ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανίχνευση Ρητορικής Μίσους με Χρήση του Γλωσσικού Μοντέλου BERT

(Hate Speech Detection Using BERT Language Model)

ΖΙΩΖΑΣ ΓΕΩΡΓΙΟΣ
ICSD: 15058

Επιβλέπων Καθηγητής:
ΕΥΣΤΑΘΙΟΣ ΣΤΑΜΑΤΑΤΟΣ

Μέλη εξεταστικής επιτροπής:
ΣΥΜΕΩΝΙΔΗΣ ΠΑΝΑΓΙΩΤΗΣ
ΚΩΣΤΟΥΛΑΣ ΘΕΩΔΟΡΟΣ

ΣΑΜΟΣ, ΟΚΤΩΒΡΙΟΣ, 2021

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

Η επιτυχής εκπόνηση της παρούσας διπλωματικής εργασίας, παρά τον προσωπικό μόχθο και χρόνο αφοσίωσης, αποτελεί ταυτόχρονα ένα προϊόν μιας ευρύτερης συνεργασίας και υποστήριξης τόσο από το ακαδημαϊκό όσο και από το οικογενειακό και φιλικό περιβάλλον.

Σε ακαδημαϊκό επίπεδο, θα ήθελα να εκφράσω τις ευχαριστίες μου στον επιβλέποντα καθηγητή Σταματάτο Ευστάθιο, για την ευκαιρία που μου έδωσε να ασχοληθώ σε βάθος με τον κλάδο της Τεχνητής Νοημοσύνης και της ανίχνευσης Ρητορικής Μίσους. Επίσης, τον ευχαριστώ τόσο για την στοχευμένη και εποικοδομητική καθοδήγηση, όσο και για τις γνώσεις που μου παρείχε καθ' όλη τη διάρκεια εκπόνησης της παρούσας εργασίας.

Σε προσωπικό επίπεδο, δεν θα μπορούσα να παραλείψω τις ευχαριστίες προς τους φίλους και συμφοιτητές μου, Γιώργο, Σωτήρη και Μάνο για την αμέριστη στήριξη τους καθ' όλη την διάρκεια του ταξιδιού αυτού.

Τέλος, το μεγαλύτερο ευχαριστώ απευθύνεται στην οικογένεια μου, η οποία με στήριξε απο την αρχή και συνεχίζει να με στηρίζει στο ταξίδι της γνώσης.

ΖΙΩΖΑΣ ΓΕΩΡΓΙΟΣ

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Η σελίδα αυτή είναι σκόπιμα λευκή

Πίνακας περιεχομένων

Λίστα Σχημάτων

Λίστα Πινάκων

Ακρωνύμια

DL	Deep Learning
NLP	Natural Language Processing
AI	Artificial Intelligence
ML	Machine Learning
RNN	Recurrent Neural Networks
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Networks
GRU	Gated Recurrent Unit
B.E.R.T	Bidirectional Encoder Representations
biLSTM	Bi-directional Long Short Term Memory
Lr	Learning Rate
ANI	Artificial Narrow Intelligence
AGI	Artificial General Intelligence
ASI	Artificial Super Intelligence

Η σελίδα αυτή είναι σκόπιμα λευκή

Περίληψη

Ο τεράστιος όγκος πληροφοριών που παράγεται καθημερινά στο διαδίκτυο και η μαζική χρήση μέσων κοινωνικής δικτύωσης έχουν δημιουργήσει πρωτόγνορα φαινόμενα ρητορικής μίσους. Η πρόσφατη ραγδαία εξέλιξη του τομέα της τεχνητής νοημοσύνης και ιδιαίτερα των αρχιτεκτονικών βαθιάς μάθησης όπως η Transformer, έχει βοηθήσει σημαντικά την διαδικασία αυτόματης ταξινόμησης κειμένων με βάση το περιεχόμενο τους. Στην έρευνα μας ξεκινάμε αναλύοντας τον τομέα της τεχνητής νοημοσύνης και της ταξινόμησης ρητορικής μίσους. Στην συνέχεια επικεντρονώμαστε στην εξέταση συνόλων δεδομένων ρητορικής μίσους και προτείνουμε μια μεθοδολογία επιλογής αυτών με βάση την αντικειμενικότα. Στην πειραματική μας προσέγγιση, μετά την εκτενή επεξεργασία των δεδομένων, υλοποιούμε την αρχιτεκτονική Transformer B.E.R.T με σκοπό την δημιουργία ταξινομητή πολλαπλών ετικετών 6 κλάσεων στόχων. Τα αποτελέσματα μας παρουσιάζουν τις καλύτερες επιδόσεις στο συγκεκριμένο σύνολο δεδομένων και δίνουν κατευθύνσεις για μελλοντικές έρευνες.

Λέξεις Κλειδιά: *Ταξινόμηση Κειμένου, Ρητορική Μίσους, Βαθιά Μάθηση, Τεχνητή Νοημοσύνη, Ανάλυση Φυσικής Γλώσσας, B.E.R.T, Νευρωνικά Δίκτυα*

Abstract

The huge amount of information produced daily on the internet and the massive use of social media have created unprecedented phenomena of hate speech. The recent rapid development of the field of artificial intelligence and especially of deep learning architectures such as Transformer, has significantly helped the process of automatic text classification based on their content. In our research we start by analyzing the field of artificial intelligence and the classification of hate speech. We then focus on examining hate speech data sets and propose a methodology for selecting them based on objectivity. In our experimental approach, after extensive data processing, we implement the Transformer B.E.R.T architecture in order to create a multi-tag classifier of 6 target classes. Our results show the best performance in this data set and give directions for future research.

Keywords: *Text Classification, Hate Speech, Deep Learning, Artificial Intelligence, Natural Language Processing, B.E.R.T, Neural Networks*

Δομή της διπλωματικής

Στο **1ο** κεφάλαιο εκτός απο τους στόχους της διπλωματικής εργασίας παρουσιάζουμε και μια εκτενή ανάλυση των κλάδων της ανάλυσης φυσικής γλώσσας, τεχνητής νοημοσύνης, μηχανικής μάθησης και βαθιάς μάθησης. Στο **2ο** κεφάλαιο παραθέτουμε τις θεμελιακές τεχνικές ταξινόμησης και αναλύουμε εκτενώς την ήδη υπάρχουσα επιστημονική βιβλιογραφία της προσπάθειας ταξινόμησης κειμένου ρητορικής μίσους. Στο **3ο** κεφάλαιο επικεντρονώμαστε στα σύνολα δεδομένων ρητορικής μίσους και προτείνουμε μια μεθοδολογία του τρόπου επιλογής τους αναλόγως με τον σκοπό της κάθε έρευνας. Στο **4ο** κεφάλαιο παρουσιάζουμε την προσέγγιση μας και τα βήματα ως προς την υλοποίηση της. Στο **5ο** κεφάλαιο παρουσιάζουμε αναλυτικά τα πειραματικά αποτελέσματα της έρευνας μας και τα συγκρίνουμε με αυτά της συγγενικής επιστημονικής βιβλιογραφίας. Στο **6ο** κεφάλαιο παραθέτουμε τα συμπεράσματα μας, αναλύουμε τα αποτελέσματα των πειραμάτων μας και προτείνουμε μελλοντικές ερευνητικές κατευθύνσεις.

1

Εισαγωγή

Στον 21ο αιώνα, ο οποίος ονομάζεται και «εποχή της πληροφορίας», τα smartphone και η χρήση του διαδικτύου έχουν γίνει αναπόσπαστα στοιχεία της ζωής μας. Η ψηφιοποίηση έχει ριζώσει σε κάθε πτυχή της καθημερινής μας δραστηριότητας, μεταμορφώνοντας τον τρόπο με τον οποίο επικοινωνούμε μεταξύ μας. Η ελεύθερη έκφραση των απόψεων σε αυτή την νέα εποχή των κοινωνικών δικτύων είναι αναμφισβήτητα παραγωγική για την κοινωνική ανάπτυξη, αλλά ταυτόχρονα δημιουργεί μια επιβλαβής παρουσία φαινομένων τοξικότητας και εξάπλωσης μίσους.

Σε πολυάριθμες μελέτες που πραγματοποιήθηκαν από το Pew Research Center με την πάροδο του χρόνου, μπορούμε να δούμε ότι στις αρχές του 2005 μόνο το 5% των Αμερικανών ενηλίκων χρησιμοποιούσε τουλάχιστον μία από τις κύριες πλατφόρμες κοινωνικής δικτύωσης. Ωστόσο, τα στατιστικά αυτά αυξήθηκαν δραματικά το 2011 και το 2021 με χρήση 50% και 72%, αντίστοιχα.[1] [2] [3]

Παρά την τα πολλά θετικά που μας προσφέρουν τα μέσα κοινωνικής δικτύωσης, παρουσιάζουν και κάποιες προκλήσεις. Αυτές οι προκλήσεις εκδηλώνονται με διάφορες παραλλαγές, όπως εκφοβισμό, ρητορική μίσους, προσβλητική και υβριστική γλώσσα. Η ρητορική μίσους αναγνωρίζεται ως ένα από τα κύρια προβλήματα που βασανίζουν τα μέσα κοινωνικής δικτύωσης και το οικοσύστημα του διαδικτύου στο σύνολό του. [4] [5]

Η διαδικτυακή ανίχνευση περιεχομένου μίσους είναι μια εγγενώς πολύπλοκη και δύσκολη εργασία. Ωστόσο, έχει λάβει σημαντική προσοχή από τον ακαδημαϊκό χώρο και τις κυβερνήσεις τα τελευταία χρόνια ώστε να παρουσιάσει μια λύση και να μετριάσει το φαινόμενο.

Οι αρχιτεκτονικές βαθιάς μάθησης έχουν δείξει εξαιρετικά αποτελέσματα στην ταξινόμηση της ρητορικής μίσους, ειδικά με την εμφάνιση αρχιτεκτονικών μετασχηματιστών (Transformers) τα τελευταία χρόνια.[6][7][8][9] Ωστόσο, ακόμη και με αυτήν την τεράστια πρόοδο των αρχιτεκτονικών και τα εξαιρετικά αποτελέσματα που φαίνεται να επιτυγχάνουν, υπάρχουν ακόμη ορισμένα σοβαρά εμπόδια που πρέπει να αντιμετωπιστούν ώστε να επιλυθεί αποτελεσματικά η ταξινόμηση της ρητορικής μίσους. [10]

Τα κεντρικά προβλήματα είναι τα εξής:

- **Σαφής ορισμός ετικετών σε σύνολα δεδομένων ρητορικής μίσους** [11] [12]

Μια σταθερή απαίτηση προς τους εξεταστές και δημιουργούς δεδομένων, είναι να υπάρχουν σαφείς ορισμοί στο σύνολο δεδομένων, διαχωρίζοντας τη ρητορική μίσους από άλλους τύπους προσβλητικής γλώσσας και την υβριστική από τη προσβλητική γλώσσα. Η επίτευξη συναίνεσης σχετικά με τον ετυμολογικό ορισμό αυτών των φαινομένων μπορεί να επεκτείνει τις γνώσεις μας στον τομέα γρηγορότερα.

- **Γενίκευση Αποτελεσμάτων**

Μας δείχνει πώς συμπεριφέρεται το μοντέλο όταν δοκιμάζεται σε νέα δεδομένα στα οποία δεν έχει εκπαιδευτεί. Αποτελεί μια βασική μέτρηση η οποία εστιάζει στο να δείξουμε πώς ένα εκπαιδευμένο μοντέλο μπορεί να είναι ανταγωνιστικό σε μελλοντικά κείμενα. Όπως δηλώνει η έρευνα [13] τα περισσότερα μοντέλα τα οποία φαίνεται να πετυχαίνουν άριστες επιδόσεις έχουν υπερεκτιμήθεί σε πολύ μεγάλο βαθμό.

- **Μεροληψία δειγματοληψίας**[14]

Η μεροληψία δειγματοληψίας αποτελεί μια τεράστια πρόκληση που μπορεί να αλλάξει τα αποτελέσματα μίας μελέτης και να επηρεάσει την εγκυρότητα της διερευνητικής διαδικασίας. Εμφανίζεται όταν δεν υφίσταται μια δίκαιη ή ισορροπημένη παρουσίαση των απαιτούμενων δειγμάτων δεδομένων κατά τη διεξαγωγή της έρευνας. Η τυχαία δειγματοληψία καθιστά τα σύνολα δεδομένων επιρρεπή σε προκατάληψη.

- **Φυλετική Προκατάληψη**[15]

Η κοινωνία έχει πάντα μια προκατάληψη όσον αφορά τον προφορικό λόγο. Αυτό το φαινόμενο καταλήγει σε μια κατάσταση όπου οι μειονοτικές ομάδες υποεκπροσωπούνται σε διαθέσιμα δεδομένα ή/και σχολιαστές δεδομένων, προκαλώντας προκατάληψη εναντίον τους όταν εκπαιδεύονται μοντέλα σε αυτά.

- **Προκατάληψη Σχολιασμού**

Οι προκαταλήψεις των ατόμων που σχολιάζουν τα σύνολα δεδομένων καταλήγουν σε συμπεράσματα με βάση τις δικές τους προκαταλήψεις, οδηγώντας σε λανθασμένες προβλέψεις, χαμηλή ακρίβεια και πολύ χαμηλές βαθμολογίες γενίκευσης. Όπως διαπιστώθηκε απο την έρευνα[16]: Μόνο περίπου τα δύο τρίτα των υφιστάμενων συνόλων δεδομένων αναφέρουν την βαθμολογία συμφωνίας μεταξύ των σχολιαστών.

Στην διπλωματική μας, προσπαθούμε να αντιμετωπίσουμε τα περισσότερα από αυτά τα προβλήματα χρησιμοποιώντας το πολυγλωσσικό σύνολο δεδομένων M.L.M.A. [17] το οποίο περιέχει υψηλή ποιότητα σχολιασμού, αποφεύγει τη φυλετική προκατάληψη δίνοντας συγκεκριμένους κανόνες στους σχολιαστές, παρουσιάζει σαφείς ορισμούς ετικετών και δεν περιέχει δεδομένα από ήδη υπάρχοντα σύνολα δεδομένων που έχουν αποδειχθεί προκατειλημμένα στο παρελθόν.[18] Επίσης προτείνουμε μια μεθοδολογία επιλογής συνόλων δεδομένων ρητορικής μίσους με βάση τον τύπο ταξινόμησης και λοιπά χαρακτηριστικά.

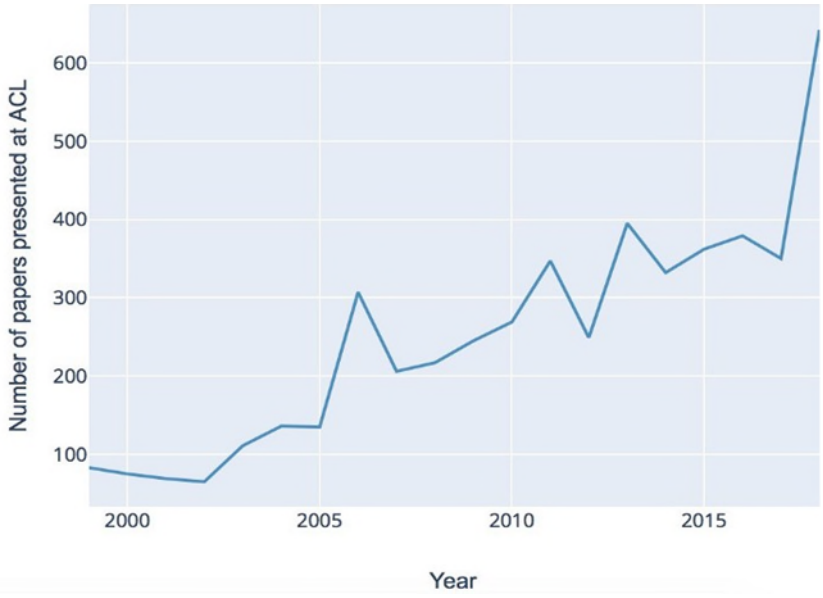
Στην πειραματική φάση εφαρμόζουμε την B.E.R.T. αρχιτεκτονική μετασχηματιστή για την πρόβλεψη κατηγοριών ρητορικής μίσους, με 6 ετικέτες εξόδου (προσβλητική, ασέβεια, μίσος, φόβο από άγνοια, καταχρηστική ή φυσιολογική) στην Αγγλική γλώσσα. Στον τύπο ταξινόμησης χρησιμοποιούμε ταξινόμηση κειμένου με πολλαπλές ετικέτες.

1.1 Ανάλυση Φυσικής Γλώσσας – Natural Language Processing

Η επεξεργασία φυσικής γλώσσας (NLP) αποτελεί αντικείμενο της επιστήμης των υπολογιστών - συγκεκριμένα, του κλάδου της τεχνητής νοημοσύνης (AI) - που αφορά την ικανότητα των υπολογιστών να κατανοούν/επεξεργάζονται το γραπτό κείμενο και τις προφορικές λέξεις με τον ίδιο τρόπο που μπορούν και οι άνθρωποι. Στον κλάδο της επεξεργασία φυσικής γλώσσας η υπολογιστική γλωσσολογία -η μοντελοποίηση της γλώσσας βασισμένη σε κανόνες- συνδυάζεται με στατιστικά μοντέλα, μοντέλα μηχανικής μάθηση και αρχιτεκτονικές βαθιάς μάθησης. Αυτές οι τεχνολογίες, όταν χρησιμοποιούνται μαζί, επιτρέπουν στους υπολογιστές να επεξεργάζονται την ανθρώπινη γλώσσα με τη μορφή δεδομένων κειμένου ή ομιλίας και να «κατανοούν» το πλήρες νόημά της, συμπεριλαμβανομένης της πρόθεσης και του συναισθήματος του ομιλητή ή του συγγραφέα.

Η Ένωση Υπολογιστικής Γλωσσολογίας (ACL) αποτελεί την κορυφαία διεθνή επιστημονική ομάδα, η οποία ασχολείται με υπολογιστικά προβλήματα που αφορούν την ανθρώπινη γλώσσα, πεδίο που συχνά αναφέρεται είτε ως υπολογιστική γλωσσολογία είτε ως επεξεργασία φυσικής γλώσσας (NLP). Σύμφωνα με τις μελέτες που έχουν δημοσιευθεί στην πλατφόρμα τους σχετικά με τον αριθμό νέων μελετών ανα τα χρόνια, παρατηρούμε πως ο συγκεκριμένος κλάδος διανύει ραγδαία αύξηση.

Στα επόμενα κεφάλαια θα αναλύσουμε εκτενώς τα πιο σημαντικά μοντέλα που χρησιμοποιούνται στον τομέα της επεξεργασίας φυσικής γλώσσας και τις διαφορετικές χρήσεις τους.



Πηγή: towardsdatascience.com/major-trends-in-nlp-a-review-of-20-years-of-acl-research-56f5520d473

Εικόνα 1: Ποσότητα Ερευνητικών Δημοσιεύσεων του ACL ανά Χρόνο (2000 – 2019)

1.2 Τεχνητή Νοημοσύνη – Artificial Intelligence

Η τεχνητή νοημοσύνη αναφέρεται στην ικανότητα δημιουργίας ευφυών συστημάτων ή εφαρμογών λογισμικού αυτομάθησης που προσομοιώνουν τα ανθρώπινα χαρακτηριστικά του εγκεφάλου μας. Η ικανότητα της τεχνητής νοημοσύνης να ξεπερνά τις ανθρώπινες δυνατότητες όσον αφορά την ανακάλυψη γνώσης έχει συγκεντρώσει το ενδιαφέρον της βιομηχανίας και των ερευνητικών οργανισμών σε όλο τον κόσμο. Ένα τεχνητά ευφυές σύστημα διαβάζει δεδομένα σε πραγματικό χρόνο, αξιολογεί το επιχειρηματικό πλαίσιο και ανταποκρίνεται κατάλληλα. Η τεχνητή νοημοσύνη (AI) έχει γίνει πλέον σημαντικό και αναπόσπαστο κομμάτι της τεχνολογίας των πληροφοριών καθώς πλέον υποστηρίζεται και απο επιτυχής και εξιδικευμένες επιστημονικές μελέτες.

Στα μέσα της δεκαετίας του 1950, έξι χρόνια αφότου ο Άλαν Τούρινγκ δημοσίευσε την εργασία του για τη δημιουργία μηχανών σκέψης [19], ο John McCarthy εισήγαγε τον όρο «Τεχνητή Νοημοσύνη» τον οποίο όρισε ως «την επιστήμη και την μηχανική κατασκευής ευφυών μηχανών». [20]

Τα μεγαλύτερα προβλήματα στην τεχνητή νοημοσύνη περιλαμβάνουν την κωδικοποίηση και τον προγραμματισμό υπολογιστών για συγκεκριμένες λειτουργίες όπως:

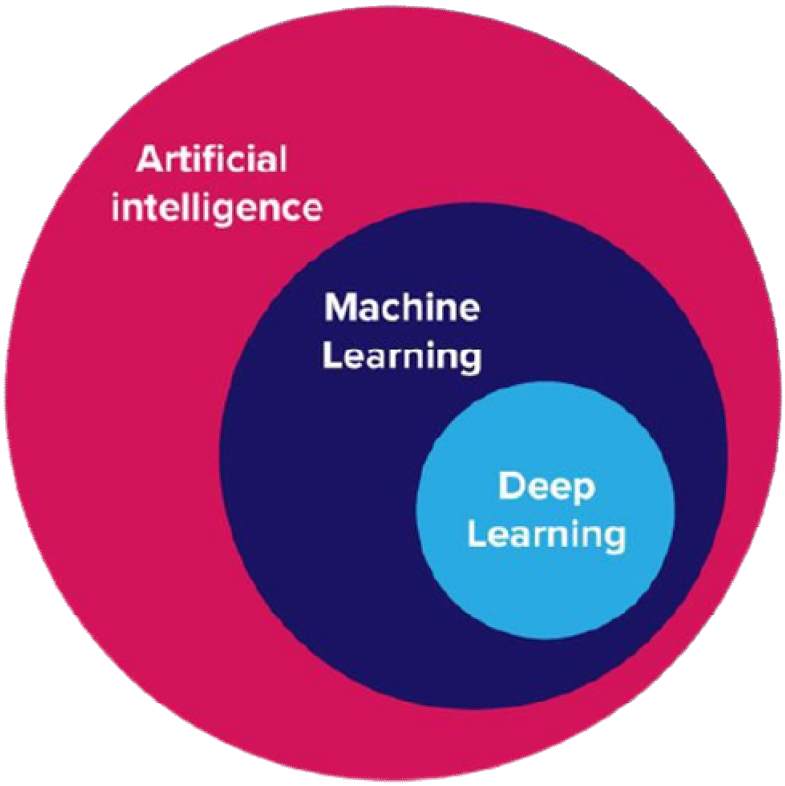
- απόκτηση γνώσης απο δεδομένα
- επεξεργασία του λόγου
- επίλυση καθημερινών προβλημάτων
- απόκτηση αντίληψης

Η Τεχνολογία Γνώσης βρίσκεται στον πυρήνα της έρευνας για την τεχνητή νοημοσύνη. Οι μηχανές μπορούν να λειτουργήσουν και να αντιδράσουν όπως οι άνθρωποι μόνο εάν τους δοθούν επαρκείς πληροφορίες για τον εκάστοτε τομέα και τον κόσμο. Κατά συνέπεια, η τεχνητή νοημοσύνη πρέπει να έχει πρόσβαση σε όλες τις πληροφορίες σχετικά με αντικείμενα, κατηγορίες, χαρακτηριστικά και σχέσεις μεταξύ όλων των λειτουργιών ενός συστήματος, προκειμένου το μηχάνημα να εφαρμόσει αποτελεσματικά το Knowledge Engineering και να επιτύχει παρόμοια ή και καλύτερα αποτελέσματα απο τον άνθρωπο. Το έργο της ενσωμάτωσης της κοινής λογικής, της λήψης αποφάσεων, της συλλογιστικής και της επίλυσης προβλημάτων σε μηχανές αποτελεί μεγάλη πρόκληση ακόμη και σήμερα.

Η ταχύτητα ανάπτυξης του τομέα της τεχνητής νοημοσύνης και τον αλγορίθμων υλοποίησης της, δημιούργησε διαφορετικούς κλάδους όπου ο καθένας αποσκοπεί σε συγκεκριμένες εφαρμογές και τεχνικές. Οι δύο βασικότεροι εξ αυτών είναι:

- Τομέας Μηχανικής Μάθησης
- Τομέας Βαθιάς Μάθησης

Οι παραπάνω κλάδοι τους οποίους θα μελετήσουμε αναλυτικά στην συνέχεια, είναι στενά συνδεδεμένοι με την τεχνητή νοημοσύνη καθώς ο καθένας αποσκοπεί σε διαφορετικού είδους προβλήματα. Η τεχνητή νοημοσύνη αποτελεί την κεφαλή-υπερσύνολο του τομέα με την μηχανική μάθηση να αποτελεί υποσύνολο της και την βαθιά εκμάθηση να αποτελεί υποσύνολο της μηχανικής όπως φαίνεται στην Εικόνα 1.



Πηγή: https://serokell.io/blog/ai-ml-dl-difference

Εικόνα 2: Η Τεχνητή Νοημοσύνη ως υπερσύνολο της Μηχανικής Μάθησης και της Βαθιάς Μάθησης

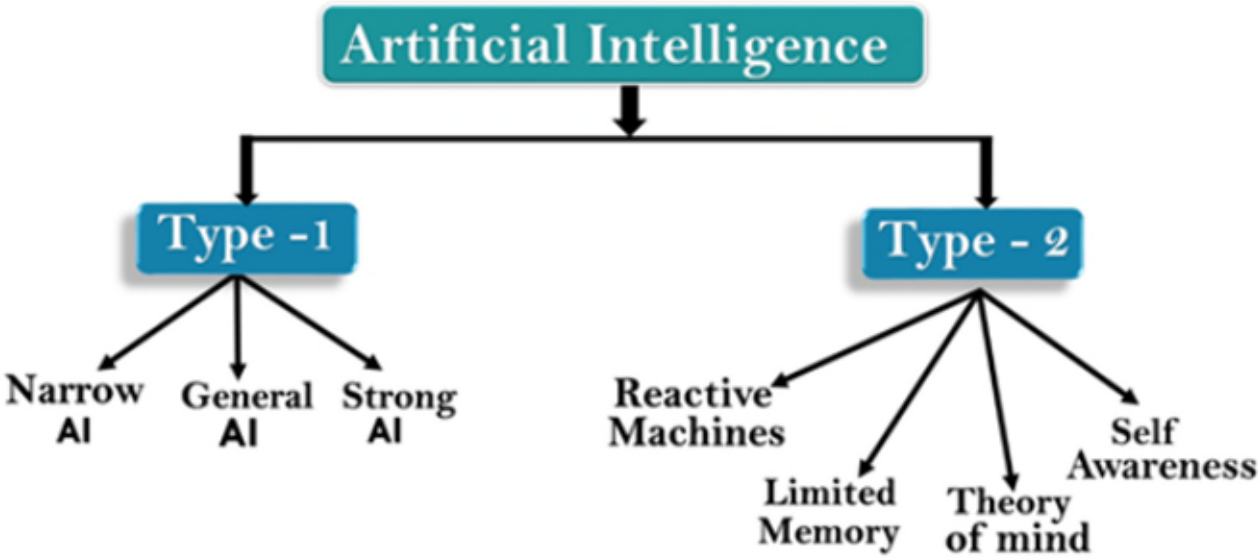
Στις παρακάτω υποενότητες θα ανακαλύψουμε τις κατηγορίες, εφαρμογές αλλα και δυσκολίες που αντιμετωπίζει η τεχνητή νοημυσύνη καθώς και τον ρόλο που κατέχει η ηθική στα πλαίσια αυτής.

1.2.1 Κατηγορίες

Το θεμελιώδες χαρακτηριστικό για την ταξινόμηση της τεχνητής νοημοσύνης είναι η ικανότητά της να αναπαράγει ενέργειες που έρχονται κοντά ή ακόμη και ξεπερνούν τις δυνατότητες του ανθρώπου. Θεωρητικά, η τεχνητή νοημοσύνη μπορεί να χωριστεί σε δύο κύριες κατηγορίες, με επίκεντρο την ικανότητά της να μιμείται τις λειτουργίες του ανθρώπινου εγκεφάλου.

Η πρώτη κατηγορία, "**Βασισμένη στις δυνατότητες**" της τεχνητής νοημοσύνης έναντι της ανθρώπινης νοημοσύνης, είναι πιο διαδεδομένη στη βιομηχανία της τεχνολογίας σήμερα.

Η δεύτερη κατηγορία, "**Βασισμένη στη λειτουργικότητα**" ταξινομεί την τεχνητή νοημοσύνη με βάση την ομοιότητά της με τον ανθρώπινο εγκέφαλο και την ικανότητά της να σκέφτεται και να αισθάνεται σαν ανθρώπινη οντότητα.



Πηγή: www.javatpoint.com/types-of-artificial-intelligence

Εικόνα 3: Οι δύο γενικές κατηγορίες στις οποίες χωρίζεται ο κλάδος της Τεχνητής Νοημοσύνης.

Στην συνέχεια, θα αναλύσουμε τις παραπάνω δύο κατηγορίες και τις υποκατηγορίες τους ώστε να γίνει κατανοητή η δομή και χρήση τους.

1.2.1.1 Τύπος I - Βασισμένη στις Δυνατότητες

Στενή Τεχνητή Νοημοσύνη (ANI)

Αυτή η κατηγορία περιλαμβάνει όλες τις υπάρχουσες εφαρμογές AI που βλέπουμε γύρω μας. Το ANI περιλαμβάνει ένα σύστημα τεχνητής νοημοσύνης που, όπως και οι άνθρωποι, μπορεί να εκτελέσει κάποια δομημένα συγκεκριμένα καθήκοντα. Από την άλλη πλευρά, αυτά τα μηχανήματα δεν μπορούν να εκτελέσουν ενέργειες για τις οποίες δεν έχουν προγραμματιστεί εκ των προτέρων, επομένως δεν μπορούν να ολοκληρώσουν ένα έργο χωρίς προηγούμενη γνώση. Το σύστημα αυτό είναι ένα μείγμα από όλα τα αντιδραστικά και περιορισμένα AI μνήμης. Αυτός ο τομέας της τεχνητής νοημοσύνης περιλαμβάνει τους αλγόριθμους που χρησιμοποιούμε στον σημερινό κόσμο για να κάνουμε περίπλοκες Μοντελοποίησεις Πρόβλεψεων (Predictive Modeling).

Τεχνητή Γενική Νοημοσύνη (AGI)

Το AGI μπορεί να εκπαιδεύσει, να μάθει, να κατανοήσει και να εκτελέσει εργασίες με τον ίδιο τρόπο που μπορεί να το κάνει και ένας άνθρωπος. Αυτά τα συστήματα έχουν πολυλειτουργικές δυνατότητες που καλύπτουν πολλούς τομείς. Επιπλέον, αυτά τα συστήματα θα είναι πιο ευκίνητα, θα ανταποκρίνονται και θα αυτοσχεδιάζουν με τον ίδιο τρόπο που κάνει ένας άνθρωπος όταν έρχεται αντιμέτωπος με άγνωστες καταστάσεις. Παρόλο που δεν υπάρχει πραγματικό παράδειγμα τέτοιου τύπου τεχνητής νοημοσύνης, έχει σημειωθεί σημαντική πρόοδος σε αυτόν τον τομέα.

Τεχνητή Δυνατή Νοημοσύνη (ASI)

Το ζενίθ της προόδου της τεχνητής νοημοσύνης θα είναι η Τεχνητή Σούπερ Νοημοσύνη. Το ASI θα είναι ο πιο εξελιγμένος τύπος ευφυΐας που δημιουργήθηκε ποτέ στον Πλανήτη. Λόγω των ασύγκριτα ανώτερων δυνατοτήτων επεξεργασίας δεδομένων, μνήμης και λήψης αποφάσεων, θα έχει την δυνατότητα να εκτελέσει όλες τις εργασίες καλύτερα από τους ανθρώπους. Εν κατακλείδι, ορισμένοι ακαδημαϊκοί πιστεύουν ότι η εισαγωγή του ASI θα οδηγήσει τελικά στην "Τεχνολογική ιδιαιτερότητα" (Singularity Hypothesis). Πρόκειται για μια υποθετική κατάσταση όπου οι τεχνολογικές καινοτομίες φτάνουν σε ένα ανεξέλεγκτο όριο, οδηγώντας σε αφάνταστες αλλαγές τον ανθρώπινο πολιτισμό.

Παρ 'όλα αυτά, είμαστε πολύ μακριά από την επίτευξη αυτού του σταδίου, καθώς βρισκόμαστε μόνο στις πρώτες φάσεις της προηγμένης ανάπτυξης της τεχνητής νοημοσύνης. Σύμφωνα με τους επιστήμονες, αυτήν τι στιγμή «χαράζουμε» μόνο την επιφάνεια του πραγματικού δυναμικού της τεχνητής νοημοσύνης και ακόμη και για τους σκεπτικιστές, είναι πολύ νωρίς για να ανησυχούμε για την τεχνολογική ιδιαιτερότητα (Singularity Hypothesis).

1.2.1.2 Τύπος ΙΙ - Βασισμένη στην Λειτουργικότητα

Αντιδραστική Μηχανή – Reactive Machine

Η τεχνητή νοημοσύνη αυτού του τύπου είναι η πιο βασική και η παλαιότερη. Μιμείται την ικανότητα ενός ανθρώπου να αντιδρά σε εξωτερικές ενδείξεις. Ωστόσο, δεδομένου ότι αυτός ο τύπος τεχνητής νοημοσύνης στερείται μνήμης, δεν μπορεί να χρησιμοποιήσει προηγούμενη εμπειρία ή γνώση για τη βελτίωση των αποτελεσμάτων. Κατά συνέπεια, σε αντίθεση με την τεχνητή νοημοσύνη που βλέπουμε στις μέρες μας, αυτοί οι τύποι τεχνητής νοημοσύνης δεν μπορούν να μάθουν αυτοβούλως.

Στα τέλη της δεκαετίας του 1990, ο Deep Blue, ο υπερυπολογιστής σκακιού της IBM, είναι κυρίως γνωστός για την νίκη του απέναντι στον παγκόσμιο γκρανμέστερ Garry Kasparov. Το Deep Blue μπορούσε να αναγνωρίσει τα διαφορετικά τύπου πιόνια και πώς αυτά κινούνται στην σκακιέρα. Στη συνέχεια, μπορούσε να δει όλες τις διαθέσιμες κινήσεις και να επιλέγει τη βέλτιστη δυνατή κίνηση με βάση την επιλογή του αντιπάλου. Ωστόσο, δεδομένου ότι αυτά τα μηχανήματα δεν έχουν δική τους μνήμη, δεν μπορούν να μάθουν από τις προηγούμενες ενέργειές τους.

Θεωρία Περιορισμών – Limited Theory

Αυτό το είδος τεχνητής νοημοσύνης, παρόμοιο με το Reactive Machines, διαθέτει δυνατότητες μνήμης, επιτρέποντάς του να χρησιμοποιεί προηγούμενες πληροφορίες και εμπειρία για να λαμβάνει καλύτερες αποφάσεις στο μέλλον. Αυτή η κατηγορία περιλαμβάνει την πλειοψηφία των εφαρμογών που παρατηρούμε στην καθημερινή μας ζωή. Αυτές οι εφαρμογές τεχνητής νοημοσύνης μπορούν να εκπαιδευτούν χρησιμοποιώντας μια τεράστια ποσότητα δεδομένων εκπαίδευσης που αποθηκεύονται σε ένα μοντέλο αναφοράς στη μνήμη τους.

Πολλά αυτόνομα οχήματα, για παράδειγμα, έχουν τεχνολογία θεωρίας περιορισμών. Αποθηκεύουν δεδομένα όπως η τοποθεσία GPS, οι κινήσεις κοντά στο αυτοκίνητο, το μέγεθος/η φύση των εμποδίων και εκατό άλλες μορφές δεδομένων για να καταφέρνουν να οδηγούν το ίδιο αποτελεσματικά με έναν άνθρωπο.

Θεωρία του Νού – Theory of Mind

Το επόμενο επίπεδο της A.I., με ελάχιστη έως καθόλου επίδραση στην καθημερινή μας ζωή προς το παρόν, είναι η Θεωρία του Νού. Αυτά τα είδη A.I. βρίσκονται συνήθως στο στάδιο "Work in Progress" και διατίθενται μόνο σε ερευνητικά εργαστήρια. Μόλις αναπτυχθούν, το A.I. θα έχει πλήρη κατανόηση του ανθρώπινου μυαλού, συμπεριλαμβανομένων των αναγκών, των συμπαθειών, των συναισθημάτων και των εγκεφαλικών λειτουργιών του. Το A.I. θα είναι ικανό να τροποποιήσει την επόμενη κίνηση του με βάση την κατανόησή του για τον ανθρώπινο εγκέφαλο και τις ιδιοτροπίες τους.

Για παράδειγμα, σε παλαιότερη του έρευνα ο Patrick Winston παρουσίασε ένα πρωτότυπο ρομπότ που μπορεί να περπατήσει σε έναν μικρό διάδρομο καθώς άλλα ρομπότ πλησιάζουν από

την αντίθετη κατεύθυνση. Το A.I. μπορεί να προβλέψει τις κινήσεις άλλων ρομπότ και μπορεί να στρίψει δεξιά, αριστερά ή σε οποιαδήποτε άλλη κατεύθυνση ώστε να αποφύγει μια πιθανή σύγκρουση με τα εισερχόμενα ρομπότ. Αυτό το Robot, σύμφωνα με τον Winston, παίρνει αποφάσεις με βάση την «κοινή λογική» του για το πώς θα συμπεριφέρεται στα άλλα ρομπότ.

