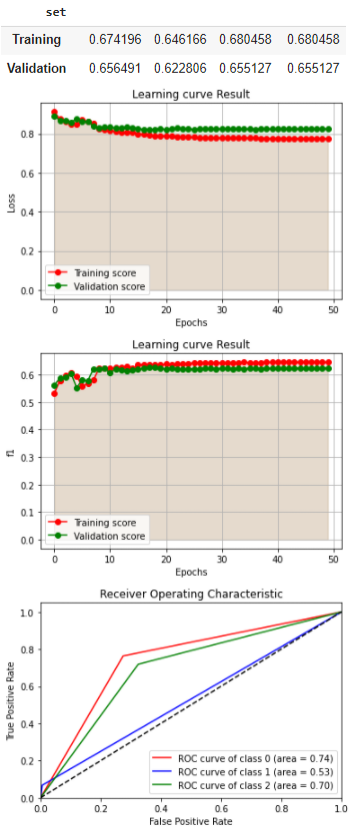
(Μετρικές από αριστερά προς τα δεξιά: Precision-f1-Recall-Accurasy) :



Δοκιμή 12



Δοκιμή 13



Δοκιμή 14

Συνδυασμένα δίκτυα LSTM-GRU Πολλαπλών Επιπέδων

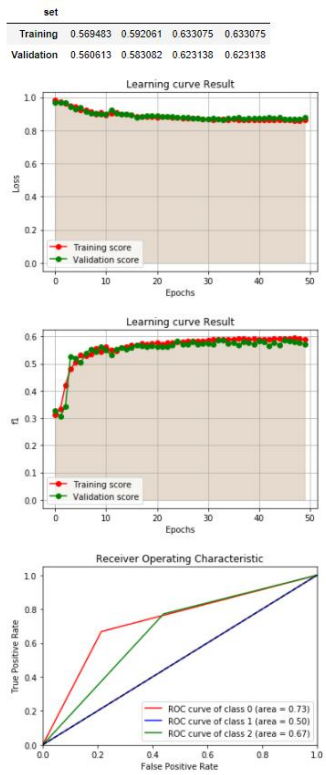
Ακολουθεί η ίδια διαδικασία για συνδυασμένα δίκτυα LSTM-GRU πολλών επιπέδων. Ο πίνακας των αποτελεσμάτων:

No	Scheme	GC	DO	Loss (Train/Val)	Score (Train/Val)	Ep	Comments
15	LSTM (100, bid) GRU (50, bid) GRU (3)	2	0.2	0.9653315/ 0.974742	0.3026444/ 0.3035751	22	• Vanishing Gradients
16	15 -> + Batch Norm	2	0.2	0.972217/ 0.9810409	0.452055/ 0.4383760	36	• Χαμηλό f1
17	LSTM (100, bid) LSTM (50, bid) GRU (3)	2	0.2	0.991864/ 0.990924	0.33748/ 0.35720	19	• Χαμηλό f1 • Ικανοποιητική Εκπαίδευση
18	GRU (100, bid) GRU (50, bid) LSTM (3)	2	0.2	0.86703/ 0.86879	0.5895/ 0.58651	33	• Καλή σύγκλιση • Καλό μοντέλο • Αρχικά έντονες μεταβολές
19	17 -> + Batch Norm	2	0.2		0.34 - 0.42		• Έντονα Μεταβαλλόμενο
20	18 -> + Batch Norm	2	0.2	0.932307/ 0.944228	0.303753/ 0.3073211	13	• Vanishing Gradients
21	GRU (100, bid) LSTM (50, bid) LSTM (3)	2	0.2	0.90307/ 0.9107	0.5530/ 0.55516	33	• Καλή σύγκλιση • Κακό f1 • Αρχικές έντονες μεταβολές
22	LSTM (80, bid) GRU (3)	2	0.2	0.85375/ 0.8818	0.3130/ 0.3145		• Vanishing Gradients
23	22 -> + Batch Norm	2	0.2	0.83938/ 0.8645	0.6103/ 0.5936	46	• Καλό μοντέλο
24	GRU (80, dim) LSTM (3)	2	0.2	0.8278/ 0.84330	0.60052/ 0.5943	43	• Καλή σύγκλιση • Καλό μοντέλο • Αρχικά έντονες μεταβολές

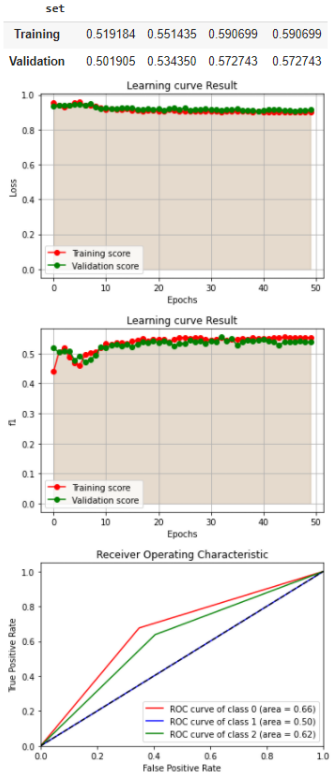
Ο συνδυασμός διαφορετικού είδους στρωμάτων **δεν επιφέρει βελτίωση** στην απόδοση των μετρικών. Φαινόμενα **vanishing gradients** αντιμετωπίστηκαν με χρήση batch normalization μειώνοντας περισσότερο την απόδοση των μετρικών. Τελικά προτιμώνται μοντέλα που δοκιμάστηκαν στις προηγούμενες ενότητες.

