Ταξινόμηση Κειμένων με Προσαρμογή Προ-εκπαιδευμένων Γλωσσικών Μοντέλων

Η Προπτυχιακή Διατριβή κατατέθηκε στο τμήμα
Μηχανικών Πληροφοριακών & Επικοινωνιακών Συστημάτων
του Πανεπιστημίου Αιγαίου
σε μερική εκπλήρωση των απαιτήσεων για το
Δίπλωμα του
Μηχανικού Πληροφοριακών και
Επικοινωνιακών Συστημάτων

UNIVERSITY OF THE AEGEAN

Αθανάσιος Μπόνης
Γεώργιος Δημόπουλος

Μέλη επιτροπής:
Ευστάθιος Σταματάτος - Αναπληρωτής Καθηγητής (επιβλέπων),
Χρήστος Γκουμόπουλος (μέλος, ΜΠΕΣ)
και Ανδρέας Παπασαλούρος (μέλος, Τμήμα Μαθηματικών) Σεπτέμβριος 2019

Text Categorization Based on Fine-tuning of Pre-trained Language Models

A thesis submitted in partial fulfillment of the requirements for the Degree in Information and Communication Systems Engineering.

UNIVERSITY OF THE AEGEAN

Athanasios Bonis
Georgios Dimopoulos

Supervisor: Efstathios Stamatatos - Associate Professor

September 2019

Δήλωση Γνησιότητας

Δηλώνουμε ότι αυτή η διατριβή είναι το πρωτότυπο έργο μας, που συγγράφηκε ειδικά για την εκπλήρωση των σκοπών και στόχων αυτής της μελέτης και δεν έχει προηγουμένως υποβληθεί σε οποιοδήποτε άλλο Πανεπιστήμιο στην Ελλάδα ή στο εξωτερικό. Δηλώνουμε επίσης ότι οι πηγές που χρησιμοποιήθηκαν για την εκπόνηση της συγκεκριμένης διατριβής αναφέρονται στο σύνολό τους, δίνοντας πλήρεις αναφορές στους συγγραφείς, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το Διαδίκτυο.

Καρλόβασι, Σεπτέμβριος 2019

Αθανάσιος Μπόνης
Γεώργιος Δημόπουλος

Υπογραφή

``I think the brain is essentially a computer and consciousness is like a computer program. It will cease to run when the computer is turned off. Theoretically, it could be re-created on a neural network, but that would be very difficult, as it would require all one's memories."

Stephen Hawking

Περίληψη

Η Ταξινόμηση Κειμένων είναι μια σημαντική μελέτη στον τομέα της εξαγωγής πληροφορίας από κείμενα (Text Mining), έχοντας ένα μεγάλο εύρος εφαρμογής. Τα τελευταία χρόνια, μέσω της εξέλιξης αλγορίθμων νευρωνικών δικτύων (Neural Networks), έχουν αναπτυχθεί πολλές τεχνικές εξαγωγής γλωσσικών μοντέλων από μεγάλες συλλογές κειμένων γνωστά ως προ-εκπαιδευμένα γλωσσικά μοντέλα (Pre-Trained Language Models), οι οποίες βρίσκουν εφαρμογή σε ποικίλες εργασίες επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP). Την συγκεκριμένη χρονική στιγμή, η βέλτιστη πρακτική για ταξινόμηση κειμένων, π.χ. αναγνώριση συγγραφέα, είναι η εφαρμογή των Pre-Trained Language Models με την κατάλληλη προσαρμογή τους (Fine-Tuning). Στην υπάρχουσα εργασία, θα αναλύσουμε και θα εφαρμόσουμε την τεχνική του Universal Language Model Fine Tuning της ερευνητικής ομάδας του fast.ai στον τομέα του NLP, σε διάφορες εφαρμογές της κατηγοριοποιήσης κειμένου, καθώς και σύγκριση με άλλες τεχνικές του Fine-Tuning

Λέξεις κλειδιά: Text-Mining, NLP, Authorship-Attribution, Fine-Tuning, ULMFiT.

Αθανάσιος Μπόνης
Γεώργιος Δημόπουλος
Τμήμα Μηχανικών Πληροφοριακών & Επικοινωνιακών Συστημάτων
Πανεπιστήμιο Αιγαίου

Abstract

Text Categorization is an important study in the field of Text-Mining, with a wide range of applications. In recent years, through the development of Neural Networks, many techniques have been developed such as pre-trained language models, which are applicable to Natural Language Processing (NLP). Currently, the best practice for categorizing texts, e.g. writer recognition, is the application of Pre-Trained Language Models through Fine-Tuning. In this research, we analyze and present the application of the Universal Language Model Fine Tuning technique (ULMFiT) in some text categorization applications, which is developed by NLP's fast.ai research team. Furthermore, we compare this technique with others, and we conclude, presenting the results of this comparison.

Keywords: Text-Mining, NLP, Authorship-Attribution, Fine-Tuning, ULMFiT.

Athanasios Bonis
Georgios Dimopoulos
Department of Information and Communication Systems Engineering
University of the Aegean

Ευχαριστίες

Για την επιτυχή ολοκλήρωση της παρούσας Διπλωματικής Εργασίας, θα θέλαμε να ευχαριστήσουμε θερμά τον κύριο Σταματάτο, ο οποίος ήταν δίπλα μας σε κάθε στάδιο της εκπόνησης της, όπως και σε κάθε απορία ή δυσκολία που συναντήσαμε. Επιπλέον, θα θέλαμε να ευχαριστήσουμε θερμά τους γονείς μας, οι οποίοι μας βοήθησαν να φτάσουμε στον τελικό μας στόχο, καθώς και όλους τους ανθρώπους οι οποίοι μας βοήθησαν κατά την διάρκεια των σπουδών μας.

Athanasios Bonis
Georgios Dimopoulos
Department of Information and Communication Systems Engineering
University of the Aegean

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Λίστα Ακρωνύμων

NLP . . . . . **N**atural **L**anguage **P**rocessing

LM . . . . . **L**anguage **M**odel

TC . . . . . **T**ext **C**lassification

FT . . . . . **F**ine **T**uning

ULMFiT . . . . . **U**niversal **L**anguage **M**odel **F**ine **T**uning

STLR . . . . . **S**lanted **T**riangular **L**earning **R**ates

DFT . . . . . **D**isciminative **F**ine **T**uning

GU . . . . . **G**radual **U**nfreezing

AI . . . . . **A** rtificial **I**ntelligence

ΚΚ . . . . . **Κ**ατηγοριοποίηση **Κ**ειμένων

ΕΦΓ . . . . . **Ε**πεξεργασία **Φ**υσικής **Γ**λώσσας

ΤΝ . . . . . **Τ**εχνητή **Ν**οημοσύνη

ΠΓΜ . . . . . **Π**ροεκπαιδευμένα **Γ**λωσσικά **Μ**οντέλα

Κεφάλαιο 1

Εισαγωγή

Το αδιάκοπο έργο των ερευνητών σε ποίκιλα θέματα Πληροφορικής, Στατιστικής και Μαθηματικών, έχει επιφέρει σημαντικές αναβαθμίσεις σε τομείς όπως είναι η Τεχνητή Νοημοσύνη και η Μηχανική Μάθηση. Μέσω αυτής της εξέλιξης, έχει επιτευχθεί η δημιουργία νέων μεθόδων δημιουργίας αξιόπιστων μοντέλων Μηχανικής Μάθησης, τα οποία βρίσκουν εφαρμογή σε ένα μεγάλο πλήθος προβλημάτων.

Ένα απο αυτά τα προβλήματα είναι και η Κατηγοριοποίηση Κειμένου την οποία η παρούσα Διπλωματική Εργασία πραγματεύεται μέσω των εφαρμογών νέων μεθόδων Μηχανικής Μάθησης, όπως είναι το Transfer Learning και τα Pretrained Models. Πριν όμως εμβαθύνουμε, θα δώσουμε μερικούς ορισμούς πάνω σε θέματα Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης.

1.1 Τεχνητή Νοημοσύνη

Ο όρος τεχνητή νοημοσύνη αναφέρεται στον κλάδο της πληροφορικής ο οποίος ασχολείται με τη σχεδίαση και την υλοποίηση υπολογιστικών συστημάτων που μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς τα οποία υπονοούν έστω και στοιχειώδη ευφυΐα: μάθηση, προσαρμοστικότητα, εξαγωγή συμπερασμάτων, κατανόηση από συμφραζόμενα, επίλυση προβλημάτων κλπ. Ο Τζον Μακάρθι όρισε τον τομέα αυτόν ως «επιστήμη και μεθοδολογία της δημιουργίας νοημόνων μηχανών».

Η τεχνητή νοημοσύνη (AI από το Artificial Intelligence) καθιστά τις μηχανές ικανές να μαθαίνουν από την εμπειρία, να προσαρμόζονται σε νέα εισαγόμενα δεδομένα και να εκτελούν ανθρωπομορφικά έργα. Τα περισσότερα παραδείγματα AI για τα οποία ακούμε σήμερα –από τους υπολογιστές που παίζουν σκάκι έως τα αυτο-οδηγούμενα αυτοκίνητα– βασίζονται σε μεγάλο βαθμό στο deep learning και την επεξεργασία φυσικής γλώσσας (ΕΦΓ). Με τη χρήση των τεχνολογιών αυτών, οι υπολογιστές μπορούν να εκπαιδευτούν ώστε να επιτελούν συγκεκριμένα καθήκοντα με επεξεργασία μεγάλων ποσοτήτων δεδομένων και αναγνώριση μορφών στα δεδομένα [1].

Οι πρώτες κιόλας αναφορές σε θέματα Τεχνητής Νοημοσύνης και Τεχνητών Οντοτήτων, εμφανίζονται ήδη απο την αρχαιότητα, σε μύθους. Μεγάλοι αρχαίοι Έλληνες, Κινέζοι, Ινδοί φιλόσοφοι και μαθηματικοί όμως, ερεύνησαν σημαντικά το πεδίο της συλλογιστικής σκέψης, παράγοντας σημαντικό έργο για τις επόμενες γενιές. Στις έρευνες αυτές βασίστηκαν αργότερα οι ερευνητές επιστήμονες του κλάδου, αναπτύσσοντας την έννοια των λογικών μηχανών και της αλγοριθμικής σκέψης.

[1] https://www.sas.com/el$_{gr}/insights/analytics/what-is-artificial-intelligence.html$

Ο Alan Turing, πατέρας της Θεωρίας Υπολογισμού έπαιξε έναν καθοριστικό ρόλο στην εξέλιξη της σύγχρονης έννοιας της Τεχνητής Νοημοσύνης. Ο Turing παρουσίασε το 1950 στην έρευνα του Computing Machinery and Intelligence, το διάσημο Turing Test το οποίο εξετάζει την ικανότητα μιας μηχανής να παρουσιάζει ευφυή συμπεριφορά, τέτοια ώστε ένας άνθρωπος να μη μπορεί να ξεχωρίσει αν είναι μηχανή ή όχι. Η δοκιμασία αυτή θεωρείται ένα σημαντικό κεφάλαιο στην φιλοσοφία της ΤΝ.

Ως σημείο αρχής της ΤΝ, θεωρείται η χρονιά του 1956, όπου έγινε η πρώτη τυπική θεμελιώση του πεδίου, σε ένα συνέδριο στο Dartmouth College, όπου συμμετείχαν οι Ray Solomonoff, John McCarthy, Allen Newell, Marvin Minsky, Herbert Simon, Arthur Samuel κ.α, οι οποίοι έγιναν επικεφαλής της έρευνας του πεδίου της Τεχνητής Νοημοσύνης.

Σχήμα 1.1: Alan Turing

1.2 Μηχανική Μάθηση

Ως Μηχανική Μάθηση ή αλλιώς (Machine Learning) ορίζουμε τον υποκλάδο της Τεχνητής Νοημοσύνης, που επικεντρώνεται στην δημιουργία τεχνικών και αλγορίθμων οι οποίοι μαθαίνουν απο τα δεδομένα που τους δίνονται, χωρίς να είναι ρητά προγραμματισμένοι να εκτελέσουν την εκάστοτε ενέργεια.

Οι αλγόριθμοι της Μηχανικής Μάθησης είναι ικάνοι να εξάγουν προβλέψεις απο τα δεδομένα που τους δίνονται και να πάρουν αποφάσεις για καταστάσεις, μέσω αναγνώρισης προτύπων στα δεδομένα. Όπως όλο το ευρύ αχανές πεδίο της Τεχνητής Νοημοσύνης, έτσι και η Μηχανική Μάθηση έχει τις βάσεις της στην επιστήμη των Μαθηματικών και της Στατιστικής. Πιο συγκεκριμένα, είναι άρρηκτα συνδεδεμένη με τον τομέα της μαθηματικής βελτιστοποίησης, η οποία παρέχει μεθόδους, τεχνικές και θεωρήματα τα οποία βελτιώνουν συνεχώς τους αλγόριθμους της.

Μέσω τις εξέλιξης και της έρευνας στο πεδίο της Μηχανικής Μάθησης, έχουν αναπτυχθεί πολλές τεχνικές οι οποίες χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος. Οι τεχνικές αυτές χωρίζονται σε τρεις κατηγορίες[2]:

- μάθηση με επίβλεψη (supervised learning) ή μάθηση με παραδείγματα (learning from examples)
- μάθηση χωρίς επίβλεψη (unsupervised learning) ή μάθηση απο παρατήρηση (learning from observation)
- ενισχυτική μάθηση (reinforcement learning)

[2]http://aibook.csd.auth.gr/

1.3 Εξόρυξη Δεδομένων (Data Mining)

Η εξόρυξη πληροφορίας (data mining) είναι ένας ερευνητικός τομέας που προσπαθεί να επιλύσει το πρόβλημα της υπερφόρτωσης πληροφοριών με τη χρησιμοποίηση διάφορων τεχνικών. Προσπαθούμε να αποκτήσουμε τις πληροφορίες που μας ενδιαφέρουν από πολλές και αχανής πηγές πληροφοριών, όπως είναι οι βάσεις δεδομένων. Εκτός από αυτό πρέπει και τα δεδομένα που θα αποκτήσουμε να είναι κατανοητά για να μπορεσουμε να τα χειριστούμε και σωστά.

Βασικός στόχος του Data Mining είναι η ανάλυση μεγάλων όγκων δεδομένων ή αλλιώς Big Data -τα οποία με την εξέλιξη της τεχνολογίας αναπτύσσονται και μεγαλώνουν ραγδαία- ώστε να κατηγοριοποιηθούν, να εντοπιστούν τυχόν ανωμαλίες και διαφοροποιημένες εγγραφές αλλά και εύρεση προτύπων (patterns). Σαν απώτερος σκοπός του συγκεκριμένου κλάδου είναι η αυτόματη ή ημιαυτόματη ανάλυση πολλών δεδομένων για την εξαγωγή κάποιου ενδιαφέροντος νέου προτύπου. Οι πολλές πηγές και οι πολλές πληροφορίες που μπορούμε να αποκτήσουμε δεν είναι πάντα θετικό μιας και ο πολύς όγκος που συλλέγουμε από αυτές μπορεί να μας οδηγήσει σε σύγχιση με αρκετά ασήμαντα στοιχεία τελικά. Για αυτό εν κατακλείδι από τα μεγάλα σύνολα δεδομένων παίρνουμε μόνο τις λίγες και σημαντικές πληροφορίες.

Υπάρχουν πολλές χρήσεις και παραδείγματα που χρησιμοποιείται η εξόρυξη δεδομένων. Για παράδειγμα στα διαδικτυακά μηνύματα αλληλογραφίας με ανεπιθύμητο περιεχόμενο (email) χρησιμοποιείται φίλτρο που υλοποιείται με κανόνες με αλγόριθμους Data Mining. Εκεί έχει μάθει από την εξέταση εκατομμυρίων μηνυμάτων, ποια έχουν χαρακτηριστεί ως ανεπιθύμητα (spam). Ένας άλλος κλάδος όπου χρησιμοποιείται είναι στις αγορές και στο ψάρεμα υποψήφιων πελατών ή στις τράπεζες και στην έγκριση τραπεζικών προϊόντων η οποία είναι βασισμένη σε στοιχεία των αιτούντων. Τέλος είναι πολύ χρήσιμη στις φορολογικές αρχές και στην εντόπιση φορολογικών δηλώσεων που είναι πιθανόν να είναι ψευδείς. Οι τομείς που χρησιμοποιούν πλέον εξόρυξη πληροφοριών είναι πολλοί και ενδεικτικά η ιατρική, οι τηλεπικοινωνίες όπως και το χρηματιστήριο και γενικά η οικονομία εφαρμόζουν το data mining για τη συλλογή των σημαντικών δεδομένων μέσα από μεγάλο όγκο στοιχείων. Κάτι το οποίο οδηγεί στη καλύτερη απόφαση και ενέργεια σύφμωνα με τις υπάρχουσες συνθήκες κάθε φορά.

1.4 Εξόρυξη Δεδομένων από Κείμενα (Text Mining)

Το Text Mining είναι παρόμοιος τομέας με το Data Mining, μία παραλλαγή της εξόρυξης δεδομένων ουσιαστικά που περιγράφτηκε στο προηγούμενο κεφάλαιο. Ενώ το Data Mining είναι μία διαδικασία βασισμένη σε αλγορίθμους για απόσπαση και ανάλυση χρήσιμων στοιχείων από διάφορης μορφής δεδομένα, το Text Mining είναι το σύνολο των διαδικασιών που απαιτούνται για τη μετατροπή αδόμητων εγγράφων, δεδομένα σε γραπτή μορφή, σε πολύτιμες και δομημένες πληροφορίες.

Τα συστήματα στο Data Mining ασχολούνται με πηγές που θεωρούνται ομογενείς και είναι εύκολες γενικά προς την κατανόησή τους, ενώ στο Text Mining έχουμε μία νέα και πιο δύσκολη πρόκληση. Και αυτό γιατί η πρόβλεψη των διαφόρων χρήσιμων αποτελεσμάτων από τις μεγάλες βάσεις γραπτών δεδομένων περιλαμβάνει το πρόβλημα της ετερογενής δύσκολης μορφής των έγγραφων και πηγών. Πχ μία πηγή μπορεί να είναι σε μορφή email, μίας δημοσίευσης σε κοινωνικό δίκτυο ή ακόμα και ενός κλασσικού μηνύματος SMS. Έχει σίγουρα περισσότερες δυσκολίες ο τομέας του Text Mining.

Η γενική εξόρυξη δεδομένων είναι αποδεδειγμένη, ισχυρή, κατανοητή και γρήγορη

τεχνολογία για πολλές δεκαετίες. Αλλά με τη πάροδο του χρόνου η συνεχή χρησιμοποίηση του διαδικτύου και οι διαφορετικές μορφές κειμένων που περιέχει κάνουν αναγκαία πλέον και την εξόρυξη δεδομένων από κείμενα. Ακόμα και αν αυτά είναι πιο δύσκολα σε κατανόηση και ανάλυση. Οι περισσότερες πληροφορίες είναι σε κάποια μορφή κειμένου και σίγουρα η απόκτηση μόνο των σημαντικών πληροφοριών από τεράστια κείμενα και δεδομένα είναι ένα μεγάλο επίτευγμά σε κέρδος κόστους και χρόνου.

Εν κατακλείδι, παρότι οι δύο προαναφερθείσες έννοιες αλληλοσυμπληρώνουν η μία την άλλη, η εξόρυξη στοιχείων από κείμενα πλέον είναι πιο χρήσιμη, χάρη στη τεχνολογία και στις υποκατηγορίες που έχει και λύνουν δύσκολα προβλήματα σε πολλούς κλάδους. Η εξόρυξη κειμένου έχει καταστεί πρακτικότερη για τους ειδικούς αλλά και τους απλούς χρήστες λόγω της ανάπτυξης μεγάλων πλατφορμών δεδομένων και αλγορίθμων βαθιάς μάθησης Deep Learning που μπορούν να αναλύσουν μαζικά σύνολα μη δομημένων δεδομένων.

1.5 Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing)

Ως Επεξεργασία Φυσικής Γλώσσας (NLP), ορίζουμε το κοινό πεδίο ανάμεσα στην επιστήμη της Γλωσσολογίας, της Πληροφορικής και της Τεχνητής Νοημοσύνης το οποίο μελετά τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των ανθρώπινων φυσικών γλωσσών, με απώτερο σκοπό την πλήρη κατανόηση της φυσικής γλώσσας από έναν υπολογιστή, ώστε να μπορεί να εξάγει νοήματα αλλά και φυσική γλώσσα από γλωσσικά δεδομένα[1]. Με άλλα λόγια, ο κύριος στόχος της Επεξεργασίας Φυσικής Γλώσσας είναι ένας υπολογίστης να μπορεί να αποκρυπτογραφήσει κάθε έννοια της ανθρώπινης γλώσσας, όπως ο ίδιος ο άνθρωπος αντιλαμβάνεται, ώστε να μπορεί να την χρησιμοποιήσει για οποιοδήποτε όφελος.

Λόγω του μεγάλου εύρους της, ο τομέας της NLP περιλαμβάνει πολλούς υποκλάδους οι οποίοι με την σειρά τους έχουν εξελιχθεί σε αυτόνομες επιστήμες με πολλές έρευνες και έργα. Έννοιες όπως η Γλωσσικής Μοντελοποίηση (Language Modeling), Ομαδοποίηση και Ενσωμάτωση Λέξεων (Clustering and Word Embeddings) και Ανάκτηση Πληροφορίας (Information Retrieval) στελεχώνουν το ευρύ πεδίο της NLP. Ο λόγος της ευρύτητας του πεδίου της ΕΦΓ, είναι η δυσκολία του προβλήματος που προκύπτει μέσα από αυτή. Το βασικό μέρος το προβλήματος, θεωρείται η πολυπλοκότητα της ανθρώπινης γλώσσας, ώστε να μπορέσει να αποκρυπτογραφηθεί συνολικά από έναν υπολογιστή. Για παράδειγμα, πολλές φορές αποδίδεται διαφορετικό νόημά σε μια πρόταση μέσω του ύφους που χρησιμοποιείται όπως είναι το σαρκαστικό, κάτι το οποίο δεν μπορεί ο υπολογιστής να αντιληφθεί.

1.6 Ανάκτηση Πληροφορίας (Information Retrieval)

Η ανάκτηση πληροφορίας έχει να κάνει με μία ευρεία περιοχή της Επιστήμης των Υπολογιστών και επικεντρώνεται κυρίως στην παροχή εύκολης πρόσβασης σε χρήσιμες πληροφορίες για το κοινό. Υπάρχουν διάφορες πηγές από τις οποίες μπορούμε να αποκτήσουμε σημαντικά στοιχεία. Μία κύρια πηγή πληροφοριών είναι τα κείμενα.

Για χιλιάδες χρόνια ο άνθρωπος έχει οργανώσει τις πληροφορίες που συλλέγει από διάφορες πηγές κειμένων για μεταγενέστερη ανάκτηση και αναζήτηση. Γύρω στο 3000 π.Χ. χτίστηκαν οι πρώτες οι βιβλιοθήκες στη Συρία σύμφωνα με ιστορικούς. Από τότε μέχρι σήμερα οι βιβλιοθήκες έχουν επεκταθεί και έχουν ακμάσει, μιας και αποτελούν τη συλλογική μνήμη του ανθρώπινου γένους.

Πλέον έχουμε και τη συνδρομή του διαδικτύου στην ανάγνωση κειμένων και απόκτηση χρήσιμων πληροφοριών από τμήματά τους. Οι χρήσιμες λειτουργίες που προσφέρονται στους χρήστες για έξυπνη αναζήτηση και άμεσο κατέβασμα δίνουν ακόμα πιο αποτελεσματικά συμπεράσματα από τις πληροφορίες που έχουν συλλεχθεί από κείμενα στο διαδίκτυο. Και η συνεχής δημιουργία τεράστιων όγκων από κείμενα και πληροφορίες έχει συνθέσει τη μεγαλύτερη ανθρώπινη πηγή γνώσης στην ιστορία. Και για αυτό το λόγο η εύρεση ενός χρήσιμου στοιχείου δεν είναι πάντα μια απλή υπόθεση. Και σίγουρα η αναζήτηση αυτή δε χρησιμοποιείται μόνο για λόγους ψυχαγωγίας, από το μέσο χρήστη, αλλά είναι και ένα σημαντικό μέρος του επιστημονικού κλάδου του NLP.

Ο κλάδος της NLP έχει δωσει τη δυνατότητα σε ειδικούς να ανακτούν πάρα πολλές και χρήσιμες πληροφορίες από διαφόρων ειδών έγγραφα. Ο πιο βασικός στόχος ούτως ή άλλως σε αυτή την επιστήμη είναι η απόκτηση σημαντικών στοιχείων με τα οποία μπορούμε περαιτέρω να εργαστούμε σε ποικίλους τομείς και επιστημονικά έργα. Αυτά τα στοιχεία τα αποκτάμε υπόψιν είτε από εξ ολοκλήρου μη οργανωμένα δεδομένα είτε από ημιδομημένα, τα οποία συνήθως είναι μεγάλοι κειμενικοί πόροι και όπως προαναφέρθηκε ο μεγάλος όγκος δεδομένων κειμένων δυσκολέυει την απόκτηση χρήσιμων στοιχείων.