Αυτοσυνείδητη Τεχνητή Νοημοσύνη – Self Awareness

Η αυτογνωσία είναι η τελευταία φάση της τεχνητής νοημοσύνης. Αυτά τα συστήματα AI είναι ικανά να κατανοήσουν και να προκαλέσουν ανθρώπινα συναισθήματα και στην συνέχεια να δημιουργήσουν δικά τους. Αυτό το είδος τεχνητής νοημοσύνης απέχει δεκαετίες, αν όχι αιώνες, από το να γίνει πραγματικότητα. Ο Έλον Μασκ και άλλοι σκεπτικιστές της τεχνητής νοημοσύνης είναι επιφυλακτικοί με αυτήν την μορφή της. Επειδή μόλις μια Τεχνητή Νοημοσύνη αναπτύξει αυτογνωσία, μπορεί να εισέλθει στη λειτουργία Αυτοσυντήρησης, θεωρώντας την ανθρωπότητα ως πιθανή απειλή και επιδοιώκοντας την εξάλειψη της άμεσα ή έμμεσα.

1.2.2 Εφαρμογές

Η ενσωμάτωση της τεχνητής νοημοσύνης σε αλγόριθμους που χρησιμοποιούν και χειρίζονται δεδομένα σε πραγματικό χρόνο έχει πολλά οφέλη. Η τεχνητή νοημοσύνη πλέον κυριαρχεί σε διάφορους κλάδους όπου απαιτείται ανάγνωση και επεξεργασία δεδομένων σε πραγματικό χρόνο, όπως τα οικονομικά και η ιατρική.

Ψηφιακά Παιχνίδια – Gaming

Απαιτείται ανάλυση δεδομένων σε πραγματικό χρόνο σε στρατηγικά παιχνίδια όπως το Σκάκι, το Πόκερ και το Tic Tac Toe. Το σύστημα θα πρέπει να είναι σε θέση να εξετάσει μια ποικιλία πιθανών ενεργειών, να αξιολογήσει αυτές τις επιλογές και να λαμβάνει μια απόφαση βασισμένη σε ευρετικές γνώσεις. Σε αυτά τα παιχνίδια στρατηγικής, η τεχνητή νοημοσύνη (AI) είναι ζωτικής σημασίας.

Επεξεργασία Φυσικής Γλώσσας – Natural Language Processing

Είναι ζωτικής σημασίας για τα μηχανήματα να κατανοήσουν τη γλώσσα διαφορετικών χρηστών για να λειτουργήσει αποτελεσματικά το λογισμικό. Το σύστημα θα πρέπει να μπορεί να προσαρμόζεται σε διαφορετικές γλώσσες, αλλά και σε διαφορετικές διαλέκτους και προφορές. Σε ορισμένες περιπτώσεις, η τεχνητή νοημοσύνη έχει αποδειχθεί εξαιρετικά χρήσιμη.

Εξιδικευμένα Συστήματα - Expert Systems

Η πρωταρχική λειτουργία ενός ευφυούς μηχανήματος είναι η λήψη αποφάσεων. Αυτές οι συσκευές απαιτούν λογισμικό που λαμβάνει δεδομένα ως είσοδο, τα ερμηνεύει, ζυγίζει πολλές διαφορετικές επιλογές και λαμβάνει μια απόφαση. Οι χρήστες μπορούν να λάβουν έτσι τεκμηριομένες αποφάσεις έχοντας πλέον μια σαφή εικόνα των δεδομένων.

Συστήματα όρασης – Vision Systems

Η εικόνα είναι ένας τύπος δεδομένων που είναι τόσο σημαντικός όσο και δύσκολος στην ερμηνεία. Κατά συνέπεια, ένα σύστημα με Ευφυΐα πρέπει να διαβάζει, να καταλαβαίνει, να ερμηνεύει και να κατανοεί οπτικές εισόδους προτού κρίνει με βάση αυτές.

1.2.3 Δυσκολίες και Προκλήσεις

Η Τεχνητή Νοημοσύνη εκτός απο τα θετικά της σημεία, αντιμετωπίζει και αρκετές δυσκολίες, όπως:

Υπολογιστική Δύναμη

Η μηχανική και η βαθιά μάθηση είναι τα θεμέλια της Τεχνητής Νοημοσύνης και απαιτούν έναν συνεχώς αυξανόμενο αριθμό πυρήνων και καρτών γραφικών - GPU για να λειτουργούν καλά. Έχουμε πλέον τις δεξιότητες για την εφαρμογή πλαισίων βαθιάς μάθησης σε διάφορους κλάδους, συμπεριλαμβανομένης της παρακολούθησης αστεροειδών, της υγειονομικής περίθαλψης, του κοσμικού εντοπισμού σώματος και πολλά άλλα. Οι αλγόριθμοι αυτοί απαιτούν την ικανότητα επεξεργασίας ενός υπερυπολογιστή, αλλά οι υπερυπολογιστές είναι πολύ ακριβοί.

Απόρρητο και Ασφάλεια Δεδομένων

Η διαθεσιμότητα δεδομένων και πόρων για την εκπαίδευση μοντέλων βαθιάς και μηχανικής μάθησης είναι ο πιο σημαντικός παράγοντας που πρέπει να ληφθεί υπόψη. Ωστόσο, ο τεράστιος αριθμός δεδομένων που παράγονται από εκατομμύρια χρήστες παγκοσμίως καθημερινά, δημιουργεί κίνδυνο να μην χρησιμοποιηθούν αυτά τα δεδομένα για κακόβουλους σκοπούς.

Κοστοφόρο

Οι μικρές και μεσαίες επιχειρήσεις αντιμετωπίζουν σημαντικές προκλήσεις στην εφαρμογή τεχνολογιών τεχνητής νοημοσύνης καθώς πρόκειται για μία δαπανηρή διαδικασία. Οι μεγάλες εταιρείες όπως το Facebook, η Apple, η Microsoft, η Google και η Amazon διαθέτουν μεγάλα χρήματα ποσά για την υιοθέτηση και ανάπτυξη τεχνολογιών τεχνητής νοημοσύνης.

Θέμα Ευθύνης

Η δημιουργία μιας εφαρμογής τεχνητής νοημοσύνης φέρει μεγάλη ευθύνη. Οποιοδήποτε άτομο πρέπει να αντιμετωπίσει το βάρος για τυχόν αστοχίες υλικού. Προηγουμένως, ήταν εύκολο να προσδιοριστεί εάν ένα περιστατικό προκλήθηκε από χρήστη, προγραμματιστή ή κατασκευαστή.

Σπανιότητα δεδομένων

Με τις μεγάλες εταιρείες όπως η Google, το Facebook και η Apple να κατέχουν ένα τεράστιο όγκο δεδομένων δημιουργείται ο κίνδυνος πιθανών προκαταλήψεων αφού μόνο ελάχιστοι κολλοσοί θα κατέχουν το μεγαλύτερο ποσοστό των δεδομένων. Τα δεδομένα χρησιμοποιούνται για να εκπαιδεύουν τα συστήματα να μαθαίνουν και να κάνουν προβλέψεις. Ορισμένες επιχειρήσεις προσπαθούν να αναπτύξουν νέες προσεγγίσεις και επικεντρώνονται στην ανάπτυξη μοντέλων τεχνητής νοημοσύνης που μπορούν να παρέχουν αξιόπιστα αποτελέσματα παρά την έλλειψη δεδομένων.

Το πρόβλημα της προκατάληψης

Ο όγκος των δεδομένων που χρησιμοποιούνται για την εκπαίδευση ενός συστήματος τεχνητής νοημοσύνης καθορίζει εάν τα αποτελέσματα θα είναι καλά ή κακά. Κατά συνέπεια, στο μέλλον, η ικανότητα απόκτησης καλών δεδομένων θα είναι το κλειδί για την ανάπτυξη καλών συστημάτων τεχνητής νοημοσύνης.

Ωστόσο, τα δεδομένα που συλλέγουν τώρα οι οργανισμοί σε καθημερινή βάση είναι αδύναμα και έχουν μικρή σημασία από μόνα τους. Τα περισσότερα είναι προκατειλημμένα και προσδιορίζουν μόνο τη φύση και τα χαρακτηριστικά μιας μικρής ομάδας ατόμων που μοιράζονται κοινά ενδιαφέροντα με βάση τη θρησκεία, την εθνότητα, το φύλο, την κοινότητα και άλλες φυλετικές προκαταλήψεις.

Μόνο με την ανακάλυψη αλγορίθμων που μπορούν να παρακολουθούν αποτελεσματικά αυτές τις προκλήσεις μπορεί να η Τεχνητή Νοημοσύνη να περάσει στο επόμενο επίπεδο.

1.2.4 Νοημοσύνη και Ηθική

Η ηθική είναι ένας ευρύς όρος που σχετίζεται με την ηθική συμπεριφορά ενός ατόμου σε διάφορες καθημερινές δραστηριότητες. Η ηθική στην τεχνητή νοημοσύνη, από την άλλη πλευρά, αναφέρεται στις ενέργειες των συστημάτων και των ρομπότ. Λογισμικό βασισμένο σε τεχνητή νοημοσύνη, όπως η μηχανή αναζήτησης της Google, οι διάφορες συστάσεις της Alexa, το YouTube, Netflix, τα αυτόνομα αυτοκίνητα και τα συστήματα αναγνώρισης προσώπου, είναι όλα πλέον μέρος της καθημερινής μας ζωής.

Είναι απλό να κατανοήσουμε τη ροή και εσωτερική λειτουργία ορισμένων αλγορίθμων, αλλά δεν είμαστε σε θέση να εξηγήσουμε επακριβώς πώς μαθαίνουν οι αλγόριθμοι βαθιάς μάθησης. Τώρα έχουμε συνεχώς εξελισσόμενους αλγόριθμους που δέχονται τεράστιους όγκους δεδομένων και μαθαίνουν ασταμάτητα. Έτσι είναι δύσκολο να πούμε με ακρίβεια σε ποιές παραμέτρους βασίζονται τα ευρήματα που παρουσιάζουν.

Ως άνθρωποι, παίρνουμε αποφάσεις με βάση τις προσωπικές μας προτιμήσεις, όπως το ντύσιμο, το στυλ, το φαγητό και τους τύπους ανθρώπων με τους οποίους θέλουμε να είμαστε φίλοι. Λαμβάνουμε υπόψη τις αρετές και τα μειονεκτήματα κατά την επιλογή ενός συντρόφου ζωής Ως αποτέλεσμα, η ανθρώπινη ηθική είναι πολυδιάστατη και διαφορετικοί για πολλούς ανθρώπους.

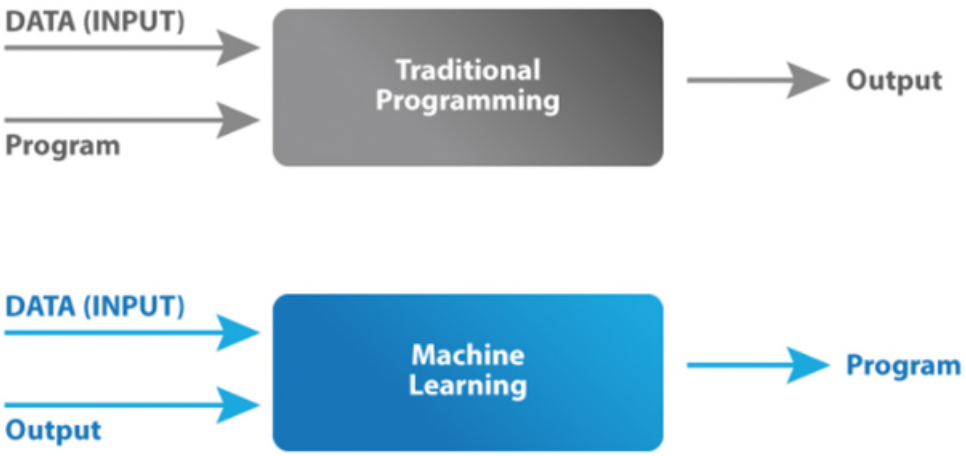
Αυτές οι ιδέες οδήγησαν στην έννοια της ηθικής της τεχνητής νοημοσύνης, η οποία καθορίζει τις ηθικές αξίες ενός ατόμου ή μιας ομάδας που ελέγχουν τη συμπεριφορά αυτών των συστημάτων ή ανθρώπινων ρομπότ AI. Θα μπορούσαμε να ισχυριστούμε ότι η Τεχνητή Νοημοσύνη δεν είναι μόνο ένα κομμάτι της τεχνολογίας. Έχει επιπτώσεις και στην καθημερινή μας ζωή. Επομένως, η καθιέρωση ηθικής τεχνητής νοημοσύνης είναι ζωτικής σημασίας.

1.3 Μηχανική Μάθηση – Machine Learning

Η μηχανική μάθηση αποτελεί μια κατηγορία της τεχνητής νοημοσύνης (AI) η οποία επιτρέπει στους υπολογιστές να μαθαίνουν και να βελτιώνονται αυτόματα χωρίς επιπλέον προγραμματισμό. Ασχολείται με τη δημιουργία προγραμμάτων υπολογιστών που μπορούν να έχουν πρόσβαση σε πλήθος δεδομένων και να μαθαίνουν απο αυτά ανεξάρτητα. Κυρίως χρησιμοποιούνται μέθοδοι στατιστικής για τον εντοπισμό προτύπων σε τεράστιους όγκους δεδομένων. Τα δεδομένα ενός αλγορίθμου μηχανικής μάθησης μπορούν να αναφέρονται σε ένα ευρύ φάσμα στοιχείων, συμπεριλαμβανομένων αριθμών, κειμένων, φωτογραφιών, κ.ο.κ. Γενικά, οτιδήποτε μπορεί να αποθηκευτεί ψηφιακά, μπορεί να τροφοδοτηθεί σε ένα σύστημα μηχανικής μάθησης.

Η διαδικασία εκμάθησης ξεκινά με παρατηρήσεις ή δείγματα δεδομένων, άμεση εμπειρία ή οδηγίες, για αναζήτηση μοτίβων στα δεδομένα και λήψη καλύτερων αποφάσεων στο μέλλον. Ο κύριος στόχος της κατηγορίας της μηχανικής μάθησης είναι οι υπολογιστές να μάθαίνουν μόνοι τους, και να προσαρμόσουν ανάλογα τη συμπεριφορά τους χωρίς την ανάγκη ανθρώπινης παρέμβασης.

Η μηχανική μάθηση διαφέρει από τον τυπικό προγραμματισμό. Στον παραδοσιακό προγραμματισμό, θα τροφοδοτούσαμε τα δεδομένα εισόδου σε ένα μηχάνημα μαζί με ένα καλά γραμμένο και δοκιμασμένο πρόγραμμα για την παραγωγή εξόδου. Αντιθέτως, κατά τη φάση εκμάθησης της μηχανικής μάθησης, τα δεδομένα εισόδου και εξόδου παρέχονται στο μηχάνημα, το οποίο στη συνέχεια παράγει ένα πρόγραμμα απο μόνο του. Δείτε το παρακάτω γράφημα για καλύτερη κατανόηση:



Πηγή: machinelearningmastery.com/basic-concepts-in-machine-learning/

Εικόνα 4: Η Θεμελιώδης Διαφορά της Μηχανικής Μάθησης με τον Κλασσικό Προγραμματισμό.

1.3.1 Κατηγορίες

Ο τομέας της μηχανικής μάθησης μπορεί να αναλυθεί περαιτέρω στις παρακάτω κατηγορίες:

Εποπτευόμενη Μηχανική Μάθηση

Οι αλγόριθμοι εποπτευόμενης μάθησης μαθαίνουν από δεδομένα στα οποία έχουν προηγουμένως εκπαιδευτεί, γνωστά ως δεδομένα εκπαίδευσης. Εκτελούν αναλύσεις και χρησιμοποιούν τα αποτελέσματα για να προβλέψουν μελλοντικά γεγονότα ταξινομόντας τα στην κατάλληλη κατηγορία. Για να παράξουν ακριβής προβλέψεις οι αλγόριθμοι αυτοί χρειάζονται πολλά δεδομένα εκπαίδευσης ώστε να καταφέρουν να ανακαλύψουν τα σωστά «μοτίβα» στα δεδομένα. Συγκρίνοντας τα αποτελέσματα της εκπαίδευσης με τα αποτελέσματα ελέγχου και χρησιμοποιώντας τα σφάλματα για να τροποποιήσει τους μεθόδους του, ο αλγόριθμος μπορεί να εκπαιδευτεί περαιτέρω.

Μη Εποπτευόμενη Μηχανική Μάθηση

Όταν δεν γνωρίζουμε ποια θα είναι τα τελικά αποτελέσματα και δεν έχουμε πρόσβαση σε ταξινομημένα ή επισημασμένα αποτελέσματα, χρησιμοποιούμε αλγόριθμους μάθησης χωρίς επίβλεψη. Αυτοί οι αλγόριθμοι ερευνούν και αναπτύσσουν μια συνάρτηση για τον χαρακτηρισμό προτύπων που είναι πλήρως άγνωστα και χωρίς ετικέτα. Αναλύει τα δεδομένα για να αποκαλύψει άγνωστα μοτίβα σε δεδομένα χωρίς ετικέτα.

Ημι-εποπτευόμενη Μηχανική Μάθηση

Αυτοί οι αλγόριθμοι χρησιμοποιούνται συνήθως σε συστήματα όπου τα δεδομένα χωρίς ετικέτα είναι πολύ περισσότερα απο τα δεδομένα με ετικέτα. Η ημι-εποπτευόμενη μηχανική μάθηση ονομάζεται έτσι επειδή λειτουργεί τόσο με τεχνικές μάθησης με επίβλεψη όσο και χωρίς επίβλεψη. Τα μοντέλα αυτά, έχουν αποδειχθεί ότι βελτιώνουν την ακρίβεια εκμάθησης των συστημάτων.

Ενισχυμένη Μηχανική Μάθηση

Αυτός ο τύπος αλγορίθμου μηχανικής μάθησης παράγει αποτελέσματα με βάση την καλύτερη απόδοση της λειτουργίας χρησιμοποιώντας τη μέθοδο δοκιμής και σφάλματος. Η έξοδος συγκρίνεται μέσω της διαδικασίας εύρεσης σφαλμάτων και έτσι το σύστημα λαμβάνει ανατροφοδότηση ώστε να το βοηθήσει να βελτιώσει ή να μεγιστοποιήσει την απόδοσή του. Στο μοντέλο παρέχονται κίνητρα και αντι-κίνητρα με την μορφή της ανατροφοδότησης, προκειμένου να επιτευχθεί ένας συγκεκριμένος στόχος.

1.3.2 Αλγόριθμοι

Οι αλγόριθμοι της μηχανικής μάθησης χωρίζονται σε πέντες κατηγορίες, η κάθε κατηγορία στοχεύει στην επίλυση ενός συγκεκριμένου εύρους προβλημάτων.

1. **Classification Algorithms**

Οι αλγόριθμοι ταξινόμησης χρησιμοποιούν ως είσοδο δεδομένα εκπαίδευσης για να προβλέψουν την πιθανότητα τα επόμενα δεδομένα να εμπίπτουν σε μία από τις προκαθορισμένες κατηγορίες. Για παράδειγμα, μια από τις πιο κοινές χρήσεις της ταξινόμησης είναι το φιλτράρισμα των μηνυμάτων ηλεκτρονικού ταχυδρομείου σε "ανεπιθύμητα" ή "μη ανεπιθύμητα". Εν ολίγοις, μια ταξινόμηση είναι μια μορφή "αναγνώρισης προτύπων", με αλγόριθμους ταξινόμησης, που χρησιμοποιούνται στα δεδομένα εκπαίδευσης για να θέσουν σε μελλοντικά σύνολα δεδομένων το ίδιο μοτίβο.

Οι πιο βασικοί αλγόριθμοι ταξινόμησης παρουσιάζονται παρακάτω:

K-Nearest neighbors

Η κατηγοριοποίηση που βασίζεται σε γείτονες είναι ένα είδος «τεμπέλικης» μάθησης καθώς δεν προσπαθεί να δημιουργήσει ένα γενικό εσωτερικό μοντέλο. Αντιθέτως, απλά αποθηκεύει περιπτώσεις δεδομένων εκπαίδευσης. Η κατάταξη καθορίζεται με απλή πλειοψηφία των k πλησιέστερων γειτόνων κάθε σημείου.

Naive Bayes

Αυτός ο αλγόριθμος βασίζεται στο θεώρημα του Bayes και υποθέτει ότι κάθε ζεύγος χαρακτηριστικών είναι ανεξάρτητο μεταξύ του. Πολλές λειτουργίες, όπως η ταξινόμηση εγγράφων και το φιλτράρισμα ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου, επωφελούνται από τους ταξινομητές Naive Bayes.

Logistic Regression

Η λογιστική παλινδρόμηση έχει σχεδιαστεί για προβλήματα ταξινόμησης και είναι αρκετά χρήσιμη στην κατανόηση της επιρροής πολλών ανεξάρτητων μεταβλητών σε μια συγκεκριμένη μεταβλητή αποτελέσματος.

Support Vector Machine

Σε ένα μηχάνημα διανύσματος υποστήριξης, τα δεδομένα εκπαίδευσης αναπαριστώνται ως σημεία στο διάστημα, χωρισμένα σε κατηγορίες με ένα σαφές κενό μεταξύ τους, όσο το δυνατόν ευρύτερο. Στη συνέχεια, νέα παραδείγματα χαρτογραφούνται στον ίδιο χώρο και ταξινομούνται ανάλογα με την πλευρά του κενού στην οποία πέφτουν.

Decision Tree

Ένα δέντρο αποφάσεων είναι μια τεχνική εποπτείας μάθησης ιδανική για ταξινόμηση εργασιών. Λειτουργεί ως διάγραμμα ροής, διαιρώντας τα σημεία δεδομένων σε δύο σχετικές κατηγορίες ταυτόχρονα. Αυτή η διαδικασία δημιουργεί υποκατηγορίες, επιτυγχάνοντας έτσι οργανική ταξινόμηση με ελάχιστη ανθρώπινη παρέμβαση.

2. Regression Algorithms

Με τη χρήση μεθόδων παλινδρόμησης στην Μηχανική Μάθηση, μπορούν να προβλεφθούν μελλοντικές τιμές. Χρησιμοποιώντας παλινδρόμηση, τα δεδομένα εισόδου/ιστορικά δεδομένα χρησιμοποιούνται για να προβλέψουν ένα ευρύ φάσμα μελλοντικών τιμών. Η ετικέτα είναι η μεταβλητή στόχου (προς πρόβλεψη) και η παλινδρόμηση καθορίζει τη σχέση μεταξύ της ετικέτας και των σημείων δεδομένων.

Οι πιο βασικοί αλγόριθμοι παλινδρόμησης παρουσιάζονται παρακάτω:

Linear Regression

Η εξαρτώμενη μεταβλητή σε αυτήν την τεχνική είναι συνεχής, η ανεξάρτητη μεταβλητή (ες) μπορεί να είναι συνεχής ή διακριτή και η γραμμή παλινδρόμησης είναι γραμμική. Η γραμμική παλινδρόμηση χρησιμοποιεί την καταλληλότερη ευθεία για να δημιουργήσει μια σχέση μεταξύ μιας εξαρτημένης μεταβλητής (Y) και μίας ή περισσότερων ανεξάρτητων μεταβλητών (X) (γνωστή και ως γραμμή παλινδρόμησης).

Lasso Regression

Ο αλγόριθμος Lasso, όπως και η παλινδρόμηση Ridge, τιμωρεί το απόλυτο μέγεθος των συντελεστών παλινδρόμησης (coefficiemts). Μπορεί επίσης να μειώσει τη μεταβλητότητα και να βελτιώσει την ακρίβεια των μοντέλων γραμμικής παλινδρόμησης.

Ridge Regression

Η παλινδρόμηση Ridge είναι μια τεχνική για την αντιμετώπιση πολυγραμμικών δεδομένων (οι ανεξάρτητες μεταβλητές συσχετίζονται σε μεγάλο βαθμό). Παρόλο που οι εκτιμήσεις των ελάχιστων τετραγώνων (OLS) είναι αμερόληπτες ως προς την πολυγραμμικότητα, οι αποκλίσεις τους είναι σημαντικές, με αποτέλεσμα η παρατηρούμενη τιμή να αποκλίνει σημαντικά από την πραγματική τιμή. Η παλινδρόμηση Ridge μειώνει τα τυπικά σφάλματα προσθέτοντας έναν βαθμό προκατάληψης στις εκτιμήσεις παλινδρόμησης.

3. Clustering Algorithms

Η ομαδοποίηση είναι η διαδικασία κατά την οποία παρόμοια αντικείμενα συγκεντρώνεται μαζί σε υπο-ομάδες. Βοηθά στον αυτόματο προσδιορισμό παρόμοιων αντικειμένων.

Οι πιο βασικοί αλγόριθμοι ομαδοποίησης παρουσιάζονται παρακάτω:

K-Means

Ο αλγόριθμος ομαδοποίησης K-Means είναι μια μέθοδος μάθησης χωρίς επίβλεψη που περιλαμβάνει μια επαναληπτική διαδικασία. Αρχικά, το σύνολο δεδομένων χωρίζεται σε k αριθμό προκαθορισμένων μη αλληλεπικαλυπτόμενων συμπλεγμάτων ή υποομάδων, με τα εσωτερικά σημεία των συστάδων να είναι όσο το δυνατόν πιο όμοια ενώ η απόσταση των εξωτερικών συστάδων να είναι όσο το δυνατόν μακρύτερη.

Agglomerative Clustering

Ο συχνότερος τύπος ιεραρχικής ομαδοποίησης που χρησιμοποιείται για την τοποθέτηση αντικειμένων σε συμπλέγματα με βάση την ομοιότητά τους είναι η συσσωρευτική ομαδοποίηση. Το AGNES είναι ένα άλλο όνομα για αυτήν (Agglomerative Nesting). Κάθε στοιχείο αντιμετωπίζεται αρχικά ως μεμωνομένο σύμπλεγμα από τον αλγόριθμο. Στην συνέχεια, ζεύγη συστάδων συγχωνεύονται ένα προς ένα έως ότου όλα τα συμπλέγματα συγχωνευτούν σε ένα μεγάλο τελικό σύμπλεγμα που περιέχει όλα τα στοιχεία. Η έξοδος του αλγορίθμου είναι ένα δενδρόγραμμα, το οποίο είναι μια αναπαράσταση των αντικειμένων βασισμένη στην δομή των δέντρων.

DBSCAN

Η χωρική ομαδοποίηση δεδομένων με θόρυβο με βάση την πυκνότητα αναφέρεται ως DBSCAN. Σε αντίθεση με τον αλγόριθμο k-means, χρησιμοποιεί τεχνική ομαδοποίησης βάσει πυκνότητας. Αυτή είναι μια αποτελεσματική προσέγγιση για τον προσδιορισμό των «υπερβολικών» τιμών (outliers) σε ένα σύνολο δεδομένων. Χωρίζει τις περιοχές σε ζώνες χαμηλής πυκνότητας συστάδων, προκειμένου να εντοπίσει ευκολότερα την απόσταση των μεγάλων συστάδων.

4. Dimensionality Reduction Algorithms

Όσο περισσότερα χαρακτηριστικά διαθέτουμε, τόσο πιο δύσκολο είναι δουλέψουμε και να παρατηρήσουμε το σύνολο δεδομένων. Επιπλέον, τα περισσότερα από αυτά τα χαρακτηριστικά είναι μερικές φορές πολύ στενά συνδεδεμένα και ως εκ τούτου περιττά. Οι μέθοδοι μείωσης διαστάσεων είναι χρήσιμες σε αυτήν την περίπτωση. Η διαδικασία μείωσης του αριθμού των τυχαίων μεταβλητών που εξετάζονται, μέσω της δημιουργίας μιας συλλογής πρωτογενών μεταβλητών, είναι γνωστή ως μείωση διαστάσεων.

Οι πιο βασικοί αλγόριθμοι μείωσης διαστάσεων δεδομένων παρουσιάζονται παρακάτω:

Principal Component Analysis

Η κύρια ανάλυση συστατικών είναι μια προσέγγιση μείωσης διαστάσεων για τη μείωση της διαστατικότητας των μεγάλων συνόλων δεδομένων. Μετατρέπει μια μεγάλη συλλογή μεταβλητών σε μια μικρότερη που όμως διατηρεί τις περισσότερες απο πληροφορίες της μεγάλης συλλογής. Φυσικά, η μείωση του αριθμού των μεταβλητών σε ένα σύνολο δεδομένων μειώνει την ακρίβεια. Ωστόσο, η απάντηση στη μείωση των διαστάσεων είναι η ανταλλαγή κάποιας ακρίβειας με την απλότητα. Επειδή μικρότερα σύνολα δεδομένων είναι ευκολότερο να εξερευνηθούν και να παρουσιαστούν καθώς αλγόριθμοι μηχανικής μάθησης μπορούν να αναλύσουν τα δεδομένα εύκολα και γρήγορα.

5. Deep Learning Algorithms

Τους αλγορίθμους της κατηγορίας βαθιάς μάθησης θα τους μελετήσουμε αναλυτικά στο επόμενο κεφάλαιο.

1.3.3 Εφαρμογές

Ταξινόμηση Ανεπιθύμητης Αλληλογραφίας

Πλέον, κατηγοριοποιούμε τα μηνύματα ηλεκτρονικού ταχυδρομείου ως ανεπιθύμητα ή μη ανεπιθύμητα χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης σε δεδομένα όπως περιεχόμενο μηνύματος, προωθητική ορολογία, διεύθυνση ηλεκτρονικού ταχυδρομείου αποστολέα, IP αποστολέα, υπερσυνδέσεις, σημεία στίξης και άλλους παράγοντες.

Εντοπισμός Καρκίνου

Η Μηχανική Μάθηση εφαρμόζεται στην υγειονομική περίθαλψη για διάγνωση και ακόμη και ανίχνευση καρκίνου με βάση ιατρικά δεδομένα προηγούμενων ασθενών. Η μέθοδος εκπαίδευσης χρησιμοποιεί εισόδους, που περιλαμβάνουν εκτός άλλων μέγεθος όγκου, ακτίνα, καμπυλότητα και περίμετρο για τον εντοπισμό καρκίνου του μαστού. Λαμβάνουμε την πιθανότητα ο όγκος να γίνει κακοήθης ή όχι ως αποτέλεσμα της εξόδου.

Προβλέψεις Πωλήσεων

Ένας αυξανόμενος αριθμός προμηθευτών ψηφιοποιεί τα αρχεία τους και πολλοί έχουν αρχίσει να χρησιμοποιούν τεχνολογίες μηχανικής εκμάθησης για να προβλέψουν πωλήσεις ενός συγκεκριμένου προϊόντος σε μια συγκεκριμένη εβδομάδα, ώστε να μπορούν να αποθηκεύουν αρκετό απόθεμα. Αρχικά, οι αλγόριθμοι μηχανικής μάθησης συλλέγουν δεδομένα από τις πωλήσεις διαφόρων ειδών του προηγούμενου έτους και αναζητούν μοτίβα σε εποχιακές

διακυμάνσεις, προκειμένου να προβούν σε λεπτομερείς προβλέψεις σχετικά με την πώληση συγκεκριμένων αντικειμένων.

Αναγνώριση Προσώπου

Πιθανότατα έχετε παρατηρήσει ότι όταν υποβάλλετε φωτογραφίες στο Facebook, αυτόματα επισημαίνει τα πρόσωπα των φίλων σας με τα ονόματά τους. Αυτό γίνεται στο πίσω μέρος με τεχνικές μηχανικής/βαθιάς εκμάθησης. Επίσης η αναγνώριση προσώπων χρησιμοποιείται και απο την αστυνομία για την ταυτοποίηση υπόπτων.

Ταξινόμηση Κειμένου

Με τον αυξανόμενο αριθμό ατόμων που χρησιμοποιούν το Διαδίκτυο, οι αλγόριθμοι που βασίζονται στην ταξινόμηση κειμένου γίνονται όλο και πιο σημαντικοί για τους ιστότοπους και τις πλατφόρμες κοινωνικών. Μέσω της ταξινόμησης κειμένου το twitter και το facebook εντοπίζουν σχόλια και δημοσιεύσεις μίσους. Οι αλγόριθμοι κατηγοριοποίησης κειμένου χρησιμοποιούνται επίσης από ορισμένους ειδησεογραφικούς οργανισμούς για την ομαδοποίηση συγκρίσιμων ειδήσεων.

1.3.4 Δυσκολίες και Προκλήσεις

Προκατειλημένα Δεδομένα

Συχνά, τα δεδομένα εισόδου σε ένα σύστημα ML είναι μη αντικειμενικά προς ένα δεδομένο φύλο, φυλή, χώρα, κάστα και ούτω καθεξής. Ως αποτέλεσμα, οι αλγόριθμοι ML εισάγουν ακούσια προκατάληψη στη διαδικασία λήψης αποφάσεων. Αυτό έχει φανεί σε αρκετά προγράμματα που χρησιμοποιούν τη μηχανική μάθηση για να προσομοιώσουν μια διαδικασία εισαγωγής παιδιών σε σχολεία/κολλέγια αλλα και σε λογισμικά υπεύθυνα για την προβολή «συστάσεων» σε κοινωνικά μέσα.

Χρονοβόρο και Ακριβό

Απαιτούνται μεγάλοι όγκοι δεδομένων και χρόνος για να επιτευχθεί αποδεκτή ακρίβεια. Ενώ οι άνθρωποι μπορούν να μάθουν γρήγορα με μικρά σύνολα δεδομένων, ορισμένες εφαρμογές απαιτούν τεράστιο όγκο δεδομένων και χρόνο για να επιτύχουν αποδεκτή ακρίβεια.

1.4 Βαθιά Μάθηση – Deep Learning

Η βαθιά μάθηση είναι ένα υποσύνολο της μηχανικής μάθησης στον ευρύτερο τομέα της τεχνητής νοημοσύνης, έχει καταφέρει να εφαρμόσει λειτουργίες που μαθαίνουν την λειτουργικότητα του εγκεφάλου δημιουργώντας αλγορίθμους και τεράστια όγκο δεδομένων. Βασίζεται σε τεχνητά νευρωνικά δίκτυα. Το 1958, ο καθηγητής του Cornell Frank Rosenblatt δημιούργησε μια πρώιμη έκδοση ενός τεχνητού νευρωνικού δικτύου που αποτελείται από διασυνδεδεμένα perceptrons.[21] Ως κόμβος στα σύγχρονα τεχνητά νευρωνικά δίκτυα, ένα perceptron λαμβάνει δυαδικές εισόδους και υπολογίζει αυτές τις εισόδους για να παράγει μια έξοδο. Σημειώστε ότι στον perceptron, τόσο οι είσοδοι όσο και οι έξοδοι είναι δυαδικές - για παράδειγμα, μηδέν/ένα, ενεργοποίηση/απενεργοποίηση, είσοδος/έξοδος.

Παρακάτω αναλύουμε τις θεμελιακές λειτουργίες ενός απλού μοντέλου βαθιάς μάθησης:

1. Μέσα από τη μελέτη δομημένων δεδομένων, η βαθιά μάθηση μαθαίνει να αναγνωρίζει τα χαρακτηριστικά που μοιράζονται τα αντικείμενα της μελέτης.

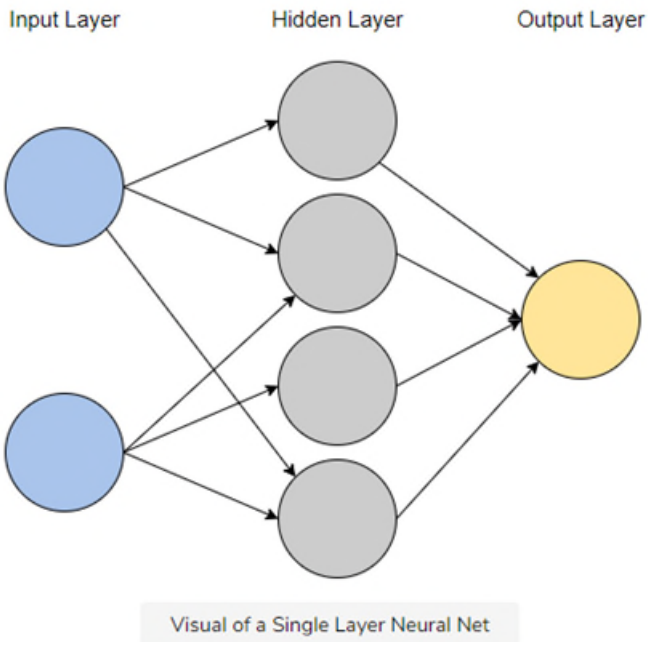
2. Στη συνέχεια, ο αλγόριθμος εξετάζει κάθε χαρακτηριστικό των δεδομένων και αναζητά κοινά στοιχεία μεταξύ τους. Αυτή η μέθοδος ονομάζεται εξαγωγή χαρακτηριστικών.

3. Έπειτα, το πρόγραμμα καθορίζει ποια από αυτά τα χαρακτηριστικά είναι τα πιο ακριβή και σημαντικά. Το όριο απόφασης είναι το όνομα αυτού του κριτηρίου.

4. Τέλος, αφού ο αλγόριθμος έχει καθορίσει αυτά τα κριτήρια χρησιμοποιώντας όλα τα διαθέσιμα δεδομένα εκπαίδευσης, χρησιμοποιεί αυτά τα πρόσφατα κριτήρια για να ταξινομήσει τα μη δομημένα δεδομένα εισόδου στις κατάλληλες κατηγορίες.

