ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Μελέτη χρήσης τεχνικών μηχανικής μάθησης για την αξιολόγηση βιογραφικών σημειωμάτων

Πτυχιακη Εργασια

του

ΨΑΡΟΥΔΑΚΗ ΓΕΩΡΓΙΟΥ

Επιβλέπων: Παναγιώτης Ζέρβας

Πάτρα 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Μελέτη χρήσης τεχνικών μηχανικής μάθησης για την αξιολόγηση βιογραφικών σημειωμάτων

Πτυχιακη Εργασια

του

ΨΑΡΟΥΔΑΚΗ ΓΕΩΡΓΙΟΥ

Επιβλέπων: Παναγιώτης Ζέρβας

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 6 Απριλίου 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Παναγιώτης Ζέρβας	Ιωάννης Τζήμας	Ιωάννης Τσακνάκης
Επίκουρος Καθηγητής	Καθηγητής	Αναπληρωτής Καθηγητής

Πάτρα 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Υπεύθυνη Δήλωση

Βεβαιώνω ότι είμαι συγγραφέας αυτής της πτυχιακής εργασίας, και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην πτυχιακή εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης, βεβαιώνω ότι αυτή η πτυχιακή εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τις απαιτήσεις του προγράμματος σπουδών του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Πελοποννήσου.

(Υπογραφή)

.....
Ψαρουδάκης Γεώργιος

Περίληψη

Η εργασία αυτή, έχει σαν σκοπό την θεωρητική μελέτη συστημάτων ταξινόμησης κειμένου μέσω μηχανικής μάθησης,την ανάπτυξη δικού μας συνόλου δεδομένων βιογραφικών από διαδικτυακές πλατφόρμες εύρεσης εργασίας, καθώς και υλοποίηση ενός συστήματος για αναγνώριση και κατηγοριοποίηση κειμένων με την χρήση της βάσης που δημιουργήθηκε.

Στο πρώτο κεφάλαιο δίνονται τα κίνητρα και ο σκοπός της μελέτης, καθώς και η περιγραφή των προσεγγίσεων για την λύση του προβλήματος της ταξινόμησης βιογραφικών στα πλαίσια μιας διαδικασίας αξιολόγησης.

Στο δεύτερο κεφάλαιο, παρουσιάζονται θεωρητικά στοιχεία και γίνεται βιβλιογραφική ανασκόπηση της έρευνας έρευνα γύρω απο τα συστήματα αξιολόγησης βιογραφικών, τα βιογραφικά σημειώματα, τεχνικές για την προεπεξεργασία, την επαύξηση και την εξαγωγή χαρακτηριστικων, καθώς και τη διαδικασία της μηχανικής μάθησης. Η θεματική ενότητα του τρίτου κεφαλαίου αφορά την ανάλυση της μεθοδολογίας δημιουργίας της βάσης δεδομένων με προεπεξεργασία και επαύξηση, όπως και την παρουσίαση των προσεγγίσεων εξαγωγής χαρακτηριστικων, των αλγόριθμων ταξινόμησης και την αξιολόγηση των αποτελεσμάτων.

Ακολούθως, στο τέταρτο κεφάλαιο γίνεται περιγραφή των αποτελεσμάτων της υλοποίησης και ανάπτυξης του συστήματός μας. Στο πεμπτο και τελευταίο, επισημαίνονται οι τελικές παρατηρήσεις και οι μελλοντικές κατευθύνσεις της έρευνας.

Λέξεις Κλειδιά

Ταξινόμηση κειμένου, επεξεργασια φυσικής γλώσσας, BERT, εξαγωγή χαρακτηριστικών, SVM, feature selection

στους γονείς μου

Ευχαριστίες

Αυτή η εργασία, δεν θα είχε ολοκληρωθεί, εάν δεν ήταν ο υπευθύνος καθηγητής μου, Παναγιώτης Ζέρβας, τον οποίο ευχαριστώ πολύ για όλη την καθοδήγηση που μου έδωσε κατά τη διάρκεια της έρευνας, ώστε να κατανοήσω ένα θέμα, το οποίο ήταν εντελώς άγνωστο σε εμένα. Επίσης θα ήθελα να ευχαριστήσω την Ανδριάνα, τον Γιάννη, τον Παναγιώτη και την Ελένη, για την βοήθεια με την εργασία και την στήριξή τους. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, που με στήριξαν όλα αυτά τα χρόνια.

Περιεχόμενα

Περίληψη **i**

Ευχαριστίες **v**

Περιεχόμενα **viii**

Κατάλογος Σχημάτων **x**

Κατάλογος Πινάκων **xi**

1 Εισαγωγή **1**

1.1 Ιστορικό και κίνητρα 1
1.2 Δήλωση του προβλήματος 2
1.3 Στόχοι της διατριβής 2
1.4 Οργάνωση της διατριβής 3

2 Βιβλιογραφική ανασκόπηση **5**

2.1 Επισκόπηση της ταξινόμησης βιογραφικών σημειωμάτων 5
2.1.1 Βιογραφικά Σημειώματα 5
2.1.2 Διαχείρηση Βιογραφικών Σημειωμάτων 6
2.2 Προηγούμενες εργασίες για την αξιολόγηση βιογραφικών σημειωμάτων με μηχανική μάθηση 8
2.3 Τεχνικές προεπεξεργασίας, επαύξησης και εξαγωγής χαρακτηριστικών κειμένου 10
2.3.1 Προεπεξεργασία κειμένου 10
2.3.2 Εξαγωγή χαρακτηριστικών κειμένου 11
2.3.3 Επαύξηση Κειμένου 12
2.4 Ταξινόμηση με αλγόριθμους μηχανικής μάθησης 13
2.4.1 Ταξινόμηση 15
2.4.2 Ταξινόμηση κειμένου 15
2.4.3 Αξιολόγηση Αλγορίθμων 17
2.4.4 Διασταυρωμένη Επικύρωση (Cross-Validation) 18

3 Μεθοδολογία **21**
3.1 Συλλογή Δεδομένων . . . 21
3.1.1 Μορφή και χαρακτηριστικά Βιογραφικών Σημειωμάτων . . . 21
3.2 Συλλογή δεδομένων από ιστότοπους βιογραφικών . . . 22
3.2.1 Πηγές αναζήτησης βιογραφικών . . . 24
3.2.2 Προεπεξεργασία κειμένων και κατασκευή βάσεων δεδομένων . . . 27
3.2.3 Επαύξηση Εγγράφων Βιογραφικών Σημειωμάτων . . . 28
3.3 Προσεγγίσεις εξαγωγής χαρακτηριστικών . . . 29
3.3.1 Αναπαραστάσεις με αραιά διανύσματα . . . 30
3.3.2 Πίνακες Συνύπαρξης . . . 31
3.3.3 TF-IDF . . . 33
3.3.4 Πυκνές Αναπαραστάσεις - Ενσωματώσεις . . . 34
3.3.5 Transformers και δυναμικές ενσωματώσεις . . . 38
3.4 Αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται για εκπαίδευση . . . 44
3.4.1 Naive Bayes . . . 45
3.4.2 Λογιστική Παλινδρόμηση . . . 46
3.4.3 Μηχανές Διανυσμάτων Υποστήριξης . . . 49
3.4.4 BERT . . . 51
3.5 Μετρικές αξιολόγησης . . . 54

4 Πειραματικά αποτελέσματα των εφαρμοζόμενων προσεγγίσεων **57**
4.1 Λογισμικό . . . 57
4.2 Επισκόπηση του συνόλου δεδομένων που χρησιμοποιήθηκε . . . 58
4.3 Πειραματικά αποτελέσματα . . . 61
4.3.1 Διαδικασία Ταξινόμησης . . . 61
4.3.2 Ταξινόμηση - Naive Bayes . . . 61
4.3.3 Ταξινόμηση - Λογιστική Παλινδρόμηση . . . 62
4.3.4 Ταξινόμηση - Μηχανές Διανυσμάτων Υποστήριξης . . . 64
4.3.5 Ταξινόμηση - BERT . . . 65
4.4 Συγκριτική Ανάλυση . . . 67
4.5 Συζήτηση των πλεονεκτημάτων και των περιορισμών του μοντέλου . . . 69

5 Συμπεράσματα **71**
5.1 Σύνοψη των ευρημάτων . . . 71
5.2 Συμβολή της διατριβής . . . 71
5.3 Περιορισμοί και μελλοντικές κατευθύνσεις της έρευνας . . . 72

Κατάλογος Σχημάτων

1.1 Ροή Διαδικασίας 3

2.1 Συστήματα Αξιολόγησης Βιογραφικών 7
2.2 Αξιολόγηση Μοντέλου 18
2.3 k-fold cross validation 19

3.1 Δημιουργία Συνόλων Δεδομένων 23
3.2 Ιστοσυγκομιδή: Indeed.com 27
3.3 Δομή Παραρτήματος Εμπειρίας 28
3.4 Επαύξηση πρότασης 29
3.5 Επαύξηση πρότασης 29
3.6 Διαδικασία Επαύξησης Εγγράφων 30
3.7 Υπολογισμος Word2vec 35
3.8 Skip-gram 35
3.9 Εκπαίδευση του μοντέλου PV-DM. Τα διανύσματα των γειτονικών λέξεων και της παραγράφου συμβάλλουν στην πρόβλεψη της λέξης στόχου. 37
3.10 Εκπαίδευση του μοντέλου PV-DBOW. 37
3.11 Ένα transformer block 39
3.12 Επίπεδο Αυτοπροσοχής 40
3.13 Πίνακες Προσοχής 40
3.14 Scaled dot-product attention 41
3.15 Multi-head attention 42
3.16 Ενσωμάτωση θέσης με τριγωνομετρικές συναρτήσεις 43
3.17 Συναρτήσεις ενσωμάτωσης θέσης 43
3.18 Σιγμοειδής συνάρτηση 48
3.19 Αρχιτεκτονική του BERT, πηγή *DLVU* 52
3.20 Masked Language Modeling 53
3.21 Πίνακας Σύγχυσης 54
3.22 Πολυωνυμικός Πίνακας Σύγχυσης 55

4.1 Ιστόγραμμα δεδομένων ελέγχου 59
4.2 WordCloud βιογραφικών 60
4.3 Naive Bayes - CountVectorizer 62

4.4 Naive Bayes - TF-IDF 62
4.5 Logistic Regression - CountVectorizer 63
4.6 Logistic Regression - TF-IDF 63
4.7 Logistic Regression - Doc2Vec 63
4.8 SVM - CountVectorizer 64
4.9 SVM - TF-IDF 64
4.10 SVM - Doc2Vec 65
4.11 Νευρωνικό δίκτυο BERT με TensorFlow 66
4.12 BERT 66

Κατάλογος Πινάκων

3.1 Πίνακας εγγράφων-λέξεων 32
3.2 Πίνακας Συνύπαρξης λέξεων-λέξεων 32

4.1 Αρχικό σύνολο εκπαίδευσης 59
4.2 Επαυξημένο σύνολο εκπαίδευσης 59
4.3 Σύγκριση αλγορίθμων - Αρχικό Σύνολο 67
4.4 Σύγκριση αλγορίθμων - Επαυξημένο Σύνολο 68

Κεφάλαιο 1

Εισαγωγή

Η παρούσα πτυχιακή εργασία επικεντρώνεται στη χρήση της μηχανικής μάθησης για την κατηγοριοποίηση βιογραφικών σημειωμάτων, σημαντική διαδικασία στη διαχείριση των αιτήσεων εργασίας σε επιχειρήσεις και οργανισμούς. Γίνεται μελέτη και ανάπτυξη ενός αυτοματοποιημένου συστήματος κατηγοριοποίησης βιογραφικών που μπορεί να βοηθήσει στη μείωση του χρόνου και του κόστους που απαιτείται για την επεξεργασία των αιτήσεων και την επιλογή των υποψηφίων.

1.1 Ιστορικό και κίνητρα

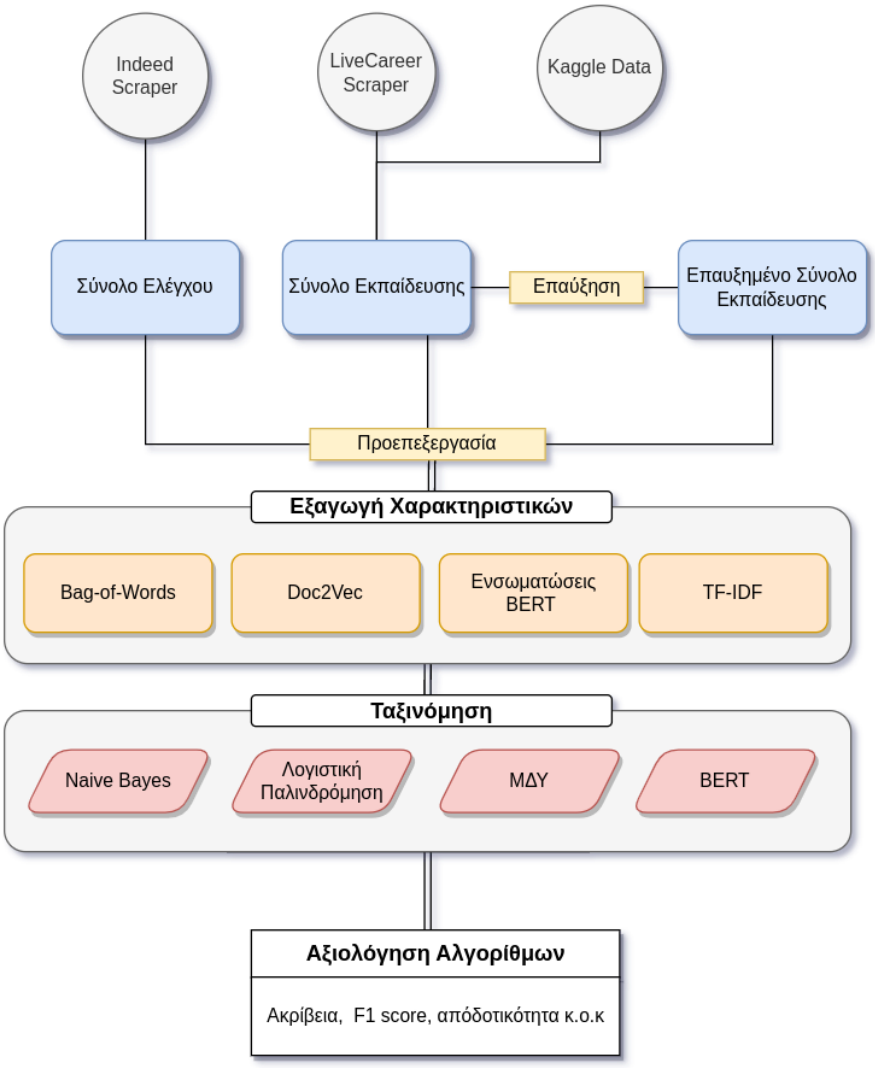
Με την αυξανόμενη παρουσία των πληροφοριακών συστημάτων και του παγκόσμιου ιστού, εταιρείες και οργανισμοί στρέφονται στο διαδίκτυο για να βρουν κατάλληλα άτομα για να καλύψουν ένα ευρύ φάσμα θέσεων. Πλατφόρμες με σκοπό να φέρουν κοντά εργοδότες και άτομα που αναζητούν εργασία εμφανίστηκαν ήδη από τις αρχές του διαδικτύου, και πλέον είναι η καθιερωμένη, κύρια δίοδος στην αγορά εργασίας για εκατομμύρια εργαζομένους. Η διαδικτυακή αγορά εργασίας ήταν και παραμένει ένας ταχέως αναπτυσσόμενος τομέας του παγκόσμιου τοπίου απασχόλησης, με ιδιαίτερη άνθιση σε αναπτυσσόμενες χώρες με μεγάλους πληθυσμούς [1]. Σημαντικό πλεονέκτημα αυτής της διαδικτυακής συγκέντρωσης της αγοράς εργασίας είναι ότι επιτρέπει στους εργοδότες να έχουν ευρύτερο δίκτυο κατά την αναζήτηση υποψηφίων, καθώς μπορούν να δημοσιεύουν λίστες θέσεων εργασίας σε διάφορους ιστότοπους αναζήτησης εργασίας ή και πλατφόρμες κοινωνικής δικτύωσης. Αντιστοίχως, διευκολύνει τα άτομα που αναζητούν εργασία να βρουν περισσότερες επαγγελματικές ευκαιρίες και να υποβάλουν αίτηση σε αυτές. Η διαδικτυακή αγορά εργασίας προσφέρει επίσης μεγάλη ευελιξία τόσο για τους εργοδότες όσο και για τους εργαζόμενους. Οι εργοδότες μπορούν να προσλάβουν απομακρυσμένους εργαζομένους που μπορούν να εργαστούν από οπουδήποτε στον κόσμο, δίνοντας τους πρόσβαση σε ένα μεγαλύτερο δίκτυο πιθανών υποψηφίων. Η ανάπτυξη της μηχανικής μάθησης και της επεξεργασίας φυσικής γλώσσας τα τελευταία χρόνια είχε ως αποτέλεσμα τη δημιουργία εργαλείων που βρίσκουν πρακτική εφαρμογή στην ανάγνωση και ανάλυση κειμένων. Η τεχνολογία αυτή μπορεί να χρησιμοποιηθεί για την αυτοματοποίηση της διαδικασίας κατηγοριοποίησης βιογραφικών.

1.2 Δήλωση του προβλήματος

Η ανάγνωση μεγάλου όγκου εισερχόμενων αιτήσεων εργασίας σε μορφή κειμένων βιογραφικών είναι μια χρονοβόρα και επαναλαμβανόμενη διαδικασία, καθώς πρέπει να αναγνωστούν όλα προσεκτικά χωρίς να παραληφθεί κάποιος πιθανός υποψήφιος. Αυτή η διαδικασία θα απασχολήσει πολύτιμο χρόνο των υπαλλήλων και απαιτεί πόρους που θα μπορούσαν να χρησιμοποιηθούν παραγωγικά, καθυστερώντας η εμποδίζοντας τα υπάρχοντα έργα μιας ομάδας. Η πρόοδος του τομέα της επεξεργασίας φυσικής γλώσσας έχει οδηγήσει στην ανάπτυξη αυτοματοποιημένων συστημάτων ανάλυσης κειμένων που μπορούν να αντικαταστήσουν την ανθρώπινη παρέμβαση επισπεύδοντας και απλοποιώντας τη διαδικασία επιλογής των πιο αξιοπρόσεκτων περιπτώσεων.

1.3 Στόχοι της διατριβής

Στόχος της παρούσας διατριβής είναι η μελέτη και υλοποίηση μοντέλων μηχανικής μάθησης για την αυτόματη ταξινόμηση εγγράφων βιογραφικών. Αρχικά πραγματοποιείται μη εποπτευόμενη επισήμανση και αποκόμιση συνόλων κειμένων για σκοπούς εκπαίδευσης και ελέγχου. Με βάση ενός εκ των συνόλων γίνεται επαύξηση για τη δημιουργία ενός πολυπληθέστερου μοντέλου δειγμάτων, ενώ το άλλο χρησιμοποιείται αποκλειστικά ως σύνολο ελέγχου. Μετά από μια προεπεξεργασία, γίνεται επιλογή χαρακτηριστικών και κωδικοποίηση με μεθόδους πυκνών και αραιών ενσωματώσεων. Τέλος, λαμβάνονται ως είσοδοι σε διάφορους αλγόριθμους ταξινόμησης, κλασσικούς και νεότερους, και παρουσιάζεται η συγκριτική ανάλυση των αποτελεσμάτων, τόσο για τα διαφορετικά σύνολα εκπαίδευσης όσο και για τους μετασχηματισμούς των δεδομένων.



Σχήμα 1.1: Ροή Διαδικασίας

1.4 Οργάνωση της διατριβής

- Στο κεφάλαιο 2: Βιβλιογραφική ανασκόπηση γίνεται βιβλιογραφική ανασκόπηση του προβλήματος της αξιολόγησης βιογραφικών με μηχανική μάθηση. Επίσης γίνεται αναφορά στις μεθόδους προεπεξεργασίας κειμένου φυσικής γλώσσας, καθώς και σε τεχνικές επαύξησης και εξαγωγής χαρακτηριστικών. Τέλος, αναλύονται έννοιες της διαδικασίας εκπαίδευσης αλγορίθμων μηχανικής μάθησης.

- Στο κεφάλαιο 3: Μεθοδολογία αναλύονται οι τεχνικές που θα χρησιμοποιηθούν στην υλοποίηση της μελέτης. Περιλαμβάνονται η διαδικασία συλλογής και επαύξησης των κειμένων, διαφορετικοί τρόποι εξαγωγής χαρακτηριστικών με αραιές και πυκνές κωδικοποιήσεις, καθώς και ανάλυση της λειτουργίας των αλγορίθμων που θα αξιολογηθούν, όπως επίσης και οι μετρικές αξιολόγησης.

- Στο κεφάλαιο 4: Αποτελέσματα παρατίθενται και σχολιάζονται τα αποτελέσματα των πειραμάτων, με συγκριτική ανάλυση των μοντέλων σχετικά με τα διαφορετικά σύνολα δεδομένων, το απλό και το επαυξημένο, στο σύνολο εκπαίδευσης.
- Στο κεφάλαιο 5: Συμπεράσματα σχολιάζονται τα ευρήματα της εργασίας και γίνεται λόγος για πιθανές προοπτικές για μελλοντική έρευνα πάνω στο αντικείμενο.

Κεφάλαιο 2

Βιβλιογραφική ανασκόπηση

2.1 Επισκόπηση της ταξινόμησης βιογραφικών σημειωμάτων

2.1.1 Βιογραφικά Σημειώματα

Το βιογραφικό σημείωμα είναι ένα έγγραφο που δημιουργεί και χρησιμοποιεί ένα άτομο για να αναδείξει το μορφωτικό επίπεδο, την επαγγελματική εμπειρία και τα επιτεύγματά του. Αν και υπάρχουν πολλές χρήσεις για τα βιογραφικά, χρησιμοποιούνται συχνότερα για την εύρεση νέων θέσεων εργασίας [2]. Ένα τυπικό βιογραφικό περιέχει μια περίληψη της σχετικής εργασιακής εμπειρίας και εκπαίδευσης. Το βιογραφικό είναι συνήθως ένα από τα πρώτα στοιχεία, μαζί με μια συστατική επιστολή και μερικές φορές μια αίτηση για απασχόληση, την οποία βλέπει ένας πιθανός εργοδότης σχετικά με το άτομο που αναζητά εργασία και χρησιμοποιείται συνήθως για τον έλεγχο των αιτούντων, ακολουθούμενη συχνά από συνέντευξη.

Ιστορικά στοιχεία βιογραφικών

Η πρώτη καταγεγραμμένη χρήση βιογραφικών εντοπίζεται στον 15ο αιώνα, όταν ο Λεονάρντο ντα Βίντσι έγραψε μια επιστολή στον Δούκα του Μιλάνου περιγράφοντας τις δεξιότητες και τα προσόντα του για δουλειά ως μηχανικός [2]. Για τα επόμενα 450 χρόνια, το βιογραφικό συνέχισε να είναι απλώς μια περιγραφή ενός ατόμου, συμπεριλαμβανομένων των ικανοτήτων και της προηγούμενης απασχόλησης. Επίσης, υπήρχαν περιπτώσεις όπου άτομα που αναζητούσαν εργασία στην Ευρώπη άρχισαν να μοιράζουν φυλλάδια για να διαφημίσουν τις δεξιότητες και τα προσόντα τους σε πιθανούς εργοδότες. Αυτά τα έγγραφα αναρτώνταν σε δημόσιους χώρους, όπως εκκλησίες και αγορές.

Τον 19ο αιώνα, η βιομηχανική επανάσταση οδήγησε στην άνοδο των εργοστασίων και στη μαζική παραγωγή. Αυτό δημιούργησε την ανάγκη για μια πιο οργανωμένη προσέγγιση στις προσλήψεις, η οποία οδήγησε στην ανάπτυξη των πρώτων εντύπων αίτησης εργασίας. Στις αρχές του 1900, τα βιογραφικά περιλάμβαναν πληροφορίες όπως το βάρος, το ύψος, την οικογενειακή κατάσταση και τη θρησκεία. Μέχρι το 1950, στις ανεπτυγμένες χώρες, τα βιογραφικά θεωρούνταν απαραίτητα για κάθε εργαζόμενο και άρχισαν να περιλαμβάνουν μεγα-

λύτερη λεπτομέρεια για τα προσόντα, καθώς και πληροφορίες όπως προσωπικά ενδιαφέροντα και χόμπι. Τη δεκαετία του 1970, την αρχή της Ψηφιακής Εποχής, τα βιογραφικά απέκτησαν μια πιο επαγγελματική εμφάνιση όσον αφορά την παρουσίαση και το περιεχόμενο. [7] Τις επόμενες δεκαετίες, το διαδίκτυο γινόταν όλο και περισσότερο ευρέως διαθέσιμο και όσοι αναζητούσαν εργασία άρχισαν να χρησιμοποιούν email για να στέλνουν τα βιογραφικά τους σε πιθανούς εργοδότες. Αναπτύχθηκαν διαδικτυακές πλατφόρμες εύρεσης εργασίας και βάσεις δεδομένων βιογραφικών, οι οποίες διευκόλυναν τους εργοδότες να αναζητήσουν κατάλληλους υποψηφίους.

Στις αρχές του 21ου αιώνα σημειώθηκε περαιτέρω εξέλιξη για τα βιογραφικά στο διαδίκτυο, καθώς τα μέσα κοινωνικής δικτύωσης βοήθησαν τους ανθρώπους να διαδώσουν τα βιογραφικά γρηγορότερα. Το 2003 κυκλοφόρησε το LinkedIn, το οποίο επέτρεπε στους χρήστες να δημοσιεύουν τα βιογραφικά και τις δεξιότητές τους στο Διαδίκτυο [8]. Σήμερα, τα βιογραφικά υποβάλλονται συχνά στο διαδίκτυο και οι εργοδότες χρησιμοποιούν συστήματα παρακολούθησης αιτούντων για να ελέγξουν τα βιογραφικά για συγκεκριμένες λέξεις-κλειδιά και προσόντα. Το σύγχρονο βιογραφικό έχει επίσης εξελιχθεί ώστε να περιλαμβάνει νέες ενότητες, όπως μια περίληψη ή μια αντικειμενική δήλωση, και μπορεί να περιλαμβάνει συνδέσμους προς το διαδικτυακό χαρτοφυλάκιο ενός υποψηφίου ή το προφίλ LinkedIn.

Μορφολογία βιογραφικών σημειωμάτων

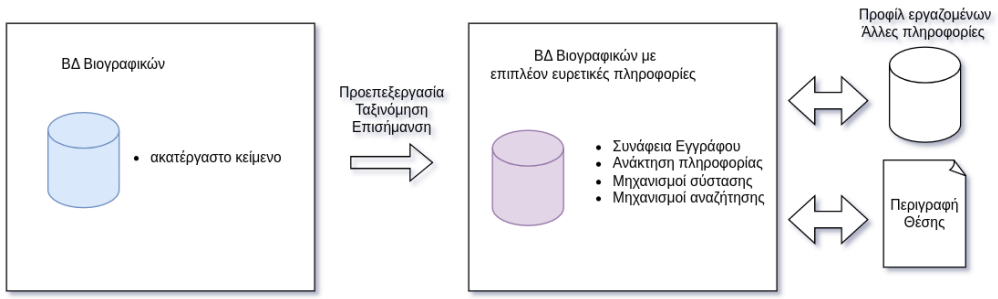
Σε πολλά περιβάλλοντα, ένα βιογραφικό περιορίζεται συνήθως σε μία ή δύο σελίδες μεγέθους A4, επισημαίνοντας μόνο εκείνες τις εμπειρίες και τα προσόντα που ο συγγραφέας θεωρεί πιο σχετικές με την επιθυμητή θέση. Πολλά βιογραφικά περιέχουν λέξεις-κλειδιά ή δεξιότητες που αναζητούν οι πιθανοί εργοδότες μέσω συστημάτων παρακολούθησης αιτούντων, κάνουν έντονη χρήση ενεργητικών ρημάτων και εμφανίζουν περιεχόμενο με κολακευτικό τρόπο. Τα ακρωνύμια και τα διαπιστευτήρια μετά το όνομα του αιτούντος θα πρέπει να γράφονται πλήρως στην κατάλληλη ενότητα του βιογραφικού για να αυξηθεί η πιθανότητα να βρεθούν σε μια ηλεκτρονική αναζήτηση λέξεων-κλειδιών [11]. Το βιογραφικό είναι ένα διαφημιστικό έγγραφο στο οποίο το περιεχόμενο προσαρμόζεται ώστε να ταιριάζει σε κάθε μεμονωμένη αίτηση εργασίας ή αιτήσεις που απευθύνονται σε έναν συγκεκριμένο κλάδο [12]. Η πολυπλοκότητα ή η απλότητα των διαφόρων μορφών βιογραφικών τείνει να παράγει αποτελέσματα που διαφέρουν από άτομο σε άτομο, για το επάγγελμα και τον κλάδο. Τα βιογραφικά ή χαρτοφυλάκια που χρησιμοποιούνται από επαγγελματίες υγείας, καθηγητές, καλλιτέχνες, και άτομα σε άλλους εξειδικευμένους τομείς μπορεί να είναι συγκριτικά μεγαλύτερα. Για παράδειγμα, το βιογραφικό ενός καλλιτέχνη, που συνήθως αποκλείει οποιαδήποτε εργασία που δεν σχετίζεται με την τέχνη, μπορεί να περιλαμβάνει εκτενείς λίστες ατομικών και ομαδικών εκθέσεων [13].

2.1.2 Διαχείριση Βιογραφικών Σημειωμάτων

Η ταξινόμηση των κειμένων βιογραφικών είναι ένα μέρος της ευρύτερης έρευνας γύρω από το πρόβλημα αντιστοίχησης θέσης εργασίας-βιογραφικού. Είναι ένα πεδίο μελέτης που ξεκίνησε

στις αρχές του διαδικτύου και έχει δει ιδιαίτερη άνθιση τα τελευταία χρόνια, συμπίπτοντας με τις τελευταίες εξελίξεις στον τομέα της επεξεργασίας φυσικής γλώσσας. Η περισσότερη νεότερη έρευνα γίνεται από οργανισμούς σε ταχέως αναπτυσσόμενες χώρες με μεγάλες αγορές εργασίας και δυνατότητες ένταξης εκατομμυρίων νέων εργαζομένων στο εργατικό δυναμικό, όπως η Ινδία. Ένα βιογραφικό σημείωμα αποτελείται από ελεύθερο κείμενο και συνεπώς η ταξινόμηση του υπάγεται στο πεδίο της ταξινόμησης ελεύθερου κειμένου, αντικείμενο της επεξεργασίας φυσικής γλώσσας. Σε αντίθεση με κοινή χρήση της φυσικής γλώσσας, στα βιογραφικά σημειώματα η χρήση της είναι σημασιολογικά περιορισμένη, με αποτέλεσμα την επιτυχή χρήση σχετικά απλών μεθόδων σε εργασίες ταξινόμησης και υπολογισμού συνάφειας.

Αρχικά, η έρευνα επικεντρωνόταν σε θέματα σχετικά με την αγορά εργασίας με πεδία μελέτης την απασχόληση, τα πληροφοριακά συστήματα, το e-recruitment, την αντιστοίχηση θέσης και τα βιογραφικά σημειώματα. Παράλληλα, άρχιζαν να εμφανίζονται αναφορές σε ποικιλία θεμάτων τεχνητής νοημοσύνης που έχουν προταθεί για την αντιμετώπιση αυτών των προβλημάτων, όπως μηχανική μάθηση, συστήματα σύστασης (recommender systems), επιστήμη δεδομένων, ανάκτηση δεδομένων και επεξεργασία φυσικής γλώσσας. Πλέον η έρευνα ασχολείται με την αντιστοίχηση θέσης (job matching), την επεξεργασία φυσικής γλώσσας στα πλαίσια των βιογραφικών σημειωμάτων και την μηχανική μάθηση, συμπίπτοντας με τις σημασιολογικές προσεγγίσεις στην πληροφορία κειμένου χάρη στην πρόοδο της ΕΦΓ.



Σχήμα 2.1: Συστήματα Αξιολόγησης Βιογραφικών

Η ταξινόμηση των βιογραφικών, μαζί με άλλες τροποποιήσεις του συνόλου των εγγράφων στη βάση δεδομένων, χρησιμοποιείται στα πλαίσια ενός πληροφοριακού συστήματος διαχείρισης βιογραφικών σημειωμάτων για τον διαχωρισμό και την οργάνωση των ακατέργαστων δεδομένων που έχουν εισαχθεί στη βάση. Μαζί με την ταξινόμηση τα κείμενα περνάνε από επιπλέον επίπεδα επεξεργασίας φυσικής γλώσσας, με σκοπό τελικά να δημιουργηθεί ένα δομημένο, ευρετηριασμένο σύνολο από τα αρχικά δεδομένα, το οποίο μπορεί να βοηθήσει αποτελεσματικά σε επόμενες ενέργειες όπως η χρήση ενός μηχανισμού προτάσεων, ανάκτηση πληροφορίας ή έλεγχος συνάφειας και shortlisting (περιορισμός των αποτελεσμάτων στα πιο αξιόλογα) σχετικά με μια περιγραφή θέσης.

Ο συνήθης τρόπος που ξεκινά αυτή τη διαδικασία είναι όταν έρχεται μια νέα απαίτηση εργασίας (που ονομάζεται REQ). Ο υπεύθυνος προσλήψεων σαρώνει τη βάση δεδομένων του για να δει εάν τα βιογραφικά κάποιου ταιριάζουν με τις απαιτήσεις. Αν όχι, προχωρούν σε αναζήτηση νέων επαγγελματιών.

2.2 Προηγούμενες εργασίες για την αξιολόγηση βιογραφικών σημειωμάτων με μηχανική μάθηση

Μια από τις πρώτες προσεγγίσεις ήταν αυτή των Malinowski et al [3], η οποία προσδιορίζει το πρόβλημα αντιστοίχησης του βιογραφικού εργασίας ως κατάλληλη εργασία για συστήματα συστάσεων, όπου οι δεξιότητες των υποψηφίων και οι απαιτήσεις εργασίας των υπαλλήλων προσλήψεων συνδυάζονται έτσι ώστε να γίνονται συστάσεις βιογραφικού, όπως ακριβώς με τον τρόπο που προτείνονται προϊόντα σε αγοραστές σε πλατφόρμες ηλεκτρονικού εμπορίου.