Η ανάκτηση πληροφοριών γενικά στηρίζεται στη θεωρία της τεχνητής νοημοσύνης και έχει εφαρμόζεται και στις γνωστές σε όλους μας μηχανές αναζήτησης.Τα βασικά βήματα για την ανάλυση κειμένων είναι η προεργασία, η αναπαράσταση και τέλος η εξαγωγή χαρακτηριστικών γνωρισμάτων των εξεταζόμενων κειμένων. Έχει μελετηθεί εκτεταμένα σε διάφορες ερευνητικές κοινότητες, συμπεριλαμβανομένης της επεξεργασίας φυσικής γλώσσας, της ανάκτησης πληροφοριών και της εξόρυξης Web. Διαθέτει ευρύ φάσμα εφαρμογών σε τομείς όπως η εξόρυξη βιοϊατρικής λογοτεχνίας και η επιχειρησιακή ευφυΐα.

1.7 Transfer Learning & Pretrained Models

Με τον όρο Transfer Learning προσδιορίζουμε την μέθοδο της Μηχανικής Μάθησης με την οποία ένα μοντέλο το οποίο είναι κατασκευασμένο για ένα συγκεκριμένο πρόβλημα, χρησιμοποιείται ως αρχικό μοντέλο ενός άλλου παρόμοιου προβλήματος. Τα τελευταία χρόνια η χρήση του Transfer Learning έχει γνωρίσει μεγάλη άνθηση μιας και λύνει δύο σημαντικά πρόβλήματα της Μηχανικής Μάθησης, την ταχύτητα εκμάθησης του μοντέλου και το μέγεθος των πληροφοριών που απαιτείται για την εκπαίδευση του.

Χρησιμοποιώντας μοντέλα τα οποία έχουν εκπαιδευτεί ήδη σε κάποιο παρόμοιο πρόβλημα με αυτό που καλούμαστε να λύσουμε (Pretrained Models), μειώνουμε, όπως αναφέραμε, σημαντικά τον χρόνο εκπαίδευσης του μοντέλου, καθώς η παραμετροποίηση του δεν ξεκινά απο το μηδέν αλλά έχει ήδη παραμετροποιηθεί με την χρήση ενός άλλου Dataset σε παρόμοιο πρόβλημα. Με αυτό τον τρόπο, μπορούμε να χρησιμοποιήσουμε ένα μικρό πλήθος δεδομένων πάνω σε ένα προεκπαιδευέμενο μοντέλο και να έχουμε ένα σημαντικό πλεονέκτημα χρόνου και κόστος στην εκπαίδευση του.

Η εφαρμογή του Transfer Learning θεωρείται η μεγάλη εξέλιξη στα προβλήματα Μηχανικής Μάθησης και γενικότερα Τεχνητής Νοημοσύνης καθώς αν παρατηρήσουμε την εκμάθηση ενός ανθρώπινου εγκεφάλου, θα διαπιστώσουμε οτι ο ανθρώπινος εγκέφαλος δεν μαθαίνει τα πάντα απο την αρχή, αλλά μεταδίδει γνώση την οποία κατέχει ήδη, σε ένα άλλο πρόβλημα το οποίο καλείται να λύσει.

1.8 Κατηγοριοποίηση Κειμένων

Ως Κατηγοριοποίηση Κειμένων (Text Classification) ορίζουμε την διαδικασία του προσδιορισμού της κατηγορίας στην οποία ανήκει ένα κείμενο. Εν ολίγοις, κατά την διαδικασία αυτή, αποδίδεται στο εκάστοτε κείμενο της συλλογής της οποίας επεξεργαζόμαστε, μια ετικέτα (label) ώστε να προσδιοριστεί σε ποια κατηγορία ανήκει σε σχέση με τα υπόλοιπα κείμενα της συλλογής.

Για να αποσαφηνιστεί καλύτερα η έννοια της Κ.Κ., μπορούμε να σκεφτούμε ένα σύνολο κειμένων τα οποία έχουν αποσπαστεί απο ιστοσελίδες όπου θα αποδόσουμε μια ετικέτα ανάλογα με το είδος στο οποίο ανήκει. Ένα ακόμη παράδειγμα Κ.Κ είναι τα κείμενα αφήγησης ιστοριών στα οποίο θα αποδόσουμε σε κάθε ένα απο αυτά, όπως αναφέραμε και παραπάνω, μια ετικέτα η οποία προσδιορίζει τον συγγραφέα του. Με αυτό τον τρόπο, καταφέρνουμε να διασπάσουμε σε κατηγορίες τα κείμενα μιας συλλογής κειμένων, ώστε να αξιοποιήσουμε την πληροφορία αυτή σε διάφορες άλλες εφαρμογές της Επεξεργασίας Φυσικής Γλώσσας. Πιο σφαιρικά μέσω των ετικετών που δημιουργούμε μπορούμε και να βρούμε ένα επιθυμητό κείμενο πιο γρήγορα αλλά και να βγάλουμε και πιο σύντομα συμπεράσματα μέσω της ταξινόμησης κειμένων που έχει δημιουργηθεί.

Για παράδειγμα στις διαδικτυακές συλλογές κειμένων (πχ στα google books) έχουν δημιουργηθεί τέτοιες ταξινομήσεις κειμένων για να διευκολύνουν τους αναγνώστες στην εύρεση του βιβλίου που επιθυμούν. Γενικά η ταξινόμηση στο διαδίκτυο δεν παρατηρείται μόνο για κείμενα και βιβλία, αλλά και σε άλλους τομείς όπως για μουσική και βίντεο, κατηγοριοποιώντας και εκεί τα κομμάτια και ταινίες με αντίστοιχα κριτήρια. Εν γένει, βλέπουμε ότι η ταξινόμηση σε διάφορες κατηγορίες και με πολλούς τρόπους εφαρμόζεται σε πολλούς τομείς και όχι μόνο σε κείμενα και συγγραφείς.

Κεφάλαιο 2

Κατηγοριοποίηση

Η ταξινόμηση ή κατηγοριοποίηση είναι η διαδικασία στην οποία διάφορα αντικείμενα ή έννοιες αναγνωρίζονται, διαφοροποιούνται μεταξύ τους και κατανοούνται από τον άνθρωπο. Βοηθάει πολλούς τομείς και επιστήμες να αποκτήσουν καλύτερη οργάνωση και απόκτηση χρήσιμων συμπερασμάτων. Η κατηγοριοποίηση (classification ή categorization) στα αγγλικά) είναι μία τεχνική για απόκτηση δεδομένων, κατά την οποία ένα στοιχείο ανατίθεται σε μία συγκεκριμένη κατηγορία ή σε ένα προκαθορισμένο σύνολο κατηγοριών. Γενικότερα, ο στόχος της διαδικασίας αυτής είναι η ανάπτυξη ενός μοντέλου, το οποίο θα μπορεί να χρησιμοποιείται πολλές φορές άμεσα για την κατηγοριοποίηση διαφόρων μελλοντικών δεδομένων.

Μπορούμε να χωρίσουμε τη κατηγοριοποίηση, ανεξαρτήτου αντικειμένου που μελετά, σε δύο βήματα: Εκμάθηση(Learning) και Αξιολόγηση(Classification). Στο πρώτο βήμα δημιουργείται το μοντέλο με βάση ένα σύνολο προκατηγοριοποιημένων παραδειγμάτων, τα δεδομένα εκπαίδευσης(training data). Τα δεδομένα αναλύονται από ένα αλγόριθμο κατηγοριοποίησης, προκειμένου να σχηματιστεί το επιθυμητό μοντέλο. Λόγω του ότι τα δεδομένα εκπαίδευσης ανήκουν σε μία προκαθορισμένη γνωστή κατηγορία, η κατηγοριοποίηση αποτελεί μέθοδος εποπτευομένης μάθησης(supervised learning). Το μοντέλο, που λέγεται και αλλιώς κατηγοριοποιητής(classifier), αναπαρίσταται με τη μορφή κανόνων κατηγοριοποίησης(classification rules), δέντρων απόφασης(decision trees) ή μαθηματικών τύπων.

Μετά την δημιουργία του μοντέλου, το επόμενο βήμα είναι η αξιολόγησή του. Για να επιτευχθεί αυτό, χρησιμοποιούμε τα δοκιμαστικά δεδομένα(test data) για να υπολογίστεί η ακρίβεια του μοντέλου και αυτό το πετυχαίνουμε με κατηγοριοποίηση των δοκιμαστικών δεδομένων. Έπειτα, η κατηγορία που σχηματίστηκε συγκρίνεται με την πρόβλεψη που έγινε για τα δεδομένα εκπαίδευσης, τα οποία υπόψιν είναι ανεξάρτητα από αυτά της δοκιμής. Η ακρίβεια του μοντέλου υπολογίζεται με μία σύγκριση: το ποσοστό των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά σε σχέση με το υπό εκπαίδευση μοντέλο. Αν το μοντέλο κριθεί αποδεκτό, τότε μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δειγμάτων, των οποίων η κατηγοριοποίηση είναι άγνωστη. Παρακάτω θα αναλυθούν εν συντομία κάποιες κατηγορίες που έχουν δημιουργηθεί στον κλάδο της κατηγοριοποίησης και κάποιες εφαρμογές που έχουν αναπτυχθεί με μία σύντομη μικρή ιστορική αναδρομή.

2.1 Ιστορική Αναδρομή

Η ιδέα της ταξινόμησης/κατηγοριοποίησης των οντότητων υπήρχε ήδη στα αρχαία χρόνια. Στην Ελλάδα ο Αριστοτέλης(384-322 πΧ) προπάθησε να δώσει ένα πρώτο ορισμό στο τομέα εδώ, ταξινομώντας διάφορα ζώα. Όρισε ότι οι κατηγορίες είναι διακριτές οντότητες που χαρακτηρίζονται από ένα σύνολο χαρακτηριστικών που μοιράζονται τα μέλη τους. Επιπροσθέτως οι κατηγορίες πρέπει να προσδιορίζονται με σαφήνεια αλλά και να αλληλοαποκλείονται. Έτσι κάθε οντότητα ανήκει ξεκάθαρα σε μία και μόνο μία από τις προτεινόμενες κατηγορίες.

Οι παρατηρήσεις του βασίστηκαν πιο πρακτικά σε παραδείγματα και πιο συγκεκριμένα στα χαρακτηριστικά των ζώων τα οποία και χρησιμοποίησε για να διαιρέσει τα ζώα σε διάφορες κατηγορίες. Αρχικά οι κατηγορίες ήταν μόνο δύο βασικές (ασπόνδυλα, με σπονδυλική στήλη ή κατά εκείνον τότε, "με κόκκινο αίμα ή χωρίς"), οι οποίες με τη σειρά τους διαιρέθηκαν σε άλλες πέντε η κάθε μία, η οποίες και αυτές με τη σειρά τους σε δύο τελικές υποκατηγορίες, μία για κάθε φύλο. Και ο Πλάτωνας ανέλυσε την επιστήμη της ταξινόμησης. Ανέφερε ότι μπορούμε να χωρίζουμε αντικείμενα μεταξύ τους σύμφωνα με κοινά χαρακτηριστικά που θα έχουν. Το συγκεκριμένο παράδειγμα, με τον διαχωρισμό του ζωικού βασιλείου, βοηθάει να καταλάβει οποιοσδήποτε την έννοια της κατηγοριοποίησης/ταξινόμησης. [**CategorizationHistory**]

Για να αναπτυχθούν περισσότερες θεωρίες και ορισμοί σε αυτόν τον κλάδο πρέπει να πάμε πρόσφατα, στις τελευταίες δεκαετίες, όπου ξεκίνησε η ενασχόληση στο κλάδο αυτό, χάρη και στην εξέλιξη της τεχνολογίας. Το 1969 παρουσιάστηκε νέα πρόταση για τον διαχωρισμό των όντων στην Γη και διατηρείται μέχρι και σήμερα. Προτάθηκε λοιπόν ο διαγωρισμός των ζώων σε 5 κατηγορίες από τον Αμερικάνο οικολόγο Whittaker. Ζώα, φυτά, μύκητες, πρώτιστα, προκαρυώτης (μονοκύτταρα). Λίγο αργότερα αναπτύχθηκε μία σύγχρονη παραλλαγή της κλασσικής προσέγγισης, η εννοιολογική ομαδοποίηση (Conceptual Clustering), κατά τη διάρκεια της δεκαετίας του '80, ως παράδειγμα μηχανής για την ανεξέλεγκτη μάθηση. Εδώ η κάθε ομάδα/κατηγορία όταν δημιουργείται, διαμορφώνει τις εννοιολογικές της περιγραφές και χαρακτηριστικά . Έπειτα ταξινομούνται οι οντότητες στη κάθε κατηγορία, σύμφωνα με τις περιγραφές τους.

Υπάρχουν τα τελευταία χρόνια πολλές παραλλαγές και παρόμιες θεωρίες στη κατηγοριποίηση οντοτήτων. Η εννοιολογική ομαδοποίηση δε θα πρέπει να συγχέεται με την έννοια του Data Clustering. Αλλά σχετίζεται στενά με τη θεωρία των ασαφών συνόλων(fuzzy set theory). Αναπτύχθηκε και εκείνη το ίδιο χρονικό διάστημα, και εδώ τα αντικείμενα μπορεί να ανήκουν σε μία ή περισσότερες ομάδες. Επίσης τότε αναπτύχθηκαν και δύο αντίθετες κατευθύνσεις της κατηγοριοποίησης, όπου στη μία περίπτωση έχουμε από την αρχή κάποιες δωσμένες ετικέτες/χαρακτηριστικά στις κατηγορίες (supervised learning) ενώ στην άλλη όχι (unsupervised learning).

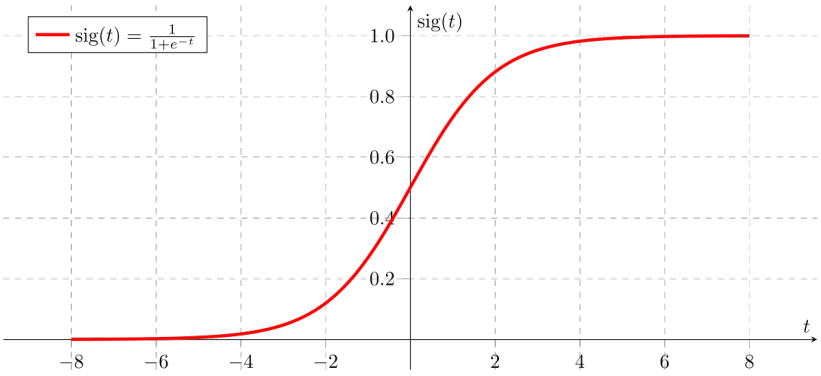
Εν κατάκλείδι μπορούμε να πούμε ότι η κατηγοριοποίηση και η μελέτη της έχει εξελιχθεί πολύ τα τελευταία χρόνια και σε διάφορους τομείς, με ίσως πιο αναπτυγμένους τους τομεία της μηχανικής όρασης και της ταξινόμησης κειμένων, όπου θα αναφερθούμε και παρακάτω πιο αναλυτικά. Υπάρχουν αρκετοί τρόποι να χαρακτηρίσουμε την ταξινόμηση κειμένων αλλά ο πιο διαδεδομένος είναι ο supervised learning/unsupervised learning.

2.2 Μεθόδοι κατηγοριοποίησης

Στον τομέα της Μηχανικής Μάθησης και της Στατιστικής υπάρχουν πολλές διαδεδομένες μέθοδοι Κατηγοριοποίησης, οι οποίες είναι κατάλληλες ανάλογα με το πρόβλημα που καλούμαστε να λύσουμε. Παρακάτω θα αναλύσουμε κάποιες απο τις βασικές μεθόδους.

2.2.1 Logistic Regression

Είναι μια στατιστική μέθοδος για την ανάλυση ενός συνόλου δεδομένων στο οποίο υπάρχουν μία ή περισσότερες ανεξάρτητες μεταβλητές που καθορίζουν ένα αποτέλεσμα. Το αποτέλεσμα μετράται με μια διχοτόμο μεταβλητή (στην οποία υπάρχουν μόνο δυο πιθανά αποτελέσματα). Ο στόχος της λογιστικής παλινδρόμησης είναι να βρεθεί το καλύτερο μοντέλο για να περιγραφεί η σχέση ανάμεσα στο διχοτόμο χαρακτηριστικό του ενδιαφέροντος (εξαρτώμενη μεταβλητή = μεταβλητή απόκρισης ή έκβασης) και ένα σύνολο ανεξάρτητων (προγνωστικών ή επεξηγηματικών) μεταβλητών. Αυτή η προσέγγιση είναι καλύτερη από άλλες δυαδικές ταξινομήσεις όπως ο πλησιέστερος γείτονας (k-means), καθώς εξηγεί επίσης ποσοτικά τους παράγοντες που οδηγούν στην ταξινόμηση.



Σχήμα 2.1: Αναπαράσταση της σιγμοειδούς συνάρτησης

2.2.2 Naive Bayes

Η μέθοδος κατηγοριοποίησης Naïve Bayes ανήκει στην κατηγορία των απλοποιημένων πιθανοτικών ταξινομητών και βασίζεται στην εφαρμογή του στατιστικού θεωρήματος Bayes. Ο Naïve Bayes έχει μελετηθεί εκτενώς από τη δεκαετία του 1960. Εισήχθη (αν και όχι κάτω από αυτό το όνομα) στην κοινότητα ανάκτησης κειμένου στις αρχές της δεκαετίας του 1960 [6] και παραμένει μια δημοφιλής μέθοδος για την κατηγοριοποίηση κειμένων. Το θεώρημα Bayes στο οποίο βασίζεται ο Απλός Ταξινομητής Bayes ή Naive Bayes, εκφράζεται απο την παρακάτω μαθηματική πιθανοτική έκφραση:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Η παραπάνω έκφραση μας δηλώνει οτι, θα βρούμε την πιθανότητα του να συμβεί το A δεδομένου οτι το B συνέβη. Η παραδοχή που γίνεται εδώ είναι ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους και συνεπώς δεν επηρεάζονται, γιαυτό και καλείται Naive Bayes.

Υπάρχουν τρείς τύποι Naive Bayes Ταξινομητών, ο Multinomial Naive Bayes, ο Bernoulli Naive Bayes και ο Gaussian Naive Bayes. Ο Multinomial βρίσκει συνήθως εφαρμογές σε Κατηγοριοποίηση Κειμένων και Εγγράφων. Τα χαρακτηριστικά τα οποία χρησιμοποιεί ο Ταξινομητής είναι η συχνότητα των λέξεων σε ένα έγγραφο. Ο Ταξινομητής Bernoulli Naive Bayes είναι παρόμοιος με τον Multinomial με την διαφορά ότι τα χαρακτηριστικά τα οποία χρησιμοποιεί είναι Boolean μεταβλητές. Τέλος ο Gaussian Naive Bayes χρησιμοποιείται όταν τα χαρακτηριστικά που θα χρησιμοποιηθούν στον ταξινομητή έχουν συνεχής τιμές και δεν είναι διακριτοί, έτσι υποθέτουμε οτι οι τιμές των χαρακτηριστικών λαμβάνονται απο μια Γκαουσιανή Κατανομή.

Γενικοτερα, ο Naive Bayes έχει αποκτήσει ιδιαίτερη δημοφιλία λόγω των πολλών χαρακτηριστικών που προσφέρει, όπως η εκπαίδευση με λίγα δεδομένα, η απλότητα και η ευκολία της υλοποίησης του καθώς και το οτι μπορεί να χρησιμοποιηθεί τόσο για δυαδικές ετικέτες αλλά και για πολλαπλές.

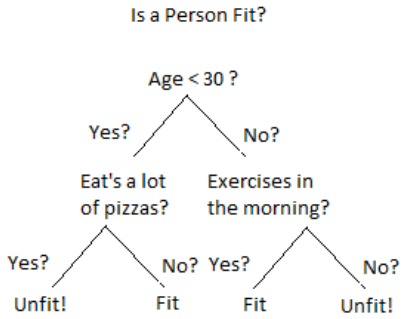
2.2.3 Δέντρα απόφασης

Το δέντρο απόφασης δημιουργεί μοντέλα ταξινόμησης ή παλινδρόμησης με τη μορφή δομής δέντρου. Διασπά ένα σύνολο δεδομένων σε μικρότερα και μικρότερα υποσύνολα ενώ παράλληλα αναπτύσσει σταδιακά ένα σχετικό δέντρο αποφάσεων. Το τελικό αποτέλεσμα είναι ένα δέντρο με κόμβους απόφασης και κόμβους φύλλων.

Ένας κόμβος απόφασης έχει δύο ή περισσότερους κλάδους και ένας κόμβος φύλλων αντιπροσωπεύει μια ταξινόμηση ή μια απόφαση. Ο κορυφαίος κόμβος απόφασης σε ένα δέντρο που αντιστοιχεί στον καλύτερο προγνωστικό παράγοντα που ονομάζεται κόμβος ρίζας. Τα δέντρα αποφάσεων μπορούν να χειριστούν τόσο τα κατηγορικά όσο και τα αριθμητικά δεδομένα.

Η απλότητα των Δέντρων Απόφασης τα καθιστά εύκολα στην κατανοήση ακόμα και απο ανθρώπους χωρίς κάποιο υπόβαθρο στην στατιστική και στα μαθηματικά, καθώς και εύκολο στην προγραμματιστική υλοποίηση τους. Επιπλέον, τα δέντρα απόφασης είναι ένας από τους πιο γρήγορους τρόπους προσδιορισμού των σημαντικότερων μεταβλητών και της σχέσης μεταξύ δύο ή περισσοτέρων μεταβλητών. Πολλές φορές τα δέντρα απόφασης χρησιμοποιούνται σε στάδιο εξερεύνησης δεδομένων.

Εκτός όμως απο τα πλεονεκτήματα τα οποία έχουν, τα δέντρα απόφασης έχουν και διάφορα μειονεκτήματα. Ένα απο αυτά είναι οτι κατά την εκμάθηση δημιουργούν υπερβολικά πολύπλοκα δέντρα που δεν γενικεύουν καλά τα δεδομένα, πρόβλημα το οποίο καλείται overfitting.



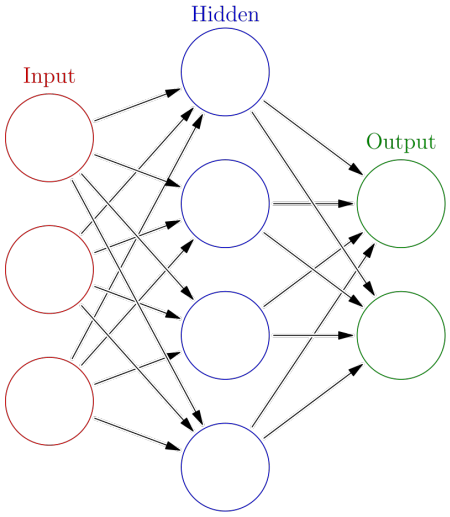
Σχήμα 2.2: Παράδειγμα αναπαράστασης ενός Δέντρου Απόφασης

2.2.4 Νευρωνικά δίκτυα

Βασισμένα στην λειτουργία της εκάθησης του ανθρώπινου εγκεφάλου, τα Τεχνητά Νευρωνικά Δίκτυα έχουν γνωρίσει μεγάλη ανάπτυξη τα τελευταία χρόνια. Ο εγκέφαλος είναι ένας πολύπλοκος, μη γραμμικός και παράλληλος υπολογιστής. Έχει την ικανότητα να οργανώνει τα δομικά συστατικά του, γνωστά ως νευρώνες, ώστε να εκτελεί ορισμένους υπολογισμούς (π.χ. αναγνώριση προτύπων, αντίληψη και έλεγχος κινητήρα) πολλές φορές ταχύτερα από τον ταχύτερο ψηφιακό υπολογιστή που υπάρχει σήμερα.[**NeuralNetworks**]

Ένα Τεχνητό Νευρωνικό Δίκτυο αποτελείται εξίσου από νευρώνες, διευθετημένους σε στρώματα, οι οποίοι μετατρέπουν ένα διάνυσμα εισόδου σε κάποια έξοδο. Κάθε μονάδα λαμβάνει μια είσοδο, εφαρμόζει μια μη γραμμική συνάρτηση σε αυτήν και στη συνέχεια μεταδίδει την έξοδο στο επόμενο στρώμα. Σε γενικές γραμμές, τα δίκτυα ορίζονται ως τροφοδοτικά: ο νευρώνας τροφοδοτεί την παραγωγή του σε όλους τους νευρώνες του επόμενου στρώματος, αλλά δεν υπάρχει ανάδραση στο προηγούμενο στρώμα. Οι διορθωτικοί συντελεστές εφαρμόζονται στα σήματα που περνούν από τον ένα νευρώνα στον άλλον και είναι αυτοί οι συντελεστές στάθμισης που συντονίζονται στη φάση της εκπαίδευσης για να προσαρμόσουν ένα νευρικό δίκτυο στο συγκεκριμένο πρόβλημα.