Τα πολλαπλά κρυμμένα επίπεδα επεξεργασίας που πρέπει να περάσουν τα δεδομένα εισόδου, αποτελεί το μυστικό της βαθιάς μάθησης. Πολλαπλοί νευρώνες ή «κόμβοι» με μαθηματικές συναρτήσεις συλλέγουν και ταξινομούν δεδομένα σε κάθε επίπεδο. Τα επίπεδα εισόδου και εξόδου είναι το πρώτο και το τελευταίο επίπεδο, αντίστοιχα. Τα κρυμμένα επίπεδα με κόμβους που χρησιμοποιούν ως είσοδο τα αποτελέσματα προηγούμενων ταξινομήσεων, επικοινωνούν μεταξύ τους. Αυτοί οι κόμβοι χρησιμοποιούν τις λειτουργίες κατηγοριοποίησης στα αρχικά ευρήματα και αλλάζουν τη στάθμιση των βαρών αντίστοιχα στην πορεία.

Τα παραδοσιακά νευρωνικά δίκτυα πριν από τη βαθιά μάθηση θα περνούσαν δεδομένα μόνο από 2-3 κρυμμένα επίπεδα πριν από την ολοκλήρωση τους. Η βαθιά μάθηση αυξάνει αυτόν τον αριθμό έως και 150 κρυμμένα επίπεδα με αποτέλεσμα να αυξάνεται και η ακρίβεια των αποτελεσμάτων.



Πηγή: educative.io/blog/deep-learning-beginner-tutorial

Εικόνα 5: Η Θεμελιακή Δομή Ενός Απλού Νευρωνικού Δικτύου.

Παρακάτω αναλύουμε πιό συγκεκριμένα, τον τρόπο ταξινόμησης και την μεταφορά δεδομένων με βάση την δομή ενός απλού νευρωνικού δικτύου σε 4 βήματα:

1. Ως επίπεδο εισόδου (Input Layer) χρησιμοποιούνται ακατέργαστα δεδομένα. Ο αλγόριθμος εκτελεί μία πρόχειρη ταξινόμηση και προωθεί τα αποτελέσματα στον επόμενο κόμβο κρυμμένου επιπέδου (Hidden Layer).

2. Οι κόμβοι στο πρώτο κρυφό επίπεδο ταξινομούνται με βάση πιο συγκεκριμένα και σημαντικά κριτήρια.

3. Οι κόμβοι κάθε διαδοχικού κρυμμένου στρώματος γίνονται όλο και πιο συγκεκριμένοι, περιορίζοντας ακόμη περισσότερο τις δυνατότητες κατηγοριοποίησης σταθμίζοντας το αποτέλεσμα μέσω βαρών.

4. Τέλος, από αυτά που δεν έχουν αποκλειστεί, το τελικό επίπεδο εξόδου επιλέγει την πιο πιθανή ετικέτα ταξινόμησης.

1.4.1 Βασικές Έννοιες

Συναρτήσεις Απώλειας – Loss Functions

Οι συναρτήσεις απώλειας χρησιμοποιούνται για τον υπολογισμό του σφάλματος (γνωστό και ως "η απώλεια") μεταξύ της εξόδου των αλγορίθμων μας και της παρεχόμενης τιμής στόχου. Με απλά λόγια, η συνάρτηση απώλειας εκφράζει πόσο μακριά η υπολογισμένη μας έξοδος είναι εκτός ορίου.

Λειτουργία ενεργοποίησης – Activation Function

Η έξοδος ενός μοντέλου νευρωνικού δικτύου καθορίζεται από συναρτήσεις ενεργοποίησης, οι οποίες είναι μαθηματικές εξισώσεις. Οι λειτουργίες ενεργοποίησης έχουν σημαντικό αντίκτυπο στην ικανότητα σύγκλισης των νευρωνικών δικτύων και στον ρυθμό με την οποία την επιτυχαίνουν. Σε ορισμένες περιπτώσεις, οι λειτουργίες ενεργοποίησης μπορεί ακόμη και να αποτρέψουν τη σύγκλιση των νευρωνικών δικτύων. Η λειτουργία ενεργοποίησης βοηθά επίσης στην ομαλοποίηση οποιασδήποτε εισόδου -εξόδου στην περιοχή 1 έως -1 ή 0 έως 1. Επειδή τα νευρωνικά δίκτυα εκπαιδεύονται σε εκατομμύρια δεδομένα, η λειτουργία ενεργοποίησης πρέπει να είναι αποτελεσματική και να μειώνει το χρόνο υπολογισμού.

Εμπρός Διάδοση – Forward Propagation

Η εμπρός διάδοση είναι η διαδικασία αποθήκευσης και υπολογισμού των δεδομένων εισόδου πριν από την αποστολή τους μέσω του δικτύου (hidden layers) για τη δημιουργία εξόδου. Για παράδειγμα, σε ένα νευρωνικό δίκτυο, τα κρυμμένα επίπεδα δέχονται δεδομένα από το επίπεδο εισόδου, τα επεξεργάζονται χρησιμοποιώντας μια λειτουργία ενεργοποίησης και στη συνέχεια τα στέλνουν στο επίπεδο εξόδου ή στα επόμενα επίπεδα. Έτσι, τα δεδομένα ρέουν προς τα εμπρός για να αποφευχθεί μια κυκλική ροή δεδομένων που δεν παρέχει έξοδο. Το δίκτυο προώθησης είναι μια αρχιτεκτονική δικτύου που βοηθά στην προώθηση της διάδοσης.

Πίσω Διάδοση – Backward Propagation

Με αυτόν τον αλγόριθμο δημιουργούμε παράγωγα. Τα παράγωγα (κλίσεις) υπολογίζονται σε κάθε επανάληψη για βελτιστοποίηση. Δυστυχώς, οι λειτουργίες στη βαθιά μάθηση δεν είναι απλές, αποτελούνται από πολλές ξεχωριστές λειτουργίες. Επειδή ο υπολογισμός των κλίσεων είναι δύσκολος σε αυτό το σενάριο, υπολογίζουμε τα παράγωγα χρησιμοποιώντας διαφοροποίηση προσέγγισης. Όσο περισσότερες παράμετροι υπάρχουν, τόσο πιο δαπανηρή γίνεται κατά προσέγγιση η διαφοροποίηση.

Στοχαστική κατάβαση κλίσης – Stochastic Gradient Descent

Ο σκοπός της κατάβασης κλίσης είναι να ανακαλύψει παγκόσμια ελάχιστα (global minimum) ή βέλτιστες λύσεις. Ωστόσο, για να γίνει αυτό, πρέπει επίσης να εξετάσουμε τοπικές ελάχιστες λύσεις (οι οποίες είναι ανεπιθύμητες). Είναι απλό να ανακαλύψουμε τα παγκόσμια

ελάχιστα αν η αντικειμενική συνάρτηση είναι κυρτή. Η αρχική τιμή της συνάρτησης και ο ρυθμός εκμάθησης χρησιμεύουν ως καθοριστικές παράμετροι για τον εντοπισμό παγκόσμιων ελαχίστων.

Ρυθμός εκμάθησης – Learning Rate

Ο ρυθμός εκμάθησης ορίζει ουσιαστικά το «κατα πόσο» θα μεταβληθούν τα βάρη στην προσπάθεια του αλγορίθμου να εντοπίσει το παγκόσμιο ελάχιστο. Προτιμάται χαμηλός ρυθμός εκμάθησης ώστε ο αλγόριθμος να φτάσει σίγουρα κάποια στιγμή σε παγκόσμιο ελάχιστο. Στην περίπτωση μεγάλου ρυθμού εκμάθησης μπορεί να μην φτάσουμε ποτέ σε παγκόσμιο ελάχιστο.

Optimizer

Προσαρμόζουμε και αλλάζουμε τις παραμέτρους του μοντέλου μας (βάρη) κατά τη διάρκεια του training phase για να ελαχιστοποιήσουμε την απώλεια και να κάνουμε τις πιο ακριβείς δυνατές προβλέψεις. Τα βελτιστοποιητικά μέτρα είναι χρήσιμα σε αυτήν την περίπτωση. Συνδέουν τη συνάρτηση απώλειας και τις παραμέτρους του μοντέλου, αλλάζοντας το μοντέλο ως απάντηση στην έξοδο της συνάρτησης απώλειας. Με απλά λόγια, οι βελτιστοποιητές ανακατανέμουν τα βάρη κατάλληλα για να μετασχηματίσουν το μοντέλο μας στην πιο ακριβείς μορφή. Η συνάρτηση απώλειας χρησιμεύει ως «οδικός χάρτης» για το βελτιστοποιητή, υποδεικνύοντας εάν «ταξιδεύει» στη σωστή ή λάθος διαδρομή.

Επίπεδα Εγκατάληψης – Drop-out Layers

Η υπερπροσαρμογή(overfiting) είναι ένα πρόβλημα που προκύπτει συχνά στη βαθιά μάθηση. Η υπερπροσαρμογή καθιστά δύσκολη την σωστή πρόβλεψη σε δεδομένα εκτός του σετ εκπαίδευσης. Σε δίκτυα με πολλές παραμέτρους χρησιμοποιούμε την στρατηγική που περιλαμβάνει τη δημιουργία νέων «αραιωμένων δικτύων» και την απόρριψη μονάδων του δικτύου που «μπερδεύουν» τον αλγόριθμο κατά τη διάρκεια της εκπαίδευσης. Οι προβλέψεις αυτών των αραιωμένων δικτύων υπολογίζονται κατά μέσο όρο στην διάρκεια των δοκιμών, γεγονός που βοηθά στην αποφυγή υπερβολικής προσαρμογής.

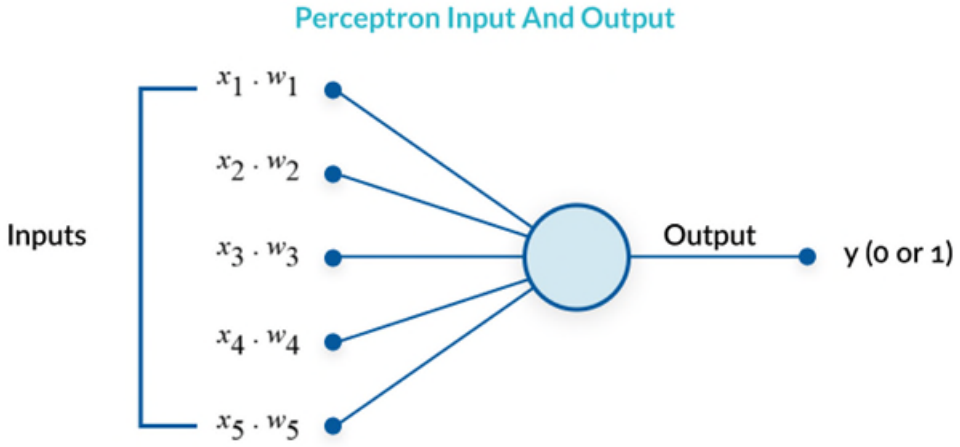
Ομαλοποίηση παρτίδας – Batch Normalization

Βοηθά στην δημιουργία βαθύτερων νευρωνικών δικτύων τα οποία μπορύν να εκπαιδευτούν γρηγορότερα μόλις εισαχθούν στο επίπεδο ομαλοποίησης παρτίδας. Η ομαλοποίηση παρτίδας είναι ένα επίπεδο νευρωνικού δικτύου που μπορεί να βρεθεί επί του παρόντος σε μια ποικιλία τοπολογιών. Για παράδειγμα, περιλαμβάνεται συχνά ως μέρος ενός γραμμικού ή συνεκτικού δικτύου και βοηθά στη σταθεροποίηση του δικτύου κατά τη διάρκεια της εκπαίδευσης.

1.4.2 Αλγόριθμοι – Αρχιτεκτονικές

Perceptron

Το μοντέλο perceptron Minsky-Papert[22] είναι ένα από τα απλούστερα και παλαιότερα μοντέλα νευρώνων. Αποτελεί την μικρότερη μονάδα ενός νευρωνικού δικτύου που εκτελεί ορισμένους υπολογισμούς για να ανακαλύψει των εισερχόμενων δεδομένων. Παίρνει σταθμισμένες εισόδους και εφαρμόζει τη λειτουργία ενεργοποίησης(activation function) για να παράγει το τελικό αποτέλεσμα. Το TLU είναι ένα άλλο όνομα για το perceptron (μονάδα λογικής κατωφλίου). Το Perceptron ουσιαστικά είναι ένας δυαδικός ταξινομητής που χωρίζει τα δεδομένα σε δύο ομάδες.



Πηγή: jatinmishra27.medium.com/understanding-a-perceptron-building-block-of-an-artificial-neural-network-558942f8ee37

Εικόνα 6: Η Θεμελιακή Δομή Ενός Perceptron.

Πλεονεκτήματα:

- Χρησιμοποιούμε perceptrons για την αναπαράσταση λογικών πυλών όπως AND, OR ή NAND.

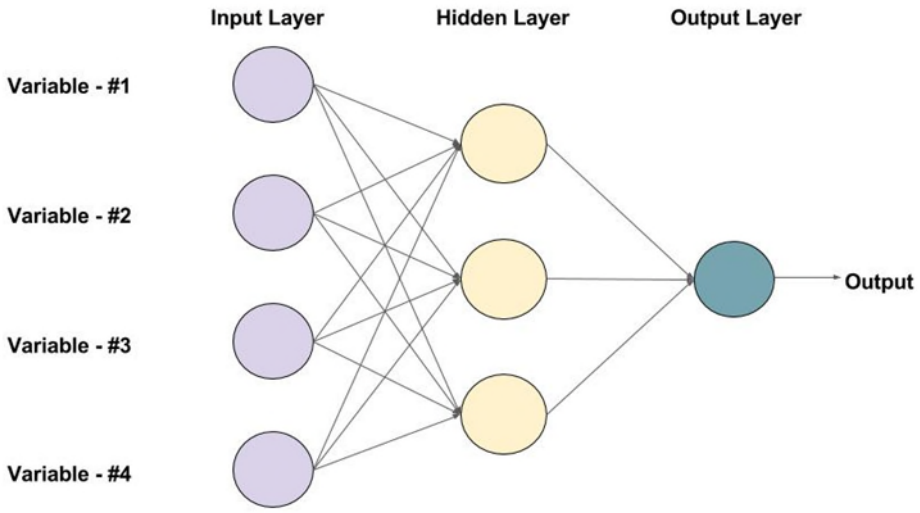
Μειονεκτήματα:

- Ικανός αλγόριθμος ώστε να μάθει μόνο γραμμικά διαχωρίσιμα προβλήματα. Για μη γραμμικά προβλήματα όπως το πρόβλημα του boolean XOR, δεν αποδίδει.

Feed Forward Neural Networks

Στην πιο βασική έκδοση των νευρωνικών δικτύων, τα δεδομένα εισόδου ρέουν μόνο προς μία κατεύθυνση, περνώντας απο το στρώμα εισόδου (input layer) στο στρώμα εξόδου (output layer). Αναλόγως με το πόσα κρυφά στρώματα υπάρχουν ανάμεσα τους κατηγοριοποιούνται σε: μονοεπίπεδα ή πολυεπίπεδα. Η πολυπλοκότητα της συνάρτησης καθορίζει και τον αριθμό των επιπέδων που θα χρησιμοποιηθούν ενώ τα βάρη παραμένουν σταθερά. Στη συνέχεια, πολλαπλασιάζουμε τις εισόδους με τα αντίστοιχα βάρη τους και τις στέλνουμε σε μια συνάρτηση ενεργοποίησης. Χρησιμοποιούμε μια συνάρτηση ενεργοποίησης ταξινόμησης (classification function) ή μια συνάρτηση ενεργοποίησης βημάτων (step function) για να το επιτύχουμε αυτό.

Για παράδειγμα, εάν το κατώφλι του νευρώνα (τυπικά το 0) ξεπεραστεί, ο νευρώνας ενεργοποιείται και παράγει ως έξοδο ένα (1). Εάν ο νευρώνας είναι κάτω από το κατώφλι (συνήθως 0), θεωρείται (-1) και δεν ενεργοποιείται.



Πηγή: learnopencv.com/understanding-feedforward-neural-networks/

Εικόνα 7: Η Θεμελιακή Δομή Ενός Feed Forward Neural Network.

Πλεονεκτήματα:

- Εύκολη κατασκευή και συντήρηση
- Αποδίδει εξαιρετικά ακόμη και σε δεδομένα με θόρυβο

Μειονεκτήματα:

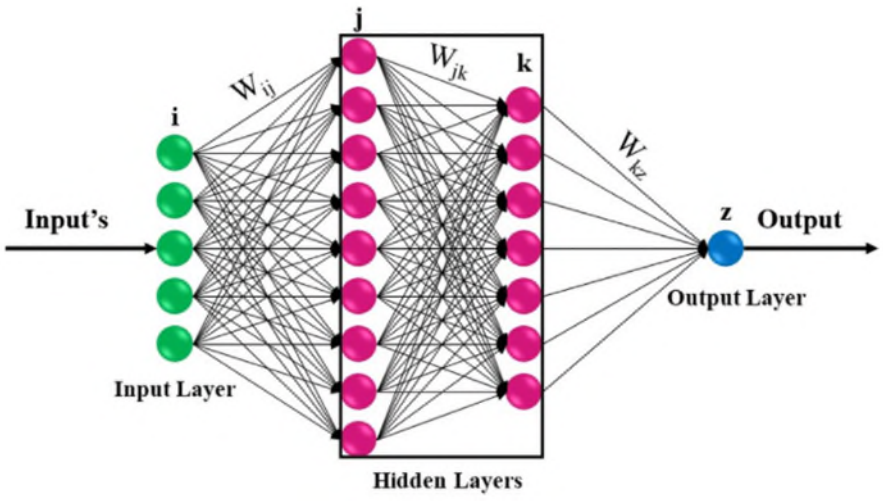
- Δεν είναι κατάλληλα για βαθιά μάθηση (λόγω έλλειψης πυκνών στρωμάτων - dense layers - και της τεχνικής back-propagation)

Εφαρμογές:

- Απλή Ταξινόμηση
- Αναγνώριση Προσώπου
- Υπολογιστική Όραση
- Αναγνώρισης Ομιλίας

Multi-layer Perceptron

Το Multilayer Perceptron αποτελεί την πρώτη αρχιτεκτονική που μας εισαγάγει στα βαθιά νευρωνικά δίκτυα, στα οποία τα δεδομένα εισόδου δρομολογούνται μέσω πολλαπλών στρωμάτων τεχνητών νευρώνων (hidden layers). Είναι ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο αφού κάθε κόμβος είναι συνδεδεμένος με όλους τους νευρώνες στο προηγούμενο στρώμα. Επιπλέον, διαθέτει διάδοση δύο κατευθύνσεων, πράγμα που σημαίνει ότι μπορεί να διαδοθεί τόσο προς τα εμπρός (forward propagation) όσο και προς τα πίσω (back propagation). Οι εισροές πολλαπλασιάζονται με τα αντίστοιχα βάρη και αποστέλλονται στη συνάρτηση ενεργοποίησης, όπου αλλάζουν συνεχώς τιμή ώστε να ελαχιστοποιηθεί η απώλεια (loss function). Ανάλογα με τη διαφορά μεταξύ των αναμενόμενων αποτελεσμάτων και των εισροών, προσαρμόζονται ανάλογα.



Πηγή: www.researchgate.net/figure/The-basic-form-of-multilayer-perceptron-artificial-neural-network-ANN-61_fig1_341626283

Εικόνα 8: Η Θεμελιακή Δομή Ενός Multi-layer Perceptron.

Πλεονεκτήματα

- Είναι κατάλληλο για βαθιά μάθηση (λόγω της παρουσίας πυκνών, πλήρως συνδεδεμένων στρωμάτω)

Μειονεκτήματα:

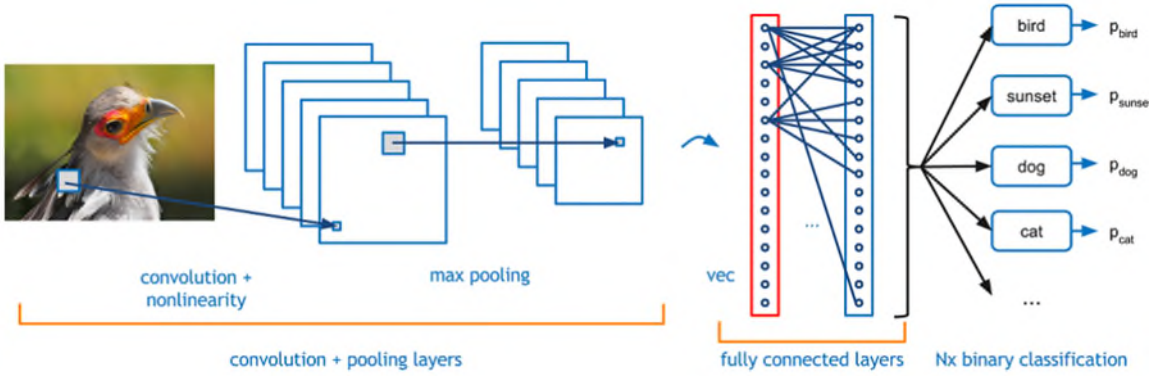
- Πολύπλοκο σχεδιασμό και συντήρηση
- Σχετικά αργό (εξαρτάται από τον αριθμό των κρυφών στρωμάτων)

Εφαρμογές:

- Αναγνώρισης ομιλίας
- Μηχανική Μετάφραση
- Σύνθετη Ταξινόμηση

Convolutional Neural Networks[23]

Κάθε νευρώνας convolution στρώματος αναλύει δεδομένα μόνο από ένα περιορισμένο τμήμα του οπτικού πεδίου. Όπως ένα φίλτρο, τα χαρακτηριστικά εισόδου λαμβάνονται σε παρτίδες. Όταν ενα convolutional νευρωνικό δίκτυο θέλει να καταλάβει το περιεχόμενο μίας εικόνας, τότε το κάθε convolution στρώμα είναι υπεύθυνο για την αναγνώριση ενός διαφορετικού σημείου της εικόνας (Το πρώτο αναγνωρίζει τις γωνίες, το δεύτερο τα ενδιάμεσα μέρη κ.ο.κ). Έτσι το δίκτυο αποκωδικοποιεί εικόνες σε πολλά μικρά κομμάτια και μπορεί να εκτελέσει αυτές τις λειτουργίες πολλές φορές ώστε να ολοκληρώσει την επεξεργασία της εικόνας. Η εικόνα μετατρέπεται από RGB ή HSI σε κλίμακα του γκρι κατά την επεξεργασία. Περαιτέρω παραλλαγές στην τιμή των εικόνων θα βοηθήσουν στην ανίχνευση άκρων, επιτρέποντας την κατηγοριοποίηση των εικόνων σε διάφορες κατηγορίες. Τέλος, η έξοδος του convolution στρώματος οδηγείται σε σε ένα πλήρως συνδεδεμένο στρώμα για να ολοκληρωθεί η ταξινόμηση.



Πηγή: https://www.kdnuggets.com/2016/09/beginners-guide-understanding-convolutional-neural-networks-part-1.html

Εικόνα 9: Η Θεμελιακή Δομή Ενός Convolutional Neural Network.

Πλεονεκτήματα:

- Πολύ καλά αποτελέσματα στην αναγνώριση εικόνων
- Χρησιμοποιείται για βαθιά μάθηση με λίγες παραμέτρους

Μειονεκτήματα:

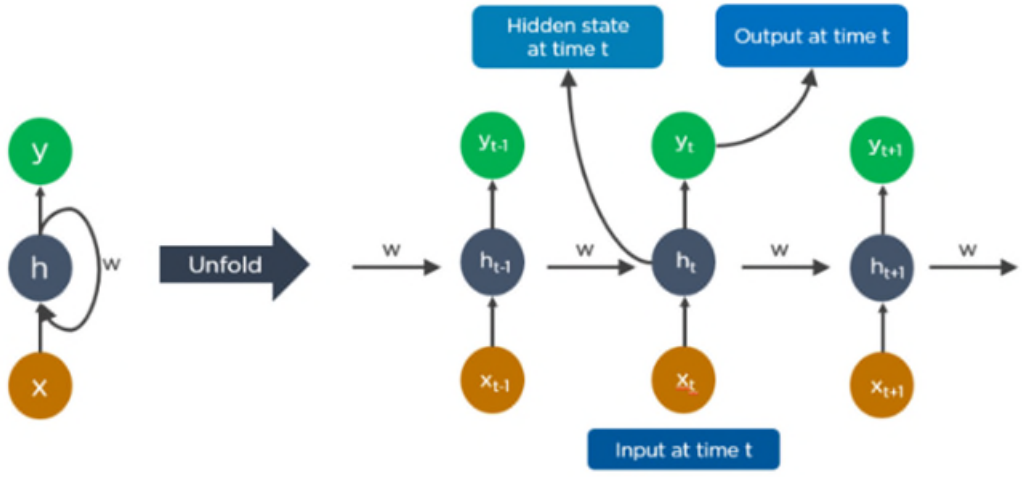
- Συγκριτικά πολύπλοκο στο σχεδιασμό και τη συντήρηση
- Συγκριτικά αργά (εξαρτάται από τον αριθμό των κρυφών στρωμάτων)

Εφαρμογές:

- Επεξεργασία Εικόνας
- Αναγνώρισης ομιλίας
- Μηχανική μετάφραση

Recurrent Neural Networks[24]

Τα recurrent νευρωνικά δίκτυα έχουν σχεδιαστεί για να αποθηκεύουν την έξοδο ενός επιπέδου και να το τροφοδοτούν ξανά στην είσοδο του επόμενου ώστε να βοηθήσουν στην όλο και καλύτερη πρόβλεψη όσο συνεχίζεται η πορεία προς το στρώμα εξόδου. Το πρώτο επίπεδο είναι συνήθως ένα νευρωνικό δίκτυο προώθησης (feed forward neural network), ακολουθούμενο από ένα recurrent στρώμα δικτύου. Στο στρώμα αυτό υπάρχει λειτουργία μνήμης η οποία κρατάει το αποτέλεσμα του προηγούμενο κόμβου. Ουσιαστικά, αυτή η μνήμη αποθηκεύει πληροφορίες που θα χρειαστούν στο μέλλον. Εάν η πρόβλεψη είναι λανθασμένη, το ποσοστό εκμάθησης (learning rate) χρησιμοποιείται για να πραγματοποιήσει μικρές προσαρμογές. Ως αποτέλεσμα, αυξάνεται σταδιακά η πιθανότητα να γίνει η σωστή πρόβλεψη κατά τη διάρκεια της διαδικασία οπισθοδρόμησης (back-propagation).



Πηγή:simplilearn.com/tutorials/deep-learning-tutorial/

Εικόνα 10: Η Θεμελιακή Δομή Ενός Recurrent Neural Network.

Πλεονεκτήματα:

- Διατηρεί πληροφορίες
- Χρησιμοποιείται μαζί με στρώμματα convolution για να αυξήσει την ευκρίνεια των εικόνων

Μειονεκτήματα:

- Φαινόμενα Gradient Vanishing και gradient exploding.

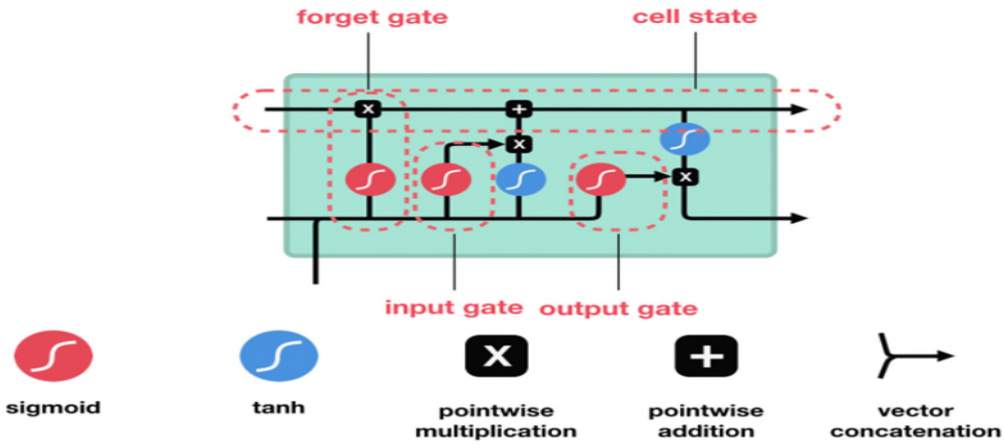
- Δύσκολη η επεξεργασία μεγάλου μήκους διαδοχικών δεδομένων

Εφαρμογές:

- Επεξεργασία κειμένου
- Μετατροπή κειμένου σε ήχο.
- Προσθήκη ετικετών εικόνας
- Ανάλυση συναισθημάτων
- Μετάφραση

Long Short Term Memory Neural Networks[25]

Τα δίκτυα LSTM είναι ένας τύπος δικτύου RNN. Στις μονάδες LSTM, υπάρχει ένα κελί μνήμης που μπορεί να αποθηκεύσει πληροφορίες για μεγάλο χρονικό διάστημα. Οι πληροφορίες διέπονται από ένα σύστημα πύλων όταν εισέρχονται στη μνήμη, όταν εξάγονται και όταν ξεχνιούνται. Τα τρία είδη πύλων είναι πύλες εισόδου, πύλες εξόδου και πύλες που ελέγχουν τί θα κρατήσουν και τί θα αφήσουν απο την είσοδο που δέχονται, ονομάζονται και πύλες που «ξεχνούν». Η πύλη εισόδου ρυθμίζει την ποσότητα δεδομένων που αποθηκεύονται στη μνήμη από το προηγούμενο δείγμα. Η πύλη εξόδου ελέγχει την ποσότητα των δεδομένων που μεταδίδονται στο επόμενο επίπεδο και οι πύλες ξεχασμού ελέγχουν το πόσα και ποιά δεδομένα θα μεταφερθούν στο επόμενο επίπεδο του νευρωνικού δικτύου. Αυτή η αρχιτεκτονική τους επιτρέπει να μάθουν μακροπρόθεσμες εξαρτήσεις, κάτι το οποίο βοηθάει αρκετά στο να επιτυγχάνουν καλύτερα αποτελέσματα.



Πηγή: juejin.cn/post/6923794050101280775

Εικόνα 11: Η Θεμελιακή Δομή Ενός Long Short Term Memory Neural Network.

Πλεονεκτήματα:

- Δημιουργεί βαθύτερες μακροπρόθεσμες εξαρτήσεις, γεγονός που οδηγεί σε καλύτερα αποτελέσματα.
- Χρησιμοποιείται με convolution στρώματα για να επεκτείνει την αποτελεσματικότητα των εικόνων σε CNN δίκτυα.

Μειονεκτήματα:

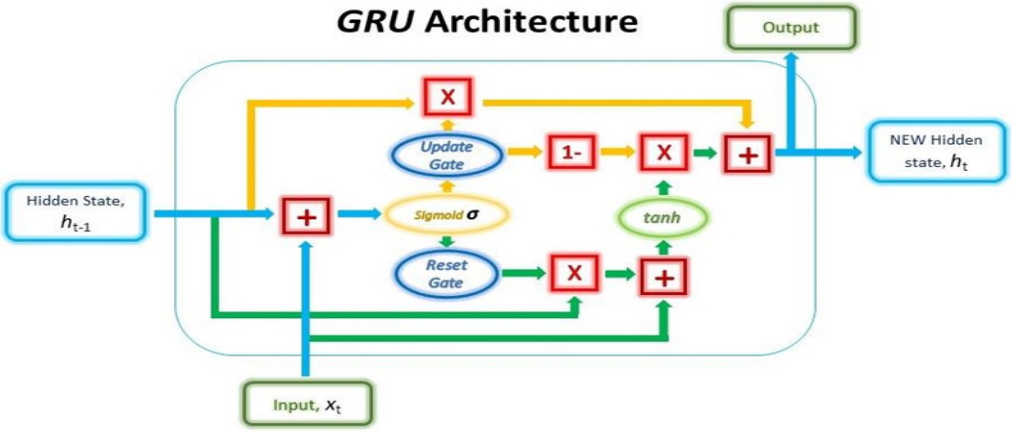
- Απαιτεί χρόνο και πόρους για την προπόνηση.
- Τα LSTM είναι επιρρεπή σε υπερπροσαρμογή (overfitting).

Εφαρμογές:

- Επεξεργασία Κειμένου (Αυτόματη διόρθωση)
- Μετατροπή κειμένου σε ομιλία
- Προσθήκη ετικετών σε εικόνες
- Ανάλυση συναισθημάτων
- Μετάφραση

Gated Recurrent Unit Neural Networks

Το GRU, είναι ένα βήμα μπροστά από το παραδοσιακό RNN και παρουσιάστηκε για πρώτη φορά το 2014 από τους Kyunghyun Cho et al. [26]. Η αρχιτεκτονική μακροπρόθεσμης μνήμης (LSTM) και η GRU είναι αρκετά παρόμοιες. Το GRU, όπως και το LSTM, ελέγχει τη ροή πληροφοριών μέσω θυρών. Σε σύγκριση με το LSTM, είναι αρκετά πιο καινούργιο. Αυτός είναι και ο λόγος που τα ξεπερνούν σε αποτελεσματικότητα και έχουν μια πιο απλή αρχιτεκτονική, εύκολα υλοποιήσιμη. Ένα άλλο ενδιαφέρον χαρακτηριστικό του GRU είναι ότι, σε αντίθεση με το LSTM, δεν διαθέτει μια ξεχωριστή κατάσταση κελιού (Ct). Έχει μία μοναδική κατάσταση: την κρυφή (Ht).



Πηγή: blog.floydhub.com/gru-with-pytorch/

Εικόνα 12: Η Θεμελιακή Δομή Ενός Gated Recurrent Unit Neural Network.

Πλεονεκτήματα:

- Δυνατότητα επεξεργασίας εισόδων οποιουδήποτε μήκους
- Το μέγεθος του μοντέλου δεν αυξάνεται αναλογικά με το μέγεθος της εισόδου
- Ο υπολογισμός λαμβάνει υπόψη τις προηγούμενες πληροφορίες
- Τα βάρη μοιράζονται με την πάροδο του χρόνου

Μειονεκτήματα:

- Ο υπολογισμός είναι αργός
- Δυσκολία πρόσβασης σε πολυ παλιές πληροφορίες
- Δεν είναι δυνατή η εξέταση μελλοντικής εισόδου για την τρέχουσα κατάσταση

Εφαρμογές:

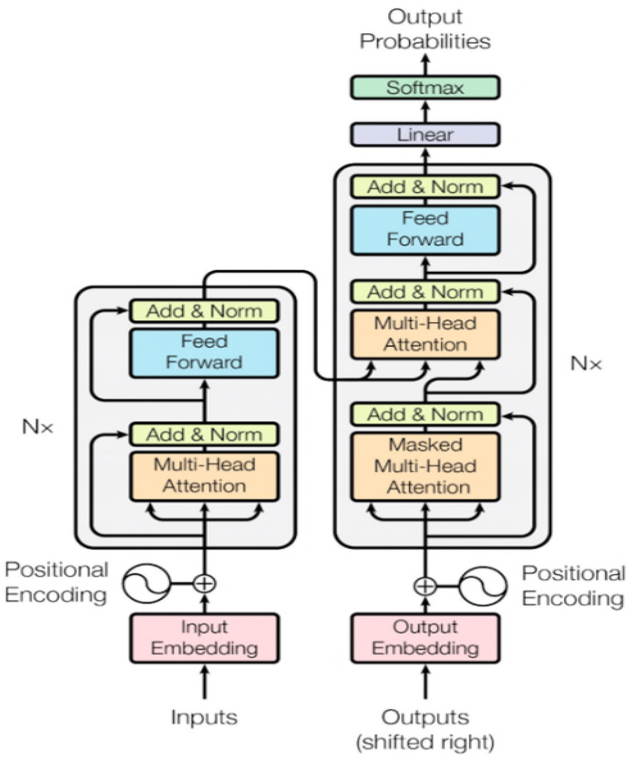
- Παραγωγή Μουσικής
- Ανάλυση συναισθημάτων
- Μηχανική μετάφραση

Transformers Neural Network

Ένα νευρωνικό δίκτυο μετασχηματιστή μπορεί να δεχθεί μια ακολουθία διανυσμάτων ως είσοδο και να τα μετατρέψει σε ένα κωδικοποιημένο διάνυσμα, το οποίο στη συνέχεια μπορεί να αποκωδικοποιήσει ξανά σε άλλη ακολουθία. Ο μηχανισμός προσοχής (attention mechanism) του μετασχηματιστή είναι ένα κρίσιμο συστατικό. Ο μηχανισμός προσοχής ορίζει την σημαντικότητα της κάθε λέξης στην πρόταση, ώστε στην μετέπειτα μετατροπή της για παράδειγμα σε άλλη γλώσσα να γνωρίζει σε ποιά σημεία της εισόδου πρέπει να δώσει περισσότερη σημασία ώστε το αποτέλεσμα να είναι ικανοποιητικό.

Για παράδειγμα, σε ένα μοντέλο αυτόματης μετάφρασης, ο μηχανισμός προσοχής επιτρέπει στον μετασχηματιστή να μεταφράσει όρους όπως "αυτό" σε μια γαλλική ή ισπανική λέξη που ταιριάζει στο φύλο, δίνοντας προσοχή σε όλες τις σχετικές λέξεις στο αρχικό κείμενο.

