ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

BUSINESS ANALYTICS
Master of Science

Ανάπτυξη Νευρωνικού Δικτύου

[Course: Machine Learning & Content Analytics]

Ηλιάκη Γεωργία (f2821903)
Κιντάκα Μιχαέλλα (f2821904)
Στεφανοπούλου Δανάη (f2821913)

Περιεχόμενα

1.Εισαγωγή 2

2. Διατύπωση Προβλήματος 2

3. Συλλογή Δεδομένων 2

4. Ανάλυση Δεδομένων 4

5. Μεθοδολογία Ανάπτυξης Κώδικα 6

5.1 Multilayer Perceptron (MLP) 7

5.2 Recurrent Neural Network (RNN) 9

5.3 Convolutional Neural Network (CNN) 12

6. Συμπεράσματα 14

7. Διάρθρωση Ομάδας 15

Βιβλιογραφία 16

Παράρτημα 18

1. Εισαγωγή

Ο όρος machine learning αναφέρεται σε αλγορίθμους οι οποίοι είναι σε θέση να μαθαίνουν και να βελτιώνονται όσο αναλύουν δεδομένα. (Harnad Stevan, 2008) Επεξηγηματικά, όσο μεγαλύτερος είναι ο όγκος των δεδομένων που τροφοδοτείται σε έναν αλγόριθμο κατά το στάδιο της ανάπτυξης κι εκπαίδευσης του, τόσο αποτελεσματικότερες θα είναι και οι διαδικασίες της κατηγοριοποίησης των δεδομένων, αναγνώρισης προτύπων (patterns) που αυτά ενέχουν και της διενέργειας προβλέψεων βάσει της ανάλυσης του παρελθόντος. (SAS, White Paper)

Η παρούσα εργασία ασχολείται με την κατηγοριοποίηση των δεδομένων, ή όπως αλλιώς ονομάζεται, με την επιβλεπόμενη μάθηση (supervised learning) κάνοντας χρήση των νευρωνικών δικτύων. Συγκεκριμένα, τα νευρωνικά δίκτυα τροφοδοτούνται από δεδομένα για τα οποία είναι γνωστό σε ποιά κατηγορία ανήκουν και αναμένεται από εκείνα να αναγνωρίσουν βάσει ποιών κριτηρίων γίνεται η κατάταξη ώστε να είναι σε θέση να τα εφαρμόσουν σε άγνωστα δεδομένα. Η ιδιαιτερότητα της εν λόγω εργασίας προκύπτει από την ίδια τη φύση των δεδομένων, τα οποία δεν είναι αριθμητικά όπως συνηθίζεται αλλά ολόκληρες προτάσεις.

Συνεπώς μετά την συλλογή των κειμένων πραγματοποιήθηκε text annotation. Τόσο το annotation όσο και η συλλογή των κειμένων μπορούν να χαρακτηριστούν μεγάλο μέρος της παρούσας ανάλυσης, καθώς έλαβε χώρα για περίπου ένα μήνα. Στη συνέχεια δομήθηκαν, νευρωνικά δίκτυα, τα οποία καλούνται να εντοπίσουν αν υπάρχουν συμπεράσματα (claims) και τα στοιχεία που το τεκμηριώνουν (evidences) σε κάθε κείμενο που επεξεργάστηκε. Η ανάπτυξή τους απασχόλησε την ομάδα μόλις 2 εβδομάδες λόγω εξωτερικών παραγόντων και παράλληλα επιδιώχθηκε η επεξήγησή τους στην παρούσα αναφορά.

2. Διατύπωση Προβλήματος

Πιο αναλυτικά, τα νευρωνικά δίκτυα αναπτύχθηκαν με σκοπό να διαχωρίζουν τα claims και τα evidences των επιστημονικών άρθρων κοινωνικού χαρακτήρα και με αυτόν τον τρόπο μπορεί να διευκρινιστεί το αν η κοινωνία εξελίσσεται προς την επίτευξη αυτών των στόχων. Η σημασία αυτού είναι ανυπολόγιστη. Δεδομένου ότι οι οποιεσδήποτε κοινωνικές ανισότητες έχουν κατά διαστήματα απασχολήσει εκτενώς την επιστημονική κοινότητα, και όχι μόνο, η σχετική αρθρογραφία είναι αστείρευτη. Η ανάγνωση και ανάλυση όλων αυτών θα ήταν αδιαμφισβήτητα μια διαδικασία εξαιρετικά χρονοβόρα. Αντ΄αυτού, τροφοδοτώντας τα αποσπάσματα αυτών στα νευρωνικά δίκτυα εκείνα απευθείας εντοπίζουν το claims εξοικονομώντας χρόνο και προσφέροντας πρόσβαση σε ποικίλες πηγές πληροφόρησης (διαφορετικά άρθρα, συγγραφείς, κανάλια κ.ο.κ.).

3. Συλλογή Δεδομένων

Τα δεδομένα που συλλέχθηκαν είναι abstracts ερευνών που άπτονται θεμάτων κοινωνικού χαρακτήρα όπως αυτό της κλιματικής αλλαγής, της ισότητας των φύλων και της φτώχειας (εφ' εξής SDGs). Τα SDGs έχουν ορισθεί από τα Ηνωμένα Έθνη και είναι 17 στον αριθμό. Πιο συγκεκριμένα συλλέχθηκαν 150 κείμενα από κάθε ομάδα για το κάθε SDG που τα μέλη της είχαν επιλέξει.

Η παρούσα ομάδα επέλεξε να συλλέξει άρθρα που άπτονταν της ισότητας των φύλων και της ενδυνάμωσης των γυναικών, δηλαδή κείμενα που σχετίζονται με το 5ο SDG των Ηνωμένων Εθνών. Η αναζήτηση αυτών πραγματοποιήθηκε μέσω των παρακάτω queries (Πίνακας 1) που προέκυψαν με βάση τα παρακάτω targets του SDG 5.

Target	Queries
5.3 Eliminate all harmful practices, such as child, early and forced marriage and female genital mutilation	"Female Genital Mutilation", "Child Marriage", "Forced Child Marriage"
5.2 Eliminate all forms of violence against all women and girls in the public and private spheres, including trafficking and sexual and other types of exploitation	"Female Genital Mutilation", "Intimate Partner Violence", "Violence against Women" , "Trafficking"

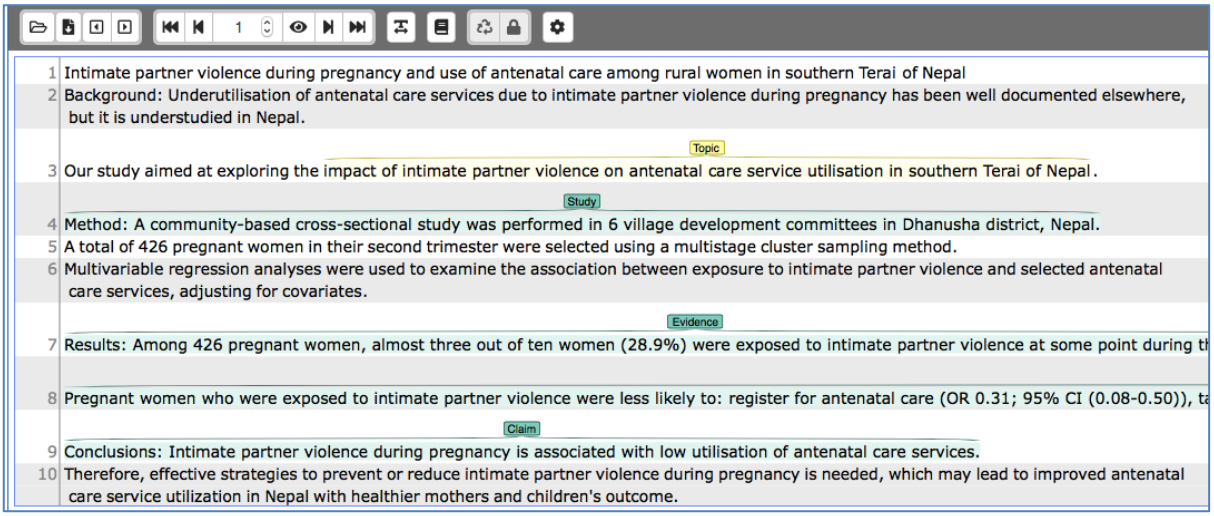
Πίνακας 1: Targets και Queries για την αναζήτηση των κειμένων