Ο βασικός στόχος των Τεχνητών Νευρωνικών Δικτύων είναι να λύνει προβλήματα με τον ίδιο τρόπο με τον οποίο λύνει ο ανθρώπινος εγκέφαλος. Παρόλαυτά, με την πάροδο του χρόνου η εφαρμογή τους εστιάστηκε σε ξεχωριστές εργασίες, οδηγώντας σε αποκλίσεις απο την προσέγγιση της βιολογίας.



Σχήμα 2.3: Η δομή ενός Νευρωνικου Δικτύου

2.3 Εφαρμογές Κατηγοριοποίησης

Η Κατηγοριοποίηση βρίσκει εφαρμογή σε δύο μεγάλα πεδία τα οποία απασχολούν την επιστημονική κοινότητα για πολλούς αιώνες, την ανάλυση εικόνας και την ανάλυση κειμένου. Αναλύοντας τα δύο πεδία, μπορούμε να παρατηρήσουμε οτι υπάρχει μια πληθώρα ερευνών και ανάπτυξης δίνοντας μας μεγάλα πλεονεκτήματα και λύσεις σε προβλήματα.

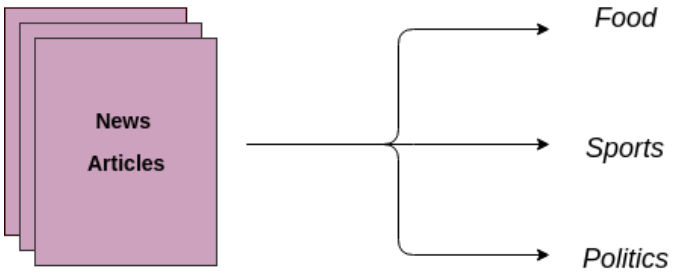
Ο διαχωρισμός των mails σε κανονικά και spam, η κατηγοριοποίηση μουσικών τραγουδιών ανάλογα με το είδος τους και ο διαχωρισμός καρκινικών κυττάρων χαρακτηρίζοντας τα ως καλοηθή ή κακοηθή είναι μερικές μόνο απο τις εφαρμογές της Κατηγοριοποίησης. Σε κάθε επιστήμη μπορούμε να βρούμε ένα μεγάλο πλήθος εφαρμογών της Κατηγοριοποίησης, καθώς βοηθά σε μεγάλο βαθμό στην κατανόηση της πληροφορίας της οποίας διαθέτουμε, μέσω της δόμησης της οποίας παρέχει. θα περιγράψουμε αναλυτικά της εφαρμογές της τόσο για το πεδίο της Κατηγοριοποίησης Εικόνας αλλά και για το πεδίο της Κατηγοριοποίησης Κειμένου το οποίο πραγματεύεται η παρούσα Διπλωματική Εργασία.

2.3.1 Κατηγοριοποίηση Εικόνας

Κατηγοριοποίηση Εικόνας αναφέρεται στην διαδικασία στην υπολογισική όραση η οποία μπορεί να ταξινομήσει μια εικόνα σύμφωνα με το περιεχόμενο της. Για παράδειγμα, ένας αλγόριθμος Κατηγοριοποίησης Εικόνας μπορεί να σχεδιαστεί για να ανιχνεύση εάν μια εικόνα περιέχει κάποιο ζώο ή όχι. Γενικότερα, στον τομέα της Υπολογιστικής Όρασης, η προσοχή εστιάζεται στην εκμάθηση ενός μοντέλου να αναγνωρίζει χαρακτηριστικά σε εικόνες και σε βίντεο επεξεργάζοντας την πληροφορία μέσω των pixels, αναπαριστώντας τα ως μαθηματικούς πίνακες. όλα τα τμήματα της Επεξεργασίας Εικόνας, Μηχανικής και Υπολογιστικής Όρασης, η Κατηγοριοποίηση βρίσκει εφαρμογή, μιας και είναι ενα βασικό στάδιο για την εκμάθηση ενός μοντέλου πρόβλεψης και ανίχνευσης χαρακτηριστικών. Εφαρμογές όπως αυτόνομη οδήγηση, ρομποτική αγρό-καλλιέργια, τραπεζικά συστήματα, βιομηχανία και υγεία έχουν ως βασικό βήμα την κατηγοριοποίηση εικόνας για την εκπαίδευση των μοντέλων μάθησης.

2.3.2 Κατηγοριοποίηση Κειμένων

Η Κ.Κ χρησιμοποιείται σε ένα ευρύ φάσμα εφαρμογών όπως είναι η ταξινόμιση τίτλων ειδήσεων ή tweets, η ταξινόμιση κριτικών πελατών σε αγορές, καθώς και η ανάλυση νομικών πλαισίων. Κ.Κ περιλαμβάνεται μεταξύ άλλων, η ανάλυση συναισθημάτων του κειμένου και η κατηγοριοποίηση του ανάλογα με το ύφος του, η απόδοση ετικέτας ανάλογα με την κατηγορία ενός κειμένου, όπως αναφέραμε σε προηγούμενο Κεφάλαιο για παράδειγμα, η αναγνώριση συγγραφέα απο κείμενο, καθώς και η ταξινόμιση ιστοσελίδων με βάση το είδος τους. Περιλαμβάνεται επίσης η αναγνώριση της γλώσσας του κειμένου αποδίδοντας μια ετικέτα ανάλογα με την γλώσσα, ακόμη και ο χαρακτηρισμός μιας είδησης αν είναι ψευδής ή οχι. επόμενο Κεφάλαιο, θα αναλύσουμε διεξοδικά την Κατηγοριοποίηση Κειμένου, ως προς την προέλευση της, τις μεθόδους εφαρμογής της και τις έρευνες που έχουν αναπτυχθεί πάνω σε αυτό το πεδίο.



Σχήμα 2.5: Παράδειγμα Κατηγοριοποίησης Κειμένου.

Κεφάλαιο 3

Κατηγοριοποίηση Κειμένων

Η Κατηγοριοποίηση Κειμένων γνωρίζει μια μεγάλη άνθηση στην σημερινή εποχή, με πολλά είδη μεθόδων και τεχνικών να καλύπτουν το φάσμα της. Σε αυτό το Κεφάλαιο, θα καλύψουμε ιστορικά την ΚΚ, θα αποσαφηνίσουμε έννοιες οι οποίες την διέπουν, καθώς και θα αναλύσουμε μεθόδους και τεχνικές οι οποίες επικρατούν αυτή την περίοδο ως βέλτιστες στον χώρο.

3.1 Ιστορική Αναδρομή

Παρότι η ανάγκη για ορθή ταξινόμηση κειμένων έχει ξεκινήσει να μελετάται παραπάνω από μισό αιώνα, τα τελευταία χρόνια και χάρη στη τεχνολογία και στο διαδίκτυο έχει παρουσιάσει εκπληκτικά αποτελέσματα. Οι εφαρμογές του (Text Categorization) έχουν γνωρίσει άνθηση τα τελευταία είκοσι περίπου χρόνια, αν και ξεκίνησαν να εφαρμόζονται στις αρχές της δεκαετίας του ‘60. Αργότερα και προς το τέλος της δεκαετίας του 90 εφαρμόστηκαν με μεγάλη επιτυχία διάφορες τεχνικές της Μηχανικής Μάθησης(Machine Learning).

Οι αρχικές τεχνικές έδωσαν τη θέση τους σιγά σιγά στις τεχνικές Μηχανικής Μάθησης. Αρκετές μέθοδοι έχουν μελετηθεί και αναπτυχθεί τα τελευταία χρόνια, από τις οποίες άλλες λιγότερο και άλλες περισσότερο σημειώνουν επιτυχία στην επίλυση του προβλήματος. Πλέον με τη παρουσία του διαδικτύου και τη συνεχή χρήση από τον μέσο χρήστη υπολογιστή, χρησιμοποιείται η κατηγοριοποίηση κειμένου σε πολλούς τομείς και περιπτώσεις που ίσως δε το έχουμε σκεφτεί. Σε μία αναζήτηση πχ στο ίντερνετ για ένα προϊόν έχουμε συνδυασμό των επιστημών Ταξινόμησης Κειμένων και της Ανάκτησης Πληροφορίας.

Ο ρυθμός αύξησης των αριθμών των κειμένων που είναι διαθέσιμα σε ηλεκτρονική μορφή είναι τεράστιος τη σημερινή εποχή. Άρα το προσφερόμενο υλικό προς εξέταση μεγαλώνει και οι ερευνητές μπορούν να αποκτήσουν και καλύτερα αποτελέσματα χάρη στη μεγάλη γκάμα συγγραμμάτων. Πολλά από τα αυτά τα συγγράμματά τίθενται στο κοινό χωρίς κάποιο αντίκτυπο. Άρα ο συνδυασμός του μεγάλου αριθμού διαθέσιμων συγγραμμάτων και η δωρεάν απόκτησή τους δίνει τη δυνατότητα σε οποιονδήποτε θέλει να ασχοληθεί με το τομέα αυτό και με καλά αποτελέσματα κιόλας.

3.2 Τύποι Κατηγοριοποίησης Κειμένων

Με τη πάροδο του χρόνου και τις διάφορες έρευνες δημιουργήθηκε ένα νέο και εξίσου σημαντικό είδος κατηγοριοποίησης. Ένα κείμενο να μπορεί να έχει πολλές ετικέτες, με άλλα λόγια στοιχεία από διάφορες κατηγορίες ταυτόχρονα (Multi-Label Classification) [1]. Πολλά συγγράμματά περιέχουν πλέον πάνω από ένα χαρακτηριστικό και δεν υπάρχει κανένας περιορισμός ως προς τον αριθμό των κατηγοριών στις οποίες μπορεί να αναγνωριστεί το έγγραφο. Ενώ στη περίπτωση μονής ετικέτας Single-Label Classification έχουμε την συνθήκη ότι για κάθε έγγραφο υπάρχει αντιστοίχηση σε μία κατηγορία και μόνο.

Υπάρχει και η δυαδική κατηγοριοποίηση κειμένου, μία υποκατηγορία της περίπτωσης της μονής ετικέτας, όπου ένα έγγραφο αντιστοιχείται σε μία κατηγορία ή στο συμπλήρωμα της. Να τονιστεί εδώ, ότι ένας αλγόριθμος για κατηγοριοποίηση πολλαπλής ετικέτας δε μπορεί να χρησιμοποιηθεί για κατηγοριοποίηση μονής ετικέτας. Ενώ ένας αλγόριθμος μονής ή δυαδικής ετικέτας μπορεί να χρησιμοποιηθεί για κείμενα με πολλές ετικέτες κατηγοριών. Απλά πρέπει να μετατραπεί σε πολλά και ανεξάρτητα προβλήματα της μονής κατηγοριοποίησης των κατηγοριών.

Αν συγκρίνουμε τις δύο μορφές κατηγοριοποίησης ως προς τον αριθμό των ετικετών, ο πιο συχνός τύπος ταξινόμησης εγγράφων είναι ο δυαδικός, μιας και είναι πιο απλός και αποτελεσματικός ως προς τη χρήση του. Δεύτερον έχει παρατηρηθεί ότι η βιβλιογραφία που έχει να κάνει με την αυτόματη κατηγοριοποίηση κειμένων περιέχει κυρίως όρους της μονής ετικέτας. Τέλος και όπως προαναφέρθηκε, ένα θετικό της δυαδικής κατηγοριοποίησης είναι ότι η επίλυσής της επιφέρει συγχρόνως και επίλυση της πολλαπλής κατηγοριοποίησης. Καταλήγουμε όμως ότι πάντα παίζουν ρόλο οι συνθήκες, η δοσμένη βάση δεδομένων με τα κείμενα και τους υποψήφιους δημιουργούς τους και τα διαθέσιμα εργαλεία που έχει στη διάθεσή του ο ερευνητής για τη τελική επιλογή ως προς το τύπος κατηγοριοποίησης κειμένων και όχι μόνο.

Η επεξεργασία φυσικής γλώσσας είναι μια από τις πιο δημοφιλείς εφαρμογές της ΤΝ. Η ιδέα του να επικοινωνεί κάποιος με τον υπολογιστή και να τον ελέγχει μιλώντας τη μητρική του γλώσσα ή κάποια ευρύτερα ομιλούμενη, όπως τα αγγλικά, είναι πολύ ελκυστική. Όμως, η φυσική γλώσσα έχει διττή φύση (ως προς τη σύνταξη και ως προς τη σημασιολογία), γεγονός που δεν εμποδίζει μεν την επεξεργασία της, αλλά δημιουργεί προβλήματα στην κατανόησή της, με αποτέλεσμα να καθίσταται το εγχείρημα της επεξεργασίας και παράλληλα της κατανόησης της ιδιαίτερα δύσκολο.

Επιπροσθέτως υπάρχουν και άλλα είδη ως προς τον τρόπο που μπορούν να συγκριθούν και να ταξινομηθούν τα συγγράμματα, όπως η κατηγοριοποίηση διαρθρωμένη κατά κείμενα ή κατά κατηγορίες, που αφορά τον τρόπο που χρησιμοποιείται ένας ο κατηγοριοποιητής κειμένων. Στη πρώτη περίπτωση θέλουμε να βρούμε όλες τις κατηγορίες από τις οποίες χαρακτηρίζεται ένα κείμενο ενώ στη δεύτερη περίπτωση θέλουμε όλα τα κείμενα που κατηγοριοποιούνται σε αυτή. Υπάρχει και η σύγκριση αυστηρής εναντίων της κατηγοριοποίησης κατάταξης όπως και η ιεραρχική κατηγοριοποίηση κειμένων εναντίων της επίπεδης. Όμως η πιο συνηθισμένη σύγκριση στο είδος ταξινόμησης κειμένων είναι κατηγοριοποίηση αναλόγως με τον αριθμό κατηγοριών που χαρακτηρίζουν ένα κείμενο, δηλαδή εάν έχει το έγγραφο μονή ή πολλαπλή ετικέτα. Στη παρούσα εργασία δε θα έχουμε να εξετάσουμε τον αριθμό κατηγοριών που χαρακτηρίζουν ένα κείμενο, μιας και το είδος είναι ένα και δοσμένο.

Επιπρόσθετα παρουσιάζεται παρακάτω, με εικόνες για καλύτερη κατανόηση, ένα πρόβλημα το οποίο ξεκινά με πολλές ετικέτες για το κάθε αντικείμενο/οντότητα και μετατρέ-

[1] https://en.wikipedia.org/wiki/Multi-label$_{c}lassification$

πεται έπειτα σε πρόβλημα μονών ετικετών. Έτσι όσες οντότητες έχουν στην αρχή τον ίδιο αριθμό κοινών ετικετών, έπειτα θα έχουν την ίδια ετικέτα [2].

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0

Σχήμα 3.1: Πίνακας επίλυσης κατηγοριοποίησης κειμένων με πολλές ετικέτες

X	y1
x1	1
x2	2
x3	3
x4	1
x5	4
x6	3

Σχήμα 3.2: Πίνακας επίλυσης κατηγοριοποίησης κειμένων με μία ετικέτα

[2]https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/

3.3 Επεξεργασία φυσικής γλώσσας(NLP)

Η Επεξεργασία Φυσικής Γλώσσας (ΕΦΓ),στα αγγλικά Natural Language Processing, είναι ένας διεπιστημονικός κλάδος της επιστήμης της πληροφορικής, της τεχνητής νοημοσύνης και της υπολογιστικής γλωσσολογίας και ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των ανθρώπινων (φυσικών) γλωσσών. Η Επεξεργασία Φυσικής Γλώσσας ξεκινά την ιστορία της στις αρχές της δεκαετίας του 1950, ως ένας συνδυασμός της Τεχνητής Νοημοσύνης και της Γλωσσολογίας [3] . Κατά συνέπεια, η ΕΦΓ συνδέεται στενά με την αλληλεπίδραση ανθρώπου-υπολογιστή. Προκλήσεις στην ΕΦΓ περιλαμβάνουν την κατανόηση φυσικής γλώσσας, δηλαδή την προσπάθεια να καταστούν ικανοί οι υπολογιστές να εξάγουν νοήματα από ανθρώπινα ή γλωσσικά δεδομένα, αλλά και την παραγωγή φυσικής γλώσσας.

Η επεξεργασία φυσικής γλώσσας είναι μια από τις πιο δημοφιλείς εφαρμογές της ΤΝ. Η ιδέα του να επικοινωνεί κάποιος με τον υπολογιστή και να τον ελέγχει μιλώντας τη μητρική του γλώσσα ή κάποια ευρύτερα ομιλούμενη, όπως τα αγγλικά, είναι πολύ ελκυστική. Όμως, η φυσική γλώσσα έχει δύο μορφές. Μία ως προς τη σύνταξη και μία ως προς τη σημασιολογία, γεγονός που δεν εμποδίζει μεν την επεξεργασία της, αλλά δημιουργεί προβλήματα στην κατανόησή της, με αποτέλεσμα να καθίσταται ιδιαίτερα δύσκολο το εγχείρημα της επεξεργασίας και της κατανόησης της [1].

Πλήθος τομέων μπορούν να επωφεληθούν από τη χρήση της, με κυριότερο την επικοινωνία ανθρώπου-μηχανής. Κάποια από τα πεδία έρευνας έχουν εφαρμογές στην απλή καθημερινή ζωή. Εδώ η χρήση της φυσικής γλώσσας επιτρέπει στους χρήστες να χρησιμοποιούν απλώς τη γλώσσα τους, όπως μιλάνε δηλαδή, και όχι κάποιο περίπλοκο και τεχνητό σύστημα όπως κάποια γλώσσα προγραμματισμού. Μία άλλη χρήση της ΕΦΓ είναι η διαχείριση πληροφορίας, όπου αυτόματες διαδικασίες ενεργοποιούνται για διαχείριση και επεξεργασία πληροφοριών με βάση το νόημα που βγάζει τα δοσμένα δεδομένα. Για παράδειγμα, αν ένα σύστημα κατανοήσει το νόημα ενός εγγράφου, θα μπορεί και να το αρχειοθετήσει μαζί με τα άλλα αντίστοιχα έγγραφα.

Προβλήματα υπάρχουν ακόμα πολλά στην επεξεργασία φυσικής γλώσσας. Η ασάφεια στη γλώσσα και οι διάφορες ερμηνείες που δίνονται σε μία πρόταση είναι ένα σημαντικό εμπόδιο. Πρώτον σε επίπεδο λεξιλογικό όπου μία λέξη μπορεί να έχει πολλές έννοιες αναλόγως τη χρήση της. Σε σημασιολογικό επίπεδο, όπου μία πρόταση έχει μεταφορική έννοια ή όχι και ο υπολογιστής δύσκολα μπορεί να το καταλάβει). Υπάρχει και η δυσκολία στο αναφορικό επίπεδο μίας πρότασης, πχ σε ποιον αναφέρεται το υποκείμενο "αυτοί" σε μία φράση. Τέλος υπάρχει και το πραγματολογικό επίπεδο. Εδώ δε προσδιορίζεται ακριβώς η σημασία μίας λέξης, είναι αυθαίρετη, για παράδειγμα όταν λέμε σε πολλά χρόνια από τώρα, πόσα εννοούμε ακριβώς; Τέλος μπορεί μία πρόταση να συνδυάζει αρκετά από τα παραπάνω επίπεδα και να δημιουργεί πολύ μεγάλη ασάφεια στη πρόταση με διφορούμενες έννοιες, άρα και η επεξεργασία της από έναν υπολογιστή γίνεται δύσκολη.

[3] https://repository.kallipos.gr/bitstream/11419/3385/1/02$_{c}hapter_{0}$7.*pdf*

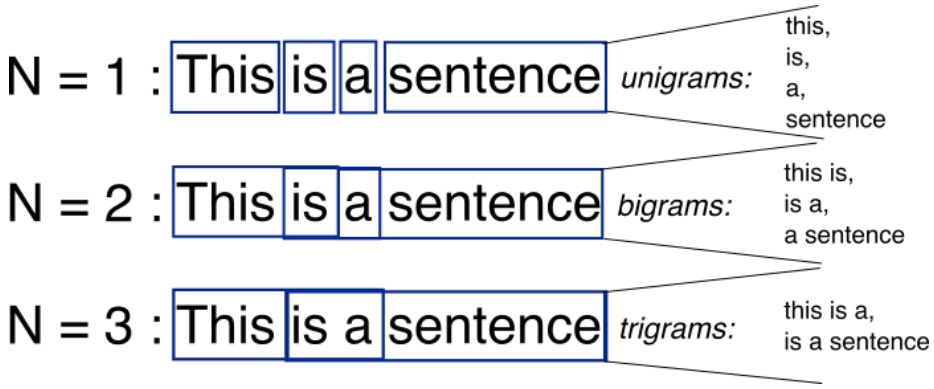
3.4 Προ-επεξεργασία Δεδομένων

3.4.1 Θεωρία N Gram

Πριν προχωρήσουμε και αναλύσουμε τη προτεινόμενη μέθοδο για αποτελεσματική εξαγωγή σημαντικών στοιχείων που υπερτερεί των παραπάνω μεθόδων ας γίνει κατανοητή η έννοια της Θεωρίας του n gram και του περιεχόμενού της. Ένα μοντέλο τύπου n gram προσπαθεί να καταλάβει το επόμενο/προηγούμενο στοιχείο σε μία δοσμένη πρόταση. Συσχετίζεται με την ακολουθία Markov, ένα μαθηματικό σύστημα, όπου η κύρια ιδέα είναι ότι για μία κατάσταση χ δε αποθηκεύονται οι προηγούμενες μεταβολές της και η επόμενη κατάσταση της εξαρτάται μόνο από τη χ. Κάτι σαν μερική αμνησία κατά πολλούς, θυμόμαστε δηλαδή για κάτι κατάσταση μόνο τη προηγούμενή της. Το εξεταζόμενο στοιχείο μπορεί να είναι μία συλλαβή ή μέχρι και ένα ζεύγος λέξεων [23].

Μέσω του N Gram μπορούμε να εξετάσουμε γενικά τη σημασία λέξεων μέσω των γειτόνων της (λέξεις που ακολουθούν μετά ή προηγήθηκαν), μιας και ο υπολογιστής δε μπορεί σε όλες τις περιπτώσεις να ξέρει την έννοια μίας λέξης όπως ο άνθρωπος. Υπάρχουν λέξεις πχ που έχουν δύο και παραπάνω έννοιες (για παράδειγμα η λέξη ουρά). Πλεονεκτήματά της μεθόδου αυτής θεωρούνται από ειδικούς η απλότητα και η κλιμάκωση. Και αυτό γιατί απλοποιούμε στο μέγιστο το πρόβλημα πρόβλεψης των επόμενων όρων (μόνο από το προηγούμενο ακριβώς!). Ο αριθμός n είναι το πλήθος των λέξεων του εξεταζόμενου μοντέλου (πχ η φράση to be or not είναι 4 gram, ο αριθμός των λέξεων που εξετάζουμε). Υπόψιν παίζουν μεγάλο ρόλο στην εξέταση μίας n gram φράσης και της σημασίας της τα σημεία στίξης όπως και οι αρνήσεις/σαρκασμοί μέσα στη πρόταση.

Τέλος το μοντέλο αυτό μοντελοποιεί ακολουθίες χρησιμοποιώντας τις στατιστικές ιδιότητες των η-grams. Εν κατακλείδι, τα n gram μοντέλα μας δίνουν μέσω κατάλληλων εκπαιδευτικών υπολογιστικών μηχανών να μελετήσουμε μία πρόταση και μέσω των λέξεων της ο υπολογιστής να παρουσιάσει το νόημα μιας εξεταζόμενης φράσης.



Σχήμα 3.3: Είδη N Gram αναλόγως με τον αριθμό των εξεταζόμενων n λέξεων
Πηγή: http://recognize-speech.com/language-model/n-gram-model/comparison

3.5 Υπάρχουσες προσεγγίσεις/Αλγόριθμοι

Όπως περιγράφηκε στο προηγούμενο Κεφάλαιο, στην Ενότητα των μεθόδων Κατηγοριοποίησης, υπάρχουν πολλά ήδη αλγορίθμων οι οποίοι εφαρμόζονται και στην περίπτωση των κειμένων: Η Λογιστική Παλινδρόμηση (Logistic Regression), τα Δέντρα Απόφασης, ο Naive Bayes, τα Νευρωνικά Δίκτυα και διάφορες ακόμη προσεγγίσεις. Η παρούσα εργασία πραγματεύεται την εφαρμογή των Νευρωνικών Δικτύων με την χρήση προηγμένων μεθόδων όπως το Transfer Learning και το Fine Tuning, τα οποία όπως έχουμε αναλύσει, έχουν φέρει επαναστατικά αποτελέσματα στον τομέα της Μηχανικής Μάθησης. Σε αυτή την ενότητα θα αναλύσουμε την χρήση διάφορων μοντέλων ταξινομητών που χρησιμοποιούνται στην Κ.Κ, καθώς και ταξινομητές οι οποίοι περιγράφηκαν στο προηγούμενο Κεφάλαιο.

3.5.1 Δέντρα Απόφασης

Όπως αναφέρθηκε και στην ανάλυση του ορισμού στο Κεφάλαιο 2, ένα δέντρο απόφασης είναι ουσιαστικά μια ιεραρχική αποσύνθεση των δεδομένων, στην οποία χρησιμοποιείται ένα κατηγόρημα ή μια συνθήκη για την τιμή του χαρακτηριστικού προκειμένου να διαιρέσει ιεραρχικά τα δεδομένα.