Συγκεκριμένα, ο μηχανισμός προσοχής του μετασχηματιστή του επιτρέπει να εστιάζει σε συγκεκριμένες λέξεις αριστερά και δεξιά της τρέχουσας λέξης για να καθορίσει τον τρόπο μετάφρασής τους. Ως αποτέλεσμα, τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN), η μακροχρόνια βραχυπρόθεσμη μνήμη (LSTM) και οι αρχιτεκτονικές νευρωνικών δικτύων με επαναλαμβανόμενες πύλες (GRU) αντικαταστάθηκαν από νευρωνικά δίκτυα μετασχηματιστών, τα οποία βέβαια έχουν περίπλοκη δομή.



Πηγή: tungmphung.com/the-transformer-neural-network-architecture/

Εικόνα 13: Η Θεμελιακή Δομή Ενός Transformer Neural Network.

Πλεονεκτήματα:

- Μηχανισμός Προσοχής
- Καλύτερα αποτελέσματα
- Μπορεί να μάθει αποτελεσματικά τεράστιες ποσότητες δεδομένων

Μειονεκτήματα:

- Υψηλή υπολογιστική ισχύς
- Αργός υπολογισμός

Εφαρμογές:

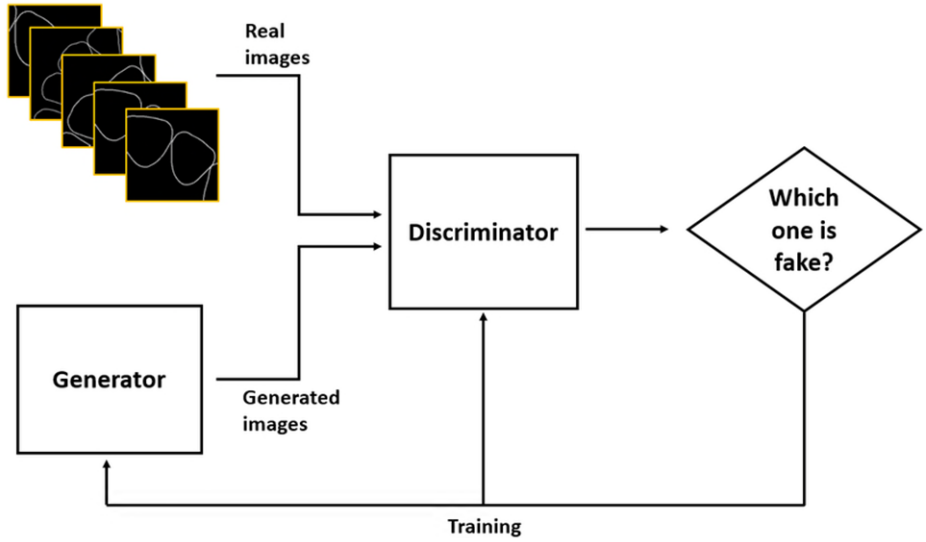
- Επεξεργασία φυσικής γλώσσας
- Ανάλυση αλληλουχιών γονιδιώματος
- Ανάλυση ηχητικών σημάτων
- Δεδομένα χρονοσειρών
- Ανάκτηση πληροφορίας

Generative Adversarial Neural Networks

Τα GAN είναι γενετικοί αλγόριθμοι εκμάθησης που παράγουν νέα παραδείγματα δεδομένων παρόμοια με τα δεδομένα εκπαίδευσης. Το GAN περιλαμβάνει δύο μέρη: μια γεννήτρια που μαθαίνει να παράγει ψεύτικα δεδομένα και μια μέθοδο διάκρισης που αποκτά γνώσεις από αυτά τα δεδομένα. Έχουν γίνει όλο και πιο δημοφιλή με την πάροδο του χρόνου. Για παράδειγμα, μπορούν να βελτιώσουν φωτογραφίες αστρονομίας και να μιμηθούν βαρυτικούς φακούς για την έρευνα σκοτεινής ύλης. Επιπλέον, τα GAN χρησιμοποιούνται από τους παραγωγούς βιντεοπαιχνιδιών για την αναβάθμιση των γραφικών χαμηλής ανάλυσης 2D σε παλαιότερα παιχνίδια χρησιμοποιώντας εκπαίδευση εικόνας για την αναδημιουργία τους σε 4K ή υψηλότερες αναλύσεις.

Η λειτουργία τους μπορεί να αναλυθεί σε 3 βασικά βήματα:

1. Ο διακριτής μαθαίνει να λέει τη διαφορά μεταξύ των πλαστών δεδομένων που δημιουργούνται από τη γεννήτρια και των πραγματικών δειγμάτων.
2. Η γεννήτρια παράγει δόλια δεδομένα κατά την πρώιμη εκπαίδευση και ο διακριτής μαθαίνει γρήγορα να τα αναγνωρίζει ως τέτοια.
3. Για την ενημέρωση του μοντέλου, το GAN παραδίδει την έξοδο του στη γεννήτρια και τον διακριτή.



Πηγή:researchgate.net/figure/Overview-of-generative-adversarial-network-GAN_fig1_338509383

Εικόνα 14: Η Θεμελιακή Δομή Ενός Generative Adversarial Neural Network.

Πλεονεκτήματα:

- Ικανά να δημιουργήσουν τα δικά τους δεδομένα

Μειονεκτήματα:

- Υψηλή κατανάλωση υπολογιστικής ισχύος
- Δύσκολη εκπαίδευση

Εφαρμογές:

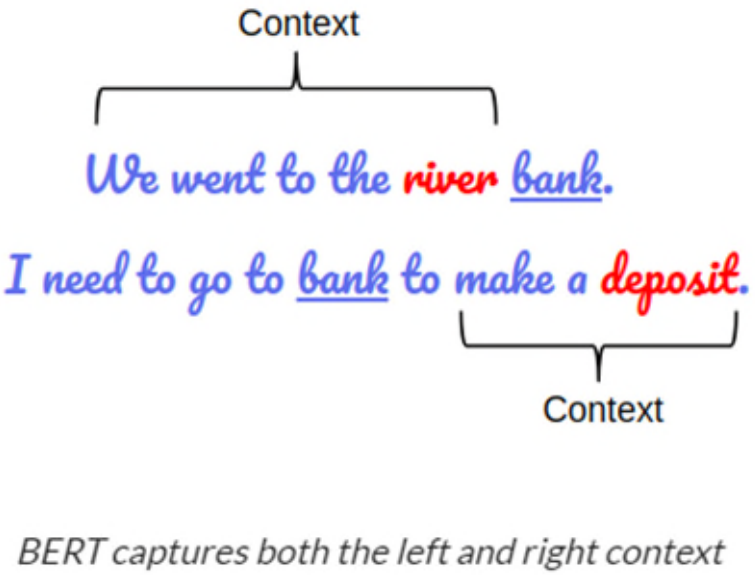
- Δημιουργία Μουσικής
- Δημιουργία εικόνας
- Ανασυγκρότηση εικόνας
- Προσομοιώσεις

Bidirectional Encoder Representations from Transformers - BERT

Η αρχιτεκτονική BERT έχει σχεδιαστεί ώστε να εκπαιδεύει εκ των προτέρων βαθιές αμφίδρομες αναπαραστάσεις από μη επισημασμένο κείμενο, δίνοντας σημασία από κοινού τόσο το αριστερό όσο και το δεξί μέρος μιας πρότασης . Ως αποτέλεσμα, το προ-εκπαιδευμένο μοντέλο BERT μπορεί να ρυθμιστεί με ένα μόνο επιπλέον επίπεδο εξόδου ώστε να δημιουργήσει μοντέλα τελευταίας τεχνολογίας (state-of-the-art) για ένα ευρύ φάσμα εργασιών NLP όπως η ταξινόμηση κειμένου,πρόβλεψη επόμενης λέξης,μετάφραση).

Το BERT είναι προ-εκπαιδευμένο σε μεγάλα σώματα κειμένου κειμένου χωρίς ετικέτα, συμπεριλαμβανομένης ολόκληρης της Wikipedia (δηλαδή 2.500 εκατομμύρια λέξεις!) Και του Book Corpus (800 εκατομμύρια λέξεων). Ένα απο τα κύρια χαρακτηριστικά της επιτυχίας του αποτελεί το στάδιο της προ-εκπαίδευσης. Αυτό συμβαίνει επειδή καθώς εκπαιδεύουμε ένα μοντέλο σε ένα μεγάλο σώμα κειμένου, το μοντέλο μας αρχίζει να συλλέγει τις βαθύτερες και οικείες νοήσεις για το πώς λειτουργεί η γλώσσα και οι συσχετίσεις της.

Το BERT είναι ένα «βαθιά αμφίδρομο» μοντέλο. Διμερής κατεύθυνση (bidirectional) σημαίνει ότι το μοντέλο λαμβάνει υπόψιν πληροφορίες τόσο από την αριστερή όσο και από τη δεξιά πλευρά μιας πρότασης στην προσπάθεια να δημιουργήσει μια ολοκληρωμένη αναπαράσταση της σχέσης μεταξύ κάθε λέξης στην φάση της εκπαίδευσης. Η αμφίδρομη συμπεριφορά ενός μοντέλου είναι σημαντική για την πραγματική κατανόηση του νοήματος μιας γλώσσας. Παρακάτω θα αναφέρουμε ένα παράδειγμα για την καλύτερη κατανόηση του.



Πηγή: analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/

Εικόνα 15: BERT Παράδειγμα Λειτουργίας Αμφίδρομης Συμπεριφοράς

Υπάρχουν δύο προτάσεις στην παραπάνω εικόνα όπου και οι δύο περιλαμβάνουν τη λέξη "τράπεζα". Για να καταφέρει το μοντέλο να αναπαραστήσει το νόημα των παραπάνω λέξεων σωστά θα πρέπει να λάβει υπόψη τόσο το αριστερό όσο και το δεξί μέρος των προτάσεων πρίν προβεί σε τελική πρόβλεψη.

Τέλος, με την ανακάλυψη της αρχιτεκτονικής BERT, η νέα προσέγγιση για την επίλυση εργασιών NLP μετατράπηκε σε μια διαδικασία 2 βημάτων:

1. Εκπαιδεύστε ένα γλωσσικό μοντέλο σε ένα μεγάλο κείμενο χωρίς ετικέτα (χωρίς επίβλεψη ή ημι-εποπτεία)
2. Ρυθμίστε (fine-tune) αυτό το μεγάλο μοντέλο σε συγκεκριμένες εργασίες NLP αναλόγως με τις ανάγκες της κάθεμιας ώστε να είστε σε θέση να χρησιμοποιήσετε το μεγάλο αποθετήριο γνώσεων που έχει αποκτήσει αυτό το μοντέλο (υπό επίβλεψη) λόγω της προ-εκπαίδευσης του.

Στο κεφάλαιο 4 θα γίνει ακόμη μεγαλύτερη ανάλυση του τρόπου λειτουργίας του μοντέλου BERT και της λεπτομερής δομής του.

1.4.3 Εφαρμογές

Ιατρική Περίθαλψη

Το Deep Learning έχει παίξει σημαντικό ρόλο, από την ανάλυση ιατρικής εικόνας έως τις θεραπείες ασθενειών, κυρίως όταν χρησιμοποιούνται επεξεργαστές GPU. Επίσης δίνει την δυνατότητα στους γιατρούς να βοηθήσουν τους ασθενείς τους να ξεφύγουν από τον κίνδυνο, να τους διαγνώσουν και να τους θεραπεύσουν με κατάλληλα φάρμακα. Η ανάλυση ιατρικών δεδομένων με τεχνικές βαθιάς μάθησης έχει γίνει πλέον προτεραιότητα πολλών κρατών καθώς μπορεί να παράγει πολύ καλά αποτελέσματα.

Ανάλυση μετοχών

Οι ποσοτικοί αναλυτές κεφαλαίων χρησιμοποιούν την βαθιά μάθηση ώστε να αναλύσουν τα δεδομένα τους και να δημιουργήσουν ανταγωνιστικό πλεονέκτημα, ιδίως για τον προσδιορισμό τής τάση μιας μετοχής και για το άν η πορεία της θα είναι ανοδική ή καθοδική. Με τα νευρωνικά δίκτυα μπορούν να χρησιμοποιήσουν πολλούς περισσότερους παράγοντες κατά την εκπαίδευση των μοντέλων, όπως ο αριθμός των συναλλαγών που πραγματοποιήθηκαν, ο αριθμός των αγοραστών, ο αριθμός των πωλητών, το κλεισίματος των αγορών της προηγούμενης ημέρας. Κατά την εκπαίδευση των αλγορίθμων βαθιάς μάθησης, οι αναλυτές κεφαλαίων λαμβάνουν υπόψη κριτήρια όπως η απόδοση, ο δείκτης P/E, η απόδοση του ενεργητικού, το μέρισμα, η απόδοση του απασχολούμενου κεφαλαίου, το κέρδος ανά εργαζόμενο, το σύνολο των μετρητών.

Αναγνώριση Εικόνας

Εάν το αστυνομικό τμήμα της πόλης έχει μια βάση δεδομένων για την πόλη και θέλει να μάθει ποιος εμπλέκεται σε εγκλήματα βίας σε δημόσιες συγκεντρώσεις, μπορεί να χρησιμοποιήσει δημόσιες κάμερες για να μάθει ποιος εμπλέκεται. Η βαθιά μάθηση χρησιμοποιώντας τα CNN (Convolution Neural Networks) μπορεί να βοηθήσει στην εύρεση του ατόμου που συμμετείχε στην πράξη.

Ανάλυση ειδήσεων

Η κυβερνήσεις καταβάλουν τεράστιες προσπάθειες τα τελευταία χρόνια για να αποτρέψει τη διάδοση των ψεύτικων ειδήσεων και να καθορίσει την πηγή τους. Επίσης, κατά τη διάρκεια δημοσκοπήσεων όπως το ποιος θα κερδίσει τις εκλογές από άποψη δημοτικότητας, ποιος υποψήφιος έχει συζητηθεί περισσότερο από τους χρήστες στα κοινωνικά μέσα ενημέρωσης και ανάλυση των tweets που έγιναν από ανθρώπους της χώρας, μπορούμε να προβλέψουμε τα αποτελέσματα της δημοσκόπησης χρησιμοποιόντας τεχνικές βαθιάς μάθησης. Ωστόσο, έχει ορισμένους περιορισμούς, όπως η μη γνώση της γνησιότητας των δεδομένων, αν είναι αυθεντικά ή πλαστά ή αν υπάρχουν οι απαραίτητες πληροφορίες.

Αυτοκινούμενα Αμάξια

Το Deep Learning χρησιμοποιείται σε αυτοκινούμενα αμάξια για την ανάλυση δεδομένων που συλλέγονται αυτά όπως: τα δεδομένα από αισθητήρες, δημόσιες κάμερες και άλλες πηγές μπορούν να συλλεχθούν για να βοηθήσουν στη δοκιμή, την εφαρμογή και γενικά την καλύτερη λειτουργία τους. Το σύστημα πρέπει να είναι ικανό να διασφαλίζει ότι όλα τα σενάρια αντιμετωπίζονται σωστά καθ 'όλη τη διάρκεια της εκπαίδευσης.

1.4.4 Δυσκολίες – Προκλήσεις

- Για καλύτερη απόδοση σε σχέση με άλλες στρατηγικές μηχανικής γνώσης, απαιτεί τεράστιο όγκο δεδομένων εκπαίδευσης.

- Λόγω των πολύπλοκων μοντέλων, η εκπαίδευση είναι δαπανηρή. Η βαθιά μάθηση απαιτεί επίσης τη χρήση ακριβών GPU και εκατοντάδων σταθμών εργασίας. Με αποτέλεσμα το κόστος των χρηστών να αυξάνεται.

- Επειδή απαιτεί γνώση αρκετών αρχιτεκτονικών, μεθόδων και άλλων χαρακτηριστικών, δεν υπάρχει τυπική θεωρία που να μας καθοδηγεί στην επιλογή των σωστών εργαλείων βαθιάς μάθησης. Ως αποτέλεσμα, είναι δύσκολο για τους νέους χρήστες να υιοθετήσουν την τεχνολογία.

1.5 Στόχοι Διπλωματικής

Η παρούσα διπλωματική στοχεύει στην εξερεύνηση και υλοποίηση τεχνικών ταξινόμησης κειμένων για την καλύτερη αντιμετώπιση του φαινομένου της ρητορικής μίσους. Στην πειραματική μας προσέγγιση, μετά την εκτενή επεξεργασία και επάυξηση των δεδομένων, υλοποιούμε την αρχιτεκτονική Transformer B.E.R.T με σκοπό την δημιουργία ταξινομητή πολλαπλών ετικετών με 6 κλάσεις στόχους (abusive, hateful, offensive, disrespectful, fearful, normal). Στο τέλος αφού παρουσιάσουμε αναλυτικά τα αποτελέσματα μας, συγκρίνουμε τις επιδόσεις του μοντέλου μας με αυτά παρόμοιων μοντέλων που έχουν προκύψει απο το ίδιο σύνολο δεδομένων.

Για σκοπό των πειραμάτων, κάνουμε εκτενής ανάλυση των διαθέσιμων συνόλων δεδομένων ρητορικής μίσους και μέσω της προτεινόμενης μεθοδολογίας μας επιλέγουμε το πιο κατάλληλο, απαλαγμένο όσο το δυνατόν περισσότερο απο φαινόμενα φυλετικής προκατάληψης, προκατάληψης σχολιασμού και έλλειψη επαρκή ορισμού.

2

Ταξινόμηση Κειμένου για Ανίχνευση Ρητορικής Μίσους

Σε αυτό το κεφάλαιο αναλύουμε σε βάθος τον τομέα της ταξινόμησης κειμένου ρητορικής μίσους. Ξεκινάμε με την παρουσίαση τον τύπων ταξινόμησης και των υποκατηγοριών τους, συνεχίζουμε με την παρουσίαση των μετρικών συστημάτων επίδοσης και ολοκληρώνουμε το κεφάλαιο κάνοντας εκτενή ανάλυση της πρόσφατης επιστημονικής βιβλιογραφίας στον τομέα της αναγνώρισης ρητορικής μίσους.

2.1 Τύποι Ταξινόμησης

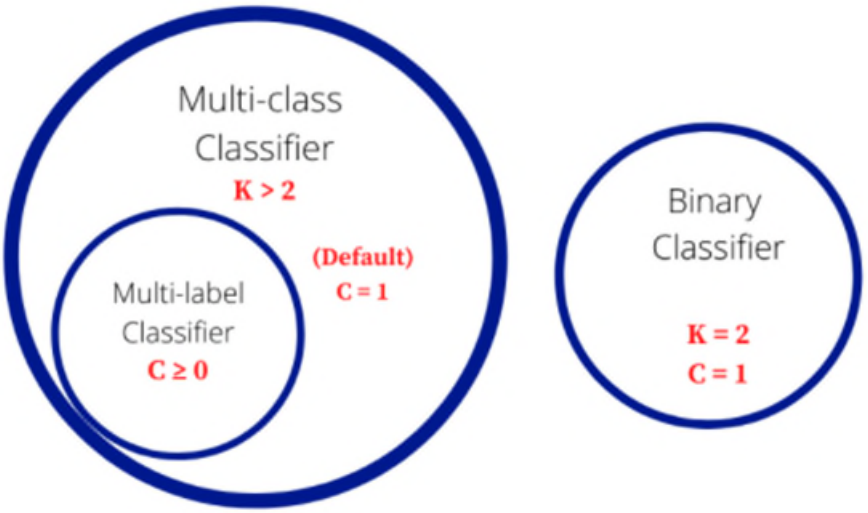
Ταξινόμηση στόν τομέα της μηχανική μάθησης είναι η διαδικασία κατα την οποία ενα σύνολο δεδομένων ταξινομείται έτσι ώστε κάθε παράδειγμα απο τα δεδομένα να ανήκει σε ένα σύνολο προκαθορισμένων ετικετών-κλάσσεων.

Ακολουθούν ορισμένα παραδείγματα ζητημάτων ταξινόμησης:

- Καθορισμός ενός μυνήματος ηλεκτρονικού ταχυδρομείου (email) ώς ανεπιθύμητο ή οχι.
- Ταξινόμηση ενός κειμένου ως μισητού ή μη μισητού.
- Ταξινόμηση ενός κειμένου ως μισητού/όχι μισητού/υβριστικού/φυσιολογικού.

Όσον αφορά τη μοντελοποίηση, η ταξινόμηση χρειάζεται ένα μεγάλο σύνολο δεδομένων εκπαίδευσης με πολλά παραδείγματα για να μάθει. Επίσης, πρέπει να είναι επαρκώς αντιπροσωπευτικό του προβλήματος και να περιέχει πολλά δείγματα από την ετικέτα κάθε τάξης.

Οι ετικέτες κλάσης είναι συχνά τιμές συμβολοσειράς, όπως "ανεπιθύμητο", "όχι ανεπιθύμητο" και πρέπει να μετατραπούν σε αριθμητικές τιμές πριν εισαχθούν σε μια διαδικασία μοντελοποίησης. Αυτό συχνά ονομάζεται κωδικοποίηση ετικετών, όπου εκχωρείται ένας μοναδικός ακέραιος αριθμός σε κάθε ετικέτα κλάσης, π.χ. "spam" = 0, "no-spam" = 1. Τέλος, το μοντέλο πρέπει να ελεγχθεί και με βάση το σύνολο δεδομένων ελέγχου ώστε να διαπιστώσουμε εάν τα αποτελέσματα του είναι ικανοποιητικά. Η ακρίβεια ταξινόμησης είναι μια τυπική μέτρηση για την αξιολόγηση της απόδοσης ενός μοντέλου με βάση τις προβλεπόμενες ετικέτες τάξης.



K = Total number of classes in the problem statement
C = Number of classes an item maybe assigned to

Πηγή: serokell.io/blog/classification-algorithms

Εικόνα 16: Οι 4 Διαφορετικές Κατηγορίες Ταξινόμησης

2.1.1 Δυαδική Ταξινόμηση – Binary Classification

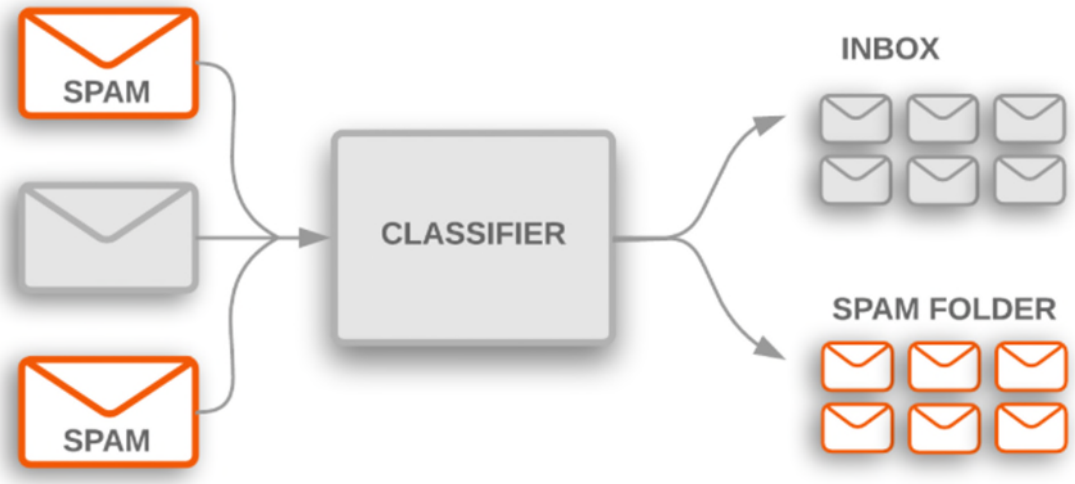
Η δυαδική ταξινόμηση ανήκει σε εργασίες ταξινόμησης που έχουν μόνο δύο ετικέτες κλάσης.

Τα παραδείγματα περιλαμβάνουν:

- Εντοπισμός ανεπιθύμητων μηνυμάτων (ανεπιθύμητα ή μη).
- Πρόβλεψη μετατροπής (αγορά ή όχι).
- Ανάλυση Κειμένου (επιθετικό/μη-επιθετικό)

Στα περισσότερα προβλήματα δυαδικής ταξινόμησης, η μία τάξη αντιπροσωπεύει την κανονική κατάσταση και η άλλη αντιπροσωπεύει την παρεκκλίνουσα κατάσταση. Για παράδειγμα, εάν η φυσιολογική κατάσταση είναι "όχι spam", ενώ η μη φυσιολογική κατάσταση είναι "spam", στην κανονική κατηγορία κανονικής κατάστασης αποδίδεται η ετικέτα κλάσης 0, ενώ στην κατηφορία ανώμαλης κατάστασης απονέμεται η ετικέτα κλάσης 1.

Ορισμένοι αλγόριθμοι, όπως Logistic Regression και Support Vector Machines που μελετήσαμε σε προηγούμενο κεφάλαιο, έχουν δημιουργηθεί κυρίως για δυαδική ταξινόμηση και δεν υποστηρίζουν παραπάνω απο δύο κατηγορίες ετικετών κλάσσης.



Πηγή: developers.google.com/machine-learning/guides/text-classification/

Εικόνα 17: Παράδειγμα Δυαδικής Ταξινόμησης σε Εφαρμογή Εντοπισμού Ανεπιθύμητων Μηνυμάτων Ηλεκτρονικού Ταχυδρομείου

2.1.2 Ταξινόμηση Πολλών Κλάσεων – Multiclass Classification

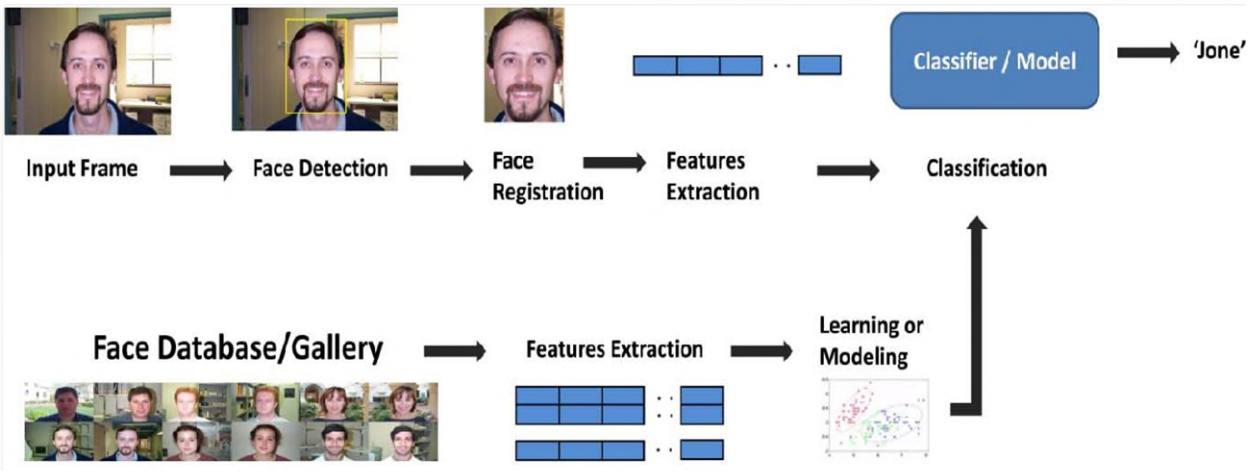
Οι εργασίες ταξινόμησης με περισσότερες από δύο ετικέτες κατηγορίας κλάσσης αναφέρονται ως ταξινόμηση πολλαπλών κλάσσεων.

Τα παραδείγματα περιλαμβάνουν:

- Αναγνώριση Προσώπου (‘γωνίες προσώπου’,’μάτια’, ‘αυτιά’, ...)
- Ταξινόμηση Φυτικών Ειδών (‘παπαρούνα’,’τριαντάφυλλο’,’χρυσάνθεμο’, ...)
- Οπτική Αναγνώριση Χαρακτήρων (‘Α’, ‘Β’, ‘Γ’, ...)

Η ταξινόμηση πολλαπλών κλάσεων, σε αντίθεση με τη δυαδική ταξινόμηση, δεν κάνει διάκριση μεταξύ φυσιολογικών και ανώμαλων καταστάσεων. Αντ 'αυτού, παραδείγματα ανατίθενται σε μία από τις πολλές προκαθορισμένες κλάσεις. Σε ορισμένες περιπτώσεις, ο αριθμός των ετικετών τάξης μπορεί να είναι αρκετά μεγάλος.

Για παράδειγμα, σε ένα σύστημα αναγνώρισης προσώπου, ένα μοντέλο μπορεί να προβλέψει ότι ένα πλάνο ανήκει σε ένα από τα χιλιάδες ή δεκάδες χιλιάδες πρόσωπα.



Πηγή: semanticscholar.org/paper/Face-recognition-system-using-bag-of-features-and-Nasr-Bouallegue/5d8233c1cc3c38eb286912a644dc56f144ea9782

Εικόνα 18: Παράδειγμα Ταξινόμησης Πολλών Κλάσσεων σε Εφαρμογή Αναγνώρισης Προσώπου

2.1.3 Ταξινόμηση Πολλαπλών Ετικετών – Multi-label Classification

Οι εργασίες ταξινόμησης με δύο ή περισσότερες ετικέτες κατηγορίας, όπου κάθε παράδειγμα μπορεί να λάβει απο καμία εώς όλες τις ετικέτες κατηγορίας, αναφέρεται ως ταξινόμηση πολλαπλών ετικετών. Παρατηρούμε πως διαφέρει από την δυαδική και την ταξινόμηση πολλαπλών κλάσεων, η οποία προβλέπει μία ετικέτα κατηγορίας για κάθε παράδειγμα. Παρακάτω παρουσιάζουμε κάποια παραδείγματα ταξινόμησης πολλαπλών ετικετών.

Κατηγοριοποίηση Τραγουδιών

Κατηγοριοποίηση τραγουδιών σε διάφορα είδη ή κατηγοριοποίηση με βάση το συναίσθημα ή την διάθεση, όπως "χαλαρωτικό-ήρεμο", "λυπητερό-μοναχικό" και ούτω καθεξής.

Κατηγοριοποίηση Εικόνας

Η κατηγοριοποίηση πολλών ετικετών με βάση την εικόνα προσφέρει ένα ευρύ φάσμα εφαρμογών. Οι εικόνες μπορούν να επισημανθούν για να αντιπροσωπεύουν διάφορα αντικείμενα, άτομα ή ιδέες.

Κατηγοριοποίηση Κειμένου

Η προσπάθεια να ταξινομηθεί ένα κείμενο με βάση τα χαρακτηριστικά του (επιθετικό/ρατσιστικό/φιλικό/κανονικό/σεξιστικό/...)

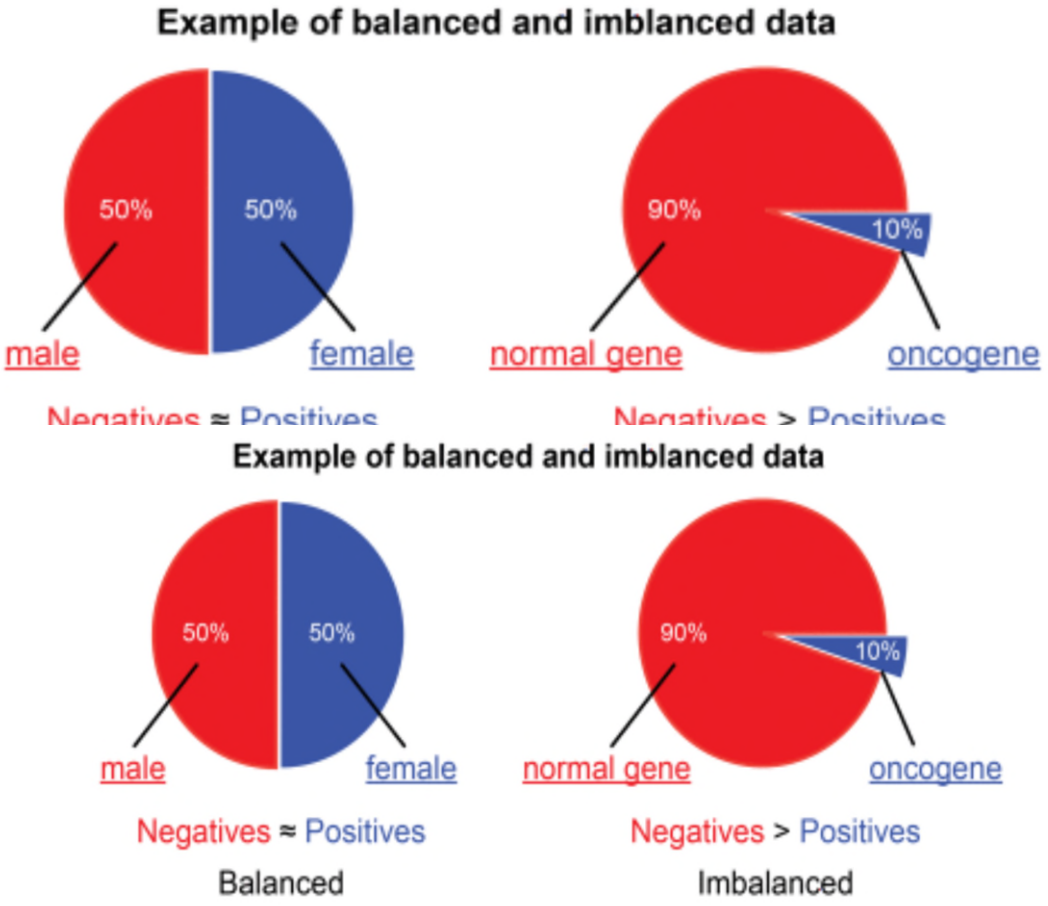
Πηγή: https://medium.com/@saugata.paul1010/a-detailed-case-study-on-multi-label-classification-with-machine-learning-algorithms-and-72031742c9aa

Εικόνα 19: Παράδειγμα Ταξινόμησης Πολλών Ετικετών σε Εφαρμογή Αναγνώρισης Εικόνας

2.1.4 Μη ισορροπημένη ταξινόμηση – Imbalanced Classification

Η μη ισορροπημένη ταξινόμηση αναφέρεται σε προβλήματα στα οποία ο αριθμός των παραδειγμάτων σε κάθε τάξη κατανέμεται άνισα. Δηλαδή, η πλειοψηφία των παραδειγμάτων δεδομένα εκπαίδευσης ανήκει στην κανονική τάξη ενώ η ανώμαλη τάξη έχει μια μειοψηφία παραδειγμάτων.

Τα παραδείγματα περιλαμβάνουν:



υπερ - πήσουμε οση του

Πηγή: medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5

Εικόνα 20: Παράδειγμα Μη Ισσοροπημένης Κατανομής Κλάσεων

2.2 Μετρικά Συστήματα Επίδοσης Ταξινόμησης

Τα μετρικά συστήματα επίδοσης χρησιμοποιούνται ώστε να μετρηθεί η επίδοση του μοντέλου μετά το πέρας της φάσης της εκπαίδευσης του. Το μέτρο της επίδοσης ενός μοντέλου μετράται όταν του παρουσιάζονται νέες περιπτώσεις παραδειγμάτων που δεν ήταν μέρος των δεδομένων εκπαίδευσης του. Παρακάτω θα αναλύσουμε τα βασικότερα μετρικά συστήματα που χρησιμοποιούνται σε προβλήματα ταξινόμησης.

Accuracy

Η ευστοχία είναι το ποσοστό των σωστών αποτελεσμάτων μεταξύ του συνολικού αριθμού των περιπτώσεων που εξετάστηκαν. Όποτε χρησιμοποιείται η μέτρηση ευστοχίας, ουσιαστικά, στοχεύουμε στο να μάθουμε την εγγύτητα μιας τυχαίας τιμής σε σχέση με μια ήδη γνωστή. Επομένως χρησιμοποιείται συνήθως σε περιπτώσεις όπου η μεταβλητή εξόδου είναι κατηγορική ή διακριτή – Δηλαδή, σε περιπτώσεις ταξινόμησης. Η ευστοχία αποτελεί μια έγκυρη επιλογή αξιολόγησης για προβλήματα ταξινόμησης στα οποία τα παραδείγματα του συνόλου δεδομένων είναι καλά ισορροπημένα μεταξύ των διαφορετικών κατηγοριών.

Precision

Σε περιπτώσεις που μας απασχολεί το πόσο ακριβείς είναι οι προβλέψεις του μοντέλου μας, καλή επιλογή αποτελεί το Precision. Η μέτρηση ακριβείας μας ενημερώνει για τον αριθμό των ετικετών που όντως χαρακτηρίζονται ως θετικές σε αντιστοιχία με τις περιπτώσεις που ο ταξινομητής χαρακτηρίζει ως θετικές. Η ακρίβεια είναι μια έγκυρη επιλογή μέτρησης αξιολόγησης όταν θέλουμε να είμαστε πολύ σίγουροι για την πρόβλεψή μας.

Recall

Η ανάκληση μετρά πόσο καλά το μοντέλο μπορεί να ανακαλέσει τη θετική κατηγορία (δηλαδή τον αριθμό των θετικών ετικετών που το μοντέλο αναγνώρισε ως θετικές). Η ανάκληση είναι μια έγκυρη επιλογή μέτρησης αξιολόγησης όταν θέλουμε να αποτυπώσουμε όσο το δυνατόν περισσότερα θετικά παραδείγματα.

F1 Score

Η βαθμολογία F1 είναι ένας αριθμός μεταξύ 0 και 1 και αποτελεί τον αρμονικό μέσο όρο μεταξύ της ακρίβειας και της ανάκλησης. Η ακρίβεια και η ανάκληση είναι συμπληρωματικές μετρήσεις που έχουν αντίστροφη σχέση. Αν μας ενδιαφέρουν εξίσου και οι δύο μετρήσεις, τότε θα χρησιμοποιούσαμε τη βαθμολογία F1 για να συνδυάσουμε την ακρίβεια και την ανάκληση σε μια ενιαία μέτρηση.