Στην εργασία των Xing et al [4], προτείνεται μια διαφορετική προσέγγιση, αντιμετωπίζοντας το πρόβλημα της αντιστοίχησης εργασίας-ατόμου χρησιμοποιώντας δομημένα μοντέλα συνάφειας. Αυτά τα μοντέλα εκτελούν ερωτήματα στα εξαγόμενα χαρακτηριστικά των βιογραφικών και των περιγραφών θέσεων εργασίας χρησιμοποιώντας ένα σύνολο σχετικών βιογραφικών, επισημασμένων για να ανακτήσουν παρόμοια από το μεγαλύτερο σύνολο βιογραφικών χωρίς επισημάνσεις.

Οι Karthik et al [5] εισήγαγαν στη διαδικασία ένα βοηθητικό εργαλείο για να βοηθήσει στην επιλογή των σωστών βιογραφικών για θέσεις εργασίας. Το σύστημα επεξεργάζεται πρώτα τις μη δομημένες πληροφορίες που περιέχονται στο βιογραφικό χρησιμοποιώντας έναν αναλυτή πινάκων (table analyser) , ενός κατατμητή (segmenter) και ενός αναγνωριστή εννοιών. Τα παραπάνω μέρη χρησιμοποιούνται για την εξαγωγή ενός συνόλου χαρακτηριστικών που περιλαμβάνουν οπτικά, λεξιλογικά καθώς και άλλα χαρακτηριστικά σχετικά με το κείμενο. Επιπλέον, ένα μοντέλο Conditional Random Fields (CRF) εκτελεί αναγνώριση ονομαστικών οντοτήτων. Με αυτόν τον τρόπο εξάγονται συνολικά 37 χαρακτηριστικά. Στη συνέχεια, το εργαλείο ταξινομεί τους υποψηφίους μέσω ενός μοντέλου βαθμολόγησης που εφαρμόζει TF-IDF (συχνότητα όρου, αντίστροφη συχνότητα εγγράφου) σε ένα ερώτημα (query) που προερχεται από τις απαιτήσεις εργασίας και τα χαρακτηριστικά που εξάγονται από τη συλλογή βιογραφικών. Είναι ένα από τα πρώτα ευφυή συστήματα σχεδιασμένα για την στήριξη σε πράκτορες προσλήψεων να ξεδιαλέξουν τις αιτήσεις εργασίας.

Σε ακόμα μια προσέγγιση βασισμένη στην ανάκτηση πληροφορίας, η εργασία του Guo et al [6] περιγράφει ένα σύστημα αντιστοίχισης εργασίας-βιογραφικού, το οποίο αναλύει μη δομημένα βιογραφικά και περιγραφές εργασιών για εξαγωγή και αντιστοίχιση επισημάνσεων σε διακριτικά χρησιμοποιώντας κανονικές εκφράσεις και τεχνικές αντιπαραβολής μοτίβων. Στη συνέχεια, τα μοντέλα θέσης εργασίας και βιογραφικού που προκύπτουν συγκρίνονται χρησιμοποιώντας οντολογίες και μέτρα ομοιότητας οντολογιών για συγκεκριμένους τομείς.

Οι Pradeep et al [1] προτείνουν ένα σύστημα βασισμένο στην εξαγωγή χαρακτηριστικών κειμένου με αραιούς πίνακες συνεμφάνισης και αναζύγισης χαρακτηριστικών TF-IDF, μετά από κάποια επίπεδα εξομάλυνσης των αρχικών κειμένων. Βασίζεται σε μια τοπική βάση δεδομένων στην οποία οι υποψήφιοι υποβάλλουν τα βιογραφικά τους μετά από την επιτυχή ολοκλήρωση μιας αυτοματοποιημένης δοκιμασίας, φιλτράροντας έτσι τους πιθανούς υποψήφιους και περιορίζοντας τα δεδομένα της βάσης στα πιο αξιοπρόσεκτα, με βάση την ικανότητα του υποψηφίου να περάσει τη δοκιμασία. Το μοντέλο στοχεύει στην απλοποίηση των εισαγόμενων κειμένων με σταδιακή, σχολαστική αφαίρεση αχρείαστων μερών του κειμένου μέχρι να μείνουν τα πιο

ειδοποιή χαρακτηριστικά, για ευκολότερη και ορθότερη ταξινόμηση. Έπειτα, δημιουργούν ένα μοντέλο Word-Document, εφαρμόζουν TF-IDF και προχωράνε στην αξιολόγηση τους με βάση τη συνάφεια ή ομοιότητα των διανυσμάτων με το αντίστοιχο διάνυσμα μιας δημοσιευμένης θέσης εργασίας. Η μέθοδος ομοιότητας είναι η εγγύτητα των διανυσμάτων κειμένου στο διάνυσμα περιγραφής θέσης σε ένα διανυσματικό χώρο χρησιμοποιώντας τον αλγόριθμο K-Nearest Neighbors.

Μια πιο εκλεπτυσμένη διαδικασία ακολουθούν οι Chirag et al [7] χρησιμοποιώντας τεχνικές ανάκτησης πληροφορίας (feature engineering) για την εξαγωγή μιας δομημένης αναπαράστασης των περιεχομένων ενός βιογραφικού γραμμένου σε φυσική γλώσσα. Εκμεταλλεύεται το γεγονός της φύσης ενός βιογραφικού σημειώματος ως ένα ημι-δομημένο τυποποιημένο έγγραφο, στο οποίο εξυπακούεται η ύπαρξη κάποιων σχετικών πληροφοριών που απαντώνται σε κάθε τέτοιου είδους έγγραφο. Σκοπός είναι η αποδόμηση αυτών των χαρακτηριστικών από το ελεύθερο κείμενο και η δημιουργία μιας συμπυκνωμένης αναπαράστασης. Μετά το βασικό βήμα του καθαρισμού, το πρώτο βήμα της προσέγγισης είναι αυτό της επιλογής και εξαγωγής χαρακτηριστικών (Feature Engineering). Χρησιμοποιούνται μεθόδοι βασισμένοι σε κανόνες για μια διεργασία αναγνώρισης οντοτήτων (Named Entity Recognition). Τα χαρακτηριστικά της δομημένης αναπαράστασης που μοντελοποιούν είναι: όνομα, email, τηλέφωνο, μια λίστα με ικανότητες, σχολή, πτυχίο και όνομα θέσης. Αυτό το στάδιο αναφέρεται στη βιβλιογραφία ως σύνοψη (summarization), ορισμός διαφορετικός από τη συνήθη ερμηνεία του όρου στην επεξεργασία φυσικής γλώσσας. Μετά την απόκτηση των παραπάνω πληροφοριών, τα κείμενα αποθηκεύονται σε αυτή τη μορφή σε δομημένα αρχεία JSON. Το επόμενο στάδιο της επεξεργασίας εμπεριέχει την μετατροπή των συμπυκνωμένων δεδομένων σε διανυσματικές μορφές με πίνακες TF-IDF, αυτή τη φορά με μικρότερης διάστασης σε σχέση με την πρώτη προσέγγιση εφ΄όσον τα κείμενα περιορίζονται μονάχα στους στοχευμένους όρους ενδιαφέροντος (θα μπορούσαμε να πούμε ότι η προσέγγιση αφορά ένα συγκεκριμένο ¨καλούπι' για τα κείμενα). Έπειτα, παράλληλα με την επανάληψη της διαδικασίας για ένα έγγραφο περιγραφής θέσης (JD), συντάσσεται μια κατάταξη των συναφέστερων εγγράφων της ΒΔ με χρήση ομοιότητας συνημιτόνου. Παρότι παραπάνω προσεγγίσεις στην ταξινόμηση των είναι υπολογιστικά αποδοτικές, υστερούν στην ισχύ αντίληψης του περιεχομένου του κειμένου διότι χρησιμοποιούν ως πλαίσιο της ανάκτησης πληροφορίας μόνο τα περιεχόμενα της βάσης δεδομένων, η οποία είναι εν γένει περιορισμένη στα λίγα δεδομένα που καταφέρνουν να φτάσουν. Τα μοντέλα αυτά δεν έχουν κατανόηση της γλώσσας και αδυνατούν να συλλάβουν νοηματικές σχέσεις με βάση τα συμφραζόμενα.

2.3 Τεχνικές προεπεξεργασίας, επαύξησης και εξαγωγής χαρακτηριστικών κειμένου

2.3.1 Προεπεξεργασία κειμένου

Κατάτμηση

Τυπικά, το πρώτο στάδιο επεξεργασίας είναι η διάσπαση του κειμένου σε μεμονωμένες λεκτικές μονάδες, οι οποίες θα χρησιμοποιηθούν ως τα βασικά στοιχεία της ανάλυσης. Ο τρόπος διάσπασης ποικίλλει ανάλογα με τον αλγόριθμο, και μπορεί να είναι σε επίπεδο λέξης, σε επίπεδο ομάδων ν λέξεων (νιάδων) ή και σε επίπεδο χαρακτήρων, ειδικά για έγγραφα γραμμένα σε γλώσσες που δεν χρησιμοποιούν κενά μεταξύ των λέξεων, ή χρησιμοποιούν κενά μέσα στις λέξεις.

Εξομάλυνση κειμένου

Επόμενο επίπεδο εξομάλυνσης των ελεύθερων κειμένων είναι η αντικατάσταση ή αφαίρεση αριθμών, σημείων στίξης και άλλων μη αλφαβητικών χαρακτήρων, έτσι ώστε να μείνουν μόνο οι λέξεις αυτές καθ΄ αυτές στο κείμενο. Συχνός στόχος αντικατάστασης αποτελούν επίσης συμβολοσειρές από αλφαβητικούς όρους που δεν είναι αποσπάσματα λόγου, όπως URL ιστοτόπων, διευθύνσεις ηλεκτρονικού ταχυδρομείου ή λέξεις με hashtags. Οι αντικαταστάσεις αυτές γίνονται με τη χρήση κανονικών εκφράσεων. Μια κανονική έκφραση (Regular Expression) είναι μια ακολουθία χαρακτήρων που καθορίζει ένα μοτίβο στο κείμενο αντιστοίχησης, και τέτοια μοτίβα χρησιμοποιούνται από αλγόριθμους αναζήτησης και αντικατάστασης συμβολοσειρών. Μια ακόμα ακόμα παρέμβαση είναι η μετατροπή όλου του κειμένου σε πεζά για την συνέπεια ανάλυσης των παραλλαγών των ίδιων λέξεων.

Αποκοπή καταλήξεων και λημματοποίηση

Ένα πρόβλημα των μοντέλων συνύπαρξης είναι πολυμορφία του ίδιου όρου σε ένα κείμενο. Η κλίση της ίδιας λέξης θεωρείται διαφορετικό χαρακτηριστικό και αυτό αποτελεί πρόβλημα για ένα μοντέλο που βασίζεται στη εμφάνιση των ειδοποιών λέξεων για να ερμηνεύσει το νόημα του κειμένου, με αποτέλεσμα πιο περίπλοκα μοντέλα με περισσότερες διαστάσεις (για παράδειγμα η λέξη ΄διατροφή΄ και ΄διατροφής΄ θα ήταν δυο διαφορετικά χαρακτηριστικά του κειμένου, με το κάθε ένα σε δικιά του διαφορετική στήλη, και τα μοντέλα θα έπρεπε να μάθουν διαφορετικές εσωτερικές παραμέτρους για κάθε περίπτωση). Μια απλοϊκή μέθοδος για την απλοποίηση των λέξεων είναι η αποκοπή των γραμματικών τους καταλήξεων (stemming) με βάση κάποιο αλγόριθμο βασισμένο σε κανόνες (stemmer). Πιο εκλεπτυσμένη μέθοδος για την εύρεση της γραμματικής ρίζας είναι η λημματοποίηση, με το μειονέκτημα της υπολογιστικής πολυπλοκότητας. Για την εύρεση του σωστού λήμματος, χρησιμοποιούνται μέθοδοι βασισμένες στην αποσαφήνιση της χρήσης των λέξεων ανάλογα με το γραμματικό και συντακτικό πλαίσιο της πρότασης, κάνοντας χρήση μεθόδων διαδοχικής επισήμανσης ακολουθιών (sequence tagging). Αυτό αποτελεί περίπτωση παρεμφερής της γενικότερης εργασίας επισήμανσης μερών

του λόγου (Part-Of-Speech Tagging).

Η αποκοπή καταλήξεων και η λημματοποίηση μπορούν να μειώσουν τον όγκο και την πολυπλοκότητα των δεδομένων σε γλωσσικά μοντέλα βασισμένα στο ΄σάκο λέξεων΄ ομαδοποιώντας τις διάφορες εκφάνσεις της κάθε λέξης στη γραμματική της ρίζα. Έτσι, το μοντέλο προσπαθεί να συμπεράνει τη σημασία της λέξης και τη φύση του κειμένου σε απλούστερους όρους.

Αφαίρεση συχνών λέξεων

Για την περαιτέρω απλοποίηση, συνήθως γίνεται η αφαίρεση συχνών λέξεων (stopwords), οι οποίες είναι μικρές, συχνά χρησιμοποιούμενες βοηθητικές λέξεις που εμφανίζονται σε κάθε κείμενο φυσικής γλώσσας, χωρίς να προσθέτουν σημασιολογικό περιεχόμενο. Παραδείγματα συχνών λέξεων είναι μόρια ή προθέσεις όπως 'θα, και, έως, κατά'. Στα εργαλεία ΕΦΓ παρέχονται συχνά λίστες με συχνές λέξεις για διάφορες γλώσσες.

2.3.2 Εξαγωγή χαρακτηριστικών κειμένου

Για να μπορέσουν τα δεδομένα να χρησιμοποιηθούν σε κάποιον αλγόριθμο μηχανικής μάθησης πρέπει πρώτα να γίνει εξαγωγή χαρακτηριστικών (feature selection). Τα χαρακτηριστικά για τη μηχανική μάθηση είναι μεμονωμένες μετρήσιμες μονάδες που προκύπτουν από τα δεδομένα, που επιλέγονται έτσι ώστε να είναι ενημερωτικά και καθοριστικά με σκοπό την πιο αποτελεσματική κατηγοριοποίηση. Στην περίπτωση που τα δεδομένα είναι κείμενο, η τεχνική που επιλέγεται για την εξαγωγή χαρακτηριστικών είναι κωδικοποίηση του κειμένου σε διανύσματα. Με την επιλογή χαρακτηριστικών του κειμένου, γίνεται μια προσπάθεια άπόσταξης΄ των πιο καθοριστικών για την ταξινόμηση όρων που εμφανίζονται στο κείμενο, ώστε να μπορεί να διαφοροποιηθεί από τα άλλα έγγραφα.

Η εξαγωγή χαρακτηριστικών από κείμενα γίνεται σε δυο πλαίσια: το πρώτο είναι η εξαγωγή μιας διανυσματικής αναπαράστασης του εγγράφου σε έναν πολυδιάστατο χώρο βασισμένο στο περιεχόμενο του. Οι αναπαραστάσεις αυτές επιτρέπουν τη δημιουργία μοντέλων συσταδοποίησης των εγγράφων, είτε για σκοπούς ταξινόμησης με βάση την απόσταση (με μεθόδους όπως K-κοντινότερων γειτόνων), είτε για τη σύγκριση της συνάφειας τους με άλλα κείμενα, όπως με τις μεθόδους ομοιότητας διανυσμάτων. Οι αναπαραστάσεις αυτές γίνονται συνήθως με βάση την ύπαρξη και τη συχνότητα όρων μέσα στο κείμενο, μετά από κάποια κατάτμηση του κειμένου στα μέρη του. Αυτές οι κατατμήσεις μπορεί να γίνονται στο επίπεδο της λέξης, χωρισμένες από τα κενά μεταξύ τους, ή και σε ν-άδες λέξεων, όπου εκφράσεις δύο, τριών ή και περισσοτέρων λεκτικών μονάδων συντάσσουν ένα όρο στο λεξιλόγιο. Τα διανύσματα που εξάγονται από τέτοια μοντέλα έχουν μέγεθος που ισούται με το λεξιλόγιο, και μπορούν να συσταθούν είτε σε σχέση με την εμφάνιση των όρων σε μια συλλογή διαφορετικών εγγράφων, με διαστάσεις έγγραφα-όροι (πίνακες εγγράφων-όρων) ή σε σχέση με τους ίδιους τους όρους (πίνακες όρων-όρων) με διαστάσεις όροι-όροι. Όταν συγκρίνονται για ομοιότητα δύο διανύσματα τέτοιου τύπου, στην ουσία προκύπτει μια μετρική εμφάνισης των ίδιων όρων σε διαφορετικά έγγραφα, και κατ΄ επέκταση η σημασιολογική τους ομοιότητα. Βασικές μέθοδοι διαχωρισμού τέτοιων διανυσμάτων στο χώρο είναι οι Μηχανές Διανυσμάτων Υποστήριξης

και οι αλγόριθμοι K-κοντινότερων γειτόνων.

Ισχυρότερες αναπαραστάσεις, βασισμένες σε μικρότερα, πυκνά διανύσματα έχουν εφευρεθεί και χρησιμοποιούνται τα τελευταία χρόνια, με τεχνικές εκμάθησης τέτοιων ΄γλωσσικών ενσωματώσεων΄ των λέξεων στο χώρο με βάση το σημασιολογικό τους περιεχόμενο, τοποθετώντας νοηματικά όμοιες λέξεις πιο κοντά μεταξύ τους. Το νόημα της κάθε διάστασης δεν είναι ξεκάθαρο, άλλα διάφορα νοηματικά μοτίβα φαίνεται να εμφανίζονται μεταξύ διανυσμάτων με οντολογικές σχέσεις του τύπου Α είναι για το Β ότι Γ είναι για το Δ΄. Π.χ. το διάνυσμα για τη λέξη ΄κουτάβι΄ έχει ίδια απόσταση με τη λέξη ΄σκύλος΄ με την απόσταση που έχει το ΄πουλάρι΄ από το διάνυσμα για το ΄άλογο΄. Με παρόμοιες μεθόδους έχουν αναπτυχθεί πυκνές διανυσματικές αναπαραστάσεις εγγράφων ανεξαρτήτου μεγέθους. Αυτές οι ενσωματώσεις μπορεί να είναι στατικές, δηλαδή ο κάθε όρος αντιστοιχίζεται σε 1 μοναδικό διάνυσμα, ενώ καινούρια μοντέλα δημιουργούν νέα διανύσματα στην κάθε εμφάνιση της λέξης μέσω μηχανισμών προσοχής.

Το δεύτερο πλαίσιο εξαγωγής χαρακτηριστικών αναφέρεται σαν διεργασία της ανάκτησης πληροφορίας. Ασχολείται με το να εντοπίζει και να επισημαίνει επιπλέον σημασιολογικές πληροφορίες στο ελεύθερο κείμενο. Η αναγνώριση ονομαστικών οντοτήτων, η επισήμανση μερών του λόγου και η εύρεση σημασιολογικών εξαρτήσεων αποτελούν τέτοιες μορφές εξαγωγής των χαρακτηριστικών, και χρησιμοποιούνται για τη δημιουργία μιας δομημένης αναπαράστασης του περιεχομένου του κειμένου.

2.3.3 Επαύξηση Κειμένου

Εμπειρικά, έχει παρατηρηθεί ότι η ποσότητα των δεδομένων εκπαίδευσης έχει μεγαλύτερη σημασία για τις επιδόσεις του μοντέλου από την ποιότητα της μεθόδου εκμάθησης για γλωσσικά μοντέλα ([8]).

Η επαύξηση δεδομένων είναι μια τεχνική που χρησιμοποιείται για την τεχνητή αύξηση του μεγέθους ενός συνόλου δεδομένων εκπαίδευσης δημιουργώντας πρόσθετα δεδομένα από τα υπάρχοντα. Στην περίπτωση δεδομένων εικόνας, αυτό μπορεί να γίνει με την εφαρμογή διαφόρων μετασχηματισμών στα αρχικά δεδομένα, όπως περιστροφή, περικοπή, κλιμάκωση ή προσθήκη θορύβου. Η ιδέα πίσω από την επαύξηση δεδομένων είναι να εκτεθεί το μοντέλο σε μια ευρύτερη ποικιλία δεδομένων κατά τη διάρκεια της εκπαίδευσης, κάτι που μπορεί να βοηθήσει στη βελτίωση της γενίκευσης και στη μείωση της υπερπροσαρμογής. Η επαύξηση είναι ιδιαίτερα χρήσιμη όταν το μέγεθος του συνόλου δεδομένων εκπαίδευσης είναι μικρό, καθώς μπορεί να βοηθήσει στη δημιουργία ενός πιο ποικιλόμορφου συνόλου παραδειγμάτων εκπαίδευσης. Μπορεί επίσης να χρησιμοποιηθεί για τη βελτίωση της απόδοσης των μοντέλων μηχανικής εκμάθησης σε μια συγκεκριμένη εργασία παρέχοντας στο μοντέλο περισσότερα παραδείγματα των τύπων εισόδου που αναμένεται να συναντήσει κατά τη δοκιμή.

Υπάρχουν πολλές διαφορετικές προσεγγίσεις για την επαύξηση δεδομένων και οι συγκεκριμένες τεχνικές που χρησιμοποιούνται θα εξαρτηθούν από τα χαρακτηριστικά των δεδομένων και την εργασία που εκτελείται. Ορισμένες συχνά χρησιμοποιούμενες τεχνικές περιλαμβάνουν την εφαρμογή τυχαίων μετασχηματισμών σε εικόνες, όπως η περιστροφή, η περικοπή και η κλιμάκωση ή η προσθήκη θορύβου σε δεδομένα ήχου ή κειμένου. Η αύξηση δεδομένων μπορεί

να υλοποιηθεί χειροκίνητα ή χρησιμοποιώντας υπάρχουσες βιβλιοθήκες ή εργαλεία.

Προκειμένου να επαυξηθούν τα δεδομένα κειμένου, ορισμένες προσεγγίσεις είναι:

- Επαύξηση σε επίπεδο λέξης, όπου μπορεί κανείς να εισάγει τυπογραφικά λάθη ως πιθανές εισόδους, και χρησιμοποιείται κυρίως σε σύνολο δεδομένων όπου τα δεδομένα είναι άτυπα, όπως κείμενο άμεσων μηνυμάτων (IM) όπου τέτοια λάθη στην πληκτρολόγηση είναι συνηθισμένα. Μερικά παραδείγματα είναι τα μηνύματα SMS, οι πλατφόρμες ανταλλαγής άμεσων μηνυμάτων (Messenger, Viber), αναρτήσεις σε μέσα κοινωνικής δικτύωσης και είσοδοι σε μηχανές αναζήτησης.

- Επαύξηση σε επίπεδο πρότασης, όπου επιχειρούμε να δημιουργήσουμε νέες, συνθετικές προτάσεις με παρόμοια σημασία και ουσία με τις αρχικές, δηλαδή διαφορετικούς τρόπους να γράψουμε το ίδιο πράγμα. Οι προσεγγίσεις περιλαμβάνουν παράφραση προτάσεων, αντικατάσταση μιας λέξης με το συνώνυμό της («χάθηκε» με «εξαφανίστηκε») καθώς και αλλαγή της δομής της πρότασης αντικαθιστώντας τις παθητικές μορφές ρημάτων με ενεργητικούς τύπους. Ένας συνδυασμός των προαναφερθεισών τεχνικών μπορεί να δημιουργήσει περισσότερα παραδείγματα από το αρχικό δείγμα. Ένας άλλος, πιο περίπλοκος τρόπος δημιουργίας μιας πρότασης με το ίδιο νόημα είναι να αποκτηθεί γνώση τομέα της γλώσσας-στόχου και να βασιστεί σε αυτήν η νέα πρόταση. Για παράδειγμα με 'phrasal verbs' και εκφράσεις στα αγγλικά, χρησιμοποιώντας βάσεις δεδομένων παράφρασης όπως το PPDB του UPenn CIS. Έτσι, η συμβολοσειρά 'escaped' μπορεί να αντικατασταθεί με την 'got away¨.

- Ένας άλλος τρόπος είναι η εκ των υστέρων μετάφραση που χρησιμοποιεί εξελιγμένα νευρωνικά μοντέλα για να δημιουργήσει μια ενδιάμεση αναπαράσταση του νοήματος μιας πρότασης χρησιμοποιώντας ένα πλαίσιο κωδικοποιητή (encoding), να την αποκωδικοποιήσει σε άλλη γλώσσα με βάση ένα μοντέλο πιθανότητας πρότασης άλλης γλώσσας (βασικά, ¨πώς θα λεγόταν αυτό στα γαλλικά¨) και μετά να αποκωδικοποιήσει, με την ίδια διαδικασία (decoding), πίσω στην αρχική γλώσσα.

- Γενετική επαύξηση, η οποία προσθέτει νέες προτάσεις χρησιμοποιώντας γενετικά μοντέλα νευρωνικών γλωσσών που έχουν εκπαιδευτεί σε μεγάλα σώματα σχετικού κειμένου, έτσι ώστε να μπορούν να προστεθούν περισσότερες προτάσεις μετά την αρχική.

2.4 Ταξινόμηση με αλγόριθμους μηχανικής μάθησης

Η μηχανική μάθηση είναι μια εφαρμογή της Τεχνητής Νοημοσύνης (AI) που δίνει τη δυνατότητα σε υπολογιστικά συστήματα να μαθαίνουν και να βελτιώνονται χωρίς τον ρητό προγραμματισμό τους. Συγκεκριμένα, η μηχανική μάθηση στοχεύει στην ανάπτυξη προγραμμάτων υπολογιστών στα οποία τους δίνεται πρόσβαση σε δεδομένα και τα χρησιμοποιούν για να βελτιωθούν από μόνα τους. Η διαδικασία της μάθησης ξεκινά με παρατηρήσεις ή δεδομένα, όπως παραδείγματα, μετρήσεις ή εντολές, ώστε το πρόγραμμα να διερευνήσει μοτίβα στα δεδομένα και να λάβει καλύτερες αποφάσεις στο μέλλον με βάση τα παρεχόμενα παραδείγματα.

Ο κεντρικός στόχος είναι να επιτραπεί στα υπολογιστικά συστήματα να μαθαίνουν χωρίς ανθρώπινη παρέμβαση ή βοήθεια και να προσαρμόζονται ανάλογα (expertsystem.com). Ο όρος Μηχανική Μάθηση΄ (Machine Learning) επινοήθηκε το 1959 απο τον Arthur Samuel [9]. Ένας συχνά αναφερόμενος, πιο επίσημος ορισμός των αλγορίθμων που μελετώνται από το πεδίο της μηχανικής μάθησης δίνεται από τον Tom M. Mitchell ως:

Definition: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E (Mitchell, 1997).

Υπάρχουν τέσσερις τύποι αλγορίθμων Μηχανικής Μάθησης:

- Η **εποπτευόμενη ή επιβλεπόμενη μάθηση** αναφέρεται σε οποιαδήποτε διεργασία μηχανικής μάθησης η οποία μαθαίνει μια συνάρτηση από έναν τύπο εισόδου σε έναν τύπο εξόδου χρησιμοποιώντας δεδομένα που αποτελούνται από τιμές εισόδου μαζί με τιμές εξόδου. Δύο κλασσικά παραδείγματα εποπτευόμενης μαθήσεως είναι η ταξινόμηση και η παλινδρόμηση. Σε αυτές τις περιπτώσεις, η τύποι εξόδου είναι κατηγορήματα (κλάσεις) και αριθμητικοί. Η εποπτευόμενη μάθηση αντιτίθεται της μη εποπτευόμενης μάθησης, η οποία προσπαθεί να βρει δομή στα δεδομένα, και στην ενισχυτική μάθηση, όπου διαδοχικές πολιτικές λήψης αποφάσεων μαθαίνονται ως επιβραβεύσεις χωρίς ρητά αποτελέσματα 'ορθής' συμπεριφοράς.

- Η **μη εποπτευόμενη μάθηση** αναφέρεται σε οποιαδήποτε διαδικασία μηχανικής μάθησης που επιδιώκει να ανακαλύψει κάποια δομή στα δεδομένα, απουσία είτε μιας σωστής εξόδου ή κάποιας καθοδήγησης. Τρία συνήθη παραδείγματα μη εποπτευόμενης μάθησης είναι η συσταδοποίηση, οι κανόνες συσχετίσεων και τα δίκτυα αυτοοργανωμένης απεικόνισης.

- Η **ενισχυτική μάθηση** περιγράφει μια μεγάλη κατηγορία από προβλήματα μάθησης σχετικά με την διάδραση αυτόνομων πρακτόρων με το περιβάλλον τους. Διαδοχικά προβλήματα λήψης αποφάσεων με την επιβράβευση να έρχεται αργότερα. Οι αλγόριθμοι ενισχυτικής μάθησης σκοπεύουν να μάθουν μια πολιτική (δηλαδή να αντιστοιχίζουν από καταστάσεις σε ενέργειες) η οποία να μεγιστοποιεί την επιβράβευση σε βάθος χρόνου. Εν αντιθέσει με τα προβλήματα εποπτευόμενης μάθησης, στην ενισχυτική μάθηση δεν υπάρχουν επισημασμένα παραδείγματα ορθής ή λανθασμένης συμπεριφοράς. Ωστόσο, σε αντίθεση με τα προβλήματα μη εποπτευόμενης μάθησης, δύναται αντιληφθεί από τον πράκτορα κάποιο σήμα επιβράβευσης.

- **Ημι-εποπτευόμενη μάθηση**, η οποία χρησιμοποιεί επισημασμένα και μη επισημασμένα δεδομένα για να εκτελέσει διεργασίες εποπτευόμενης ή μη διεργασίας.

(Sammut C., 2010)

2.4.1 Ταξινόμηση

Στη συνήθη του χρήση, ο όρος ΄ταξινόμηση΄ αναφέρεται στην ανάθεση αντικειμένων σε κατηγορίες, δηλαδή σε κάποια χρήσιμη ομαδοποίηση τους. Ανθρωπίνως, το κάνουμε αυτό διότι αντικείμενα σε μια ομάδα, η λεγόμενη ΄κλάση΄ στην μηχανική μάθηση, μοιράζονται κοινά χαρακτηριστικά. Εάν γνωρίζουμε την κλάση του αντικειμένου, γνωρίζουμε πολλά για αυτό. Στη μηχανική μάθηση, ο όρος ΄ταξινόμηση΄ συνήθως σχετίζεται με έναν συγκεκριμένο τύπο εκπαίδευσης όπου παραδείγματα από μια ή περισσότερες κλάσεις, επισημασμένα με το όνομα της κλάσης, παρέχονται στον αλγόριθμο μάθησης. Τα δεδομένα εισόδου μιας διαδικασίας ταξινόμησης είναι ένα σύνολο από εγγραφές. Κάθε εγγραφή, γνωστή και ως παράδειγμα χαρακτηρίζεται από μια πλειάδα (x, y) όπου x είναι τα χαρακτηριστικά και y είναι ένα ειδικό χαρακτηριστικό, που ορίζεται ως η ετικέτα ή το σημάδι της κλάσης, γνωστό και ως χαρακτηριστικό-στόχος. Αυτό το χαρακτηριστικό πρέπει να είναι στοιχείο ενός συνόλου πεπερασμένων κατηγοριών. Επίσης, αυτό το χαρακτηριστικό διαφοροποιεί την ταξινόμηση από την παλινδρόμηση, μια διαδικασία προγνωστικής μοντελοποίησης όπου το g είναι ένας αριθμός (Pang Ning Tan). Επίσης η ταξινόμηση ορίζεται στο βιβλίο ως η διαδικασία εκμάθησης μιας συνάρτησης, η οποία αντιστοιχίζει κάθε σύνολο χαρακτηριστικών x σε μια από κάποιες ορισμένες κλάσεις y [7]. Η συνάρτηση στόχος είναι γνωστή ανεπίσημα και ως μοντέλο ταξινόμησης. Ένα μοντέλο ταξινόμησης χρησιμεύει για τους παρακάτω σκοπούς:

- **Περιγραφική μοντελοποίηση**: Ένα πρότυπο ταξινόμησης το οποίο χρησιμεύει σαν ένα επεξηγηματικό εργαλείο για το διαχωρισμό μεταξύ αντικειμένων διαφορετικών κλάσεων.

- **Προγνωστική μοντελοποίηση**: Ένα πρότυπο ταξινόμησης το οποίο χρησιμεύει για να εκτιμήσει την κλάση αγνώστων παραδειγμάτων. Ένα πρότυπο ταξινόμησης μπορεί να σαν ένα αδιαφανές κουτί το οποίο αναθέτει αυτομάτως μια ετικέτα όταν του παρουσιαστούν τα χαρακτηριστικά ενός άγνωστου δείγματος.

Ένας κανόνας ταξινόμησης είναι ένας συλλογισμός IF-THEN. Η συνθήκη του κανόνα (το σώμα του κανόνα ή προηγούμενο) αποτελείται από έναν συνδυασμό δυαδικών όρων, ο καθένας εκ των οποίων αποτελεί έναν περιορισμό ο οποίος πρέπει να ικανοποιηθεί από κάποιο παράδειγμα. Εάν όλοι οι περιορισμοί ικανοποιηθούν, ο κανόνας ενεργοποιείται και το παράδειγμα θεωρείται ότι καλύπτεται από τον κανόνα. Η κεφαλή του κανόνα (το συμπέρασμα) αποτελείται από μια μοναδική τιμή της κλάσης, η οποία εκτιμάται στην περίπτωση της ενεργοποίησης του κανόνα. Αυτό έρχεται σε αντίθεση με τους κανόνες συσχέτισης, οι οποίοι μπορούν να δεχτούν πολλαπλά χαρακτηριστικά στην κεφαλή. (Sammut C., 2010)

2.4.2 Ταξινόμηση κειμένου

Μια συνήθης διαδικασία στα πλαίσια της επεξεργασίας φυσικής γλώσσας είναι η κατηγοριοποίηση κειμένου, η ανάθεση δηλαδή μιας ή περισσοτέρων κατηγοριών (labels-ετικέτες) σε ένα έγγραφο. Μερικά παραδείγματα εφαρμογής της ταξινόμησης κειμένου παρατίθενται στη συνέχεια.

Εφαρμογές Ταξινόμησης Κειμένου

Η **ανάλυση συναισθήματος** (sentiment analysis) είναι η ταξινόμηση ενός κειμένου με βάση το συναίσθημα (την ουδέτερη, αρνητική ή θετική στάση του συγγραφέα) προς κάποιο αντικείμενο. Κριτικές ταινιών, βιβλίων και προϊόντων ή σχόλια χρηστών σε ιστοσελίδες όπως το YouTube αποτελούν στόχους ανάλυσης συναισθήματος. Ακόμα μια βασική εφαρμογή είναι η εξαγωγή της κοινής γνώμης από δημοσιεύσεις και αναρτήσεις σε μέσα κοινωνικής δικτύωσης όπως το Twitter, με την πληροφορία να είναι χρήσιμη τόσο σε εταιρείες marketing όσο και σε πολιτικούς αναλυτές [6].