Τα παραγόμενα διαγράμματα

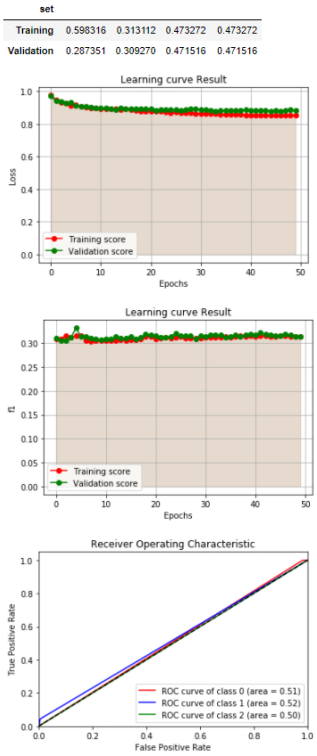
(Μετρικές από αριστερά προς τα δεξιά: Precision-f1-Recall-Accurasy) :



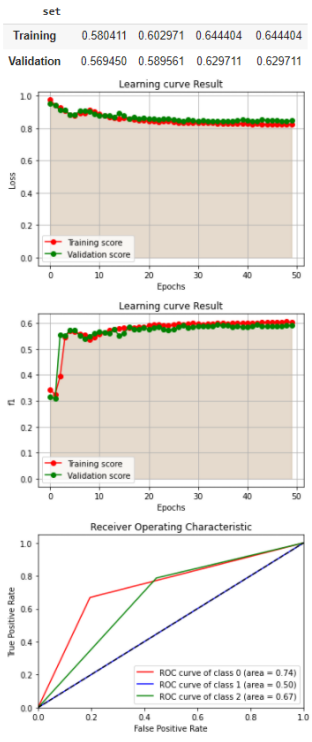
Δοκιμή 18



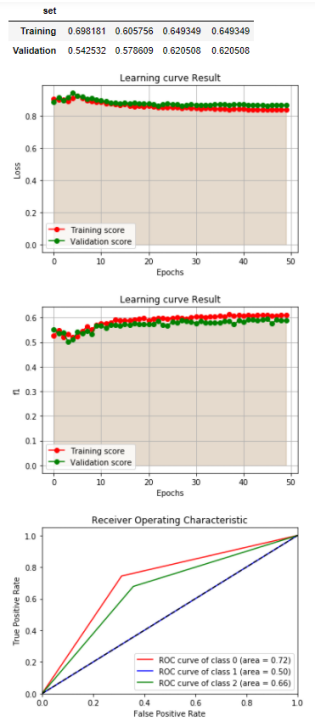
Δοκιμή 21



Δοκιμή 22



Δοκιμή 23



Δοκιμή 24

Μεμονωμένα Δίκτυα LSTM/GRU με χρήση Skip Layers

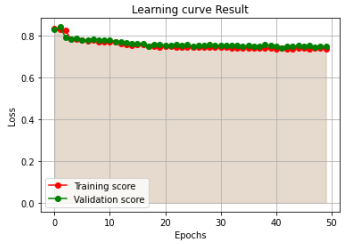
Εφαρμόστηκαν δοκιμές χρησιμοποιώντας skipping layers σε δίκτυα που περιέχουν LSTM ή GRU layers. Στην επόμενη ενότητα παρουσιάζονται πολυπλοκότερα μοντέλα με συνδυασμούς των δύο ειδών.

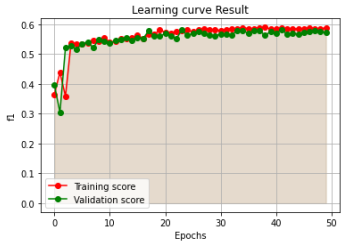
No	Scheme	GC	DO	Loss (Train/Val)	Score (Train/Val)	Ep	Comments
42	a=LSTM (80, bid) b=LSTM (80. bid) Addition (a, b) LSTM (3)	4	0.2	0.7394/ 0.7416	0.58668/ 0.58299	42	• Καλή σύγκλιση • Καλό μοντέλο
43	a=LSTM (80, bid) b=LSTM (80. bid) a=Addition (a, b) b=LSTM (80. bid) Addition (a, b) LSTM (3)	3	0.2	0.8226/ 0.8220	0.4293/ 0.4281	35	• Χαμηλό f1 score και υψηλό σφάλμα
44	a=GRU (80, bid) b=GRU (80. bid) Addition (a, b) GRU (3)	3	0.2	0.75135/ 0.76209	0.56577/ 0.57006	47	• Καλό μοντέλο • Χαμηλό score
45	a= GRU (80, bid) b= GRU (80. bid) a=Addition (a, b) b= GRU (80. bid) Addition (a, b) GRU (3)	3	0.2	0.8080/ 0.8070	0.48211/ 0.49407	48	• Χαμηλό score • Καλό μοντέλο

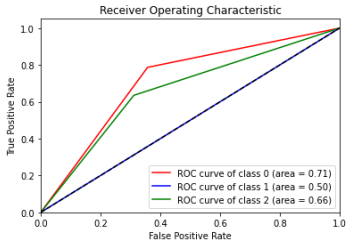
Οι δοκιμές επέφεραν ικανοποιητικά αποτελέσματα όπως φαίνεται από τις δοκιμές 42 και 44. Πολυπλοκότερα δίκτυα όπως των δοκιμών 43 και 45 μειώνουν την απόδοση των μετρικών. Τελικά τα αποτελέσματα των παραπάνω δοκιμών δεν είναι υψηλότερα των παραπάνω ενοτήτων.

Τα παραγόμενα διαγράμματα:

set	precision	f1	recall
Training Class 0	0.680229	0.734452	0.798069
Training Class 1	0.000000	0.000000	0.000000
Training Class 2	0.599087	0.633311	0.671683
Validation Class 0	0.658288	0.716852	0.786854
Validation Class 1	0.000000	0.000000	0.000000
Validation Class 2	0.579782	0.606218	0.635179