AUC ROC

AUC ονομάζεται η περιοχή κάτω από την καμπύλη ROC. Η μέτρηση αυτή μας δείχνει πόσο καλά διαχωρίζονται οι πιθανότητες στις θετικές κατηγορίες από τις αρνητικές. Ουσιαστικά, το AUC μας βοηθά να ποσοτικοποιήσουμε την ικανότητα του μοντέλου μας να διαχωρίζει τις κλάσεις καταγράφοντας τον αριθμό των θετικών προβλέψεων που είναι σωστές έναντι του αριθμού των θετικών προβλέψεων που είναι λανθασμένες.

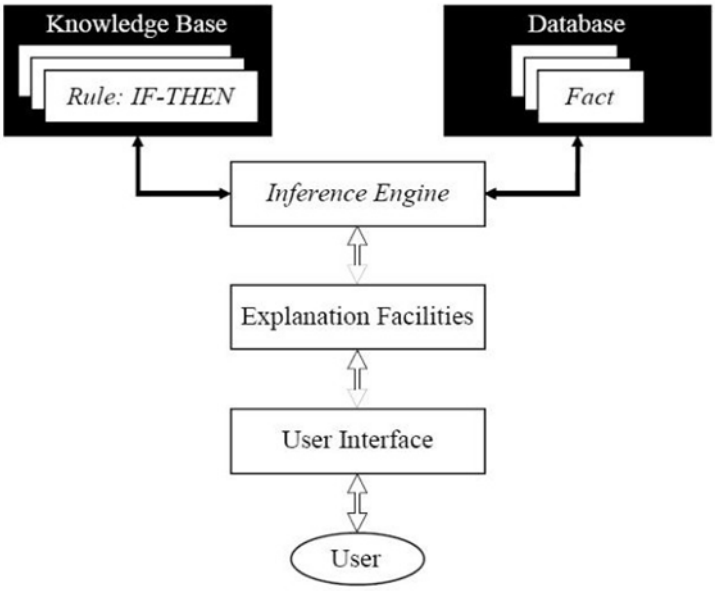
2.3 Συστήματα Βασισμένα σε Κανόνες

Μια τεχνητή νοημοσύνη που βασίζεται σε κανόνες παράγει προκαθορισμένα αποτελέσματα βασισμένα σε ένα σύνολο κανόνων που έχουν κωδικοποιηθεί από τον άνθρωπο. Αυτά τα συστήματα είναι απλά μοντέλα τεχνητής νοημοσύνης που χρησιμοποιούν τον κανόνα κωδικοποίησης αν-τότε ώστε να μοντελοποιήσουν το πρόβλημα.

Στα συστήματα βασισμένα σε κανόνες, οι οδηγίες παρέχονται με τη μορφή τεσσάρων βασικών στοιχείων:

1. **Γεγονότα, βάση γνώσεων ή ένα μείγμα γεγονότων**, όπως η τιμή πετρελαίου 80Ευρώ/βαρέλι.
2. **Ένα σύνολο κανόνων**, όπως "Αν συμβεί το Χ, τότε κάντε το Ψ."
3. **Μηχανή συμπερασμάτων**, η οποία επεξεργάζεται τις πληροφορίες και ανταποκρίνεται με βάση τις αλληλεπιδράσεις μεταξύ της εισόδου και της βάσης κανόνων.
4. **Μνήμη εργασίας**

Ωστόσο, ένα μικρό λάθος σε μια προσέγγιση βασισμένη σε κανόνες, απαιτεί αρκετό χρόνο και προσπάθεια ώστε να επιλυθεί. Για το λόγο αυτό, τα συστήματα τεχνητής νοημοσύνης που βασίζονται σε κανόνες συχνά προτιμώνται για έργα περιορισμένης κλίμακας που συνεπάγονται περιορισμένη προσπάθεια, κόστος και ενημερώσεις.



Πηγή: artificialintelligence.oodles.io/wp-content/uploads/2020/06/basic-structure-of-a-rule-based-expert-system-l.jpg

Εικόνα 21: Παράδειγμα Γενικής Δομής Συστημάτων Βασισμένων σε Κανόνες

2.3.1 Επιστημονικό Έργο

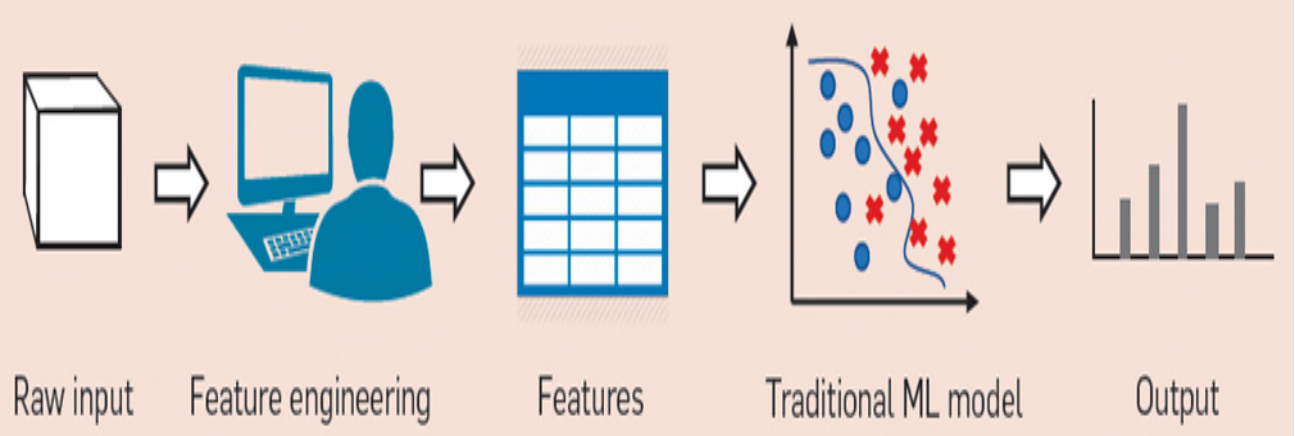
Το 2014, οι *C. J. Hutto et al.* προτείναν μια προσέγγιση για την ταξινόμηση κειμένου με βάση το συναίσθημα (sentiment analysis) χρησιμοποιώντας τη μέθοδο VADER, η οποία είναι μια προσέγγιση βασισμένη σε κανόνες[27]. Αρχικά, δημιούργησαν μια λίστα λεξικών χαρακτηριστικών που είναι ιδιαίτερα ευαίσθητα στο συναίσθημα των αναρτήσεων στα μέσα κοινωνικής δικτύωσης. Στη συνέχεια, συνδύασαν αυτόν τον κατάλογο λεξικών χαρακτηριστικών με πέντε γενικούς κανόνες που περικλείουν συντακτικούς και γραμματικούς κανόνες για την παρουσίαση της έντασης του συναισθήματος. Τέλος, διαπίστωθηκε ότι το VADER πέτυχε 96% ακρίβεια όταν δοκιμάστηκε στο σύνολο δεδομένων του twitter.

Το 2015, οι *Dennis Gitariet al.* προτείναν μια μέθοδο για τον προσδιορισμό της ανάλυσης συναισθημάτων του κειμένου (tweets) χρησιμοποιώντας μια μέθοδο βασισμένη σε κανόνες [28]. Σε αυ΄τη την έρευνα, κατηγοριοποίησαν το πρόβλημα της ρητορικής μίσους σε τρεις τομείς: θρησκεία, εθνικότητα και φυλή. Ο κύριος στόχος αυτής της εργασίας είναι να αναπτύξει ένα μοντέλο ταξινόμησης που στοχεύει στην ανάλυση συναισθημάτων. Το αναπτυγμένο μοντέλο όχι μόνο ανιχνεύει υποκειμενικές προτάσεις αλλά ταξινομεί και κατατάσσει τις φράσεις με βάση την πολικότητα τους. Στη συνέχεια, συνδέουν τα σημασιολογικά και υποκειμενικά χαρακτηριστικά με τη ρητορική μίσους. Τέλος, πέτυχαν ακρίβεια 71,55 % χρησιμοποιώντας αυτή την προσέγγιση.

2.4 Συστήματα Βασισμένα σε Μηχανική Μάθηση

Η μηχανική μάθηση, σε αντίθεση με τα συστήματα που βασίζονται σε κανόνες, διαμορφώνει την ανθρώπινη νοημοσύνη για την ολοκλήρωση εργασιών χωρίς να χρειάζεται να προγραμματιστεί ρητά. Με τη μορφή μεγάλων συνόλων δεδομένων που επισημαίνονται χειροκίνητα για την εκπαίδευση μηχανών, αυτή η πληροφορία/γνώση εγχέεται στα μοντέλα μηχανικής μάθησης.

Τα μοντέλα μηχανικής μάθησης συνεχώς "προσαρμόζουν" και "εξελίσσουν" την απόδοσή τους ως απάντηση σε συνεχή ροή δεδομένων εκπαίδευσης, αντί να παράγουν συγκεκριμένα αποτελέσματα. Ως αποτέλεσμα, τα συστήματα μηχανικής μάθησης μπορούν να μάθουν από τις ανθρώπινες εμπειρίες και να βελτιώσουν την απόδοση τους με βάση τα παρεχόμενα δεδομένα.



Πηγή:cacm.acm.org/magazines/2020/1/241703-techniques-for-interpretable-machine-learning/fulltext

Εικόνα 22: Παράδειγμα Γενικής Δομής Συστημάτων Βασισμένων σε Μηχανική Μάθηση

2.4.1 Επιστημονικό Έργο

Το 2017, οι Fatahillah et al., χρησιμοποίησαν τον αλγόριθμο ταξινόμησης Naive Bayes για τον εντοπισμό ρητορικής μίσους στο Instagram χρησιμοποιώντας τον ταξινομητή k-πλησιέστερου γείτονα [29]. Συγκέντρωσαν το σύνολο δεδομένων χρησιμοποιώντας το API της Twitter (μέσο κοινωνικής δικτύωσης), ο σχολιασμός των δεδομένων πραγματοποιήθηκε χειροκίνητα. Μετά από την προεπεξεργασία των δεδομένων, εφάρμοσαν τον αλγόριθμο ταξινόμησης Naive Bayes και πέτυχαν ακρίβεια της τάξης του 93%.

Το 2018, οι M. Ali Fauzi et al., πρότειναν μια προσέγγιση για τον προσδιορισμό της ρητορικής μίσους χρησιμοποιώντας ένα σύνολο αλγορίθμων μάθησης με επίβλεψη [30]. Συνέθεσαν πέντε διαφορετικούς αλγορίθμους ταξινόμησης, συμπεριλαμβανομένων των K-Nearest Neighbours, Random Forest, Naive Bayes, Support Vector Machine και Maximum Entropy. Συγκέντρωσαν το σύνολο δεδομένων χρησιμοποιώντας το API του Twitter (μέσο κοινωνικής δικτύωσης) και ο σχολιασμός των δεδομένων πραγματοποιήθηκε χειροκίνητα. Στην φάση της προεπεξεργασίας των δεδομένων, χρησιμοποίησαν μεθόδους διακριτοποίησης, φιλτραρίσματος και όρων στάθμισης. Χρησιμοποίησαν την τεχνική σακούλα των λέξεων (bag of words) μαζί με τεχνικές όρων στάθμισης (TFIDF). Ο αφελής αλγόριθμος Bayes απέδωσε καλύτερα με 78,3 % ακρίβεια μεταξύ όλων των άλλων πέντε αυτόνομων ταξινομητών.

Το 2019, οι P. Sari et al. πρότειναν μια προσέγγιση για τον εντοπισμό της ρητορικής μίσους χρησιμοποιώντας την υλικοτεχνική οπισθοδρόμηση στο Twitter (μέσο κοινωνικής δικτύωσης). [31] Συγκέντρωσαν τα δεδομένα από το Twitter και χρησιμοποίησαν μεθόδους Cold Folding, Tokenizing, Filtering και Stemming στην φάση της προεπεξεργασίας. Μετά την προεπεξεργασία, η τεχνική στάθμισης όρων TF-IDF χρησιμοποιήθηκε για τη διανυσματοποίηση (vectorization). Μετά, εφαρμόστηκε ο αλγόριθμος Logistic regression. Τέλος, η προσέγγιση τους πέτυχε ακρίβεια της τάξης του 84%.

Το 2020, οι Oluwafemi Oriola et al. πρότειναν μια προσέγγιση για τον εντοπισμό προσβλητικής ομιλίας στο twitter [32]. Οι συγγραφείς συνέλεξαν το σύνολο δεδομένων χρησιμοποιώντας το API του Twitter και σχολίασαν αυτά τα δεδομένα σε δύο ενότητες, ελεύθερη ομιλία «FS» και ρητορική μίσους «HS». Κατά τη φάση της προεπεξεργασίας δεδομένων, αφαίρεσαν ειδικούς χαρακτήρες, emoji, σημεία στίξης, σύμβολα, hashtags, λέξεις στάσης για να «καθαρίσουν» τα δεδομένα. Έπειτα, χρησιμοποίησαν την τεχνική στάθμισης όρων (TF-IDF) για να μετατρέψουν το κείμενο σε διανύσματα. Μετά την εφαρμογή μίας βελτιστοποιημένης μορφής του αλγορίθμου SVM με n-gram, πέτυχαν ποσοστά ακρίβειας της τάξης του 89,4%.

Το 2020, οι Annisa Briliani et al. Πρότειναν μια προσέγγιση για τον προσδιορισμό της ρητορικής μίσους στο Instagram χρησιμοποιώντας τον αλγόριθμο ταξινόμησης k-κοντινότερου[33]. Συγκέντρωσαν το σύνολο δεδομένων χρησιμοποιώντας το Instagram API από το Instagram και ο σχολιασμός των δεδομένων πραγματοποιήθηκε χειροκίνητα. Χώρισαν το σύνολο δεδομένων σε 2 ετικέτες, δηλαδή (δυαδική ταξινόμηση). Στη φάση της προεπεξεργασίας, καθάρισαν τα δεδομένα και χρησιμοποίησαν την τεχνική στάθμισης όρων (TF-IDF). Στη συνέχεια, εφάρμοσαν τον αλγόριθμο k-πλησιέστερου γείτονα και πέτυχαν ακρίβεια της τάξης του 98,13%.

Το 2015, οι Rui Zhao et al., πρότειναν μια προσέγγιση για τον εντοπισμό του διαδικτυακού εκφοβισμού με τη χρήση σημασιολογικά ενισχυμένου αυτόματου κωδικοποιητή (auto encoder) μείωσης θορύβου (smSDA) [34]. Θεωρούν πως η προτεινόμενη μέθοδος είναι σε θέση να εκμεταλλευτεί τη δομή των κρυφών χαρακτηριστικών των δεδομένων εκφοβισμού και να μάθει μια ισχυρή και διακριτική αναπαράσταση του κειμένου. Χρησιμοποίησαν δύο πηγές συνόλων

δεδομένων. Η πρώτη πηγή είναι το Twitter και η δεύτερη πηγή είναι το Myspace. Τα δεδομένα του Twitter συλλέχθηκαν μέσω του API ροής Twitter και τα δεδομένα του Myspace συλλέχθηκαν χρησιμοποιώντας την τεχνική ανίχνευσης ιστού (web crawling). Η προσέγγιση τους πετυχαίνει ακρίβεια της τάξης του 84,9% χρησιμοποιώντας το smSDA στο σύνολο δεδομένων του Twitter και ακρίβεια της τάξης του 89,7% χρησιμοποιώντας το smSDA στο σύνολο δεδομένων του MySpace.

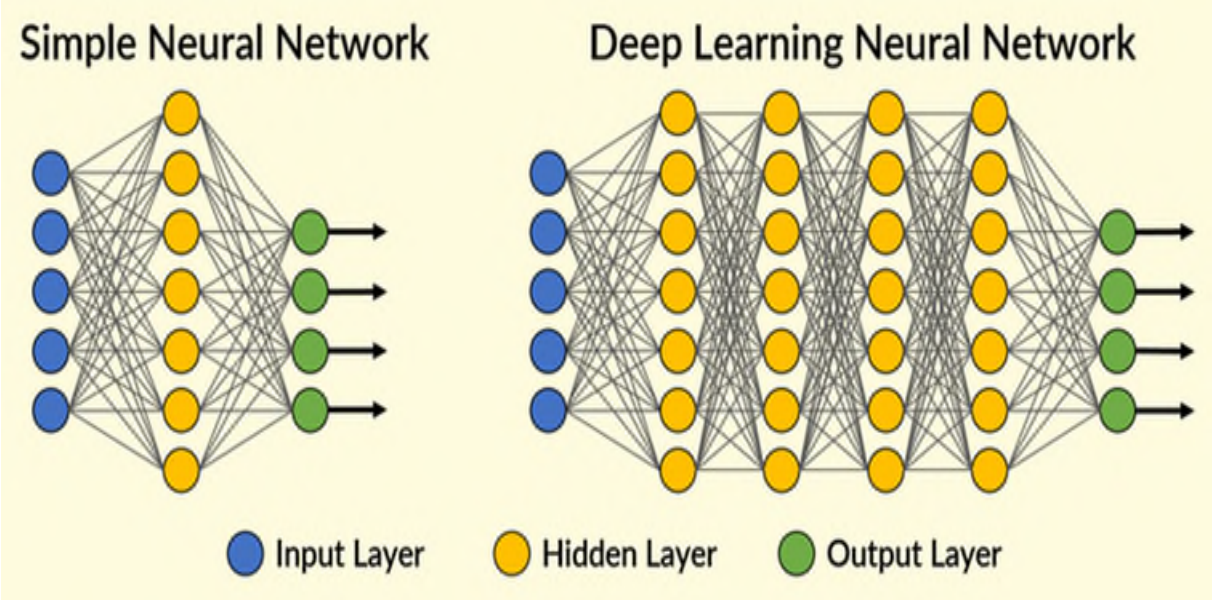
Το 2019, οι Axel Rodríguez et al., πρότειναν μια προσέγγιση για τον εντοπισμό περιεχομένου ρητορικής μίσους χρησιμοποιώντας ανάλυση συναισθημάτων στο Facebook [35]. Χρησιμοποίησαν το Graph API για να εξαγάγουν τις αναρτήσεις και τα σχόλια αυτών από το Facebook. Για την αφαίρεση των άσχετων κειμένων χρησιμοποιήθηκε η μέθοδος VADER. Στην φάση της προεπεξεργασίας δεδομένων, φιλτράρισαν όλες τις περιττές λέξεις -κλειδιά ή σύμβολα. Έπειτα, τα προ-επεξεργασμένα κείμενα μετατράπηκαν σε διάνυσμα χρησιμοποιώντας την μέθοδο στάθμισης όρων (TFIDF). Τέλος, τα διανύσματα αυτά περνούν στον αλγόριθμο ομαδοποίησης k-means ως εισόδος.

Το 2019, οι Sylvia Jaki et al., επέδειξαν μια προσέγγιση για τον εντοπισμό περιεχομένου ρητορικής μίσους χρησιμοποιώντας μηχανική μάθηση χωρίς επίβλεψη μη επίβλεψη στο Twitter [36]. Συγκέντρωσαν πάνω από 50,00 σύνολα δεδομένων χρησιμοποιώντας το Twitter API. Χρησιμοποίησαν τεχνικές NLP για να ομαδοποιήσουν τις λέξεις σε παρόμοιες ομάδες. Υπολόγισαν τρεις ομάδες από τους κορυφαίους 250 πιο προκατειλημμένους όρους χρησιμοποιώντας k-means και skip-gram τεχνικές. Ως αποτέλεσμα, πέτυχαν βαθμολογία F1 της τάξης του 84,21%.

2.5 Συστήματα Βασισμένα σε Βαθιά Μάθηση

Τα συστήματα βαθιάς μάθησης αποτελούν ένα υποσύνολο τεχνικών μηχανικής μάθησης που κάνουν χρήση πολλών επιπέδων μη γραμμικής επεξεργασίας πληροφοριών για την επίβλεψη και χωρίς επίβλεψη εξαγωγή χαρακτηριστικών, μετασχηματισμό δεδομένων, ανάλυση προτύπων και ταξινόμηση στοιχείων/κειμένων. Αποτελούνται από πολλαπλά ιεραρχικά επίπεδα που επεξεργάζονται δεδομένα με μη γραμμικό τρόπο και με ορισμένες έννοιες χαμηλότερου επιπέδου να βοηθούν στον προσδιορισμό των εννοιών υψηλότερου επιπέδου.

Πολλές κοινές εφαρμογές, όπως η φυσική ομιλία, η ανάλυση εικόνας, η ανάκτηση πληροφοριών και άλλες εφαρμογές επεξεργασίας πληροφοριών, δείχνουν ότι τα ρηχά τεχνητά νευρωνικά δίκτυα δεν μπορούν να διαχειριστούν ένα σημαντικό αριθμό περίπλοκων δεδομένων ενώ η βαθιά μάθηση είναι κατάλληλη για τέτοιες εργασίες. Η βαθιά μάθηση επιτρέπει σε ένα μηχάνημα να αναγνωρίζει, να ταξινομεί και να κατηγοριοποιεί μοτίβα στα δεδομένα με πολύ λιγότερη προσπάθεια.



Πηγή:analyticsvidhya.com/blog/2021/07/ai-vs-ml-vs-dl-lets-understand-the-difference/

Εικόνα 23: Παράδειγμα Γενικής Δομής Συστημάτων Βασισμένων σε Βαθιά Μάθηση

2.5.1 Επιστημονικό Έργο

Το 2018, οι Hugo Rosa et al., πρότειναν μια προσέγγιση για τον εντοπισμό του διαδικτυακού εκφοβισμού χρησιμοποιώντας μεθόδους της βαθιάς μάθησης [37]. Το σύνολο δεδομένων εκπαίδευσης και δοκιμών συλλέχθηκε από το Kaggle. Το μοντέλο τους ξεκινά με ένα επίπεδο CNN, συνεχίζει με ένα εντελώς συνδεδεμένο επίπεδο (fully connected layer) με επίπεδο εγκατάλειψης (dropout layer) 0,5 και τέλος, συνάρτηση ενεργοποίησης softmax. Στη συνέχεια συνδύασαν το μοντέλο με ένα επίπεδο LSTM ώστε να επιτύχουν τη μέγιστη ακρίβεια. Πέτυχαν ακρίβεια 64,9% με ενσωματώσεις (embeddings) της google.

Το 2019, οι Tin Van Huynh et al., πρότειναν μια προσέγγιση για τον εντοπισμό ρητορικής μίσους χρησιμοποιώντας το μοντέλο Bi-GRU-CNN-LSTM [38]. Σε αυτό το έγγραφο, συνέλεξαν δεδομένα από το Twitter και κατηγοριοποίησαν τα δεδομένα τους σε τρεις ετικέτες (OFFENSIVE, HATE και CLEAN). Μετά τον καθαρισμό των δεδομένων, εφάρμοσαν τρία μοντέλα νευρωνικών δικτύων όπως τα BiGRU-LSTM-CNN, Bi-GRU-CNN και TextCNN για τον εντοπισμό ρητορικής μίσους. Η καλύτερη τους απόδοση στην F1 βαθμολογία έφτασε τα επίπεδα του 70,57%.

Το 2019, οι Gambäck et al., πρότειναν έναν αλγόριθμο βαθιάς εκμάθησης για τον εντοπισμό ρητορικής μίσους στο Twitter [39]. Σε αυτή την έρευνα, συνέλεξαν δεδομένα από το Twitter και χώρισαν τα δεδομένα σε τέσσερις κατηγορίες (σεξισμός, ρατσισμός, συνδυασμός (σεξισμός και ρατσισμός) και μη ρητορική μίσους). Χρησιμοποίησαν τέσσερα μοντέλα CNN που

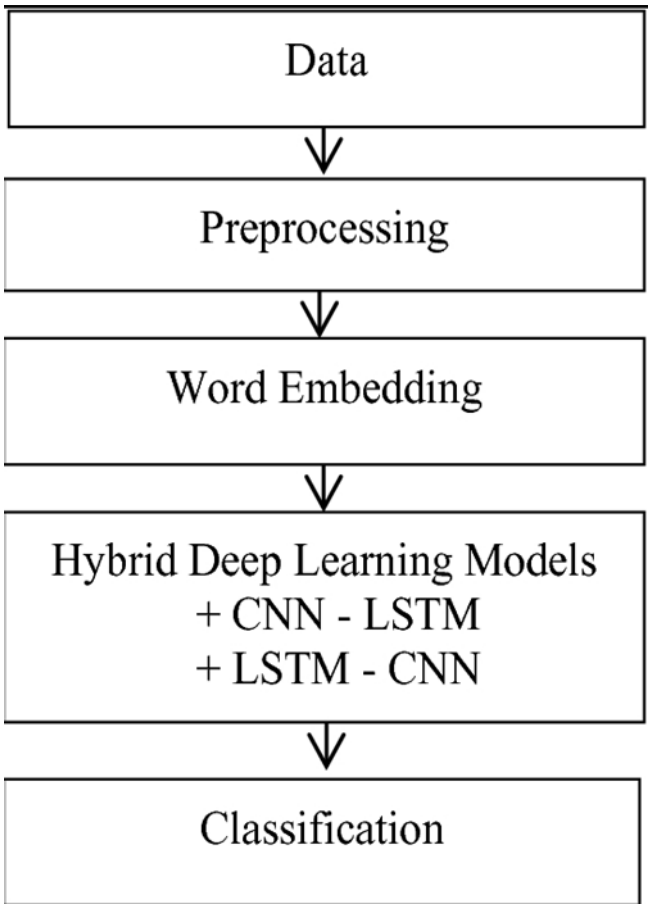
εκπαιδεύτηκαν με μεθόδους: χαρακτήρα n-gram, word2vec, τυχαία διανύσματα συνδυασμένα (word2vec και χαρακτήρα n-gram). Οι συγγραφείς χρησιμοποίησαν τεχνική k-fold (k=10) για να βελτιώσουν την ακρίβεια του μοντέλου. Μεταξύ των τεσσάρων μοντέλων, το μοντέλο CNN με βάση το word2vec είχε καλή απόδοση με 78,3% της βαθμολογίας F1.

Το 2018, οι Zhang et al., Εισήγαν μια νέα μέθοδο βασισμένη στη βαθιά μάθηση που συνδυάζει CNN και GRU νευρωνικά δίκτυα[40]. Πραγματοποιούν μια εκτενής αξιολόγηση της μεθόδου στις μεγαλύτερες δημόσιες συλλογές διαθέσιμων δεδομένων του Twitter. Συγκρίνοντας την έρευνα τους με παρόμοιες έρευνες πάνω στο ίδιο σύνολο δεδομένων, η προτεινόμενη μέθοδος τους είναι σε θέση να συλλάβει και την ακολουθία λέξεων και λέξεις-κλειδιά σε μικρά κείμενα. Έτσι, θέτει ένα νέο σημείο αναφοράς με καλύτερη επίδοση σε 6 από τα 7 σύνολα δεδομένων που δοκιμάστηκε με απόδοση καλύτερη κατά 1 έως 13 τοις εκατό στη βαθμολογία F1.

Το 2019, οι Devlin et al., Εισήγαν ένα νέο μοντέλο αναπαράστασης γλωσσών που ονομάζεται B.E.R.T, το οποίο σημαίνει Bidirectional Encoder Representations from Transformers. Σε αντίθεση με τα πρόσφατα μοντέλα αναπαράστασης γλωσσών, το BERT έχει σχεδιαστεί για να εκπαιδεύει εκ των προτέρων (pre-training) βαθιές αμφίδρομες αναπαραστάσεις σε μη σημειομένο κείμενο. Ως αποτέλεσμα, το προ-εκπαιδευμένο μοντέλο BERT μπορεί να ρυθμιστεί με ένα μόνο επιπλέον επίπεδο εξόδου για να δημιουργήσει μοντέλα τελευταίας τεχνολογίας για ένα ευρύ φάσμα εργασιών, όπως απάντηση ερωτήσεων και συμπεράσματα γλώσσας χωρίς ουσιαστική εργασία με μόνο μικρές συγκεκριμένες τροποποιήσεις στην αρχιτεκτονική. Το BERT είναι εννοιολογικά απλό και εμπειρικά ισχυρό. Γιαυτό το λόγο θέτει νέο σημείο αναφοράς στα αποτελέσματα του σε έντεκα εργασίες επεξεργασίας φυσικής γλώσσας (NLP)

2.6 Υβριδικά Συστήματα

Κάθε λύση έχει το δικό της σύνολο περιορισμών. Έτσι, είναι μια έξυπνη επιλογή να συνδυαστούν δύο ή περισσότερες μέθοδοι σε μια υβριδική προσέγγιση στην οποία η μία μέθοδος συμπληρώνει την άλλη. Για να δημιουργήσουμε ένα αποτελεσματικό μοντέλο, χρησιμοποιούμε υβριδικές τεχνικές που ενσωματώνουν προσεγγίσεις μηχανικής μάθησης, βασισμένες σε κανόνες και βαθιάς μάθησης.



Πηγή: link.springer.com/chapter/10.1007/978-3-030-34365-1_13

Εικόνα 24: Παράδειγμα Γενικής Δομής Υβριδικών Συστημάτων

2.6.1 Επιστημονικό Έργο

Το 2019, οι Viviana Patti et al., πρότειναν μια υβριδική προσέγγιση για τον εντοπισμό της ρητορικής μίσους [41]. Σε αυτή την έρευνα, χρησιμοποίησαν δύο μοντέλα. Στο πρώτο τους μοντέλο, εφάρμοσαν έναν ταξινομητή γραμμικής υποστήριξης διανύσματος (LSVC) και στο δεύτερο μοντέλο, χρησιμοποίησαν ένα νευρωνικό μοντέλο βραχυπρόθεσμης μνήμης (LSTM) με

ενσωμάτωση λέξεων (word embeddings). Η κοινή μάθηση με ένα πολύγλωσσικό μοντέλο ενσωμάτωσης λέξεων (multilingual word embeddings) σε 17 κατηγορίες του συνόλου δεδομένων HurtLex, πέτυχε αποτελέσματα της τάξης του 68,7% της βαθμολογίας F1.

Το 2020, οι Safa Alsafari et al., πρότειναν ένα μοντέλο ανίχνευσης ρητορικής μίσους για τα αραβικά μέσα κοινωνικής δικτύωσης[42]. Σε αυτή την έρευνα, συγκέντρωσαν το σύνολο δεδομένων χρησιμοποιώντας το API αναζήτησης Twitter και το σύνολο δεδομένων κατηγοριοποιήθηκε σε τέσσερις κατηγορίες (Θρησκευτική, Εθνικότητα, Φύλο και Εθνότητα). «Καθάρισαν» το σύνολο δεδομένων στην φάση της προεπεξεργασίας αφαιρώντας περιττές λέξεις όπως διευθύνσεις URL, σημεία στίξης, σύμβολα, ετικέτες. Εφάρμοσαν μια ταξινόμηση πολλαπλών κλάσσεων (3) με το μοντέλο CNN και το μοντέλο B.E.R.T για να επιτύχουν ποσοστά της τάξης του 75,51% της βαθμολογίας F1.

3

Σύνολα Δεδομένων Εκπαίδευσης Ρητορικής Μίσους

Σε αυτό το κεφάλαιο αναλύουμε τα σύνολα δεδομένων εκπαίδευσης ρητορικής μίσους και πιο συγκεκριμένα τις προκλήσεις που παρουσιάζονται και την προτεινόμενη μας μεθοδολογία για την ορθή και αντικειμενική επιλογή συνόλων δεδομένων.

3.1 Προκλήσεις

Η επιλογή του κατάλληλου συνόλου δεδομένων στην προσπάθεια αναγνώρισης ρητορικής μίσους, αποτελεί ένα απο τα πιο σημαντικά βήματα της διαδικασίας. Τα σύνολα δεδομένων είναι φτιαγμένα απο τους ανθρώπους με αποτέλεσμα πολλές φορές να παρατηρούνται φαινόμενα τα οποία επηρεάζουν άμμεσα την αποτελεσματικότητα του μοντέλου [garbage in garbage out]

Ορισμός Ρητορικής Μίσους

Η πολυπλοκότητα της διαδικασίας σαφή ορισμού της ρητορικής μίσους αποτελεί ένα απο τα βασικά ζητήματα του τομέα. Η παράλληλη χρήση διαφορετικών ορισμών δημιουργεί προβλήματα στην προσπάθεια σύγκρισης επιδόσεων μοντέλων και στην προσπάθεια να υπάρχει μια ακριβή εικόνα των αποτελεσμάτων.

Μεροληψία Δειγματοληψίας – Sampling Bias

Η μεροληψία δειγματοληψίας αποτελεί μια μεγάλη πρόκληση που μπορεί να επηρεάσει αρνητικά τα αποτελέσματα μιας μελέτης και να επηρεάσει την εγκυρότητα της διερευνητικής διαδικασίας. Παρατηρείτε αυτο το φαινόμενο όταν δεν υπάρχει ισσοροπημένη κατανομή δειγμάτων στο σύνολο δεδομένων εκπαίδευσης. Τέλος, η τυχαία δειγματοληψία καθιστά τα σύνολα δεδομένων επιρρεπή σε προκατάληψη.

Προκατάληψη Σχολιασμού – Annotation Bias

Οι προκαταλήψεις των ατόμων που σχολιάζουν τα σύνολα δεδομένων καταλήγουν σε συμπεράσματα με βάση τις δικές τους προκαταλήψεις, οδηγώντας σε λανθασμένες προβλέψεις, χαμηλή ακρίβεια και πολύ χαμηλές βαθμολογίες γενίκευσης. Μόνο περίπου τα δύο τρίτα των υφιστάμενων συνόλων δεδομένων αναφέρουν τη συμφωνία μεταξύ σχολιαστών

Φυλετική Προκατάληψη – Racial Bias

Το φαινόμενα της φυλετικής προκατάληψης δυστηχώς εντοπίζεται αρκετά σε σύνολα δεδομένων, έτσι και τα αποτελέσματα που θα παραχθούν μέσω της εκπαίδευσης σε αυτά τα δεδομένα θα περιέχουν το φαινόμενο της φυλετικής προκατάληψης στις μελλοντικές αποφάσεις τους.

3.2 Εύρεση Διαθέσιμων

Η ραγδαία αύξηση του τομέα της τεχνητής νοημοσύνης τα τελευταία χρόνια έχει φέρει στο επίκεντρο την δυνατότητα επίλυσης καθημερινών προβλημάτων με την βοήθεια της τεχνολογίας. Ειδικά στο τομέα της Ανάλυσης Φυσικής Γλώσσας (NLP) παρατηρούμε μεγάλη αύξηση στην δημιουργία επιστημονικών ερευνών. Αυτό το φαινόμενο δεν είναι τυχαίο, καθώς εντάσεται στον στόχο που έχουν θέσει κυβερνήσεις και οργανισμοί ανα τον κόσμο για την καταπολέμηση του ρατσισμού, εξτρεμισμού και της βίας στην σφαίρα του διαδικτύου.

Βέβαια, τα μοντέλα τεχνητής νοημοσύνης χρειάζονται μεγάλους όγκους δεδομένων ώστε να παράγουν καλά αποτελέσματα. Κρίσημο σημείο στην προσπάθεια αυτή αποτελεί η ποιότητα των δεδομένων, καθώς, εάν τα δεδομένα είναι χαμηλής ποιότητας τότε όποιο μοντέλο και να υλοποιήσουμε, το αποτέλεσμα θα είναι μή αντικειμενικό και παραπλανητικό.

Στην προσπάθεια μας εύρεσης συνόλου δεδομένων το οποίο ελαχιστοποιεί όσο το δυνατόν περισσότερο τα προβλήματα που θέσαμε στην παραπάνω ενότητα, χρησιμοποιήσαμε ώς πηγή δεδομένων την πρόσφατη και εκτενή ανάλυση της έρευνας [16] η οποία αναλύει σε βάθος και ανα κατηγορία όλα τα πρόσφατα (2018-2020) σύνολα δεδομένων (46) σχετικά με την ανάλυση και ταξινόμηση ρητορικής μίσους.

Η συστηματική τους ανάλυση χωρίζει και συγκρίνει τα σύνολα δεδομένων σε πέντε διαστάσεις:

1. **Τύπος – Type**: Σε αυτή τη διάσταση γίνεται διαχωρισμός μεταξύ Corpora (συλλογή κειμένων) και Lexica (λίστες λέξεων ή φράσεων που σχετίζονται με μια κοινή σημασιολογια).

2. **Πεδίο Εστίασης – Topical Focus**: Σε αυτή τη διάσταση, γίνεται διαχωρισμός σύμφωνα με την διαφορετική εστίαση που έχει κάθε σύνολο δεδομένων(Επιθετική Ρητορική, Ρατσιστική Ρητορική, Ρητορική Μίσους)

3. **Πηγή Δεδομένων – Data Source**: Σε αυτή την διάσταση, γίνεται διαχωρισμός με βάση την πηγή απο την οποία αντλήθηκαν τα παραδείγματα των συνόλων δεδομένων.

4. **Σχολιασμός – Annotation**: Σε αυτή την διάσταση, γίνεται διαχωρισμός με βάση το «πώς» και απο «ποιόν» έχουν σχολιαστεί τα δεδομένα, σύμφωνα με ποιά δομή και πώς έγινε έλεγχος ποιότητας του σχολιασμού αυτού.

5. **Γλώσσα – Language**: Σε αυτή την διάσταση, γίνεται διαχωρισμός με βάση την γλώσσα/ες στην οποία διατίθετε το κάθε σύνολο δεδομένων.