Οι πηγές αυτές στην πλειοψηφία τους επιβεβαίωσαν την διαιώνιση της ανισότητας των φύλων παρουσιάζοντας έρευνες σχετικά με την ενδοοικογενειακή βία και τις αιτίες που την προκαλούν και με την ευρύτερη παραβίαση των δικαιωμάτων των γυναικών (αποκλεισμό από τη σεξουαλική ζωή μέσω της κλειτοριδεκτομής, αποκλεισμό από οποιαδήποτε οικονομική δραστηριότητα, στέρηση εκπαίδευσης). Βάσει των εν λόγω πηγών, οι παθογένειες αυτές εντοπίζονται κυρίως στις χώρες της Αφρικής και της Μέσης Ανατολής, χωρίς να περιορίζονται αποκλειστικά σε αυτές. (United Nations)

Ακολούθησε το text annotation μέσω της πλατφόρμας Athena για το καθένα από τα κείμενα. Το text annotation αποτελεί ένα από τα στάδια της επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP). Το NLP αποτελεί μία προσπάθεια επεξεργασίας κειμένου με σκοπό την διεξαγωγή συμπερασμάτων, την ανάλυση του νοήματος και την αποσαφήνιση των συναισθημάτων που διέπουν το εκάστοτε κείμενο. Επιχειρείται, δηλαδή, η δόμηση του κειμένου, το οποίο είναι εκ φύσεως αδόμητο, με σκοπό να εντοπιστεί το θέμα του κειμένου ή η κατηγοριοποίησή του, που επί τω προκειμένω είναι και το ζητούμενο. Σε αυτό το πλαίσιο το text annotation περιγράφει τη διαδικασία σχολιασμού των προτάσεων των κειμένων (research theme, topic, evidence, claim). Η διαδικασία αυτή έχει ως αποτέλεσμα την ανάθεση χαρακτηρισμού (label) των συστατικών αυτών.

Για περαιτέρω αποσαφήνιση, παρατίθεται το εξής παράδειγμα από την πλατφόρμα Athena.

Η παρακάτω εικόνα παρουσιάζει μία περίληψη ενός άρθρου, όπως αυτή υποστηρίζεται από την πλατφόρμα. Οι προτάσεις διαφοροποιούνται μεταξύ τους και απαριθμούνται διαδοχικά. Έπειτα είναι ευθύνη του χρήστη να κατατάξει κάθε πρόταση, ή έστω ένα κομμάτι αυτής στα προαναφερόμενα στοιχεία. Η κατάταξη αυτή είναι το text annotation.



Εικόνα 1: Παράδειγμα annotation σε ένα κείμενο μέσω της πλατφόρμας clarin

Για χάρη βαθύτερης ανάλυσης τα στοιχεία αυτά (labels) επεξηγούνται εκτενέστερα παρακάτω:

- Topic: Το θέμα της έρευνας / άρθρου
- Research Theme: Ο τρόπος διεξαγωγής της έρευνας, μπορεί να είναι μια συνέντευξη, μια στατιστική έρευνα, μια κλινική μελέτη, ένα διαγνωστικό εργαλείο κ.α.
- Evidence: Τα στοιχεία που τεκμηριώνουν το αποτέλεσμα της έρευνας / άρθρου, που είθισται να είναι μετρήσιμο
- Claim: Το συμπέρασμα της έρευνας / άρθρου όπως αυτό προκύπτει από το evidence

Αξίζει να σημειωθεί ότι σε περίπτωση που κάποιο απόσπασμα δεν περιλάμβανε claim, τότε οριζόταν ως missing argument. Αφού, λοιπόν, το annotation είχε ολοκληρωθεί και από τους 3 annotators ελέγχθηκε η συμφωνία μεταξύ τους και επιλέχθηκε εκείνο που είχε οριστεί από την πλειοψηφία. Αν για παράδειγμα 2 στους 3 annotators δεν εντόπισαν argument στο κείμενο, τότε δεν λήφθηκε υπόψη.

Όταν τα δεδομένα ήταν πλέον έτοιμα προς επεξεργασία, δηλαδή μπορούσαν να λειτουργήσουν ως input ενός νευρωνικού δικτύου, ακολούθησε η ανάπτυξη αυτών. Αναπτύχθηκαν 3 διαφορετικά είδη νευρωνικών δικτύων τα οποία αξιολογήθηκαν ως προς την ικανότητα τους στην εκμάθηση των δεδομένων και στην ικανότητα κατάταξης αποσπασμάτων άρθρων τα οποία δεν είχαν χρησιμοποιηθεί κατά τη διάρκεια της εκπαίδευσης.

4. Ανάλυση Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν είναι 890 CSV αρχεία δύο στηλών. Κάθε αρχείο αντιστοιχεί στην περίληψη ενός άρθρου. Στην δεύτερη στήλη είναι η κάθε μία πρόταση όπως αυτή έχει διαχωριστεί από το Athena (βλέπε Εικόνα 1) και στην πρώτη το αντίστοιχο label της. Τα labels προς ανάλυση είναι “claim”, “evidence” και “no label”. Το “no label” αναφέρεται στις προτάσεις του κάθε άρθρου που είτε δεν πήραν κανένα label κατά το text annotation, είτε πήραν οποιοδήποτε άλλο πλην των claim και evidence. Επομένως, δημιουργήθηκε ένα pandas dataframe το οποίο απαρτίζεται από 11.604 σειρές, όσο και συνολικό το πλήθος των προτάσεων των 890

αποκομμάτων. Αξίζει να αναφερθεί πως ως input δεν χρησιμοποιήθηκαν δεδομένα αναφορικά με την ισότητα των φύλων αλλά από όλα τα SDGs.

Η συχνότητα των labels στις 11.064 προτάσεις έχει ως εξής:

Label	Αριθμός Προτάσεων
NO LABEL	8.588 / 11.064 = 77.62%
EVIDENCE	1.486 / 11.064 = 13.43%
CLAIM	990 / 11.064 = 8.95%

Πίνακας 2: Συχνότητα των labels

Δεδομένου ότι η παρούσα εργασία ασχολείται με supervised learning, ο παραπάνω πίνακας ενέχει ιδιαίτερο ενδιαφέρον. Έχει ήδη αναφερθεί πως το επιλεχθέν νευρωνικό δίκτυο θα κληθεί να χαρακτηρίσει μία πρόταση ενός επιστημονικού άρθρου. Η ικανότητα του να κατατάσσει τις προτάσεις εξαρτάται από την ποιότητα της εκπαίδευσης του. Παρατηρείται, λοιπόν, πως το 77.62% των προτάσεων δεν είναι ούτε evidence ούτε claim. Αυτό πρακτικά σημαίνει πως το μοντέλο αναπόφευκτα εκπαιδεύεται ενδελεχώς στην αναγνώριση no label προτάσεων. Αντίθετα, δεν υπάρχουν πολλές προτάσεις που να αναφέρονται στα άλλα δυο. Αυτό σημαίνει πως αφού το input υστερεί, ενδέχεται το μοντέλο να μην είναι σε θέση να διαχωρίσει τα άλλα δύο (unbalanced data). Εναλλακτικά, ακόμη και αν κατασκευαστεί ένα νευρωνικό δίκτυο το οποίο σε κάθε πρόταση που του δίνεται αναγνωρίζεται το no label και τίποτα άλλο, τότε η ικανότητα του θα ισούταν με 77.62%, το οποίο σαν score είναι ικανοποιητικό γενικά, παραπλανητικό ειδικά.