Από τις πρώτες εφαρμογές ταξινόμησης κειμένου είναι το **spam detection**, ο διαχωρισμός μηνυμάτων που λαμβάνονται στο ηλεκτρονικό ταχυδρομείο σε ανεπιθύμητα και επιθυμητά (spam και non-spam). Είναι μια από τις πλέον διαδεδομένες και τυπικές διαδικασίες δυαδικής ταξινόμησης και χρησιμοποιείται από υπηρεσίες ηλεκτρονικού ταχυδρομείου για τον εντοπισμό ανεπιθύμητων μηνυμάτων και τη μεταφορά τους σε έναν ξεχωριστό φάκελο.

Η **απόδοση συγγραφής** (authorship attribution) είναι η εφαρμογή της ταξινόμησης κειμένου που χρησιμοποιείται ώστε να οριστεί εάν ένα κείμενο γράφτηκε από κάποιον συγκεκριμένο συγγραφέα. Παράλληλα, βοηθάει στον καθορισμό της κατάστασης ενός κειμένου ως πρωτότυπο ή μη διότι εντοπίζει περιπτώσεις που το κείμενο ενός συγγραφέα εμφανίζεται κάπου αλλού. Λόγω αυτού χρησιμοποιείται στον **εντοπισμό λογοκλοπής**, ο οποίος έχει αξία σε οργανισμούς που ασχολούνται με δημοσιεύσεις, όπως επιστημονικά περιοδικά, εκδοτικές εταιρείες κ.ο.κ.

Η γνώση του τι αφορά ένα κείμενο αποτελεί σημαντικό μέρος της ανεύρεσης πληροφορίας (Information Retrieval). Στην παρούσα εργασία, γίνεται αναφορά στη πράξη της ταξινόμησης κειμένου σχετικά με το *είδος* του περιεχομένου του σε ένα πεπερασμένο σύνολο (set) από κατηγορίες (ή κλάσεις).

Ταξινόμηση Πολλών κλάσεων

Ένα βιογραφικό, στα πλαίσια της αντιστοίχησης του με μια θέση εργασίας, μπορεί να ταξινομηθεί με βάση μια γενική τάση των περιεχομένων του. Αυτό γιατί τα βιογραφικά περιλαμβάνουν διάφορες θέσεις, αρμοδιότητες και εμπειρίες σε διαφορετικά πόστα. Έτσι ένα βιογραφικό μπορεί να αναφέρεται κατά 60% σε ένα επάγγελμα και κατά 40% σε ένα άλλο. Το γεγονός αυτό περιπλέκει την ταξινόμηση, και μετατρέπει το πρόβλημα σε αυτό της ταξινόμησης πολλών κλάσεων, όπου κάθε παράδειγμα μπορεί να λαμβάνει παραπάνω από μια κλάση. Σε τέτοια προβλήματα ταξινόμησης κειμένου συγκαταλέγονται η αναγνώριση γλώσσας και η ανάλυση συναισθήματος.

Υπάρχουν πολλές διαφορετικές προσεγγίσεις που μπορούν να εφαρμοστούν για την αντιμετώπιση προβλημάτων ταξινόμησης πολλαπλών κλάσεων και το ποια προσέγγιση είναι καλύτερη εξαρτάται από τα χαρακτηριστικά των δεδομένων και τις επιθυμητές ιδιότητες του ταξινομητή. Παρακάτω γίνεται αναφορά σε μερικές μεθόδους ταξινόμησης πολλαπλών κλάσεων:

- Μέθοδος ταξινόμησης one-versus-all (OvA): Στην ταξινόμηση OvA, εκπαιδεύεται ένας ξεχωριστός δυαδικός ταξινομητής για κάθε κλάση, με την κατηγορία να ταξινομείται ως

¨θετική' και όλες οι άλλες κατηγορίες να ταξινομούνται ως ¨αρνητικές¨. Κατά τον έλεγχο, ο ταξινομητής που επιτυγχάνει την υψηλότερη βαθμολογία χρησιμοποιείται για την εκτίμηση της ετικέτας της κλάσης. Το OvA είναι μια απλή και αποτελεσματική μέθοδος που μπορεί να λειτουργήσει καλά με ένα ευρύ φάσμα ταξινομητών, αλλά μπορεί να είναι ευαίσθητη σε μη ισορροπημένες κατανομές κλάσεων. Ένα παράδειγμα αλγορίθμου για τέτοιες περιπτώσεις είναι η λογιστική παλινδρόμηση.

- Παρόμοια μέθοδος είναι η one-versus-one (OvO) όπου εκπαιδεύεται ένας δυαδικός ταξινομητής για κάθε ζεύγος κλάσεων, και κατα τον συμπερασμό, η ετικέτα της κλάσης προβλέπεται από τον ταξινομητή που επιτυγχάνει την υψηλότερη βαθμολογία σε όλους τους ταξινομητές. Παρότι το OvO μπορεί να είναι υπολογιστικά απαιτητικότερο, είναι πιο ανθεκτικό σε μη ισορροπημένες κατανομές κλάσεων από το OvA.

- Ορισμένοι ταξινομητές μπορούν να εκπαιδευτούν απευθείας στο πρόβλημα πολλών κλάσεων, χωρίς να χρειάζεται να μετασχηματιστεί το πρόβλημα σε πολλαπλές εργασίες δυαδικής ταξινόμησης. Παραδείγματα τέτοιων ταξινομητών περιλαμβάνουν δέντρα αποφάσεων (decision trees), τυχαία δάση (random forest) και διάφορες αρχιτεκτονικές νευρωνικών δίκτύων (νευρωνικά δίκτυα εμπροσθοδιάδωσης, sequence-to-label αναδρόμικά νευρωνικά δίκτυα, transformers κ.ο.κ). Αυτοί οι ταξινομητές μπορούν να εκπαιδευτούν για τη βελτιστοποίηση μιας αντικειμενικής συνάρτησης που σχετίζεται άμεσα με το πρόβλημα ταξινόμησης πολλών κλάσεων, όπως η απώλεια διασταυρούμενης εντροπίας (cross-entropy loss).

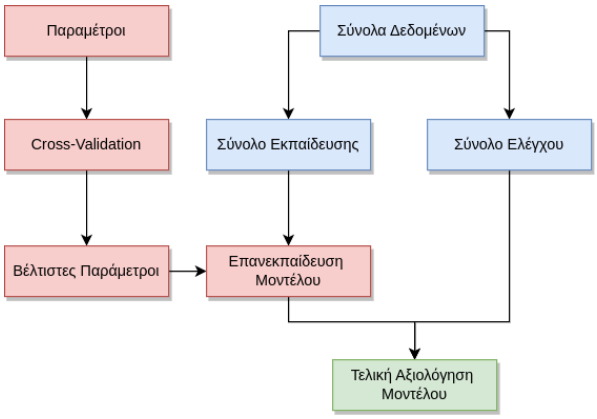
- Τέλος, υπάρχουν πολλές άλλες μέθοδοι που μπορούν να χρησιμοποιηθούν για ταξινόμηση πολλαπλών κλάσεων, όπως μηχανές boosting, bagging και Support Vector Machines. Αυτές οι μέθοδοι μπορεί να είναι αποτελεσματικές, αλλά μπορεί να απαιτούν πιο προσεκτικό συντονισμό του μοντέλου ή/και των χαρακτηριστικών εισόδου προκειμένου να επιτευχθεί καλή απόδοση.

Η αφαίρεση συχνών λέξεων και η λημματοποίηση, σε αντίθεση με τα απλά μοντέλα ταξινόμησης, έχουν αρνητικό αντίκτυπο στην απόδοση προεκπαιδευμένων μοντέλων κατανόησης της γλώσσας, όπως του transformer. Αυτό διότι οι κλίσεις, τα άρθρα και οι φράσεις στο κείμενο ενημερώνουν το μοντέλο για τη χρήση της γλώσσας με αποτέλεσμα να παράγει ακριβέστερες και σχετικότερες αναπαραστάσεις από ότι αν έλειπαν. Η συμπεριφορά αυτή οφείλεται στο ότι περιέχει πληροφορίες χρήσης αδόμητου, ελεύθερου κειμένου και κατανόηση γλώσσας, η οποία χρησιμοποιεί αυτες τις απλές λέξεις για να εμπλουτίσει το εκφραζόμενο νόημα.

2.4.3 Αξιολόγηση Αλγορίθμων

Η προσαρμογή των παραμέτρων μιας συνάρτησης εκτίμησης και ο έλεγχος της στα ίδια δεδομένα συνιστούν μεθοδολογικό σφάλμα: ένα μοντέλο που θα επαναλάμβανε απλώς τις ετικέτες των δειγμάτων που μόλις είδε θα είχε τέλεια ακρίβεια, αλλά δεν θα μπορούσε χρησιμεύσει για προβλέψεις σε άγνωστα δεδομένα. Αυτή η περίπτωση ονομάζεται 'υπερπροσαρμογή' του μοντέλου στα δεδομένα. Προκειμένου να αποφευχθεί, είναι κοινή πρακτική όταν εκτελείται ένα

εποπτευόμενο πείραμα μηχανικής εκμάθησης να διατηρείται μέρος των διαθέσιμων δεδομένων ως σύνολο ελέγχου. Ακολουθεί ένα διάγραμμα ροής τυπικής ροής εργασίας διασταυρούμενης επικύρωσης στην εκπαίδευση μοντέλων.



Σχήμα 2.2: Αξιολόγηση Μοντέλου

Κατά την αξιολόγηση διαφορετικών ρυθμίσεων ('υπερπαραμέτρων', hyperparameters) προγνωστικών αλγορίθμων, όπως η παράμετρος C που πρέπει να ρυθμιστεί από το χρήστη για μια ΜΔΥ, εξακολουθεί να υπάρχει κίνδυνος υπερπροσαρμογής στο σύνολο ελέγχου, επειδή οι παράμετροι μπορούν να τροποποιηθούν έως ότου ο αλγόριθμος φτάσει τη βέλτιστη απόδοση. Έτσι, η πληροφορία από το σύνολο ελέγχου μπορεί να «διαρρεύσει» στο μοντέλο και οι μετρήσεις αξιολόγησης να μην αναφέρουν πλέον την ακρίβεια σε γενικευμένα δεδομένα. Προς λύσιν αυτού του προβλήματος, ένα άλλο μέρος του συνόλου δεδομένων μπορεί να διατηρηθεί ως ένα λεγόμενο ΄σύνολο επικύρωσης΄ (validation set): η εκπαίδευση γίνεται στο σύνολο εκπαίδευσης, μετά την οποία το μοντέλο αξιολογείται στο σύνολο επικύρωσης και όταν βρεθούν οι βέλτιστες παράμετροι, η τελική αξιολόγηση εκτελείται στο σύνολο ελέγχου.

Ωστόσο, αυτός ο διαχωρισμός των διαθέσιμων δεδομένων σε τρία σύνολα μειώνει δραστικά τον αριθμό των δειγμάτων που χρησιμοποιούνται για την εκπαίδευση του μοντέλου με τα αποτελέσματα να εξαρτώνται από μια τυχαία επιλογή των συνόλων (εκπαίδευσης, επικύρωσης).

Μια λύση σε αυτό το πρόβλημα είναι μια διαδικασία που ονομάζεται διασταυρούμενη επικύρωση (cross-validation). Ένα σύνολο ελέγχου θα πρέπει να κρατηθεί για την τελική αξιολόγηση, αλλά το σύνολο επικύρωσης είναι προαιρετικό. Η βασική προσέγγιση, γνωστή και ως k-fold cross validation, το σύνολο εκπαίδευσης χωρίζεται σε k υποσύνολα (άλλες προσεγγίσεις περιγράφονται παρακάτω, αλλά γενικά ακολουθούν τις ίδιες αρχές). Στη συνέχεια παρατίθεται η διαδικασία για καθεμία από τις k «αναδιπλώσεις».(Scikit-learn: Machine Learning in Python, 2011) (API design for machine learning software: experiences from the scikit-learn project, 2013)

2.4.4 Διασταυρωμένη Επικύρωση (Cross-Validation)

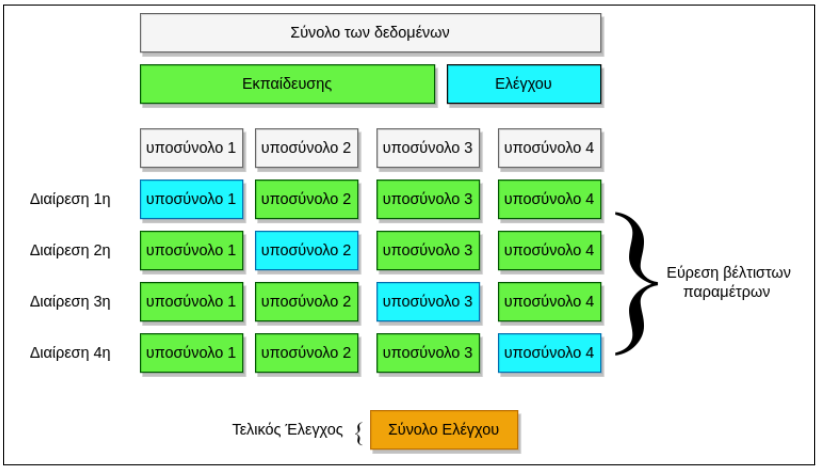
Η διασταυρούμενη επικύρωση είναι μια διαδικασία επαναδειγματοληψίας που χρησιμοποιείται για την αξιολόγηση μοντέλων μηχανικής εκμάθησης σε ένα περιορισμένο σύνολο δεδο-

μένων. Η διαδικασία έχει μια μοναδική παράμετρο k η οποία αναφέρεται στον αριθμό των κλάσεων στις οποίες πρόκειται να χωριστεί ένα δεδομένο σύνολο δεδομένων. Ως εκ τούτου, η διαδικασία ονομάζεται συχνά k-fold cross-validation. Όταν επιλέγεται μια συγκεκριμένη τιμή για το k, μπορεί να χρησιμοποιηθεί στη θέση του κ στην αναφορά στο μοντέλο, όπως το $k = 10$ να γίνει δεκαπλάσια διασταυρούμενη επικύρωση (10-fold Cross Validation).

Η διασταυρούμενη επικύρωση χρησιμοποιείται κυρίως στην εφαρμοσμένη μηχανική μάθηση για την εκτίμηση των δυνατοτήτων ενός μοντέλου μηχανικής μάθησης όταν κληθεί να ταξινομήσει άγνωστα δεδομένα. Δηλαδή, να χρησιμοποιηθεί ένα περιορισμένο δείγμα προκειμένου να εκτιμηθεί πώς αναμένεται να αποδώσει το μοντέλο γενικά όταν χρησιμοποιείται για να κάνει προβλέψεις σε δεδομένα που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση του. Είναι μια δημοφιλής μέθοδος επειδή είναι απλή στην κατανόηση και επειδή γενικά οδηγεί σε μια λιγότερο προκατειλημμένη ή λιγότερο αισιόδοξη εκτίμηση (υπερπροσαρμογή) της ικανότητας του μοντέλου από άλλες μεθόδους, όπως μια απλή διαίρεση των δεδομένων σε σύνολα εκπαίδευσης/δοκιμής.

Σε γενικές γραμμές, η διαδικασία έχει ως εξής:

1. Τυχαία ανάμιξη των περιεχόμενων του συνόλου.
2. Διαχωρισμός του συνόλου σε κ υποσύνολα
3. Για κάθε υποσύνολο:
   (α′) Αντιμετώπιση του υποσυνόλου ως συνόλου ελέγχου
   (β′) Αντιμετώπιση των υπολοίπων ως σύνολο εκπαίδευσης
   (γ′) Προσαρμογή του μοντέλου στο σύνολο εκπαίδευσης και έλεγχος στο σύνολο ελέγχου
   (δ′) Διατήρηση μόνο του αποτελέσματος για κάθε επανάληψη
4. Σύνοψη των δυνατοτήτων του μοντέλου με βάση τα αποτελέσματα των δοκιμών



Σχήμα 2.3: k-fold cross validation

Κάθε παράδειγμα στο δείγμα δεδομένων ομαδοποιείται σε ένα μοναδικό υποσύνολο και παραμένει σε σε αυτό κατά τη διάρκεια της διαδικασίας. Άρα κάθε δείγμα χρησιμοποιείται στο σύνολο ελέγχου 1 φορά και στο σύνολο εκπαίδευσης του μοντέλου $k - 1$ φορές (An Introduction to Statistical Learning, 2013).

Παρά τις καλύτερες προσπάθειες των στατιστικών μεθοδολόγων, οι χρήστες συχνά μολύνουν τα αποτελέσματά τους κοιτάζοντας κατά λάθος τα δεδομένα ελέγχου (Artificial Intelligence: A Modern Approach (3rd Edition), 2009). Τα αποτελέσματα μιας διασταυρούμενης επικύρωσης συχνά συνοψίζονται χρησιμοποιώντας τον μέσο όρο της απόδοσης του μοντέλου. Καλή πρακτική επίσης θεωρείται η συμπερίληψη κάποιου μέτρου της διακύμανσης της απόδοσης, όπως η τυπική απόκλιση ή το τυπικό σφάλμα. [7]

Κεφάλαιο 3

Μεθοδολογία

3.1 Συλλογή Δεδομένων

3.1.1 Μορφή και χαρακτηριστικά Βιογραφικών Σημειωμάτων

Συνήθως, τα βιογραφικά που κυκλοφορούν στους ιστότοπους αναζήτησης εργασίας είναι μικρότερου μεγέθους, αλλά αυτό μπορεί να διαφέρει ανάλογα με τον κλάδο. Για παράδειγμα, τα βιογραφικά των επιστημόνων θα είναι μεγαλύτερα, ειδικά αν συνεχίζουν να αναλαμβάνουν ερευνητικές εργασίες. Πιθανές καριέρες μπορεί να απαιτούν περισσότερες πληροφορίες στο βιογραφικό, καθώς οι προαγωγές και οι αλλαγές θέσεων μπορεί να παίξουν σημαντικό ρόλο στο καταληκτικό μέγεθος του.

Κύκλος ζωής ενός βιογραφικού

Δομικά, ένα βιογραφικό αποτελείται από παραρτήματα σχετικά με πληροφορίες για τον κάτοχο οι οποίες θα ήταν χρήσιμες για έναν οργανισμό που αναζητά ανθρώπους με τα κατάλληλα προσόντα ώστε να καλύψουν τις αρμοδιότητες που χρειάζεται για να λειτουργήσει.

Ένα βιογραφικό σημείωμα τυπικά συντίθεται όταν το άτομο αρχίζει να ψάχνει για εργασία. Συνήθως αυτό γίνεται μετά το πέρας των σπουδών, έτσι το βιογραφικό ξεκινάει μόνο με το πτυχίο και τη σχολή του ατόμου και ίσως λίγα λόγια για τον κάτοχο. Το παράρτημα λοιπόν με τη μόρφωση του ατόμου συχνά θα μείνει στατικό, εκτός από την περίπτωση του μεταπτυχιακού, διδακτορικού η την επανένταξη του ατόμου σε κάποια δεύτερη σχολή. Επίσης, οι περισσότερες σχολές περιλαμβάνουν μια περίοδο πρακτικής άσκησης, όπου το άτομο θα αποκτήσει μια περιστασιακή εργασιακή εμπειρία ως πρακτικάριος σε κάποια εταιρεία σχετική με το αντικείμενο του, και έτσι θα έχει γεμίσει το πρώτο μέρος της εργασιακής εμπειρίας στο βιογραφικό. Από εκεί και πέρα το κυρίως μεταβαλλόμενο μέρος του βιογραφικού είναι εκείνο που αναφέρεται στα πόστα και τα εργασιακά καθήκοντα του κάθε πόστου, καθώς και το παράρτημα της εργασιακής εμπειρίας. Αυτό το παράρτημα τυπικά έχει μεγαλύτερη ταχύτητα αλλαγής τα πρώτα χρόνια της ένταξης του ατόμου στην αγορά εργασίας, όπου στην περίπτωση των μισθωτών, το άτομο ξεκινάει από θέσεις μικρής ευθύνης και ανά διαστήματα πιθανόν να προάγεται σε υψηλότερες θέσεις στο ίδιο πόστο. Δεν είναι σπάνιο επίσης να αλλάζει πόστο

εντός αυτής της ανέλιξης. Αυτό θα συνεχιστεί μέχρι να φτάσει στην 'incompetence level'[1] το επίπεδο όπου το άτομο δεν μπορεί πλέον να ανελιχθεί περαιτέρω και να μεταβάλλει τη θέση του, το οποίο αντικατοπτρίζεται στο βιογραφικό, με θέσεις πλέον να κρατάνε χρόνια μέχρι και δεκαετίες.

Σε περιπτώσεις ελεύθερων επαγγελματιών παρατηρείται ένα παρόμοιο μοτίβο, όπου ανάλογα με τον κλάδο (και κυρίως σε επαγγέλματα μικρής εξειδίκευσης όπως οδηγός, εργάτης, κηπουρός, σερβιτόρος ή χειριστής μηχανήματος) θα εμφανιστεί πολλές φορές μια αλλαγή στο επάγγελμα, μέχρι κάποιο σημείο όπου το άτομο θα έχει εξειδικευθεί αρκετά ώστε να παραμείνει σε μια θέση (π.χ. στην περίπτωση όπου ένας εργάτης ανελιχθεί σε προϊστάμενο). Άλλη περίπτωση είναι οι ελεύθεροι επαγγελματίες που στην ουσία συνεργάζονται, με τον ίδιο ρόλο, με πολλούς διαφορετικούς πελάτες ανά τα έτη.

Ακόμα μια περίπτωση είναι οι καθηγητές-ερευνητές όπου πέρα από τις θέσεις που αναλαμβάνουν σε διάφορους οργανισμούς, ασχολούνται και με ερευνητικά έργα τα οποία αναγράφονται στο βιογραφικό και μπορεί αυτό, ενώ το παράρτημα των θέσεων να μην αλλάζει, να αυξάνεται σε όγκο λόγω των ερευνητικών έργων.

3.2 Συλλογή δεδομένων από ιστότοπους βιογραφικών

Στη διαδικτυακή εποχή, τα βιογραφικά απαντώνται είτε σε προσωπικές ιστοσελίδες (π.χ. στην ιστοσελίδα της σχολής, στο προφίλ του κάθε καθηγητή) είτε στους προσωπικούς υπολογιστές των κατόχων.

Κυρίως βρίσκονται συγκεντρωμένα σε ιστοτόπους με μεγάλες, κεντρικές βάσεις δεδομένων που κάνουν διευκολύνουν την εύρεση και διαλογή βιογραφικών από πράκτορες προσλήψεων και εργαζόμενους ανθρώπινων πόρων. Αυτές οι πλατφόρμες συσσωρεύουν μεγάλο όγκο βιογραφικών σε κυκλοφορία στην αγορά εργασίας. Οι πλατφόρμες που θα ασχοληθούμε σε αυτή τη εργασία είναι το livecareer.com και indeed.com τα οποία είναι ιδανικά για την εργασία μας για τους λόγους ότι περιλαμβάνουν ολόκληρα βιογραφικά και είναι ανοιχτά στους ανθρώπους που ψάχνουν, σε διαφορά με το LinkedIn. Επιλέχθηκαν οι παρακάτω 18 κατηγορίες βιογραφικών ως queries:

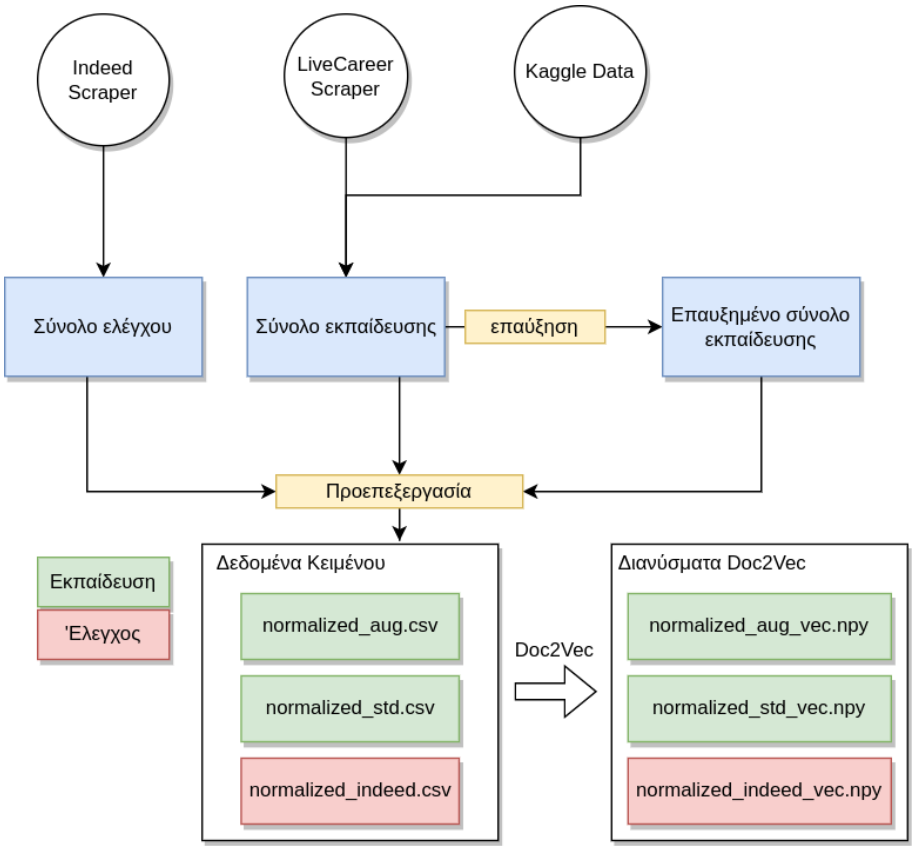
- Κηπουρός
- Αρχιτέκτονας
- Τραπεζίτης
- Πολιτικός μηχανικός
- Μάγειρας
- Εξυπηρέτηση πελατών
- Οδοντίατρος
- Οδηγός
- Ηλεκτρολόγος μηχανικός
- Μηχανικός αυτοκινήτων
- Επιστήμη Δεδομένων
- Ανθρωπίνων Πόρων
- Ψυκτικός
- IT Admin

[1] https://en.wikipedia.org/wiki/Peter_principle

- Νομικός
- Marketing
- Σιδηρουργός / Ηλεκτροσυγκολλητής
- Web Developer

Οι επιλογές αυτές είναι αυθαίρετες, αλλά γίνεται η προσπάθεια να συλλάβουμε ένα όσο το δυνατόν γίνεται μεγαλύτερο φάσμα της αγοράς εργασίας. Μια ιδιαιτερότητα που εμφανίστηκε ήταν στο επάγγελμα του αρχιτέκτονα, μιας και ο όρος αρχιτέκτονας εμφανίζεται σε ποικίλες θέσεις, και όχι μόνο ως σχεδιαστής κτηρίων, όπως solutions architect, software architect. Συμπληρωματικά, προσθέτουμε και ένα μικρό σύνολο που βρέθηκε στην ιστοσελίδα επιστήμης δεδομένων kaggle. Η κατηγοριοποίηση γίνεται με βάση μια μόνο ετικέτα, πράγμα το οποίο περιορίζει τις δυνατότητες του μοντέλου μας, ιδιαίτερα σε περιπτώσεις όπου το βιογραφικό δεν έχει κάποια γενική εξειδίκευση και περιλαμβάνει διάφορες θέσεις άσχετες μεταξύ τους, ή θέσεις που μοιάζουν και υπάρχει περίπτωση να μπερδεύουν το μοντέλο λόγω παρόμοιων καθηκόντων.

Η εκπαίδευση αλγορίθμων ταξινόμησης προϋποθέτει ότι τα δεδομένα έχουν επισημάνσεις ή ετικέτες. Για τη δημιουργία νέου επισημασμένου συνόλου δεδομένων θα ακολουθήσουμε μια μη εποπτευόμενη διαδικασία επισήμανσης, χρησιμοποιώντας τεχνικές ιστοσυγκομιδής για να εξάγουμε και παράλληλα να επισημάνουμε τα δεδομένα που συλλέγονται.



Σχήμα 3.1: Δημιουργία Συνόλων Δεδομένων

3.2.1 Πηγές αναζήτησης βιογραφικών

Η συλλογή των βιογραφικών έγινε με μια αυτοματοποιημένη διαδικασία συγκομιδής δεδομένων ιστού. Ξεκινάμε θέτοντας ερωτήματα (queries) σε μηχανές αναζήτησης βιογραφικών, τα οποία επιστρέφουν μια λίστα με αποτελέσματα. Έπειτα, αποθηκεύουμε το περιεχόμενο των αποτελεσμάτων στη βάση δεδομένων μας, με ετικέτα την είσοδο που μας επέστρεψε το αποτέλεσμα. Επαναλαμβάνουμε τη διαδικασία μέχρις ότου έχει συλλεχθεί επαρκής όγκος δεδομένων.

Τα αποτελέσματα της αναζήτησης βασίζονται σε αλγόριθμους σύστασης των σχετικότερων αποτελεσμάτων (sorted by relevance). Δεν υπάρχει ξεκάθαρος τρόπος να ταχθούν και έτσι οι ιστοσελίδες δίνουν την δυνατότητα να προσαρμοστεί η αναζήτηση με διάφορα φίλτρα σχετικά με την τοποθεσία, το χρόνο και την κατηγορία, επιστρέφοντας διαφορετικές διατάξεις. Η ακριβής λειτουργία των αλγόριθμων προτάσεων δεν είναι γνωστή, αλλά φαίνεται ότι βασίζονται στη συνάφεια σε σχέση με το περιεχόμενο (ταιριάζοντας μάλιστα στις ακριβείς λέξεις της αναζήτησης με τα κείμενα που επιστρέφει, υπογραμμισμένες μέσα στο κάθε κείμενο) και του πόσο πρόσφατα αναρτήθηκε το βιογραφικό, ή πόσο πρόσφατα έγινε η τελευταία τροποποίηση, δείχνοντας ότι έγινε κάποια αλλαγή στην επαγγελματική κατάσταση του κατόχου.

Η τελευταία τροποποίηση στους ιστότοπους βιογραφικών (resume aggregators) είναι μια αρκετά καλή έμμεση ένδειξη του αν το άτομο ψάχνει ενεργά για εργασία, και χρησιμοποιείται από το Indeed στην πρόταση των βιογραφικών των χρηστών. Μια άλλη προσέγγιση είναι τα προφίλ του LinkedIn, όπου το άτομο έχει ένα μόνιμο προφίλ, και ανακοινώνει τη διάθεση του για εύρεση εργασίας ρητά, με την ενεργοποίηση μιας ρύθμισης στο προφίλ του χρήστη και το keyword "opentowork.

Εξαγωγέας δεδομένων από LiveCareer.com

Ο ιστότοπος Livecareer περιέχει μια μεγάλη βάση δεδομένων απο τυποποιημένα παραδείγματα βιογραφικών, στα οποία έχουν αντικατασταθεί οι προσωπικές πληροφορίες του ατόμου με σύμβολα, αποκρύπτοντας έτσι ονόματα, ημερομηνίες, γεωγραφικές περιοχές και χρονικές περιόδους. Παρόλα αυτά, παραμένουν τα μέρη που περιγράφουν τις θέσεις και τις περιγραφές των καθηκόντων της κάθε θέσης. Η σελίδα αυτή είναι ιδιαίτερα βολική για την εξαγωγή δεδομένων μιας και προσφέρει τα βιογραφικά σε στατικές ιστοσελίδες στο διαδίκτυο στις οποίες ένα συγκεκριμένο μέρος του κώδικα HTML περιλαμβάνει όλο τα δεδομένα κείμενου του βιογραφικού. Έπισης δεν χρησιμοποιεί πολλά μέτρα προστασίας από προσπαθήσεις εξαγωγής των δεδομένων και τέλος είναι δωρεάν και διαθέσιμα για όλους.

Έτσι, χρησιμοποιώντας μόνο εργαλεία για την εξαγωγή του κειμένου από HTML, δημιουργούμε τα αρχεία ακολουθώντας την παρακάτω διαδικασία:

- Κάνουμε ένα request στη σελίδα χρησιμοποιώντας μια εκ των κατηγόριων στο URL της μηχανής αναζήτησης, και μας επιστρέφεται μια σελίδα η οποία περιέχει τη λίστα με σχετικά βιογραφικά.

- Για κάθε μέρος της λίστας, αποθηκεύουμε το URL σε μια ενα csv αρχείο, μαζί με την

κατηγορία.

- Το παραπάνω επαναλαμβάνεται για όσες σελίδες επιλέξουμε εμείς, στην παρούσα περίπτωση με 100 βιογραφικά ανά κατηγορία, δηλαδή 2 σελίδες με 50.
- Έχοντας αυτό το σύνολο δεδομένων, περνάμε από κάθε σύνδεσμο και για κάθε έναν, ζητάμε τη σελίδα του βιογραφικού σε μορφή HTML.
- Βρίσκουμε το στοιχείο (element) με το κείμενο των βιογραφικών και εξάγουμε όλα τα στοιχειά με κείμενο.
- Με χρήση της βιβλιοθήκης BeautifulSoup4 εξάγουμε από τον κώδικα τα κείμενα που μας ενδιαφέρουν από τα δομικά μέρη του HTML (elements).
- Η εξαγόμενη πληροφορία αποθηκεύεται στην ίδια γραμμή, σε μια νέα στήλη resume.
- Μετά το πέρας όλων των γραμμών και την ολοκλήρωση της στήλης με τα βιογραφικά, μπορούμε να διαγράψουμε τη στήλη με τους συνδέσμους.

Έτσι μας μένει ένα αρχείο με δυο στήλες: ΄resume΄ και ΄category΄.