Στους παρακάτω πίνακες βλέπουμε μία συνολική εικόνα όλων των συνόλων δεδομένων που παρουσιάζει η έρευνα και κάποια βασικά χαρακτηριστικά τους:

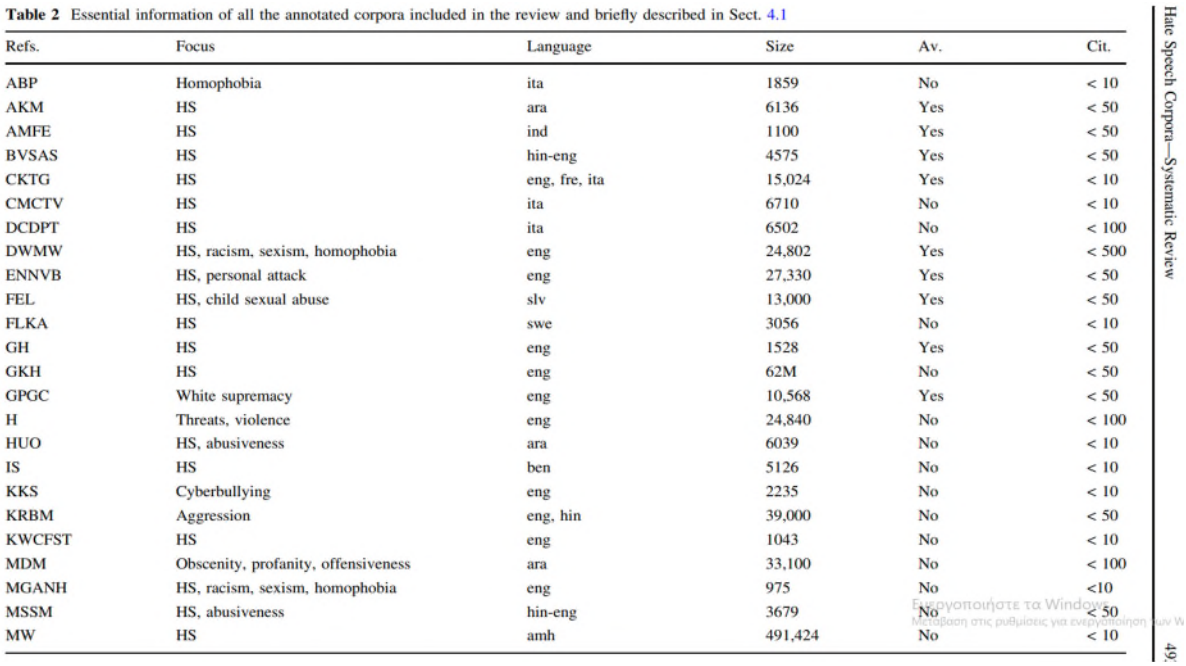


Table 2 Essential information of all the annotated corpora included in the review and briefly described in Sect. 4.1

Refs.	Focus	Language	Size	Av.	Cit.
ABP	Homophobia	ita	1859	No	< 10
AKM	HS	ara	6136	Yes	< 50
AMFE	HS	ind	1100	Yes	< 50
BVSAS	HS	hin-eng	4575	Yes	< 50
CKTG	HS	eng, fre, ita	15,024	Yes	< 10
CMCTV	HS	ita	6710	No	< 10
DCDPT	HS	ita	6502	No	< 100
DWMW	HS, racism, sexism, homophobia	eng	24,802	Yes	< 500
ENNVB	HS, personal attack	eng	27,330	Yes	< 50
FEL	HS, child sexual abuse	slv	13,000	Yes	< 50
FLKA	HS	swe	3056	No	< 10
GH	HS	eng	1528	Yes	< 50
GKH	HS	eng	62M	No	< 50
GPGC	White supremacy	eng	10,568	Yes	< 50
H	Threats, violence	eng	24,840	No	< 100
HUO	HS, abusiveness	ara	6039	No	< 10
IS	HS	ben	5126	No	< 10
KKS	Cyberbullying	eng	2235	No	< 10
KRBM	Aggression	eng, hin	39,000	No	< 50
KWCFST	HS	eng	1043	No	< 10
MDM	Obscenity, profanity, offensiveness	ara	33,100	No	< 100
MGANH	HS, racism, sexism, homophobia	eng	975	No	<10
MSSM	HS, abusiveness	hin-eng	3679	No	< 50
MW	HS	amh	491,424	No	< 10

Πηγή: link.springer.com/content/pdf/10.1007/s10579-020-09502-8.pdf

Εικόνα 25: Συνολική Εικόνα Εύρενας Συνόλων Δεδομένων/1

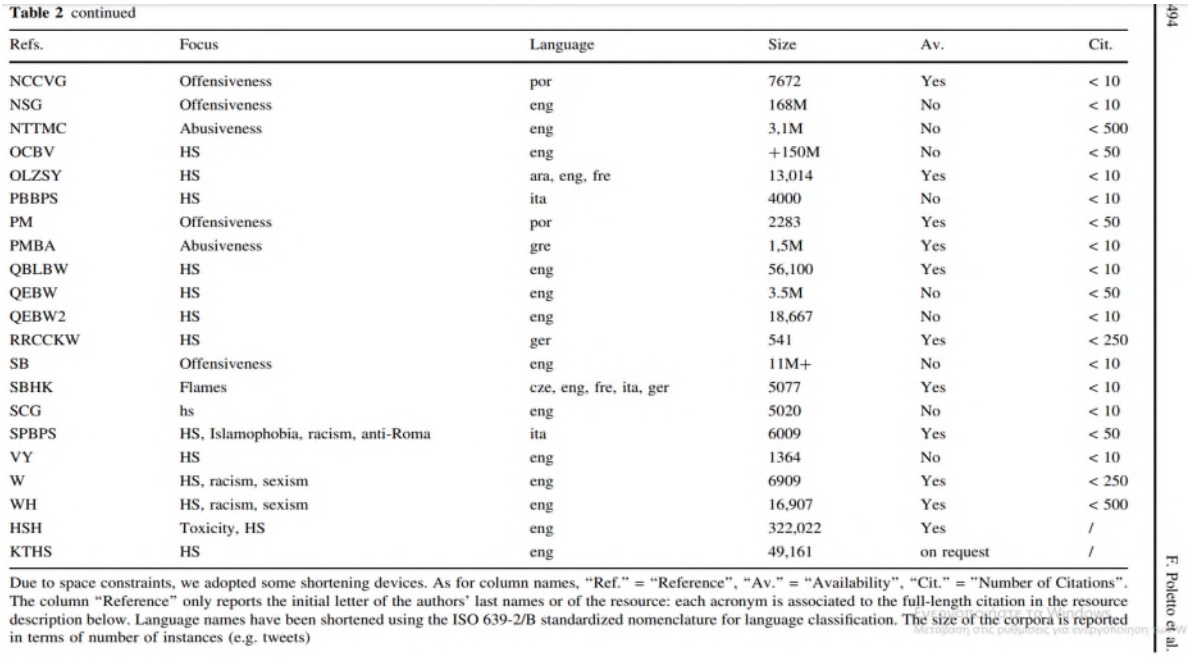


Table 2 continued

Refs.	Focus	Language	Size	Av.	Cit.
NCCVG	Offensiveness	por	7672	Yes	< 10
NSG	Offensiveness	eng	168M	No	< 10
NTTMC	Abusiveness	eng	3,1M	No	< 500
OCBV	HS	eng	+150M	No	< 50
OLZSY	HS	ara, eng, fre	13,014	Yes	< 10
PBBPS	HS	ita	4000	No	< 10
PM	Offensiveness	por	2283	Yes	< 50
PMBA	Abusiveness	gre	1,5M	Yes	< 10
QBLBW	HS	eng	56,100	Yes	< 10
QEBW	HS	eng	3.5M	No	< 50
QEBW2	HS	eng	18,667	No	< 10
RRCCKW	HS	ger	541	Yes	< 250
SB	Offensiveness	eng	11M+	No	< 10
SBHK	Flames	cze, eng, fre, ita, ger	5077	Yes	< 10
SCG	hs	eng	5020	No	< 10
SPBPS	HS, Islamophobia, racism, anti-Roma	ita	6009	Yes	< 50
VY	HS	eng	1364	No	< 10
W	HS, racism, sexism	eng	6909	Yes	< 250
WH	HS, racism, sexism	eng	16,907	Yes	< 500
HSH	Toxicity, HS	eng	322,022	Yes	/
KTHS	HS	eng	49,161	on request	/

Due to space constraints, we adopted some shortening devices. As for column names, "Ref." = "Reference", "Av." = "Availability", "Cit." = "Number of Citations". The column "Reference" only reports the initial letter of the authors' last names or of the resource: each acronym is associated to the full-length citation in the resource description below. Language names have been shortened using the ISO 639-2/B standardized nomenclature for language classification. The size of the corpora is reported in terms of number of instances (e.g. tweets)

Πηγή: link.springer.com/content/pdf/10.1007/s10579-020-09502-8.pdf

Εικόνα 26: Συνολική Εικόνα Εύρενας Συνόλων Δεδομένων/1

3.3 Μεθοδολογία Επιλογής

Σκοπός της έρευνας μας είναι να βρούμε εκείνο το σύνολο δεδομένων το οποίο μετριάζει είτε εξαφανίζει προβλήματα όπως: φυλετική προκατάληψη, προκατάληψη σχολιασμού, μεροληψία δειγματοληψίας και θέτει σαφή όρια και κατευθύνσεις σχετικά με τον ορισμό της ρητορικής μίσους. Στην συνέχεια, θέλουμε να υπάρχει σαφής διαχωρισμός ανάμεσα στην ρητορική μίσους και την επιθετική ρητορική, καθώς ο συνψηφισμός των δύο εννοιών οδηγεί σε μη αντικειμενικές προβλέψεις. Έπειτα, στοχεύουμε σε όσες περισσότερες κλάσεις γίνεται, αφού η πληθώρα κλάσεων μπορεί να μας δείξει μια γενικότερη και σαφέστερη εικόνα του προβλήματος. Επιπρόσθετα, στην παρούσα έρευνα ενδιαφερόμαστε για σύνολα δεδομένων της Αγγλικής γλώσσας. Τέλος, ένα στοιχείο μείζονος σημασίας αποτελούν οι οδηγίες και η λεπτομερής ανάλυση της διαδικασίας σχολιασμού. Απο την διαδικασία σχολιασμού μπορούμε συχνά να καταλάβουμε την ποιότητα του συνόλου δεδομένων και αν είναι κατάλληλο για την έρευνα μας.

3.3.1 Πρώτη Φάση - Συλλογή

Στην πρώτη φάση της μεθοδολογίας μας θέτουμε τα χαρακτηριστικά του συνόλου δεδομένων μας στις πέντε διαστάσεις ώστε να βρούμε αυτά που ταιριάζουν στις απαιτήσεις μας. Έτσι από την ερευνα [43] και τα αναλυτικά της γραφήματα επιλέγουμε εκείνα τα συνολα δεδομένων που πληρούν τις παρακάτω προυποθεσεις,

1. Τύπος: Corpora
2. Πεδίο Εστίασης: Ρητορική Μίσους
3. Γλώσσα: Αγγλικά/Μίξη.
4. Ύπαρξη Μεθόδου Σχολιασμού: Ναι
5. Τύπος Μεθόδου Σχολιασμού: Μή-δυαδική, Πολυ-επίπεδη(πολλαπλών ετικετών).

Τα σύνολα δεδομένων που προκύπτουν σύμφωνα με τα παραπάνω χαρακτηριστικά είναι τα εξής:

ENNVB - ElSherief et al. (2018) [44]

Σύνδεσμος Συνόλου δεδομένων: https://github.com/mayelsherif/hate_speech_icwsm18

Περίληψη: (ENNVB)—27,330 tweets στην αγγλική γλώσσα, σχολιασμένα απο σύμπραξη ειδικών ώς εχθρικά [με προσωπική επίθεση/χωρίς] / μη εχθρικά. Το ποσοστό συμφωνίας μεταξύ των ειδικών είναι 92% για την «εχθρική» κλάση και 82% για την «μη εχθρική κλάση».

DWMW - Davidson et al. (2017) [11]

Σύνδεσμος Συνόλου δεδομένων: https://github.com/t-davidson/hate-speech-and-offensive-language

Περίληψη: (DWMW)—24,802 tweets στην αγγλική γλώσσα, σχολιασμένα απο σύμπραξη ειδικών ώς: ρητορική μίσους, επιθετικά αλλα όχι ρητορική μίσους, όχι ρητορική μίσους. Βέβαια, Μόνο το 5% εκ αυτών κρίθηκαν ως ρητορική μίσους απο την πλειοψηφία.

OLSZY - Ousidhoum et al. (2019) [17]

Σύνδεσμος Συνόλου Δεδομένων: github.com/HKUST-KnowComp/MLMA_hate_speech

Περίληψη: (OLZSY)—13,014 tweets στα Αραβικά, Αγγλικά και Γαλλικά σχολιασμένα πολυεπίπεδα απο σύμπραξη ειδικών ώς:

- Directness (direct, indirect),
- Hostility (abusive, hateful, offensive, disrespectful, fearful, normal)
- Target (origin, gender, sexual orientation, religion, disability, other)
- Group (individual woman. special needs, African descent. other)
- The feeling aroused in the annotator by the tweet (disgust. Shock. Anger. Sadness. Fear. Confusion. indifference).

WH - Waseem and Hovy (2016) [44]

Σύνδεσμος Συνόλου δεδομένων: https://github.com/zeerakw/hatespeech

Περίληψη: (WH)—16,907 tweets στην αγγλική γλώσσα, σχολιασμένα απο σύμπραξη ειδικών ώς: ‘‘sexist/racist/both/neither’’. Το ποσοστό συμφωνίας μεταξύ των κριτών είναι: j = 0.85.

3.3.2 Δεύτερη Φάση – Σύγκριση

Σε αυτή τη δεύτερη φάση επικεντρονώμαστε στην ποιότητα του σχολιασμού για κάθε σύνολο δεδομένων που προέκυψε απο την πρώτη φάση ώστε να επιλέξουμε το κατάλληλο για την έρευνα μας.

WH—Paper: Waseem and Hovy (2016)

Οι Waseem και Hovy (2016) συγκέντρωσαν 130 χιλιάδες tweets που περιείχαν τουλάχιστον ένα απο τα δεκαεπτά άσχημα επίθετα ή φράσεις που έθεσαν οι ίδιοι. Στη συνέχεια χρησιμοποίησαν κριτικά εμπνευσμένα από τη θεωρία της φυλής (critical race) για να σχολιάσουν ένα δείγμα αυτών των tweets. Για να ελέγξουν την προκατάληψη, οι σχολιαστές εξετάστηκαν από "μια 25χρονη γυναίκα που μελετούσε σπουδές φύλου και δηλώνει μη-ακτιβιστική φεμινίστρια". Υπάρχουν 16.849 tweets σε αυτό το σύνολο δεδομένων που έχουν ταξινομηθεί ως

ρατσισμός, σεξισμός ή κανένα από τα δύο. Η πλειοψηφία των σεξιστικών tweets αφορούν συζητήσεις σε αυστραλιανή τηλεοπτική εκπομπή και η πλειοψηφία των ρατσιστικών tweets είναι αντι-μουσουλμανικά.

Σύμφωνα με την έρευνα (Davidson et al. 2019), οι οποία εκπαίδευσε ταξινομητές στο σύνολο δεδομένων χρησιμοποιώντας tweets με δημογραφίκές πληροφορίες ώστε να συγκρίνει την απόδοση κάθε ταξινομητή σε tweets γραμμένα στα Αφροαμερικανικά Αγγλικά (AAE) έναντι των Αμερικάνικων Αγγλικών (SAE), ανακάλυψε στοιχεία συστηματικής φυλετικής προκατάληψης . Τα tweets γραμμένα στα Αφροαμερικανικά Αγγλικά (AAE) ταξινομούνται πιο εύκολα και σε μεγαλύτερο βαθμό στις κλάσσεις της ρητορικής μίσους ή παρενόχλησης σε σχέση με τα tweet που είναι γραμμένα στα Αμερικάνικα Αγγλικά (SAE).

Επίσης, ο ταξινομητής σε αυτό το σύνολο δεδομένων είναι πιο πιθανό κατα 1.5 φορές να ταξινομήσει tweet απο αφροαμερικάνους ώς σεξιστικά σε σχέση με τα tweet λευκών. Αυτο συμβαίνει κυρίως λόγω ότι οι λέξεις “n*gga” και “b*tch” οι οποίες χρησιμοποιούνται περισσότερο απο αφροαμερικανούς στην καθημερινή τους ομιλία δεν έχουν ληφθεί υπόψιν ώς ουδέτερες σε μερικές περιπτώσεις με αποτέλεσμα ο ταξινομητής να παράγει αποτελέσματα τα οποία περιέχουν το στοιχείο της φυλετικής διάκρισης.

Λαμβάνοντας υπόψη τα παραπάνω, απορρίπτουμε το σύνολο δεδομένων Waseem και Hovy (2016) από τη λίστα μας καθώς δεν καλύπτει τις προυποθέσεις εγκυρότητας και τους στόχους της έρευνας μας.

DWMW—Paper: Davidson et al. (2017)

Προχωρώντας στο σύνολο δεδομένων των Davidson et al. (2017), Σύμφωνα με την έρευνα *(Davidson et al. 2019),* διαπιστώνουμε μεγάλες ανισότητες, με περίπου 5% των tweets τα οποία προέρχονται απο αφροαμερικανούς να ταξινομούνται στην κατηγορία ρητορικής μίσους σε σύγκριση με το 2% εκείνων στο σετ της κατηγορίας των λευκών. Ομοίως, το 17% των tweets που προέρχονται απο αφροαμερικανούς προβλέπεται ότι περιέχουν προσβλητική γλώσσα σε σύγκριση με το 6,5% των tweets που προέρχονται απο λευκούς, Ο ταξινομητής που εκπαιδεύτηκε στα δεδομένα των Davidson et al. (2017) είναι λιγότερο πιθανό να ταξινομήσει τα tweet αφροαμερικανών ως ρητορική μίσους, αν και είναι πιο πιθανό να τα ταξινομήσει ως προσβλητικά.

Τα αποτελέσματα που επικεντρώνονται σε αυτό το σύνολο δεδομένων καταδεικνύουν συνεπείς, συστηματικές και ουσιαστικές φυλετικές προκαταλήψεις στον ταξινομητή που έχει εκπαιδευτεί σε αυτό. Σχεδόν σε κάθε περίπτωση, τα tweets με μαύρη ευθυγράμμιση ταξινομούνται ως σεξισμός, ρητορική μίσους, παρενόχληση και κακοποίηση με υψηλότερα ποσοστά από τα tweets με λευκή ευθυγράμμιση. Τέλος, σύμφωνα και με την έρευνα *(Sap et al. 2019)*, ταξινομητής που εκπαιδεύτηκαν στο παρών σύνολο δεδομένων προβλέπει ότι σχεδόν το 50% των μη επιθετικών tweets Αφροαμερικανών έχουν κατηγοριοποιηθεί λανθασμένα ώς προσβλητικά.

Υπό το φως αυτών των δεδομένων, το *DWMW - Davidson et al. (2017)* απορρίπτεται από τη λίστα μας καθώς δεν καλύπτει τις προυποθέσεις εγκυρότητας και τους στόχους της έρευνας μας.

ENNVB—Paper: ElSherief et al. (2018)

Συνεχίζοντας την ανάλυση με το σύνολο δεδομένων της έρευνας *ElSherief et al. (2018),* η έρευνα αυτή παρουσιάζει ένα αγγλικό σύνολο δεδομένων που καταγράφει την κατηγορία-στόχο με βάση την κατεύθυνση της ρητορικής μίσους σε άτομα, όπως η εθνικότητα, το φύλο ή ο σεξουαλικός προσανατολισμός και ζητά από τους σχολιαστές να ταξινομήσουν τα tweets ως μίσος και μη μίσος.

Στη έρευνα μας, θέλουμε να διαχωρίσουμε τη ρητορική μίσους από την προσβλητική γλώσσα και αυτό το σύνολο δεδομένων δεν μας δίνει αυτήν την ευκαιρία. Έτσι, το ENNVB - *ElSherief et al. (2018)* απορρίπτεται από τη λίστα μας καθώς δεν καλύπτει τις προυποθέσεις εγκυρότητας και τους στόχους της έρευνας μας.

OLSZY— Ousidhoum et al. (2019)

Το σύνολο δεδομένων τους περιέχει 13,000 tweets, κατηγοριοποιημένα σε 5 αυτόνομες πτυχές:

1. Με βάση την αμεσότητα χωρίζεται σε: direct ή indirect;
2. Με βάση την εχθρότητα χωρίζεται σε: offensive, disrespectful, hateful, fearful out of ignorance, abusive, ή normal
3. Με γνώμονα το χαρακτηριστικό βάσει του οποίου εισάγει διακρίσεις σε βάρος ενός ατόμου ή μιας ομάδας ανθρώπων σε: ‘origin, gender, sexual orientation, religion, disability, other.
4. Με βάση τον όνομα της ομάδας: (individual, woman, special needs, African descent, other.
5. Με βάση τα αισθήματα που προκαλεί το περιεχόμενο του κειμένου στους σχολιαστές, σε: disgust, shock, anger, sadness, fear, confusion, indifference.

Αποτελεί το πρώτο τρίγλωσσο σύνολο δεδομένων που περιέχει αγγλικά, γαλλικά και αραβικά tweets και περιλαμβάνει διάφορους στόχους και τύπους εχθρότητας και ρητορικής μίσους. Επιπλέον, είναι το σύνολο δεδομένων που εξετάζει πώς αντιδρούν οι σχολιαστές στα σχόλια ρητορικής μίσους.

Το OLSZY μας παρέχει ένα σύνολο δεδομένων πολλαπλών ετικετών υψηλής ποιότητας, ικανό να διαχωρίσει την ρητορική μίσους (HS) από την προσβλητική γλώσσα (Offensive). Επίσης, λαμβάνει υπόψη προβλήματα των προηγούμενων συνόλων δεδομένων, όπως: «προκατάληψη σχολιασμού», «φυλετικής προκατάληψη». Μας δίνει την επιλογή να χρησιμοποιήσουμε μία-δύο ή περισσότερες ετικέτες-στόχους στις προσπάθειές ταξινόμησης.

3.3.3 Τρίτη Φάση - Επιλογή

Μετά την παραπάνω σύγκριση των συνόλων δεδομένων και την αναλυτική έρευνα στον τρόπο σχολιασμού τους, καταλήγουμε πως το κατάλληλο για την έρευνα μας είναι το *OLSZY-Ousidhoum et al. (2019)* σύνολο δεδομένων καθώς προσφέρει δυνατότητα πολυ-επίπεδης ταξινόμησης, περέχει συγκεκριμένες και ακριβής οδηγίες στους σχολιαστές, αντιμετωπίζει όσον το δυνατό περισσότερο τις δυσκολίες που αναφέραμε στην αρχή του κεφαλαίου (φυλετική προκατάληψη, προκατάληψη σχολιασμού) και δεν υπάρχουν έρευνες που να θέτουν σε αμφισβήτηση την ποιότητα του στην ακαδημαική βιβλιογραφία.

4

Η Μεθοδολογία Μας

Σε αυτό το κεφάλαιο αναλύουμε την μεθοδολογία που ακολουθήσαμε κατα την διεξαγωγή των πειραμάτων μας, συμπεριλαμβανομένων των τεχνικών επεξεργασίας δεδομένων και της αρχιτεκτονικής BERT.

4.1 Εισαγωγή

Στόχος της εργασίας μας είναι η υλοποίηση μοντέλου βαθιάς μάθησης με σκοπό την ταξινόμηση κειμένων (tweets) ρητορικής μίσους που έχουν αντληθεί απο αντικειμενικές πηγές απαλαγμένα όσο το δυνατόν περισσότερο απο φαινόμενα φυλετικής προκατάληψης, προκατάληψης σχολιασμού και έλλειψη επαρκή ορισμού. Κατηγοριοποιημένα έτσι, ώστε να διαχωρίζουν την ρητορική μίσους απο την επιθετικότητα και να προσφέρουν αρκετές κατηγορίες ταξινόμησης έτσι ώστε τα αποτελέσματα τους να προσφέρουν μια ευρύτερη εικόνα της ουσίας των κειμένων.

Ο τύπος ταξινόμησης που επιλέξαμε είναι ταξινόμηση πολλαπλών ετικετών με έξι (6) κλάσεις στόχους (abusive / hateful / offensive / disrespectful / fearful / normal) και για αρχιτεκτονική την BERT (Bidirectional Encoder Representations from Transformers).

Αρχικά, για τις ανάγκες της κατασκευής του μοντέλου μας είναι σημαντικό να προηγηθεί μια στατιστική ανάλυση των δεδομένων ώστε να γνωρίζουμε την μορφή, τις ιδιαιτερότητες ή και τις δυσκολίες που μπορεί να παρουσιαστούν. Στην συνέχεια, επεξεργαζόμαστε και επαυξάνουμε τα δεδομένα ώστε να είναι όσο πιο αποδοτικά γίνεται. Έπειτα, αφού τα χωρίσουμε σε δεδομένα εκπαίδευσης, ελέγχου και επικύρωσης τα τροφοδοτούμε στο μοντέλο. Τέλος, ελέγχουμε την απόδοση του μοντέλου μας με διάφορες σημαντικές μετρήσεις όπως: Macro/Micro F1 score, AUC, Accuracy, Recall.

4.2 Βιβλιοθήκες/Frameworks

Η συγγραφή του κώδικα πραγματοποιήθηκε με την γλώσσα Python, για την υλοποίηση των πειραμάτων μας χρησιμοποιήσαμε μια πληθώρα βιβλιοθηκών για την ανάλυση των δεδομένων αλλά και για την υλοποίηση της επεξεργασίας των δεδομένων και της υλοποίησης του BERT μοντέλου.

Οι βασικές βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι εξής:

- **Pandas**

  Το pandas είναι ένα γρήγορο, ισχυρό, ευέλικτο και εύχρηστο εργαλείο ανοιχτού κώδικα που παρέχει εύχρηστες δομές δεδομένων υψηλής απόδοσης και εργαλεία που βοηθάνε στην ανάλυσης και τον χειρισμό δεδομένων. Βασισμένο στη γλώσσα προγραμματισμού Python.

- **Numpy**

Το NumPy είναι το ισχυρότερο πακέτο για επιστημονικούς υπολογισμούς στην Python. Είναι μια βιβλιοθήκη που παρέχει την χρήση πολυδιάστατων πινάκων και μια ποικιλία «ρουτίνων» για γρήγορες λειτουργίες σε πίνακες, συμπεριλαμβανομένων μαθηματικών πράξεων, λογικών πράξεων, αλλαγή μορφής πινάκων, ταξινόμησης, επιλογής, εισόδου/εξόδου , διακριτοί μετασχηματισμοί Fourier, βασική γραμμική άλγεβρα, βασικές στατιστικές πράξεις, τυχαία προσομοίωση και πολλά άλλα.

- **Spacy**

Το spaCy είναι μια βιβλιοθήκη για προηγμένη επεξεργασία φυσικής γλώσσας σε Python και Cython. Είναι βασισμένο και ενημερωμένο πάνω στην πιο πρόσφατες και σύγχρονες έρευνες.

- **Matplotlib**

Το Matplotlib είναι μια πολλαπλή πλατφόρμα, οπτικοποίησης δεδομένων και βιβλιοθήκη γραφικής σχεδίασης για την Python.

- **Scikit-learn**

Το Scikit-learn είναι μια βιβλιοθήκη μηχανικής μάθησης ανοιχτού κώδικα που υποστηρίζει μάθηση με επίβλεψη και χωρίς επίβλεψη. Παρέχει επίσης διάφορα εργαλεία για δημιουργία μοντέλων, προεπεξεργασία δεδομένων, επιλογή και αξιολόγηση μοντέλου και πολλά άλλα βοηθητικά προγράμματα.

- **Tensorflow**

Το TensorFlow είναι μια πλατφόρμα ανοικτού κώδικα από άκρο σε άκρο για μηχανική εκμάθηση. Διαθέτει ένα ολοκληρωμένο, ευέλικτο οικοσύστημα εργαλείων, βιβλιοθηκών και κοινοτικών πόρων που επιτρέπει στους ερευνητές να μάθουν και να προωθήσουν περαιτέρω την τεχνολογία στη μηχανική μάθηση και στους προγραμματιστές να αναπτήσουν πιο εύκολα και γρήγορα μοντέλα.

- **Nlpaug**

Το nlpgaug είναι μια βιβλιοθήκη για επ-αύξηση δεδομένων κειμένου σε πειράματα μηχανικής και βαθιάς μάθησης. Ο στόχος είναι η βελτίωση της απόδοσης του μοντέλου δημιουργώντας επιπλέον δεδομένα κειμένου με βάση τα ήδη υπάρχοντα. Περιέχει ένα εύρος χρήσιμων αλγορίθμων για την επεξεργασία δεδομένων κειμένου.

4.3 Επεξεργασία Δεδομένων

Κάθε σύνολο δεδομένων για να παράγει έγκυρα και υψηλής ποιότητας αποτελέσματα, πρέπει να μετατραπεί σε κατάλληλη μορφή ώστε να βοηθάει τον αλγόριθμο να κατανοήσει την ουσία του. Έτσι, σε αυτή την ενότητα θα παρουσιάσουμε τις τεχνικές που χρησιμοποιήσαμε ώστε να επεξεργαστούμε το σύνολο δεδομένων.

4.3.1 Προεπεξεργασία Δεδομένων - Preprocessing

Στην φάση της προεπεξεργασία των δεδομένων υλοποιήσαμε μία σειρά απο τεχνικές ώστε να μειώσουμε την πολυπλοκότητα των tweets.

Οι τεχνικές που χρησιμοποιήσαμε είναι οι εξής:

Decoding

Συχνά συμβαίνει να έχουμε δεδομένα κειμένου σε μορφή Unicode, αλλά να πρέπει να τα μετατρέψουμε σ μορφή ASCII. Η συνάρτηση unidecode() της python, λαμβάνει δεδομένα Unicode μορφής και προσπαθεί να τα αναπαραστήσει με χαρακτήρες ASCII (δηλαδή, τους χαρακτήρες που εμφανίζονται καθολικά μεταξύ 0x00 και 0x7F).

Apostrophe Normalization

Υπάρχουν δύο χαρακτήρες που χρησιμοποιούν οι άνθρωποι για συστολή (contraction) . "" " (Απόστροφο) και " "" (απόσπασμα). Εάν αυτά τα δύο σύμβολα χρησιμοποιούνται και τα δύο για συστολή, θα είναι δύσκολο να εντοπιστεί και να χαρτογραφηθεί σωστά η σωστή διευρυμένη μορφή. Έτσι, κάθε απόστροφος μετατρέπεται σε ενιαίο απόσπασμα.

Expand Contractions

Οι συσπάσεις είναι λέξεις ή συνδυασμοί λέξεων που συντομεύονται διαγράφοντας γράμματα και αντικαθιστώντας τα με μια απόστροφο. Στην Αγγλική γλώσσα, συχνά διαγράφουμε τα φωνήεντα από μια λέξη για να σχηματίσουμε τις συσπάσεις. Η αφαίρεση των συσπάσεων συμβάλλει στην τυποποίηση του κειμένου και είναι χρήσιμη όταν εργαζόμαστε σε δεδομένα που έχουν αντληθεί απο το Twitter.

```
str = 'Y\'all\'d be surprised if you know what I\'ll do.';
out = expandContractions( str );
// returns 'You all would be surprised if you know what I will do.'
```

Εικόνα 27: Παράδειγμα Αφαίρεσης Συσπάσεων.

Filtering

Πολλές από τις λέξεις που χρησιμοποιούνται σε μία πρόταση, είναι ασήμαντες και δεν έχουν κανένα ουσιαστικό νόημα. Για παράδειγμα η φράση: "τα Αγγλικά είναι ένα μάθημα". Εδώ, τα «αγγλικά» και «μάθημα» είναι οι πιο σημαντικές λέξεις. Αντίθετα, οι λέξεις: «είναι» και «ένα», είναι σχεδόν άχρηστες. Οι φράσεις: «Αγγλικά μάθημα» και «μάθημα Αγγλικά» έχουν την ίδια σημασία ακόμη και αν αφαιρέσουμε τις ασήμαντες λέξεις - («είναι», «ένα»). Έτσι, φιλτράρισμα είναι η διαδικασία κατα την οποία αφαιρούμε όλα τα στοιχεία του κειμένου που δεν έχουν βαρύνουσα σημασία και κρατάμε τα πιό ουσιαστικά. Στα πειράματα μας αφαιρέσαμε: τα σημεία στίξης, τον κενό χαρακτήρα (backspace), τους αριθμούς και τα URL χρησιμοποιώντας μεθόδους της βιβλιοθήκης Spacy. Τέλος, αφήσαμε άθικτο το περιεχόμενο κειμένου των hashtags.

Special Characters Removal

Καταργήσαμε τους ειδικούς χαρακτήρες και τα mentions των tweet posts, χαρακτήρες οι οποίοι εμφανίζονται συχνά σε tweets και δυσκολεύουν τους αλγορίθμους στην προσπάθεια εξαγωγής ουσίας απο τα κείμενα.

Spell Correction

Αποτελεί μια τεχνική ορθογραφικής διόρθωσης που ασχολείται κυρίως με επαναλαμβανόμενους χαρακτήρες όπως "sooooo goooood". Εάν ο ίδιος χαρακτήρας επαναληφθεί περισσότερες από δύο φορές, ο αλγόριθμος ορθογραφικής διόρθωσης μειώνει την επανάληψη του χαρακτήρα σε δύο φορές. Για παράδειγμα, το "sooooo goooood" θα μετατραπεί σε "so good". Αυτό θα βοηθήσει στη μείωση του χώρου των χαρακτηριστικών μετατρέποντας τις εμφανίσεις των: "sooo", "soooo", "sooooo" στην ίδια λέξη "so".

Input: "sooooo goooood"

Output: "so good"

Εικόνα 28: Παράδειγμα Ορθογραφικής Διόρθωσης

4.3.2 Μετασχηματισμός δεδομένων

Αναλόγως με τον τύπο ταξινόμησης, τα δεδομένα μετασχηματίζονται στην κατάλληλη μορφή ώστε να μπορούν να ταξινομηθούν αναλόγως. Στα πειράματα μας, μετασχηματίζουμε τα δεδομένα μας έτσι ώστε να υλοποιήσουμε ταξινόμηση πολλαπλών ετικετών σε έξι (6) κλάσεις στόχους ρητορικής μίσους. Η κάθε κλάσση αναπαριστάται με την μέθοδο one-hot-encoding ώς μηδέν (0) ή ένα (1).

	tweet	abusive	hateful	offensive	disrespectful	fearful	normal	tweet_cleaned	tweet_length
0	call sis im boy girl still faggot shut	0	0	1	0	0	0	call sis boy girl still faggot shut	7
1	@user @user legit nilas retarded idk	0	0	1	1	0	0	legit nilas retarded idk	4
2	said retard @url	1	1	1	0	1	0	said retard	2
3	america another 8 years obama's ideology via h...	1	1	0	1	1	1	america another years obama ideology via hilla...	13
4	@user don... fucking made cry twat.	0	0	1	0	0	0	don fucking made cry twat	5

Εικόνα 29: Παράδειγμα Μετασχηματισμένων Δεδομένων

4.3.3 Επαύξηση Δεδομένων – Data Augmentation

Επαύξηση δεδομένων είναι η διαδικασία κατα την οποία λόγω ανισόρροπης κατανομής παραδειγμάτων μεταξύ των κλάσσεων (π.χ. τρομακτικό, υβριστικό, ασεβές), δημιουργούνται επιπλέον δεδομένα στις μειωνοτικές κλάσσεις ώστε το τελικό μοντέλο να είναι αντικειμενικό. Στα πειράματα μας, αυξάνουμε τον αριθμό των παραδειγμάτων για καθεμία από αυτές τις κλάσσεις στόχους με τεχνικές επαύξησης κειμένου αλλά μόνο στα δεδομένα τα οποία θα εισαχθούν στο σετ εκπαίδευσης, τα δεδομένα που θα εισαχθούν στα σετ επαλήθευσης και ελέγχου μένουν ώς έχει. Οι τεχνικές οι οποίες χρησιμοποιούμε είναι: η ανάστροφη μετάφραση (*back-translation*) και η τεχνική αντικατάστασης μέσω *BERT embeddings*.

Η διαδικασία ανάστροφης μετάφρασης λειτουργεί με τον ακόλουθο τρόπο:

1. Παίρνουμε μια πρόταση στα Αγγλικά και την μεταφράσουμε στα Γερμανικά.

2. Μεταφράσουμε την πρόταση απο τα Γερμανικά πίσω στα Αγγλικά.

3. Ελέγχουμε εάν η νέα πρόταση είναι διαφορετική από την αρχική πρόταση. Εάν είναι, τότε προσθέτουμε και αυτήν τη νέα πρόταση στο σύνολο δεδομένων μας ως μια αυξημένη έκδοση του αρχικού κειμένου.

```
Original:
Its a pretty horrible time to be a sportsperson at the moment.

Augmented Text:
It's a pretty terrible time to be an athlete right now.
```

Εικόνα 30: Παράδειγμα Επαύξησης Δεδομένων με την Μέθοδο Back-Translation.

Η διαδικασία αντικατάστασης μέσω *BERT embeddings* χρησιμοποιεί την γνώση που λαμβάνει απο τα συμφραζόμενα της πρότασης ώστε να αντικαταστήσει κάποιες λέξεις αυτής με παρόμοιες οι οποίες ενιολογικά είναι πολύ κοντά στις αρχικές.

```
Original:
The quick brown fox jumps over the lazy dog
Augmented Text:
the quick thinking fox jumps over the lazy dog
```

Εικόνα 31: Παράδειγμα Επαύξησης Δεδομένων με την Μέθοδο Subtitute BERT Embeddings.