Από αυτές τις 11.064 γραμμές, οι 643 είναι διπλοεγγραφές, των οποίων η κατανομή είναι:

Label	Αριθμός Προτάσεων
NO LABEL	624
EVIDENCE	10
CLAIM	9

Πίνακας 3: Duplicates

Αυτό πρακτικά σημαίνει ότι υπάρχουν 9 άρθρα από τα 890, τα οποία όχι μόνο καταλήγουν στο ίδιο συμπέρασμα, αλλά το διατυπώνουν και με τον ίδιο τρόπο.

Τα δεδομένα υποβλήθηκαν σε μερική επεξεργασία πριν την χρήση τους. Αυτή περιλαμβάνει την αφαίρεση ειδικών συμβόλων και αριθμών μιας και αυτά όχι μόνο δεν συμβάλλουν στην κατάταξη αλλά αντιθέτως δημιουργούν θόρυβο. Επιπλέον, όλα τα κεφαλαία γράμματα γίνανε μικρά με την λογική ότι ο υπολογιστής οι λέξεις πρέπει να είναι γραμμένες ακριβώς με τον ίδιο τρόπο για να τις αναγνωρίσει ο υπολογιστής ως ίδιες.

Πέραν αυτού, θόρυβο προκαλούν και τα επονομαζόμενα stopwords. Τα stopwords στο machine learning είναι οι λέξεις οι οποίες αφαιρούνται από μία πρόταση αφού δεν προσφέρουν παραπάνω πληροφορία ή στην προκειμένη περίπτωση δεν συμβάλλουν στο supervised learning (Sai Teja). Η python, στην οποία χτίστηκαν τα νευρωνικά δίκτυα, περιέχει ήδη μία λίστα με 127 αγγλικά stopwords. Σε αυτή περιλαμβάνονται προθέσεις, σύνδεσμοι, αντωνυμίες κι επιρρήματα. (Sean Bleier) Τα stopwords όμως δεν περιορίζονται εκεί. Είναι και όσες λέξεις αναφέρονται πάνω από ένα ορισμένο από το χρήση threshold βάσει του εξεταζόμενου θέματος. Παραδείγματος χάρη, αν όλα τα άρθρα αναφέρονταν στην ισότητα των φύλων η λέξη "ισότητα" θα αποτελούσε stopword αφού θα εμφανίζονταν τόσο συχνά και στα 3 labels που δε θα προσέφερε αρωγή στην κατάταξη. Επομένως, κάθε μία λέξη που εμπεριέχεται σε οποιαδήποτε πρόταση / σειρά, απομονώνεται (token) και υπολογίζεται η σχετική συχνότητά της.

5. Μεθοδολογία Ανάπτυξης Κώδικα

Έχει γίνει πλέον αντιληπτό πως το πρόβλημα είναι ένα multi-class text classification το οποίο θα προσεγγιστεί με την ανάπτυξη των τριών (3) ειδών μοντέλων νευρωνικών δικτύων.

Ως input στα παρακάτω μοντέλα δόθηκε ένα τμήμα των δεδομένων που αναλύθηκαν προηγουμένως. Επιπλέον, δόθηκαν 99 κείμενα που χρησιμοποιήθηκαν στην αξιολόγηση των μοντέλων ως blind test. Προκειμένου να εξασφαλιστεί η αμεροληψία των μοντέλων, ο διαχωρισμός των προτάσεων των κειμένων έλαβε χώρα βάσει της μεθόδου stratified sampling (Adam Hayes). Το μεγαλύτερο μέρος αυτών, δηλαδή το 65%, χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων, στο 15% ανατέθηκε ο έλεγχος τους (test dataset) ενώ το υπόλοιπο 20% χρησίμευσε στην αξιολόγησή του (validation dataset). Στην ουσία, η ικανότητα των μοντέλων ως προς την κατάταξη αξιολογείται στα test και validation datasets. Επεξηγηματικά, πρόκειται για δεδομένα τα οποία δεν χρησιμοποιήθηκαν κατά την εκπαίδευση, ωστόσο η κατάταξη είναι γνωστή. Επομένως τα τρία μοντέλα νευρωνικών δικτύων καλούνται να κατατάξουν τις προτάσεις κι έπειτα ο χρήστης ελέγχει εάν τα αποτελέσματα είναι σωστά δημιουργώντας ένα confusion matrix.

Έχει ήδη αναφερθεί πως τα μοντέλα που αναπτύχθηκαν είναι τρία. Αναλυτικότερα, πρόκειται για ένα Multilayer Perceptron (MLP), της κατηγορίας feedforward artificial neural network (ANN), ένα Recurrent Neural Network (RNN) και Convolutional Neural Network (CNN). Κατά την διάρκεια της εκπαίδευσης των μοντέλων χρησιμοποιήθηκε ένα callback function (μορφής checkpoint), με σκοπό να αποθηκευτεί το μοντέλο με την υψηλότερη ακρίβεια πρόβλεψης (εφ΄εξής accuracy). Πρακτικά αυτό σημαίνει ότι αποθηκεύεται το μοντέλο όπου μεγιστοποιήθηκε το accuracy στην αντίστοιχη εποχή[1]. (Supratim Haldar,2019)

Αξίζει να σημειωθεί πως το στοιχείο μεταξύ των 3 αυτών μοντέλων είναι η ίδια τους η σύνθεση (compiling). Σε κάθε μία προσπάθεια πρόβλεψης του μοντέλου, υπάρχει μία συνάρτηση η οποία μετρά την απόκλιση της πρόβλεψης αυτής και των πραγματικών δεδομένων (loss function). (Μηχανική Μάθηση) Για τις ανάγκες της παρούσας εργασίας η loss function που χρησιμοποιήθηκε ονομάζεται categorical cross entropy ή εναλλακτικά Softmax Loss. Για κάθε μία από τις προτάσεις προσδίδεται μία πιθανότητα για κάθε label. Η τελική κατάταξη λαμβάνει χώρα βάσει της υψηλότερης πιθανότητας υπό περίπτωση.

[1] Κάθε μία εποχή (epoch) είναι ένας πλήρης κύκλος εκπαίδευσης κατά τον οποίο όλες οι μεταβλητές έχουν χρησιμοποιηθεί μία φορά για την ενημέρωση των βαρών των νευρωνικών δικτύων. (Α. Λύκας)

Επιπλέον, με σκοπό την βελτιστοποίηση του τρόπου "κατάβασης" στο βέλτιστο σημείο χρησιμοποιείται ο Adam optimizer, οποίος συμβάλει στον σκοπό αυτό με την ενημέρωση των βαρών του δικτύου με βάση τα δεδομένα εκπαίδευσης, (Jason Brownlee, 2017). Συνεπώς, ο εν λόγω αλγόριθμος ανανεώνει επαναληπτικά τα βάρη των νευρώνων με βάσει τα δεδομένα που δίνονται κατά την εκπαίδευση.