Στην περίπτωση των δεδομένων κειμένου (text datasets), τέτοια κατηγορήματα είναι συνθήκες σχετικά με την παρουσία ή την απουσία μιας ή περισσότερων λέξεων στο έγγραφο. Η κατανομή των δεδομένων πραγματοποιείται αναδρομικά στο δέντρο των αποφάσεων, μέχρις ότου οι κόμβοι των φύλλων να περιέχουν έναν ορισμένο ελάχιστο αριθμό εγγραφών. Η ετικέτα κλάσης στον κόμβο φύλλων χρησιμοποιείται για τους σκοπούς της ταξινόμησης (Classification).

Στην περίπτωση μιας δεδομένης δοκιμής, εφαρμόζουμε την ακολουθία των κατηγορημάτων στους κόμβους, προκειμένου να διασχίσουμε μια διαδρομή του δέντρου με τρόπο από πάνω προς τα κάτω και να προσδιορίσουμε τον σχετικό κόμβο των φύλλων. Προκειμένου να μειωθεί η περίπτωση του overfitting, μερικοί από τους κόμβους του δέντρου μπορούν να αφαιρεθούν (κοπούν) κρατώντας ένα μέρος των δεδομένων, τα οποία δεν χρησιμοποιούνται για την κατασκευή του δέντρου. Ειδικότερα, εάν η κατανομή κλάσης στα δεδομένα εκπαίδευσης (Training data) που χρησιμοποιήθηκε για την κατασκευή του δέντρου απόφασης είναι πολύ διαφορετική από την κατανομή κλάσης στα δεδομένα εκπαίδευσης που χρησιμοποιείται για κλάδεμα, τότε θεωρείται ότι ο κόμβος υπερκαλύπτει (overfit) τα δεδομένα εκπαίδευσης. Σε μια τέτοια περίπτωση ο κόμβος μπορεί να αφαιρεθεί.

Στη περίπτωση δεδομένων κειμένου, τα κατηγορήματα για τους κόμβους τυπικά ορίζονται στους όρους της συλλογής κειμένων. Για παράδειγμα, ένας κόμβος μπορεί να χωριστεί στους κόμβους των παιδιών του, ο οποίος εξαρτάται από την παρουσία ή την απουσία ενός συγκεκριμένου όρου στο έγγραφο.[12]

3.5.2 Ταξινομητές βασισμένη σε μοτίβα (Rule-based Classifiers)

Σε ταξινομητές που βασίζονται σε κανόνες, ο χώρος δεδομένων διαμορφώνεται με ένα σύνολο κανόνων, όπου η αριστερή πλευρά αποτελεί προϋπόθεση για το υποκείμενο σύνολο χαρακτηριστικών και η δεξιά πλευρά είναι η ετικέτα κλάσης.

Το σύνολο κανόνων είναι ουσιαστικά το μοντέλο που παράγεται από τα δεδομένα εκπαίδευσης. Για μια δεδομένη δοκιμαστική περίπτωση, καθορίζουμε το σύνολο των κανόνων για τους οποίους το παράδειγμα δοκιμής ικανοποιεί την προϋπόθεση στην αριστερή πλευρά του κανόνα. Καθορίζουμε την προβλεπόμενη ετικέτα κλάσης ως συνάρτηση των ετικετών κλάσης των κανόνων που ικανοποιούνται από την δοκιμαστική παρουσία.

Στην πιο γενική μορφή της, η αριστερή πλευρά του κανόνα είναι μια δυαδική κατάσταση, η οποία εκφράζεται σε Διαζευκτική Κανονική Μορφή (Disjunctive Normal Form). Ωστόσο, στις περισσότερες περιπτώσεις, η κατάσταση στην αριστερή πλευρά είναι πολύ απλούστερη και αντιπροσωπεύει ένα σύνολο όρων, το σύνολο των οποίων πρέπει να υπάρχει στο έγγραφο για να πληροίτε η προϋπόθεση.[12]

3.5.3 Naive Bayes

Όπως περιγράψαμε και στην ενότητα που αναφερθήκαμε στους αλγορίθμους ταξινόμησης, ο ταξινομητής Naive Bayes είναι ίσως ο απλούστερος και επίσης ο συνηθέστερα χρησιμοποιούμενος γενετικός ταξινομητής. Μονελοποιεί τη κατανομή των εγγράφων σε κάθε τάξη χρησιμοποιώντας ένα πιθανοτικό μοντέλο με τον ισχυρισμό οτι είναι τα δεδομένα ανεξάρτητα. Δύο κατηγορίες μοντέλων χρησιμοποιούνται συνήθως για την ταξινόμηση του Naive Bayes. Και στις δύο περιπτώσεις τα δύο μοντέλα υπολογίζουν ουσιαστικά την πιθανότητα μιας κλάσης, με βάση τη κατανομή των λέξεων στο έγγραφο. Αυτά τα μοντέλα αγνοούν την πραγματική θέση των λέξεων στο έγγραφο και λειτουργούν με την μέθοδο "bag of words". Η κύρια διαφορά μεταξύ αυτών των δύο μοντέλων είναι η παραδοχή όσον αφορά τη λήψη (ή μη λήψη) των συχνοτήτων λέξεων και την αντίστοιχη προσέγγιση για τη δειγματοληψία του χώρου πιθανότητας:

- **Πολυμερές μοντέλο Bernoulli:** Σε αυτό το μοντέλο, χρησιμοποιούμε την παρουσία ή την απουσία λέξεων σε ένα έγγραφο κειμένου ως χαρακτηριστικά που αντιπροσωπεύουν ένα έγγραφο. Επομένως, οι συχνότητες των λέξεων δεν χρησιμοποιούνται για τη μοντελοποίηση ενός εγγράφου και τα χαρακτηριστικά των λέξεων στο κείμενο θεωρούνται δυαδικά, με τις δύο τιμές να δείχνουν την παρουσία ή την απουσία μιας λέξης στο κείμενο. Δεδομένου ότι τα χαρακτηριστικά που θα μοντελοποιηθούν είναι δυαδικά, το μοντέλο για τα έγγραφα σε κάθε κατηγορία είναι ένα πολυπαραγοντικό μοντέλο Bernoulli.

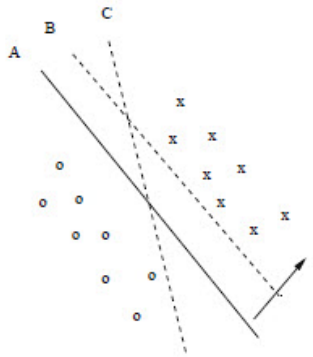
- **Πολυωνυμικό Μοντέλο:** Σε αυτό το μοντέλο, αποτυπώνουμε τις συχνότητες των όρων σε ένα έγγραφο, αντιπροσωπεύοντας ένα έγγραφο με την μέθοδο "bag of words". Τα έγγραφα σε κάθε κλάση μπορούν στη συνέχεια να διαμορφωθούν ως δείγματα που προέρχονται από πολυωνυμική κατανομή λέξεων. Ως αποτέλεσμα, η υποθετική πιθανότητα ενός εγγράφου που δίνεται σε μια κλάση είναι απλά ένα προϊόν της πιθανότητας κάθε παρατηρούμενης λέξης στην αντίστοιχη κλάση.

3.5.4 Support Vector Machines

Η βασική αρχή των Support Vector Machines είναι να προσδιορίσουν διαχωριστές στον χώρο αναζήτησης που μπορούν να διαχωρίσουν καλύτερα τις διαφορετικές κλάσεις των δεδομένων.

Για παράδειγμα, μπορούμε να παρατηρήσουμε το Σχήμα 3.3, στο οποίο έχουμε δύο κλάσεις που υποδηλώνονται με 'x' και 'o' αντίστοιχα. Έχουμε υποδείξει τρία διαφορετικά διαχωριστικά υπερπλαίσια (hyperplanes), τα οποία συμβολίζονται με A,B και C αντίστοιχα. Μπορούμε εύκολα να διαπιστώσουμε οτι το διαχωριστικό υπερπλαίσιο (hyperplane) A παρέχει τον αποδοτικότερο διαχωρισμό των δεδομένων, επειδή η κανονική απόσταση οποιουδήποτε από τα σημεία δεδομένων από αυτό είναι η μεγαλύτερη. Επομένως, το υπερπλαίσιο (hyperplane) A αντιπροσωπεύει το μέγιστο περιθώριο διαχωρισμού. Σημειώνουμε ότι ο κανονικός διάνυσμα σε αυτό το υπερπύρηνο (που αναπαριστάται από το βέλος στο σχήμα) είναι μια κατεύθυνση στο χώρο των χαρακτηριστικών, κατά μήκος του οποίου έχουμε τη μέγιστη διάκριση. [12]

Ένα πλεονέκτημα της μεθόδου SVM είναι ότι από τη στιγμή που προσπαθεί να καθορίσει τη βέλτιστη κατεύθυνση της διάκρισης στο χώρο χαρακτηριστικών εξετάζοντας τον κατάλληλο συνδυασμό χαρακτηριστικών, είναι αρκετά ανθεκτικό σε υψηλή διαστασιολόγηση (high dimensionality). Όπως έχει αναφερθεί στην έρευνα [17], τα δεδομένα κειμένου είναι ιδανικά για ταξινόμηση που παρέχουν τα SVM, λόγω της αραιής υψηλής διαστασιολόγησης που έχουν τα κείμενα απο την φύση τους. [12]



Σχήμα 3.4: Ένα σύνολο δεδομένων το οποίο διαχωρίζεται σε δύο κλάσεις
Πηγή: Mining Text Data 2012

3.5.5 Λογιστική Παλινδρόμηση

Η Λογιστική Παλινδρόμηση είναι μία τεχνική σχεδιασμένη για την πραγματοποίηση ανάλυσης δεδομένων που αφορούν την μελέτη και την πρόβλεψη τιμών κάποιας κατηγορικής εξαρτημένης μεταβλητής και χρησιμοποιεί ποσοτικές και ποιοτικές ανεξάρτητες μεταβλητές [28]. Ουσιαστικά η μέθοδος αυτή γενικεύει τα γραμμικά μοντέλα, έτσι ώστε η εξαρτημένη μεταβλητή να ακολουθεί την εκθετική οικογένεια κατανομών.

Για να κατανοήσουμε καλύτερα τη λογιστική παλινδρόμηση ας δούμε ένα παράδειγμα από την ανάλυση που υπάρχει στις διαλέξεις του Φωκιανού [28]: "Πραγματοποιήθηκε έρευνα με εργάτες μίας αμερικάνικης εταιρείας στη βιομηχανία βαμβακιού. Η εταιρεία θέλει να εξετάσει αν κάποιος εργάτης της πάσχει από κάποια συγκεκριμένη ασθένεια του πνεύμονα. Για αυτό το λόγο δημιουργήθηκαν πέντε κριτήρια, με την αντίστοιχη μεταβλητή/τιμή του το καθένα: φυλή (1.λευκός, 2. άλλο), φύλο (1. άνδρας, 2. γυναίκα), κάπνισμα (1.ναι, 2. όχι), διάρκεια εργασίας (1. λιγότερο από 10 χρόνια, 2. 10 με 20 χρόνια) και ποσοστό σκόνης στον εργασιακό χώρο (1. υψηλό, 2. μέτριο, 3.χαμηλό). Το πρόβλημα για αυτά τα δεδομένα είναι το να εξακριβωθεί κατά πόσο οι παραπάνω επεξηγηματικές μεταβλητές είναι σημαντικές για την εμφάνιση αυτής της ασθένειας. Αν δηλαδή ποιες αυτά τα κριτήρια μπορούν να χρησιμοποιηθούν για να προβλέψουν κατά πόσο ένας εργάτης θα πάσχει από αυτή ασθένεια του πνεύμονα και επειδή η ανεξάρτητη μεταβλητή είναι δυαδική, χρησιμοποιείται η λογιστική παλινδρόμηση για την ανάλυση.

Ο τύπος για να λυθεί το προαναφερθέν πρόβλημα μέσω της λογιστικής παλινδρόμησης δίνεται παρακάτω (η πιο διαδεδομένη έκφραση της εξίσωσης της Λογιστικής Παλινδρόμησης):

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}$$

Σχήμα 3.5: Τύπος λογιστικού μετασχηματισμού για την εύρεση πιθανότητας
Πηγή: http://www.lib.teiher.gr/webnotes/seyp/SPSS/Kef12.pdf

όπου p η πιθανότητα ένας εργάτης να πάσχει από την ασθένεια που εξετάζουμε στο παράδειγμά μας και k οι μεταβλητές/κριτήρια

Όπως βλέπουμε, αντί να χρησιμοποιηθεί ένα γραμμικό μοντέλο για να εξεταστεί η εξάρτηση της πιθανότητας εμφάνισης της ασθένειας του πνεύμονα από τις επεξηγηματικές μεταβλητές, χρησιμοποιείται ο λογιστικός μετασχηματισμός, ο οποίος ορίζεται ως εξής:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + \ldots + \beta_k X_k.$$

Σχήμα 3.6: Τελικός τύπος για την εύρεση πιθανότητας
Πηγή: http://www.lib.teiher.gr/webnotes/seyp/SPSS/Kef12.pdf

Να τονιστεί εδώ ότι υπάρχει και η διαχωριστική ανάλυση, παρόμοια έννοια της λογιστικής παλινδρόμηση. Και οι δύο μέθοδοι χρησιμοποιούνται για να κατατάξουν τις παρα-

τηρήσεις μας, σε γνωστές ομάδες αλλά και για να προβλέψουν σε ποια ομάδα θα τοποθετήσουμε τις καινούριες παρατηρήσεις. Μάλιστα είναι από τις πιο διαδεδομένες μεθόδους κατάταξης. Βέβαια έχουν πολλές διαφορές μεταξύ τους. Στη λογιστική παλινδρόμηση δε χρειάζονται οι περίπλοκες υποθέσεις όπως χρειάζονται να γίνουν στη διαχωριστική ανάλυση. Δηλαδή στη λογιστική παλινδρόμηση δε μας ενδιαφέρει αν οι ανεξάρτητες μεταβλητές ακολουθούν κανονική κατανομή, αν σχετίζονται γραμμικά ή και αν έχουν ίσες διασπορές για τον κάθε ένα πληθυσμό όπως συμβαίνει στη διαχωριστική ανάλυση. Επίσης η λογιστική παλινδρόμηση δεν κάνει καμία υπόθεση για τις ανεξάρτητες μεταβλητές. Για το λόγο αυτό η λογιστική παλινδρόμηση χρησιμοποιείται πιο συχνά [24].

3.5.6 Νευρωνικά Δίκτυα

3.5.6.1 Ανάλυση Νευρωνικών Δικτύων

Νευρωνικό δίκτυο ονομάζεται ένα κύκλωμα διασυνδεδεμένων νευρώνων. Η αρχική έννοια αφορούσε τα βιολογικά νευρωνικά δίκτυα, που υπάρχουν στη φύση. Έπειτα και στο τομέα της επιστήμης εμφανίστηκαν οι τεχνητοί νευρώνες, έχουν αναπτυχθεί μόνο κατά τα τελευταία σαράντα περίπου χρόνια.Ένα νέο είδος νευρωνικών δικτύων, τα τεχνητά. Η λειτουργία τους είναι εμπνευσμένη από τον τρόπο λειτουργίας των Βιολογικών Νευρικών δικτύων, τα οποία αποτελούν δομικά συστατικά των εγκεφάλων των ζώων και των ανθρώπων. Άρα ένα Τεχνητό Νευρωνικό Δίκτυο (Artificial Neural Network) είναι ένα υπολογιστικό σύστημα υλικού και λογισμικού [4] . Ένα ιδιαίτερο χαρακτηριστικό είναι ότι οι επιστήμονες στην περιοχή των νευρωνικών δικτύων προέρχονται σχεδόν από όλες τις περιοχές των φυσικών επιστημών, όπως την ιατρική, την επιστήμη μηχανικών, την φυσική, την χημεία, τα μαθηματικά, την επιστήμη υπολογιστών, ηλεκτρολογία, κλπ.

Σχήμα 3.7: Βιολογικό και Τεχνητό Νευρωνικό δίκτυο
Πηγή: https://elearning.teicm.gr/file.php/472/P202$_N eural_1 .pdf$

Οι νευρώνες τους είναι τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες. Οι νευρώνες εισόδου δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συναπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη.

Τα τελευταία χρόνια έχει υπάρξει μία έκρηξη ενδιαφέροντος για τα νευρωνικά δίκτυα καθώς εφαρμόζονται με μεγάλη επιτυχία σε ένα ασυνήθιστα μεγάλο φάσμα τομέων της επιστήμης και της τεχνολογίας, όπως τα χρηματοοικονομικά, η ιατρική, η επιστήμη μηχανικού, η γεωλογία, η φυσική, η ρομποτική, η επεξεργασία σήματος κτλ. Στην πραγματικότητα, τα νευρωνικά δίκτυα εισάγονται οπουδήποτε τίθεται θέμα πρόβλεψης, ταξινόμησης

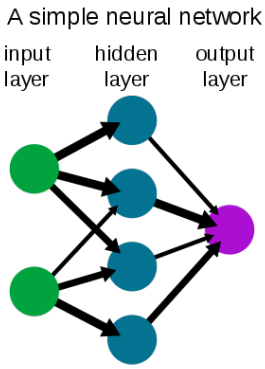
[4]https://en.wikipedia.org/wiki/Artificial$_n eural_n etwork$

ή ελέγχου. Η σαρωτική αυτή επιτυχία, μπορεί να αποδοθεί σε δύο βασικά στοιχεία: την ισχύ και την ευχρηστία.

Ένα τεχνητό νευρωνικό δίκτυο έχει είσοδο, επεξεργασία και έξοδο. Οι είσοδοι των τεχνητών νευρώνων τροφοδοτούνται μέσω των συναπτικών βαρών, είτε από τις εισόδους του δικτύου (στο στρώμα εισόδου), είτε από τις εξόδους άλλων νευρώνων (στα κρυφά στρώματα). Άρα η κάθε τιμή εισόδου επηρεάζεται έπειτα στη διαδικασία επεξεργασίας από το βάρος που της δίνεται και κουβαλάει μαζί της. Οι έξοδοι του δικτύου είναι οι έξοδοι των νευρώνων του στρώματος εξόδου.

Τα στρώματα/επίπεδα που υπάρχουν μέσα στην επεξεργασία του συστήματος, μεταξύ των στρώσεων εισόδου και εξόδου, ονομάζονται hidden layers και λαμβάνουν ένα σύνολο εισόδων και παράγουν μια έξοδο μέσω μιας συνάρτησης ενεργοποίησης. Εκεί η έξοδος ενός στρώματος είναι η είσοδος ενός άλλου στρώματος. Ουσιαστικά αυτά τα ενδιάμεσα επίπεδα πρέπει να μετατρέψουν επιτυχώς τις εισόδους που δέχονται και τροποποιούν σε κάτι που μπορεί να χρησιμοποιήσει το στρώμα εξόδου. Το κάθε ένα layer δέχεται το αποτέλεσμα του προηγούμενου επιπέδου, το επεξεργάζεται και το στέλνει στο επόμενο. Τέλος ο αριθμός των κρυμμένων νευρώνων πρέπει να είναι μεταξύ του μεγέθους του στρώματος εισόδου και του μεγέθους του στρώματος εξόδου και επίσης να ισούται με τα 2/3 του μεγέθους του στρώματος εισόδου, συν το μέγεθος της στρώσης εξόδου. Ο αριθμός των κρυμμένων νευρώνων θα πρέπει να είναι μικρότερος από το διπλάσιο του μεγέθους του στρώματος εισόδου [5] .

Για παράδειγμα, μπορεί να θέλουμε με τη βοήθεια ενός υπολογιστή να εξετάσουμε σε μία φωτογραφία αν υπάρχει ένα λεωφορείο. Να δώσουμε τα σωστά εργαλεία και να διδάξουμε επιτυχώς στον υπολογιστή να καταφέρνει τη παραπάνω επιθυμία μας. Ο ανιχνευτής μας μπορεί να αποτελείται από ανιχνευτή τροχών (ώστε να μας πει ότι πρόκειται για όχημα), από ανιχνευτή κιβωτίων (δεδομένου ότι το λεωφορείο έχει σχήμα μεγάλης θήκης) και ανιχνευτή μεγέθους (ώστε να μας πει ότι είναι πολύ μεγάλο για να είναι ένα κλασσικό επιβατικό αυτοκίνητο). Αυτά είναι τα τρία στοιχεία του κρυμμένου επιπέδου (hidden layer). Δεν είναι μέρος του input layer δηλαδή, είναι εργαλεία που σχεδιάσαμε για να προσδιορίσουμε τη δομή των λεωφορείων. Αν ενεργοποιηθούν και οι τρεις από αυτούς τους ανιχνευτές, τότε υπάρχει μια καλή πιθανότητα να έχετε ένα λεωφορείο μπροστά σας. Τα νευρικά δίχτυα είναι χρήσιμα επειδή υπάρχουν καλά εργαλεία για την κατασκευή πολλών ανιχνευτών και την τοποθέτηση τους μαζί.



Σχήμα 3.8: Νευρωνικό δίκτυο
Πηγή: Wikipedia

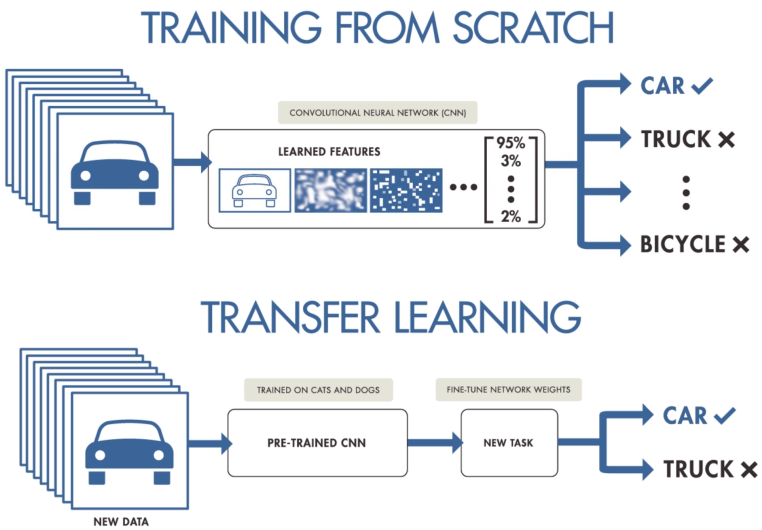
[5]https://stats.stackexchange.com/questions/181/how-to-choose-the-number-of-hidden-layers-and-nodes-in-a-feedforward-neural-netw

3.5.6.2 Μεταφορά Μάθησης (Transfer Learning)

Στο πεδίο της επεξεργασίας φυσικής γλώσσας (NLP), τα βαθιά νευρωνικά δίκτυα έχουν βελτιώσει την απόδοση των μοντέλων σε πολλά διαφορετικά προβλήματα/tasks. Η εκμάθηση μεταφοράς είναι ένα ερευνητικό πρόβλημα στη μηχανική μάθηση που επικεντρώνεται στην αποθήκευση της γνώσης που αποκτάται κατά την επίλυση ενός προβλήματος και στην εφαρμογή του σε ένα διαφορετικό αλλά σχετικό πρόβλημα. Για παράδειγμα, οι γνώσεις που αποκτήθηκαν κατά την εκμάθηση της αναγνώρισης αυτοκινήτων θα μπορούσαν να εφαρμοστούν όταν προσπαθούσαν να αναγνωρίσουν τα φορτηγά.

Το κύριο χαρακτηριστικό των νευρωνικών δικτύων είναι η εγγενής ικανότητα μάθησης. Ως μάθηση μπορεί να οριστεί η σταδιακή βελτίωση της ικανότητας του δικτύου να επιλύει κάποιο πρόβλημα (π.χ. η σταδιακή προσέγγιση μίας συνάρτησης). Η μάθηση επιτυγχάνεται μέσω της εκπαίδευσης, μίας επαναληπτικής διαδικασίας σταδιακής προσαρμογής των παραμέτρων του δικτύου (συνήθως των βαρών και της πόλωσής του) σε τιμές κατάλληλες ώστε να επιλύεται με επαρκή επιτυχία το προς εξέταση πρόβλημα. Για αυτό η εκμάθηση ενός καινούργιου μοντέλου για κάθε διαφορετικό πρόβλημα απαιτεί πληθώρα δεδομένων με ετικέτες. Δεν είναι πάντα όλα όμως τόσο ιδανικά και με απόλυτη επιτυχία. Για παράδειγμα μπορεί αν μην έχουμε επαρκή αριθμό διαθέσιμων ετικετών και αρχικών δεδομένων γενικά. Αντιμετωπίζεται συχνά το παραπάνω θέμα με το να εφαρμόσουμε τη λύση ενός προβλήματος με επαρκή δεδομένα/ετικέτες σε ένα διαφορετικό αλλά παρόμοιο πρόβλημα [8].