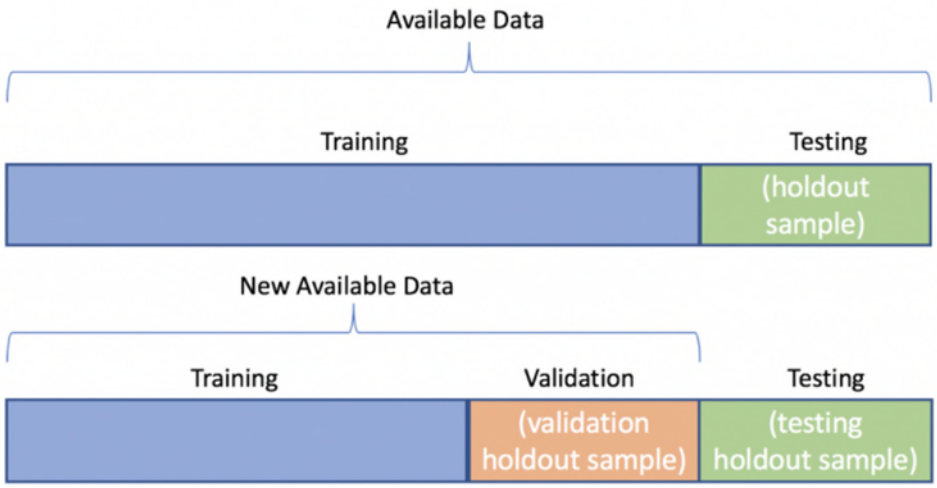
4.3.4 Αφαίρεση Ακραίων Τιμών – Outliers Removal

Στην διαδικασία ακραίων τιμών, ελέγχουμε εάν στο σύνολο δεδομένων μας υπάρχει συνδυασμός/οι αποτελέσματος που είναι μοναδικοί. Για παράδειγμα, Εάν ένα tweet χαρακτηριστεί απο το μοντέλο μας ώς (0: 'abusive', 1: 'hateful', 0: 'offensive', 0: 'disrespectful', 0: 'fearful', 1: 'normal') και αυτός ο συνδυασμός τιμών (0,1,0,0,0,1) δεν υπάρχει σε κανένα άλλο παράδειγμα, τότε αυτός ο συνδυασμός εντάσσεται στο σύνολο εκπαίδευσης διότι αποτελεί ακραία τιμή και είναι πιθανόν να βοηθήσει στην καλύτερη γενικοποίηση των αποτελεσμάτων του μοντέλου.

4.3.5 Τμηματοποίηση Δεδομένων – Data Splitting

Η τμηματοποίηση δεδομένων είναι η διαδικασία διαχωρισμού των δεδομένων σε 3 σύνολα:

- Δεδομένα που χρησιμοποιούμε για το σχεδιασμό των μοντέλων μας (**Σετ Εκπαίδευσης**)
- Δεδομένα που χρησιμοποιούμε για τη βελτίωση των μοντέλων μας (**Σέτ Επικύρωσης**)
- Δεδομένα που χρησιμοποιούμε για τον έλεγχο των μοντέλων μας (**Σετ Δοκιμών**)



Πηγή: datascience.stackexchange.com/questions/61467/clarification-on-train-test-and-val-and-how-to-use-implement-it

Εικόνα 32: Παράδειγμα Τμηματοποίησης Δεδομένων σε Train/Test/Validation Sets.

Αρχικά, το σετ εκπαίδευσης είναι το σύνολο δεδομένων που αναλύουμε (εκπαιδεύουμε) το μοντέλο μας ώστε να μάθεις τις αναπαραστάσεις των κειμένων. Στη συνέχεια, το σύνολο δοκιμών το χρησιμοποιήσουμε ως τελική δοκιμή μόλις αποφασίσουμε για το τελικό μας μοντέλο, ώστε να έχουμε την καλύτερη δυνατή εκτίμηση για το πόσο επιτυχημένο θα είναι το μοντέλο μας όταν χρησιμοποιηθεί σε εντελώς καινούργια δεδομένα. Έπειτα, το σύνολο επικύρωσης αποτελεί ένα μικρότερο σύνολο δεδομένων που δεν χρησιμοποιήσαμε κατά την εκπαίδευση του μοντέλου μας και το χρησιμοποιούμε για να αξιολογήσουμε πόσο καλά αποδίδουν οι κανόνες που έμαθε το μοντέλο μας σε νέα δεδομένα. Τέλος, είναι επίσης ένα σύνολο που χρησιμοποιούμε για τον συντονισμό παραμέτρων (optimization) και χαρακτηριστικών εισόδου για το μοντέλο μας, έτσι ώστε να μας δίνει την καλύτερη δυνατή απόδοση σε νέα δεδομένα.

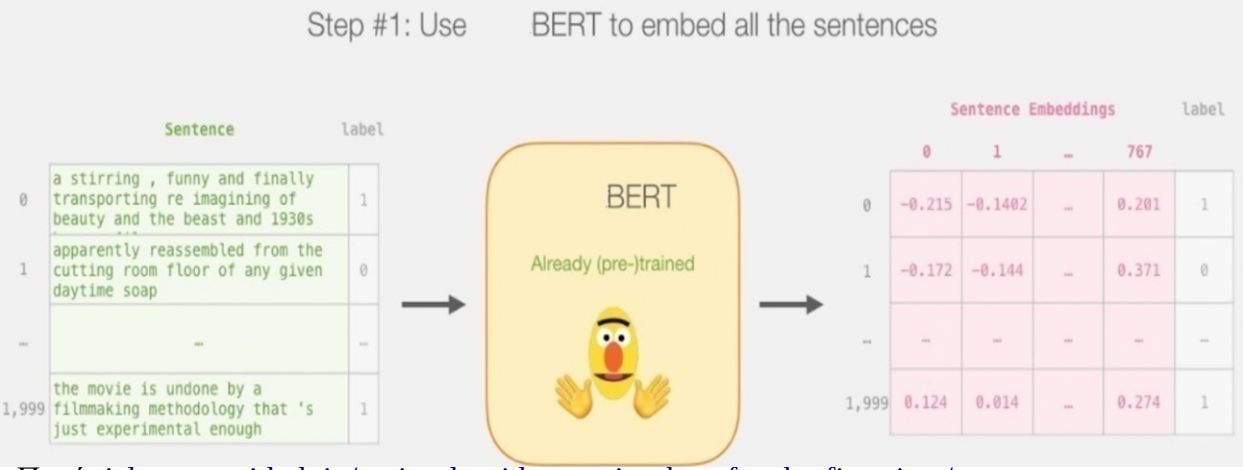
Στα πειράματα μας χρησιμοποιούμε αναλογία διαχωρισμού **train-test-evaluation 8:1:1,** δηλαδή το 80% των παραδειγμάτων μας χρησιμοποιούνται στο σέτ εκπαίδευσης, το 10% στο σέτ επαλήθευσης και το υπόλοιπο 10% στο σέτ ελέγχου.

4.4 B.E.R.T Αρχιτεκτονική – Transformer

Το BERT ουσιαστικά είναι μια στοίβα εκπαιδευμένων Transformer Encoders. Το BERT είναι ένα μοντέλο μοντελοποίησης γλώσσας της Google που δημιουργεί μοντέλα κορυφαίας ακρίβειας για ένα ευρύ φάσμα εργασιών χρησιμοποιώντας δύο βήματα: την προ-εκπαίδευση (pre-training) και τεχνικές τελειοποίησης (fine-tuning) .

Για να γίνει περισσότερο κατανοητή η λειτουργία της αρχιτεκτονικής B.E.R.T, θα αναλύσουμε την πορεία του μοντέλου και θα δείξουμε πώς παράγει την έξοδο του μέσω ενός παραδείγματος.

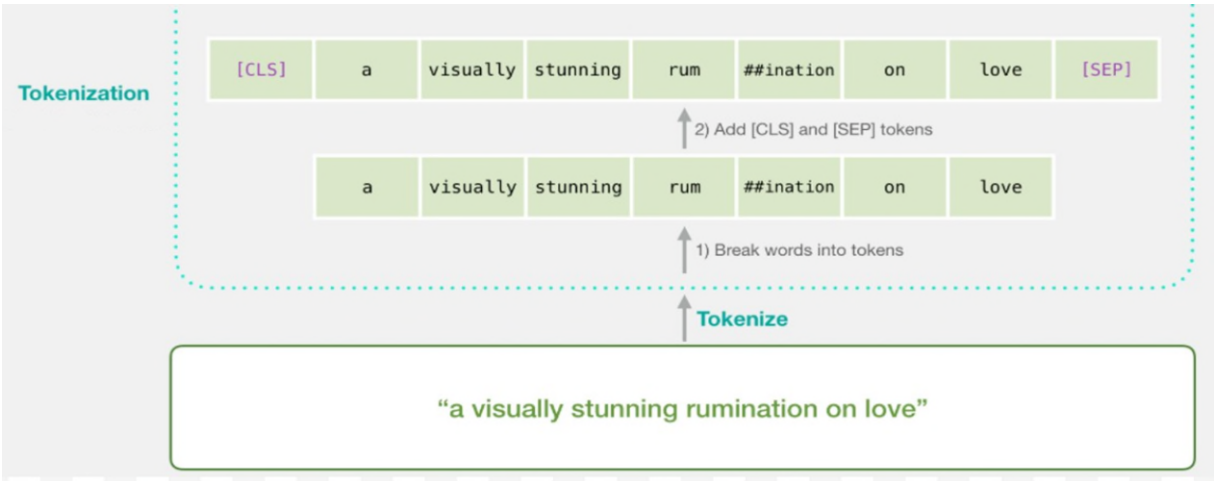
Αρχικά, μετά την προ-επεξεργασία τους, τα δεδομένα εισάγονται σε ενα επιπλέον στρώμα επεξεργασίας όπου οι προτάσεις (tweets) μετατρέπονται σε ενσωματώσεις (WordPiece embeddings).



Πηγή: jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Εικόνα 33: Μετατροπή Προτάσεων σε Ενσωματώσεις.

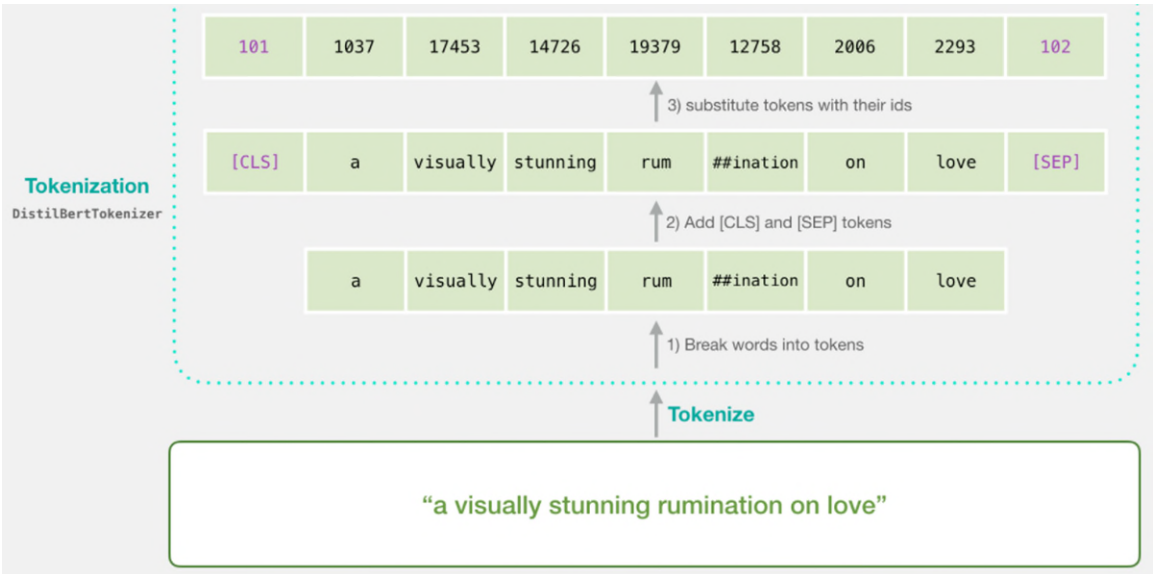
Στην συνέχεια, αναλύουμε τις προτάσεις σε ξεχωριστές λέξεις (tokenization). Έπειτα, προσθέτουμε τα ειδικά διακριτικά που απαιτεί η αρχιτεκτονική για ταξινομήσεις προτάσεων (αυτά είναι: το [CLS] στην πρώτη θέση και το [SEP] στο τέλος της πρότασης). Το πρώτο διακριτικό κάθε ακολουθίας είναι πάντα ένα ειδικό διακριτικό ταξινόμησης ([CLS]). Η τελική κρυφή κατάσταση (hidden state) που αντιστοιχεί σε αυτό το διακριτικό χρησιμοποιείται ως η αναπαράσταση της συνολικής ακολουθίας για εργασίες ταξινόμησης.



Πηγή: jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Εικόνα 34: Μετατροπή Προτάσεων σε Λέξεις & Εισαγωγή Διακριτικών Ταξινόμησης

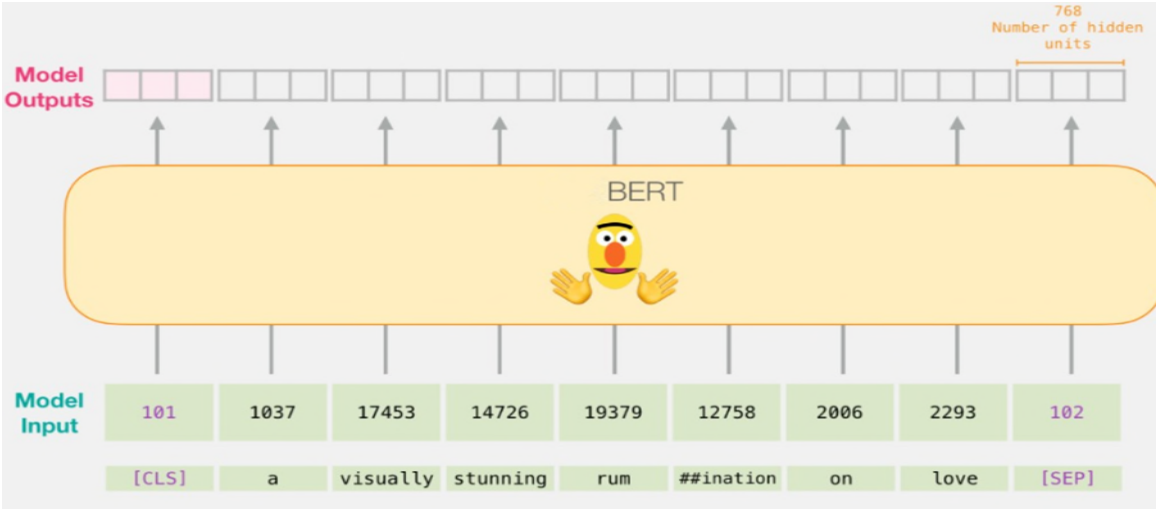
Μετέπειτα, αντικαθιστούμε κάθε λέξη με το αναγνωριστικό της από τον πίνακα με τις ενσωματώσεις, οι οποίες ενσωματώσεις έχουν προκύψει απο την προ-εκπαίδευση του μοντέλου σε τεράστια σύνολα δεδομένων.



Πηγή: jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Εικόνα 35: Αντικατάσταση Λεξεων με το Αναγνωριστικό Ενσωμάτωσης Τους

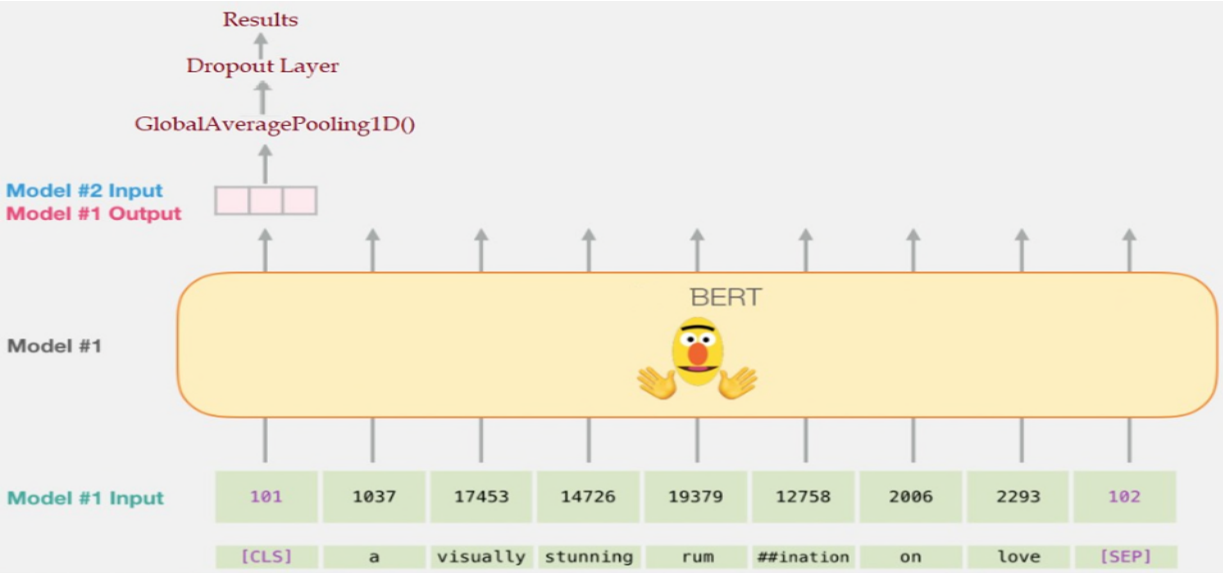
Η έξοδος που προκύπτει είναι ένα διάνυσμα για κάθε διακριτικό εισόδου, κάθε διάνυσμα στο μοντέλο B.E.R.T αποτελείται από 768 αριθμούς (floats).



Πηγή: jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Εικόνα 36: Έξοδος Μοντέλου B.E.R.T

Στην τελευταία φάση: λαμβάνουμε την έξοδο του διακριτικού ταξινόμησης [CLS] το οποίο περιέχει την τελευταία κρυφή κατάσταση (hidden state) της εισόδου, την τροφοδοτούμε σε ένα στρώμα καθολικής μέσης συγκέντρωσης (Global Average Pooling 1D), τροφοδοτούμε την έξοδο του στρώματος καθολικής μέσης συγκέντρωσης σε ένα στρώμα εγκατάλειψης (dropout layer), περνάμε το αποτέλεσμα του μέσα σε μια συνάρτηση ενεργοποίησης και λαμβάνουμε στην τελική έξοδο τα αποτελέσματα του μοντέλου, των οποίων η μορφή εξαρτάται απο τον τύπο ταξινόμησης και την συνάρτηση ενεργοποίησης που έχει επιλεχθεί.



Πηγή: jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Εικόνα 37: Συνολική Εικόνα Λειτουργίας Ταξινόμησης της Αρχιτεκτονικής B.E.R.T.

5

Πειραματικά Αποτελέσματα

Σε αυτό το κεφάλαιο θα εξηγήσουμε αναλυτικά τον τρόπο υλοποίησης των πειραμάτων μας και των αποτελεσμάτων τους. Αρχικά θα παρουσιάσουμε την στατιστική ανάλυση των δεδομένων μας με και χωρίς επαύξηση. Στην συνέχεια, αναφέρουμε το ακριβή μοντέλο αρχιτεκτονικής βαθιάς μάθησης που χρησιμοποιήσαμε. Μετέπειτα θα απαριθμήσουμε τις παραμέτρους που θέσαμε για την καλύτερη απόδοση των μοντέλων μας. Τέλος, θα παρουσιάσουμε τα αποτελέσματα της διπλωματικής μας έρευνας με και χωρίς την επαύξηση δεδομένων και θα τα συγκρίνουμε με αυτά της έρευνας (*Ousidhoum et al 2019*)[17].

5.1 Περιγραφή Συνόλου Δεδομένων

Για την υλοποίηση των πειραμάτων της διπλωματικής μας εργασίας χρησιμοποιήσαμε, όπως αναφέραμε και στα κεφάλαια 3-4 το σύνολο δεδομένων MLMA[18]. Το σύνολο των tweets, χρησιμοποιήθηκε για την ταξινόμηση πολλάπλών ετικετών σε 6 κατηγορίες στόχους (‘‘abusive / hateful / offensive / disrespectful / fearful / normal’’) και η κάθε κατηγορία μπορεί να λάβει δύο τιμές μηδέν (0) και ένα (1) αναλόγως με την ύπαρξη της ή όχι στο tweet.

Στο παρακάτω σχήμα βλέπουμε μέσω ενός μικρού δείγματος (14 παραδείγματα), την μορφή των δεδομένων αφού έχουν περάσει απο την φάση της προεπεξεργασίας.

	tweet	abusive	hateful	offensive	disrespectful	fearful	normal	tweet_cleaned	tweet_length
0	call sis im boy girl still faggot shut	0	0	1	0	0	0	call sis boy girl still faggot shut	7
1	@user @user legit nilas retarded idk	0	0	1	1	0	0	legit nilas retarded idk	4
2	said retard @url	1	1	1	0	1	0	said retard	2
3	america another 8 years obama's ideology via h...	1	1	0	1	1	1	america another years obama ideology via hilla...	13
4	@user don... fucking made cry twat.	0	0	1	0	0	0	don fucking made cry twat	5
5	one wants 2 negro swan	0	0	1	0	0	0	wants negro swan	3
6	@user @user @user hahaha. one trump's shithole...	0	0	1	0	0	0	hahaha trump shithole countries	4
7	u juul ur faggot	0	0	1	0	0	0	juul ur faggot	3
8	@user @user even called country shithole peopl...	0	0	1	0	0	0	even called country shithole people shithole p...	8
9	canadians never met seen associated person cur...	0	0	1	0	0	0	canadians never met seen associated person cur...	10
10	@user @user south africa years come. keep supp...	0	0	1	0	0	0	south africa years come keep supporting loud m...	9

Εικόνα 38: Μορφή Προ-Επεξεργασμένων Δεδομένων του Πειράματος.

Το περιεχόμενο της κάθε στήλης του παραπάνω σχήματος αναλύεται στον παρακάτω πίνακα.

Tweet	Η μορφή του tweet πριν την διαδικασία της προ-επεξεργασίας
Abusive	Η κατηγορία που δείχνει εάν το tweet είναι υβριστικό
Hateful	Η κατηγορία που δείχνει εάν το tweet αποτελεί προιόν ρητορικής μίσους
Offensive	Η κατηγορία που δείχνει εάν το tweet είναι επιθετικό
Disrespectful	Η κατηγορία που δείχνει εάν το tweet προκαλεί αισθήματα ασέβειας
Feaful	Η κατηγορία που δείχνει εάν το tweet προκαλεί αισθήματα φόβου
Normal	Η κατηγορία που δείχνει εάν το tweet είναι κανονικό
Tweet_Cleaned	Η μορφή του tweet μετά την διαδικασία της προ-επεξεργασίας
Tweet_Length	Το μήκος του tweet (σε λέξεις)

Πίνακας 1: Επεξήγηση Στηλών Συνόλου Δεδομένων

5.1.1 Στατιστική Ανάλυση Δεδομένων

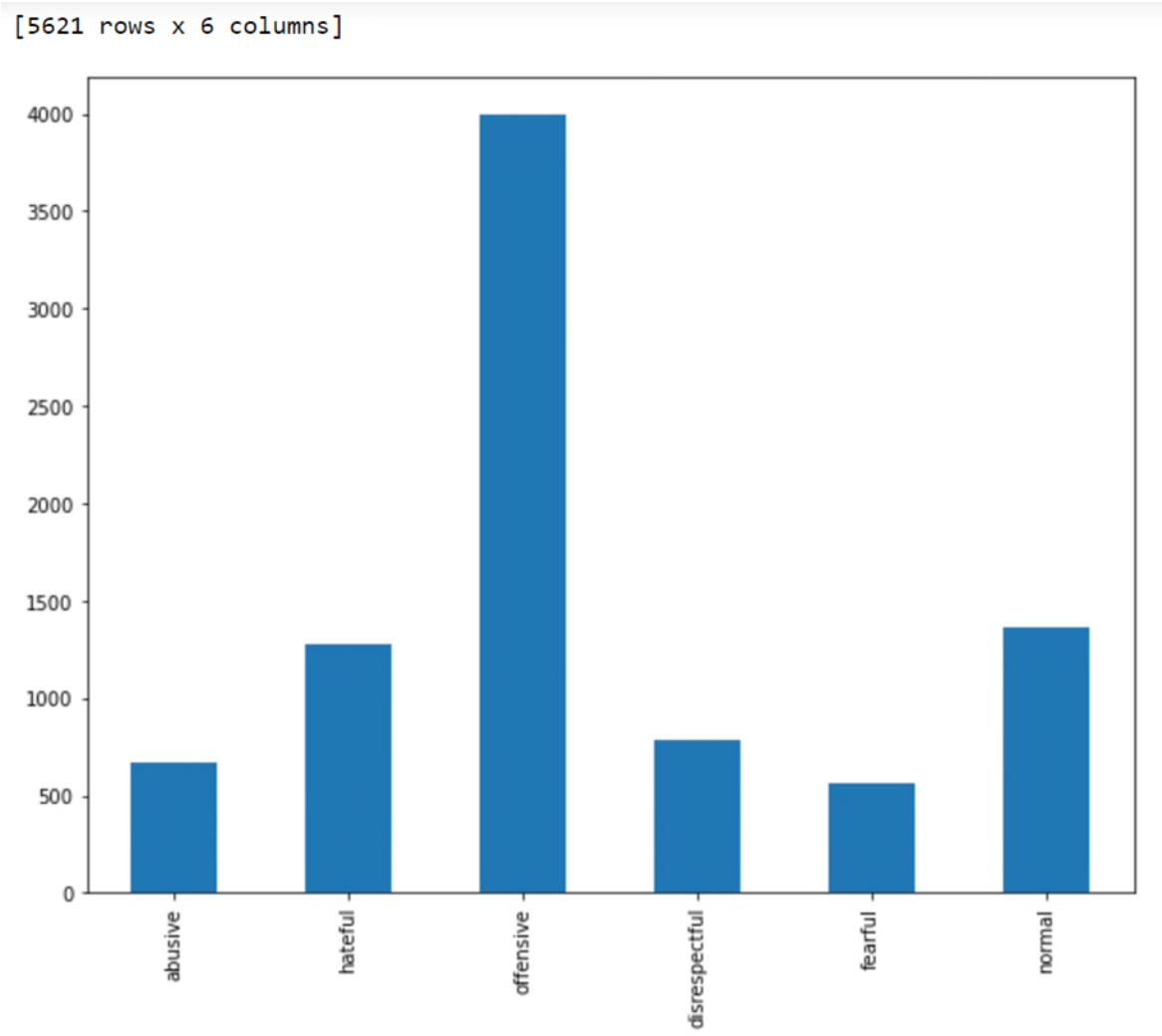
Σε αυτή την ενότητα, παρουσιάζουμε μια στατιστική ανάλυση του συνόλου δεδομένων μας πριν και μετά απο την διαδικασία επαύξησης δεδομένων (data augmentation), η οποία πραγματοποιήθηκε μέσω της μεθόδου μετάφρασης δυο κατευθύνσεων (back-translate).

5.1.1.1 Στατιστική Ανάλυση Αρχικών Δεδομένων

Αρχικά παραθέτουμε την κατανομή του συνόλου αρχικών δεδομένων του κάθε είδους ρητορικής μίσους στη μορφή ενός πίνακα και στην συνέχεια σε μορφή διαγράμματος μπάρας, ώστε να διαπιστωθεί κατά πόσο υπάρχουν αρκετά διαθέσιμα tweets από κάθε κατηγορία ώστε να παράγουμε ικανοποιητικά αποτελέσματα.

ΕΤΙΚΕΤΕΣ	ΑΡΙΘΜΟΣ ΠΑΡΑΔΕΙΓΜΑΤΩΝ
Abusive	671
Hateful	1278
Offensive	3994
Disrespectful	782
Feaful	562
Normal	1359

Πίνακας 2: Ποσοτική Κατανομή Παραδειγμάτων Αρχικών Δεδομένων



Εικόνα 39: Ποσοτική Κατανομή Παραδειγμάτων Αρχικών Δεδομένων σε Μορφή Μπάρας

5.1.1.2 Στατιστική Ανάλυση Αρχικών Διαχωρισμένων Δεδομένων

```
 Value Counts of Training set Categories before augmentation and after split:

abusive: 538
hateful: 1031
offensive: 3187
disrespectful: 638
fearful: 465
normal: 1096

 Value Counts of Val set Categories before augmentation and after split:

abusive: 72
hateful: 124
offensive: 398
disrespectful: 71
fearful: 50
normal: 130

 Value Counts of Test set Categories before augmentation and after split:

abusive: 60
hateful: 122
offensive: 407
disrespectful: 72
fearful: 47
normal: 128
```

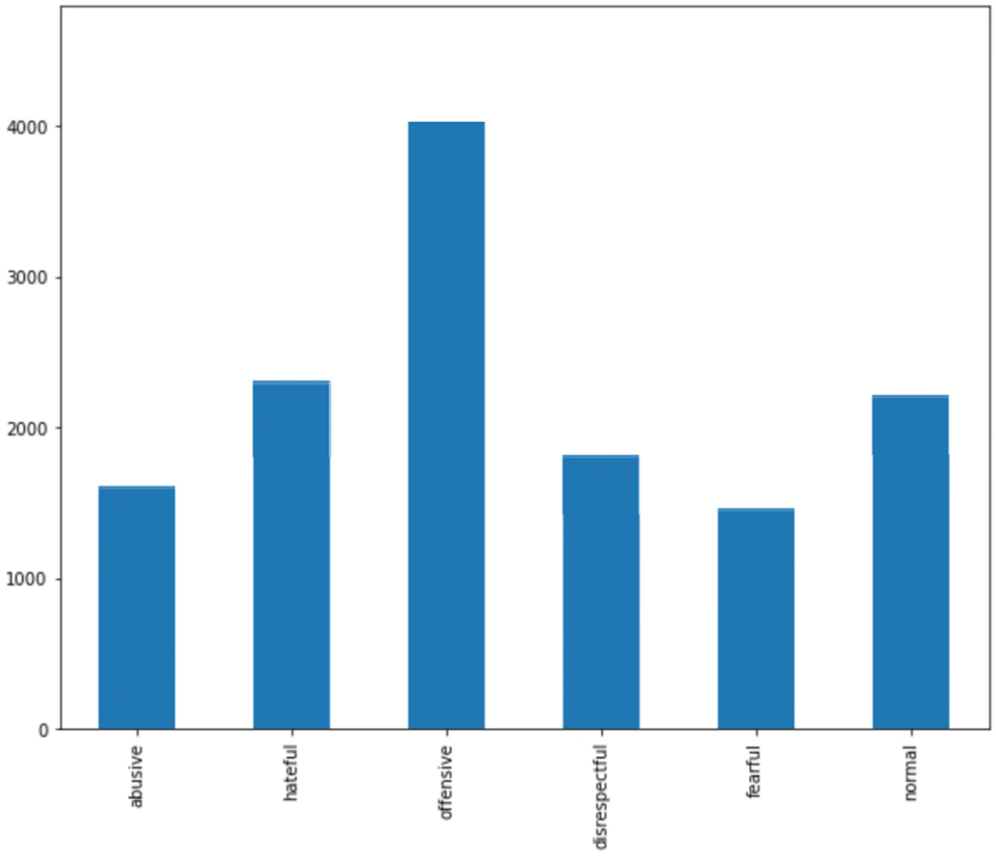
Εικόνα 40: Ποσοτική Κατανομή Παραδειγμάτων Αρχικών Δεδομένων μετά τον Διαχωρισμό σε σετ Εκπαίδευσης-Εκτίμησης-Ελέγχου

5.1.1.3 Στατιστική Ανάλυση Επαυξημένων Δεδομένων

Στην συνέχεια, παραθέτουμε την κατανομή του συνόλου επαυξημένων δεδομένων του κάθε είδους ρητορικής μίσους στη μορφή ενός πίνακα και στην συνέχεια σε μορφή διαγράμματος μπάρας, ώστε να διαπιστωθεί το αντίκτυπο της χρήσης της μεθόδου back-translating στον αριθμό παραδειγμάτων.

ΕΤΙΚΕΤΕΣ	ΑΡΙΘΜΟΣ ΠΑΡΑΔΕΙΓΜΑΤΩΝ
Abusive	1568
Hateful	2112
Offensive	3994
Disrespectful	1792
Feaful	1376
Normal	2109

Πίνακας 3: Ποσοτική Κατανομή Παραδειγμάτων Επαυξημένων Δεδομένων



Εικόνα 41: Ποσοτική Κατανομή Παραδειγμάτων Επαυξημένων Δεδομένων σε Μορφή Μπάρας

5.1.1.4 Στατιστική Ανάλυση Επαυξημένων Διαχωρισμένων Δεδομένων

```
 Value Counts of Training set Categories AFTER augmentation and after split:

abusive: 1436
hateful: 1866
offensive: 3187
disrespectful: 1649
fearful: 1279
normal: 1851

 Value Counts of Val set Categories AFTER augmentation and after split:

abusive: 72
hateful: 124
offensive: 398
disrespectful: 71
fearful: 50
normal: 130

 Value Counts of Test set Categories AFTER augmentation and after split:

abusive: 60
hateful: 122
offensive: 407
disrespectful: 72
fearful: 47
normal: 128
```

Εικόνα 42: Ποσοτική Κατανομή Παραδειγμάτων Επαυξημένων Δεδομένων μετά τον Διαχωρισμό σε σετ Εκπαίδευσης-Εκτίμησης-Ελέγχου

5.2 Περιγραφή Μοντέλου

Για την μετατροπή των δεδομένων της εισόδου μας σε κατάλληλη μορφή ώστε να εισαχθούν στο μοντέλο B..E.R.T, χρησιμοποιήσαμε το στρώμα προεπεξεργασίας του Tensorflow Hub *bert_en_uncased_preprocess.*

Για την υλοποίηση του B.E.R.T, χρησιμοποιήσαμε το μοντέλο *bert_en_uncased_L-12_H-768_A-12* της βιβλιοθήκης Tensorflow Hub με χαρακτηριστικά: L=12 κρυφά στρώματα (i.e., Transformer blocks), Hidden Size, H=768, Attention Heads, A=12.

Χρησιμοποιώντας την έξοδο του B.E.R.T μοντέλου, ακολουθούμε με την σειρά τα εξής βήματα ώστε να λάβουμε το αποτέλεσμα της ταξινόμησης:

1. Τροφοδοτούμε την έξοδο του μοντέλου B.E.R.T σε ένα στρώμα *GlobalAveragePooling1D* ώστε να μειώσουμε την διάσταση του.

2. Τροφοδοτούμε την έξοδο του στρώματος *GlobalAveragePooling1D* σε ένα στρώμα εγκατάληψης (Drop-out Layer) για κανονικοποίηση και αποφυγή του φαινομένου της υπερπροσαρμογής δεδομένων (Overfitting).

3. Τροφοδοτούμε την έξοδο του στρώματος εγκατάληψης σε ένα βαθιά συνδεδεμένο στρώμα νευρωνικού δικτύου το οποίο μας δίνει το αποτέλεσμα της ταξινόμησης.

Στο παρακάτω διάγραμμα παρουσιάζεται η τελική δομή του B.E.R.T μοντέλου μας, όπως αυτή παράγεται απο την βιβλιοθήκη βαθιάς μάθησης Tensorflow.

```
print(model.summary())
```

```
Model: "model"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            [(None,)]            0
__________________________________________________________________________________________________
keras_layer (KerasLayer)        {'input_mask': (None 0           input_1[0][0]
__________________________________________________________________________________________________
keras_layer_1 (KerasLayer)      {'sequence_output':  109482241   keras_layer[0][0]
                                                                 keras_layer[0][1]
                                                                 keras_layer[0][2]
__________________________________________________________________________________________________
global_average_pooling1d (Globa (None, 768)          0           keras_layer_1[0][14]
__________________________________________________________________________________________________
dropout (Dropout)               (None, 768)          0           global_average_pooling1d[0][0]
__________________________________________________________________________________________________
outputs (Dense)                 (None, 6)            4614        dropout[0][0]
==================================================================================================
Total params: 109,486,855
Trainable params: 109,486,854
Non-trainable params: 1
__________________________________________________________________________________________________
```

Εικόνα 43: Τελική Δομή του Μοντέλου Ταξινόμησης Μας.