5.1 Multilayer Perceptron (MLP)

Πριν την πραγματοποίηση της εκπαίδευσης του dataset στο MLP μοντέλο τα labels των κειμένων, που αποτελούν την ανεξάρτητη μεταβλητή (y), κωδικοποιήθηκαν βάσει της μεθόδου One-Hot Encoder, ενώ οι προτάσεις των κειμένων, εξαρτημένη μεταβλητή (x) έγιναν encoded με την μέθοδο Count Vectorizer. Η μέθοδος One Hot Encoder είναι μία τεχνική η οποία εφαρμόζεται σε κατηγορικές μεταβλητές και εν προκειμένω στα 3 labels τα οποία σχηματίζουν ένα διάνυσμα μήκους 3. Η κάθε μία κατηγορία (no label, claim και evidence) αντιπροσωπεύεται σε κάθε πρόταση από το 0 και το 1. Εάν μία πρόταση έχει χαρακτηριστεί δηλαδή ως evidence, η αντίστοιχη θέση στο διάνυσμα θα πάρει την τιμή 1 ενώ στις θέσεις που αντιστοιχούν το no label και claim θα υπάρχει η τιμή 0. Όσον αφορά το count vectorizer, πρόκειται για μία τεχνική η οποία άπτεται του one hot encoding. Δεδομένου ενός λεξικού, εν προκειμένω των προτάσεων προς κατάταξη, όλες οι λέξεις απομονώνονται και η ύπαρξη τους σε μία πρόταση δηλώνεται με 1 ενώ η ανυπαρξία τους με 0. Αξίζει να σημειωθεί πως η συνάρτηση του Count Vectorizer εμπεριέχει κάποια ορίσματα που συμβάλλουν στο text processing (max_df, stopwords και λοιπά).

Όσον αφορά, λοιπόν, τη δομή του τελικού MLP μοντέλου περιέχει 1 hidden dense layer με ReLu activation function[2] και dropout rate 20%. Στην προκειμένη περίπτωση επιλέχθηκε η Rectified Linear Unit (ReLU), καθώς επιταχύνει σημαντικά τη σύγκλιση της στοχαστικής καθόδου στο βέλτιστο σημείο, συνεπώς επιτυγχάνεται ταχύτερη μάθηση. Το dropout rate, το οποίο ορίζει το ποσοστό των νευρώνων που δεν θα χρησιμοποιηθούν στο επόμενο hidden dense layer είναι μια από τις παραμέτρους που επιλέχθηκε μετά από tuning hyper-parameters. Το γεγονός ότι το 20% των νευρώνων από κάθε layer δεν θα χρησιμοποιείται σαν input στο επόμενο πρακτικά χρησιμεύει στην αποφυγή του overfitting του μοντέλου. Βεβαία στο output layer ως activation function ορίζεται η softmax, καθώς, αποτελεί μια συνάρτηση που παίρνει τις τελικές output τιμές και τις μετατρέπει σε πιθανότητα. Λειτουργεί αποτελεσματικά σε προβλήματα όπως το παρόν αφού πρόκειται για multiclass classification, και επιστρέφει ως αποτέλεσμα ένα «σκορ εμπιστοσύνης» για κάθε class, δηλαδή για κάθε πιθανό label της πρότασης που εξετάζεται (evidence, claim ή no label). Το παρόν μοντέλο εκπαιδεύτηκε στις 40 εποχές στο training dataset.

Τέλος, ακολουθεί το compile του μοντέλου όπως αυτό περιγράφηκε παραπάνω και το fitting αυτού στα training data. Κατά τη διάρκεια του fitting εφαρμόζονται βάρη (class weights) στο μοντέλο, ώστε να μην ληφθεί υπόψη ότι τα δεδομένα είναι unbalanced και να μην επηρεαστεί από το μεγαλύτερο αριθμό no label προτάσεων.

Αξίζει να αναφερθεί πως πριν την επιλογή του παραπάνω μοντέλου δοκιμάστηκαν παρόμοια μοντέλα σε 30, 40, 50 εποχές, με διαφορετικό αριθμό dense layer και διαφορετικά dropout rates (π.χ. 30%, 40%). Άξιο προσοχής αποτελεί το γεγονός ότι το παρόλο που το accuracy αυξανόταν

[2] Activation function : μαθηματικές εξισώσεις που καθορίζουν την έξοδο ενός νευρωνικού δικτύου. Η συνάρτηση συνδέεται με κάθε νευρώνα στο δίκτυο και καθορίζει εάν θα πρέπει να ενεργοποιηθεί ή όχι, με βάση το εάν η είσοδος κάθε νευρώνα είναι σχετική με την πρόβλεψη του μοντέλου

όσο αυξάνονταν τα hidden layers, στο confusion matrix δεν αντικατοπτρίζονταν η αντίστοιχη απόδοση. Χαρακτηριστικά το accuracy ενός μοντέλου με 5 hidden dense layers και χωρίς βάρη στο test dataset έφτανε στο 78.072 % , αλλά το confusion matrix (Εικόνα 2) αναγνώριζε λιγότερα claim και evidence, από ότι το τελικό MLP που επιλέχθηκε με 1 hidden layer και βάρη αλλά 74% accuracy (Εικόνα 3). Αυτή η διαφορά στο accuracy μπορεί να εξηγηθεί καθώς το dataset έχει παραπάνω no labels, όπως ήδη αναφέρθηκε. Ποαρότι, λοιπόν, λαμβάνεται ως επιτυχία να βρίσκει περισσότερα “No labels” δεν προφέρεται ιδιαίτερη πληροφορία στο μοντέλο.

	CLAIM	EVIDENCE	NO LABEL
CLAIM	31	10	108
EVIDENCE	8	34	181
NO LABEL	19	38	1231

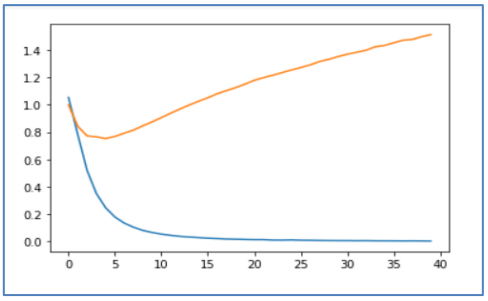
Εικόνα 2: Confusion Matrix αρχικού μοντέλου

	CLAIM	EVIDENCE	NO LABEL
CLAIM	45	14	90
EVIDENCE	15	68	140
NO LABEL	55	106	1127

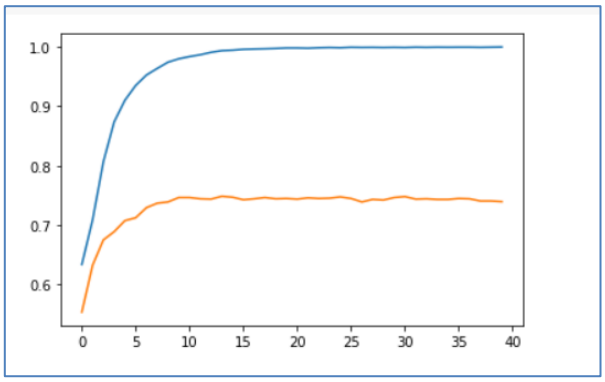
Εικόνα 3: Confusion Matrix τελικού μοντέλου

Αξιολόγηση

Παράλληλα, με την παρακολούθηση του confusion matrix εξετάζεται το Loss και Accuracy plot των μοντέλων. Από τα παρακάτω διαγράμματα μπορεί να συμπεράνει κανείς ότι μετά την 5η εποχή περίπου το μοντέλο δεν εκπαιδεύεται πλέον, καθώς το validation loss (Εικόνα 4 - πορτοκαλί γραμμή) ολοένα και αυξάνεται και παράλληλα το accuracy (Εικόνα 5) περίπου μετά την 5η εποχή αρχίζει να σταθεροποιείται.