Άρα για να έχουμε ικανοποιητικά αποτελέσματα, τα μοντέλα αυτά τυπικά χρειάζεται να εκπαιδευτούν σε εκατομμύρια δεδομένα και παρόμοια προβλήματα, με ειδικές ετικέτες για κάθε υποπρόβλημα κλπ. Αφού ένα δίκτυο εκπαιδευτεί, οι παράμετροί του συνήθως σταματάνε στις κατάλληλες τιμές και από εκεί κι έπειτα είναι σε λειτουργική κατάσταση για άλλα προβλήματα. Το ζητούμενο είναι το λειτουργικό δίκτυο να χαρακτηρίζεται από μία ικανότητα γενίκευσης: αυτό σημαίνει πως δίνει ορθές εξόδους για πολλές και διαφορετικές εισόδους από αυτές με τις οποίες εκπαιδεύτηκε. Η μεταφορά μάθησης συνήθως οδηγεί σε γρηγορότερη και υψηλότερη απόδοση από αυτήν που θα είχε το μοντέλο, αν είχε εκπαιδευτεί μόνο σε ένα μικρό σύνολο δεδομένων [6].



Σχήμα 3.9: Παράδειγμα χρήσης Μεταφορά Μάθησης σε Μοντέλα
Πηγή: https://www.pinterest.com/pin/672232681856247783/

[6]https://en.wikipedia.org/wiki/Artificial$_n eural_n etwork$

3.5.6.3 Ρυθμοί Εκμάθησης (Learning Rates)

Σαν ρυθμό εκμάθησης (Learning Rate) ορίζουμε μια υπερ-παράμετρο που ελέγχει για μας πόσο μπορούμε να προσαρμόζουσε τα βάρη του δικτύου μας σε σχέση με loss gradient. Ένα υψηλό ποσοστό εκμάθησης (high learning rate) σημαίνει ότι το δίκτυο αλλάζει τη "θεωρία" του και πιο γρήγορα. Αυτό μπορεί να είναι καλό αλλά και κακό συχνά. Το Learning Rate δείχνει πόσο γρήγορα ένα δίκτυο εγκαταλείπει παλιές πεποιθήσεις για νέες. Θέλουμε να βρεθεί ένας ρυθμός εκμάθησης που να είναι αρκετά χαμηλός ώστε το δίκτυο να συγκλίνει σε κάτι χρήσιμο, αλλά αρκετά υψηλός ώστε να μην χρειάζεται να ξοδευτεί πολύς χρόνος για να το εκπαιδεύσουμε.

Για να το καταλάβουμε καλύτερα με ένα παράδειγμα, ας πούμε ότι ένα παιδάκι θέλει να καταλάβει πως είναι οι γάτες για να μπορεί αν τις ξεχωρίζει. Του δείχνουμε μία γάτα η οποία έχει άσπρο χρώμα. Αμέσως το παιδί συμπεραίνει ότι όλες οι γάτες είναι άσπρες και όταν ξανά κοιτάξει ζώα θα ελέγξει αν είναι άσπρες για να πει ότι είναι ίσως γάτες. Όμως αν του πουν οι γονείς του την επόμενη φορά ότι μία γάτα μπορεί να είναι και μαύρη (supervised learning περίπτωση), τότε θα καταλάβει το παιδί ότι το χρώμα δεν είναι το σημαντικότερο στοιχείο για να ορίσουμε ένα ζώο ως γάτα (μεγάλο/high learning rate). Αν δε του το πούνε τότε το παιδί θα συνεχίζει ότι οι γάτες είναι άσπρες (μικρό/low learning rate). Στο παραπάνω παράδειγμα το μεγάλο learning rate είναι καλό και μας βοηθά να έχουμε καλύτερα συμπεράσματα. Άλλα δε πρέπει να είναι και πάρα πολύ μεγάλο, γιατί μπορεί το παιδί να αρχίσει να πιστεύει ότι όλες οι γάτες είναι μαύρες παρόλο που είχε δει περισσότερες άσπρες γάτες από μαύρες [7].

Ο ρυθμός εκμάθησης έχει οριστεί επίσης σαν μια υπερ-παράμετρος(hyper-parameter). Ουσιαστικά μια παράμετρος της οποίας η τιμή έχει οριστεί πριν ξεκινήσει η διαδικασία εκμάθησης. Ελέγχει πόσο προσαρμόζουμε τα βάρη του δικτύου μας σε σχέση με την πιθανή κλίση της απώλειας. Αν και αυτό μπορεί να είναι μια καλή ιδέα (χρησιμοποιώντας ένα χαμηλό ρυθμό εκμάθησης) όσον αφορά τη διασφάλιση ότι δεν θα χαθεί κανένα τοπικό ελάχιστο, θα μπορούσε επίσης να σημαίνει ότι θα μας πάρει και πολύ χρόνο για να συγκλίνουμε [8]. Ο τύπος που μας παρουσιάζει τη σχέση των βαρών και του learning rate παρατίθεται παρακάτω και μέσω του γραφήματος οι διαφορές μεγάλου με μικρό learning rate (και του ιδανικού):

- newWeight = existingWeight — learningRate * gradient

3.5.6.4 Fine Tuning

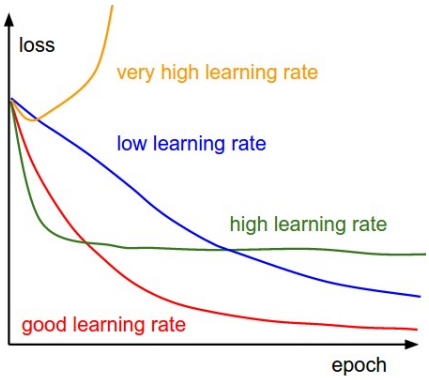
Το "Fine tuning" ομαδοποιεί ένα πολύ μεγάλο σύνολο τεχνικών που προσπαθούν να βρουν τους καλύτερους δυνατούς υπερπαραμετρικούς παράγοντες για τον αλγόριθμό και το πρόβλημα. Το Fine tuning είναι τεχνική που χρησιμοποιείται κυρίως στο πλαίσιο της supervised learning. Πιο πρακτικά, από τον ορισμό του Fine tuning, το τελευταίο στρώμα αφαιρείται/αποσπάται. [9]. Επίσης θεωρείται καi τεχνική που χρησιμοποιείται στο πλαίσιο της επίβλεψης και της ενίσχυσης της μάθησης. Είναι πολύ δημοφιλές και στο τομέα του computer vision και στο τομέα NLP γενικά [10]. Η "εκμάθηση μεταφοράς" μπορεί να περιλαμβάνει το "fine tuning" του προ-εκπαιδευμένου μοντέλου.

[7] https://www.quora.com/What-is-the-learning-rate-in-neural-networks

[8] https://towardsdatascience.com/understanding-learning-rates-and-how-it-improves-performance-in-deep-learning-d0d4059c1c10

[9] https://www.quora.com/What-is-the-difference-between-transfer-learning-and-fine-tuning

[10] http://wiki.fast.ai/index.php/Fine$_tuning$



Σχήμα 3.10: Παράδειγμα γραφήματος με διάφορα learning rates
Πηγή: https://towardsdatascience.com/understanding-learning-rates-and-how-it-improves-performance-in-deep-learning-d0d4059c1c10

Συχνά συγχέεται με το transfer learning. Η "εκμάθηση μεταφοράς" (transfer learning) είναι ένα μοντέλο που αναπτύσσεται για μια υπάρχουσα εργασία και επαναχρησιμοποιείται για ένα μοντέλο σε μία δεύτερη παρόμοια εργασία. Το fine tuning από την άλλη είναι η προσέγγιση που γίνεται για αυτή τη εκμάθηση μεταφοράς. Με άλλα λόγια μπορούμε να πούμε ότι το fine tuning είναι υποκλάδος του transfer learning. Στο Transfer learning γενικά η βασική ιδέα είναι ότι οι πληροφορίες που προέρχονται από εργασίες παρόμοιες με τις εργασίες στόχου (δηλαδή η ερώτηση που προσπαθεί επί του παρόντος να επιλύσει ο χρήστης) μπορεί να είναι χρήσιμες για την επιτάχυνση της μαθησιακής διαδικασίας γενικά. Φυσικά, το μεγάλο πρόβλημα είναι να καταλάβουμε πότε είναι χρήσιμο να μεταφέρουμε τις πληροφορίες, αποφεύγοντας ταυτόχρονα την γνωστή ως αρνητική μεταφορά. Δηλαδή τη μεταφορά πληροφοριών που μπορεί να οδηγήσουν σε επιβράδυνση της μαθησιακής διαδικασίας για το στόχο.

Κατά τη διάρκεια της εκμάθησης μεταφοράς πχ μπορούμε να απελευθερώσουμε το προ-καταρτισμένο μοντέλο και να το αφήσουμε να προσαρμοστεί περισσότερο στο τρέχον πρόβλημα. Ας πούμε ότι έχουμε ένα σύνολο δεδομένων, και χρησιμοποιούμε το 90% της στην εκπαίδευση/training. Στη συνέχεια εκπαιδεύουμε το ίδιο μοντέλο με το υπόλοιπο 10%. Συνήθως, αλλάζουμε τον ρυθμό εκμάθησης σε μικρότερο, έτσι ώστε να μην έχει σημαντική επίδραση στα ήδη προσαρμοσμένα βάρη. Το Fine tuning πιο συγκεκριμένα:

- Αντικαθιστά το τελευταίο επίπεδο (εξόδου) με ένα επίπεδο που αναγνωρίζει τον αριθμό των κλάσεων που χρειάζεται όμως ο χρήστης.
- Το νέο στρώμα εξόδου που είναι συνδεδεμένο με το μοντέλο έχει εκπαιδευτεί ώστε να λαμβάνει τα χαρακτηριστικά του προηγούμενού του επιπέδου από το μπροστινό μέρος του δικτύου και να τα αντιστοιχίζει στις επιθυμητές τάξεις εξόδου.

Κεφάλαιο 4

Τομείς Έρευνας

tΚαθώς η εργασία αναλύει την Κ.Κ με Προσαρμογή (Fine Tuning) Προ-εκπαιδευμένων Γλωσσικών Μοντέλων, οι τομείς που θα εφαρμόσουμε την τεχνική αυτή είναι δύο, η αναγνώριση συγγραφέα απο κείμενα Fanfiction, καθώς και η αναγνώριση είδους ιστοσελίδας απο το κείμενο κάθε σελίδας χωρίς τα HTML Tags. Σε αυτή την ενότητα θα αναλύσουμε τις προσεγγίσεις που υπάρχουν στους δύο αυτούς τομείς, τα αποτελέσματα που έχουν, καθώς και μια εισαγωγή στο πως θα τους προσεγγίσουμε εμείς.

4.1 Αναγνώριση Συγγραφέα

Η αναγνώριση συγγραφέα από κείμενα, αποτελεί ένα σημαντικό πεδίο της επιστήμης της *Επεξεργασίας Φυσικής Γλώσσας* (NLP). Ως αναγνώριση συγγραφέα, ορίζουμε την αυθεντικοποίηση και ταυτοποίηση ενός συγγράμματος ως προς ένα μοναδικό πρόσωπο.

Το να μπορούμε άμεσα και με επιτυχία να αποδίδουμε οτιδήποτε είναι σε γραπτή μορφή σε ένα συγκεκριμένο φυσικό πρόσωπο ή ομάδα προσώπων ήταν, είναι και θα είναι σίγουρα μία μεγάλη πρόκληση. Και όχι μόνο για ηθικούς λόγους. Το ύφος του συγγραφέα μπορεί να τον βοηθήσει να δημιουργήσει ένα ολόκληρο νέο και αποκλειστικά δικό του πεδίο στη τέχνη και επιπροσθέτως να βγάλει κάποιο σημαντικό κέρδος χάρη στη δουλειά που έχει απαιτηθεί για τη δημιουργία των έργων του.

Οποιοσδήποτε λοιπόν συγγραφέας χρειάζεται να ασφαλίσει τα δημιουργήματα του ως προς κακόβουλους ανταγωνιστές του ή και από απλούς θαυμαστές του, οι οποίοι εμπνεύστηκαν από τα συγγράμματα του και πιθανόν να θέλουν να αναπτύξουν οποιοδήποτε στοιχείο έχει αναπτύξει ο συγγραφέας, όπως συχνά είναι το συγγραφικό ύφος. Και σίγουρα αυτός ο στόχος μπορεί να επιτευχθεί με στοιχεία από πολλούς από τους προαναφερθέντες κλάδους (Ανάκτηση Πληροφορίας, Τεχνητή Νοημοσύνη, Εξόρυξη από Κείμενα κλπ).

4.1.1 Ιστορική Αναδρομή

Η ανάγκη για την διερεύνηση συγγραφικής πατρότητας, ξεκινά στα μέσα του 19ου αιώνα με την έρευνα Thomas Corwin Mendenhall (1887), καθώς και στιγματίστηκε απο την μελέτη των Mosteller και Wallace το 1964, πάνω στην αναγνώριση του συγγραφέα των ομοσπονδιακών κειμένων των ΗΠΑ (Federalist Papers), η οποία βασίστηκε στην χρήση του στατιστικού μοντέλου Bayes πάνω στην χρήση συχνών λέξεων όπως για παράδειγμα 'και', 'το', 'σε' κλπ[10]. Αυτή η έρευνα ουσιαστικά πυροδότησε την αρχή για νέες μεθόδους αναγνώρισης συγγραφέων σε αντίθεση με τις μέχρι τότε προσεγγίσεις οι οποίες βασίζονταν σε ανθρώπινη κρίση. Απο τότε και στο εξής έγιναν πολλές προσπάθειες για

εξαγωγή χαρακτηριστικών, μέσω της ανάλυσης του στυλ, με μεθόδους όπως η μέτρηση μήκους λέξεων, συχνότητα εμφάνισης των λέξεων, συχνότητα χαρακτήρων και πλούτος λεξιλογίου[10].

Τα τελευταία χρόνια, η ραγδαία εξέλιξη της τεχνολογίας και η έλευση της τεχνητής νοημοσύνης και της μηχανικής μάθησης στο προσκήνιο, έχει επηρεάσει σημαντικά αυτό το πεδίο, φέρνοντας νέες μεθόδους και τεχνικές για την βελτιστοποίηση των λύσεων. Ο παράγοντας του διαδικτύου παίζει και αυτός έναν σημαντικό ρόλο και καθιστά αναγκαία την υπεράσπιση της συγγραφικής πατρότητας. Το χάος που υπάρχει εκεί και η δυσκολία συνεχούς ελέγχου οποιοδήποτε άρθρου που ανεβαίνει καθιστά ακόμα πιο δύσκολη τη προσπάθεια αυθεντικοποίησης ως προς ένα φυσικό πρόσωπο ή ομάδα.

Όπως φαίνεται και στη παρακάτω εικόνα με διάφορες γενικές πληροφορίες, κάθε ένα λεπτό ανεβαίνει ένας τεράστιος αριθμός δεδομένων στο ίντερνετ από πάρα πολλούς χρήστες, για τους οποίους συχνά δεν καθίσταται δυνατό να ελεγχθεί άμεσα και εγκαίρως το περιεχόμενό τους και η ασφάλεια του. Το ίδιο ισχύει και για τα άρθρα που διατίθενται στο κοινό διαδικτυακά και τα δικαιώματα των δημιουργών τους.

Data per Minute		
Platform	Number	Entity
Wikipedia	0.5	contents
Blogs	1800	blogs
Articles	850	texts
Twitter	347.222	tweets
Emails	187.000.000	emails
Reddit	18.327	votes

Πίνακας 4.1: Ογκος δεδομένων ανά λεπτό διαδικτυακά
Πηγή: https://www.iflscience.com/technology/amount-data-internet-generates-every-minute-crazy/

4.1.2 Είδη αναγνώρισης συγγραφέα

Ο αυτόματος εντοπισμός του συγγραφέα ενός κειμένου μπορούμε να πούμε ότι χωρίζεται σε 4 διαφορετικά είδη. Πρώτον υπάρχουν τα κλειστά προβλήματα, όταν πρέπει να βρούμε τον πραγματικό συγγραφέα μέσα από μία κλειστή λίστα με γνωστούς συγγραφείς (γνωστό και το ύφος τους). Ουσιαστικά ρωτάμε "ποιος από τους Α,Β,Γ έγραψε αυτό το κείμενο;".

Δεύτερον έχουμε τα ανοιχτά προβλήματα, όπου έχουμε στη κατοχή μας κάποιους για ένα κείμενο, οι οποίοι μπορεί και να μην καν υποψήφιοι για το δοσμένο κείμενο, άρα εδώ αναρωτιόμαστε "είναι πχ ο Α υποψήφιος για αυτό το ΄σύγγραμα;". Τρίτον έχουμε τα προβλήματα στα οποία μας ενδιαφέρει να εντοπίσουμε συγκεκριμένα χαρακτηριστικά για έναν συγγραφέα. Δεν επιδιώκουμε εδώ να ταυτοποιήσουμε ένα κείμενο με ένα φυσικό πρόσωπο αλλά να συλλέξουμε διάφορου είδους χαρακτηριστικά για αυτό το άτομο.

Τέλος έχουμε προβλήματα υφομετρικής ομοιογένειας, στα οποία πρέπει να γίνει εντοπισμός της κακόβουλης χρήσης ή λογοκλοπής όπως πχ η αλλοίωση του περιεχομένου μίας ιστοσελιδας. Εδώ πρέπει να μελετηθεί η κανονικότητα του υφομετρικού προφίλ και η χρήση ποσοτικών μεθόδων για την αξιολόγησή του. Στη δική μας περίπτωση θα ασχοληθούμε με το πρώτο είδος συγγραφικής πατρότητας.

4.1.3 Υφομετρικά Χαρακτηριστικά

Σημαντικό ρόλο στην αναγνώριση του συγγραφέα ενός κειμένου κατέχει ο τρόπος ο οποίος αντιλαμβάνεται και χρησιμοποιεί την γλώσσα ένας συγγραφέας. Μέσω του τρόπου γραφής δημιουργούνται υφολογικά χαρακτηριστικά που μπορούν να ποσοτικοποιηθούν, τα οποία ονομάζονται υφομετρικοί δείκτες.

4.1.3.1 Λεκτικά Χαρακτηριστικά

Η εξαγωγή των λεκτικών χαρακτηριστικών είναι η πιο απλή μέθοδος για να αναλύσουμε ένα κείμενο και να εξάγουμε πληροφορία απο αυτό. Μπορεί να θεωρηθεί ως μια σειρά συμβόλων, χωρισμένων σε προτάσεις.

Όπως αναφέραμε παραπάνω, οι πρώτες προσεγγίσεις στην αναγνώριση πατρότητας ήταν οι μετρήσεις των αριθμών προτάσεων και λέξεων, πράγμα το οποίο μπορεί να εφαρμοστεί εύκολα σε κάθε γλώσσα και σε κάθε κείμενο χωρίς περιορισμούς. Υπάρχει ένα πλήθος εργαλείων τα οποία έχουν αναπτυχθεί με στόχο την ορθή εξαγωγή λεκτικών χαρακτηριστικών.[10]

Ένα απο τα σημαντικά και απαραίτητα εργαλεία θεωρείται ο tokenizer. Με την χρήση του tokenizer μπορούμε να διαχωρίσουμε το κείμενο σε tokens οργανώνοντας με αυτό τον τρόπο σε πρώτο στάδιο την πληροφορία του κειμένου.

Ένα ακόμη σημαντικό εργαλείο το οποίο βοηθά στην επεξεργασία της πληροφορίας είναι ο stemmer, ο οποίος ευθύνεται για την αφαίρεση των καταλήξεων με την μέθοδο η οποία ονομάζεται stemming, με αποτέλεσμα να κρατάμε μόνο το περιεχόμενο της λέξης και όχι τις διαφορετικές μορφές της. Βέβαια δεν μπορεί να θεωρηθεί πως ο stemmer λύνει όλα τα προβλήματα μας, καθώς πολλές φορές η λανθασμένη αφαίρεση των καταλήξεων μπορεί να οδηγήσει σε σφάλματα κατανόησης ως προς το μοντέλο μάθησης.

Αντίστοιχα με τον stemmer υπάρχει το εργαλείο lemmatizer το οποίο ευθύνεται για την λημματοποίηση (Lemmatising) των λέξεων, δηλαδή την μετατροπή τους στην απλή αρχική τους μορφή, καθώς και εργαλεία τα οποία μετατρέπουν τα κεφαλαία γράμματα σε μικρά, με στόχο την απαλοιφή σφαλμάτων.

Stemming	
Word	Stem
studies	stud
adjustable	adjust
continuously	continou

Πίνακας 4.2: Παράδειγμα αφαίρεσης καταλήξεων
Πηγή:A Survey of Modern Authorship Attribution Methods

Lemmatisation	
Word	Lemma
studies	stud
adjustable	adjust
continuously	continou

Πίνακας 4.3: Παράδειγμα Λημματοποίησης
Πηγή:A Survey of Modern Authorship Attribution Methods

Ένα ακόμη λεκτικό χαρακτηριστικό θεωρείται ο πλούτος του λεξιλογίου που χρησιμοποιεί ο συγγραφέας. Είναι ένας τρόπος ποσοτικοποίησης της ποικιλίας του λεξιλογίου. Ένας τρόπος αναπαράστασης της ποσότητας του πλούτου είναι ο ονομαζόμενος type-token V/N, όπου V είναι το μέγεθος του λεξιλογίου που χρησιμοποιεί ο συγγραφέας (μοναδικές λέξεις) και όπου N το συνολικό μέγεθος του κειμένου. Το αρνητικό με την συγκεκριμένη μέθοδο είναι ότι το μέγεθος του λεξιλογίου που χρησιμοποιείται εξαρτάται άμεσα από το μέγεθος του κειμένου. Τέλος, μια από τις πιο διαδεδομένες προσεγγίσεις είναι η αναπαράσταση του κειμένου ως διανύσματα καθώς και μια πληθώρα ακόμη προσεγγίσεων[10].

4.1.3.2 Χαρακτηριστικά βασισμένα στους χαρακτήρες

Σε αντίθεση με τα λεκτικά χαρακτηριστικά, τα χαρακτηριστικά βασισμένα στους χαρακτήρες μιας λέξης αντιμετωπίζουν ένα κείμενο ως μια ακολουθία χαρακτήρων. Σύμφωνα με αυτή την μέθοδο, μπορούν να οριστούν διάφορα είδη μέτρων στο επίπεδο των χαρακτήρων, όπως είναι ο αριθμός των χαρακτήρων ενός αλφαβήτου, ο αριθμός πεζών και κεφαλαίων, καθώς και τα σημεία στίξης.

4.1.3.3 Συντακτικά Χαρακτηριστικά

Άλλη μια μέθοδος ανάλυσης κειμένων, η οποία θεωρείται λίγο πιο περίπλοκη είναι η ανάλυση της σύνταξης. Κάθε συγγραφέας, όπως έχουμε αναφέρει και σε προηγούμενη ενότητα, αντιλαμβάνεται και χρησιμοποιεί την γλώσσα με τον δικό του τρόπο, πράγμα που τον διαφοροποιεί -έστω και σε μικρό βαθμό- από τους υπόλοιπους συγγραφείς. Με αυτό τον τρόπο τείνει να δημιουργεί σχέδια (patterns) ασυνείδητα.

4.1.4 Fanfiction

Τις τελευταίες δεκαετίες στη συγγραφική τέχνη έχει δημιουργηθεί ένα ενδιαφέρον φαινόμενο και είδος γραφής και αυτό είναι η δημιουργία διαφόρων παραλλαγών αφηγημάτων απο θαυμαστές του αυθεντικού βιβλίου, τα οποία ορίζονται επίσημα ως Fanfiction. Για παράδειγμα, πόλλοι θαυμαστές των βιβλίων Harry Potter της J. K Rowling έχουν παραχαράξει την ιστορία των βιβλίων και έχουν δημιουργήσει κάτι δικό τους δανείζοντας στοιχεία από το αυθεντικό έργο. Πως θα είχε εξελιχθεί για παράδειγμα η ιστορία του συγκεκριμένου βιβλίου αν ο πρωταγωνιστής είχε αντιδράσει διαφορετικά;

Αυτός είναι ο κύριος τρόπος που γίνεται συνήθως, ο δημιουργός του Fanfiction αλλάζει διάφορες καταλήξεις γεγονότων που υπάρχουν μέσα στα αυθεντικά συγγράμματα ακολουθώντας την δική τους τροπή και αλλάζοντας στοιχεία της ιστορίας. Εκτός από αυτόν τον τρόπο, μπορεί να γράψει κάτι τελείως διαφορετικό εξ αρχής, μία άλλη ιστορία, η οποία όμως θα βρίσκεται στο ίδιο σύμπαν/περιβάλλον με αυτή του βιβλίου. Το περιβάλλον/σύμπαν που έχει δημιουργηθεί η διαφορετική ιστορία του συνήθως είναι κοινό με αυτό του αυθεντικού συγγράμματος. Επίσης μπορεί η αντιγραφή του θαυμαστή να συσχετίζεται εν τέλει απλά με το συγγραφικό ύφος και τίποτα άλλο. Αυτοί οι τρεις τρόποι δημιουργίας κειμένων εμπλέκονται συνήθως στο Fanfiction. Το κάθε τέτοιο έργο ονομάζεται fanfic.