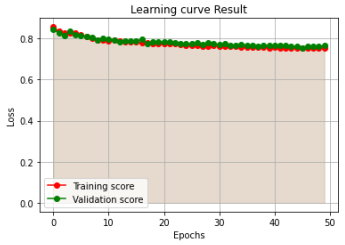


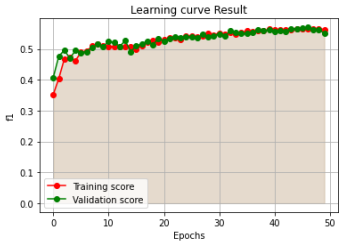


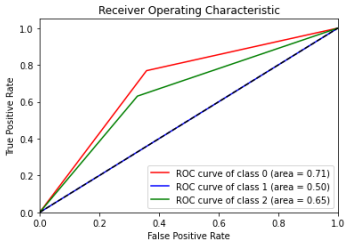


Δοκιμή 42

set	precision	f1	recall
Training Class 0	0.662105	0.715597	0.778493
Training Class 1	0.000000	0.000000	0.000000
Training Class 2	0.585542	0.618224	0.654771
Validation Class 0	0.652590	0.706034	0.769014
Validation Class 1	0.000000	0.000000	0.000000
Validation Class 2	0.565725	0.596509	0.630836

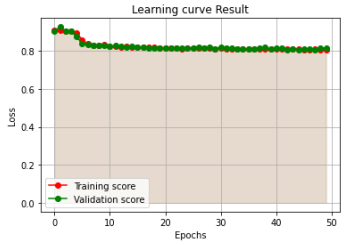


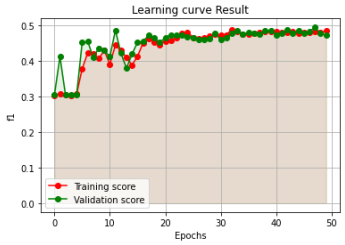


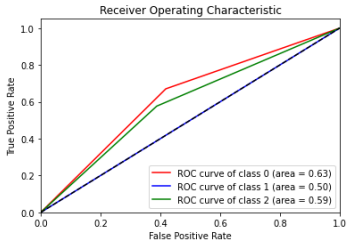


Δοκιμή 43

set	precision	f1	recall
Training Class 0	0.584555	0.628249	0.679002
Training Class 1	0.000000	0.000000	0.000000
Training Class 2	0.507452	0.539468	0.575795
Validation Class 0	0.583810	0.624126	0.670423
Validation Class 1	0.000000	0.000000	0.000000
Validation Class 2	0.501416	0.536364	0.576547





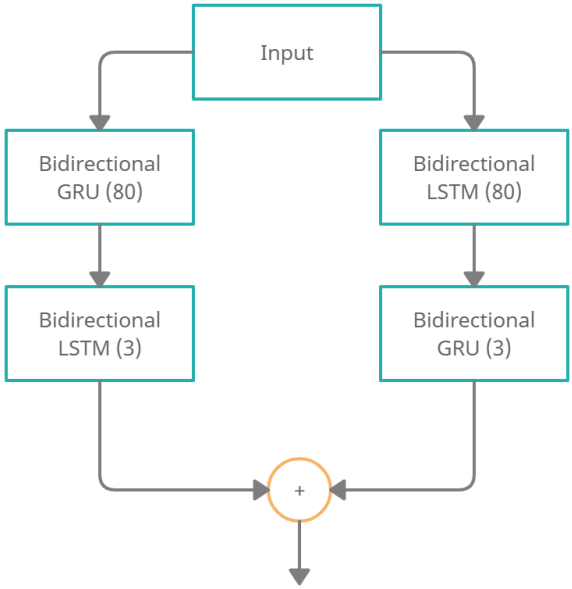


Δοκιμή 45

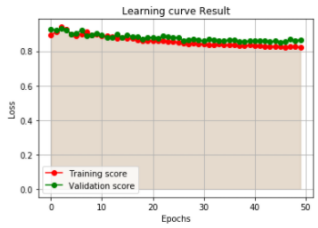
Συνδυασμένα δίκτυα LSTM-GRU με χρήση Skip Layers

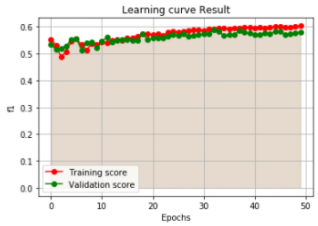
Πραγματοποιήθηκαν δοκιμές σε πολυπλοκότερα δίκτυα συνδυασμένων LSTM-GRU layers με χρήση skipping layers. Παρακάτω φαίνονται η σχεδίαση και τα αποτελέσματα των μετρικών.

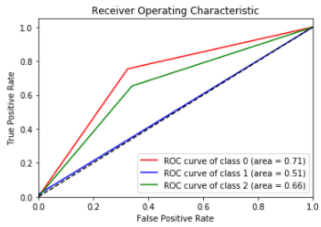
Πείραμα 25:



set				
Training	0.620225	0.600923	0.639397	0.639397
Validation	0.583550	0.580284	0.617879	0.617879





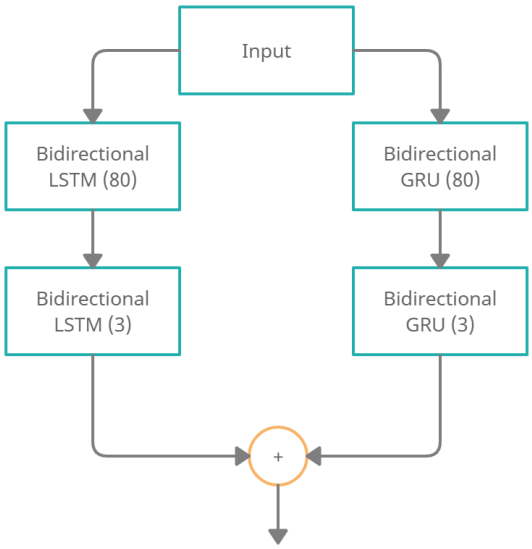


Αποτελέσματα:

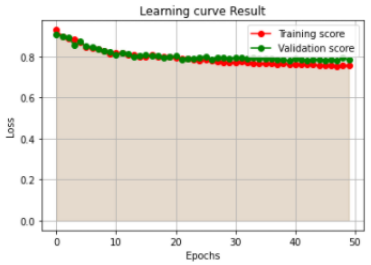
- Clip Gradients Norm: 2
- Dropout: 0.2
- Training Loss: 0.832539
- Validation Loss: 0.852747
- Training f1 score: 0.59584
- Validation f1 score: 0.58669
- Epoch: 38
- Im: 18

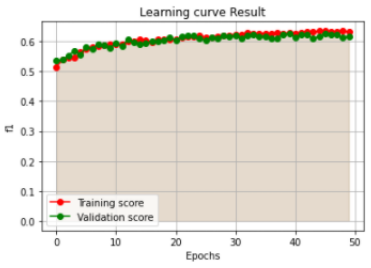
Το πείραμα 25 εμφανίζει ικανοποιητικά αποτελέσματα. Δεν παρατηρείται φαινόμενο overfitting και η εκπαίδευση είναι ορθή. Ωστόσο δεν αποτελεί την καλύτερη απόδοση των μετρικών.