Εικόνα 4: Πορεία Loss στο validation και train dataset



Εικόνα 5: Πορεία Accuracy στο validation και train dataset

Η απόδοση ενός αλγορίθμου Classification, όπως το MLP, μπορεί επίσης να αναγνωριστεί μέσω του Area Under Curve. Αυτό αντιπροσωπεύει το εμβαδόν κάτω από το ROC Curve (βλέπε Παράρτημα 1) και μετρά το πόσο καλά κατατάσσονται οι προβλέψεις. Η ποιότητα των προβλέψεων σε αυτό το metric απορρέει από την σύγκριση των True Positive και False Negative τιμών. Στην προκειμένη περίπτωση υπολογίστηκε 0.722, μια τιμή ικανοποιητική για την επιλογή του συκεριμένου MLP μοντέλου έναντι άλλων MLP που δοκιμάστηκαν.

Τέλος, το classification report[3] του παρόντος μοντέλου (Εικόνα 6) παρουσιάζει μια συνολική εικόνα του μοντέλου. Το recall, ορίζει ότι το 30% των προτάσεων βρέθηκαν ως claim και evidence έναντι του συνολικού αριθμού αυτών και κατά 88% αναγνωρίστηκαν τα no labels. Επιπλέον, με το precision μπορεί κανείς να διακρίνει ότι το ποσοστό ορθής αναγνώρισης στην κάθε κατηγορία label. Για παράδειγμα, το 0.39 των προτάσεων που έγιναν classify ως claim ταξινομήθηκαν ορθά σε αυτή την κατηγορία. Τέλος, στο support παρουσιάζεται ο πραγματικός αριθμός των προτάσεων που ανήκουν σε κάθε κατηγορία.

	precision	recall	f1-score	support
CLAIM	0.39	0.30	0.34	149
EVIDENCE	0.36	0.30	0.33	223
NO LABEL	0.83	0.88	0.85	1288
accuracy			0.75	1660
macro avg	0.53	0.49	0.51	1660
weighted avg	0.73	0.75	0.74	1660

Εικόνα 6 : MLP Classification Report

5.2 Recurrent Neural Network (RNN)

Μια διαφορά μεταξύ των RNN και των feed forward networks, όπως το MLP που αναλύθηκε προηγουμένως, είναι ο χρόνος που απαιτούν για την εκπαίδευσή τους. Συγκεκριμένα, το output ενός hidden layer σε ένα RNN μοντέλο αυτο-τροφοδοτείται συνεχώς. (Andrew Thomas). Ο τρόπος, λοιπόν, που λειτουργούν τα RNN είναι διαφορετικός από ότι τα MLP. Ως προέκταση αυτού, ακόμη

[3] Το f1-score αποτελεί ένα weighted average του precision και του recall για αυτό και δεν αναλύεται περαιτέρω.

και οι προτάσεις προς ταξινόμηση (εξαρτημένη μεταβλητή χ) απαιτούν διαφορετική κωδικοποίηση και όχι το Count Vectorizer που περιγράφηκε παραπάνω.

Τα RNN, όπως και τα CNN που θα αναλυθούν αργότερα (βλέπε 5.3), αποτελούν είδη μοντέλων που λειτουργούν με embeddings και η κωδικοποίηση των προτάσεων (εξαρτημένη μεταβλητή - x) πραγματοποιείται με το tokenizer function.

Αναφέρθηκε προηγουμένως πως το tokenization αναφέρεται σε μία διαδικασία κατά την οποία ένα αδόμητο κείμενο σπάει σε μεμονωμένες λέξεις οι οποίες ονομάζονται tokens. (Techopedia) Από το δοθέν training dataset προέκυψε ότι η μεγαλύτερη σε μήκος πρόταση εμπεριέχει 770 λέξεις, tokens. Πρόκειται για ένα σημαντικό όγκο πληροφορίας και συνεπώς τίθεται θέμα αποδοτικότητας. Προκειμένου η διαδικασία εκπαίδευσης του μοντέλου να μην είναι εξαιρετικά χρονοβόρα οι λέξεις που επιλέχθηκαν είναι 250 ανά πρόταση.

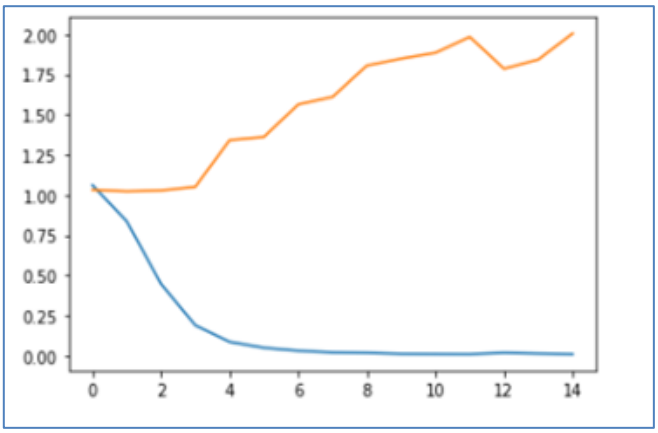
Το επιλεγμένο μοντέλο RNN, χρησιμοποιεί 100 embeddings dimensions, μόλις 15 εποχές και dropout rate 30%. Κάθε embedding του μοντέλου λαμβάνει ως είσοδο 250 λέξεις. Ο αριθμός αυτός προέκυψε μετά από δοκιμές (hyper-parameter tuning) . Το μοντέλο χρησιμοποιεί τις 128 προτάσεις ως δείγμα (από την 1η έως την 128η) ως πρώτο training dataset ώστε να εκπαιδευτεί το δίκτυο. Στη συνέχεια εκπαιδεύει το νευρωνικό δίκτυο με βάση τις επόμενες 128 (από την 129η έως 256 κλπ). Αυτή είναι η διαδικασία εκπαίδευσης στο RNN μοντέλο μέχρι να χρησιμοποιηθούν όλες οι προτάσεις στο training dataset. (David Israwi) Επιπλέον, χρησιμοποιήθηκε μια Flatten function, ένα dense layer με 64 νευρώνες και ως activation function η ReLu. Στο output layer ορίστηκε softmax activation function για του λόγους που αναλύθηκαν και παραπάνω στην ανάπτυξη του MLP μοντέλου.

Όταν πρόκειται για ανάλυση κειμένου, σημαντικές είναι οι λέξεις που περικλείουν συνολικά η μία άλλη και όχι μόνο όσες προηγούνται αυτής. Γι' αυτό το λόγο οι προτάσεις αναλύθηκαν bidirectionally. Έτσι επιτυγχάνεται η ολιστική αξιολόγηση του περιεχομένου.