Πρέπει να τονιστεί εδώ ότι πρόκειται για μη εμπορευματοποιημένη λογοτεχνική παραγωγή συνήθως. Ερασιτέχνες απλοί θαυμαστές δημιουργούν τέτοιου είδους κείμενα σαν χόμπι και χωρίς κάποιο απώτερο σκοπό ή κέρδος. Σπάνια εξουσιοδοτείται ή δημοσιεύεται από τον δημιουργό και εξίσου σπανίως δημοσιεύεται επαγγελματικά. Όσο για τις αντιδράσεις των δημιουργών, αυτές ποικίλουν. Από αδιάφορες μέχρι ενθαρρυντικές ή απορριπτικές. Η τελευταία κατάληξη συνήθως υπάρχει όταν δημιουργείται η ευκαιρία για εκμετάλλευση, χρηματική κυρίως.

Πάντα υπάρχουν άτομα που επιδιώκουν κέρδος από κάτι που ξεκινάει σαν ενασχόληση/χόμπι. Για αυτό και στις τρεις περιπτώσεις πρέπει να υπάρχει κάποιος έλεγχος, διότι μπορεί κάποιος να έχει προσωπικό κέρδος από μία τέτοια πράξη αντιγραφής μέρους της ιστορίας του συγγράμματος ή απλά του ύφους του αυθεντικού κειμένου. Στο παρελθόν υπήρξαν συχνές καταγγελίες από τους δημιουργούς οι οποίοι έχουν απαντήσει περιστασιακά με νομικές ενέργειες. Ένα ζήτημα προβληματικής κατάστασης δημιουργείται όταν πολλοί αναγνώστες μπερδεύουν τα αυθεντικά κείμενα με διάφορες αντιγραφές, δημιουργώντας σύγχυση για την πατρότητα και αυθεντικότητα του συγγράμματος που διαβάζουν. Και εκεί πάλι έρχεται το θέμα των δικαιωμάτων.

Έτσι καταλήγουμε στο συμπέρασμα, ότι οι δύο προαναφερθέντες λόγοι κάνουν κρίσιμο το ζήτημα της αναγνώρισης αυθεντικότητας ενός βιβλίου. Η εξέλιξη της τεχνολογίας όπως αναφέρθηκε βοηθάει στην εξάπλωση του παραπάνω προβλήματος, αλλά στον αντίποδα μπορεί να βοηθήσει και στη καταπολέμηση του. Στο παρακάτω πίνακα παρουσιάζονται οι δέκα πιο δημοφιλείς ιστορίες Fanfiction, αντίστοιχα από δέκα γνωστά συγγράμματά για βιβλία ή σειρές[1].

[1] https://www.dailydot.com/parsec/archive-of-our-own-fanfiction-pairing-census/

#	Ship	Fandom
1	Sherlock Holmes/John Watson	Sherlock (TV)
2	Castiel/Dean Winchester	Supernatural
3	Derek Hale/Stiles Stilinski	Teen Wolf (TV)
4	Harry Styles/Louis Tomlinson	One Direction (Band)
5	Dean Winchester/Sam Winchester	Supernatural
6	Merlin/Arthur Pendragon	Merlin (TV)
7	Draco Malfoy/Harry Potter	Harry Potter - J. K. Rowling
8	Steve Rogers/Tony Stark	The Avengers (Marvel Movies)
9	Sherlock Holmes & John Watson	Sherlock (TV)
10	Blaine Anderson/Kurt Hummel	Glee

Σχήμα 4.1: Οι πιο δημοφιλή ιστορίες στο διαδίκτυο για Fanfiction
Πηγή:https://www.dailydot.com/parsec/archive-of-our-own-fanfiction-pairing-census/

4.1.5 Μέθοδοι αντιμετώπισης της πληροφορίας

Σε όλη την πορεία ανάπτυξης της αναγνώρισης συγγραφέα έχουν αναπτυχθεί, όπως έχουμε παρατηρήσει, πολλές προσεγγίσεις λύσεων. Όλες οι προσεγγίσεις όμως έχουν κάποια κοινά χαρακτηριστικά μεταξύ τους τα οποία δεν μεταβάλλονται και θεωρούνται βασικά για κάθε πρόβλημα αναγνώρισης συγγραφέα.

Οι απαιτήσεις αυτές είναι οι εξής: ένα σύνολο από συγγραφείς με αντιστοιχία ένα σύνολο κειμένων για τον κάθε έναν, καθώς και ένα σύνολο από άγνωστα κείμενα στα οποία σε κάθε ένα από αυτά θα αντιστοιχηθεί ένας υποψήφιος συγγραφέας. Επίσης, υπάρχουν δύο τρόποι προσέγγισης ενός συνόλου δεδομένων απο κείμενα συγγραφέων οι οποίοι χρησιμοποιούνται και θα αναλύσουμε, η προσέγγιση με βάση το προφίλ του συγγραφέα (Profile-based approaches) καθώς και η στιγμιαία πρόσέγγιση (Instance-based approaches).

Κατά καιρούς πολλές από τις μεθόδους της K.K που αναφέραμε στο Κεφάλαιο 3 έχουν εφαρμοστεί στις δύο αυτές προσεγγίσεις δίνοντας σημαντικά αποτελέσματα, τα οποία θα αναλύσουμε σε αυτή την ενότητα, με εκτενέστερη ανάλυση στην Instance-Based approach.

4.1.5.1 Προσέγγιση με βάση το προφίλ του συγγραφέα (Pofile-based approach)

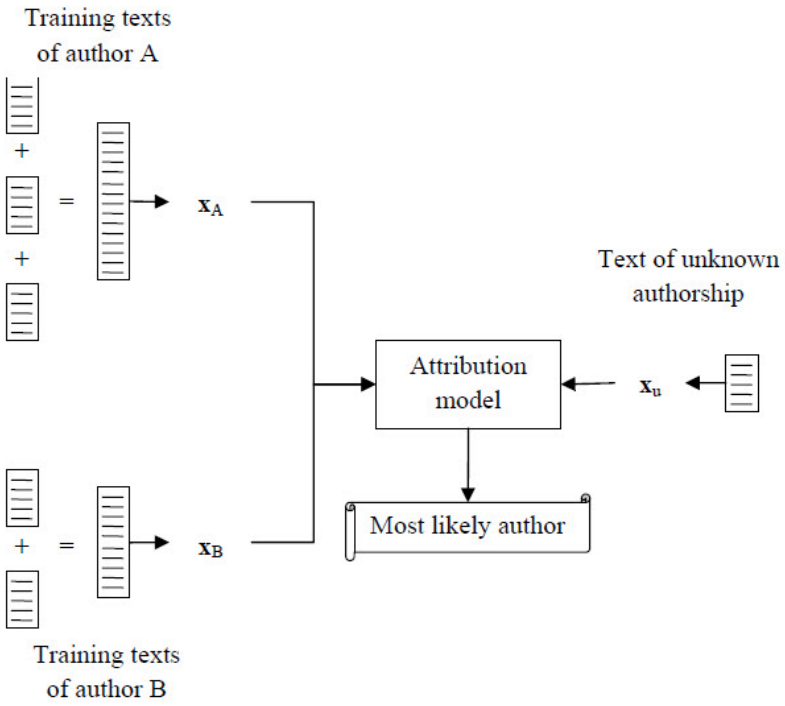
Η ειδοποιός διαφορά μεταξύ της Profile-based και της Instance-based προσέγγισης είναι η αντιμετώπιση του συνόλου των δεδομένων (dataset) που αφορούν τα κείμενα των συγγραφέων. Πιο συγκεκριμένα στην προσέγγιση με βάση το προφίλ, τα δεδομένα κειμένου τα οποία είναι ξεχωριστά αρχεία γίνονται σύμπτυξη σε ένα μεγάλο κείμενο.

Ο σκοπός πίσω από την δημιουργία αυτής της μεθόδου είναι η απόκτηση ενός συνολικού ύφους του κάθε συγγραφέα, μέσω της ανάλυσης του μεγάλου αρχείου που έχει δημιουργηθεί από το σύνολο των κειμένων. Στη συνέχεια, για να προβλεφθεί ο συγγραφέας ενός κειμένου χωρίς ετικέτα, το κείμενο συγκρίνεται με το κάθε ένα μεγάλο κείμενο κάθε διαφορετικού συγγραφέα και μέσω της απόστασης τους υπολογίζεται σε ποιόν συγγραφέα ταιριάζει περισσότερο.

Βέβαια, αν κανείς αναλύσει και συγκρίνει τις στυλομετρικές μετρήσεις του συνολικού κειμένου κάθε συγγραφέα και του κάθε κειμένου ξεχωριστά θα παρατηρήσει σημαντικές διαφορές μεταξύ τους. 'Ενα τυπικό παράδειγμα εφαρμογής της προσέγγισης αυτής μπορεί να παρατηρηθεί στο παρακάτω σχήμα.

Η διαδικασία της εκπαίδευσης ενός μοντέλου είναι σχετικά απλή. Αρχικά εξάγει αποτελέσματα χαρακτηριστικών απο ένα μεγάλο κείμενο για κάθε συγγραφέα και στη συνέχεια χρησιμοποιείται η μέθοδος της απόστασης, όπως αναφέραμε, μεταξύ των διανυσμά-

των του άγνωστου κειμένου και των υποψήφιων συγγραφέων.[10]



Σχήμα 4.2: Κλασσική αρχιτεκτονική προσέγγισης με βάση το προφίλ (Profile-based approach)

Πηγή: A Survey of Modern Authorship Attribution Methods, Stamatatos

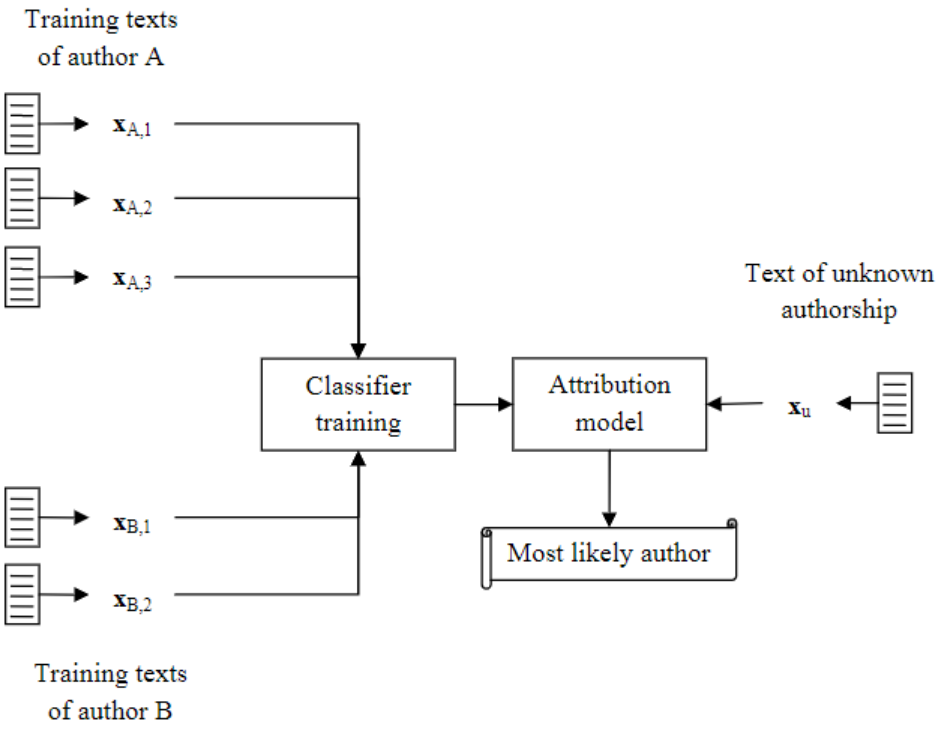
4.1.5.2 Στιγμιαία προσέγγιση (Instance-based approach)

Καθώς η εργασία μας πραγματεύεται τις μοντέρνες μεθόδους αναγνώρισης πατρότητας ενός κειμένου, η πιο διαδεδομένη μέθοδος αυτή την χρονική περίοδο είναι η ξεχωριστή αντιμετώπιση των κειμένων κάθε συγγραφέα (Instance-based) σε αντίθεση με την προηγούμενη προσέγγιση.

Συγκεκριμένα, η έρευνα και η τεχνική η οποία εφαρμόσαμε στην παρούσα εργασία έχει χρησιμοποιήσει αυτού του είδους την προσέγγιση εκπαιδεύοντας έναν ταξινομητή (classifier) με κάθε ένα κείμενο απο την συλλογή κάθε συγγραφέα, όπως θα αναλύσουμε στο επόμενο Κεφάλαιο. Γενικότερα, ένας ταξινομητής για να είναι αποδοτικός θα πρέπει να εκπαιδευτεί σε μια πληθώρα διαφορετικών κειμένων.

Στην περίπτωση όπου υπάρχει μόνο ένα μεγάλο κείμενο για κάθε συγγραφέα, όπως για παράδειγμα ένα ολόκληρο βιβλίο, η τεχνική που τείνει να εφαρμόζεται είναι η διάσπαση σε διαφορετικά κομμάτια. Επιπλέον, τα κείμενα τα οποία θα εκπαιδευτεί ένας ταξινομητής θα πρέπει να είναι ίσου μεγέθους[13] ως προς τον αριθμό λέξεων και στην περίπτωση όπου δεν είναι, θα πρέπει να εφαρμοστεί η τεχνική της κανονικοποίησης (normalization).

Στο παρακάτω σχήμα, μπορούμε να παρατηρήσουμε μια κλασσική αρχιτεκτονική της instance-based προσέγγισης όπου ένας ταξινομητής εκπαιδεύεται ξεχωριστά με κάθε κείμενο. Παρακάτω θα αναλύσουμε μεθόδους με τις οποίες εφαρμόζεται η συγκεκριμένη προσέγγιση μέσω των τεχνικών Κατηγοριοποίησης και συγκεκριμένα Κειμένου πάνω στο dataset στο οποίο έγινε η μελέτη της συγκεκριμένης εργασίας.



Σχήμα 4.3: Κλασσική αρχιτεκτονική στιγμιαίας προσέγγισης (Instance-based approach)
Πηγή: A Survey of Modern Authorship Attribution Methods, Stamatatos

4.1.5.3 Υβριδικές προσεγγίσεις

Ως Υβριδικές παρουσιάζονται οι προσεγγίσεις οι οποίες περιέχουν τεχνικές τόσο profile-based αλλά και instance-based. Η μέθοδος αυτή παρουσιάστηκε από τον Hans van Halteren το 2007 [14]. Συγκεκριμένα, τα κείμενα του συνόλου δεδομένων αναπαριστώνται το καθένα ξεχωριστά, όπως συμβαίνει στην Instance-based προσέγγιση, ενώ στη συνέχεια τα διανύσματα τα οποία παράγονται από το κάθε κείμενο ξεχωριστά, παράγουν ένα νέο διάνυσμα μέσω του υπολογισμού του Μέσου Όρου τους, το οποίο χαρακτηρίζει το στύλ-ύφος του συγγραφέα όπως είδαμε στην profile-based προσέγγιση. Εκτός απο την προσέγγιση του van Halteren, μια ακόμη παρόμοια προσέγγιση υπήρξε και απο τον Grieve το 2007 [15].

4.1.5.4 Σύγκριση

Στον παρακάτω πίνακα θα παρουσιάσουμε μια σύγκριση των δύο μεθόδων που αναπτύχθηκαν.

Προσεγγίσεις		
	Profile-Based	Instance-Based
Αναπαράσταση Κειμένου	Μια συσσωρευμένη αναπαράσταση απο όλα τα κείμενα συνολικά	Κάθε κείμενο αντιμετωπίζεται ξεχωριστά
Υφολογικά Χαρακτηριστικά	studies	stud
Ταξινόμηση	studies	stud
Χρονικό κόστος εκπαίδευσης	studies	stud
Χρονικό κόστος εκτέλεσης	studies	stud
Ανισσοροπία κλάσης	studies	stud

Πίνακας 4.4: Συγκρίσεις προσεγγίσεων
Πηγή: A Survey of Modern Attribution Methods, Stamatatos

4.1.6 Fanfiction Dataset - PAN18

Όπως αναλύσαμε σε προηγούμενη ενότητα, η έννοια του fanfiction έχει γίνει ευρέως γνωστή τα τελευταία χρόνια δημιουργώντας μια σειρά από προβλήματα που τίθενται προς λύση. Σε αυτή την ενότητα θα αναλυθεί το σύνολο δεδομένων (dataset) το οποίο χρησιμοποιήθηκε σε αυτή την εργασία, οι υποβολές μεθόδων-λύσεων που έγιναν από υποψήφιους καθώς και παρουσίαση των αποτελεσμάτων τους ώστε να παρουσιαστεί συγκριτικά η εφαρμογή της δικής μας μεθόδου πάνω στο ίδιο θέμα. Το dataset το οποίο έχει χρησιμοποιηθεί αφορά τον τον τομέα του fanfiction.[11]

4.1.6.1 Ανάλυση Dataset

Αρχικά, στον πρώτο πίνακα παρουσιάζονται αναλυτικά οι παράμετροι οι οποίοι απαρτίζουν το dataset που χρησιμοποιήθηκε. Στην φάση της ανάπτυξης παρατηρούμε οτι δόθηκαν 2 προβλήματα ανα γλώσσα, με σύνολο 5 γλώσσες (Αγγλικά, Γαλλικά, Ιταλικά, Πολωνικά, Ισπανικά). Πρόσθετα, δόθηκαν 5 και 20 σύνολα συγγραφέων (subsets) καθώς και απο 7 κείμενα σε κάθε έναν απο αυτούς για εκπαίδευση και μια διαφορετική ποικιλία για έλεγχο σε κάθε γλώσσα. Τέλος, βλέπουμε οτι τα κείμενα είναι κανονικοποιημένα (normalized) ως προς το μέγεθος τους, κάτι πολύ βασικό για την εκπαίδευση.

Όσον αφορά την φάση της αξιολόγησης χρησιμοποιήθηκαν 4 σετ προβλημάτων σε κάθε διαφορετική γλώσσα καθώς και ένα διαφορετικό πλήθος υποσυνόλων δεδομένων συγγραφέων. Τα κείμενα που χρησιμοποιήθηκαν για εκπαίδευση του μοντέλου παρέμειναν ίδια ως προς τον αριθμό, καθώς και τα δεδομένα για τεστ ποικίλουν με διαφορετικό τρόπο όπως παρατηρείται στον πίνακα. Τέλος, και σε αυτή την φάση το μέγεθος των κειμένων είναι κανονικοποιημένο σε κάθε γλώσσα ξεχωριστά.

	Language	Problems	Authors	Texts per author		Text length
			(subsets size)	training	test	(avg. words)
Development	English	2	5,20	7	1-22	795
	French	2	5,20	7	1-10	796
	Italian	2	5,20	7	1-17	795
	Polish	2	5,20	7	1-21	800
	Spanish	2	5,20	7	1-21	832
Evaluation	English	4	5,10,15,20	7	1-17	820
	French	4	5,10,15,20	7	1-20	782
	Italian	4	5,10,15,20	7	1-29	802
	Polish	4	5,10,15,20	7	1-42	802
	Spanish	4	5,10,15,20	7	1-24	829

Σχήμα 4.4: Άποψη του corpus το οποίο χρησιμοποιήθηκε
Πηγή: Overview of the Author Identification Task at PAN-2018

4.1.6.2 Προσεγγίσεις πάνω στο Dataset στο PAN18

Δοθέντος ενός βασικού κορμού ανάπτυξης ενός κώδικα, υπήρξαν 9 υποβολές απο ομάδες ερευνητών οι οποίες ανέπτυξαν μεθόδους για την αποδοτικότερη λύση. Στον παρακάτω πίνακα μπορούμε να παρατηρήσουμε ότι στις περισσότερες μεθόδους κυριαρχεί η μέθοδος n-grams για την αναπαράσταση των κειμένων.

Στις περισσότερες προσεγγίσεις επίσης, παρατηρείται η χρήση του SVM (Support Vector Machine) για την ταξινόμηση των κειμένων (Classifier) ενώ επίσης φαίνεται ξεκάθαρα η απουσία της χρήσης της μεθόδου Νευρωνικών Δικτύων, η οποία χρησιμοποιήθηκε μόνο από δύο ερευνητικές ομάδες, την ομάδα του Gagala και την ομάδα του Schaetti η οποία χρησιμοποίησε μια πιο συγκεκριμένη προσέγγιση η οποία ονομάζεται Echo-State Network.

Τέλος, για την κανονικοποίηση των κειμένων χρησιμοποιήθηκε κυρίως η τεχνική του TF-IDF, σε αντίθεση με την τεχνική των word embeddings η οποία χρησιμοποιήθηκε μόνο μια φορά. Κάτι που επίσης πρέπει να τονιστεί στις προσεγγίσεις των ερευνητικών ομάδων είναι η έλλειψη χρήσης profile-based μεθόδων σχεδόν απο το σύνολο των ομάδων, καθώς χρησιμοποιήθηκε μόνο από μια ερευνητική ομάδα. [11]

Submission		**Features**	**Weighting / Normalization**	**Paradigm**	**Classifier**	**Parameter settings**
Team	Reference					
Custódio and Paraboni	[6]	char & word n-grams	TF-IDF	i-b	ensemble	global
Gagala	[8]	various n-grams	none	i-b	NN	global
Halvani and Graner	[14]	compression	none	p-b	similarity	global
López-Anguita et al.	[25]	complexity	L2-norm.	i-b	SVM	l-s
Martín dCR et al.	[4]	various n-grams	log-entropy	i-b	SVM	l-s
Miller et al.	[13]	various n-grams & stylistic	TF-IDF & TF	i-b	SVM	global
Murauer et al.	[28]	char n-grams	TF-IDF	i-b	SVM	local
PAN18-BASELINE		char n-grams	TF	i-b	SVM	global
Schaetti	[41]	tokens	embeddings	i-b	ESN	local
Yigal et al.	[13]	various n-grams & stylistic	TF-IDF & TF	i-b	SVM	global

Σχήμα 4.5: Προσεγγίσεις πάνω στο dataset
Πηγή: Overview of the Author Identification Task at PAN-2018

4.1.6.3 Επιδόσεις προσεγγίσεων-ομάδων με διάφορες μετρήσεις

Submission	Macro F1	Macro Precision	Macro Recall	Micro Accuracy	Runtime
Custódio and Paraboni	**0.685**	**0.672**	**0.784**	**0.779**	00:04:27
Murauer et al.	0.643	0.646	0.741	0.752	00:19:15
Halvani and Graner	0.629	0.649	0.729	0.715	00:42:50
Mosavat	0.613	0.615	0.725	0.721	00:03:34
Yigal et al.	0.598	0.605	0.701	0.732	00:24:09
Martín dCR et al.	0.588	0.580	0.706	0.707	00:11:01
PAN18-BASELINE	0.584	0.588	0.692	0.719	00:01:18
Miller et al.	0.582	0.590	0.690	0.711	00:30:58
Schaetti	0.387	0.426	0.473	0.502	01:17:57
Gagala	0.267	0.306	0.366	0.361	01:37:56
López-Anguita et al.	0.139	0.149	0.241	0.245	00:38:46
Tabealhoje	0.028	0.025	0.100	0.111	02:19:14

Σχήμα 4.6: Επιδόσεις πάνω στο dataset με διαφορετικές μετρήσεις
Πηγή: Overview of the Author Identification Task at PAN-2018

4.1.6.4 Επιδόσεις προσεγγίσεων-ομάδων

Submission	Overall	English	French	Italian	Polish	Spanish
Custódio and Paraboni	**0.685**	0.744	**0.668**	0.676	0.482	**0.856**
Murauer et al.	0.643	**0.762**	0.607	0.663	0.450	0.734
Halvani and Graner	0.629	0.679	0.536	**0.752**	0.426	0.751
Mosavat	0.613	0.685	0.615	0.601	0.435	0.731
Yigal et al.	0.598	0.672	0.609	0.642	0.431	0.636
Martín dCR et al.	0.588	0.601	0.510	0.571	**0.556**	0.705
PAN18-BASELINE	0.584	0.697	0.585	0.605	0.419	0.615
Miller et al.	0.582	0.573	0.611	0.670	0.421	0.637
Schaetti	0.387	0.538	0.332	0.337	0.388	0.343
Gagala	0.267	0.376	0.215	0.248	0.216	0.280
López-Anguita et al.	0.139	0.190	0.065	0.161	0.128	0.153
Tabealhoje	0.028	0.037	0.048	0.014	0.024	0.018

Σχήμα 4.7: Επιδόσεις πάνω στο dataset
Πηγή: Overview of the Author Identification Task at PAN-2018

4.1.6.5 Επιδόσεις προσεγγίσεων σε διαφορετικά μεγέθη δεδομένων

Submission	20 Authors	15 Authors	10 Authors	5 Authors
Custódio and Paraboni	**0.648**	**0.676**	**0.739**	**0.677**
Murauer et al.	0.609	0.642	0.680	0.642
Halvani and Graner	0.609	0.605	0.665	0.636
Mosavat	0.569	0.575	0.653	0.656
Yigal et al.	0.570	0.566	0.649	0.607
Martín dCR et al.	0.556	0.556	0.660	0.582
PAN18-BASELINE	0.546	0.532	0.595	0.663
Miller et al.	0.556	0.550	0.671	0.552
Schaetti	0.282	0.352	0.378	0.538
Gagala	0.204	0.240	0.285	0.339
López-Anguita et al.	0.064	0.065	0.195	0.233
Tabealhoje	0.012	0.015	0.030	0.056

Σχήμα 4.8: Επιδόσεις πάνω στο dataset πάνω σε κάθε διαφορετικό μέγεθος dataset
Πηγή: Overview of the Author Identification Task at PAN-2018

4.1.6.6 Συμπεράσματα

Από τους παραπάνω πίνακες που παρουσιάστηκαν, μπορούμε να διαπιστώσουμε όπως αναφέραμε και παραπάνω ότι η χρήση SVM κυριαρχεί στις περισσότερες προσεγγίσεις και επιφέρει σημαντικά καλά αποτελέσματα, σε σχέση με την χρήση άλλων μεθόδων. Επιπλέον, μπορούμε να παρατηρήσουμε ότι σε κάθε γλώσσα τα καλύτερα αποτελέσματα φαίνεται να ποικίλουν απο προσέγγιση σε προσέγγιση καθώς και οι επιδόσεις ανάλογα με τον αριθμό των συγγραφέων.