Εξαγωγέας δεδομένων απο Indeed.com

Η εξαγωγή δεδομένων από τον ιστότοπο Indeed αποδείχτηκε μια δυσκολότερη διαδικασία, λόγω διάφορων μέτρων που έχουν ληφθεί για να μην δύνανται οποιοσδήποτε να κατεβάσει ολόκληρη την βάση δεδομένων, μιας και τα βιογραφικά είναι πραγματικά, με προσωπικές πληροφορίες όπως ονόματα και διευθύνσεις. Για να αποκτήσει κάποιος πρόσβαση στη βάση δεδομένων, χρειάζεται να είναι εγγεγραμμένος ως εργοδότης, όπως επίσης και να πληρώνει συνδρομή στα διαφορετικά πλάνα που παρέχονται. Παρέχεται η δυνατότητα δωρεάν δοκιμής της υπηρεσίας για 14 ημέρες. Έχουν ληφθεί μέτρα ώστε να μην επιτρέπεται να δει κάποιος τις σελίδες αν δεν είναι συνδεδεμένος, το οποίο εξ αρχής δυσκολεύει μιας και είναι αθέμιτο να συνδεθείς από το selenium σε λογαριασμό google που απαιτείται, ενώ παράλληλα υπάρχουν μέτρα τα οποία παρατηρούν τον χρήστη για τυχών ύποπτη συμπεριφορά. Περόν αυτού, τα δεδομένα δεν είναι στατικά και αλλάζουν κάθε φορά που πατάει ο χρήστης κλικ στη σελίδα, με συνεχώς μεταβαλλόμενα, δυναμικά css classes τα οποία αποφεύγουν την σύλληψη με απλούς τρόπους, πως επίσης και με αυστηρή καταγραφή των συνεδριών στην ιστοσελίδα με cookies όταν κάποιος ζητάει μια σελίδα. Όταν παρατηρηθεί ύποπτη συμπεριφορά σε ένα από τα στρωματά ασφάλειας, ενεργοποιείται το σύστημα captcha. Περαιτέρω, οι λογαριασμοί για να δημιουργηθούν χρειάζονται επιβεβαίωση με αριθμό τηλεφώνου.

Παρόλα αυτά, με την σωστή χρήση των εργαλείων αυτοματοποίησης, κατασκευάστηκε ενα πρόγραμμα το οποίο κατάφερε να συλλέξει και να επισημάνει ικανοποιητικές ποσότητες δεδομένων για τους σκοπούς της εργασίας, με τρόπο εξομοίωσης ανθρωπινής συμπεριφοράς.

1. Έχοντας το URL σαν βάση κάνουμε requests στην κάθε σελίδα με διάφορες παραμέτρους. Στη συγκεκριμένη περίπτωση έχουμε τις παραμέτρους επάγγελμα και: περιοχή

New York, 50mile radius, και ζητάμε τα πρώτα 100 αποτελέσματα. Χρησιμοποιούμε αλγορίθμους που παράγουν ομοιογενείς κατανομές χρόνου για να αποφύγουμε τα απότομα click που ειδοποιούν την ασφάλεια για ΄μηχανική΄ συμπεριφορά.

2. Βρίσκουμε την στατική λίστα που επιστρέφεται και χρησιμοποιώντας actuators που παρέχονται με το selenium πατάμε click στο πρώτο element.

3. Δημιουργείται μια δυναμική σελίδα στα δεξιά του ιστότοπου με τα δεδομένα του βιογραφικού σε μορφή απλού κειμένου. Είναι ένα δυναμικό HTML iFrame που αλλάζει κάθε φορά που πατάμε click σε κάποιο μέλος της λίστας. Λόγω του ότι είναι iFrame, δεν εμφανίζεται στο HTML της σελίδας και από άποψη εξαγωγής μέσω του αρχικού HTML είναι αόρατο.

4. Χρησιμοποιώντας εργαλεία εξομοίωσης κίνησης του ποντικιού, τοποθετούμε τον κέρσορα σε ένα γειτονικό στοιχείο το οποίο υπάρχει διπλά από το iFrame στο αρχικό HTML της σελίδας, και το θέτουμε να μετακινηθεί μέσα στο στοιχείο που βρίσκεται το δυναμικό περιεχόμενο.

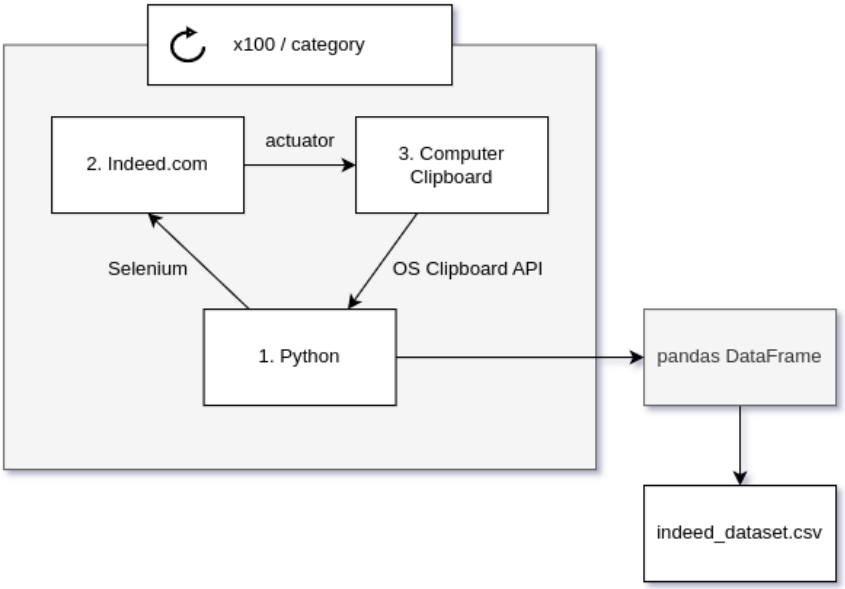
5. Από αυτό το σημείο, προσομοιώνουμε την συμπεριφορά ανθρωπίνου χρήστη, στέλνοντας στον ιστότοπο σειρές πιέσεων πλήκτρων του πληκτρολογίου εκμεταλλευόμενοι το γεγονός ότι παράλληλα αυτά ενεργοποιούν λειτουργίες του λειτουργικού συστήματος.

6. Αρχικά στέλνουμε το click, έπειτα τον συνδυασμό ctrl-a το οποίο συλλαμβάνει μόνο ότι υπάρχει μέσα στο iFrame, και μετα ctrl-c. Αυτό αντιγράφει ό,τι βρίσκεται επιλεγμένο στη μνήμη (clipboard) του λειτουργικού συστήματος.

7. Έπειτα μέσω βιβλιοθηκών για πρόσβαση ενός προγράμματος python στη μνήμη του λειτουργικού, προσπελαύνουμε το clipboard και προσθέτουμε τα περιεχόμενα του σε μια λίστα, μαζί με την ετικέτα κατηγορίας σε μορφή πλειάδας.

8. Η διαδικασία επαναλαμβάνεται για όλες τις κατηγορίες. Σε κάθε νέο request μας ζητείται captcha, το οποίο συμπληρώνεται χειροκίνητα.

9. Τέλος, περνάμε τη λίστα python σε ένα Pandas Dataframe, και το αποθηκεύουμε στο δίσκο.



Σχήμα 3.2: Ιστοσυγκομιδή: Indeed.com

3.2.2 Προεπεξεργασία κειμένων και κατασκευή βάσεων δεδομένων

Μέτα τη συλλογή των δεδομένων και την αποθήκευση τους σε αρχεία csv, το επόμενο βήμα είναι να τα επεξεργαστούμε ώστε να μετατραπούν σε μια μορφή κατάλληλη για την περαιτέρω χρήση τους. Το πρώτο βήμα σε οποιαδήποτε διαδικασία εξομάλυνσης δεδομένων είναι η αφαίρεση διπλοτύπων και τυχών ανεπιθύμητων ανωμαλιών.

Σκοπός της επεξεργασίας είναι να απλοποιηθούν τα δεδομένα μας καθαρίζοντας τα από θόρυβο, ο οποίος κάνει την πληροφορία δυσνόητη και κατ΄ επέκταση δυσκολεύει την ερμηνεία του περιεχομένου. Για δεδομένα κειμένου, ένας τρόπος είναι να εξαλείψουμε τη διαφορά των ίδιων γραμμάτων σε πεζά-κεφαλαία, κάνοντας τα όλα πεζά, και να αφαιρέσουμε από το κείμενο τις λεκτικές μονάδες που δεν είναι λέξεις, όπως σημεία στίξης, αριθμούς και ειδικούς χαρακτήρες. Έτσι στο καινούργιο κείμενο μένουμε μόνο με τις λέξεις αυτές καθ΄ αυτές ώστε στη συνέχεια τα μοντέλα μας να πρέπει να ερμηνεύσουν μόνο τις ίδιες τις λέξεις και τις συντακτικές τους σχέσεις.

Άλλη πληροφορία που θέλουμε να αφαιρέσουμε είναι τα τυποποιημένα κείμενα που προήλθαν από την ιστοσελίδα livecareer.com. Για την ομαλοποίηση αυτών των κειμένων χρησιμοποιούμε μια σειρά από διαδοχικά περάσματα του κειμένου, αντικαθιστώντας ένα συγκεκριμένο ανεπιθύμητο μοτίβο με κενό χαρακτήρα.

Το μόνο σημείο στίξης που επιτρέπουμε στο κείμενο είναι η τελεία, λόγω του ότι είναι απαραίτητη για τον χωρισμό των προτάσεων, μιας εργασίας που θα μας φανεί χρήσιμη στη συνέχεια, όταν θα χρειαστεί να εξάγουμε προτάσεις στο στάδιο της επαύξησης δεδομένων.

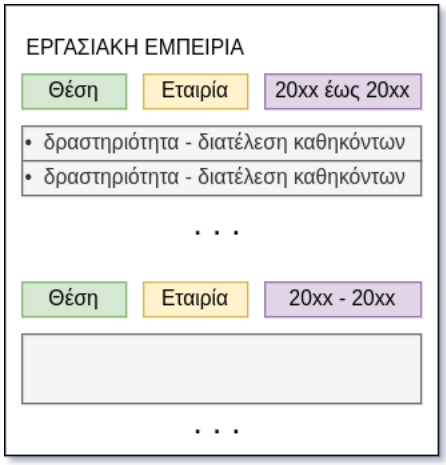
Η παραπάνω διαδικασία μπορεί να εφαρμοστεί ως μια απλή συνάρτηση python στη στήλη με τα κείμενα, εφαρμόζοντας αντικατάσταση με τη χρήση κανονικών εκφράσεων (Regular Expressions).

Τυπικά, στα συστήματα ΕΦΓ, η εργασία της προ-επεξεργασίας κειμένου υλοποιείται ως το πρώτο μέρος μιας σταδιακής διαδικασίας επεξεργασίας κειμένου ως μια στρώση από την οποία περνάει κάθε νέο κείμενο εισόδου. Για λόγους οικονομίας χρόνου και του γεγονότος ότι ο όγκος των δεδομένων που συλλέχθηκαν στα πλαίσια της εργασίας παραμένει σταθερός, επιλέχθηκε να γίνει η προεπεξεργασία στο σύνολο των δεδομένων εισόδου και στα δύο dataset, αποθηκεύοντας τα εκ νέου στην καθαρισμένη μορφή. Οι υπόλοιποι μετασχηματισμοί του κειμένου, όπως αφαίρεση συχνών λέξεων ή λημματοποίηση για περαιτέρω απλοποίηση θα γίνονται στο επίπεδο της υλοποίησης, καθώς τέτοιες μετατροπές είναι τόσο συχνά εφαρμοζόμενες που παρέχονται στις βιβλιοθήκες που θα χρησιμοποιήσουμε.

3.2.3 Επαύξηση Εγγράφων Βιογραφικών Σημειωμάτων

Η ιδέα της επαύξησης δεδομένων κείμενων βιογραφικών σημειωμάτων βασίζεται στην υπόθεση ότι η κάθε θέση εργασίας μπορεί να θεωρηθεί ως ένα πεπερασμένο σύνολο σχετικών αρμοδιοτήτων οι οποίες χαρακτηρίζουν την συγκεκριμένη θέση και την ξεχωρίζουν από τις άλλες.

Στο παράρτημα των βιογραφικών με τις εμπειρίες του ατόμου, αυτή η υπόθεση απαντάται με το καθιερωμένο μοτίβο ΄τίτλος θέσης - εταιρία - χρονικό διάστημα΄, ενώ ακολουθεί μια μη αριθμημένη λίστα με τις διάφορες αρμοδιότητες που επιτέλεσε ο εργαζόμενος κατά την παραμονή του σε αυτή τη θέση.



Σχήμα 3.3: Δομή Παραρτήματος Εμπειρίας

Σκοπός της επαύξησης δεδομένων είναι η δημιουργία ενός συνθετικού συνόλου δεδομένων από τα αρχικά παραδείγματα. Στην παρούσα υλοποίηση, ο συγγραφέας εμπνεύστηκε από μια τέτοια διαδικασία για την επαύξηση δεδομένων σε εικόνες, η οποία αναφέρεται στη βιβλιογραφία ως 'albumentation"[2]. Εφόσον τα καθήκοντα μιας θέσης είναι στο σύνολο τους πεπερασμένα, μπορούμε να δημιουργήσουμε νέες ΄περιπτώσεις΄ των ίδιων προτάσεων που περιγράφουν τις αρμοδιότητες κοιτώντας τες από μια διαφορετική 'οπτική γωνία'. Στην υλοποίηση μας, χρησιμοποιούμε την βιβλιοθήκη επαύξησης κειμένου NLPAug, με την μέθοδο επαύξησης

[2]https://www.kaggle.com/code/shonenkov/nlp-albumentations

προτάσεων Back-Translation βασισμένη σε υλοποιήσεις νευρωνικών δικτύων σχεδιασμένες για μετάφραση κειμένων από δεδομένα του Facebook.

Αρχικά, παίρνουμε όλα τα κείμενα από κάθε κατηγορία και τα ενώνουμε σε ένα συνολικό, μεγάλο αρχείο, τη ΄μάνα΄, το οποίο περιέχει όλες τις περιγραφές των αρμοδιοτήτων της θέσης από όλα τα βιογραφικά. Έπειτα, χρησιμοποιώντας τον SpaCy Sentencizer , σπάμε το κείμενο σε πολλές προτάσεις που περιγράφουν τις εμπειρίες και τις δραστηριότητες του εργαζομένου σε αυτή τη θέση, και τις αποθηκεύουμε σε ένα νέο σύνολο δεδομένων προτάσεων.

Υπάρχουν πολλοί τρόποι για την επαύξηση κειμένου σε επίπεδο χαρακτήρα, λέξης ή πρότασης, όπως για παράδειγμα σε δεδομένα ΄πρόχειρου΄ κειμένου όπως μηνύματα ή εισαγωγή τυπογραφικών λαθών. Ωστόσο σε κείμενα βιογραφικών σημειωμάτων υποθέτουμε ότι δεν θα παρουσιάζουν τέτοια διακύμανση λόγω τυπογραφικών λαθών ή παρόμοιων τεχνικών. Θεωρούμε ότι τα κείμενα είναι ομαλά και περιορισμένα στη χρήση της γλώσσας.

Για τη δημιουργία νέων, συνθετικών προτάσεων, χρησιμοποιήθηκαν τεχνικές επαύξησης κειμένου σε επίπεδο πρότασης. Εφαρμόζουμε την επαύξηση στις προτάσεις που έχουμε και μας επιστρέφονται νέες, αλλαγμένες προτάσεις οι οποίες κρατάνε το σημασιολογικό περιεχόμενο.

```
consistently maintained high levels of cleanliness organization storage and sanitation of food and beverage products to ensure quality.
['The storage and hygiene of food and beverages was always at a high level clean to ensure quality.']
```

Σχήμα 3.4: Επαύξηση πρότασης

```
used proper cleaning supplies and methods to disinfect counters where raw meat poultry fish and eggs had been prepared.
['proper detergents and methods were used to disinfect counters containing raw poultry meat, fish and eggs.']
```

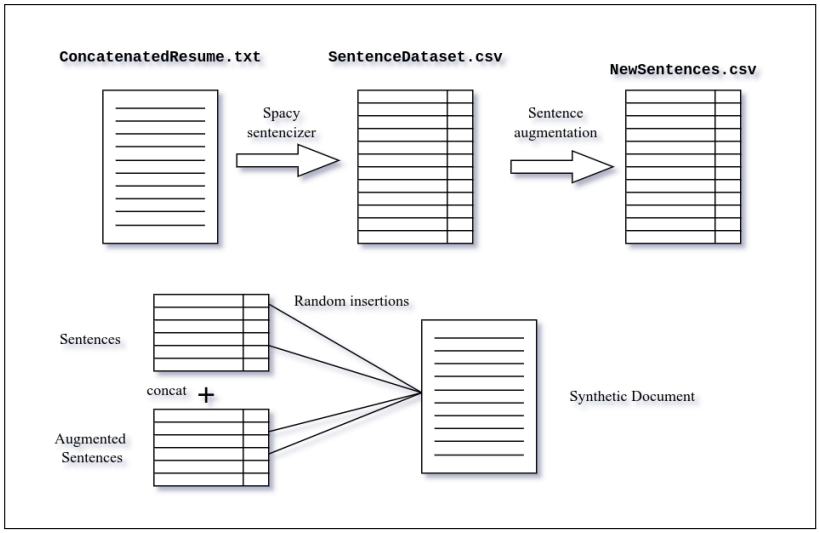
Σχήμα 3.5: Επαύξηση πρότασης

Έτσι μας επιστρέφεται ένα σύνολο επαυξημένων προτάσεων για την κάθε κατηγορία. Έχοντας αυτά τα σύνολα μπορούμε να προχωρήσουμε στη σύνθεση νέων κειμένων. Για τη σύνθεση νέων κειμένων αρχικά βρίσκουμε το μέσο όρο του μεγέθους ενός βιογραφικού σε προτάσεις, όπως και την τυπική απόκλιση. Με τη χρήση της βιβλιοθήκης random, επιλέγουμε να επιστρέφεται, για κάθε νέο βιογραφικό σε κάθε κατηγορία ένας τυχαίος αριθμός από προτάσεις, και η κάθε πρόταση επιλέγεται τυχαία από το σετ των επαυξημένων προτάσεων. Επιλέγουμε να γίνει η επαύξηση μόνο στο σύνολο δεδομένων της εκπαίδευσης, για να το συγκρίνουμε με την επίδοση του πριν και μετά την επαύξηση.

Σε αυτό το σημείο πρέπει να επιλέξουμε πόσα νέα έγγραφα θέλουμε να συνθέσουμε. Για τους σκοπούς της εργασίας επιλέγουμε 1000 νέα βιογραφικά για την κάθε κατηγορία. Τέλος, αποθηκεύουμε τα νέα βιογραφικά σε ένα νέο csv και τα ενώνουμε με το αρχικό σύνολο εκπαίδευσης.

3.3 Προσεγγίσεις εξαγωγής χαρακτηριστικών

Τα ειδοποιά χαρακτηριστικά ενός κειμένου, από την άποψη της εξαγωγής χαρακτηριστικών της μηχανικής μάθησης, είναι οι λέξεις που το συνθέτουν. Ένα λεξιλόγιο (vocabulary) είναι



Σχήμα 3.6: Διαδικασία Επαύξησης Εγγράφων

ένα πεπερασμένο σύνολο από όλους τους μοναδικούς λεξιλογικούς όρους που εμφανίζονται στα κείμενα ενός σώματος κειμένου (corpus).

Η ιδέα της λεξιλογικής διανυσματικής αναπαράστασης (vector semantics) αφορά στην αναπαράσταση κάθε λέξης ενός λεξιλογίου ως ένα σημείο (διάνυσμα) σε έναν πολυδιάστατο σημασιολογικό χώρο που προκύπτει (συνήθως) από τις κατανομές των γειτονικών της λέξεων σε γλωσσικά κείμενα. Η αναπαράσταση των λέξεων με διανύσματα ονομάζεται ενσωμάτωση. Ιδανικές ενσωματώσεις λέξεων παράγουν διανυσματικές αναπαραστάσεις όπου σημεία με παρόμοιο νόημα βρίσκονται κοντά, ενώ λέξεις με μη σχετικό νόημα βρίσκονται μακριά στο χώρο.

Distributional Hypothesis

Η υπόθεση της 'κατανομής των λέξεων' (distributional hypothesis) είναι ένα μια αρχή της γλωσσολογίας η οποία υποστηρίζει ότι λέξεις που εμφανίζονται σε παρόμοια πλαίσια τείνουν να έχουν παρόμοια νοήματα (Firth (1957)) [7] ή πιο απλά ότι το νόημα μιας λέξης εξαρτάται από τα συμφραζόμενα. Η κεντρική ιδέα εκφράζεται ως ότι 'μια λέξη χαρακτηρίζεται από τις παρέες της', διαδόθηκε από τον Firth [10] και υπονοείται στη συζήτηση περί αποσαφήνισης όρων του Weaver [11]. Είναι η βάση για την Στατιστική Σημασιολογία, και ενώ αναπτύχθηκε σαν εργαλείο για το πεδίο της γλωσσολογίας, οι ιδέες της έχουν επηρεάσει και την γνωστική επιστήμη [12].

3.3.1 Αναπαραστάσεις με αραιά διανύσματα

Μοντέλα Bag-of-words

Το Bag of words (BoW) είναι μια τεχνική που χρησιμοποιείται στην επεξεργασία φυσικής γλώσσας (NLP) για την αναπαράσταση δεδομένων κειμένου σε διανυσματική μορφή που μπορεί εύκολα να υποβληθεί σε επεξεργασία από αλγόριθμους μηχανικής μάθησης. Η βασική

ιδέα πίσω από το BoW είναι να αντιμετωπίζουμε κάθε έγγραφο ή κομμάτι κειμένου ως μια συλλογή λέξεων, αγνοώντας τη γραμματική και το συντακτικό των λέξεων, αλλά μετρώντας τη συχνότητα κάθε λέξης στο κείμενο. Αυτές οι καταμετρήσεις χρησιμοποιούνται στη συνέχεια ως χαρακτηριστικά για την αναπαράσταση του κειμένου σε διανυσματική μορφή. Σε μια προσέγγιση που χρησιμοποιεί το γλωσσικό μοντέλο Bag-of-words θεωρούμε ότι, για να χαρακτηριστεί το νόημα ενός κειμένου, αρκεί να γνωρίζουμε ποιές λέξεις χρησιμοποιηθήκαν στο κείμενο, χωρίς ενδιαφέρον για τη συντακτική τους διάταξη. Είναι μια απλοϊκή προσέγγιση, καθώς αγνοεί την πληροφορία που δίνεται από τη χρήση των λέξεων σε μια πρόταση και ομαδοποιεί μαζί ίδιες λέξεις που έχουν διαφορετικό νόημα και διαφορετική χρήση. Ως κλασσικό παράδειγμα αυτής της αδυναμίας αναφέρεται το γνωστό λογοπαίγνιο:

"time flies like an arrow, fruit flies like a banana"

όπου οι λέξεις "flies" και "like" χρησιμοποιούνται διαφορετικά στις δύο προτάσεις, αλλά για ένα μοντέλο BoW θα θεωρούνταν ίδιες. Παρόλα αυτά, λόγω των περιορισμών του νοήματος στη χρήση των λέξεων συγκεκριμένα στην περίπτωση κειμένου σε βιογραφικά σημειώματα, και λόγω της τυποποιημένης χρήσης του λόγου, μπορούμε να περιμένουμε χρήσιμα αποτελέσματα από τέτοιες τεχνικές, μιας και, ειδικά για κείμενα βιογραφικών, οι όροι που περιγράφουν εξειδικευμένες δεξιότητες χαρακτηρίζουν σε μεγάλο βαθμό το κάθε έγγραφο.

3.3.2 Πίνακες Συνύπαρξης

Μια διανυσματική αναπαράσταση κειμένου είναι αυτή των πινάκων συνύπαρξης (co-occurence matrices) λέξεων-εγγράφων ή λέξεων-λέξεων.

Σε αυτά τα μοντέλα, κάθε διάσταση αντιστοιχεί σε μια λέξη του λεξιλογίου V και τα κελιά περιέχουν τις καταμετρήσεις της κάθε λέξης. Συγκεκριμένα, στον πίνακα δίνονται από την εξίσωση συχνότητας όρου (term frequency, tf), όπου t η λέξη, d το έγγραφο και $count()$ η καταμέτρηση των εμφανίσεων της λέξης στο κείμενο.

$$tf_{t,d} = count(t, d) \tag{3.1}$$

Στην περίπτωση του πίνακα λέξεων-εγγράφων, δεδομένου λεξιλογίου N λέξεων και D εγγράφων, προκύπτει ένας πίνακας A διαστάσεων $N \times D$, με μια γραμμή για κάθε λέξη (ή όρο) του λεξιλογίου και μια στήλη για κάθε έγγραφο του corpus. Σε κάθε κελί A_{ij} βρίσκεται το πλήθος της εμφάνισης της λέξης i στο έγγραφο j.

Στον κάθετο άξονα (γραμμές) εμφανίζονται όλες οι λέξεις από όλα τα κείμενα του corpus με κάποια σειρά, όπως αλφαβητικά ή με τη συχνότητα. Μάλιστα, πολλές φορές στη δημιουργία ενός τέτοιου μοντέλου, διαγράφουμε τις πιο συχνά χρησιμοποιούμενες λέξεις, μιας και χαρακτηρίζουν ελάχιστα ένα κείμενο, αλλά μας παρέχουν καλύτερη υπολογιστική απόδοση.

Στην περίπτωση του πίνακα λέξεων-λέξεων προκύπτει ένας $N \times N$ πίνακας A, όπου στο κελί A_{ij} αυτή τη φορά βρίσκεται η καταμέτρηση των εμφανίσεων της λέξης j ως συμφραζόμενο

	Βιογραφικό 1	**Βιογραφικό 2**	**Βιογραφικό 3**	**...**	**Βιογραφικό 142**
εμπειρία	4	1	1	...	2
react	2	0	0	...	0
Ιωάννης	1	0	0	...	0
...	...	...	...	...	...
οργάνωση	0	4	2	...	4
excel	0	2	1	...	4
το	8	5	7	...	6

Πίνακας 3.1: Πίνακας εγγράφων-λέξεων

της λέξης i στο σύνολο των εγγράφων του corpus που μελετάται, με βάση κάποιο παράθυρο μεγέθους L με L λέξεις δεξιά και αριστερά της λέξης στόχου να θεωρούνται συμφραζόμενες.

Το μέγεθος του παραθύρου είναι αυθαίρετο ανάλογα με το σκοπό. Τα ΄κομμάτια΄ κειμένου ανάλογα με κάποιο παράθυρο ονομάζονται n-grams (ν-άδες) και τα μοντέλα που βασίζονται σε αυτά είναι ανάμεσα στα πιο χρησιμοποιούμενα στην επεξεργασία φυσικής γλώσσας.

	αριθμός	**ο**	**έχω**	**...**	**χρόνια**	**εμπειρία**
αριθμός	0	1	1	...	2	3
ο	2	0	0	...	0	2
έχω	1	0	0	...	7	5
...	...	...	...	...	...	...
διευθυντής	0	4	2	...	4	1
εμπειρία	0	2	1	...	0	1
χρόνο	8	5	7	...	6	0

Πίνακας 3.2: Πίνακας Συνύπαρξης λέξεων-λέξεων

Έτσι το κάθε έγγραφο ή η κάθε λέξη μπορεί να αναπαρασταθεί ως διάνυσμα είτε σε σχέση με άλλες λέξεις είτε σε σχέση με κάποιο έγγραφο.

Λόγω του ότι το λεξιλόγιο ενός εγγράφου μπορεί να είναι χιλιάδες λέξεις, και ο αριθμός αυτός αυξάνεται με κάθε νέο έγγραφο στο corpus μας, και ότι δεν υπάρχει κάθε λέξη σε κάθε έγγραφο, ο πίνακας παίρνει τη μορφή μιας αραιής μήτρας (sparse matrix), με διαστάσεις ανάλογες του λεξιλογίου και του πλήθους των εγγράφων, όπου οι περισσότερες τιμές είναι 0.

Ένα αρνητικό στοιχείο των παραπάνω αναπαραστάσεων είναι ότι με την αύξηση του πλήθους του αλφαβήτου των λέξεων και των εγγραφών οι αναπαραστάσεις αποκτούν τεράστιο μέγεθος. Επίσης, η απόλυτη τιμή συχνότητας εμφάνισης δεν δίνει πάντα χρήσιμα συμπεράσματα για σχέσεις μεταξύ των λέξεων. Για παράδειγμα έχουμε τις βασικές, μικρές και συχνά χρησιμοποιούμενες λέξεις (stopwords) όπως ΄και΄, ΄το΄, ’η’, κτλ. που εμφανίζονται συνήθως στο σύνολο των εγγράφων (και ως συμφραζόμενες στο σύνολο των υπόλοιπων

λέξεων), χωρίς να προσδίδουν ιδιαίτερο σημασιολογικό περιεχόμενο.

3.3.3 TF-IDF

Για να μετριαστεί αυτό το αρνητικό χαρακτηριστικό που οφείλεται στην χρήση της απόλυτης συχνότητας εμφάνισης, μπορούμε να ανα-ζυγίσουμε το σημασιολογικό βάρος της κάθε λέξης σε ένα έγγραφο χρησιμοποιώντας την μέθοδο TF-IDF . Η ιδέα του είναι ότι όσο περισσότερες φορές εμφανιστεί μια λέξη στο συγκεκριμένο έγγραφο, και όσο λιγότερο εμφανιστεί στα υπόλοιπα έγγραφα, τόσο περισσότερο χαρακτηρίζει το κείμενο.

Οι πίνακες TF-IDF έχουν τις ίδιες διαστάσεις με έναν πίνακα λέξεων-εγγράφων, με τη διαφορά ότι αντί για την απλή συχνότητα εμφάνισης, στο κάθε κελί υπολογίζεται το γινόμενο δύο όρων: την συχνότητα εμφάνισης του όρου στο κείμενο (term frequency) [13] και την αντίστροφη συχνότητα εγγράφου (inverse document frequency).

Ο πρώτος αφορά στην απόλυτη συχνότητα εμφάνισης της εκάστοτε λέξης σε ένα συγκεκριμένο έγγραφο:

$$tf_{t,d} = count(t, d) \tag{3.2}$$

Ο δεύτερος όρος (idf) χρησιμοποιείται για να αυξήσει το βάρος των λέξεων που εμφανίζονται σπάνια και είναι χρήσιμες για το διαχωρισμό των εγγράφων στα οποία ανήκουν. Η αντίστροφη συχνότητα εμφάνισης ενός όρου t ορίζεται ως N/df_t με N τον αριθμό των εγγράφων του corpus και df_t τον αριθμός των εγγράφων όπου εμφανίζεται ο όρος t.

Συνήθως γίνεται μια λογαριθμική συμπίεση σε αυτό τον όρο λόγω του μεγάλου πλήθους D εγγράφων που περιλαμβάνουν τα περισσότερα corpora. Ο ορισμός που προκύπτει για τον idf είναι:

$$idf_t = \log_{10}\left(\frac{N}{df_t}\right) \tag{3.3}$$

Τελικά το $tf - idf$ προκύπτει από τον πολλαπλασιασμό του όρου tf με τον όρο idf για ένα κελί $M_{i,j}$, και το μέγεθος του εκφράζει το κατ΄ αναλογία βάρος της λέξης στο έγγραφο.

	Βιογραφικό 1	**Βιογραφικό 2**	**Βιογραφικό 3**	**...**	**Βιογραφικό 142**
εμπειρία	0.03	4.23	0.008	...	2.55
react	3.92	0	0	...	0.009
Ιωάννης	0.63	0	0	...	0.06
...	...	...	...	...	...
οργάνωση	0	5.33	2	...	2.19
excel	0.93	1.008	1.03	...	1.18
το	0.004	0.032	0.001	...	0.008

3.3.4 Πυκνές Αναπαραστάσεις - Ενσωματώσεις

Η αναπαράσταση των λέξεων με μεγάλα αραιά διανύσματα εμφανίζει αδυναμίες τόσο στη δυνατότητα ανίχνευσης του νοήματος μιας λέξης, όσο και στην κλιμάκωσή τους σε μεγάλα σύνολα λέξεων και κειμένων.

Ένας διανυσματικός χώρος λεξικών ενσωματώσεων είναι ένας μικρόκοσμος από έννοιες και νοήματα σε ένα συγκεκριμένο σημασιολογικό πλαίσιο. Στην εκπαίδευση νευρωνικών γλωσσικών μοντέλων, το δίκτυο μαθαίνει τις σχέσεις μεταξύ αυτών των εννοιών.

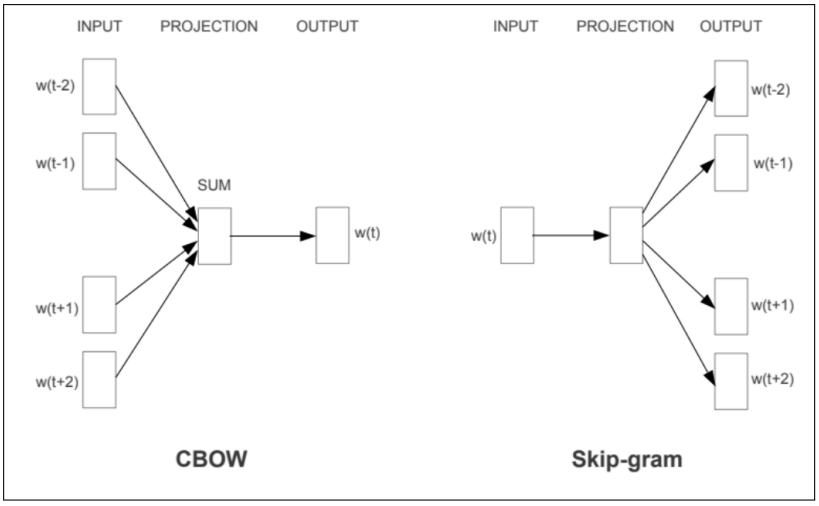
Word2Vec

Μεγάλη συνεισφορά προς τη συγκεκριμένη κατεύθυνση παρουσιάστηκε και με την δημοσίευση του Word2vec [14], ενός πακέτου λογισμικού με τεχνικές παραγωγής γλωσσικών ενσωματώσεων λέξεων, σχετικά γρήγορα και αποδοτικά, καθιστώντας το συχνά χρησιμοποιούμενη μέθοδο ενσωμάτωσης σε corpus με μεγάλο πλήθος λέξεων και εγγράφων.

Υπάρχουν δύο διαφορετικές εκδοχές του αλγορίθμου Word2Vec. Το μοντέλο του Συνεχούς Σάκου Λέξεων (Continuous Bag of Words - CBOW) και το μοντέλο Παράλειψης N-άδων (Skip-Gram) [14].