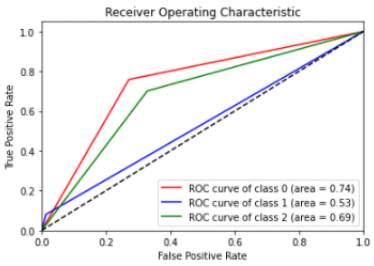
Πείραμα 26:



set				
Training	0.653758	0.636374	0.667752	0.667752
Validation	0.630940	0.619194	0.647677	0.647677





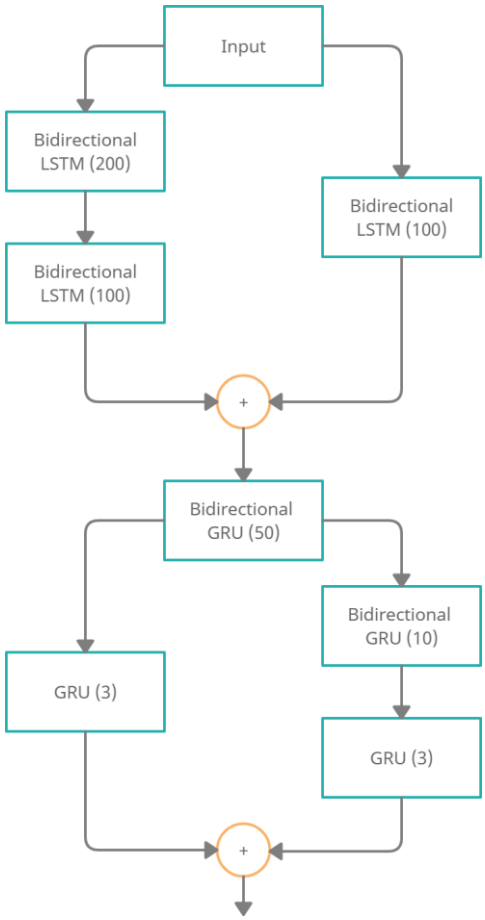


Αποτελέσματα:

- Clip Gradients Norm: 2
- Dropout: 0.2
- Training Loss: 0.758902
- Validation Loss: 0.782089
- Training f1 score: 0.634616
- Validation f1 score: 0.62566
- Epoch: 43
- Im: 17

Τα αποτελέσματα είναι επίσης ικανοποιητικά. Η σύγκλιση των καμπυλών μάθησης είναι πολύ καλή και το τελικό μοντέλο είναι αποδοτικό.

Πείραμα 27:



Αποτελέσματα:

- Clip Gradients Norm: 2
- Dropout: 0.2
- Training Loss: 0.98942534
- Validation Loss: 0.98810594
- Training f1 score: 0.3020103
- Validation f1 score: 0.303968
- Epoch: 15
- Σχόλια: Εμφάνιση Vanishing gradients.

Το μοντέλο εμφανίζει το φαινόμενο των vanishing gradients. Το επόμενο πείραμα αφορά την προσθήκη batch normalization προς αποφυγή αυτού του φαινομένου.

Πείραμα 28: 27 -> Batch Normalization

- Clip Gradients Norm: 2
- Dropout: 0.2
- Training Loss: `0.98796`
- Validation Loss: 0.9875
- Training f1 score: `0.302010`
- Validation f1 score: `0.30396`
- Epoch: 39
- Comments: Αρχικές μεταβολές της f1 τιμής και τελικά εμφάνιση vanishing gradients

Η προσθήκη batch normalization στο μοντέλο του πειράματος 27 δεν επέφερε αλλαγή στα αποτελέσματα. Το μοντέλο δεν είναι ικανοποιητικό.

Βελτιστοποίηση του τελικού μοντέλου

Πραγματοποιείται βελτιστοποίηση του καλύτερου μοντέλου των παραπάνω δοκιμών (δοκιμή 9).

Συγκεκριμένα χρησιμοποιήθηκαν:

- **Weight balancing**

  Με χρήση της συνάρτησης get_weights για να αντιμετωπιστεί η ανισορροπία των κλάσεων στο σετ δεδομένων. Αποτέλεσμα είναι η μείωση του σφάλματος, αλλά και η μείωση του f1 score.

- **Reset lr**

  Εκτέλεση περισσότερων εποχών με επαναρχικοποίηση της τιμής learning rate για την αποφυγή τοπικών ελάχιστων.

- **Διαφορετικές τιμές της νόρμας gradient clipping**

  Η τιμή 3 δείχνει να φέρει τα βέλτιστα αποτελέσματα (δοκιμή 34).

- **Αλλαγή των bidirectional στρωμάτων σε non bidirectional**

  Δεν επέφερε βελτίωση

- **Επιλογή του τελικού hidden sequence αντί του μέσου όρου των hidden sequences τελευταίου layer για επιστροφή από το δίκτυο.**

  Επέφερε μείωση στην απόδοση των μετρικών.

- **Ascending/descending sorted dataset**

  Επιστροφή του διανυσματοποιημένου σετ δεδομένων με ταξινομημένα sequences ως προς το πλήθος των embeddings που αναγνωρίστηκαν για ελαχιστοποίηση του padding πριν την εισαγωγή στο δίκτυο.