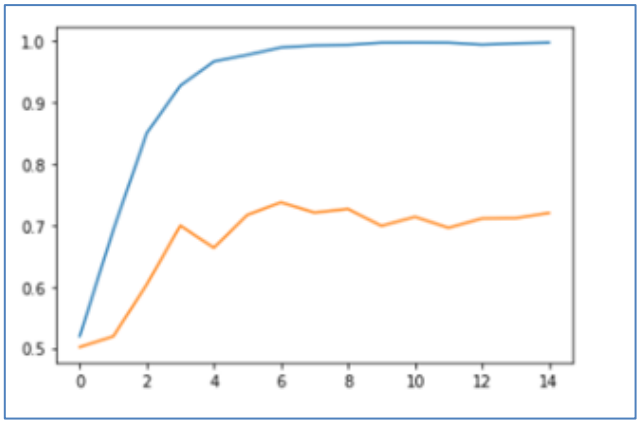
Το μέγεθος του batch δοκιμάστηκε σε πολλαπλές τιμές συνήθως πολλαπλάσιες του 2, αλλά δεν φάνηκε να επηρεάζει ιδιαίτερα την απόδοση του μοντέλου, οπότε ορίστηκε σε 128. Επίσης, το embedding δοκιμάστηκε σε διάφορες τιμές (π.χ. 250,200,150) και επιλέχθηκε το 100 (emb_dim) ως εκείνη που συνδυαστικά με τις υπόλοιπες παραμέτρους του μοντέλου είχε ως αποτέλεσμα το καλύτερο το confusion matrix. Με τον ίδιο τρόπο επιλέχθηκε και το μέγεθος του λεξικού που στο RNN και CNN μοντέλου στις 15.000 λέξεις.

Αξιολόγηση

Όπως και στο MLP μοντέλο για την αξιολόγηση του μοντέλου παρατίθενται το Loss και Accuracy plot. Από τα παρακάτω διαγράμματα μπορεί να συμπεράνει κανείς ότι μετά την πρώτη κιόλας εποχή το μοντέλο δεν εκπαιδεύεται πλέον καθώς το validation loss (Εικόνα 7 - πορτοκαλί γραμμή) ολοένα και αυξάνεται. Όσο για το accuracy (Εικόνα 8) περίπου έπειτα την 5η εποχή περίπου αρχίζει να σταθεροποιείται, όσο στο validation τόσο και στο train dataset.



Εικόνα 7: Πορεία Loss στο validation και train dataset - RNN



Εικόνα 8: Πορεία Accuracy στο validation και train dataset - RNN

Επιπλέον, το AUC score στο RNN μοντέλο βρέθηκε να φτάνει το 0.70, κάτι που σε συνδυασμό με το confusion matrix καθιστά το MLP μοντέλο ως το καλύτερο μέχρι στιγμής. Τέλος, το classification report του παρόντος μοντέλου παρουσιάζει μια συνολική εικόνα του μοντέλου. (Εικόνα 9) Παραπάνω αναφέρθηκε ότι μέσω του precision, μπορεί κανείς να διακρίνει το ποσοστό ορθής αναγνώρισης για την κάθε κατηγορία label. Όντως στις κατηγορίες των claim και evidence τα ποσοστά αυτά αντιστοιχούν σε 0.33% κι 0.32%. Το μοντέλο δεν αποτυγχάνει να αναγνωρίσει προτάσεις που είναι evidence και claims καθώς αναγνωρίζει 70 στις 223 και 34 στις 149 προτάσεις αντίστοιχα (Εικόνα 10). Συνοψίζοντας, μπορεί το συγκεκριμένο RNN μοντέλο να έχει μόλις 71% accuracy, όμως επιτυγχάνει να αναγνωρίζει το μεγαλύτερο αριθμό evidences προτάσεων και από τα τρία μοντέλα που περιγράφονται.

	precision	recall	f1-score	support
CLAIM	0.33	0.23	0.27	149
EVIDENCE	0.32	0.31	0.32	223
NO LABEL	0.83	0.86	0.85	1288
accuracy			0.73	1660
macro avg	0.50	0.47	0.48	1660
weighted avg	0.72	0.73	0.72	1660

Εικόνα 9 : RNN Classification Report

	CLAIM	EVIDENCE	NO LABEL
CLAIM	34	25	90
EVIDENCE	15	70	138
NO LABEL	53	121	1114

Εικόνα 10 : RNN Confusion Matrix

5.3 Convolutional Neural Network (CNN)

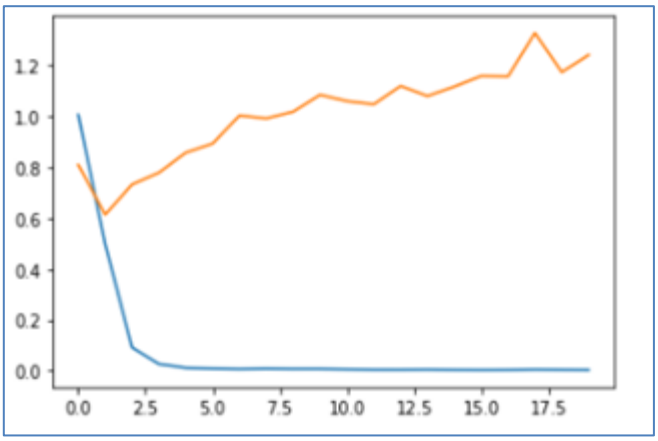
Σε ένα παραδοσιακό μοντέλο νευρωνικού δικτύου όπως το MLP συνδέουμε κάθε νευρώνα εισόδου με κάθε νευρώνα εξόδου στο επόμενο στρώμα. Στα CNN αντ 'αυτού, χρησιμοποιούμε convolutions (φίλτρα) πάνω στο input για να υπολογίσουμε την έξοδο συνδυάζοντας τα αποτελέσματά τους.

Κατά τη διάρκεια της εκπαίδευσης, ένα CNN μαθαίνει με βάση την εργασία για την οποία αναπτύχθηκε. Για παράδειγμα, στο image recognition ένα CNN στο πρώτο φίλτρο εκπαιδεύεται από ακατέργαστα pixels, στη συνέχεια ανιχνεύει απλά σχήματα και όσο αυξάνεται ο αριθμός των φίλτρων έχει την δυνατότητα να αναγνωρίσει high level χαρακτηριστικά, όπως σχήματα προσώπου. Κάτι παρεμφερές μπορεί να πραγματοποιηθεί και σε αδόμητο κείμενο, όπου τα pixels αντικαθίστανται από τις προτάσεις. Πιο συγκεκριμένα, κάθε πρόταση μετατρέπεται σε μια σειρά από vectors. (Denny Britz) Στο επιλεχθέν μοντέλο ορίζεται ότι το συνολικό λεξικό που χρησιμοποιείται είναι 15000 λέξεων (input_dim), και το μέγιστο πλήθος λέξεων που μια πρόταση περιέχει είναι 250 λέξεις. Οι παραπάνω αριθμοί είναι αποτέλεσμα δοκιμών για το με ποιες παραμέτρους ανταποκρίνεται καλύτερα το μοντέλο. Αφού λοιπόν οριστούν τα embeddings του μοντέλου, εφαρμόζονται τα convolutions με kernel size 5, τα οποία πρακτικά είναι το μεγεθος του 1D convolution window και τα 128 φίλτρα, τα οποία προέκυψαν επίσης από hyper-paremeters tuning. (δοκιμάστηκαν επίσης 64, 256 filters) Στα hidden layers διατηρούμε ως activation function την ReLu.