CBOW

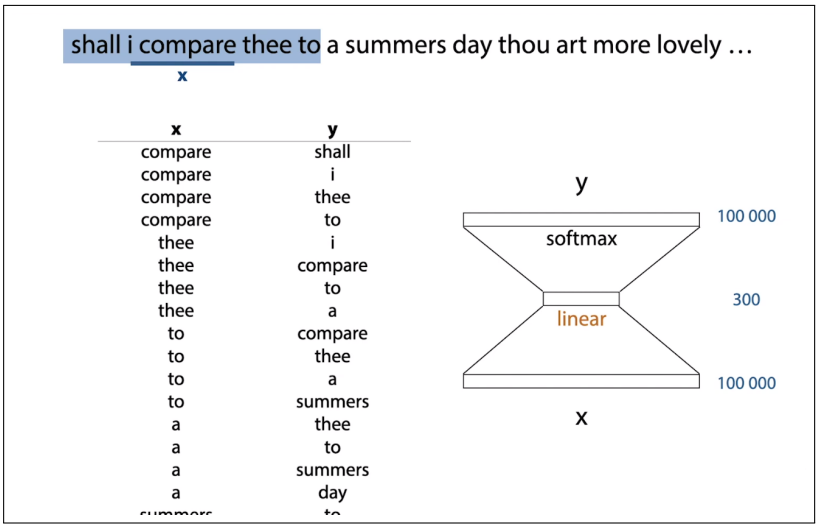
Η πρώτη είναι αυτή της πρόβλεψης της εκάστοτε λέξης από τα συμφραζόμενά της (continuous bag of words – CBOW model). Υλοποιείται ως ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο, το οποίο αποτελείται από ένα επίπεδο εισόδου, ένα επίπεδο προβολής και ένα επίπεδο εξόδου. Σκοπός της εκπαίδευσης είναι η πρόβλεψη της τρέχουσας λέξης βάσει των γειτονικών της λέξεων. Στο επίπεδο εισόδου, κωδικοποιούνται N γειτονικές λέξεις με χρήση της κωδικοποίησης one-hot. Έπειτα το επίπεδο εισόδου προβάλλεται στο επίπεδο προβολής P χρησιμοποιώντας ένα διαμοιραζόμενο πίνακα προβολών (shared projection matrix). Όλες οι λέξεις προβάλλονται στην ίδια θέση και υπολογίζεται ο μέσος όρος των αντίστοιχων διανυσμάτων (averaging).



Σχήμα 3.7: Υπολογισμος Word2vec

Skip-gram

Η δεύτερη είναι το μοντέλο skip-gram, μια ρηχή αρχιτεκτονική νευρωνικού δικτύου, μέσω της οποίας ο στόχος για κάθε λέξη είναι να προβλεφθούν οι γειτονικές της, μεγιστοποιώντας τη λογαριθμική πιθανότητα των λέξεων στο συνολικό corpus [15]:



Σχήμα 3.8: Skip-gram

όπου $nb(t)$ είναι το σύνολο των γειτονικών λέξεων μιας λέξης w_t και $p(w_j|w_t)$ είναι ιεραρχικά softmax των υπό μελέτη διανυσμάτων $v \cdot w_j$ και $v \cdot w_t$. Η αρχιτεκτονική του μοντέλου Skip-gram είναι παρόμοια με την αρχιτεκτονική CBOW. Η διαφορά τους έγκειται στον αντικειμενικό σκοπό της εκπαίδευσης. Εν αντιθέσει με το μοντέλο CBOW, το οποίο προβλέπει την κεντρική λέξη στηριζόμενο στις λέξεις που την περιβάλλουν, το μοντέλο Skip-gram προσπαθεί να μεγιστοποιήσει την πιθανότητα εμφάνισης μίας γειτονικής λέξης σε μία πρόταση, βασιζόμενο στην κεντρική λέξη.

Μια ιδιότητα μοντέλων όπως το Word2vec είναι να μοντελοποιούν σχέσεις αναλογίας ανάμεσα σε όμοιες οντότητες. Για παράδειγμα, σε έναν διανυσματικό χώρο ενσωμάτωσης και έχοντας όρους "Ελλάδα", "Αθήνα", "Ιταλία", παρατηρούμε ότι με βάση τις αναπαραστάσεις των πρώτων τριών όρων, μπορεί να βρεθεί ενας τέταρτος, με απλές πράξεις γραμμικής άλγεβρας, συγκεκριμένα ότι vec("Αθήνα") – vec("Ελλάδα") + vec("Ιταλία") περίπου ισούται με vec("Ρώμη"). Αυτό που παρατηρήθηκε είναι ότι με εκπαίδευση σε μεγάλα σύνολα δεδομένων πολλών λέξεων είναι δυνατό να βρεθούν λεπτές σημασιολογικές σχέσεις μεταξύ των λέξεων.

Τέτοιου είδους μοντέλα έχουν το μειονέκτημα ότι αναπαριστούν κάθε λέξη με ένα μοναδικό διάνυσμα, ανεξάρτητα με τα συμφραζόμενα από τα οποία περιβάλλεται κάθε φορά, κάτι το οποίο δεν παρατηρείται στην πραγματικότητα. Νεότερες προσεγγίσεις επιχειρούν να αντιμετωπίσουν αυτό το ζήτημα, παράγοντας διαφορετικά διανύσματα για την εκάστοτε λέξη αναλόγως των συμφραζόμενων ή του σημασιολογικού πλαισίου. Ένα ακόμα μειονέκτημα αυτών των μοντέλων είναι ότι δεν καταφέρνουν να ενσωματώσουν τα στατιστικά όλου του set δεδομένων, αφού κατά την εκπαίδευση, χρησιμοποιούνται μόνο τοπικές γειτονιές λέξεων κάθε φορά.

GloVe

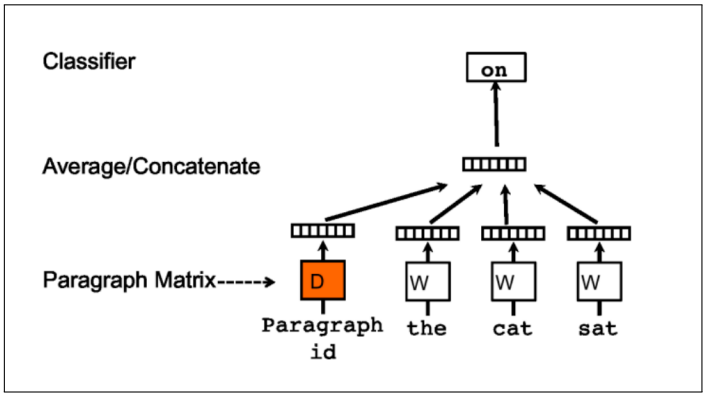
Οι στηριζόμενες σε παράθυρα μέθοδοι (window-based methods) αδυνατούν να αξιοποιήσουν τη στατιστική πληροφορία του σώματος κειμένων, καθώς δεν αξιοποιούν τα στατιστικά στοιχεία συνεμφάνισης των λέξεων σε ολόκληρο το σώμα. Αντιθέτως, σαρώνουν ΄παράθυρα΄ περιεχομένου σε όλη την έκταση του σώματος, με αποτέλεσμα να μην εκμεταλλεύονται τον τεράστιο όγκο επανάληψης των δεδομένων. Τα μοντέλα CBOW και SkipGram στηρίζονται στη χρήση παραθύρων και έτσι αδυνατούν να συλλέξουν σφαιρική πληροφορία. Ο αλγόριθμος GloVe (GlobalVectors) [16] είναι μια βελτίωση του Word2vec σχεδιασμένη να καλύψει αυτή την αδυναμία. Πρόκειται για ένα μοντέλο ελαχίστων τετραγώνων (least squares) το οποίο εκπαιδεύεται χρησιμοποιώντας τις καθολικές μετρήσεις συνεμφάνισης δύο λέξεων. Η βασική ιδέα της μεθόδου είναι να εστιάσει στις πιθανότητες συνεμφάνισης των λέξεων στο πλαίσιο ενός σώματος κειμένων. Με άλλα λόγια, ο αλγόριθμος ελέγχει κατά πόσο συχνά μία λέξη j εμφανίζεται στις λέξεις που περιβάλλουν τη λέξη i, λαμβάνοντας υπόψιν ολόκληρο το σώμα κειμένων.

Doc2Vec

Ο paragraph vectors [17] ή εναλλακτικά Doc2Vec, είναι ένας αλγόριθμος μη επιβλεπομένης μάθησης που χρησιμοποιείται για την παράγωγη ενσωματώσεων (διανυσμάτων σταθερού μήκους) εγγράφων. Ένα σημαντικό χαρακτηριστικό του αλγορίθμου είναι ότι δέχεται ως είσοδο έγγραφα οποιουδήποτε μεγέθους, από φράσεις, προτάσεις, μέχρι ολόκληρα έγγραφα. Η εκπαίδευση του μοντέλου γίνεται με δύο τρόπους. Ο πρώτος είναι παρόμοιος με αυτόν που παρουσιάστηκε στο μοντέλο Word2vec όπου στόχος είναι η πρόβλεψη μια λέξης σε μια παράγραφο δοσμένων των διανυσματικών αναπαραστάσεων γειτονικών λέξεων ως εισόδων. Εδώ όμως, στην είσοδο εισάγεται κι ένα επιπλέον ΄διάνυσμα παραγράφου΄ που καλείται να συμβάλλει στην παραπάνω πρόβλεψη, ενσωματώνοντας παράλληλα το γενικότερο νόημα της παραγράφου

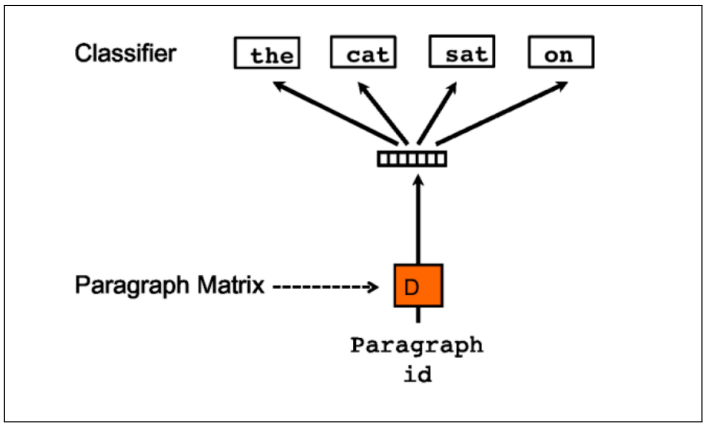
που βρίσκεται η εκάστοτε λέξη.

Το μοντέλο αυτό φέρει το όνομα Μοντέλο Διανεμημένης Μνήμης Διανυσμάτων Παραγράφου (distributed memory model of paragraph vectors, PV-DM). Η γειτονιά λέξεων είναι σταθερού μήκους, ενώ το διάνυσμα της εκάστοτε παραγράφου είναι κοινό για όλες τις λέξεις-στόχους της ίδιας παραγράφου, και διαφέρει μόνο σε λέξεις που ανήκουν σε άλλες παραγράφους. Η εκπαίδευση γίνεται με την κλασική μέθοδο στοχαστικής κατάβασης κλίσης (stochastic gradient descent). Κατά το στάδιο του συμπερασμού, πρέπει να υπολογιστεί το διάνυσμα της παραγράφου εισόδου, την οποία στη γενική περίπτωση δεν έχει συναντήσει το μοντέλο στη διάρκεια της εκπαίδευσης και ο υπολογισμός γίνεται, επίσης, με τον ίδιο τρόπο (gradient descent), μόνο που εδώ, τα βάρη του τελικού επιπέδου softmax και τα διανύσματα των λέξεων διατηρούνται σταθερά.



Σχήμα 3.9: Εκπαίδευση του μοντέλου PV-DM. Τα διανύσματα των γειτονικών λέξεων και της παραγράφου συμβάλλουν στην πρόβλεψη της λέξης στόχου.

Η άλλη μορφή που υιοθετήθηκε για το μοντέλο ονομάζεται ¨Διανεμημένος σάκος λέξεων' (distributed bag of words of paragraph vector, PV-DBOW). Με αυτή τη μέθοδο, πραγματοποιείται μια αντίστροφη διαδικασία, όπου το σύστημα καλείται να προβλέψει μια γειτονιά λέξεων με είσοδο το διάνυσμα μιας τυχαίας λέξης της γειτονιάς και του διανύσματος της παραγράφου, μέσα από μια διαδικασία ταξινόμησης.



Σχήμα 3.10: Εκπαίδευση του μοντέλου PV-DBOW.

Το μοντέλο paragraph vectors παρουσίασε καλύτερα αποτελέσματα σε διάφορες εργασίες από απλούστερα μοντέλα CBOW, εφόσον λαμβάνει υπόψη τη σειρά των λέξεων σε μια πρόταση, καταφέρνοντας να εντοπίσει σημασιολογικές συσχετίσεις σε αυτές. Στα set δεδομένων που δοκιμάστηκε το μοντέλο PV-DM μόνο του, φάνηκε να τα πηγαίνει πολύ καλά, αλλά το πλήρες μοντέλο που το συνδυάζει με το PV-DBOW (παράθεση των διανυσμάτων που παράγει το κάθε υπομοντέλο) φαίνεται να έχει μεγαλύτερη συνέπεια στις επιδόσεις σε μεγάλη ποικιλία από set δεδομένων και προβλήματα.[19]

3.3.5 Transformers και δυναμικές ενσωματώσεις

Το μοντέλο του transformer

Transformer ονομάζεται είναι ένας τύπος αρχιτεκτονικής νευρωνικών δικτύων βαθιάς μάθησης που πρωτοδημοσιεύτηκε από τους ἄσωανι ετ αλ [18].

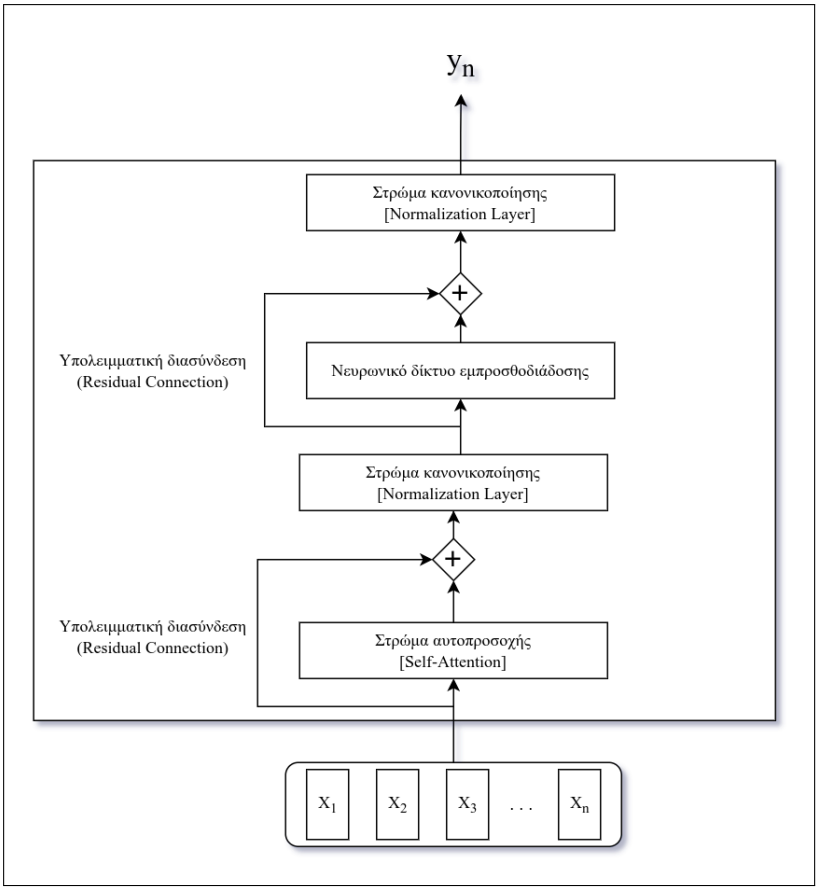
Η απλή αρχιτεκτονική transformer αποτελείται από μια διάταξη από μονάδες, όπου η κάθε μια είναι ένα πολυεπίπεδο δίκτυο επεξεργασίας συνδυάζοντας γραμμικούς μετασχηματισμούς, απλά νευρωνικά δίκτυα ανατροφοδότησης (feed forward) και κυρίως τη βασική καινοτομία του transformer, τη χρήση στρωμάτων επεξεργασίας που εφαρμόζουν μηχανισμούς αυτοπροσοχής (self-attention mechanisms). Αυτές οι μονάδες ονομάζονται ΄transformer blocks΄. Υπάρχουν διάφοροι τρόποι να συσταθούν αυτές οι μονάδες και έχουν υλοποιηθεί διάφορα μοντέλα. Ο αρχικός transformer χρησιμοποιούσε μια αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή για τους σκοπούς της μετάφρασης.

Δεδομένης μιας ακολουθίας εισόδου διανυσμάτων σταθερής αναπαράστασης $(x_1, ..., x_n)$, μια μονάδα transformer θα επεξεργαστεί αυτή την είσοδο παράγοντας μια ακολουθία διανυσμάτων εξόδου $(y_1...y_n)$ ίδιου μήκους. Σε διαφορά με άλλα δίκτυα επεξεργασίας ακολουθιών, όπως τα αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks) ή τα LSTM (Long-Short Term Networks) το transformer δέχεται ολόκληρη την ακολουθία εισόδου ταυτόχρονα.

Μηχανισμός Attention

Στην καρδιά μιας προσέγγισης που βασίζεται σε μηχανισμούς attention είναι η δυνατότητα **σύγκρισης** ενός στοιχείου ενδιαφέροντος με ένα σύνολο άλλων στοιχείων με τρόπο που να αποκαλύπτει τη συνάφειά τους στο τρέχον σημασιολογικό πλαίσιο. Στην περίπτωση της αυτοπροσοχής, οι συγκρίσεις γίνονται μεταξύ ενός στοιχείου με τα άλλα στοιχεία μιας δεδομένης ακολουθίας. Το αποτέλεσμα αυτών των συγκρίσεων χρησιμοποιείται στη συνέχεια για τον υπολογισμό της εξόδου για την τρέχουσα είσοδο. Ανάλογα με το αν η προσοχή μπορεί να βρίσκεται σε όλες τις μονάδες της ακολουθίας ή μόνο στης μονάδες πριν η προσοχή μπορεί να είναι άιτιοκρατική΄ (causal) ή ΄μη αιτιοκρατική΄. Τα διαφορετικά αυτά είδη προσοχής χρησιμοποιούνται για μοντέλα με διαφορετικούς σκοπούς, ανάλογα με το αν χρειάζεται να εμπλουτίζεται με πληροφορία από όλη την ακολουθία ή μόνο γνωρίζοντας ότι έχει ήδη εισαχθεί.

Το πρώτο βήμα μιας στρώσης attention είναι η δημιουργία ενός ΄πίνακα συσχέτισης΄. Τα περιεχόμενα του πίνακα αποτελούνται από τιμές που εκφράζουν την σχέση του κάθε διανύσμα-



Σχήμα 3.11: Ένα transformer block

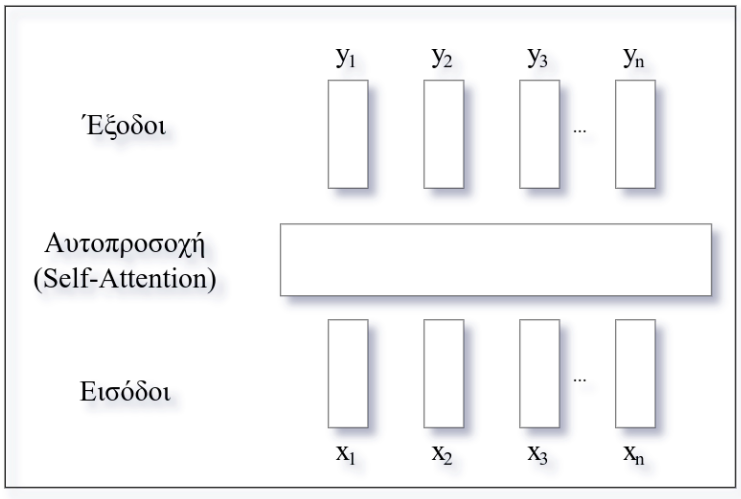
τος εισόδου με όλα τα υπόλοιπα (δίνοντας του μέγεθος $N \times N$), χρησιμοποιώντας μετρικές ομοιότητας διανυσμάτων. Συγκεκριμένα, υπολογίζουμε την ομοιότητα του κάθε διανύσματος με τα υπόλοιπα στην ακολουθία εισόδου εφαρμόζοντας το εσωτερικό γινόμενο

$$a'_{ij} = x_i^T x_j$$

Έτσι για κάθε διάνυσμα, έχουμε ένα νέο διάνυσμα που σε κάθε του διάσταση περιγράφεται η σχέση του με τα άλλα. Το αποτέλεσμα του εσωτερικού γινομένου είναι μια πραγματική τιμή, η οποία όσο μεγαλύτερη είναι τόσο πιο όμοια είναι τα διανύσματα που συγκρίνονται. Στην πράξη, το εσωτερικό γινόμενο διαιρείται στη συνέχεια με την τετραγωνική ρίζα της διάστασης των εισόδων. Αυτό διασφαλίζει ότι η είσοδος και η έξοδος της λειτουργίας αυτοπροσοχής έχουν παρόμοια διακύμανση, βοηθώντας την εκπαίδευση τους αργότερα.

$$a'_{i,j} = \frac{x_i^T x_j}{\sqrt{k}}$$

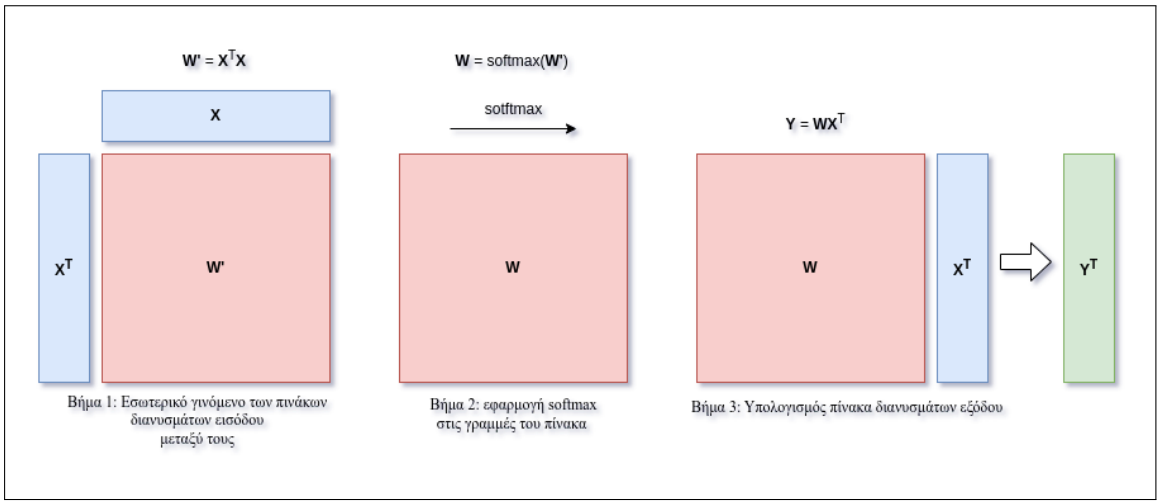
Για να χρησιμοποιήσουμε αποτελεσματικά αυτές τις βαθμολογίες, θα τις κανονικοποιήσουμε με την εφαρμογή της συνάρτησης $softmax$ για τη δημιουργία ενός διανύσματος βαρών, a_{ij}, που εκφράζει την σχετική συνάφεια κάθε εισόδου με το στοιχείο εισόδου i που είναι η



Σχήμα 3.12: Επίπεδο Αυτοπροσοχής

τρέχουσα εστίαση της προσοχής.

$$a_{ij} = \frac{e^{w'_{ij}}}{\sum e^{w'_{ij}}}$$



Σχήμα 3.13: Πίνακες Προσοχής

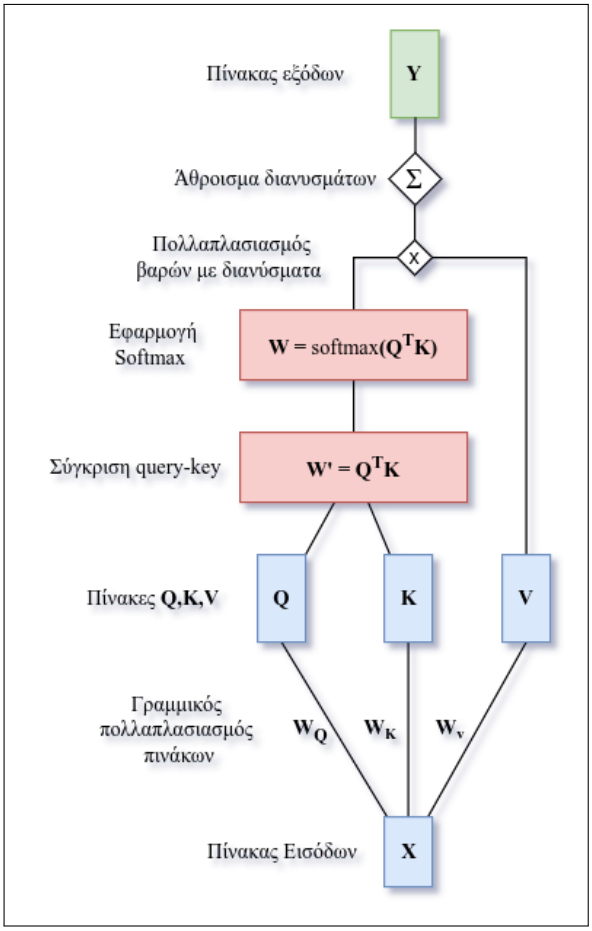
Εκπαίδευση απλής αυτοπροσοχής

Υπάρχουν τρεις διαφορετικοί ρόλοι που αναλαμβάνει κάθε ενσωμάτωση εισόδου κατά τη διάρκεια της επεξεργασίας σε ένα στρώμα αυτοπροσοχής. Αρχικά, ως το ΄κέντρο της προσοχής΄, όταν συγκρίνεται με όλες τις υπόλοιπες εισόδους και ο ρόλος αναφέρεται ως ’ερώτημα’ (query). Σαν δεύτερο ρόλο, ως ένα από τα μέρη της εισόδου που συγκρίνεται με το τρέχων ΄κέντρο της προσοχής΄, και αναφέρεται ως ’κλειδί’ (key) και τέλος ως μια τιμή η οποία χρησιμοποιείται για τον τελικό υπολογισμό της εξόδου του τρέχοντος κέντρου της προσοχής,

αναφερόμενο ως ΄τιμή΄ (value). Για να διατελέσουν αυτούς τους τρεις διαφορετικούς ρόλους, οι μετασχηματιστές προσθέτουν ενα επίπεδο γραμμικού μετασχηματισμού εισάγοντας εκπαιδεύσιμους πίνακες βαρών $\mathbf{W_Q}$, $\mathbf{W_K}$ και $\mathbf{W_V}$ με διαστάσεις ίδιες με το μέγεθος της εισόδου .

Αυτά τα βάρη θα χρησιμοποιηθούν για τον μετασχηματισμό του κάθε διανύσματος εισόδου x_i σε μια αναπαράσταση του ρόλου του ως κλειδιού, ερωτήματος ή τιμής.

Ο μηχανισμός προσοχής λειτουργεί με τον υπολογισμό του εσωτερικού γινόμενου του 'ερωτήματος'(query) και των διανυσμάτων κλειδιών (key) για κάθε ζευγάρι διανυσμάτων από την ακολουθία εισόδου.



Σχήμα 3.14: Scaled dot-product attention

Έπειτα, περνά από μια συνάρτηση $softmax$ για να ληφθούν τα βάρη για κάθε ζευγάρι διανυσμάτων. Τα βάρη προσοχής υποδεικνύουν πόσο πρέπει να ΄μετρήσει΄ κάθε λεκτική μονάδα στην τελική έξοδο. Στη συνέχεια, οι τιμές χρησιμοποιούνται για τη κλιμάκωση των διανυσμάτων ΄τιμών΄ και το κλιμακωμένο άθροισμα των διανυσμάτων τιμών χρησιμοποιείται για να ληφθεί η έξοδος του κάθε επιπέδου προσοχής.

Τελικά, η έξοδος του στρώματος προσοχής διατυπώνεται ως μια συνάρτηση των πινάκων

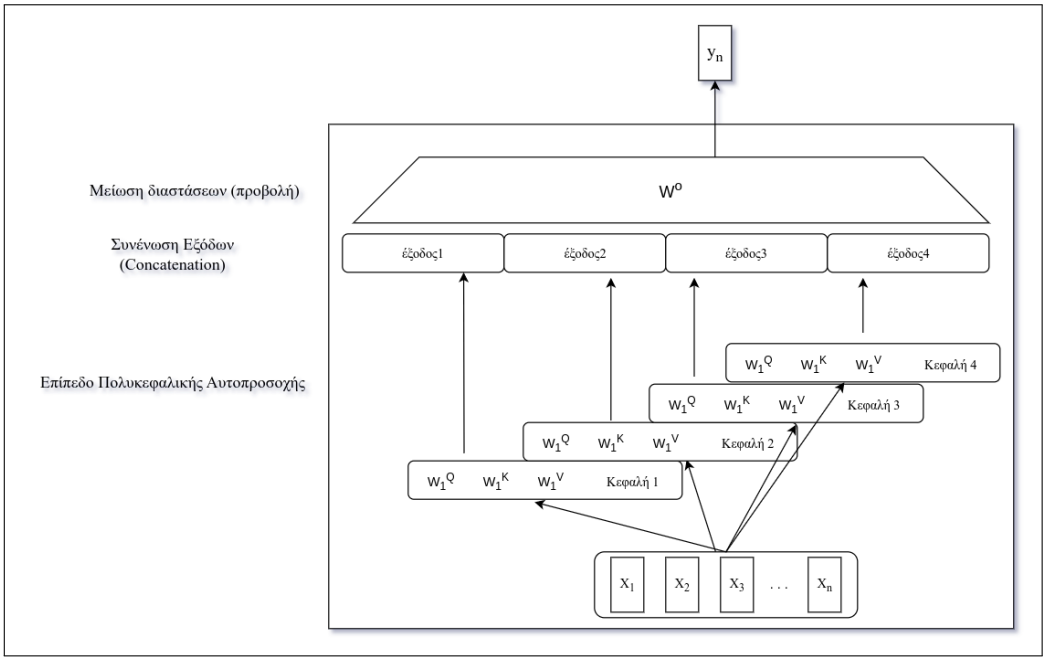
ερωτημάτων, κλειδιών και τιμών:

$$Attention(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V$$

Η χρήση μηχανισμών προσοχής στην αρχιτεκτονική του transformer επιτρέπει στο μοντέλο να επεξεργάζεται αποτελεσματικά μεγάλες ακολουθίες εισόδου εξάγοντας μια πλούσια αναπαράσταση των νοηματικών εξαρτήσεων μεταξύ τους. Αυτό έχει καταστήσει τους transformer ιδιαίτερα χρήσιμους για επεξεργασία φυσικής γλώσσας, όπου η μορφή των τελικών αναπαραστάσεων της έκφρασης εξαρτάται από τις συντακτικές, γραμματικές και νοηματικές σχέσεις μεταξύ των λέξεων.

Multi-head attention

Θα ήταν δύσκολο έως αδύνατο ένα μοναδικό transformer block να συλλάβει όλα τα είδη παράλληλων νοημάτων και συσχετίσεων μεταξύ των εισόδων του. Για τη λύση αυτού του προβλήματος, οι transformers χρησιμοποιούν στρώσεις βασισμένες στην αρχιτεκτονική της πολυ-κεφαλική αυτοπροσοχής (multihead self-attention). Είναι συστάδες από στρώσεις αυτοπροσοχής, τις κεφαλές, που βρίσκονται σε παραλληλία στο ίδιο βάθος του μοντέλου, η κάθε μια με τις δικές της εσωτερικές παραμέτρους. Κάθε κεφαλή μπορεί να μαθαίνει διαφορετικές απόψεις των συσχετίσεων που υπάρχουν μεταξύ των εισόδων. Για την εφαρμογή αυτής της ιδέας, κάθε ένα από τα επίπεδα αυτοπροσοχής πολλαπλών κεφαλών παρέχεται με τους δικούς του πίνακες βαρών κλειδιού, ερωτήματος και τιμής :$\mathbf{W_i^Q}$, $\mathbf{W_i^K}$ και $\mathbf{W_i^V}$.



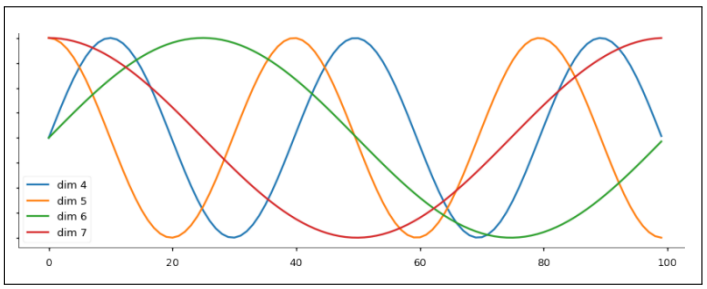
Σχήμα 3.15: Multi-head attention

Η διαδρομή της πληροφορίας ξεκινάει από την είσοδο μίας ακολουθίας λεκτικών μονάδων στο μοντέλο. Στη συνέχεια, τα διανύσματα των λεκτικών μονάδων μπαίνουν παράλληλα σε

h στρώσεις και οι έξοδοι αυτών συνδυάζονται στη στρώση μείωσης διαστάσεων. Συνεπώς συνολικά έχουμε $(3 \cdot \mathrm{h})$ Linear μπλοκ και h στρώσεις scaled dot-product attention, όπου το h θεωρείται υπέρ-παράμετρος. Οι έξοδοι από κάθε ένα από τα στρώματα ενώνονται και στη συνέχεια προβάλλονται στο d, παράγοντας έτσι μια έξοδο του ίδιου μεγέθους με την είσοδο, ώστε τα στρώματα να μπορούν να στοιβάζονται.