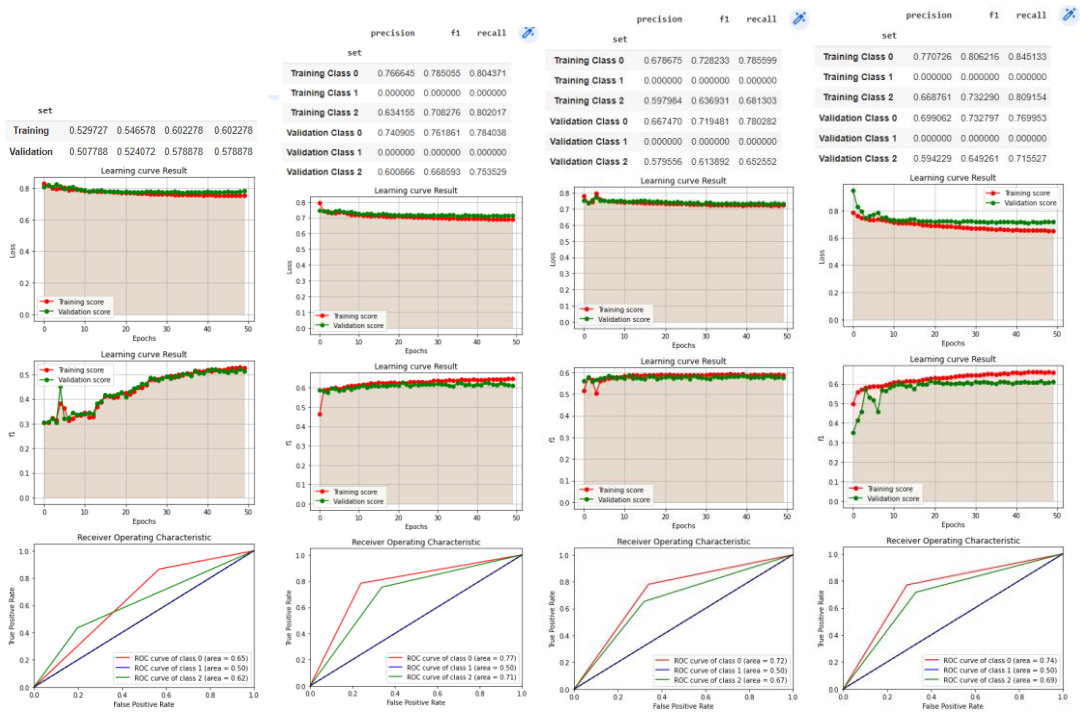
  Η descending τεχνική επέφερε ικανοποιητικά αποτελέσματα, αλλά όχι τα βέλτιστα.

Τα αποτελέσματα:

No	Scheme	GC	DO	Loss (Train/Val)	Score (Train/Val)	Ep	Comments
29	9 -> + weight balance	1	0.2	0.75281/ 0.7788	0.52852/ 0.52068	49	• Καλή σύγκλιση • Μεγάλη εναλλαγή κέρδους αλλά όχι σφάλματος. • Καλύτερο ως προς το σφάλμα μοντέλο
30	29 -> + more epochs + reset LR	1	0.2	0.745487/ 0.767176	0.587623/ 0.576243	44	• Καλή σύγκλιση
31	30-> + more epochs + reset LR	1	0.2	0.75936/ 0.7687	0.58129/ 0.5820	39	• Καλό μοντέλο
32	31-> + more epochs + reset LR	1	0.2	0.75004/ 0.76633	0.587584/ 0.58212	36	• Αμετάβλητο
33	29-> No reset LR	1	0.2	0.70946/ 0.722208	0.620402/ 0.611640	100	• Βέλτιστο σφάλμα μέχρι στιγμής • Καλό μοντέλο
34	29-> + gradient clipping norm =3	3	0.2	0.70536/ 0.71558	0.62864/ 0.62826	22	• Καλό μοντέλο • Καλή σύγκλιση • Ελάχιστο σφάλμα και βέλτιστο f1 σκορ
35	34-> + return max hidden state of final layer	3	0.2	0.75680/ 0.76665	0.3020/ 0.303961	27	• Χαμηλό f1 σκορ • Εμφάνιση vanishing gradients
36	34-> +attention	3	0.2	0.7363/ 0.7457	0.593833/ 0.5849	46	• Καλό μοντέλο • Καλή σύγκλιση
37	36-> - bidirect lstm	3	0.2	0.73255/ 0.73943	0.59686/ 0.59336	36	• Κακό σκορ - μέτριο σφάλμα
38	34 -> + batch size = 8	3	0.2	0.69982/ 0.7338	0.64718/ 0.6190	50	• Καλό μοντέλο • Ελάχιστη απόκλιση καμπυλών • Χειρότερο αποτέλεσμα από το αναμενόμενο
39	34 -> + ascending sorted dataset for minimized padding	3	0.2	0.7521/ 0.78398	0.3053/ 0.30731	26	• Άσχημο σκορ • Πιθανό bug

No	Scheme	GC	DO	Loss (Train/Val)	Score (Train/Val)	Ep	Comments
40	34 -> + descending sorted dataset for minimized padding	3	0.2	0.6908/ 0.7208	0.62736/ 0.61396	20	• Καλό μοντέλο • Έντονες μεταβολές κατά την αρχή της εκπαίδευσης
41	29 -> + gradient clipping = 4	4	0.2	0.6941/ 0.71384	0.61315/ 0.60482	47	• Καλό μοντέλο • Πολύ καλή σύγκλιση
42	34 -> + selected last sequence per layer rather than mean of sequences	3	0.2	0.83623/ 0.8396	0.31857/ 0.3168	15	• Χαμηλό σκορ και υψηλό σφάλμα

Τα παραγόμενα διαγράμματα:



Δοκιμή 19 Δοκιμή 24 Δοκιμή 27 Δοκιμή 29

Εναλλακτική Προεπεξεργασία Δεδομένων

Το βέλτιστο μοντέλο των παραπάνω δοκιμών (δοκιμή 34) δοκιμάζεται με εναλλακτικές μορφές προεπεξεργασίας του σετ δεδομένων (Π1~Π8) που παρουσιάστηκαν στην πρώτη ενότητα. Τα αποτελέσματα των τελικών συγκρίσεων:

No	Scheme	GC	DO	Loss (Train/Val)	Score (Train/Val)	Ep	Comments
1	Μοντέλο 34 με προεπεξεργασία δεδομένων Π2	3	0.2	0.6835/ 0.7110	0.65149/ 0.62560	39	• Καλό μοντέλο • Καλή σύγκλιση • Υψηλό σκορ και χαμηλό σφάλμα
2	Μοντέλο 34 με προεπεξεργασία δεδομένων Π3	3	0.2	0.7008/ 0.7308	0.63269/ 0.5990	49	• Καλό μοντέλο • Καλή σύγκλιση
3	Μοντέλο 34 με προεπεξεργασία δεδομένων Π5	3	0.2	0.67689/ 0.7116	0.65198/ 0.61822	47	• Καλό μοντέλο • Καλό σκορ και σφάλμα
4	Μοντέλο 34 με προεπεξεργασία δεδομένων Π6	3	0.2	0.6687/ 0.7025	0.6427/ 0.6207	40	• Καλό μοντέλο • Ελάχιστη απόκλιση καμπυλών
5	Μοντέλο 34 με προεπεξεργασία δεδομένων Π7	3	0.2	0.6736/ 0.7167	0.63542/ 0.60413	49	• Καλό μοντέλο • Ελάχιστη απόκλιση καμπυλών
6	Μοντέλο 34 με προεπεξεργασία δεδομένων Π8	3	0.2	0.7524/ 0.77130	0.30647/ 0.30545	24	• Χαμηλό σκορ και Υψηλό σφάλμα • Πολύ χαμηλότερη από τη αναμενόμενη απόδοση

Το μοντέλο 34 με την προεπεξεργασία Π1 (εφαρμόστηκε στα πειράματα των παραπάνω ενοτήτων) παρουσιάζει τη βέλτιστη απόδοση.

Το πείραμα 34 σε συνδυασμό με τη προεπεξεργασία Π8 παρουσιάζει χαμηλά αποτελέσματα σε σχέση με τα αναμενόμενα. Η αντιστοίχιση των ειδικών tokens δεν συνείσφερε στην διαφοροποίηση των δεδομένων της κάθε κλάσης, ενώ αντιθέτως ομαδοποιούσε τα ειδικά tokens σαν να είναι κοινά.

Δοκιμή διαφορετικών μεγεθών διανυσμάτων Embeddings

Δοκιμάστηκαν τα σετ από embeddings glove.6B μεγέθους 300, 200, 100, 50 συνιστωσών. Τα βέλτιστα αποτελέσματα συναντώνται στο σετ των 300 διαστάσεων (εξετάστηκε στις παραπάνω ενότητες).

Τα αποτελέσματα:

No	Model	Dimensions	Loss	Score	Ep	Comments
1	34	200	0.6860/ 0.719	0.6273/ 0.6052	49	• Καλό μοντέλο • Καλή σύγκλιση
2	34	100	0.7215/ 0.7200	0.6049/ 0.61707	19	• Καλό μοντέλο • Καλή σύγκλιση
3	34	50	0.7136/ 0.7337	0.62522/ 0.6036	48	• Καλό μοντέλο • Καλή σύγκλιση

Βελτιστοποίηση Loss function και Optimizer

Δοκιμάστηκαν διαφορετικές τιμές παραμέτρων για την Loss function Cross Entropy και τον optimizer Adam. Οι βέλτιστες τιμές παραμέτρων παρουσιάζονται στην **δοκιμή 49**, όπου φέρει τα καλύτερα αποτελέσματα μετρικών.

Τα αποτελέσματα:

No	Scheme	GC	DO	Loss (Train/Val)	Score (Train/Val)	Ep	Comments
43	34 -> + label smoothing = 1e-4	3	0.2	0.6647/ 0.7138	0.6701/ 0.6239	45	• Ελαφρύ overfitting
44	34 -> + weight decay = 1e-4	3	0.2	0.6299/ 0.6581	0.65472/ 0.6388	23	• Καλό μοντέλο • Έντονο overfitting μετά την εποχή 23
45	34 -> + weight decay = 1e-3	3	0.2	0.6325/ 0.66047	0.6580/ 0.63918	28	• Καλό μοντέλο • Έντονο overfitting μετά την εποχή 28

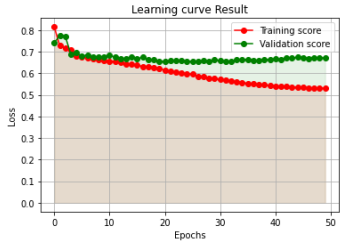
No	Scheme	GC	DO	Loss (Train/Val)	Score (Train/Val)	Ep	Comments
46	34 -> + weight decay = 1e-2	3	0.2	0.853/ 0.854	0.30317/ 0.3039	19	• Vanishing gradients
47	45 -> + betas = (0.8,0.9)	3	0.2	0.65981/ 0.6698	0.633665/ 0.635842	20	• Καλό μοντέλο • Έντονο overfitting μετά την εποχή 20
48	45 -> + betas = (0.7,0.8)	3	0.2	0.6402/ 0.67552	0.65410/ 0.6328	25	• Καλό μοντέλο • Έντονο overfitting μετά την εποχή 25
49	47 -> + amsgrad	3	0.2	0.6432/ 0.6662	0.65066/ 0.6402	14	• Καλό μοντέλο • Βέλτιστο f1 score και σφάλμα • Έντονο overfitting μετά την εποχή 20
50	49 -> + attention	3	0.2	0.7269/ 0.7268	0.57847/ 0.5900	38	• Χαμηλό σκορ και υψηλό σφάλμα
51	49 -> + relu	3	0.2	0.6368/ 0.6945	0.6709/ 0.6148	25	• Εμφάνιση overfitting σε χαμηλό σκορ και υψηλό σφάλμα