4.1.7 C10 Dataset

4.1.7.1 Ανάλυση Dataset

Σε αντίθεση με το dataset της προηγούμενης ενότητας,το συγκεκριμένο σύνολο δεδομένων περιέχει δεδομένα τα οποία εξετάζουν τις δυνατότητες του Γλωσσικού Μοντέλου πάνω σε ίδιου τομέα δεδομένα, είναι επομένως ένα Simple Domain Σύνολο Δεδομένων. Εδώ είχαμε 10 δημιουργούς, για αυτό το λόγο και ο αριθμός 10 στο όνομα C10. Να αναφερθεί ότι αυτή η συλλογή δεν συλλέχτηκε αρχικά για σκοπό εκπόνησης μελετών κατανομής συγγραφικού ύφους.

Υποσύνολο του Reuters Corpus v.1 όπου για τον κάθε έναν από τους 10 συγγραφείς υπήρχαν 100 κείμενα (19 φράσεις ανά κείμενο και 425 λέξεις). Η θεματική κατηγορία είναι η CCAT (εταιρικά και βιομηχανικά νέα). Έχει χρησιμοποιηθεί και σε αρκετές μελέτες ενώ το μέγεθος του corpus για το testing και training είναι ίσου μεγέθους. Οι συγγραφείς της CCAT-10 γράφουν κυρίως κείμενα που μεταδίδονται online όπως ειδήσεις, θέματα χρηματοπιστωτικών αγορών και άλλες παρόμοιων κλάδων πληροφορίες. Και αυτά τόσο στο train όσο και test σώμα. Στο CCAT-10 υπάρχουν λέξεις που είναι χαρακτηριστικές για τον συγγραφέα (author-specific) και τείνουν να εμφανίζονται σε πολλά έγγραφα του. Έτσι, αυτές οι λέξεις είναι πολύ χρήσιμες στο συγκεκριμένο corpus κειμένων.

Επιπρόσθετα το n gram σε αυτό το dataset στην αρχή χρησιμοποιήθηκε αρκετά, με 3 gram τύπο μιας και έδινε τα πιο αποτελεσματικά συμπεράσματα. Για κάθε κατηγορία, εξετάσαμε μόνο τα 3-grams που εμφανίζονται τουλάχιστον πέντε φορές στα εκπαιδευτικά έγγραφα. Δε βοήθησε πρακτικά εν τέλει στο C10, μιας και μερικοί από τους συγγραφείς της συλλογής χρησιμοποίησαν συστηματικά "υπογραφές" στο τέλος των εγγράφων τους και ορισμένοι χρησιμοποίησαν αναφορές πολύ συχνά. Άρα για τη μέτρηση συχνοτήτων των λέξεων στα εξεταζόμενα κείμενα κλπ, δε είναι χρήσιμη η θεωρία των N grams που έχουμε αναφέρει σε προηγούμενο κεφάλαιο[7].

Να αναφερθεί εδώ ότι παρόμοιο dataset που χρησιμοποιήθηκε σε έρευνες και ήταν συχνά σε σύγκριση με το C10/ CCAT ήταν το guardian (με κριτικές σε άρθρα από τη γνωστή αγγλική εφημερίδα Guardian από 14 συγγραφείς σε 4 διαφορετικά θέματα, Κόσμος, UK, Κοινωνικά, και Πολιτικά), το οποίο όμως ήταν για cross domain έρευνες πάνω στην αναγνώριση συγγραφικού ύφους, με καλύτερα αποτελέσματα στις συγκρίσεις να υπήρχαν στο CCAT dataset (δεδομένου κιόλας ότι απορρίφθηκαν τα mid-word n-grams, που ήταν ένας από τους καλύτερους n-gram τύπους)!

4.1.7.2 Προσεγγίσεις πάνω στο Dataset

Πάνω στο συγκεκριμένο dataset έχουν γίνει πρόσφατα διάφορες μελέτες. Πρώτον υπήρξε η μελέτη των Sapkota και Bethard. Παρουσιάζεται στο paper με τίτλο "Not All Character N-grams Are Created Equal: A Study in Authorship Attribution" [7]. Αναφέρεται στους χαρακτήρες n-grams, που έχουν αναγνωριστεί ως το ένα επιτυχημένο εργαλείο τόσο σε μία simple όσο και σε μια cross domain αναγνώριση συγγραφέα. Αναγνωρίζονται στην έρευνα οι υποομάδες n-grams και αντιστοιχήθηκαν σε κάποια γλωσσικά χαρακτηριστικά όπως η "μορφοσύνταξη", θεματικό περιεχόμενο, στυλ κλπ. Αξιολογήθηκε έπειτα η προβλεψιμότητα καθεμιάς από αυτές τις ομάδες στις προαναφερθείσες ρυθμίσεις (simple και cross domain). Τέλος συλλέχθηκαν τα ngrams χαρακτήρων που πήραν τις σωστές πληροφορίες σχετικά με τις επιγραφές και τη στίξη που αντιπροσωπεύουν σχεδόν όλη τη δύναμη του χαρακτήρα n-grams ως χαρακτηριστικά.

Η μελέτη αυτή εν γένει μπορεί να συνεισφέρει σε μελλοντικές εργασίες και άλλες ερ-

γασίες ταξινόμησης νέες γνώσεις σχετικά με τη χρήση των n-grams γενικά. Πέρα από το C10 dataset χρησιμοποιήθηκε το Guardian για το οποίο αναφέρθηκαν κάποια χαρακτηριστικά παραπάνω. Σε αυτά τα δύο datasets έγινε μελέτη στο συγκεκριμένο paper με πιο ορθά αποτελέσματα να δίνει το C10.

Οι δύο συλλογές χρησιμοποιούνται και στο paper του Σταματάτου με τίτλο "Authorship Attribution Using Text Distortion" [18]. Εδώ, για το C10 (CCAT) dataset γίνεται εξαγωγή των ιδιοτήτων από τα εξεταζόμενα κείμενα είτε σε μορφή χαραλτήρων n-gram είτε σε token n-grams και στη συνέχεια χρησιμοποιείται SVM ταξινομητής [31]. Από τα αποτελέσματα παίρνουμε τα καλύτερα κείμενα για το testing, όπου ήταν στη προκείμενη μελέτη του paper για τα baseline models. Θα πρέπει να σημειωθεί ότι τα baseline n-gram μοντέλα χαρακτηριστιών στις περισσότερες περιπτώσεις είναι πιο αποτελεσματικά από τα baseline tokens n-gram. Για τη συλλογή Guardian χρησιμοποιήθηκε παρόμοιος τρόπος μελέτης και εξόρυξης χρήσιμων στοιχείων.

Εν γένει, στο CCAT-10 σε σύγκριση με το Guardian από τα 2 papers καταλήγουμε ότι υπήρχαν περισσότερες λέξεις που είναι author-specific και τείνουν να εμφανίζονται σε πολλά έγγραφα από αυτόν τον συγγραφέα (τόσο τα tf όσο και τα df τους είναι υψηλά). Άρα, οι λέξεις αυτές είναι χρήσιμοι δείκτες της συγγραφής για αυτό το συγκεκριμένο σώμα, με το θέμα του να είναι βιομηχανικά και εταιρικά νέα σε σύγκριση με το Guardian dataset που περιείχε άρθρα από την εφημερίδα Guardian.

4.2 Αναγνώριση ύφους ιστοσελίδας

4.2.1 Ύφος αρθρογράφου

Εκτός από διαφύλαξη συμφερόντων για μυθιστορήματα και γενικά βιβλία, δικαιώματα χρίζουν και πολλά άρθρα και ρεπορτάζ που γίνονται από δημοσιογράφους ή αρθρογράφους. Ο χρήστης προσεγγίζει το θέμα του τεκμηριωμένα, χρησιμοποιεί πολλές φορές ειδικό λεξιλόγιο και συχνά υιοθετεί το δικό του ύφος (σοβαρό, ουδέτερο, αυστηρό). Το κείμενο του άρθρου μπορεί να πλαισιώνεται από φωτογραφίες, στατιστικά στοιχεία, γραφικές παραστάσεις, συνεντεύξεις[2]. Στοιχεία που έχει αποκτήσει ο δημοσιογράφος με δικό του κόπο και για αυτό θέλει να προστατέψει τα συμφέροντά του για διαφόρους λόγους.

Η εξέλιξη της τεχνολογίας και η ραγδαία ανάπτυξη του ίντερνετ έχουν βοηθήσει πολλούς επιτήδειους να βγάλουν κέρδος με αντιγραφή άρθρων ή κομματιών από άρθρα μιας και παρουσιάζουν στους δικούς τους ιστότοπους σαν προσωπικά τους αποτελέσματα. Ακριβώς παρόμοια κατάσταση μπορεί να έχουμε και για το ύφος ενός αρθρογράφου.Ύφος το οποίο μπορεί να αντιγράφει από κάποιο τρίτο πρόσωπο, χωρίς την άδεια ή ακόμα και την ενημέρωση του πρώτου. Είναι σίγουρα πολύ δύσκολα να κατηγορήσει κάποιος έναν άλλον ότι τον αντιγράφει στο ύφος, όταν μιλάμε για σύντομα άρθρα ή ειδήσεις, σε αντίθεση με μυθιστορήματα με τεράστιες πλοκές σε μία μεγάλη ιστορία κλπ. Αλλά ακόμα και σε αυτό το τομέα των άρθρων μπορεί και έχει παρατηρηθεί αντιγραφή ύφος ενός δημοσιογράφου ή εν μέρει κλοπή στοιχείων του ύφους του αυθεντικού.

Το κάθε website μπορεί να κερδίσει τον χρήστη είτε από το περιεχόμενό του, είτε από το πως παρουσιάζεται στο κοινό, το ύφος του δηλαδή. Είναι σημαντικοί παράγοντες και για αυτό χρίζουν μελέτης. Και οι δύο προαναφερθείς λόγοι μελετώνται εκτενώς μιας και πολλά ζητήματα θέτονται πάνω στη λειτουργία μίας ιστοσελίδας και τα δικαιώματα που δημιουργούνται, για τους δημιουργούς τους.

4.2.2 Τρόποι Κατανομής Ιστοσελίδων

Τα περιεχόμενα των ιστοσελίδων μπορούμε να τα χωρίσουμε πρώτον αναλόγως με το είδος των θεμάτων τους. Μπορεί να περιέχουν αθλητικό, πολιτικό ή χρηματιστηριακό περιεχόμενο για παράδειγμα. Έτσι το λεξιλόγιο των άρθρων προσαρμόζεται στο κοινό του, οι διαφημίσεις που υπάρχουν αφορούν το αντίστοιχο θέμα και άλλα πολλά στοιχεία που συσχετίζονται πάντα με τη θεματική της ιστοσελίδας. Αυτός είναι ο ένας τρόπος που μπορούμε να διαχωρίσουμε τα websites.

Δεύτερον μπορούμε να τα χωρίσουμε αναλόγως με το ύφος τους, το στιλ δηλαδή που έχουν συγκροτηθεί τα δομικά τους στοιχεία, πχ αν είναι ένα blog ή e-shop η ιστοσελίδα. Σε αυτό το κριτήριο, το είδος του ύφους, μπορούμε να πούμε ότι έχουμε να κάνουμε με μία ομάδα ιστοσελίδων που έχουν πολλές κοινές στιλιστικές ομοιότητες, παρά θεματικές. Το πως παρουσιάζεται η ροή της κεντρικής σελίδας στον αναγνώστη ή ο τρόπος ανάπτυξης ενός άρθρου είναι κάποια στοιχεία του συγκεκριμένου είδους αναγνώρισης. Οι μελέτες σχετικά με την αναγνώριση του ύφους της ιστοσελίδας επικεντρώνονται στον ορισμό των κατάλληλων χαρακτηριστικών κειμένου που μετράνε ποσοτικά τις στιλιστικές ιδιότητες των ειδών.

[2]https://en.wikipedia.org/wiki/Article$_{(publishing)}$

4.2.3 Είδη/Στιλ Ιστοσελίδων

Συχνά υπάρχουν διαφωνίες για το καθαρό ορισμό ενός είδους μίας ιστοσελίδας. Για παράδειγμα άτομα από διαφορετικές χώρες μπορούν να έχουν διαφορετικές απόψεις ως προς το χτίσιμο της δομής ενός website για να χαρακτηριστεί πχ wiki ή marketing. Και επιπροσθέτως με τη πάροδο του χρόνου μπορούν να δημιουργηθούν και νέα είδη ιστοσελίδων.

Εκτός από αυτό, μία ιστοσελίδα μπορεί να περιέχει και κατάστημα και μηχανή αναζήτησης και chat για αυτό μπορεί να μην χαρακτηρίζεται από μία ταμπέλα και μόνο. Εδώ θα μπορούσαν να βοηθήσουν single-label ανιχνευτές, αν εξετάζουν το κάθε κομμάτι της ιστοσελίδας ξεχωριστά [16].

Εν γένει, μία σχετικά αντικειμενική λίστα που έχει προταθεί με τα είδη (genres) ιστοσελίδων αναλόγως με το ύφος τους παρουσιάζεται παρακάτω:

1. Portal
2. News
3. Wiki
4. Blog
5. Personal
6. Advocacy
7. Content Aggregator
8. Educational
9. Marketing/Business
10. Social Network
11. Informational
12. Web Application

Πηγή: https://www.slideshare.net/mrblueofld s/11-types-of-web-sites

4.2.4 Μελέτες πλήθους ειδών

Η αναζήτηση ενός χρήστη για μία ιστοσελίδα στο διαδίκτυο μπορεί να διευκολυνθεί αν έχουν χωριστεί προηγουμένως οι υποψήφιες ιστοσελίδες, αναλόγως με το ύφος ή το θέμα/περιεχόμενό τους. Μπορούν να συνδυαστούν εννοείται και οι δύο παράγοντες για να χαρακτηριστεί μία ιστοσελίδα ως προς το στιλ της αλλά ακόμα οι περισσότερες μηχανές αναζήτησης δε μπορούν να τους συνδυάσουν για τη αναζήτησης ενός website τελείως αποτελεσματικά [16].

Και σίγουρα είναι ένα θέμα το οποίο πρέπει να λυθεί γιατί θα βοηθήσει όλους τους χρήστες να βρουν άμεσα και αποτελεσματικά στο ίντερνετ αυτό που επιθυμούν, πχ προτείνοντας ερωτήματα που περιγράφουν καλύτερα τις ανάγκες του χρήστη. Με δύο τρόπους μπορεί να γίνει αυτό, είτε να επιλέγει ο χρήστης στην αρχή στην αναζήτηση το είδος της ιστοσελίδας, blog, e-shop κλπ (Rosso, 2005), ή τα αποτελέσματα αναζήτησης, μέσω των

λέξεων-κλειδιά, να μπορούν να ομαδοποιηθούν ανάλογα με το είδος (Bekkerman,2008). Πληροφορίες βασισμένες σε HTML λαμβάνονται κάποιες φορές υπόψιν, όπως πρότειναν οι Meyer zu Eissen Stein το 2004, οι Boese Howe το 2005 και ο Santini το 2007.

Ιδανικά, πριν το κάθε ψάξιμο σε μία μηχανή αναζήτησης θα μπορούσε να επιτευχθεί η ολοκληρωτική αυτοματοποιημένη εξαγωγή των χαρακτηριστικών της κάθε υποψήφιας σελίδας που διαμορφώνουν το στιλ της. Όμως έχουν παρατηρηθεί ως τώρα και διάφορα εμπόδια. Η συνεχής ανάπτυξη του διαδικτύου δε βοηθάει τους (χειροκίνητους) classifier των websites μιας και πρέπει συνεχώς να προσαρμόζονται στις νέες και τροποποιημένες συνθήκες. Συν ότι πολλά, νέα διαφορετικά είδη μπορούν να δυσκολέψουν έναν classifier, που έχει μάθει να λειτουργεί καλύτερα για ένα είδος.

Τέλος απαιτείται συχνά και η εφαρμογή εργαλείων επεξεργασίας φυσικής γλώσσας, επομένως οι επιδόσεις εξαρτώνται και από τη γλώσσα και ότι η κάθε μελέτη επικεντρώνεται σε ένα διαφορετικό σύνολο ειδών. Παρακάτω θα δούμε περιληπτικά κάποιες από τις προσεγγίσεις που έγιναν στο παρελθόν και τα αποτελέσματα που έδωσαν.

Πρώτες Προτάσεις		
Ερευνητές	Αριθμός Ειδών	Έτος
Lee, Myaeng	-	2002
M.Eissen, Stein	8	2004
Lim	15	2005
Kim, Ross	24	2007
Santini	7	2007
Vidulin	20	2007
Levering	4	2008

Πίνακας 4.5: Προτεινόμενοι Αριθμοί για Είδη Websites
Πηγή: Learning to Recognize Webpage Genres (Σταματάτος Καναρης)

Όπως φαίνεται και από τον παραπάνω πίνακα, υπήρχαν μέσα σε λίγα χρόνια μεγάλο ενδιαφέρον και πολλές προτάσεις για το πως να ξεχωρίζουμε καλύτερα τις ιστοσελίδες, αναλόγως τα είδη τους. Δε συμφωνούσαν οι έρευνες, κάτι όντως υποκειμενικό, και δεν ήταν σαφές από κοινού ο αριθμός των ειδών. Γενικά όμως, σημαντικά στοιχεία για τη μελέτη ύφους ιστοσελίδων είναι για παράδειγμα η συχνότητα ενός όρου ή ολόκληρης φράσης, συμβόλων και σημείων στίξης ενώ υπάρχουν και πιο περίπλοκα και λεπτομερή βοηθήματα. Τέτοια εργαλεία εισάγουν και θόρυβο στην εκτίμηση των χαρακτηριστικών, μιας και είναι ακόμα ατελείς.

Μελέτη επίσης πραγματοποιήθηκε για την ανάλυση και των δομικών στοιχείων των HTML websites (πλήθος των tags, εικόνων, scripts, URL, links κλπ), που είναι και αυτά πολύ χρήσιμα για την εξαγωγή συμπερασμάτων για μία ιστοσελίδα, με τη προϋπόθεση να είναι HTML websites. Επιπροσθέτως παρατηρήθηκε ότι αν χωρίσουμε σε μέρη μία ιστοσελίδα (title, head, body κλπ), τα πιο σημαντικά στοιχεία για το περιεχόμενο και την αναγνώριση του είδους μίας ιστοσελίδας υπάρχουν στα anchor texts (αυτά που οδηγούν σε υπερσύνδεσμο, δηλαδή σε άλλες σελίδες/πληροφορίες) και στο κύριο σώμα.

Μία άλλη προσέγγισή είναι η ανάλυση των οπτικών χαρακτηριστικών (visual features) ενός website οδηγούν σε καλύτερα αποτελέσματα της κατηγοριοποίησης ιστοσελίδων ενώ άλλες προσεγγίσεις επίσης χρησιμοποιούν ένα αυθαίρετο σύνολο χαρακτηριστικών, συνήθως μικρού μεγέθους αλλά με πολλές χρήσιμες λειτουργίες. Προϋπόθεση εδώ η συνεχή και χειροκίνητη ενημέρωση όλων των λειτουργιών ώστε να είναι σε θέση να προσαρμοστούν στις τρέχουσες συνθήκες των ειδών.

Τέλος μία άλλη σημαντική πρόταση πρότεινε στην αρχή να υπάρχει ένα μεγάλο σύνολο χαρακτηριστικών που θα περιλαμβάνει πάρα πολλά χαρακτηριστικά και στη συνέχεια να χρησιμοποιόταν ένας αλγόριθμος επιλογής χαρακτηριστικών για την εξαγωγή των πιο χρήσιμων χαρακτηριστικών. Αυτό όμως θα είχε πραγματικό νόημα σε ένα corpus, άρα δε μπορεί να χρησιμοποιηθεί ταυτόχρονα σε δύο διαφορετικά. Για αυτό χρειάζεται μια γενική αυτοματοποιημένη μεθοδολογία για την εξαγωγή στοιχείων που δεν εξαρτώνται από σωματίδια, είδη και φυσικές γλώσσες, άρα για αρκετά corpus ταυτόχρονα.

4.2.5 Μεθοδολογίες

4.2.5.1 Προ-επεξεργασία Δεδομένων Προσέγγιση με βάση τις λέξεις(bag of words)

Το μοντέλο bag-of-words είναι μία από τις πιο δημοφιλείς μεθόδους αντιπροσώπευσης για την κατηγοριοποίηση αντικειμένων, μια απλουστευμένη αναπαράσταση. Χρησιμοποιείται στην επεξεργασία της φυσικής γλώσσας και στην ανάκτηση πληροφοριών (Information Retrieval). Σε αυτό το μοντέλο, ένα κείμενο (όπως μια πρόταση ή ένα έγγραφο) αναπαρίσταται ως ένα σύνολο/multiset με λέξεις, αγνοώντας τη γραμματική στη πρόταση και τη τάξη των λέξεων αλλά όχι τη πολλαπλότητα τους [3]. Αυτές οι εξεταζόμενες λέξεις χρησιμοποιούνται σε μοντέλα με τη συχνότητα εμφάνισης τους να χρησιμοποιείται ως χαρακτηριστικό γνώρισμα για την εκπαίδευση ενός ταξινομητή.

Επιπροσθέτως η προσέγγιση αυτή χρησιμοποιείται πολύ και στο image/object categorization, για εικόνες δηλαδή στην επιστήμη του computer vision, όπου εδώ βοηθητικό ρόλο εδώ παίζει ο αλγόριθμος K-Means [4]. Παρόλο που αρκετές μελέτες έχουν δείξει ενθαρρυντικά αποτελέσματα, άλλες μελέτες με το μοντέλο bag-of-words είναι σχεδόν ανέγγιχτες και απέτυχαν. Συσχετίζεται πολύ με το N Gram, έχουν πολλές κοινές ιδιότητες.

Για να γίνει πιο κατανοητή η απόδοση της συγκεκριμένης προσέγγισης ας δούμε ένα παράδειγμα παρακάτω (υπάρχει αναλυτικά στη wikipedia). Παρακάτω έχουμε δύο προτάσεις προς εξέταση:

- John likes to watch movies. Mary likes movies too.
- John also likes to watch football games.

Αντιπροσωπεύοντας το κάθε multiset ως ένα αντικείμενο JSON και αποδίδοντάς το σε javascript μεταβλητή, όπου για κλειδί/key έχουμε τη λέξη και για value τη συχνότητα της λέξης στη φράση, έχουμε το παρακάτω αποτέλεσμα:

- BoW2 = "John":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1;
- BoW1 = "John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1;

Εάν ενώσουμε τα δύο παραπάνω σε ένα ενιαίο set προτάσεων με όνομα πχ Bow3, τότε η νέα javascript παρουσίαση θα είναι η ακόλουθη:

- BoW3 = "John":2,"likes":3,"to":2,"watch":2,"movies":2,"Mary":1,"too":1, "also":1,"football":1,"games":1;

[3] https://en.wikipedia.org/wiki/Bag-of-words$_{m}odel$

[4] https://el.wikipedia.org/wiki/Ομαδοποίηση −

Άρα όπως παρατηρούμε η ένωση των Bow2 και Bow1 δίνει το Bow3 όπως εξηγείται περαιτέρω στη θεωρία του Disjoint Union. Αν μετατρέψουμε με τα παραπάνω βήματα μία φράση σε bag of words, τότε μπορούμε να υπολογίσουμε διάφορα χαρακτηριστικά του κειμένου, πχ τη συχνότητα των λέξεων μεμονωμένα. Η προσέγγιση αυτή εφαρμόζεται σε πολλές εφαρμογές με μία διαδεδομένη να είναι στα emails και στο φιλτράρισμα που τους γίνεται.