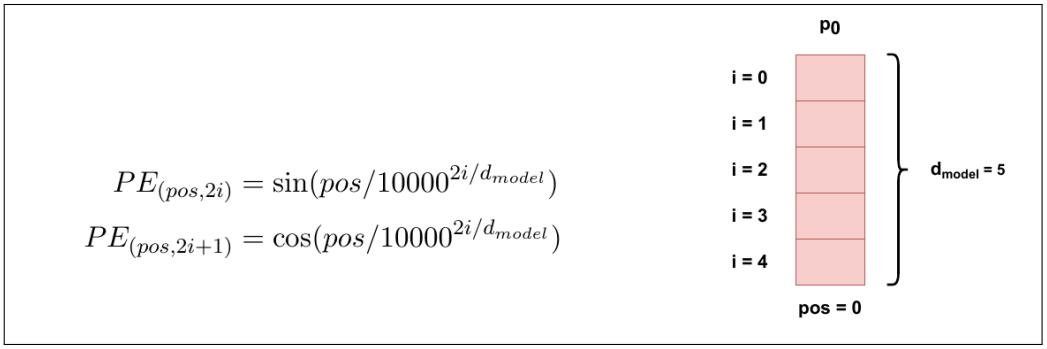
Ενσωμάτωση Θέσης - Positional Encoding

Με βάση τα παραπάνω έγινε κατανοητό ότι ο transformer και πιο συγκεκριμένα το μοντέλο του κωδικοποιητή (που κυρίως μας απασχολεί) δέχεται τις λεκτικές μονάδες εισόδου όλες μαζί ταυτόχρονα. Το γεγονός αυτό υπονοεί ότι το μοντέλο δεν έχει ιδέα για τη σειρά των εισερχομένων λέξεων. Συνεπώς, πρέπει με κάποιο τρόπο τα διανύσματα των λεκτικών μονάδων εισόδου πριν εισαχθούν στο μοντέλο να περάσουν από μια επεξεργασία, ώστε τα διανύσματα να αποκτήσουν και συντακτικές πληροφορίες. Αυτή η διαδικασία ονομάζεται κωδικοποίηση θέσης και εφαρμόζεται προσθέτοντας σε κάθε διάνυσμα της ακολουθίας εισόδου ένα ίσου μεγέθους διάνυσμα θέσης (positional encoding). Στον αρχικό transformer του χρησιμοποιήθηκε ένας συνδυασμός από τριγωνομετρικές συναρτήσεις ημίτονου και συνημιτόνου με διαφορετικές συχνότητες.



Σχήμα 3.16: Ενσωμάτωση θέσης με τριγωνομετρικές συναρτήσεις

Για τη λεκτική μονάδα στην θέση pos και με διάσταση διανύσματος d, η κωδικοποίηση θέσης ορίζεται με τους τύπους που φαίνονται στο σχήμα, ανάλογα με το αν η θέση του είναι κατ' αντιστοιχία ζυγός η περιττός αριθμός.



Σχήμα 3.17: Συναρτήσεις ενσωμάτωσης θέσης

WordPiece

Οι tokenizers είναι αλγόριθμοι κατάτμησης οι οποίοι χωρίζουν το κείμενο, τις προτάσεις ή γενικότερα ένα σύνολο δεδομένων σε μικρότερες επιμέρους λεκτικές μονάδες. Αυτές οι μονάδες είναι οι είσοδοι του μοντέλου. Το μοντέλο tokenizer που χρησιμοποιεί το BERT είναι το WordPiece [19]. Το WordPiece ανήκει στην κατηγορία των αλγόριθμων κατάτμησης σε υπό-λέξεις (sub-word tokenization algorithms). Οι αλγόριθμοι αυτοί βασίζονται στην ιδέα ότι σε ένα σύνολο δεδομένων κειμένου, οι πιο συχνά χρησιμοποιούμενες λέξεις δεν πρέπει να διαχωρίζονται, ενώ οι λέξεις που χρησιμοποιούνται πιο σπάνια πρέπει να διασπώνται σε μικρότερες ακολουθίες (wordpieces) από τα συστατικά τους μέρη για ευκολότερη ερμηνεία τους από το μοντέλο. Ο αρχικός σκοπός ανάπτυξης του WordPiece ήταν η λύση του προβλήματος διαχωρισμού ιαπωνικών και κορεάτικων δεδομένων από φράσεις σε συστήματα αναγνώρισης ομιλίας. Το βασικό πρόβλημα με αυτές τις γλώσσες είναι ότι υπήρχαν πολλές νέες λέξεις τις οποίες έβλεπε για πρώτη φορά το μοντέλο που χρησιμοποιούνταν καθώς δεν υπήρχαν στο λεξιλόγιο του. Ένα παράδειγμα διαχωρισμού λέξης με τον WordPiece είναι η λέξη 'χιονόμπαλα' η οποία (αν υποθέσουμε ότι είναι μία σπάνια λέξη) μπορεί να χωριστεί στα τμήματα ΄χιόνο΄, ΄μπάλα΄. Ο ειδικός χαρακτήρας της κάτω παύλας δηλώνει την αρχή της λέξης. Με την παραπάνω διαμέριση το μοντέλο έχει τη δυνατότητα να αποφασίσει πιο εύκολα για το νόημα της άγνωστης λέξης ΄χιονόμπαλα΄ με βάση τα συνθετικά της μέρη. Ο WordPiece κατευθύνεται αποκλειστικά από τα δεδομένα και εγγυάται ντετερμινιστικό διαχωρισμό για οποιαδήποτε πιθανή ακολουθία χαρακτήρων [20].

Η εκπαίδευση του WordPiece μοντέλου γίνεται άπληστα (greedy) με τα παρακάτω βήματα όπως περιγράφουν οι (Schuster and Nakajima) [19].

1. Αρχικοποίηση του λεξιλογίου με όλους τους βασικούς χαρακτήρες unicode συμπεριλαμβανομένων και του ASCII.

2. Δημιουργία ενός γλωσσικού μοντέλου πάνω στα δεδομένα εκπαίδευσης χρησιμοποιώντας το λεξιλόγιο από το βήμα 1.

3. Δημιουργία ενός συνδυασμού χαρακτήρων χρησιμοποιώντας δύο συνδυασμούς από το ήδη υπάρχον λεξιλόγιο και προσθήκη αυτού στο λεξιλόγιο. Ο συνδυασμός που επιλέγεται για να προστεθεί, είναι αυτός που αυξάνει την πιθανοφάνεια στα δεδομένα εκπαίδευσης (όχι ο πιο συχνός) όταν προστεθεί στο μοντέλο.

4. Επανάληψη από το βήμα 2 και παρακάτω μέχρι να δημιουργηθεί ένα λεξιλόγιο με το επιθυμητό μέγεθος λεκτικών μονάδων (tokens) ή μέχρι η πιθανοφάνεια των δεδομένων να πέσει κάτω από ένα προκαθορισμένο κατώφλι.

3.4 Αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται για εκπαίδευση

Σε αυτό το σημείο θα γίνει αναφορά στις διαδικασίες και μεθοδολογίες εκπαίδευσης των μοντέλων που θα χρησιμοποιηθούν στα πλαίσια της εργασίας για την ταξινόμηση των εξα-

γόμενων χαρακτηριστικών κειμένου. Η επιλογή των αλγορίθμων μηχανικής μάθησης έχει γίνει έτσι ώστε να περιληφθούν οι διάφορες μεθοδολογίες ταξινόμησης.

3.4.1 Naive Bayes

Ο αλγόριθμος Naive Bayes (Αφελής Bayes) είναι ένας πιθανολογικός ταξινομητής, δηλαδή δίνει μια πιθανότητα για κάθε κλάση που θα μπορούσε να ανήκει το κάθε δείγμα. Δέχεται σαν είσοδο ένα μοντέλο Bag-of-words, ένα σύνολο μοναδικών, μη στοιχισμένων συντακτικά λέξεων στο οποίο κρατείται για κάθε λέξη μόνο η συχνότητα της στο κείμενο.

Είναι πιθανολογικός ταξινομητής, δηλαδή για ένα έγγραφο d για όλες τις κλάσεις $c \in C$ επιστρέφει την κλάση $\hat{c}$ με την μεγαλύτερη πιθανότητα να ταιριάζει στο έγγραφο.

$$\hat{c} = \text{argmax} P(c|d) \tag{3.4}$$

Χρησιμοποιεί τον κανόνα Bayes , ο οποίος μας επιτρέπει να σπάσουμε οποιαδήποτε δεσμευμένη πιθανότητα (conditional probability) P(x|y) σε τρεις απλούστερες πιθανότητες.

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{3.5}$$

Στην περίπτωση της ταξινόμησης εγγράφων, ο παραπάνω κανόνας απλοποιείται καθώς η πιθανότητα για ένα συγκεκριμένο έγγραφο d παραμένει ίδια. Τελικά, η πιθανότητα για ένα έγγραφο να ανήκει σε μια κλάση ορίζεται ως:

$$\hat{c} = \underset{c \in C}{\text{argmax}} P(c|d) = \underset{c \in C}{\text{argmax}} P(d|c)P() \tag{3.6}$$

Η ιδέα της ταξινόμησης με χρήση του ταξινομητή bayes μπορεί να εκφραστεί ως ¨Από όλες τις εμφανίσεις αυτής της λέξης, σε όλα τα κείμενα, σε τί ποσοστό εμφανίζεται στα κείμενα αυτής της κατηγορίας·'

Υπολογίζουμε την πιο πιθανή κατηγορία c για κάποιο έγγραφο d επιλέγοντας το μεγαλύτερο αποτέλεσμα των γινομένων δύο πιθανοτήτων, της προηγούμενης (prior probability) της κατηγορίας c να εμφανιστεί και της πιθανότητας (likelihood) του εγγράφου να ανήκει σε αυτή την κατηγορία $P(d|c)$.

Αναπαριστούμε το έγγραφο d ως ένα σύνολο από χαρακτηριστικά, με κάθε λέξη να είναι και ένα χαρακτηριστικό.

$$c = argmax P(f1, fn|c) \cdot P(c)$$

Σε αυτό το σημείο κάνουμε δυο απλουστευτικές υποθέσεις για τη φύση των δεδομένων: Η πρώτη είναι η υπόθεση του ¨σακιού των λέξεων¨, άρα θεωρούμε ότι τα χαρακτηριστικά $f_1, f_2 ... f_n$ χαρακτηρίζουν μόνο την ταυτότητα της λέξης και όχι τη θέση της, και η δεύτερη είναι ότι θεωρούμε οι πιθανότητες εμφάνισης των χαρακτηριστικών f είναι ανεξάρτητες η μία από την άλλη δεδομένης της κατηγορίας c, ώστε να μπορούμε να τις συνδυάσουμε μέσω πολλαπλασιασμού.

Εφαρμόζοντας έναν δείκτη i από 1 μέχρι τον αριθμό των λέξεων του εγγράφου n, για κάθε λέξη w στο έγγραφο ως τα χαρακτηριστικά του κειμένου έχουμε:

$$P(f_1, f_2, f_3...f_n|c) = P(f_1|c) \cdot P(f_2|c)...P(f_n|c) = \prod_1^n P(w_i|c) \tag{3.7}$$

Εφόσον τα χαρακτηριστικά ενός κειμένου μεγέθους i είναι οι λέξεις $w_1..w_i$, η απόφαση για την κλάση που ανήκει ένα παράδειγμα ορίζεται από την παρακάτω εξίσωση:

$$\text{argmax}P(c) \cdot \prod P(w_i|c) \tag{3.8}$$

Εκπαίδευση Naive Bayes

Για την εκπαίδευση του αλγορίθμου Naive Bayes χρησιμοποιούμε συχνότητες που απαντώνται στα δεδομένα μας. Για την προηγούμενη πιθανότητα κλάσης $P(c)$ (prior) υπολογίζουμε τον λόγο των εγγράφων της κάθε κλάσης ως προς τον αριθμό του συνόλου των εγγράφων.

$$P(c) = \frac{N_c}{N_{doc}} \tag{3.9}$$

Για την πιθανότητα του κάθε χαρακτηριστικού $P(f_i|c)$ θεωρούμε ότι τα χαρακτηριστικά f είναι λέξεις στο σύνολο bag of words του συγκεκριμένου εγγράφου, άρα $P(w_i|c)$ που υπολογίζεται ως ο λόγος της συχνότητας εμφάνισης της λέξης w_i σε όλα τα έγγραφα της κλάσης c. Ενώνουμε όλα τα έγγραφα της κατηγορίας c σε ένα μεγάλο κείμενο και καταμετρούμε την συχνότητα της συγκεκριμένης λέξης. Έπειτα χρησιμοποιούμε αυτή τη συχνότητα για να υπολογίσουμε την πιθανότητα του όρου:

$$\hat{P}(w_i|c) = \frac{count(w_i, c)}{\sum_{w \in V}^n count(w, c)} \tag{3.10}$$

Εδώ εμφανίζεται το πρόβλημα ότι λέξεις οι οποίες δεν εμφανίζονται στο λεξιλόγιο αυτής της κλάσης θα αποκλείουν την ανάθεση της σε αυτή αφού έχουν πιθανότητα 0 όταν πολλαπλασιάζονται με τις υπόλοιπες. Για την αποφυγή αυτής της περίπτωσης, είτε οι νέες λέξεις αφαιρούνται, είτε γίνεται μια τεχνητή ομαλοποίηση των λεξικών με πρόσθεση συχνοτήτων τέτοιων λέξεων, γνωστή και ως Laplace Smoothing. Άλλη μια παρέμβαση για τη βελτίωση του μοντέλου είναι η αφαίρεση των συχνών λέξεων από το λεξιλόγιο (stopwords) εφόσον μπορεί να δυσκολέψουν την ταξινόμηση.

3.4.2 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση (logistic regression) είναι ένα πιθανολογικό μοντέλο ταξινόμησης. Είναι ένας αλγόριθμος εποπτευόμενης μάθησης που λαμβάνει ένα σύνολο δεδομένων και κάνει προβλέψεις σχετικά με το αποτέλεσμα μιας εξαρτημένης μεταβλητής με βάση τις τιμές μιας ή περισσότερων ανεξάρτητων μεταβλητών. Η εξαρτημένη μεταβλητή στην λογιστική παλινδρόμηση είναι δυαδική, που σημαίνει ότι μπορεί να λάβει μόνο δύο πιθανές τιμές, 0 ή 1. Είναι

ένα παράδειγμα διαχωριστικού μοντέλου και στο σενάριο της ταξινόμησης κειμένου επιχειρεί να υπολογίσει την πιθανότητα της κλάσης το παράδειγμα δ να ανήκει στην κλάση c, $P(c|d)$.

Όντας πιθανολογικός ταξινομητής επιβλεπομένης μάθησης, ο αλγόριθμος της λογιστικής παλινδρόμησης χρειάζεται ένα σύνολο δεδομένων εκπαίδευσης από n ζεύγη εισόδων - εξόδων (x_i, y_i). Τα τέσσερα μέρη που αποτελούν ένα σύστημα μηχανικής μάθησης για ταξινόμηση είναι:

- Μια αναπαράσταση των χαρακτηριστικών της εισόδου: Για κάθε παράδειγμα x, θα έχουμε ένα διάνυσμα χαρακτηριστικών $[x_1, x_2, x_3...x_n]$
- Μια συνάρτηση ταξινόμησης η οποία υπολογίζει την έξοδο $\hat{y}$, την κλάση που υπολογίστηκε ότι ανήκει το παράδειγμα.
- Μια αντικειμενική συνάρτηση για τη μάθηση (objective function), με σκοπό την ελαχιστοποίηση του σφάλματος στα παραδείγματα εκπαίδευσης. Για τη λογιστική παλινδρόμηση χρησιμοποιείται η συνάρτηση απώλειας cross-entropy.
- Έναν αλγόριθμο για την βελτίωση της συνάρτησης αξιολόγησης, όπως ο Stochastic Gradient Descent.

Η λογιστική παλινδρόμηση αποτελείται από δυο στάδια: εκπαίδευση: Μεταβάλλουμε τις εσωτερικές παραμέτρους του μοντέλου (τα βάρη w και b) με τη χρήση Stochastic Gradient Descent σε συνδυασμό με έλεγχος: Δεδομένου παραδείγματος ξ υπολογίζουμε την πιθανότητα $P(y|x)$ και επιστρέφουμε την κλάση με την μεγαλύτερη πιθανότητα $y = 1$ ή $y = 0$.

Εκπαίδευση λογιστικής Παλινδρόμησης

Σκοπός του αλγορίθμου είναι να δημιουργηθεί ένα μοντέλο που προβλέπει την πιθανότητα ότι ένα νέο παράδειγμα ανήκει σε μια συγκεκριμένη κλάση. Δεδομένης μίας εισόδου x με χαρακτηριστικά $(x_1, x_2, ...x_n)$ η έξοδος του ταξινομητή μπορεί να είναι είτε 1, δείχνοντας ότι το παράδειγμα ανήκει σε μια συγκεκριμένη κλάση, είτε 0, ότι δεν ανήκει. Ψάχνουμε την πιθανότητα $P(y = 1|x)$ ότι η είσοδος ανήκει στην κλάση.

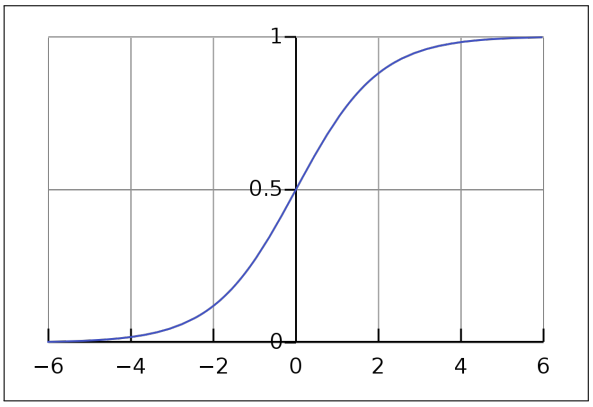
Για να απαντήσει σε αυτό το πρόβλημα, η λογιστική παλινδρόμηση μαθαίνει, από ένα σύνολο δεδομένων εισόδου, ένα διάνυσμα βαρών και ένα bias term (παράγοντας μεροληψίας). Κάθε βάρος w είναι ένας πραγματικός αριθμός και αντιστοιχίζεται σε ένα χαρακτηριστικό εισόδου x_i . Το κάθε βάρος αναπαριστά το βαθμό που το κάθε το κάθε χαρακτηριστικό επηρεάζει την τελική επιλογή της κλάσης. Ο παράγοντας bias είναι επίσης ένας πραγματικός αριθμός ο οποίος προστίθεται στο άθροισμα των βαρυνομένων εισόδων.

Για να αποφανθεί για την κλάση ενός παραδείγματος, μετά την εκπαίδευση, ο ταξινομητής πολλαπλασιάζει κάθε χαρακτηριστικό εισόδου x_i με το βάρος του w_i, αθροίζει τα γινόμενα και προσθέτει τον όρο μεροληψίας. Ο αριθμός που προκύπτει από αυτή τη διαδικασία εκφράζει πόσο ο ταξινομητής εκτιμά ότι το παράδειγμα πρέπει να ανήκει, ή όχι, στην κλάση.

$$z = \sum_{1}^{i} x_i w_i \tag{3.11}$$

Για να μετατραπεί αυτός ο πραγματικός αριθμός σε πιθανότητα, περνάει σαν είσοδος στην λογιστική συνάρτηση ή σιγμοειδή. Η λογιστική συνάρτηση, είναι μια καμπύλη σχήματος S που αντιστοιχίζει οποιονδήποτε πραγματικό αριθμό σε μια τιμή μεταξύ 0 και 1. Αυτό την καθιστά κατάλληλη για την πρόβλεψη πιθανοτήτων. Η προβλεπόμενη πιθανότητα μπορεί στη συνέχεια να μετατραπεί σε δυαδική έξοδο θεωρώντας όλες τις τιμές πάνω από ένα συγκεκριμένο κατώφλι σε μια κλάση και όλες τις τιμές κάτω από το κατώφλι στην άλλη.

$$P(y = 1) = \sigma(w \cdot x + b) = \frac{1}{1 + e^{-(z)}} \tag{3.12}$$



Σχήμα 3.18: Σιγμοειδής συνάρτηση

Διωνυμική ταξινόμηση

Για να ταξινομήσουμε παραδείγματα, στην απλή περίπτωση της λογιστικής παλινδρόμησης, πρέπει να ορίσουμε ένα κατώφλι ή ΄όριο΄ επιλογής: δεδομένου παραδείγματος x, εάν η πιθανότητα $P(y = 1|x)$ που μας επιστρέφει ο ταξινομητής είναι μεγαλύτερη από μια τιμή (π.χ. 0.5) αναθέτουμε την κλάση 1, ειδάλλως την κλάση 0.

$$\text{κλάση}(\xi) = \begin{cases} 1 & \text{εαν } P(y = 1|x) \geq 0.5 \\ 0, & \text{ειδάλλως} \end{cases}$$

Πολυωνυμική λογιστική παλινδρόμηση

Για να ταξινομήσουμε παραδείγματα σε περισσότερες από δυο κλάσεις, χρησιμοποιούμε την πολυωνυμική λογιστική παλινδρόμηση ή παλινδρόμηση softmax. Σκοπός είναι να αναθέσουμε σε κάθε παράδειγμα x μια κλάση k από ένα σύνολο K κλάσεων.Για την μοντελοποίηση των κλάσεων, ορίζουμε ότι για κάθε είσοδο x η έξοδος y είναι ένα διάνυσμα μεγέθους K. Εάν η κλάση c είναι η σωστή, θέτουμε $y_c = 1$, και όλα τα άλλα στοιχεία του διανύσματος 0. Ο ταξινομητής επιδιώκει να υπολογίσει ένα διάνυσμα εκτιμήσεων κλάσης $\hat{y}$. Για κάθε κλάση k, η τιμή εξόδου y_k είναι ο υπολογισμός της πιθανότητας $P(y_k = 1|x)$ από τον ταξινομητή.

Συνάρτηση Softmax

Ο πολυωνυμικός ταξινομητής χρησιμοποιεί μια γενίκευση της σιγμοειδούς, γνωστή ως softmax, για τον υπολογισμό της $P(y_k = 1|x)$. Δέχεται ως είσοδο ένα διάνυσμα $z = [z_1, z_2, ..., z_K]$ από K πραγματικές τιμές και τις αντιστοιχίζει σε μια κατανομή πιθανοτήτων,με κάθε τιμή ανάμεσα στο 0 και το 1, με άθροισμα όλων των τιμών το 1.

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K$$

Συνάρτηση σφάλματος cross-entropy

Για την αλλαγή των εσωτερικών παραμέτρων του μοντέλου μέσω μάθησης, χρειάζεται ένα μέτρο σύγκρισης του πόσο διαφέρει η πρόβλεψη της κλάσης y από την πραγματική κλάση $\hat{y}$ (απώλεια). Στη λογιστική παλινδρόμηση χρησιμοποιείται η συνάρτηση διασταυρωμένης εντροπίας (cross-entropy loss). Είναι μια συνάρτηση που επιχειρεί να μεγιστοποιήσει την λογαριθμική πιθανότητα των παραδειγμάτων εισόδου να ανήκουν στη σωστή κλάση μεταβάλλοντας τις παραμέτρους w και b.

$$L_{CE}(\hat{y}, y) = -[ylog\hat{y} + (1 - y)log(1 - \hat{y})]$$

Η ελαχιστοποίηση της συνάρτησης cross-entropy γίνεται με αλγορίθμους βελτιστοποίησης όπως η κάθοδος βασισμένη στην κλίση (gradient descent.)

3.4.3 Μηχανές Διανυσμάτων Υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης ή ΜΔΥ (Support Vector Machines, SVM) προτάθηκαν από τον Vladimir Vapnik ως γραμμικοί ταξινομητές, το 1963. Απέκτησαν δημοσιότητα μετά το 1992, όταν ενισχύθηκαν από τον ίδιο και τους συνεργάτες του με το κόλπο του πυρήνα (kernel trick), που επέτρεψε τη χρήση τους και σε μη γραμμικώς διαχωρίσιμα προβλήματα. Οι SVM βασίζονται στη θεωρία στατιστικής μάθησης (statistical learning theory)[16].

Μέχρι πρότινος, πριν την άνθιση των νευρωνικών δικτύων, τα SVM ήταν η βέλτιστη επιλογή για σύνολα δεδομένων πολλών διαστάσεων, με εφαρμογές σε συστήματα αναγνώρισης γραφής (handwriting recognition), ταξινόμησης κειμένων (text categorization) ή ταξινόμησης δεδομένων έκφρασης γονιδίων (gene expression data).

Στην απλούστερη περίπτωση της δυονυμικής ταξινόμησης, τα SVM βρίσκουν ένα υπερεπίπεδο (η υπερεπιφάνεια) που διαχωρίζει τις δύο κατηγορίες δειγμάτων με όσο το δυνατόν μεγαλύτερο περιθώριο. Ως αποτέλεσμα η ακρίβεια του μοντέλου γενικεύεται ευνοϊκά σε άγνωστα δεδομένα και επιτρέπει εξειδικευμένες μεθόδους βελτιστοποίησης που βοηθούν στο SVM να εκπαιδεύεται σε μεγάλο όγκο δεδομένων.

Αυτή η διατύπωση οδηγεί σε ένα διαχωριστικό υπερεπίπεδο που εξαρτάται μόνο από το (συνήθως μικρό υποσύνολο) σημείων δεδομένων που βρίσκονται στο περιθώριο, τα οποία ονομάζονται διανύσματα υποστήριξης. Ως εκ τούτου, ολόκληρος ο αλγόριθμος ονομάζεται μηχανή διανυσμάτων υποστήριξης. Επιπλέον, δεδομένου ότι τα προβλήματα ανάλυσης δεδομένων του πραγματικού κόσμου συχνά περιλαμβάνουν μη γραμμικές εξαρτήσεις, τα SVM μπορούν εύκολα να επεκταθούν για να μοντελοποιήσουν τέτοια μη γραμμικότητα μέσω θετικών ημι-καθορισμένων συναρτήσεων πυρήνων.(aibook.gr)

Τα SVM έχουν ισχυρότερο μαθηματικό υπόβαθρο από άλλες μεθόδους μηχανικής μάθησης όπως τα νευρωνικά δίκτυα και σχετίζεται με καθιερωμένα θεωρήματα της στατιστικής. Ως γραμμικό μοντέλο, όχι μόνο προσπαθεί να ταξινομήσει σωστά τα δεδομένα εκπαίδευσης αλλά επίσης μεγιστοποιεί το περιθώριο για καλύτερη απόδοση γενίκευσης.

Εκπαίδευση ΜΔΥ

Δεδομένου συνόλου L δειγμάτων εκπαίδευσης με D χαρακτηριστικά (D διαστάσεων) όπου το κάθε δείγμα ανήκει σε μία απο δυό κλάσεις $y_i = +1$ η -1 έχουμε το σύνολο: $x_i, y_i, i = 1..L$ όπου $y_i \in -1, 1$ Θεωρούμε ότι το πρόβλημα είναι γραμμικά διαχωρίσιμο, δηλαδή ότι υπάρχει μια γραμμή σε ένα γράφημα διαστάσεων x_1 και x_2 η οποία διαχωρίζει πλήρως τις δύο κλάσεις όταν $D = 2$, ενώ για διαστάσεις $D > 2$ υπάρχει μια υπερεπιφάνεια. Η εξίσωση της επιφάνειας θα έχει τη μορφή $w_1x_1 + w_2x_2 + b = 0$ ή $w \cdot x + b = 0$ αν υιοθετήσουμε διανυσματική γραφή για τα βάρη και το σημείο, όπου το w είναι κάθετο στο υπερεπίπεδο και το $b/|w|$ είναι η κάθετη απόσταση από το υπερεπίπεδο στην αρχή των αξόνων. Εφαρμόζοντας τη σχέση αυτή για δύο σημεία α και β που βρίσκονται πάνω στο σύνορο και αφαιρώντας τις δύο σχέσεις προκύπτει:

$$\mathbf{w} \cdot (\mathbf{x_a} - \mathbf{x_b}) = \mathbf{0}$$

Σε γραμμικώς διαχωρίσιμα προβλήματα, μπορούμε να ορίσουμε δύο επιπλέον σύνορα εκατέρωθεν του $w\&x + b = 0$ που επίσης χωρίζουν τις δύο κλάσεις και απέχουν όσο γίνεται περισσότερο μεταξύ τους. Τα διανύσματα υποστήριξης είναι τα δείγματα πλησιέστερα στο διαχωριστικό σύνορο και σκοπός των μηχανών διανυσμάτων υποστήριξης είναι να προσανατολίσουν αυτό το υπερεπίπεδο έτσι ώστε να βρίσκεται όσο πιο μακριά γίνεται από τα κοντινότερα μεταξύ τους μέλη των δυο κλάσεων. Για ένα σημείο z στο χώρο του προβλήματος, η κλάση y του σημείου θα είναι:

$$\mathbf{y} = \begin{cases} y = 1 & \text{εαν } w \cdot z + b \geq 0 \\ y = -1 & \text{εαν } w \cdot z + b \leq 0 \end{cases}$$

Η απόσταση αυτών των δυο συνόρων από την υπερεπιφάνεια $d_1 = d_2$ ονομάζεται το περιθώριο. Για να προσαρμοστεί η υπερεπιφάνεια ώστε να βρίσκεται όσο το δυνατόν μακρύτερα από τα διανύσματα υποστήριξης, πρέπει να μεγιστοποιηθεί. Καθώς το διάνυσμα ω είναι κάθετο στο ζητούμενο σύνορο, διαιρώντας το με το μήκος του $||w||$ προκύπτει ένα μοναδιαίο διάνυσμα που πολλαπλασιαζόμενο (εσωτερικό γινόμενο) με το διάνυσμα της διαφοράς $(x_+ - x_-)$, δίνει το ζητούμενο d

$$(\mathbf{x_+} - \mathbf{x_-}) \cdot \frac{\mathbf{w}}{||\mathbf{w}||} = \mathbf{d}$$

Σε αυτή τη σχέση, για το $x_- \cdot w$, η σχέση υπολογισμού του y δίνει $1 - b$ ενώ για το $x_- \cdot w$ δίνει $1 + b$. Αντικαθιστώντας στην προηγούμενη σχέση προκύπτει:

$$\mathbf{d} = \frac{\mathbf{2}}{||\mathbf{w}||}$$

Άρα η μεγιστοποίηση του d ισούται με την ελαχιστοποίηση του $||w||$. Άρα η διαδικασία εκπαίδευσης της SVM μπορεί να οριστεί μαθηματικά ως η μεγιστοποίηση του d και ισούται με την ελαχιστοποίηση του $||w||$. Η ελαχιστοποίηση του $|w|$ ισούται με την ελαχιστοποίηση του $1/2|w|^2$ και η χρήση αυτού του παράγοντα επιτρέπει τη βελτιστοποίηση με μεθόδους Quadratic Programming. Πρέπει λοιπόν να βρεθεί:

$$\mathbf{min}\frac{\mathbf{1}}{\mathbf{2}}||\mathbf{w}||^{\mathbf{2}}$$

Η επίλυση αυτού του προβλήματος γίνεται με τη βοήθεια των πολλαπλασιαστών Lagrange (Lagrange multipliers) και προκύπτει ότι τα w και b που ορίζουν το σύνορο απόφασης εξαρτώνται μόνο από διανύσματα (σημεία) που βρίσκονται πάνω στις ευθείες b_1 και b_2, και από εσωτερικά γινόμενα με αυτά.

(Βλαχάβας, Κεφαλάς, Βασιλειάδης Κόκκορας, Σακελλαρίου, aibook.gr)

3.4.4 BERT

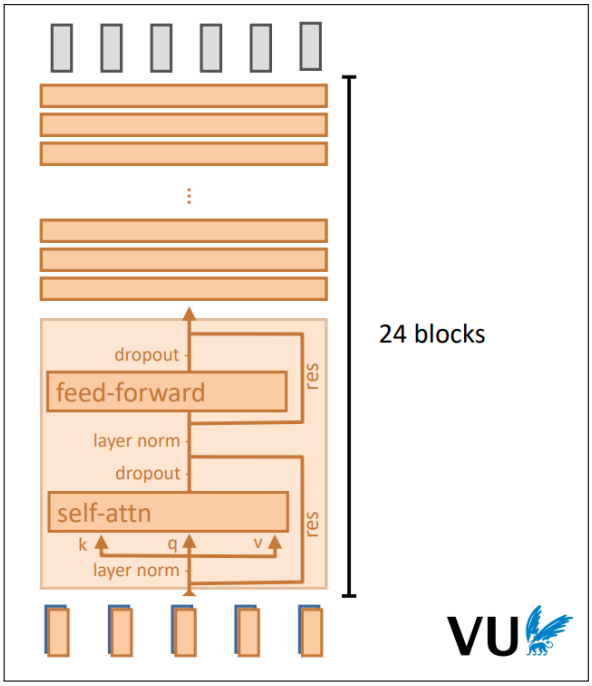
Το BERT (Bi-Directional Encoding Representations from Transformers) είναι μια οικογένεια από μοντέλα transformer βασισμένα στην αρχιτεκτονική του κωδικοποιητή. Χαρακτηρίζεται από την αξιοποίηση πληροφορίας από το σύνολο των συμφραζόμενων ενός όρου και έχει καταφέρει τις βέλτιστες επιδόσεις σε δοκιμασίες επεξεργασίας φυσικής γλώσσας.

Η ολοκληρωμένη εκπαίδευση του BERT και γενικά των μοντέλων transformer γίνεται σε δυο φάσεις: την προεκπαίδευση του μοντέλου σε έναν τεράστιο όγκο μη επισημασμένων κειμένων που περιέχουν πληθώρα θεματικών αντικειμένων με σκοπό το μοντέλο να αποκτήσει μια όσο το δυνατόν σφαιρικότερη κατανόηση της γλώσσας, με τους αλγόριθμους Masked Language Modeling. Όταν αποκτηθεί μεταβαίνει στη φάση της προσαρμογής (fine-tuning) σε συγκεκριμένα σημασιολογικά πλαίσια χρήσης της γλώσσας, με βάση τα κείμενα ενδιαφέροντος. Αυτή η μεθοδολογία της εκπαίδευσης ενός μοντέλου σε κάποια δεδομένα και η χρήση του ίδιου μοντέλου σε κάποιο άλλο πεδίο εφαρμογής αναφέρεται στη μηχανική μάθηση ως ¨μεταφορά γνώσης' (transfer learning).

Τα δύο βασικά μοντέλα που παρουσιάστηκαν το 2018 ήταν το $BERT_{BASE}$ και το $BERT_{LARGE}$. Στον παρακάτω πίνακα φαίνονται τα δύο αυτά μοντέλα μαζί με τα χαρακτηριστικά τους.