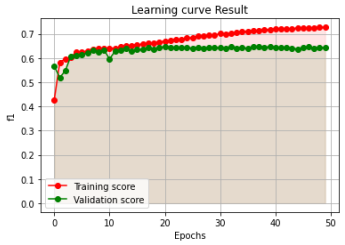
Προσθήκη Attention

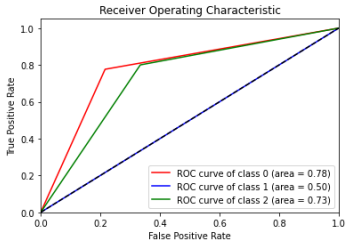
Η προσθήκη της ρουτίνας attention μείωσε τα αποτελέσματα των μετρικών, ενώ αύξησε το σφάλμα, σύμφωνα με τη **δοκιμή 50**. Η συνδυασμένη πληροφορία του ισοσταθμισμένου αποτελέσματος των hidden states που προσφέρει η attention και του αποτελέσματος του κάθε αναδρομικού στρώματος μείωσε την απόδοση. Πιθανώς τα αποτελέσματα να ήταν ιδανικότερα με απλή χρήση του αποτελέσματος της attention κι όχι της ένωσης με το αποτέλεσμα των hidden states. Διαφορετικά, η απλοποίηση της τελικής αναπαράστασης πιθανώς να ανεδείκνυε τη λειτουργικότητα της attention. Ένα bidirectional στρώμα έχει το διπλάσιο μήκος ενός απλού, ενώ με τη προσθήκη attention το μέγεθος τετραπλασιάζεται. Τελικά, η προσθήκη της ρουτίνας attention στο συγκεκριμένο πρόβλημα δεν συνεισφέρει στη λύση του.

Τα διαγράμματα πριν και μετά τη χρήση του attention (πειράματα 49 και 50):

set	precision	f1	recall
Training Class 0	0.845984	0.858445	0.871279
Training Class 1	0.000000	0.000000	0.000000
Training Class 2	0.704762	0.793216	0.907060
Validation Class 0	0.758716	0.767517	0.776526
Validation Class 1	0.000000	0.000000	0.000000
Validation Class 2	0.618289	0.697586	0.800217

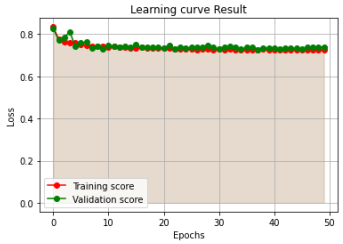


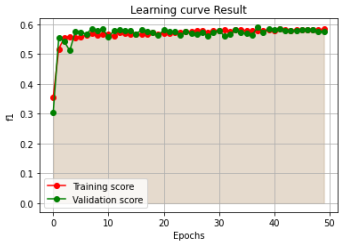


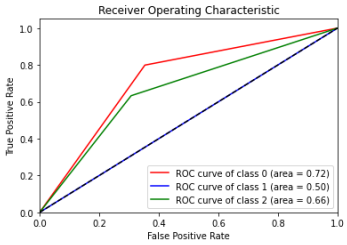


Δοκιμή 49 (Βελτιστη εποχή 14)

set	precision	f1	recall
Training Class 0	0.679349	0.725375	0.778091
Training Class 1	0.000000	0.000000	0.000000
Training Class 2	0.579231	0.620506	0.668115
Validation Class 0	0.664325	0.725490	0.799061
Validation Class 1	0.000000	0.000000	0.000000
Validation Class 2	0.582418	0.606660	0.633008





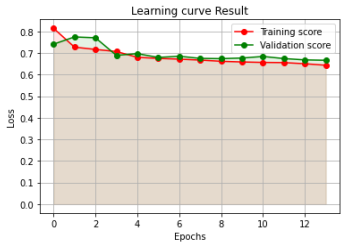


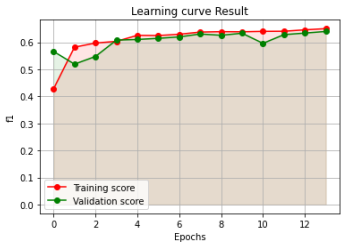
Δοκιμή 50

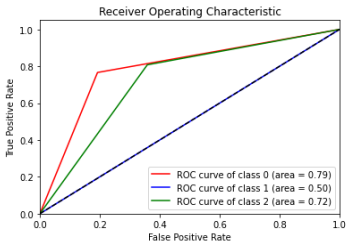
Το Τελικό Μοντέλο

Το τελικό μοντέλο που επιλέχθηκε παρουσιάζεται στη **δοκιμή 49** και φέρει την καλύτερη απόδοση στις μετρικές που αξιολογείται. Τα διαγράμματα μάθησης του παρουσιάζονται παρακάτω:

set	precision	f1	recall
Training Class 0	0.789718	0.794402	0.799142
Training Class 1	0.000000	0.000000	0.000000
Training Class 2	0.639222	0.724486	0.835997
Validation Class 0	0.777143	0.771631	0.766197
Validation Class 1	0.000000	0.000000	0.000000
Validation Class 2	0.603896	0.691129	0.807818







Παρατηρήσεις

Παρουσιάζεται πολύ καλή σύγκλιση στο μοντέλο, αποφυγή του φαινομένου overfitting και μια καλύτερη περιγραφή του μοντέλου από της ROC καμπύλες.

Συγκριμένα, το μοντέλο σταματά την διαδικασία του training την κατάλληλη στιγμή, όπου αποφεύγεται είτε η διαφορά στο κόστος μεταξύ training και validation, είτε η στασιμότητα των αποτελεσμάτων.

Παρατηρούμε πως η ποσότητα των true positive προβλέψεων είναι διαρκώς αρκετά μεγαλύτερη των false positive προβλέψεων. Οι καμπύλες των αποτελεσμάτων τείνουν προς την πάνω αριστερά κορυφή όπου περιγράφει το ιδανικό μοντέλο. Το εμβαδό των καμπυλών με τη διαγώνιο μας δίνει ικανοποιητικά αποτελέσματα, όπως αναμενόταν, εκτός της κλάσης 1. Για την κλάση 1, τα αποτελέσματα είναι κατώτερα των υπόλοιπων κλάσεων, το οποίο είναι αναμενόμενο, αφού οι εγγραφές που κατατάσσονται στην κλάση 1 αποτελούν μόλις το 15% του σετ δεδομένων εμφανίζοντας ανισορροπία.