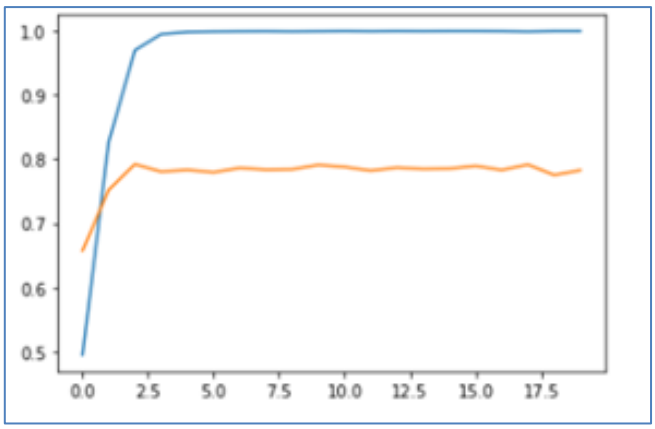
Ένα ακόμα χαρακτηριστικό των convolutional μοντέλων είναι η συνάρτηση pooling, η οποία εφαρμόζεται μετά τα convolutional layers με σκοπό να εξάγει ένα subsample από το output τους και να το τροφοδοτήσει στο επόμενο dense layer. Στο επιλεχθέν μοντέλο εφαρμόστηκε ως η max pooling συνάρτηση, έναντι της average pooling. Τέλος, αξίζει να σημειωθεί ότι το dropout rate διατηρήθηκε στο 30%, όπως και στο RNN μοντέλο και ως τελική activation function παρέμεινε η softmax λόγω της φύσης του προβλήματος, όπως έχει ήδη αναφερθεί.

Αξιολόγηση

Στο παρόν μοντέλο τα παρακάτω διαγράμματα αποδεικνύουν ότι η αποδοτική εκπαίδευση του μοντέλου σταματάει από τις πρώτες κιόλας εποχές, καθώς η λεχιστοποιήσει του loss και στα δυο dataset πραγματοποιείται πριν την 3η εποχή. (Εικόνα 11) Παρόμοια συμπεριφορά παρουσιάζει και το accuracy (Εικόνα 12), αφού σταθεροποιείται πριν την 3η εποχή τόσο στο validation όσο και στο train dataset.



Εικόνα 11: Πορεία Loss στο validation και train dataset - CNN



Εικόνα 12: Πορεία Accuracy στο validation και train dataset - CNN

Πέρα, λοιπόν, από το loss και το accuracy υπολογίζεται και το AUC score, το οποίο όπως έχει ήδη αναφερθεί υπολογίζει την πιθανότητα μεταξύ των True Positive και False Negative προβλέψεων. Το γεγονός ότι το CNN επιτυγχάνει το υψηλότερο score (0.82) και από τα δύο προηγούμενα μοντέλα είναι αρκετά ενθαρρυντικό για την επιλογή του ως το καλύτερο. Επιπροσθέτως, αξίζει να σημειωθεί ότι το accuracy στο test dataset φτάνει το 80%.

Προχωρώντας, μελετήθηκε το classification report του CNN. (Εικόνα 13) Πιο συγκεκριμένα, το 37% των προτάσεων αναγνωρίζονται ως claims και το 68% αυτών έχουν σωστά τοποθετηθεί σε αυτή την κατηγορία. Αντίστοιχα, μπορεί μόλις το 29% των προτάσεων να έχουν αναγνωριστεί ως evidences αλλά έχουν τοποθετηθεί σωστά σε ποσοστό 47%. Τέλος, το confusion matrix του CNN μοντέλου αποδεικνύει ότι από τα 149 claims τα 55 αναγνωρίστηκαν σωστά, από τα 223 evidences αναγνωρίστηκαν ορθά τα 65 και από τις 1288 προτάσεις “no label” οι 1214 κατατάχθηκαν σωστά σε αυτήν την κατηγορία. Συνεπώς, 1334 προτάσεις ταξινομήθηκαν σωστά έναντι των 326. Αυτός είναι και ο μεγαλύτερος αριθμός σωστών προβλέψεων μεταξύ των μοντέλων. (Εικόνα 14)

	precision	recall	f1-score	support
CLAIM	0.68	0.37	0.48	149
EVIDENCE	0.47	0.29	0.36	223
NO LABEL	0.84	0.94	0.89	1288
accuracy			0.80	1660
macro avg	0.66	0.53	0.58	1660
weighted avg	0.78	0.80	0.78	1660

Εικόνα 13 : CNN Classification Report

	CLAIM	EVIDENCE	NO LABEL
CLAIM	55	15	79
EVIDENCE	11	65	147
NO LABEL	15	59	1214

Εικόνα 14: CNN Confusion Matrix

6. Συμπεράσματα

Ως αποτέλεσμα της παραπάνω ανάλυσης και έπειτα από σύγκριση των confusion matrix αλλά και του accuracy επιλέχθηκε ως καλύτερο μοντέλο το CNN, καθώς αυτό είναι το μοντέλο που όπως αποδεικνύεται κατατάσσει τις περισσότερες προτάσεις στην σωστή τους κατηγορία με βάση το confusion matrix. (Πίνακας 4) Η τελευταία φάση της παρούσας εργασίας και έπειτα από την καθοδήγηση των διδασκόντων το καλύτερο μοντέλο, αφού είχε αποθηκευτεί φορτώθηκε από την αρχή και ελέγχθηκε σε ένα blind test dataset 99 abstracts, όπου καμία πρόταση δεν έχει label. Ακολουθήθηκε text processing και encoding και στην συνέχεια ήρθε σε μια μορφή όπου κάθε σειρά του νέου dataset έχει το κείμενο και αν βρέθηκε evidence ή claim σε κάποια από τις προτάσεις που περιλαμβάνει. Τα αποτελέσματα αυτά θα δοθούν στους διδάσκοντες και αν πλησιάζουν το πραγματικό annotation των κειμένων, μπορεί κανείς να συμπεράνει ότι το συγκεκριμένο μοντέλο οδηγεί τους ερευνητές ένα βήμα πιο κοντά στην εύκολη και γρήγορη αναγνώριση της προόδου ή όχι των SDGs, με την αναγνώριση των evidence και claims, όπως άλλωστε ήταν ο σκοπός του.

Μοντέλο	True Positive	False Negative
MLP	1240	420
RNN	1218	442
CNN	1334	326

Πίνακας 4: Συγκεντρωτικός πίνακας σύγκρισης confusion matrices

Αυτό που πέτυχε η ομάδα αυτή είναι ότι σε έναν κόσμο που κατακλύζεται από πληροφορία, δοθέντων ενός όγκου άρθρων κοινωνικού χαρακτήρα, κάποιος δε χρειάζεται να διαβάσει ολόκληρα τα άρθρα για να καταλάβει τι λένε. Το CNN επιστρέφει το claim και το evidence.

7. Διάρθρωση Ομάδας

Η ομάδα απαρτίζεται από 3 άτομα την Ηλιάκη Γεωργία, την Κιντάκα Μιχαέλλα και την Δανάη Στεφανοπούλου. Τα μέλη της εν λόγω ομάδας εκτός από το ότι έχουν συνεργαστεί αρκετές φορές στο παρελθόν κάτω από συνθήκες πίεσης, γνωρίζονται και σε προσωπικό επίπεδο. Αυτό συνέβαλε στην κατανομή των καθηκόντων βάσει προσωπικών ικανοτήτων, στην εξασφάλιση ειλικρινούς επικοινωνίας και στην ανιδιοτελή προσφορά εντός της ομάδας. Συνεπώς δεν αναφέρθηκε καμία δυσκολία στο επίπεδο συνεργασίας. Όσον αφορά το τεχνικό κομμάτι της εργασίας, οι κώδικες του εργαστηρίου ήταν αρκετά βοηθητικοί. Αν κάτι ήταν τροχοπέδη, η προθεσμία της εργασίας. Πιο συγκεκριμένα η κατανομή των καθηκόντων έγινε ως εξής:

Ηλιάκη Γεωργία (f2821903) - Contact Person

Η Ηλιάκη Γεωργία έχει σπουδάσει Μηχανικών Παραγωγής και Διοίκησης στα Χανιά. Εκείνη ήταν το contact person στη διάρκεια του project. Ήταν υπεύθυνη για τη σύνταξη του κώδικα, το έλεγχο της εν λόγω αναφοράς και της παρουσίασης.
Τηλ.: 6982010616
Email: georgia_ies@hotmail.com.