	Λ	Η	Α	Αριθμός Παραμέτρων
$BERT_{BASE}$	12	768	12	110M
$BERT_{LARGE}$	24	1024	16	340M

Πιο συγκεκριμένα, το L αναφέρεται στον αριθμό των στρωμάτων των κωδικοποιητών, το H στο μέγεθος του διανύσματος εξόδου (embedding length), το A στον αριθμό των κεφαλών



Σχήμα 3.19: Αρχιτεκτονική του BERT, πηγή *DLVU*

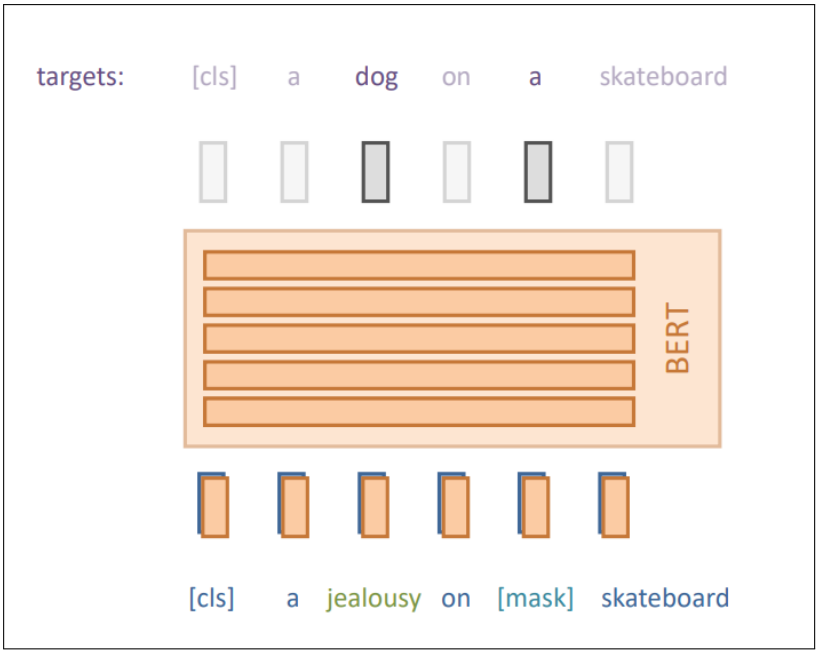
προσοχής ανά μονάδα κωδικοποιητή και οι παράμετροι στον συνολικό αριθμό των παραμέτρων εκπαίδευσης.

Ο tokenizer που χρησιμοποιεί το BERT είναι ο WordPiece, περιέχοντας στο λεξιλόγιό του περίπου 30,000 λεκτικές μονάδες. Η πρώτη λεκτική μονάδα μιας ακολουθίας είναι πάντα η ειδική λεκτική μονάδα [CLS] η οποία σχετίζεται με προβλήματα ταξινόμησης όπως το masked language modeling που χρησιμοποιεί το BERT για να εκπαιδευτεί. Όταν στο BERT χρειάζεται να εισαχθούν ως είσοδοι δύο προτάσεις τότε εισέρχονται ακολουθιακά η μία μετά την άλλη. Το BERT όμως πρέπει να καταλάβει ότι η είσοδος αποτελείται από πάνω από μία πρόταση, επομένως η διαφοροποίηση αυτή γίνεται με δύο τρόπους. Ο πρώτος τρόπος, στο τέλος κάθε πρότασης προσθέτει την ειδική λεκτική μονάδα διαχωρισμού [SEP], η οποία προέρχεται από την αγγλική λέξη separator. Με τον δεύτερο τρόπο, σε κάθε λεκτική μονάδα εισόδου προστίθεται ένα ειδικό διάνυσμα (segment embedding) το οποίο δείχνει και την πρόταση στην οποία ανήκει. Επιπλέον, προστίθεται και το διάνυσμα θέσης το οποίο προσδιορίζει συντακτικά την λεκτική μονάδα. Τελικά, για μια συγκεκριμένη λεκτική μονάδα η αναπαράσταση της εισόδου της υπολογίζεται ως το άθροισμα των τριών παραπάνω διανυσμάτων.

Προεκπαίδευση με την Τεχνική Masked Language Modeling

Σε αυτήν την τεχνική προεκπαίδευσης μερικές λεκτικές μονάδες από την ακολουθία εισόδου αντικαθίστανται με την λεκτική μονάδα [MASK]. Σκοπός του μοντέλου είναι να προβλέψει αυτή τη λεκτική μονάδα που αντικαταστάθηκε με βάση τα συμφραζόμενα. Η παραπάνω αντικατάσταση συμβαίνει με βάση μία πιθανότητα, παραδείγματος χάριν 15%.

Ένα πρόβλημα που προκύπτει με την παραπάνω διαδικασία είναι ότι το BERT χρησιμο-



Σχήμα 3.20: Masked Language Modeling

ποιεί τη λεκτική μονάδα [MASK] στη διαδικασία της προεκπαίδευσης ενώ στη διαδικασία του συμπερασμού και της προσαρμογής (fine-tuning) δε θα ξανασυναντήσει ποτέ αυτήν τη λεκτική μονάδα. Συνεπώς, ένας τρόπος για να ελαττωθεί αυτό το πρόβλημα είναι ο εξής: Στις λεκτικές μονάδες που επιλέγονται τυχαία (με πιθανότητα 15%) να μη γίνεται εξ ολοκλήρου αντικατάσταση με την λεκτική μονάδα [MASK], αλλά να σπάει η πιθανότητα στις εξής τρεις πιθανότητες:

- 80% αντικατάσταση με τη λεκτική μονάδα [MASK].
- 10% αντικατάσταση με μία τυχαία λεκτική μονάδα από το λεξιλόγιο του μοντέλου
- 10% να μη γίνει αντικατάσταση.

Τελικά, για να υλοποιηθεί η παραπάνω τεχνική προεκπαίδευσης, στην έξοδο του BERT συνδέεται ένα επίπεδο πλήρους συνδεδεμένου νευρωνικού δικτύου με αριθμό νευρώνων ίσο με τον αριθμό των λεκτικών μονάδων του λεξιλογίου του μοντέλου (περίπου 30,000), ώστε να υπολογιστεί η πιθανότητα της κρυμμένης λέξης μετά την εφαρμογή ενος επιπέδου softmax.

Προεκπαίδευση με την Τεχνική της Πρόβλεψης της Επόμενης Πρότασης

Ως δεύτερη τεχνική προεκπαίδευσης του μοντέλου BERT χρησιμοποιείται η πρόβλεψη της επόμενης πρότασης (next sentence prediction). Αυτή η μέθοδος είναι πολύ σημαντική για προβλήματα που χρειάζονται κατανόηση των σχέσεων μεταξύ των προτάσεων. Ένα τέτοιο πρόβλημα είναι το πρόβλημα της απάντησης ερωτήσεων (question answering).

Ειδικότερα, η είσοδος του μοντέλου αποτελείται από ένα ζεύγος προτάσεων, την πρόταση A και την πρόταση B. Αν η πρόταση B ακολουθεί στην πραγματικότητα την πρόταση A, τότε η

δυαδική ετικέτα είναι IsNext, αλλιώς NotNext. Τα ζευγάρια που εισήχθησαν στο BERT κατά το στάδιο της προεκπαίδευσης του είχαν ετικέτα κατά 50% IsNext, με το υπόλοιπο 50% να είναι NotNext. Συνεπώς, γίνεται λόγος για ένα δυαδικό πρόβλημα (binary task). Η δυαδική αυτή έξοδος του μοντέλου για να φανεί εκμεταλλεύεται το [CLS] token, το οποίο συνδέεται με ένα επίπεδο πλήρους συνδεδεμένου νευρωνικού δικτύου, που προβάλλει την ενσωμάτωση του [CLS], με μέγεθος 768 (για το BERTBASE), σε ένα διάνυσμα δύο διαστάσεων (IsNext-NotNext), ενώ, τέλος, με μία στρώση softmax εξάγεται το τελικό αποτέλεσμα.

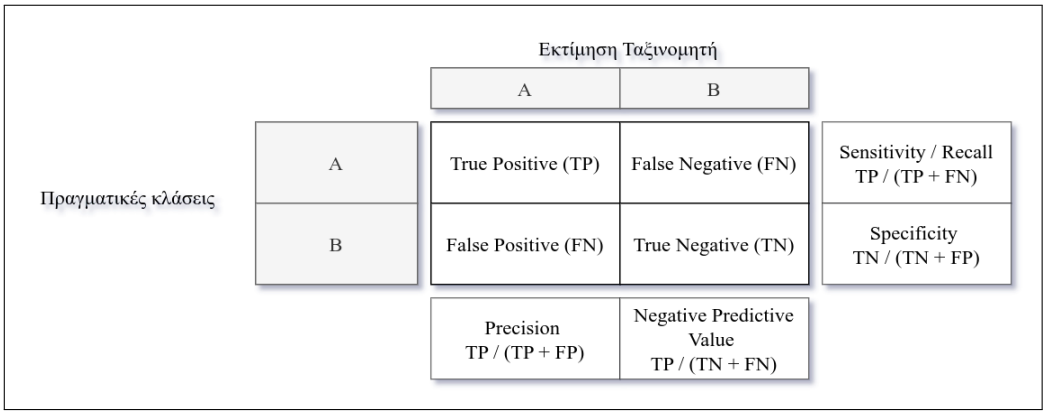
3.5 Μετρικές αξιολόγησης

Κατά την διαδικασία του συμπερασμού, ένας ταξινομητής μπορεί να κάνει δύο τύπους σφαλμάτων: θετικό αποτέλεσμα για τη λάθος κλάση (η λανθασμένη άποψη ότι ΄το παράδειγμα είναι της κατηγορίας που προσπαθώ να προβλέψω΄) και αρνητικό αποτέλεσμα στη σωστή κλάση (η λανθασμένη άποψη ότι ΄το παράδειγμα δεν είναι της κατηγορίας που προσπαθώ να προβλέψω΄).

Είναι στο ενδιαφέρον μας να προσδιοριστεί ποιος από αυτούς τους δύο τύπους σφαλμάτων γίνονται, ώστε να μπορούμε να προσδιορίσουμε εάν ο ταξινομητής είναι υπερευαίσθητος σε κάποια χαρακτηριστικά και πιάνει πολλές περιπτώσεις ως θετικές, ή δεν πιάνει περιπτώσεις που έπρεπε. Ένας πίνακας ή μήτρα σύγχυσης είναι ένας τρόπος για να έχουμε μια εικόνα για αυτές τις πληροφορίες.

Ένας πίνακας σύγχυσης συνοψίζει την απόδοση ταξινόμησης ενός ταξινομητή σε σχέση με ορισμένα δεδομένα ελέγχου. Είναι ένας δισδιάστατος πίνακας, που έχει ως δείκτες στη μία διάσταση την πραγματική κλάση ενός αντικειμένου και στην άλλη από την κλάση που αναθέτει ο ταξινομητής.

Παρατίθεται ένα πρότυπο πίνακα σύγχυσης για έναν ταξινομητή δύο κλάσεων:



Σχήμα 3.21: Πίνακας Σύγχυσης

Η διαδικασία μπορεί εύκολα να επεκταθεί στην περίπτωση περισσότερων από 2 κλάσεις:

		Εκτίμηση Ταξινομητή		
		Κλάση Α	Κλάση Β	Κλάση Γ
Πραγματικές κλάσεις	Κλάση Α	10	1	1
	Κλάση Β	0	3	0
	Κλάση Γ	0	2	7

Σχήμα 3.22: Πολυωνυμικός Πίνακας Σύγχυσης

Ακρίβεια (Precision)

Η ακρίβεια ορίζεται ως ο λόγος του αριθμού των σωστών εκτιμήσεων True Positive δια το συνολικό αριθμό των θετικών εκτιμήσεων για μια κλάση από τα συμπεράσματα ενός μοντέλου. Την ορίζουμε σε σχέση με των αριθμό σωστών θετικών προβλέψεων για μια κλάση (True Positive) και λανθασμένων θετικών προβλέψεων.

$$Precision = \frac{TP}{TP + FP} \tag{3.13}$$

Αξιοπιστία (Accuracy)

Η αξιοπιστία αναφέρεται στο βαθμό που οι προβλέψεις του μοντέλου αντικατοπτρίζουν την πραγματικότητα. Χρησιμοποιείται σαν μέτρο εκτίμησης της ικανότητας μιας μεθόδου ταξινόμησης να προσδιορίζει ή αποκλείει σωστά τα χαρακτηριστικά των παραδειγμάτων. Είναι ο λόγος των σωστών προβλέψεων (σωστές θετικές προβλέψεις συν σωστές αρνητικές προβλέψεις) προς τον συνολικό αριθμό των περιπτώσεων που εξετάζουμε.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.14}$$

Μέτρηση F (F-measure)

Η μέτρηση F είναι μια συνδυαστική μετρική απόδοσης ταξινόμησης η οποία αξιολογεί την ισορροπία μεταξύ ακρίβειας και ανάκλησης. Είναι ο αρμονικός μέσος όρος ακρίβειας και ανάκλησης και υπολογίζεται ως εξής:

$$FMeasure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.15}$$

Η μέτρηση F λαμβάνει υπόψη τόσο την ακρίβεια όσο και την ανάκληση, παρέχοντας έναν μόνο αριθμό που συνοψίζει την απόδοση ενός ταξινομητή.

Κεφάλαιο 4

Πειραματικά αποτελέσματα των εφαρμοζόμενων προσεγγίσεων

Στο κεφάλαιο αυτό παρουσιάζονται η διαδικασία και τα αποτελέσματα της υλοποίησης των μοντέλων που δημιουργήθηκαν με τις τεχνικές που αναλύθηκαν στα προηγούμενα κεφάλαια.

4.1 Λογισμικό

Το κύριο εργαλείο της υλοποίησης στη μελέτη μας είναι η γλώσσα προγραμματισμού Python (έκδοση 3.10). Η Python είναι μια υψηλού επιπέδου ερμηνευόμενη γλώσσα. Δημιουργημένη από τον Guido von Rossum το 1991, η φιλοσοφία της Python δίνει έμφαση στην αναγνωσιμότητα και εύκολη ερμηνεία του κώδικα. Οι δομές δεδομένων που παρέχει μαζί με την αντικειμενοστρεφής προσέγγιση βοηθούν τους προγραμματιστές στην συγγραφή ξεκάθαρων προγραμμάτων με ένα λογικό, δηλωτικό στυλ (declarative programming). Αυτό στηρίζεται από τη χρήση δυναμικών τύπων δεδομένων και garbage collection. Υποστηρίζει πολλαπλές προγραμματιστικές μεθοδολογίες, με στοιχεία από διαδικασιακό, αντικειμενοστρεφή και συναρτησιακό προγραμματισμό. Το πιο σημαντικό όμως χαρακτηριστικό που κάνει την Python την ιδανική επιλογή για χρήση στα πεδία της επιστήμης δεδομένων, της στατιστικής και της επεξεργασίας φυσικής γλώσσας είναι η πληθώρα πανίσχυρων βιβλιοθηκών κατασκευασμένων από την τεράστια κοινότητα της Python. Για τους σκοπούς της εργασίας χρησιμοποιήθηκαν τα κάτωθι εργαλεία Python:

- **Jupyter Lab 3.6.1** Διαδραστικό περιβάλλον προγραμματισμού, το οποίο μας παρέχει ένα περιβάλλον με κελιά Python με το καθένα να εκτελεί ένα Read-Evaluate-Print-Loop ενός ξεχωριστού Python Kernel, μαζί με πολλές πρόσθετες λειτουργίες σχεδιασμένες για την εισαγωγή δεδομένων, την αξιολόγηση της απόδοσης των υπολογισμών και την οπτικοποίηση δεδομένων. Βασίζεται στο Project Jupyter, μια μη κερδοσκοπική, ανοιχτού κώδικα προσπάθεια, η οποία ξεκίνησε με το IPython Project (διαδραστική Python) το 2014 και έχει εξελιχθεί στην μεγαλύτερη πλατφόρμα για την υποστήριξη της διαδραστικής επιστήμης δεδομένων και της επιστημονικής υπολογιστικής σε πολλές γλώσσες προγραμματισμου.

- **NumPy** Η θεμελιώδης βιβλιοθήκη για την επιστημονική υπολογιστική στην Python. Παρέχει ισχυρές και υπολογιστικά αποδοτικές δομές αριθμητικών δεδομένων όπως το n-dimensional array, δομές τυχαιότητας, πράξεις γραμμικής άλγεβρας και εργαλεία για την βελτιστοποίηση των υπολογισμών διαλειτουργώντας με κλήσεις σε C/C++.

- **Scikit-Learn** Βιβλιοθήκη μηχανικής μάθησης που παρέχει μια πλήρη γκάμα από αλγορίθμους για παλινδρόμηση, ταξινόμηση και συσταδοποίηση, καθώς και διάφορα βοηθητικά εργαλεία εξαγωγής χαρακτηριστικών, εξομάλυνσης δεδομένων και αξιολόγησης. Χρησιμοποιεί ένα κοινό, συνεκτικό API και έχει σχεδιαστεί για τη διαλειτουργικότητα με άλλες επιστημονικές βιβλιοθήκες όπως η NumPy και η SciPy [21].

- **Pandas 1.5.3** Βιβλιοθήκη διαχείρισης δεδομένων. Παρέχει εύχρηστες, υψηλής απόδοσης δομές και εργαλεία ανάλυσης δεδομένων.

- **TensorFlow 2.11** Βιβλιοθήκη βαθιάς μάθησης ανεπτυγμένη από τη Google. Παρέχει πληθώρα από δομές δεδομένων, αλγόριθμους και εργαλεία αιχμής για την κατασκευή αρχιτεκτονικών βαθέων νευρωνικών δικτύων, συμπεριλαμβανομένων και μοντέλων transformers όπως το BERT, ενώ επίσης παρέχει προεκπαιδευμένα σύνολα δεδομένων για μεταφορά μάθησης.

- **SpaCy** Ολοκληρωμένη, προσαρμόσιμη βιβλιοθήκη επεξεργασίας φυσικής γλώσσας που μας δίνει τη δυνατότητα να εξατομικεύσουμε ένα σύστημα σταδιακή επεξεργασία κειμένου, παρέχοντας εργαλεία όπως tokenizers, POS tagger, dependency parsing και λειτουργίες παρουσίασης με την υποβιβλιοθήκη displacy.

- **MatPlotLib** Βιβλιοθήκη γραφικών παραστάσεων 2D, οι οποία παράγει κάθε είδους ποιοτικά και προσαρμόσιμα γραφήματα, και έχει σχεδιαστεί για να συνεργάζεται αδιάλειπτα με τις άλλες βιβλιοθήκες ανάλυσης δεδομένων στο κομμάτι της παρουσίασης.

- **Natural Language ToolKit** Γενική βιβλιοθήκη εργαλείων για την επεξεργασία φυσικής γλώσσας. Σε αντίθεση με τη SpaCy, περιλαμβάνει κάθε λογής εργαλεία χωρίς την αυστηρή απαίτηση της τεχνολογίας αιχμής ή της βελτιστοποιημένης απόδοσης, με έμφαση στην εκπαίδευση και τη μελέτη ακόμα και απλούστερων μεθόδων. Παρέχει σώματα κειμένου όπως το Brown Corpus, λίστες συχνών λέξεων σε πολλές γλώσσες, λημματοποίηση WordNet, tokenizers κ.α.

4.2 Επισκόπηση του συνόλου δεδομένων που χρησιμοποιήθηκε

Αφού έχει συλλεχθεί και ομαλοποιηθεί το σύνολο δεδομένων των βιογραφικών και έχουν δημιουργηθεί τρία αρχεία csv, δυο σύνολα εκπαίδευσης από το livecareer.com και ένα για το σύνολο ελέγχου από το indeed.com, σε αυτή την ενότητα θα γίνει μία στατιστική ανάλυση.

Το επαυξημένο σύνολο εκπαίδευσης περιέχει περίπου 18000 παραδείγματα από 18 κατηγορίες, ενώ το αρχικό 875. Αυτά είναι ο συνδυασμός των dataset του livecareer και του

μικρού συνόλου από το kaggle. Για μια πρώτη ματιά στα δεδομένα μας, μπορούμε να δούμε σε τι ποσοστό καλύπτει η κάθε μια στο σύνολο. Κάποιες κατηγορίες έχουν περισσότερα παραδείγματα από άλλες, απόρροια της διαδικασίας καθαρισμού των δεδομένων.

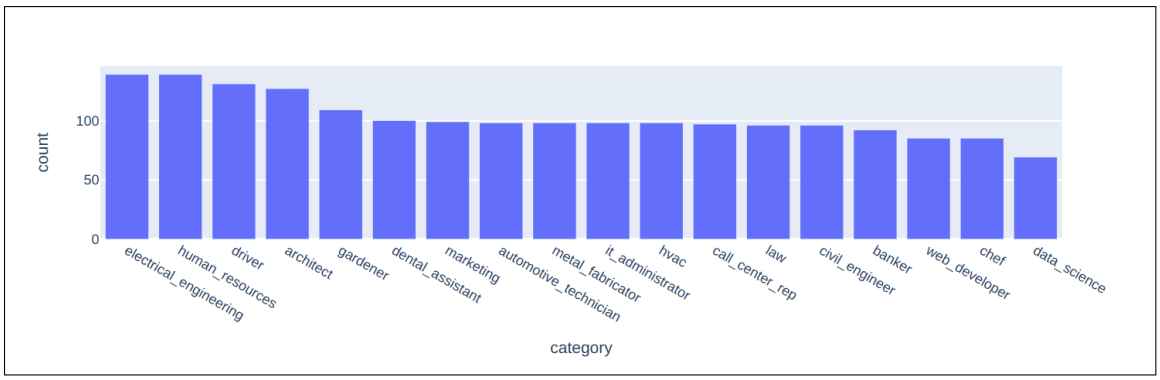
Κατηγορία	#
gardener	50
chef	50
metal_fabricator	50
driver	50
web_developer	50
civil_engineer	50
banker	50
electrical_engineering	49
architect	49
human_resources	49
law	49
automotive_technician	49
data_science	48
hvac	48
call_center_rep	47
dental_assistant	46
marketing	46
it_administrator	45

Πίνακας 4.1: Αρχικό σύνολο εκπαίδευσης

Κατηγορία	#
gardener	1060
driver	1060
metal_fabricator	1060
electrical_engineering	1060
web_developer	1060
chef	1060
civil_engineer	1060
banker	1060
call_center_rep	1046
architect	1046
human_resources	1046
hvac	1046
law	1046
automotive_technician	1046
data_science	1032
marketing	1032
dental_assistant	1046
it_administrator	1046

Πίνακας 4.2: Επαυξημένο σύνολο εκπαίδευσης

Το σύνολο ελέγχου αποτελείται από τα δεδομένα από το indeed, με τις ίδιες 18 κλάσεις σε 1900 παραδείγματα. Οι κλάσεις του πρώτου έχουν παρόμοια μεγέθη λόγω της εσκεμμένης ισορρόπησης τους, ενώ οι κλάσεις στα δεδομένα ελέγχου είναι ελαφρώς διαφορετικές σε όγκο.

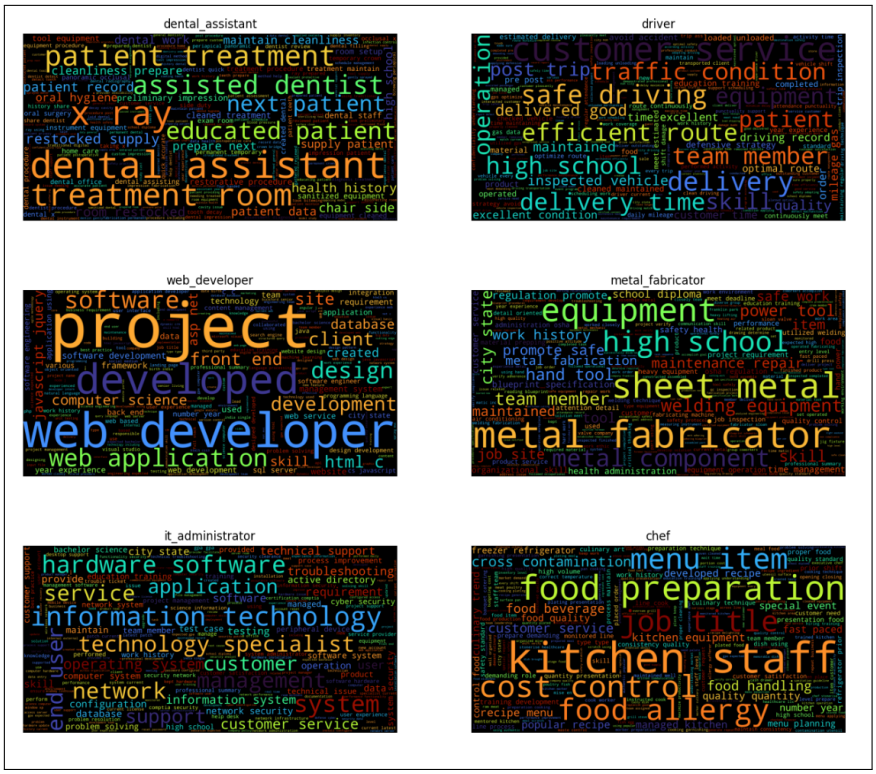


Σχήμα 4.1: Ιστόγραμμα δεδομένων ελέγχου

Για μια εικόνα του περιεχομένου των βιογραφικών μπορούμε να βρούμε τις πιο συχνά χρησιμοποιούμενες λέξεις για κάθε κατηγορία. Μιας και ασχολούμαστε για ελεύθερο κείμενο,

γνωρίζουμε ότι υπάρχει ένα σύνολο από συχνές, συνήθως βοηθητικές λέξεις οι οποίες θα εμφανίζονται πολύ συχνότερα. Εφόσον μας ενδιαφέρουν οι λέξεις που χαρακτηρίζουν καλύτερα τα κείμενα, θα τις αφαιρέσουμε, και για περαιτέρω απλοποίηση, θα αφαιρέσουμε και τις κλιτές μορφές των ίδιων λέξεων (inflections). Για αυτό το σκοπό λημματοποιούμε τα κείμενα μας, δηλαδή, αντικαθιστούμε την κάθε λέξη με την ριζική της μορφή, ανατρέχοντα σε λεξιλόγιο. Κάνουμε χρήση της βιβλιοθήκης WordNetLemmatizer. Για παράδειγμα,οι λέξη were θα γίνει be, οι λέξη positions θα γίνει position κ.ο.κ.

Εύπεπτος τρόπος να παρουσιάσουμε τη συχνότητα των λέξεων είναι με ένα Word Cloud (νέφος από λέξεις) με το μέγεθος της γραμματοσειράς να αναδεικνύει την συχνότητα της κάθε λέξης. Μπορούμε έτσι εύκολα να διακρίνουμε ποιές φράσεις εμφανίζονται συχνότερα, άρα είναι βασικότερες γιατί απαντώνται στο μεγαλύτερο μέρος των κειμένων της κάθε κατηγορίας. Μπορούμε από εδώ να διακρίνουμε μοτίβα από τις ΄προκαταλήψεις΄ που εμφανίζονται στα δεδομένα μας, όπως για παράδειγμα ότι σε ανειδίκευτες θέσεις, όπως ο οδηγός ή ο κηπουρός, εμφανίζεται ο όρος του απολυτηρίου λυκείου ΄high school΄, ή το ότι σε δουλειές που έχουν απ΄ ευθείας επαφή με τους πελάτες εμφανίζεται συχνά ο όρος customer. Εάν δεν είχαμε κάνει τον καθαρισμό των κειμένων δεν θα μπορούσαμε να βρούμε αυτά τα μοτίβα, διότι θα κυριαρχούσαν οι συχνοί και τυποποιημένοι όροι, όπως resume.



Σχήμα 4.2: WordCloud βιογραφικών

4.3 Πειραματικά αποτελέσματα

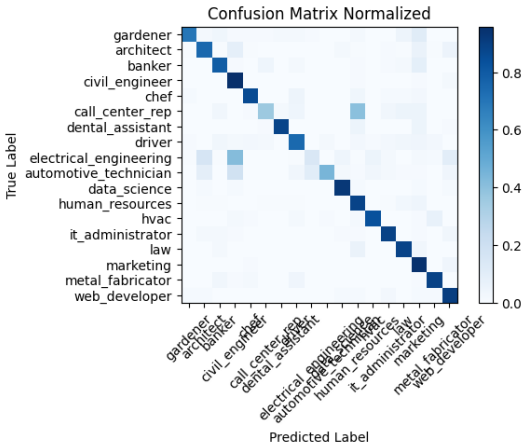
4.3.1 Διαδικασία Ταξινόμησης

Για την δημιουργία του λεξιλογίου και εξαγωγής χαρακτηριστικών των μοντέλων BoW χρησιμοποιήθηκαν οι δομές CountVectorizer και TfIdfVectorizer, από την υποβιβλιοθήκη εξαγωγής χαρακτηριστικών κειμένου του SciKit-Learn. Σε κάθε περίπτωση, προσαρμόζουμε τους διανυσματοποιητές στα λεξιλόγια των δεδομένων και εισάγουμε στους αλγόριθμους τα χαρακτηριστικά ως διανύσματα, και τις ετικέτες ως μια λίστα. Η κωδικοποίηση των συμβολοσειρών είναι σε UTF-8. Για την αρχικοποίηση του CountVectorizer, θέτουμε τις πιο βασικές παραμέτρους του πειράματος. Ορίζουμε το επίπεδο της ανάλυσης να είναι ανά λέξη, χρησιμοποιώντας τον tokenizer που παρέχεται από τη βιβλιοθήκη nltk. Εφόσον τα δεδομένα μας είναι εξαρχής ομαλοποιημένα, δεν εισάγουμε κάποια συνάρτηση καθαρισμού. Από τις δοκιμές μας, παρατηρήθηκε μια σημαντική αύξηση στην ακρίβεια των μοντέλων Logistic Regression με την αφαίρεση συχνών λέξεων χρησιμοποιώντας το αγγλικό λεξικό stopwords του nltk. Περιορίζουμε το μέγεθος του λεξιλογίου σε 3000 λεκτικές μονάδες μοναδικών λέξεων (υνιγραμς). Με αυτές τις παραμέτρους, προσαρμόζουμε τον CountVectorizer στα κείμενα του απλού και του επαυξημένου συνόλου. Για τον TfIdfVectorizer χρησιμοποιείται η ίδια ρύθμιση, με επιπλέον τη αγνόηση λέξεων με καθολική συχνότητα μικρότερη του 2, οι οποίες συχνά είναι τυπογραφικά λάθη.

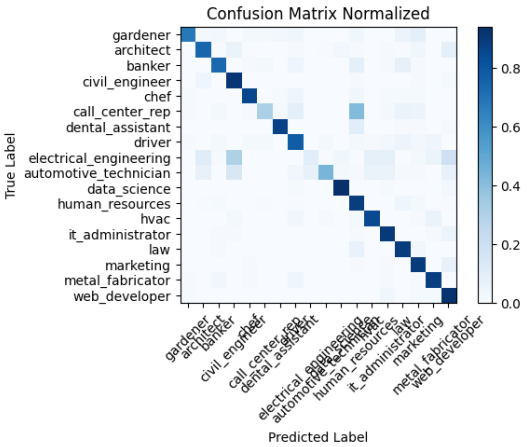
Θα δοκιμαστούν οι υλοποιήσεις των αλγορίθμων μηχανικής μάθησης που παρέχονται από τη βιβλιοθήκη SciKit-Learn. Όντας εκτιμητές (estimators), χρησιμοποιούν ενα παρόμοιο API με τις συναρτήσεις fit και predict. Η fit χρησιμοποιείται για την προσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης και δέχεται δυο εισόδους: εναν πίνακα $x \times n$ με n χαρακτηριστικά και μια λίστα με επισημάνσεις. Μετά το πέρας της διαδικασίας εκπαίδευσης μπορεί να χρησιμοποιηθεί η συνάρτηση predict, με την οποία το μοντέλο δέχεται έναν πίνακα παραδειγμάτων ίδιας μορφής και εξάγει μια λίστα με τις εκτιμώμενες επισημάνσεις. Έπειτα, μπορούμε να χρησιμοποιήσουμε αυτές τις εκτιμήσεις για να τις συγκρίνουμε με τις γνωστές ετικέτες του συνόλου εκπαίδευσης, ώστε να εξάγουμε τα αποτελέσματα των αλγορίθμων, με τη μορφή μετρικών όπως ακρίβεια, recall, F1 score με τη χρήση βοηθητικών συναρτήσεων αξιολόγησης.

Στη συνέχεια θα παρατεθούν τα αποτελέσματα της ταξινόμησης στη μορφή του πίνακα σύγχυσης για κάθε μια κλάση των εγγράφων, πριν και μετά την επαύξηση, με διαφορετικές μεθόδους εξαγωγής χαρακτηριστικών. Σημειωτέον, η ταξινόμηση με βάση την εξαγωγή διανύσματος εγγράφου δεν εφαρμόζεται στον Naive Bayes διότι περιέχει διανύσματα πραγματικών αριθμών. Συνεπώς, χρησιμοποιούνται αλγόριθμοι διαχωρισμού ενός μοναδικού διανύσματος στον πολυδιάστατο χώρο. Οι αλγόριθμοι που μπορούν να δεχτούν τέτοιες αναπαραστάσεις είναι οι SVC και Logistic Regression. Τα δεδομένα εισόδου εξάχθηκαν με τη χρήση doc2vec από τη βιβλιοθήκη SpaCy, και αποθήκευση σε ένα αντικείμενο numpy.