Επιπρόσθετα, το μοντέλο παρουσιάζει χαμηλή απόδοση στη μετρική precision σε σχέση με τη μετρική recall. Συμπέρασμα αυτού είναι η ύπαρξη περισσότερων false positives σε σχέση με την ύπαρξη των false negatives δειγμάτων. Η κυρίαρχη αιτία του προβλήματος είναι η μη καλή μάθηση των δειγμάτων της κλάσης 1, αφού πολλά δείγματα εσφαλμένα κατατάσσονται ως μέλη της κλάσης 0 ή μέλη της κλάσης 2. Η αντιμετώπιση της ανισορροπίας αυτής πιθανώς να ενίσχυε τα αποτελέσματα της μετρικής precision, συνεπώς τη μάθηση στη κλάση 1 και τη συνολική απόδοση του δικτύου.

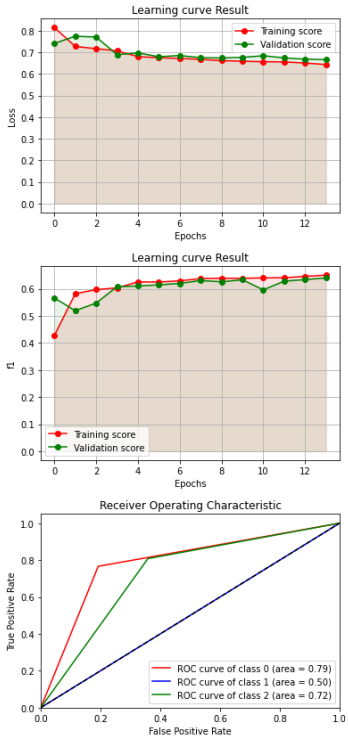
Αντιμετώπιση της Ανισορροπίας Κλάσεων

Το σετ δεδομένων αντιμετωπίζει έντονη ανισορροπία στις κλάσεις. Συγκεκριμένα, τα δείγματα της κλάσης 1 είναι πολύ λιγότερα των υπολοίπων, δημιουργώντας προβλήματα στη μάθηση. Η εφαρμογή των weights κατά την loss function cross entropy τείνει να αντιμετωπίσει το πρόβλημα αυτό. Ως αποτέλεσμα υπήρξε μείωση του σφάλματος, αλλά δεν μεταβλήθηκε η απόδοση των μετρικών και του ROC curve για την κλάση 1, διατηρώντας αδύνατα αποτελέσματα. Εναλλακτική λύση για τη διαχείριση της ανισορροπίας των δεδομένων αποτελεί η μέθοδος data augmentation όπου δε δοκιμάστηκε.

Σύγκριση με το μοντέλο DNN της Εργασίας 2

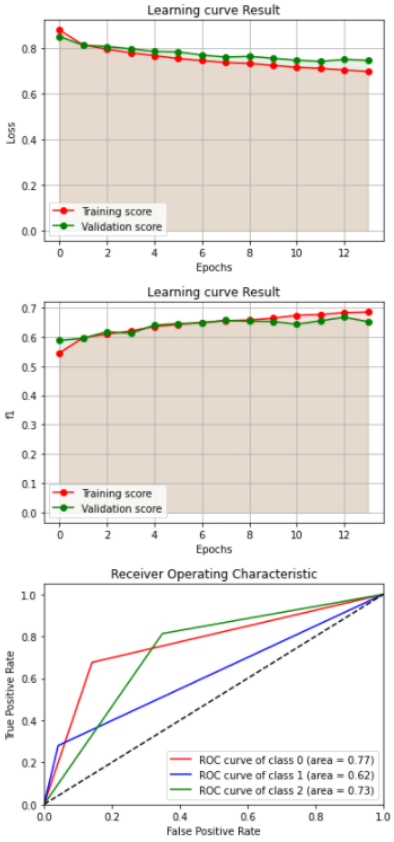
Παρακάτω παρουσιάζεται η απόδοση των μοντέλων στο σετ δεδομένων:

set	precision	f1	recall
Training Class 0	0.789718	0.794402	0.799142
Training Class 1	0.000000	0.000000	0.000000
Training Class 2	0.639222	0.724486	0.835997
Validation Class 0	0.777143	0.771631	0.766197
Validation Class 1	0.000000	0.000000	0.000000
Validation Class 2	0.603896	0.691129	0.807818



RNN μοντέλο

set	precision	f1	recall	accuracy
Training	0.707495	0.689743	0.698360	0.698360
Validation	0.688531	0.671867	0.680105	0.680105



DNN μοντέλο

Εμφανώς τα βέλτιστα αποτελέσματα παρουσιάζει το μοντέλο DNN-Embeddings που παρουσιάστηκε στην Εργασία 2.

Με μια πρώτη σκέψη, ένα πιο πολύπλοκο σύστημα όπως το αναδρομικό νευρωνικό δίκτυο που χρησιμοποιήθηκε, θα έπρεπε να προσαρμόζεται καλύτερα στα δεδομένα του προβλήματος και να παράγει καλύτερα αποτελέσματα. Παρά τις πολλαπλές δοκιμές αυτό δεν επιτυγχάνεται, ενώ πολλοί παράγοντες ενδέχεται να επηρεάζουν το αποτέλεσμα αυτό, όπως:

- Μπορεί το πρόβλημα να προσαρμόζεται καλύτερα με ένα πιο απλό (multinomial logistic regression/DNN) ή πιο σύνθετο (BERT) μοντέλο, από το μοντέλο της εργασίας αυτής.
- Μπορεί το μοντέλο της εργασίας να απαιτεί ένα πιο ισχυρό σύστημα ως προς την βελτιστοποίηση της ισορροπίας των κλάσεων ώστε να αποδείξει τη δυναμικότητα του.