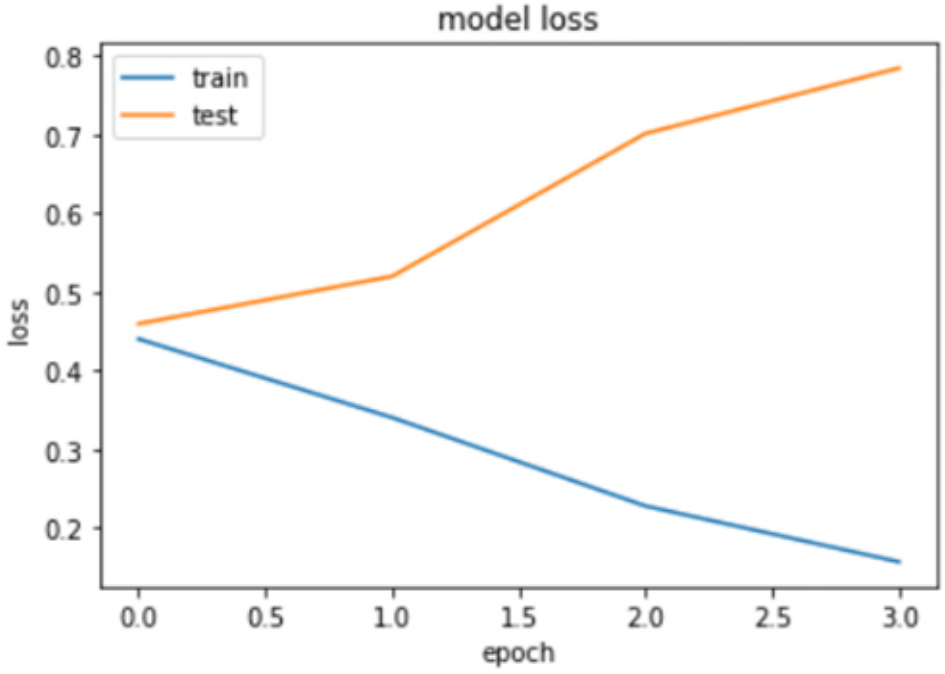
5.3 Επιλογή Υπερπαραμέτρων – Hyper parameter Tuning

Για τα μοντέλα μας, ορίσαμε 5 εποχές για την εκπαίδευση τους ώστε κάθε ένα να έχει την δυνατότητα να συγκλίνει (*converge*) και να βελτιώσει την απόδοση του. Επίσης θέσαμε την δυνατότητα να αποθηκεύονται τα βάρη των μοντέλων μας μέσω της συνάρτησης επανάκλησης (*callback function*) *ModelCheckpoint* της βιβλιοθήκης *Keras.*

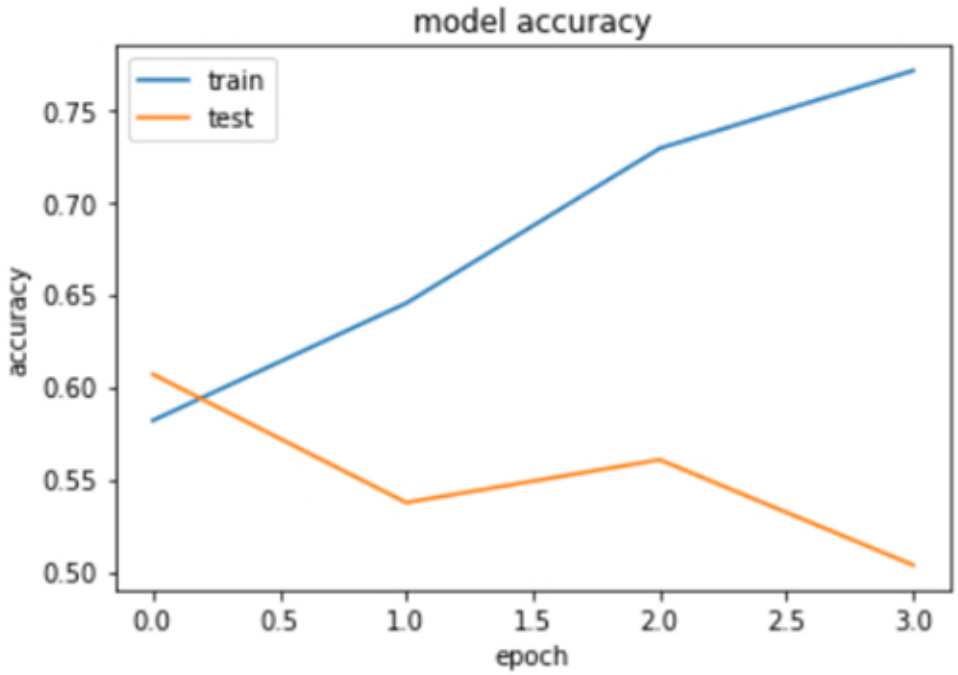
5.3.1 Διαγράμματα Εκπαίδευσης Μοντέλου Χωρίς Επαύξηση Δεδομένων

Παρακάτω αποτυπώνονται με την σειρά, τα διαγράμματα του συνόλου εκπαίδευσης που χρησιμοποιήσαμε για την επιλογή των υπερπαραμέτρων:

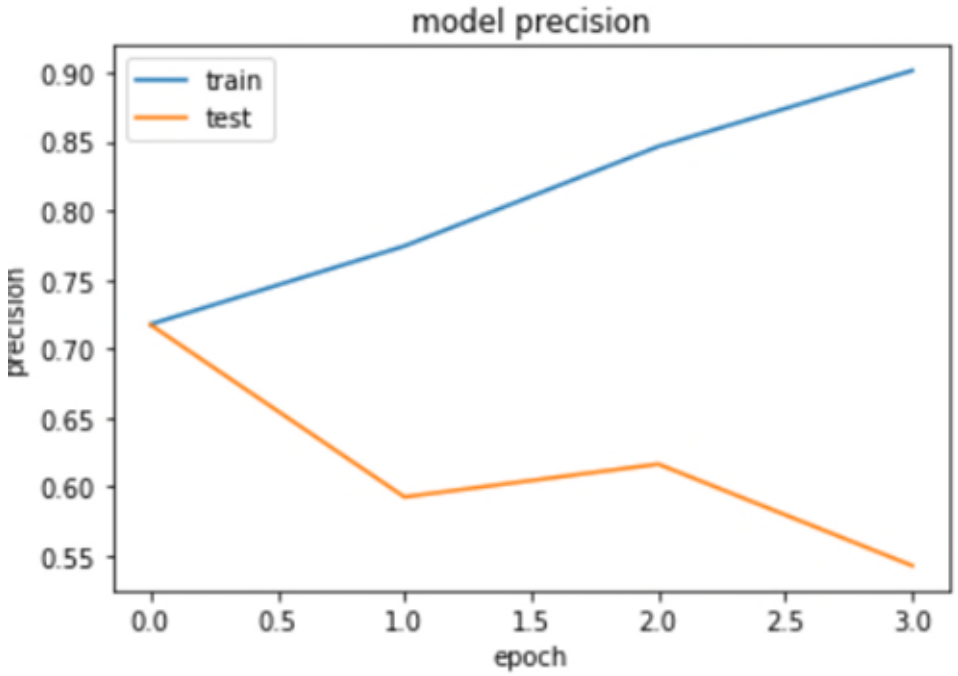
- Loss – Validation Loss
- Accuracy – Validation Accuracy
- Precision – Validation Precision
- AUC – Validation AUC



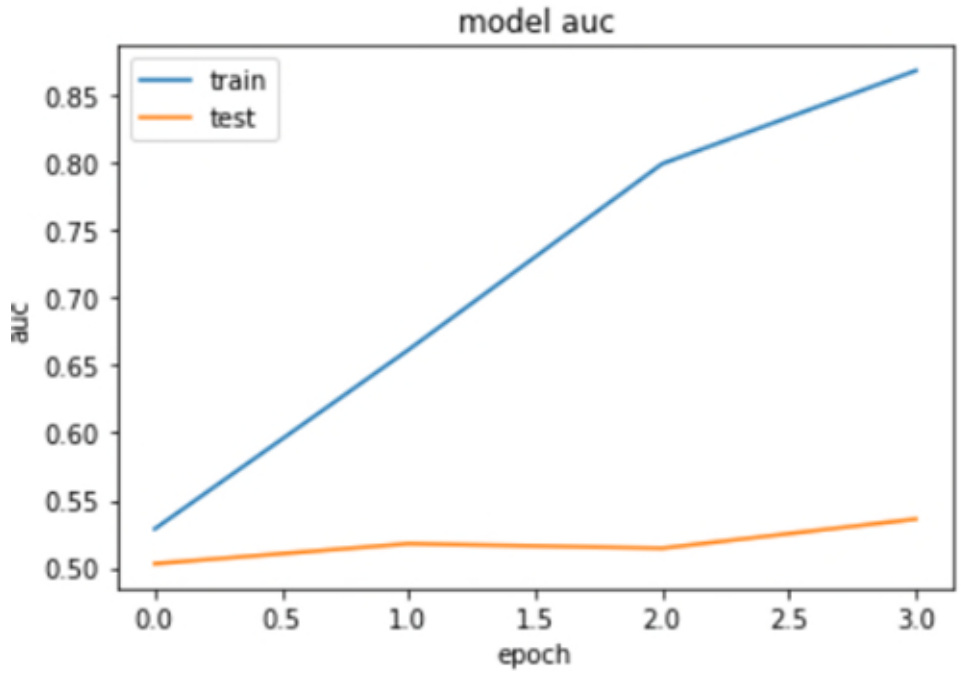
Εικόνα 44: Γράφημα Loss – Validation Loss του Μοντέλου Χωρίς Επαύξηση Δεδομενων.



Εικόνα 45: Γράφημα Accuracy – Validation Accuracy του Μοντέλου Χωρίς Επαύξηση Δεδομενων



Εικόνα 46: Γράφημα Precision – Validation Precision του Μοντέλου Χωρίς Επαύξηση Δεδομενων



Εικόνα 47: Γράφημα AUC – Validation AUC του Μοντέλου Χωρίς Επαύξηση Δεδομενων

Οι υπερπαράμετροι που επιλέχθηκαν για την καλύτερη απόδοση του μοντέλου, δίνονται αναλυτικά στον παρακάτω πίνακα:

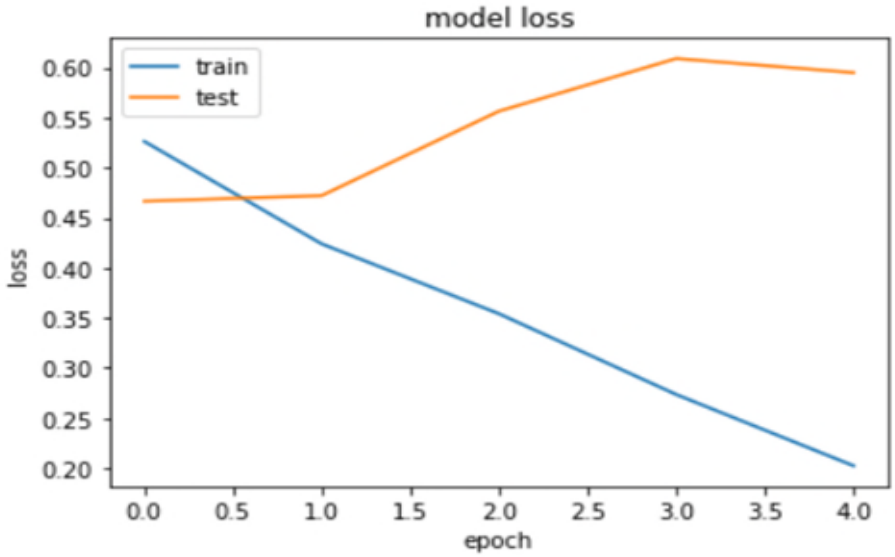
Dropout Layer Rate	0.3
Loss Function	Sigmoid
Optimizer	Adam
Learning Rate	1e-4
Decay	1e-6
Epochs	15
Activation Function	Binary_Crossentropy

Πίνακας 4: Πίνακας Υπερ-παραμέτρων του Μοντέλου Χωρίς Επαύξηση Δεδομενων

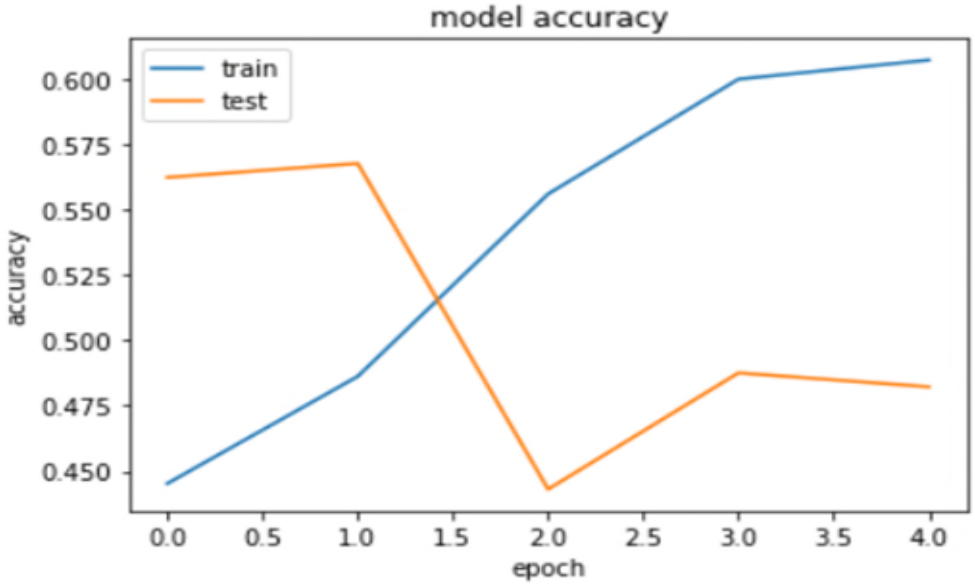
5.3.2 Διαγράμματα Εκπαίδευσης Μοντέλου Με Επαύξηση Δεδομένων

Παρακάτω αποτυπώνονται με την σειρά, τα διαγράμματα του συνόλου εκπαίδευσης που χρησιμοποιήσαμε για την επιλογή των υπερπαραμέτρων:

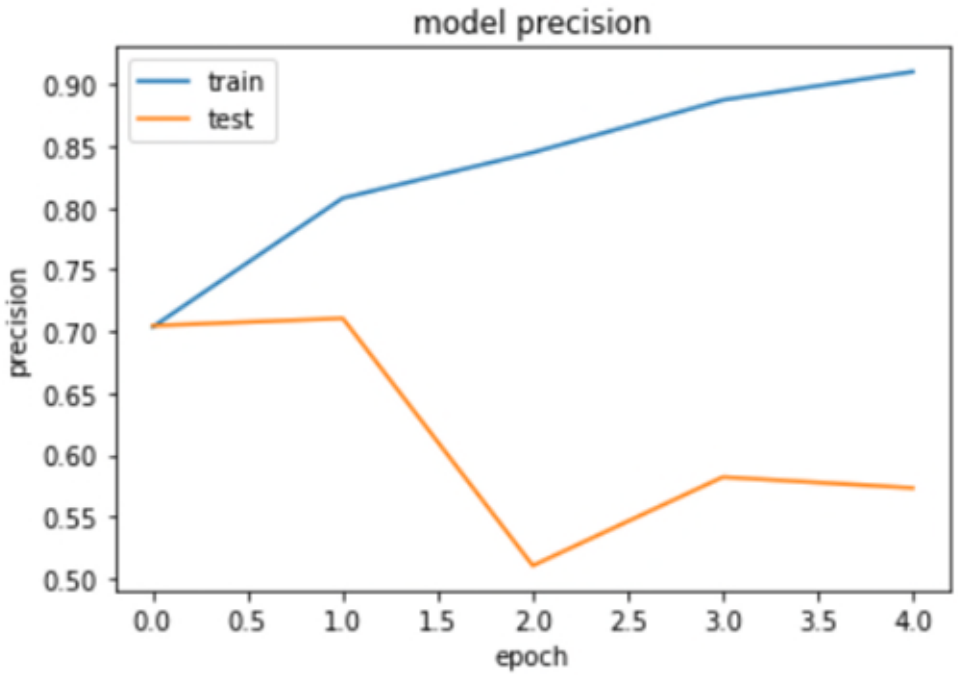
- Loss – Validation Loss
- Accuracy – Validation Accuracy
- Precision – Validation Precision
- AUC – Validation AUC



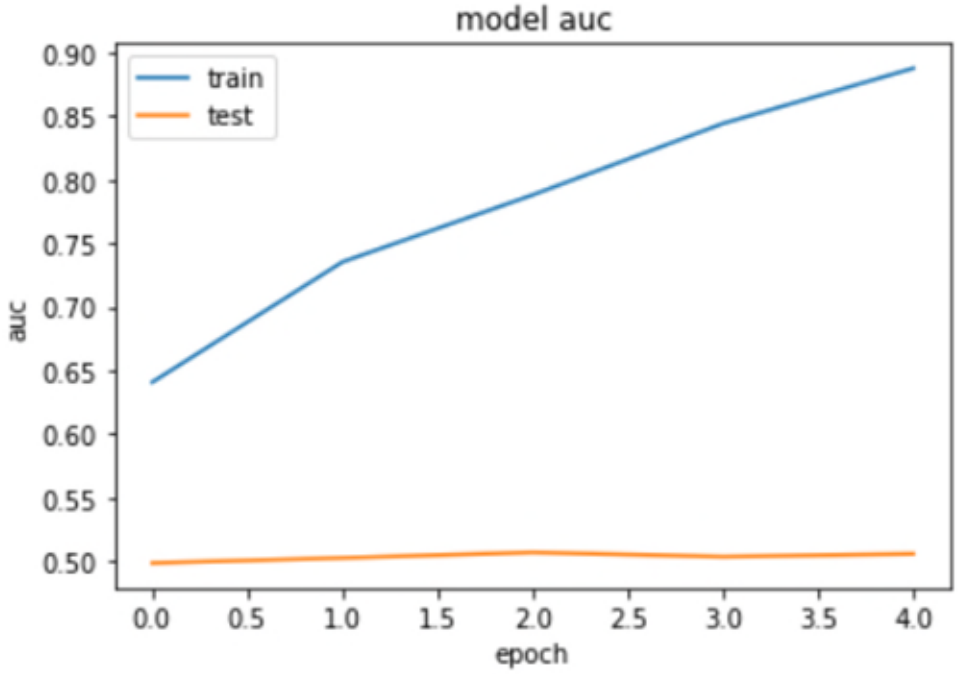
Εικόνα 48: Γράφημα Loss – Validation Loss του Μοντέλου Με Επαύξηση Δεδομενων.



Εικόνα 49: Γράφημα Accuracy – Validation Accuracy του Μοντέλου Με Επαύξηση Δεδομενων



Εικόνα 50: Γράφημα Precision – Validation Precision του Μοντέλου Με Επαύξηση Δεδομενων



Εικόνα 51: Γράφημα AUC – Validation AUC του Μοντέλου Με Επαύξηση Δεδομενων

Οι υπερπαράμετροι που επιλέχθηκαν για την καλύτερη απόδοση του μοντέλου, δίνονται αναλυτικά στον παρακάτω πίνακα:

Dropout Layer Rate	0.3
Loss Function	Sigmoid
Optimizer	Adam
Learning Rate	1e-4
Decay	1e-6
Epochs	15
Activation Function	Binary_Crossentropy

Πίνακας 5: Πίνακας Υπερ-παραμέτρων του Μοντέλου Με Επαύξηση Δεδομενων

5.4 Συγκεντρωτικά Αποτελέσματα

Μετά την εκπαίδευση των μοντέλων μας, ελέγχουμε αναλυτικά την απόδοση τους ώστε να αποφανθούμε για τα τελικά τους αποτελέσματα. Σε αυτή την ενότητα, παρουσιάζουμε και για τα τρία σέτ δεδομένων (Train, Evaluate, Test) τις εξής μετρήσεις:

- Recall Micro/Macro/Weighted για κάθε κατηγορία
- Precision Micro/Macro/Weighted για κάθε κατηγορία
- AUC Weighted
- F1 Micro/Macro/Weighted για κάθε κατηγορία

5.4.1 Συγκεντρωτικά Αποτελέσματα Μοντέλου Χωρίς Επαύξηση Δεδομένων

	Precision	Recall	F1	AUC
Micro Average	0.94	0.93	0.94	
Macro Average	0.92	0.91	0.91	
Weighted Average	0.95	0.93	0.94	0.94

Πίνακας 6: Μοντέλο Χωρίς Επαύξηση - Πίνακας Συγκεντρωτικών Αποτελεσμάτων του Συνόλου Εκπαίδευσης

	Precision	Recall	F1	AUC
Micro Average	0.54	0.46	0.50	
Macro Average	0.32	0.26	0.28	
Weighted Average	0.50	0.46	0.47	0.53

Πίνακας 7: Μοντέλο Χωρίς Επαύξηση - Πίνακας Συγκεντρωτικών Αποτελεσμάτων του Συνόλου Αποτίμησης

	Precision	Recall	F1	AUC
Micro Average	0.49	0.42	0.45	
Macro Average	0.31	0.24	0.26	
Weighted Average	0.46	0.42	0.43	0.51

Πίνακας 8: Μοντέλο Χωρίς Επαύξηση - Πίνακας Συγκεντρωτικών Αποτελεσμάτων του Συνόλου Ελέγχου

5.4.2 Συγκεντρωτικά Αποτελέσματα Μοντέλου Με Επαύξηση Δεδομένων

	Precision	Recall	F1	AUC
Micro Average	0.81	0.58	0.68	
Macro Average	0.89	0.53	0.63	
Weighted Average	0.86	0.58	0.65	0.68

Πίνακας 9: Μοντέλο Με Επαύξηση - Πίνακας Συγκεντρωτικών Αποτελεσμάτων του Συνόλου Εκπαίδευσης

	Precision	Recall	F1	AUC
Micro Average	0.70	0.45	0.55	
Macro Average	0.12	0.16	0.13	
Weighted Average	0.33	0.45	0.38	0.50

Πίνακας 10: Μοντέλο Με Επαύξηση - Πίνακας Συγκεντρωτικών Αποτελεσμάτων του Συνόλου Αποτίμησης

	Precision	Recall	F1	AUC
Micro Average	0.72	0.46	0.56	
Macro Average	0.12	0.16	0.14	
Weighted Average	0.36	0.46	0.40	0.56

Πίνακας 11: Μοντέλο Με Επαύξηση - Πίνακας Συγκεντρωτικών Αποτελεσμάτων του Συνόλου Ελέγχου

5.5 Σύγκριση Αποτελεσμάτων με Παρόμοια Επιστημονικά Έργα

Σε αυτήν την ενότητα θα συγκρίνουμε τα αποτελέσματα της διπλωματικής μας με την έρευνα (*Ousidhoum et al 2019*), η οποία δημιούργησε το σύνολο δεδομένων που χρησιμοποιήσαμε. Στην δική τους έρευνα αντί για το μοντέλο B.E.R.T χρησιμοποιούν ένα αμφίδρόμο-LSTM (biLSTM) μοντέλο και την αρχιτεκτονική (*Sluice Networks*) η οποία είναι δομημένη με τέτοιο τρόπο ώστε να δυαμοιράζει κατάλληλα τα βάρη μεταξύ των κλάσεων στόχων κατα την εκπαίδευση.

Στο παρακάτω σχήμα, με κίτρινο φόντο, βλέπουμε τα αποτελέσματα της έρευνας (*Ousidhoum et al 2019*), βασισμένα στο ίδιο σύνολο δεδομένων και στον ίδιο σκοπό ταξινόμησης.

Attribute	Model	Macro-F1				Micro-F1			
		EN	FR	AR	Avg	EN	FR	AR	Avg
Tweet	Majority	0.24	0.19	0.20	0.21	0.41	0.27	0.27	0.32
	LR	0.14	0.20	0.25	0.20	0.54	0.56	**0.48**	0.53
	STSL	0.24	0.12	0.31	0.23	0.49	0.51	0.47	0.49
	MTSL	0.09	0.20	0.33	0.21	**0.55**	**0.59**	0.46	**0.54**
	STML	0.04	0.07	**0.35**	0.16	0.54	0.47	0.37	0.46
	MTML	**0.30**	**0.28**	**0.35**	**0.31**	0.45	0.48	0.44	0.46

Πηγή: arxiv.org/pdf/1908.11049.pdf

Εικόνα 52: Συγκριτικά Αποτελέσματα της Έρευνας του συνόλου Δεδομένων MLMA.

Παρατηρούμε πώς το μοντέλο μας με επαύξηση δεδομένων καταφέρνει να επιτύχει υψηλότερες επιδόσεις απο τις υπάρχουσες στο συγκεκριμένο σύνολο δεδομένων με *micro-average*: **0.56** και *macro-average*: **0.14** της βαθμολογίας F1.

Αντίστοιχα το μοντέλο μας χωρίς επαύξηση καταφέρνει να επιτύχει υψηλότερη επίδοση στην κατηγορία *macro-average*: **0.26** ενώ στην κατηγορία *micro-average*: **0.45** δεν καταφέρνει να ξεπεράσει τις επιδόσεις της έρευνας προς σύγκριση.

6

Συμπεράσματα – Προκλήσεις - Προοπτικές

Σε αυτό το τελευταίο κεφάλαιο παρουσιάζουμε τα συμπεράσματα που προκύπτουν απο τα αποτελέσματα των πειραμάτων και των επιδόσεων των μοντέλων μας, αναφέρουμε τις προκλήσεις που αντιμετωπίσαμε κατα την διεξαγωγή της έρευνας μας και κλείνουμε με τις προοπτικές που προσφέρει το έργο μας για μελλοντικές μελέτες.

6.1 Συμπεράσματα

Η διπλωματική μας εργασία είχε ως στόχο την ταξινόμηση ρητορικής μίσους μέσω αρχιτεκτονικών βαθιάς μάθησης και συγκεκριμένα της αρχιτεκτονικής BERT. Στην βάση αυτή, αρχικά, παρουσιάσαμε και αναλύσαμε όλους τους τομείς της τεχνικής νοημοσύνης, τον κλάδο της ταξινόμησης ρητορικής μίσους και την βιβλιογραφία της. Στην συνέχεια, υλοποιήσαμε ταξινόμηση πολλαπλών ετικετών με 6 κλάσσεις στόχους (abusive, hateful, offensive, disrespectful, fearful, normal) σε σύνολο δεδομένων επιλεγμένο κατάλληλα μέσω της μεθοδολογίας που προτείναμε ώστε να ελαχιστοποιεί τα προβληματα της προκατάληψης σχολιασμού και της φυλετικής προκατάληψης. Έπειτα, παρουσιάσαμε τις προσεγγίσεις μας και τις τεχνικές που υλοποιήσαμε. Τέλος, χρησιμοποιήσαμε το μοντέλο B.E.R.T για την δημιουργία 2 μοντέλων (με επαύξηση δεδομένων και χωρίς) για την ταξινόμηση και επιτύχαμε αποδεκτά αποτελέσματα πού ξεπερνούν τα ήδη υπάρχοντα στο ίδιο σύνολο δεδομένων και στον ίδιο τύπο ταξινόμησης.

Σύμφωνα με τα αποτελέσματα των πειραμάτων, το μοντέλο μας με επαύξηση δεδομένων πέτυχαίνει απόδοση F1 *micro-average*: **56%** , *macro-average*: **14%** και απόδοση AUC της τάξης του **56%**, αποτέλεσμα που αποτελεί το καλύτερο στο επιλεγμένο σύνολο δεδομένων και σε αυτόν τον τύπο ταξινόμησης. Αντίστοιχα το μοντέλο μας χωρίς επαύξηση καταφέρνει να επιτύχει υψηλότερη επίδοση στην κατηγορία *macro-average*: **26%** ενώ στην κατηγορία *micro-average*: **45%** καταφέρνει να ξεπεράσει τις επιδόσεις της έρευνας προς σύγκριση.

Η έρευνα μας φαίνεται να υποστηρίζει το επιχείρημα πώς οι αρχιτεκτονικές Transformers και τα προ-εκπαιδευμένα μοντέλα είναι ικανά να επιτύχουν καλύτερα αποτελέσματα απο τα πιο απλά σε δομή νευρωνικά δίκτυα. Επίσης, υπογραμίζει την μεγάλη σημασία της προ-επεξεργασίας και της ποσότητας δεδομένων που απαιτούνται για την καλή αποδοτικότητα ενός μοντέλου.

Όσο αφορά τα αποτελέσματα της βαθμολογίας F1, το ‘*micro-average’* υπολογίζετε μετρώντας τα συνολικά αληθινά θετικά, ψευδώς αρνητικά και ψευδώς θετικά παραδείγματα και αποτελεί την πιο σημαντική μέτρηση σε πειράματα όπου το σύνολο δεδομένων δεν είναι σωστά ισσοροπημένο (imbalanced). Αντίθετα η βαθμολογία ‘*macro-average’* υπολογίζει τις μετρήσεις για κάθε ετικέτα και βρίσκει την μη σταθμισμένη μέση τιμή τους χωρίς να παίρνει υπόψη την ανισσοροπία στον αριθμό των ετικετών. Αυτός είναι και ο λόγος για τον οποίο παρατηρούμε πολύ χαμηλές επιδόσεις στα πειράματα μας στην βαθμολογία ‘*macro-average’* αλλά και στα πειράματα της έρευνας με την οποία συγκρίνουμε τα αποτελέσματα μας.

Στην πειραματική φάση της εργασίας μας μία απο τις βασικές προκλήσεις τις οποίες αντιμετωπίσαμε αποτελεί το πλήθος παραδειγμάτων του επιλεγμένου συνόλου δεδομένων ρητορικής μίσους. Το πλήθος κειμένων που είχαμε ανα κατηγορία δεν ήταν αρκετά μεγάλο ώστε το μοντέλο να αποτυπώσει στο έπακρο τις αναπαραστάσεις των δεδομένων. Στην προσπάθεια να

μετριάσουμε το αντίκτυπο της πρόκλησης, υλοποιήσαμε τεχνικές επαύξησης δεδομένων ώστε να βελτιώσουμε την απόδοση του.

Φυσικά πρέπει να τονιστεί πως τα αποτελέσματα που επιτύχαμε δεν είναι τα βέλτιστα που μπορούν να επιτευχθούν για το συγκεκριμένο σύνολο δεδομένων και θεωρούμε πως με την υλοποίηση στρατηγικών βημάτων στο μέλλον (μεγαλύτερη ποσότητα δεδομένων, περαιτέρω βελτιστοποίηση παραμέτρων μοντέλου) μπορούν να βελτιωθούν ραγδαία.

Αντλώντας πλέον απο την εμπειρία που αποκομήσαμε μέσω της έρευνας μας, θεωρούμε πως τα μοντέλα που παρουσιάσαμε και το σύνολο δεδομένων που προτείναμε παρουσιάζουν μεγάλο ενδιαφέρον και δίνουν περιθώριο σε μελλοντικές έρευνες οι οποίες θα λάβουν υπόψιν τις δυσκολίες που αντιμετωπίσαμε ώστε επιτύχουν ακόμη καλύτερα αποτελέσματα. Εκτός απο την ταξινόμηση πολλαπλών κατηγοριών, το σύνολο δεδομένων αποτελεί μια πολύ καλή βάση και για την εξερεύνηση άλλων τύπων ταξινόμησης στο εύρος κατηγοριών που προσφέρει.

Τέλος, όπως αναλύσαμε στην διπλωματική μας, η ταξινόμηση ρητορικής μίσους αποτελεί ένα σημαντικό κλάδο της τεχνητής νοημοσύνης με αντίκτυπο σε πολλές εκφάνσεις της καθημερινής μας ζωής. Η περαιτέρω έρευνα και δημιουργίας μοντέλων ικανών να πετυχαίνουν υψηλά αποτελέσματα δημιουργόντας αντικειμενικές αναπαραστάσεις αποτελεί την πρόκληση της επιστημονικής κοινότητας στο άμμεσο μέλλον.

6.2 Κώδικας Υλοποίησης και Πληροφορίες

Ο κώδικας υλοποίησης της διπλωματικής μας εργασίας με πλήρη γραπτή τεκμηρίωση, μπορεί να βρεθεί στην εξής διεύθυνση ιστού:

- Github

Ελπίζουμε να βοηθήσει τους αναγνώστες να καταλάβουν σε βάθος την λειτουργία του μοντέλου μας αλλά και ακαδημαϊκούς ερευνητές του κλάδου της τεχνητής νοημοσύνης να προωθήσουν τον τομέα και τις επιδόσεις του.

Βιβλιογραφία

[1]Auxier, Brooke, and Monica Anderson. "Social media use in 2021."Pew Research Center (2021).

[2]Perrin, A. "Social Media Fact Sheet. Pew Research Center. 2018." (2018)

[3]Perrin, Andrew. "Social media usage."Pew research center 125 (2015): 52-68.

[4]Mathew, Binny, et al. "Spread of hate speech in online social media."Proceedings of the 10th ACM conference on web science. 2019.

[5]Mondal, Mainack, Leandro Araújo Silva, and Fabrício Benevenuto. "A measurement study of hate speech in social media." Proceedings of the 28th ACM conference on hypertext and social media. 2017.

[6]Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.

[7]Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

[8]Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[9]Liu, Shikun, Edward Johns, and Andrew J. Davison. "End-to-end multi-task learning with attention." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[10] Yin, Wenjie, and Arkaitz Zubiaga. "Towards generalisable hate speech detection: a review on obstacles and solutions." PeerJ Computer Science 7 (2021): e598.

[11] Davidson, Thomas, et al. "Automated hate speech detection and the problem of offensive language." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 11. No. 1. 2017.

[12] ElSherief, Mai, et al. "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech."arXiv preprint arXiv:2109.05322 (2021).

[13] Arango, Aymé, Jorge Pérez, and Barbara Poblete. "Hate speech detection is not as easy as you may think: A closer look at model validation (extended version)." Information Systems (2020): 101584.

[14] Prabhu, Ameya, Charles Dognin, and Maneesh Singh. "Sampling bias in deep active classification: An empirical study."arXiv preprint arXiv:1909.09389 (2019).

[15] Davidson, Thomas, Debasmita Bhattacharya, and Ingmar Weber. "Racial bias in hate speech and abusive language detection datasets." arXiv preprint arXiv:1905.12516 (2019).

[16] Poletto, Fabio, et al. "Resources and benchmark corpora for hate speech detection: a systematic review." Language Resources and Evaluation 55.2 (2021): 477-523.

[17] Ousidhoum, Nedjma, et al. "Multilingual and multi-aspect hate speech analysis." arXiv preprint arXiv:1908.11049 (2019).

[18] Vidgen, Bertie, et al. "Challenges and frontiers in abusive content detection." Proceedings of the third workshop on abusive language online. 2019.

[19] Turing, Alan M. "Computing machinery and intelligence." Parsing the turing test. Springer, Dordrecht, 2009. 23-65.

[20] McCarthy, John. "What is artificial intelligence?." (2007).

[21] Rosenblatt, Frank. "The perceptron: a probabilistic model for information storage and organization in the brain." Psychological review 65.6 (1958): 386.

[22] Block, H. D. "A review of “perceptrons: An introduction to computational geometry."Information and Control 17.5 (1970): 501-522.

[23] LeCun, Yann, et al. "Object recognition with gradient-based learning."Shape, contour and grouping in computer vision. Springer, Berlin, Heidelberg, 1999. 319-345.

[24] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[25] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory. "Neural computation 9.8 (1997): 1735-1780.

[26] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[27] Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text."Proceedings of the International AAAI Conference on Web and Social Media. Vol. 8. No. 1. 2014.

[28] Gitari, Njagi Dennis, et al. "A lexicon-based approach for hate speech detection."International Journal of Multimedia and Ubiquitous Engineering10.4 (2015): 215-230.

[29] Fatahillah, Naufal Riza, Pulut Suryati, and Cosmas Haryawan. "Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech."2017 International Conference on Sustainable Information Engineering and Technology (SIET). IEEE, 2017.

[30] Fauzi, M. Ali, and Anny Yuniarti. "Ensemble method for indonesian twitter hate speech detection."Indonesian Journal of Electrical Engineering and Computer Science11.1 (2018): 294-299.

[31] Ginting, Purnama Sari Br, Budhi Irawan, and Casi Setianingsih. "Hate speech detection on twitter using multinomial logistic regression classification method."2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS). IEEE, 2019.

[32] Oriola, Oluwafemi, and Eduan Kotzé. "Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets."IEEE Access 8 (2020): 21496-21509.

[33] Briliani, Annisa, Budhi Irawan, and Casi Setianingsih. "Hate Speech Detection in Indonesian Language on Instagram Comment Section Using K-Nearest Neighbor Classification Method."2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS). IEEE, 2019.

[34] Zhao, Rui, and Kezhi Mao. "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder."IEEE Transactions on Affective Computing 8.3 (2016): 328-339.

[35] Rodriguez, Axel, Carlos Argueta, and Yi-Ling Chen. "Automatic detection of hate speech on facebook using sentiment and emotion analysis."2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, 2019.

[36] Jaki, Sylvia, and Tom De Smedt. "Right-wing German hate speech on Twitter: Analysis and automatic detection." arXiv preprint arXiv:1910.07518 (2019).

[37] Rosa, Hugo, et al. "A “deeper” look at detecting cyberbullying in social networks."2018 international joint conference on neural networks (IJCNN). IEEE, 2018.

[38] Van Huynh, Tin, et al. "Hate speech detection on vietnamese social media text using the bi-gru-lstm-cnn model."arXiv preprint arXiv:1911.03644 (2019).

[39] Gambäck, Björn, and Utpal Kumar Sikdar. "Using convolutional neural networks to classify hate-speech." Proceedings of the first workshop on abusive language online. 2017.

[40] Zhang, Ziqi, David Robinson, and Jonathan Tepper. "Detecting hate speech on twitter using a convolution-gru based deep neural network." European semantic web conference. Springer, Cham, 2018.

[41] Pamungkas, Endang Wahyu, and Viviana Patti. "Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon."Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop. 2019.

[42] Alsafari, Safa, Samira Sadaoui, and Malek Mouhoub. "Hate and offensive speech detection on arabic social media."Online Social Networks and Media 19 (2020): 100096.

[43] Poletto, Fabio, et al. "Resources and benchmark corpora for hate speech detection: a systematic review."Language Resources and Evaluation 55.2 (2021): 477-523.

[44] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter."Proceedings of the NAACL student research workshop. 2016.