4.3.2 Ταξινόμηση - Naive Bayes

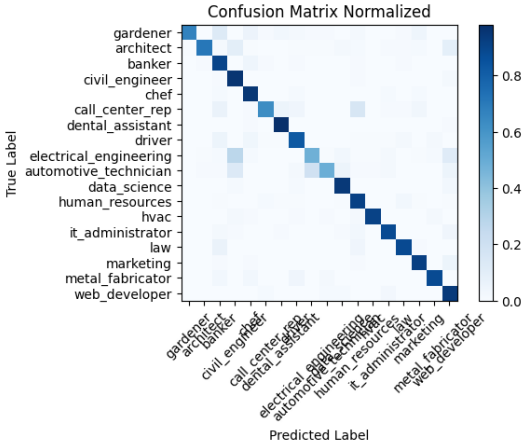


(α′) Αρχικό Σύνολο - Ακρίβεια: 0.753

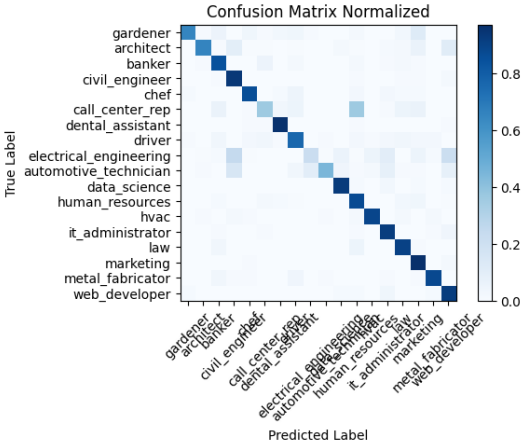


(β′) Επαυξημένο Σύνολο - Ακρίβεια: 0.741

Σχήμα 4.3: Naive Bayes - CountVectorizer



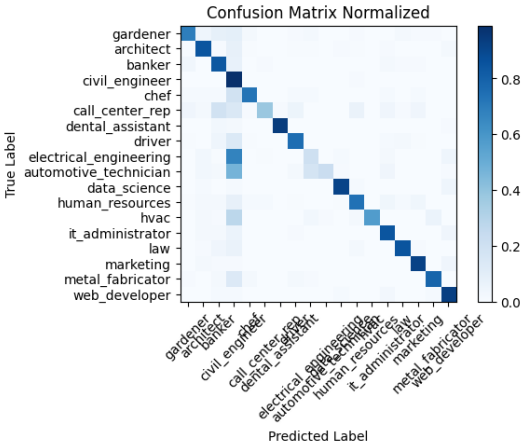
(α′) Αρχικό Σύνολο - Ακρίβεια: 0.808



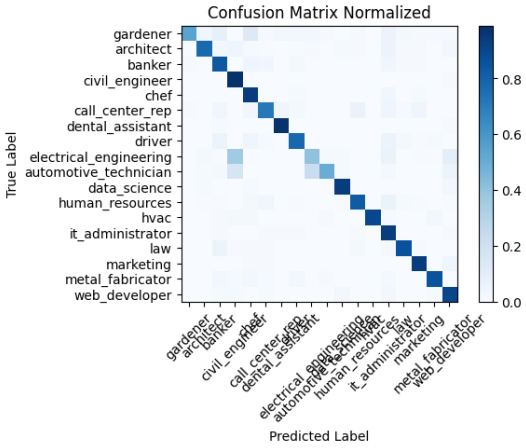
(β′) Επαυξημένο Σύνολο - Ακρίβεια: 0.753

Σχήμα 4.4: Naive Bayes - TF-IDF

4.3.3 Ταξινόμηση - Λογιστική Παλινδρόμηση

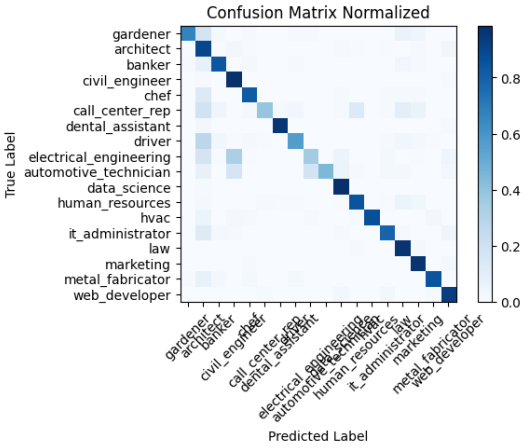


(α′) Αρχικό Σύνολο - Ακρίβεια: 0.724

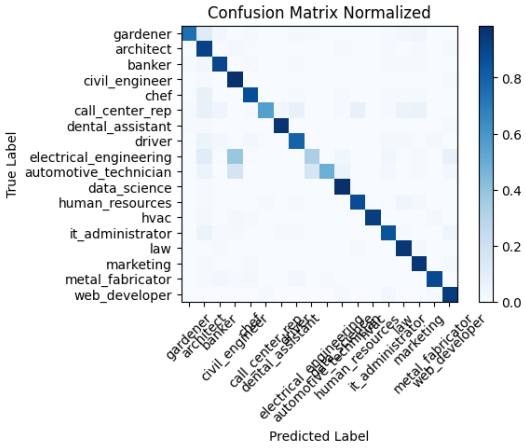


(β′) Επαυξημένο Σύνολο - Ακρίβεια: 0.795

Σχήμα 4.5: Logistic Regression - CountVectorizer

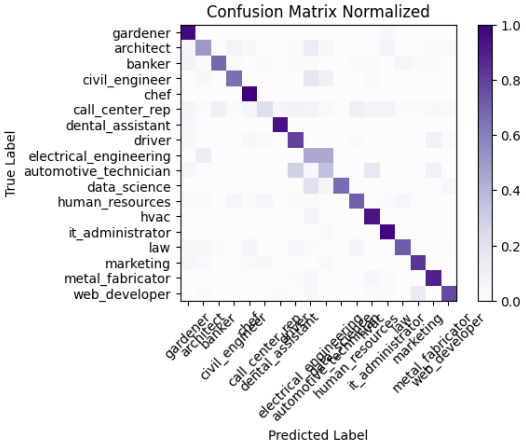


(α′) Αρχικό Σύνολο - Ακρίβεια: 0.775

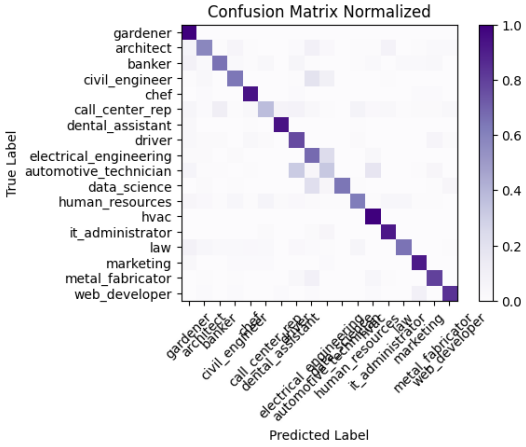


(β′) Επαυξημένο Σύνολο - Ακρίβεια: 0.821

Σχήμα 4.6: Logistic Regression - TF-IDF



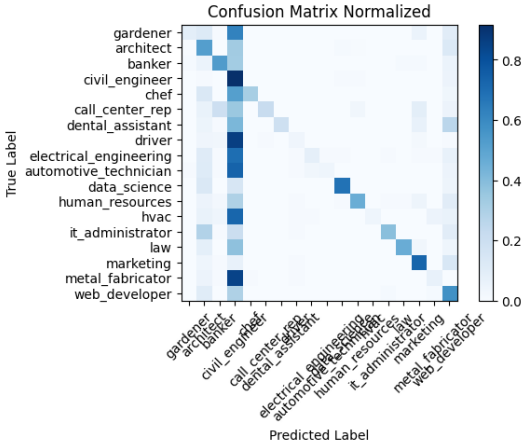
(α′) Αρχικό Σύνολο - Ακρίβεια: 0.587



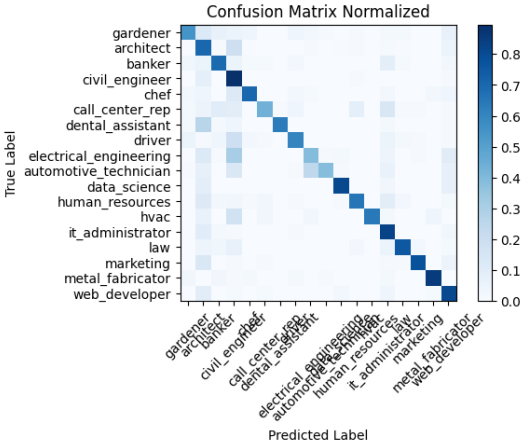
(β′) Επαυξημένο Σύνολο - Ακρίβεια: 0.654

Σχήμα 4.7: Logistic Regression - Doc2Vec

4.3.4 Ταξινόμηση - Μηχανές Διανυσμάτων Υποστήριξης

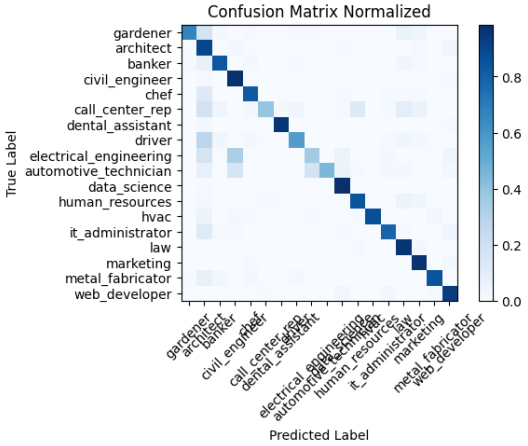


(α′) Αρχικό Σύνολο - Ακρίβεια: 0.346

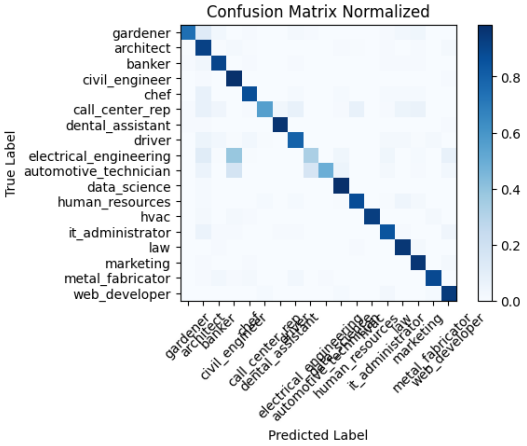


(β′) Επαυξημένο Σύνολο - Ακρίβεια: 0.664

Σχήμα 4.8: SVM - CountVectorizer

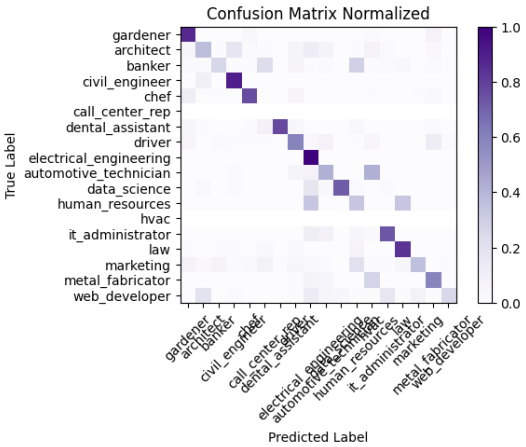


(α′) Αρχικό Σύνολο - Ακρίβεια: 0.614

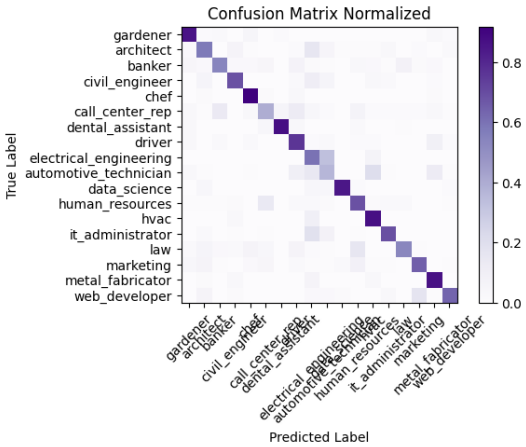


(β′) Επαυξημένο Σύνολο - Ακρίβεια: 0.756

Σχήμα 4.9: SVM - TF-IDF



(α′) Αρχικό Σύνολο - Ακρίβεια: 0.463



(β′) Επαυξημένο Σύνολο - Ακρίβεια: 0.645

Σχήμα 4.10: SVM - Doc2Vec

4.3.5 Ταξινόμηση - BERT

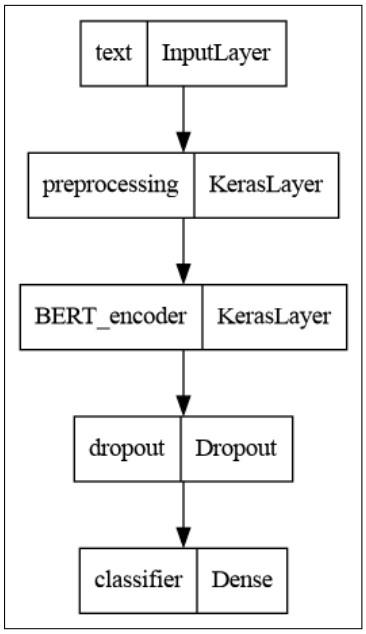
Επειδή το BERT δεν είναι αλγόριθμος ταξινόμησης αλλά ένα μεγάλο προεκπαιδευμένο γλωσσικό μοντέλο για τους σκοπούς της εργασίας πρέπει να το προσαρμόσουμε στη συγκεκριμένη περίπτωση χρήσης, την ταξινόμηση κειμένου. Για να το κάνουμε αυτό, κατασκευάζουμε μια αρχιτεκτονική νευρωνικού δικτύου και χρησιμοποιούμε σαν δομικά στοιχεία τις δομές που μας παρέχονται από το BERT small. Για την υλοποίηση, θα χρησιμοποιήσουμε το μοντέλο BERT small όπως παρέχεται από την υλοποίηση της βιβλιοθήκης Tensorflow. Η πρώτη, βοηθητική δομή που παρέχεται είναι το επίπεδο προεπεξεργασίας που αναλαμβάνει τη μετατροπή των εισόδων σε μορφή κατάλληλη για την είσοδο τους στα επόμενα επίπεδα επεξεργασίας του BERT. Αποτελείται περίπου από 30.000 λεκτικές μονάδες οι οποίες έχουν εξαχθεί από άρθρα στην αγγλική Wikipedia και το Books Corpus με χρήση του αλγορίθμου WordPiece.

Για τα κύρια επίπεδα του transformer, χρησιμοποιείται η αρχιτεκτονική που αναφέρεται στην αρχική υλοποίηση του BERT [22], υλοποιημένη για συμβατότητα με τις δομές των βιβλιοθηκών TensorFlow 2 και Keras. Είναι η μικρότερη έκδοση του BERT που προορίζεται για συστήματά με μικρότερους υπολογιστικούς πόρους. Απαρτίζεται από 4 κρυφά επίπεδα transformer (blocks), με μέγεθος ενσωματώσεων 512, και 8 κεφαλές αυτοπροσοχής. Οι παράμετροι του συγκεκριμένου μοντέλου έχουν προεκπαιδευτεί επίσης στην αγγλική Wikipedia και στο BooksCorpus[1].

Στο νευρωνικό δίκτυο έχουμε αρχικά ένα επίπεδο εισόδου (Input Layer), το οποίο δέχεται τις ανεπεξέργαστες εισόδους, και έχει μέγεθος 128 λεκτικές μονάδες. Ακολουθεί το βοηθητικό επίπεδο προεπεξεργασιας του BERT, τα οποία ετοιμάζουν τις εισόδους για εισαγωγή στα λανθάνοντα επίπεδα του transformer. Έπειτα, προσθέτουμε 2 επίπεδα για να προσαρμόσουμε την εκπαίδευση του δικτύου στην ταξινόμηση των εγγράφων μας. Για την βελτίωσή των επιδόσεων και την αποφυγή υπερπροσαρμογής, προσθέτουμε ένα dropout layer, με πιθανότητα 10%, ενώ το τελευταίο είναι ένα πλήρως συνδεδεμένο(πυκνό) επίπεδο με 18 μονάδες, μια για

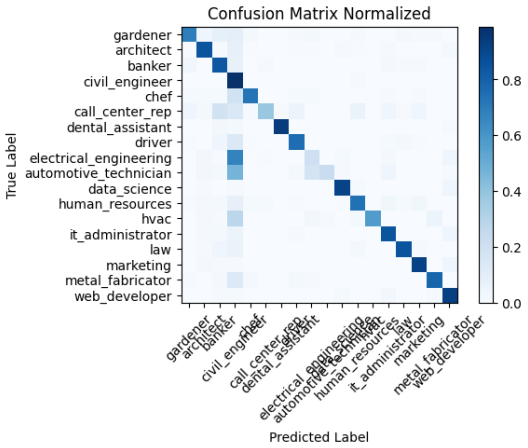
[1]https://paperswithcode.com/dataset/bookcorpus

κάθε κλάση.

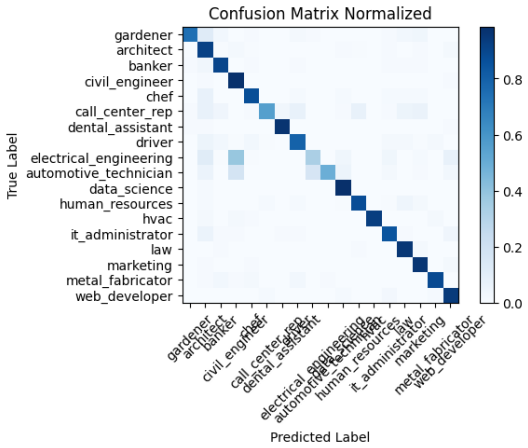


Σχήμα 4.11: Νευρωνικό δίκτυο BERT με TensorFlow

Σκοπός είναι το μοντέλο να εκπαιδευτεί στο να ξεχωρίζει τις διαφορετικές κατηγορίες βιογραφικών.Η αντικειμενική συνάρτησή κόστους, στην περίπτωσή μας με πολλές διαφορετικές κλάσεις ορίζεται η Sparse Categorical Crossentropy Loss. Τέλος, επιλέγουμε κάποιες αρχικές υπερπαραμέτρους. Στην υλοποίησή επιλέξαμε τα epochs να είναι 5, με αρχικό ρυθμό εκπαίδευσης 0.00003 και βελτιστοποίηση τον αλγόριθμο AdamW όπως προτείνεται στη βιβλιογραφία. Επίσης, για το σύστημά που χρησιμοποιήθηκε, θέσαμε τα βήματα προετοιμασίας (warmup steps) σε 10.



(α′) Αρχικό Σύνολο - Ακρίβεια: 0.453



(β′) Επαυξημένο Σύνολο - Ακρίβεια: 0.891

Σχήμα 4.12: BERT

4.4 Συγκριτική Ανάλυση

Τα αποτελέσματα της εκπαίδευσης των μοντέλων στο αρχικό, μη επαυξημένο σύνολο δίνονται στον παρακάτω πίνακα:

	Accuracy	Precision	Recall	F1 Score	Training Time (s)
NB - CountVec	0.757	0.792	0.776	0.743	0.536
NB - TFIDF	0.815	0.842	0.833	0.813	0.539
Logistic Regression - CountVec	0.744	0.831	0.749	0.754	8.36
Logistic Regression - TFIDF	0.813	0.832	0.818	0.804	2.35
Logistic Regression - Doc2Vec	0.677	0.714	0.686	0.661	0.127
SVM - CountVec	0.464	0.725	0.468	0.499	3.65
SVM - TFIDF	0.704	0.805	0.701	0.697	3.92
SVM - Doc2Vec	0.620	0.661	0.633	0.602	0.247
BERT	0.453	0.570	0.431	0.399	14.4

Πίνακας 4.3: Σύγκριση αλγορίθμων - Αρχικό Σύνολο

Ο αλγόριθμος Naive Bayes εκπαιδεύεται γρηγορότερα, και πάρα την απλότητα του, παρέχει ίδια ή καλύτερη ακρίβεια από άλλες μεθόδους. Εδώ παρατηρείται ένα ενδιαφέρον φαινόμενο. Η αναζύγιση με TF-IDF αυξάνει αισθητά την ακρίβεια στην περίπτωση του αλγορίθμου NB αλλά η χρήση μεγαλύτερου σύνολου δεδομένων φαίνεται να μπερδεύει τον ταξινομιτή σε πολλές περιπτώσεις. Πιθανόν οι αλλαγές στην κατανομής των λέξεων στα λεξιλόγια να δυσκολεύουν την ταξινόμηση, με μια μικρή διαφορά 1.2%. Είναι η μοναδική περίπτωση όπου το απλούστερο σύνολο δεδομένων ξεπερνά σε ακρίβεια το επαυξημένο. Επίσης παρατηρείται μια μεγάλη πτώση στη βελτίωση του χρόνου εκπαίδευσης με τη χρήση χαρακτηριστικών TF-IDF στον αλγόριθμο λογιστικής παλινδρόμησης, η ακρίβεια του οποίου είναι αντίστοιχη του Naive Bayes για μικρά σετ δεδομένων αλλά τον ξεπερνάει σε ακρίβεια με την εκπαίδευση στο επαυξημένο σύνολο. Ενώ στο μικρό σετ η ακρίβεια αυξάνει κατά μεγάλο ποσοστό με την αναζύγιση TF-IDF, στο μεγάλο σετ παρουσιάζει μια μικρότερη βελτίωση. Δραματική διαφορά στην ακρίβεια του SVC εμφανίζεται με τη χρήση TF-IDF με μια 52% αύξηση στην ακρίβεια. Η διαφορές στην ταχύτητα εκπαίδευσης σε σχέση με τα μεγέθη των συνόλων φαίνεται να αυξάνονται γραμμικά. Η καλύτερη απόδοση κλιμάκωσης, για ένα 10 φορές μεγαλύτερο dataset, είναι της λογιστικής παλινδρόμησης με 4.5 φορές. Ο αποδοτικότερος αλγόριθμος παραμένει ο Naive Bayes και η κλιμάκωση είναι γραμμική με περίπου 9 φορές περισσότερο χρόνο. Οι αλγόριθμοι SVC και BERT εκπαιδεύονται πιο αργά, με περίπου 21 φορές αύξηση του χρόνου εκπαίδευσης. Η διάφορα είναι η βελτίωση στην ακρίβεια, με τον SVC να βελτιώνει την ακριβεία κατά 10% ενώ το BERT 100%. με εξαίρεση τη δραματική αύξηση στον αλγόριθμο $\Sigma^{\text{ες}}$ ο οποίος εκπαιδεύτηκε 21 φορές πιο αργά. Το μοντέλο με την μεγαλύτερη ακρίβεια αποδεικνύεται το BERT, με μειονέκτημα την υπολογιστική πολυπλοκότητα, καθώς κάθε εποχη εκπαίδευσης αυξάνει τον χρόνο εκπαίδευσης. Το μοντέλο BERT διπλασίασε την ακρίβεια από 0.45 σε 0.89 με τη χρήση

περισσότερων δεδομένων.

	Accuracy	Precision	Recall	F1 Score	Training Time (s)
NB - CountVec	0.733	0.772	0.752	0.712	4.51
NB - TFIDF	0.752	0.786	0.775	0.738	4.59
Logistic Regression - CountVec	0.821	0.843	0.834	0.818	26.4
Logistic Regression - TFIDF	0.838	0.852	0.844	0.832	18.2
Logistic Regression - Doc2Vec	0.678	0.715	0.684	0.661	1.86
SVC - CountVec	0.693	0.755	0.703	0.700	80
SVC - TFIDF	0.797	0.833	0.805	0.791	112
SVC - Doc2Vec	0.693	0.710	0.699	0.678	5.91
BERT	0.891	0.891	0.893	0.887	313

Πίνακας 4.4: Σύγκριση αλγορίθμων - Επαυξημένο Σύνολο

4.5 Συζήτηση των πλεονεκτημάτων και των περιορισμών του μοντέλου

Σε τελική ανάλυση το ακριβέστερο μοντέλο φαίνεται να είναι το προεκπαιδευμένο μοντέλο αιχμής BERT προσαρμοσμένο στα δεδομένα μας. Λόγω του πλήθους των παραμέτρων χρειάζεται πολύ μεγαλύτερους υπολογιστικούς πόρους για την εκπαίδευση. Επίσης, για να φτάσει τις μέγιστες δυνατότητες χρειάζεται έναν επαρκή όγκο δεδομένων, αλλά η υπολογιστική πολυπλοκότητα του αυξάνει εκθετικά με το μέγεθος των εισόδων. Η επαύξηση δεδομένων σε συνδυασμό με το μοντέλο παράγει τα βέλτιστα αποτελέσματα σε αυτή την περίπτωση. Πολύ καλά αποτελέσματα έδειξε η λογιστική παλινδρόμηση, με υψηλή ακρίβεια από την αρχή και υψηλό λόγο απόδοσης προς υπολογιστικής πολυπλοκότητας, ακόμα και για μεγάλο όγκο δεδομένων. Βελτίωση είδαμε τόσο με την χρήση TF-IDF σε σχέση με τα αρχικά δεδομένα, όσο και από τη χρήση του επαυξημένου συνόλου. Βέβαια, όσο αυξάνεται ο αριθμός των κλάσεων, τόσο δυσκολότερη και πολυπλοκότερη η ταξινόμηση με δυαδικά μοντέλα. Η τεχνική πυκνού διανύσματος απέδωσε φτωχά αποτελέσματα στην ταξινόμηση τόσο με ΜΔΥ όσο και Λογιστική Παλινδρόμηση υπολειτουργώντας σε σχέση με τις άλλες μεθόδους, με το κόστος για τον υπολογισμό των ενσωματώσεων να είναι πολύ μεγάλο για την τελική τους απόδοση.

Ένας περιορισμός των μοντέλων είναι ότι ξεχωρίζουν τα δείγματα δεδομένου ότι αυτά θα είναι κείμενα βιογραφικών σημειωμάτων. Σε περίπτωση που υπήρχαν άλλου είδους έγγραφά, δεν έχουν τρόπο να τα ξεχωρίσουν από ένα βιογραφικό και θα ταξινομηθούν σε κάποια κατηγορία που το κάθε μοντέλο συμπεραίνει ότι βρίσκεται πλησιέστερα.

Ακόμα ένας περιορισμός της παρούσας μεθοδολογίας είναι η ταξινόμηση με μια μοναδική κλάση. Αυτός ο περιορισμός δεν είναι αδύνατο να λυθεί, καθώς τα μοντέλα, με εξαίρεση το μοντέλο SVC (Naive Bayes, Linear Regression, και το BERT) μας επιστρέφουν τις πιθανότητες για όλες τις κλάσεις, και θα γινόταν να προσαρμοστεί για ταξινόμηση πολλών κλάσεων.

Τέλος το σύστημα βασίζεται σε ένα ορισμένο σύνολο από κατηγορίες ενδιαφέροντος. Μια καλύτερη μέθοδος θα μπορούσε να εξάγει αυτόματα τις κατηγορίες με τεχνικές λανθάνουσας μοντελοποίησης θεμάτων, όπως τον αλγόριθμο Latent Dirichlet Allocation, ώστε να βρίσκει τις κατηγορίες από τα περιεχόμενα των εγγράφων και να δημιουργεί νέες όσο τα δεδομένα αλλάζουν, προσαρμόζοντας τες ανά πάσα στιγμή και παρέχοντας πληροφορίες για τις τάσεις του περιεχομένου των εγγράφων.

Κεφάλαιο 5

Συμπεράσματα

5.1 Σύνοψη των ευρημάτων

Σε γενικές γραμμές, οι κλασσικές μέθοδοι εξαγωγής χαρακτηριστικών με απλά αραιά διανύσματα πετυχαίνουν ικανοποιητικές επιδόσεις και στις περισσότερες περιπτώσεις είναι επαρκείς για την ταξινόμηση απλών κειμένων όπως τα βιογραφικά, με ακρίβεια που αγγίζει το 80%. Στα απλά διανύσματα, η αναζύγιση είναι καθολικά ευεργετική και βελτιώνει σε κάθε περίπτωση την ποιότητα των δεδομένων. Παρόμοια βελτίωση , με μοναδική εξαίρεση τον αλγόριθμο Naive Bayes, παρατηρείται με τη χρήση της γενικευμένης επαύξησης των δεδομένων , δεδομένου ενός ομοιογενούς νοηματικού πλαισίου, και θα πρέπει να χρησιμοποιείται προσεκτικά για να αποδώσει αποτελέσματα, καθώς η υπολογιστικές απαιτήσεις της επαύξησης καθώς και οι απαιτήσεις μνήμης δεν είναι αμελητέες, με το μέγεθος του μοντέλου να δεκαπλασιάζεται για μια συγκριτικά μικρότερη βελτίωση στην απόδοση. Παρατηρούμε, λοιπόν, ότι ο NB ξεπερνάει σε επιδόσεις όλους τους άλλους αλγορίθμους όταν το σύνολο εκπαίδευσης είναι μικρότερο. Η λογιστική παλινδρόμηση και οι ΜΔΥ έδειξαν βελτίωση τόσο με τα καλύτερα χαρακτηριστικά όσο και με το επαυξημένο μοντέλο, με τις επιδόσεις να είναι μέγιστες με συνδυασμό των δύο, παρουσιάζοντας παρόμοιες επιδόσεις με την λογιστική παλινδρόμηση να είναι ελάχιστα ισχυρότερη. Τα μοντέλα Naive Bayes και Logistic Regression κλιμακώνονται πολύ καλύτερα σε σχέση με το μέγεθος του συνόλου εισόδων από τις ΜΔΥ και των παραμέτρων του BERT. Τις βέλτιστες επιδόσεις επέδειξε η χρήση του γλωσσικού μοντέλου BERT με προσαρμογές και τροποποιήσεις για την περίπτωση χρήσης της εργασίας, με σημαντική διαφορά, με το μειονέκτημα της ανάγκης μεγάλου όγκου δεδομένων για την εκπαίδευση και την συγκριτικά πολύ μεγαλύτερη απαίτηση υπολογιστικών πόρων, ακόμα και στο μικρότερο μοντέλο.

5.2 Συμβολή της διατριβής

Δόθηκε έμφαση στην εισαγωγή του μοντέλου BERT στη διαδικασία ταξινόμησης βιογραφικών, της τελευταίας εξέλιξης στο πεδίο της ΕΦΓ χρησιμοποιώντας την αρχιτεκτονική του transformer. Αυτά τα μοντέλα είναι εφοδιασμένα με μια γενικότερη αντίληψη της γλωσσάς και στην εργασία μελετήθηκε η εφαρμογή τους για την ταξινόμηση κειμένου, και πως διαφέρει με

τις κλασσικές μεθοδολογίες μηχανικής μάθησης για τον διαχωρισμό εγγράφων βιογραφικών σημειωμάτων. Η μέθοδος επαύξησης κειμένου μπορεί να βοηθήσει σε περίπτωση έλλειψης επαρκούς όγκου δεδομένων η δυσκολία απόκτησης αυτών, και φαίνεται να λειτουργεί πολύ αποδοτικά σε συνδυασμό με μεγάλα γλωσσικά μοντέλα όπως το BERT. Παρότι η εκπαίδευση του μοντέλου είναι υπολογιστικά απαιτητική και χρονοβόρα, εφόσον εκπαιδευτεί μπορεί να αποθηκευτεί και να επαναχρησιμοποιηθεί σε συσκευές με περιορισμένους υπολογιστικούς πόρους. Δεδομένου ότι λόγω της προϋπόθεσης της μοναδικής ετικέτας ταξινόμησης σε κάθε κείμενο, ως ερώτημα στη μηχανή αναζήτησης για μη επιβλεπόμενη ανάθεση ετικετών στο κείμενο, θα ήταν πρακτικά αδύνατο να ταξινομηθούν τέλεια μιας και κάποια από αυτά, στο περιεχόμενο τους, δίνουν περισσότερα στοιχειά για το ότι θα άνηκαν σε κάποια άλλη κλάση. Ως εκ τούτου, θεωρούμε οι επιδόσεις που κατάφερε το μοντέλο είναι πολύ ικανοποιητικές.

5.3 Περιορισμοί και μελλοντικές κατευθύνσεις της έρευνας

Σε μελλοντική ερευνά θα μπορούσαν να ερευνηθούν μεθοδολογίες μη εποπτευόμενης εξόρυξης ετικετών για την καλύτερη συσταδοποιηση των αποτελεσμάτων μιας και η μέθοδος shortlisting με ταξινόμηση μιας ετικέτας περιορίζει τα αποτελέσματα σε κάποιο βαθμό. Επίσης θα μπορούσε να βελτιωθεί η υπολογιστική αποδοτικότητα μειώνοντας το μέγεθος των εισόδων στο BERT με τεχνικές περίληψης κειμένου (summarization). Η ερευνά σε αυτόν τον τομέα αποτελεί ανοικτό πρόβλημα[21]. Επίσης θα μπορούσαν να συνδυαστούν με μεγαλύτερα μοντέλα η μοντέλα μετάφρασης για την ταξινόμηση πολύγλωσσων βιογραφικών. Άλλη μια ιδέα θα ήταν η εξόρυξη γνώσης με τη χρήση δημιουργικών μοντέλων transformer όπως η οικογένεια GPT.

Βιβλιογραφία

[1] Pradeep Roy, Sarabjeet Chowdhary και Rocky Bhatia. A machine learning approach for automation of resume recommendation system. *Procedia Computer Science*, 167:2318–2327, 2020.

[2] The skills of leonardo da vinci, χ.χ.

[3] J. Malinowski, T. Keim, O. Wendt και T. Weitzel. Matching people and jobs: A bilateral recommendation approach. Στο *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, τόμος 6, σελίδες 137ς–137ς, 2006.

[4] Xing Yi, James Allan και W. Bruce Croft. Matching resumes and jobs based on relevance models. Στο *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, σελίδα 809–810, New York, NY, USA, 2007. Association for Computing Machinery.

[5] Rose Catherine, Karthik Visweswariah, Vijil Chenthamarakshan και Nanda Kambhatla. Prospect: A system for screening candidates for recruitment. σελίδες 659–668, 2010.

[6] Shiqiang Guo, Folami Alamudun και Tracy Hammond. Résumatcher: A personalized résumé-job matching system. *Expert Systems with Applications*, 60:169–182, 2016.

[7] Chirag Daryani, Gurneet Chhabra, Harsh Patel, Indrajeet Chhabra και Ruchi Patel. An automated resume screening system using natural language processing and similarity. σελίδες 99–103, 2020.

[8] Alon Halevy, Peter Norvig και Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24:8–12, 2009.

[9] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959.

[10] J. R. Firth. Applications of general linguistics. *Transactions of the Philological Society*, 56(1):1–14, 1957.

[11] Nancy Ide και J. Véronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24:1–41, 1998.

[12] Scott McDonald και Michael Ramscar. Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity. 2001.

[13] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957.

[14] Tomas Mikolov, Kai Chen, Greg Corrado και Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[15] Matt Kusner, Yu Sun, Nicholas Kolkin και Kilian Weinberger. From word embeddings to document distances. Στο *Proceedings of the 32nd International Conference on Machine Learning*Francis Bach και David Blei, επιμελητές, τόμος 37 στο *Proceedings of Machine Learning Research*, σελίδες 957–966, Lille, France, 2015. PMLR.

[16] Jeffrey Pennington, Richard Socher και Christopher Manning. GloVe: Global vectors for word representation. Στο *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, σελίδες 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics.

[17] Quoc V. Le και Tomas Mikolov. Distributed representations of sentences and documents, 2014.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser και Illia Polosukhin. Attention is all you need, 2017.

[19] Mike Schuster και Kaisuke Nakajima. Japanese and korean voice search. Στο *International Conference on Acoustics, Speech and Signal Processing*, σελίδες 5149–5152, 2012.

[20] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes και Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

[21] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt και Gaël Varoquaux. Api design for machine learning software: experiences from the scikit-learn project, 2013.

[22] Iulia Turc, Ming Wei Chang, Kenton Lee και Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*, 2019.