Κιντάκα Μιχαέλλα (f2821904)

Η Κιντάκα Μιχαέλλα έχει σπουδάσει Οργάνωση και Διοίκηση Επιχειρήσεων στο ΟΠΑ. Έχει εργαστεί στο κομμάτι του accounting & finance, ενώ τώρα εργάζεται ως data scientist. Ήταν υπεύθυνη για τον έλεγχο του κώδικα, για τη συγγραφή της εν λόγω αναφοράς και τη δημιουργία της παρουσίασης.

Στεφανοπούλου Δανάη (f2821913)

Η Στεφανοπούλου Δανάη έχει αποφοιτήσει με άριστα από το Οικονομικό του ΠΑΠΕΙ. Στο παρελθόν εργάστηκε ως accountant assistant ενώ τελειώνοντας το μεταπτυχιακό ξεκίνησε να εργάζεται ως Business Intelligence Analyst. Η Δανάη ήταν υπεύθυνη για τον έλεγχο του κώδικα, για τη συγγραφή της εν λόγω αναφοράς και τη δημιουργία της παρουσίασης.

Βιβλιογραφία

Αναφορά:

Harnad, Stevan (2008), "The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence", University of Southampton Institutional Repository

Jason Brownlee, 2017, "Gentle Introduction to the Adam Optimization Algorithm for Deep Learning"

Athena Research Center, "Natural Language Processing"

Adam Hayes, 2020, "Stratified Random Sampling"

Sai Teja, 10th June 2020, "Stop Words in NLP"

Kimberly Nevala, "The Machine Learning Primer", SAS Best Practices
https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/machine-learning-primer-108796.pdf

United Nations (UN), "Goal 5: Achieve gender equality and empower all women and girls"
https://www.un.org/sustainabledevelopment/gender-equality/

Sean Bleier, "NLTK's list of english stopwords", github
https://gist.github.com/sebleier/554280

Suptatim Haldar, 2019, "How to stop training a neural-network using callback?"
https://towardsdatascience.com/neural-network-with-tensorflow-how-to-stop-training-using-callback-5c8d575c18a9

Α. Λύκας, "Τεχνητά Νευρωνικά Δίκτυα" (Διαφάνειες), Παν. Ιωαννίνων
http://www.cs.uoi.gr/~arly/courses/nn/slides/K2.pdf

"Why ReLU is better than the other activation functions"
https://datascience.stackexchange.com/questions/23493/why-relu-is-better-than-the-other-activation-functions

"How to interpret classification report of scikit-learn?"
https://datascience.stackexchange.com/questions/64441/how-to-interpret-classification-report-of-scikit-learn

Andrew Thomas, "Recurrent neural networks and LSTM tutorial in Python and TensorFlow"
https://adventuresinmachinelearning.com/recurrent-neural-networks-lstm-tutorial-tensorflow/

David Israwi , "RNN Hyper parameters"
https://gist.github.com/DavidIsrawi/6c45744c12a4f8fc08bd5b8f7f9e06d8

Denny Britz, 2015, " Understanding Convolutional Neural Networks for NLP"
http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Techopedia, 2012, "Definition - What does *Tokenization* mean?"
https://www.techopedia.com/definition/13698/tokenization

"ΚΕΦΑΛΑΙΟ 4 –Μηχανική Μάθηση"
https://repository.kallipos.gr/bitstream/11419/3382/1/02_chapter_04.pdf

Κώδικας:

Saad Arshad, 2019, "Sentiment Analysis / Text Classification Using CNN (Convolutional Neural Network)"
https://towardsdatascience.com/cnn-sentiment-analysis-1d16b7c5a0e7

Jason Brownlee, 2019, "A Gentle Introduction to Sparse Matrices for Machine Learning"
https://machinelearningmastery.com/sparse-matrices-for-machine-learning

Jason Brownlee, 2019, "A Gentle Introduction to the Rectified Linear Unit (ReLU)"
https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/

DENNY BRITZ, 2015,"Understanding Convolutional Neural Networks for NLP"
http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

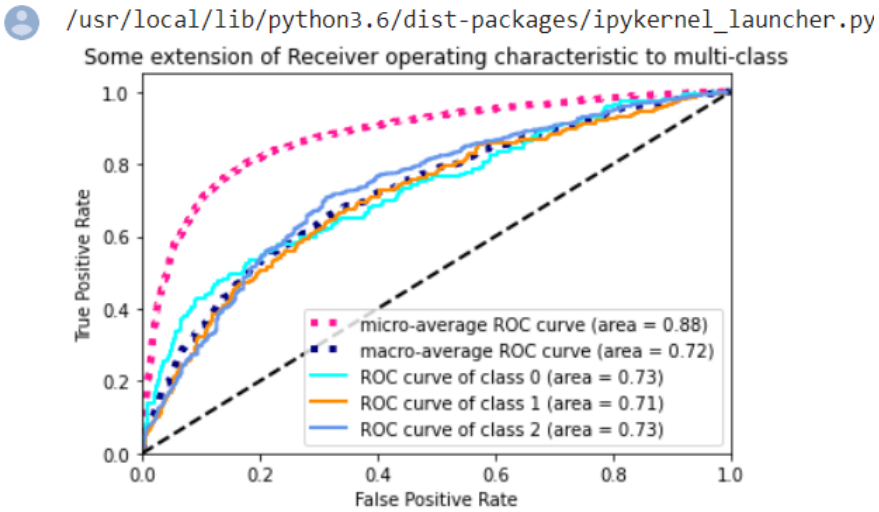
Nikolai Janakiev,"Practical Text Classification With Python and Keras"
https://realpython.com/python-keras-text-classification/

Jason Brownlee, 2019,"How to Save and Load Your Keras Deep Learning Model"
https://machinelearningmastery.com/save-load-keras-deep-learning-models/

Παράρτημα

1. Διάγραμμα AUC-ROC

Η απόδοση ενός αλγορίθμου Classification μπορεί επίσης να αναγνωριστεί με την καμπύλη AUC η οποία σχεδιαζεται για τα τρια πιθανά labels, καθώς και το micro και macro average ROC. Καλό θα ήταν μεταξύ των δυο να λάβουμε υπόψη περισσότερο το micro average ROC, το οποίο μάλιστα είναι και καλύτερο καθώς, εκείνος συγκεντρώνει τις συνεισφορές όλων των τάξεων για να υπολογίσει τη μέση μέτρηση λαμβάνοντας υπόψη την ανισορροπία που εντοπιστηκε στο διαφορετικό αριθμό labels. Αντιθέτως, το macro average θα υπολογίσει τη μέτρηση ανεξάρτητα για κάθε τάξη και στη συνέχεια θα λάβει τον μέσο όρο, θα λάβει υπόψη εξίσου όλες τις τάξεις.



ROC Curve MLP