Επειδή όμως πολλές λέξεις χρησιμοποιούνται συχνά σε μία πρόταση αλλά δεν έχουν κάποια έννοια σημαντική για τη κατανόηση του κείμενου, πχ άρθρα, είναι καλό συχνά να δίνεται ένα συγκεκριμένο βάρος, συνήθως όσο πιο συχνή είναι η λέξη τόσο πιο μικρό να είναι το βάρος(αντίστροφο της συχνότητας δηλαδή). Εν κατακλείδι η συγκεκριμένη προσέγγιση για την ανάλυση ιστοσελίδων και εξόρυξη δεδομένων μέσω των κειμένων και εικόνων (image classification) που περιέχουν εφαρμόζεται συχνά και με μεγάλη επιτυχία.

Βέβαια ο συγκεκριμένος τρόπος επειδή μετράει μόνο το πλήθος εμφάνισης των λέξεων σε μία φράση, δεν μπορεί για παράδειγμα να βγάλει το συμπέρασμα ότι ένα ρήμα ακολουθεί ένα ουσιαστικό (πχ στα παραπάνω παραδείγματα Bow η λέξη likes ακολουθείται από το όνομα κάποιου ατόμου, John likes ή Mary likes). Εδώ θα αναλάβει δράση το N Gram όπως αναλύεται παρακάτω. Υπόψιν μπορούν να συνδυαστούν οι δύο αυτές προσεγγίσεις.

4.2.5.2 Προ-επεξεργασία Δεδομένων Προσέγγιση με βάση το N Gram

Στη πρόταση του parer Learning to Recognize Webpage Genres [16], με προεργασία το n gram, χρησιμοποιήθηκαν 2 προσφερόμενα εργαλεία, HTML tags και συνεχείς ακολουθίες χαρακτήρων από ένα δοσμένο δείγμα κειμένου που επεξεργάζεται μέσω n gram μεθοδολογίας. Εισάγεται μια σειρά από πειράματα για να δειχθεί η αξιοπιστία της προσέγγισής αυτής μέσω εκπαιδευμένου μοντέλου. Μέσω συγκρίσεων του κάθε n gram με παρόμοιά του, βρίσκουμε τα κυρίαρχα με βάση τη συχνότητα εμφάνισής τους n grams.

Πρώτα αφαιρέθηκαν όποιες πληροφορίες συσχετίζονται με HMTL κώδικα για να καθαριστεί το κείμενο. Έπειτα μετρήθηκαν οι πληροφορίες που συλλέχθηκαν. Για την αναγνώριση ταυτότητας, προσδιορισμός ιστοσελίδας στη περίπτωση μας, προτείνεται μέθοδος επιλογής χαρακτηριστικών για n-grams στοιχεία μεταβλητού μήκους. Λειτουργούν ανεξαρτήτως γλώσσας, προσαρμόζονται στα κείμενα στις σελίδες εύκολα και αποτελεσματικά, πχ ενάντια στον θόρυβο/σπαμ που περιέχουν πολλά websites και μπορούν άμεσα να αποσπάσουν χρήσιμες πληροφορίες, ειδικά συγκριτικά με άλλες μεθόδους.

Η λειτουργία που εκφράζει τη σημασία ενός n gram ορίζεται ως glue. Όσο μεγαλύτερη είναι το glue ενός n-gram, τόσο πιθανότερο είναι να συμπεριληφθεί στο σύνολο των κυρίαρχων n-γραμμών (πχ το glue του n gram |the...| είναι μεγαλύτερου του |thea...|). Τα χαρακτηριστικά που εξάγονται από τον προτεινόμενο αλγόριθμο από ένα δοσμένο corpus/σώμα μπορούν να μεταφερθούν σε άλλα corpa διαφορετικού είδους διαχώρισης ύφους ιστοσελίδων.

Για την εξαγωγή n gram στοιχείων χρησιμοποιούμε τον αλγόριθμο LocalMaxs (αλγόριθμος που υπολογίζει τα τοπικά μέγιστα που συγκρίνουν παρόμοια n grams μεταξύ τους) [26]. Η λογική του παρουσιάζεται παρακάτω:

$$if\,(C.length > 3)$$
$$g(C) \geq g(ant(C)) \wedge g(C) > g(succ(C)),\ \forall ant(C), succ(C)$$
$$if\,(C.length = 3)$$
$$g(C) > g(succ(C)),\ \forall succ(C)$$

Σχήμα 4.9: Κανόνας Αλγόριθμου LocalMaxs για το κυρίαρχο n gram
Πηγή: Learning to Recognize Webpage Genres

όπου g(c) ο όρος glue για το C n gram, ant(C) ο προηγούμενος όρος (συντομότερη σειρά/string με n-1 όρους) και ο succ(C) ο επόμενος όρος με n+1 όρους(C συν έναν ακόμη χαρακτήρα). Σύμφωνα με τον προτεινόμενο αλγόριθμο, τα 3 n grams συγκρίνονται μόνο με τα επόμενα/succ n grams με όριο τα 5 n grams, άρα ο αλγόριθμος αυτός θα ευνοήσει τα 3 και τα 5 n grams έναντι αυτών με τα 4.

Για τον υπολογισμό του glue του n gram χρησιμοποιείται ο τύπος Symmetrical Conditional Probability του Silva [27]. Ένα απλό παράδειγμα να είναι το ακόλουθο: η 3 gram λέξη "the" να έχει 2 σημεία διασποράς, το th*e και το t*he. Μικρότερο SCP συνεπάγεται με μικρότερο αρχικό n gram:

$$SCP(x,y) = p(x \mid y) \cdot p(y \mid x) = \frac{p(x,y)}{p(x)} \cdot \frac{p(x,y)}{p(y)} = \frac{p(x,y)^2}{p(x) \cdot p(y)}$$

Σχήμα 4.10: Κανόνας Αλγόριθμου LocalMaxs για το κυρίαρχο n gram
Πηγή: Learning to Recognize Webpage Genres

Η δομική ανάλυση μιας ιστοσελίδας είναι σε HTML κώδικα όπως αναφέρθηκε και παραπάνω και αναλύεται μέσω του BOW τύπου, ένα δωρεάν πρότυπο bootstrap HTML5. Πρώτα εξαγάγουμε μία λίστα με όλες τις ετικέτες HTML που εμφανίζονται τουλάχιστον τρεις φορές σε ολόκληρη τη συλλογή των ιστοσελίδων και έπειτα παρουσιάζουμε τη κάθε ιστοσελίδα με διάνυσμα χρησιμοποιώντας τη συχνότητα εμφάνισης αυτών των ετικετών HTML στη σελίδα.

Στη συνέχεια, χρησιμοποιούμε πάνω στο αποτέλεσμα τον ReliefF αλγόριθμο (Robnik-Sikonja Kononenko, 2003) για τη μείωση της διαστασιολόγησης, δηλαδή της μείωσης του αριθμού των τυχαίων μεταβλητών [29]. Με αυτό τον τρόπο, αποφεύγουμε κάθε χειροκίνητο ορισμό χρήσιμων ετικετών HTML και η διαδικασία είναι όλη πλήρως αυτοματοποιημένη και οι εξαγόμενες ετικέτες προσαρμόζονται στις συγκεκριμένες ιδιότητες ενός corpus.

Παρότι δεν υπάρχει πολύ μεγάλο διαθέσιμο υλικό σε ιστοσελίδες και τα είδη της προς εξέταση και όπως αναφέρθηκε σε προηγούμενο κεφάλαιο υπάρχουν διαφορετικές προσεγγίσεις και αποτελέσματα ως προς τον ορισμό των ειδών τους, υπάρχουν διάφορα μικρά corpa ιστοσελίδων κατάλληλα για τέτοια μελέτη (Rehm, 2008).

4.2.5.3 Χρήση του SVM(Support Vector Machine)

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) αποτελούν μία σύγχρονη αποτελεσματική προσέγγιση της επίλυσης ζητημάτων κατηγοριοποίησης κειμένων και εικόνων

όπως αναλύθηκε σε παραπάνω κεφάλαιο. Μπορούν να χρησιμοποιηθούν αποτελεσματικά και στην αναγνώριση των ειδών websites, μιας και περιέχουν μεγάλου όγκου κείμενα και πολλές εικόνες. Με το SVM εφαρμόζουμε έναν classifier ιστοσελίδων, με το corpus να το αποτελούν οι πιο συχνές λέξεις που βρέθηκαν στο website.

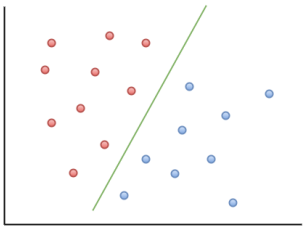
Τα προτερήματα του SVM πάνω σε ολόκληρο το τομέα του text classification γενικά (άρα και του website recognition) είναι αρκετά. Η μεθοδολογία ξεκίνησε προέκυψε από τη βαθύτερη ανάλυση της στατιστικής θεωρίας μάθησης (statistical learning theory) [5]. Είναι supervised μοντέλα (ξέρουμε συνήθως που κυμανθεί το output, labeled δηλαδή, για ένα συγκεκριμένο input). Έχουν τη δυνατότητα να αναγνωρίσουν και χειρόγραφες προτάσεις που υπάρχουν μέσα σε εικόνες στο διαδίκτυο.

Πρακτικά, οι Μηχανές Διανυσμάτων Υποστήριξης ΜΔΥ (Support Vector Machines) χωρίζουν τα δεδομένα σε δύο υποκατηγορίες αν και έχουν γίνει και προσεγγίσεις με πολλέςς υποκατηγορίες (multiple labels). Δυαδικά, με 1 για τα δεδομένα που ταιριάζουν και -1 για αυτά που δε ταιριάζουν. Αντικείμενο ορίζουμε μία γραμμή πίνακα (ΔΙΑΝΥΣΜΑ) που έχει ένα συγκεκριμένο πλήθος χαρακτηριστικών, χαρακτηριστικά ενός αντικειμένου είναι π.χ. πλάτος, ύψος και το βάρος ενός τραπεζιού [31]. Αν το σκεφτούμε γραφικά, σαν ένα γράφημα, τις δύο υποκατηγορίες τις χωρίζει μία γραμμή που κάνει κατανοητή στο χώρο τη διαμοίραση των στοιχείων. Οι φορείς που είναι πιο κοντά στη γραμμή αυτή είναι support vectors. Να τονιστεί ότι είναι καλό η συγκεκριμένη γραμμή ή υπερεπίπεδο(hyperplane) πρέπει να μας δίνει μεγάλο περιθώριο/κενό ανάμεσα στις θετικά και αρνητικά αποτελέσματα για να έχουμε ορθά αποτελέσματα και κρίση της διαχώρισης.

Επίσης στη διαχώριση των δεδομένων (μίας ιστοσελίδας στη περίπτωσή μας) δεν είναι δυνατόν άμεσα να έχουμε πολλές υπό-ομάδες για τα data μας (multiclass classification), αλλά μόνο δύο. Μπορούμε όμως να εκπαιδεύσουμε SVM πολλών "one versus all" μοντέλων και να τα συνδυάσουμε για να εφαρμόσουμε μια πολυκλαδική ταξινόμηση στο τέλος. Ουσιαστικά να ενώσουμε όλες σχεδόν τις ομάδες και αν τις παρουσιάσουμε σαν μία μεγάλη. Εκτός από μία που θα μείνει σαν τη δεύτερη ομάδα και θα συγκριθεί με τη μεγάλη ομάδα που δημιουργήθηκε και αποτελείται από τις υπόλοιπες.

Έτσι μέσω του SVM μπορούμε να δούμε σε γραφικό χώρο τα training data, ιστοσελίδων στη περίπτωσή μας, χωρισμένα και να γίνουν πιο εύκολα κατανοητά. Η πηγή μας είναι οι προτάσεις που υπάρχουν στην εξεταζόμενη ιστοσελίδα μας όπως και διάφορες εικόνες. Μπορούν και οι εικόνες να χωριστούν στις δύο (ή και παραπάνω σε κάποιες πιο περίπλοκες περιπτώσεις) υποκατηγορίες μέσω του SVM. Άρα και στο κλάδο αναγνώρισης είδους/ύφους ιστοσελίδας μπορούμε να χωρίσουμε τα στοιχεία του website που συλλέξαμε και να τα χωρίσουμε επιτυχώς μέσω ενός SVM. Παρακάτω φαίνεται γραφικά ένα πολύ απλό παράδειγμα διαχώρισης data με δύο υποκατηγορίες για τα data μας:

[5] https://en.wikipedia.org/wiki/Statistical$_l earning_t heory$



Σχήμα 4.11: A gentle introduction to support vector machines using R
Πηγή:https://eight2late.wordpress.com/2017/02/07/a-gentle-introduction-to-support-vector-machines-using-r/

4.2.6 Προσεγγίσεις και Συγκρίσεις Datasets

Όλες οι προσεγγίσεις, όχι μόνες τους δε μπορούν να δώσουν ικανοποιητικά αποτελέσματα και να καλύψουν όλο το εύρος του διαδικτύου και των ειδών των websites. Το κάθε collection (dataset) χωρίζει τα websites σε δικές τους κατηγορίες, που τις περισσότερες φορές δεν είναι ίδιες με κατηγορίες άλλων datasets. Στατιστικά, οι δύο ετικέτες που είναι συνεπείς και κοινές στις περισσότερες συλλογές είναι οι Συχνές Ερωτήσεις(FAQ) και το Κατάστημα.

Να αναφερθεί εδώ ότι και άλλα datasets έχουν πολύ καλά αποτελέσματα στο προσδιορισμού είδους ιστοσελίδας, με τα καλύτερα αποτελέσματα να φτάνουν στο 96,5/100 για το SANTINIS (Santini, 2005). Ο τύπος των στοιχείων στις διάφορες ιστοσελίδες που εξετάζουν τα datasets δεν είναι το ίδιο. Πχ το format του SANTINIS και του KI-04 είναι μόνο HTML κώδικας. Δε συμπεριλαμβάνονται εικόνες της εκάστοτε ιστοσελίδας για παράδειγμα, όπως συμβαίνει στη συλλογή του MGC (Vidulin, 2007).

Επίσης, εξαιτίας του γεγονότος ότι το κάθε dataset χρησιμοποιεί το δικό τους σύνολο κατηγοριών και ετικετών, σε πολλές περιπτώσεις απαιτείται ταίριασμα (mapping n προς n) ανάμεσα στα datasets και στις κατηγορίες του [32]. Πχ η κατηγορία MGC μπορεί να περιέχει διάφορες καταχωρήσεις ή σημειώσεις, κάθε μία από τις οποίες πρέπει να αντιστοιχηθεί στις κατηγορίες των άλλων datasets. Δεύτερον υπάρχει πάντα και το ενδεχόμενο να μη διατίθεται η εξεταζόμενη ετικέτα της μίας συλλογής σε κάποια άλλη ή να είναι αρκετές ετικέτες ενός είδους σε κάποια συλλογή αρκετά λεπτομερώς ορισμένες (πχ front Page στο SANTINIS), ενώ σε άλλα datasets χρησιμοποιούνται αρκετά ευρείες έννοιες στις ετικέτες (Informative στο MGC πχ). Τρίτον κάποιες συλλογές/datasets επιλέγουν τυχαία τις ιστοσελίδες προς εξέταση ενώ άλλες τις μελετάνε και αποφασίζουν ποιες θα εξετάζουν.

Source	# texts	# genres	Format
HGC (Stubbe and Ringlstetter, 2007)	1412	34	HTML only
I-EN-Sample (Sharoff, 2010)	250	7	TXT from HTML
KI-04 (Meyer zu Eissen and Stein, 2004)	1205	8	HTML only
KRYS I (Berninger et al., 2008)	6200	70	PDF
MGC (Vidulin et al., 2007)	1536	20	HTML with images
SANTINIS (Santini, 2010)	1400	7	HTML only
Combined (Santini and Sharoff, 2009)	9849	8	TXT from HTML
Brown Corpus (Kučera and Francis, 1967)	500	10	TXT
BNC (Lee, 2001)	4053	70	TXT

Σχήμα 4.12: Datasets ταξινόμησης είδους των webpages και ο τρόπος συλλογής στοιχείων
Πηγή: The Web Library of Babel: evaluating genre collections

Ωστόσο ακόμη και στη περίπτωση του Καταστήματος υπάρχει dataset (HGC, Stubbe και Ringlstetter, 2007) που δε την έχει ή το MGC τη χωρίζει σε δύο υποκατηγορίες (Αγορές και Προωθητικές). Σε κάποια datasets οι κατηγορίες των websites μπορεί να χωριστούν ιεραρχικά (μία κατηγορία να είναι υποκατηγορία κάποιας άλλης) επίσης. Επιπρόσθετα το KI-04 εικάζει το είδος μέσα από μεγάλα σύνολα σελίδων που επιλέγονται τυχαία στο ίντερνετ.

Γενικά είναι αναγκαία η διαδικασία του mapping μεταξύ των datasets και ας μην είναι στο σύνολό της αλάνθαστη. Υπάρχει πάντα η πιθανότητα απώλειας κάποιας πληροφορίας πχ ή σύγχυσης των ετικετών που μπορεί να οδηγήσει σε λανθασμένο χειροκίνητο

remapping (καλύτερα στη περίπτωση αυτή να απορρίψουμε τελείως τα αμφιλεγόμενα έγγραφα από τα να τοποθετήσουμε σε κάποια κατηγορία). Να τονιστεί επίσης ότι σε πολλά datasets δεν λαμβάνονται υπόψιν τα HTML tags, μιας και έχει αποδειχθεί ότι δε βοηθάνε πολύ στη ταξινόμηση των ιστοσελίδων αναλόγως με το είδος (Κανάρης και Σταματάτος, 2007).

Κάθε ιστοσελίδα σε κάθε συλλογή μετατρέπεται στην αρχή σε απλό κείμενο(.txt) χρησιμοποιώντας διάφορα απλά εργαλεία όπως το lynx ή pdftotext [6], οτιδήποτε κεφαλαίο μετατρέπεται σε μικρό ενώ οι ετικέτες POS (Part of Speech) για POS n-grams παρήχθησαν από το εργαλείο TreeTagger (Helmut Schmid, (1995)) [33].

4.2.7 Προσεγγίσεις 7genre Dataset και KI-04 Dataset

4.2.7.1 Ορισμοί

Η συλλογή 7genre δημιουργήθηκε στις αρχές του 2005. Αποτελούνταν από 1.400 αγγλικές ιστοσελίδες οι οποίες χωρίστηκαν σε 7 είδη ιστοσελίδων(αναφέρονται παρακάτω), ακολουθώντας τα κριτήρια του «σχολιασμού από αντικειμενικές πηγές» και της «συνεκτικής γενικότητας». Δημιουργός του ο Santini. Θεωρείται ένα ισορροπημένο σώμα/corpus, με την έννοια ότι οι ιστοσελίδες είναι ισόποσα κατανεμημένες μέσα στα είδη, άρα 200 σελίδες ανά είδος.

Το KI-04 corpus (2004) από την άλλη περιείχε 1295 αγγλικές σελίδες. Λιγότερες όμως και συγκεκριμένα 800 εξετάστηκαν και ταξινομήθηκαν σε 8 διαφορετικά είδη από τους Meyer zu Eissen and Stein (100 ανά είδος) [32]. Σε αντίθεση με το 7genre, εδώ η διανομή των εξεταζόμενων ιστοσελίδων στα είδη δεν είναι ισορροπημένη. Βέβαια το KI-04 έχει κάποιες υποκατηγορίες ιστοσελίδων που είναι παρόμοιες με εκείνες του 7genre, όπως για παράδειγμα η E-SHOP του 7genre είναι ίδια με τη SHOP του KI-04 [16].

7Genre		KI-04	
Genres	**Pages**	**Genres**	**Pages**
BLOG	200	ARTICLE	127
E-SHOP	200	DOWNLOAD	151
FAQs	200	LINK COLLECTION	205
ONLINE NEWSPAPER FRONTPAGE	200	PORTRAYAL-PRIVATE	126
LISTING	200	DISCUSSION	127
PERSONAL HOME PAGE	200	HELP	139
SEARCH PAGE	200	PORTRAYAL-NON PRIVATE	163
		SHOP	167

Σχήμα 4.13: Κατηγορίες των KI-04 και 7genre datasets
Πηγή: Learning to Recognize Webpage Genres

4.2.7.2 Εργαλεία Lemmatization και Stemming

Η διαφορά μεταξύ του Stemming και του Lemmatization είναι ότι η Lemmatization εξετάζει το περιεχόμενο της λέξης και τη μετατρέπει στην ουσιαστική μορφή βάσης, ενώ η Stemming απλώς αφαιρεί τους τελευταίους χαρακτήρες της λέξης, οδηγώντας βέβαια συχνά σε λανθασμένες έννοιες και ορθογραφικά λάθη.

[6] https://pdftotext.com

4.2.7.3 Αποτελέσματα Πειραμάτων

Πειράματα που χρησιμοποίησαν τις δύο παραπάνω συλλογές/collections ήταν του Sharoff και του Σταματάτου με Κανάρη. Οι τεχνικές που χρησιμοποιήθηκαν ήταν Lemmatization, Stemming ενώ στη μελέτη του Sharoff λήφθηκαν υπόψιν και τα HTML tags των ιστοσελίδων [32]. Καλύτερα ποσοστά μας έδωσε το 7genre dataset σε σύγκριση με το KI-04. Αυτό έγινε πιθανότατα επειδή στο 7genre ήταν ισόποσα μοιρασμένες οι ιστοσελίδες για τη κάθε κατηγορία ιστοσελίδας (ίδιος αριθμός ιστοσελίδων δηλαδή) ενώ στο KI-04 η κάθε κατηγορία είχε διαφορετικό αριθμό ιστοσελίδων που εξέταζε.

Παρακάτω βλέπουμε τα αποτελέσματα που δίνουν κάποιες μελέτες πάνω στις δύο συλλογές που είδαμε και τα αποτελέσματα είναι καλύτερα πάνω στο 7genre dataset σε όλες τις μελέτες όπως φαίνεται παρακάτω. Κάποιες προσεγγίσεις υπόψιν δεν εξέτασαν και τα δύο dataset αλλά το ένα από τα δύο:

Approach	**7Genre**	**KI-04**
(Meyer zu Eissen & Stein, 2004)	-	70.0%
(Boese & Howe, 2005)	-	74.8%
(Santini, 2007)	90.6%	68.9%
(Kim & Ross, 2007)	92.7%	-
(Mason, *et al*., 2009)	94.6%	-
Proposed in this paper		
Character *n*-grams – Binary	96.2%	82.8%
Character *n*-grams – TF	92.5%	79.6%
Words – Binary	95.5%	82.0%
Words – TF	95.1%	81.8%
Textual + structural	**96.5%**	**84.1%**

Σχήμα 4.14: Αποτελέσματα για τα datasetes KI-04 και 7genre

Κεφάλαιο 5

Προ-Εκπαιδευμένα Γλωσσικά Μοντέλα

Όπως περιγράψαμε στο Κεφάλαιο 3, η είσοδος της έννοιας του Transfer Learning και του Fine Tuning στην Μηχανική Μάθηση έχει επηρεάσει σημαντικά τους τομείς της Υπολογιστικής Όρασης αλλά και της Ανάλυσης και Κατηγοριοποίησης Κειμένου. Ακολουθώντας σε θεωρητικό στάδιο την λογική λειτουργίας του εγκεφάλου, ο οποίος μεταφέρει την γνώση του ώστε να λύσει ένα νέο πρόβλημα, η Μεταφορική Μάθηση έχει ήδη αποδώσει σημαντικά θετικά αποτελέσματα στο πεδίο της Υπολογιστικής Όρασης.

Η εκμάθηση μεταφοράς, στο πλαίσιο του NLP, η ικανότητα δηλαδή κατάρτισης ενός μοντέλου σε ένα σύνολο δεδομένων και στη συνέχεια η προσαρμογή αυτού του μοντέλου ώστε να εκτελούνται διαφορετικές λειτουργίες NLP σε διαφορετικό αλλά παρόμοιο σύνολο δεδομένων και η μεγάλη άνοδος γενικά του τομέα NLP έχει συμβεί σε μεγάλο βαθμό χάρη στην ιδέα της εκμάθησης της μεταφοράς. Οι βασικοί λόγοι που είναι χρήσιμα τα προεκπαιδευμένα μοντέλα είναι ότι πρώτον ο πρωτουργός έχει ήδη καταβάλει προσπάθεια να σχεδιάσει ένα πρότυπο αναφοράς για τους υπόλοιπους, άρα και για εμάς. Αντί να κατασκευάσουμε ένα μοντέλο από την αρχή για να λύσουμε ένα παρόμοιο πρόβλημα NLP, μπορούμε να χρησιμοποιήσουμε αυτό το προρυθμισμένο και δοσμένο μοντέλο στο δικό μας σύνολο δεδομένων NLP. Θα χρειαστεί μεν λίγο Fine Tuning, αλλά σίγουρα εξοικονομούμε πολύ χρόνου και υπολογιστικούς πόρους. Η μάθηση αυτή παρέχεται μέσω των pretrained models που θα αναλυθούν παρακάτω. Στο σημείο αυτό, η έρευνα με τίλο Universal Language Model Fine-tuning for Text Classification[9] των Jeremy Howard και Sebastian Ruder, έρχεται να δώσει την εφαρμογή της Μεταφορικής Μάθησης και στο πεδίο της Κατηγοριοποίησης Κειμένου.

Στο Κεφάλαιο αυτό θα αναλύσουμε την εφαρμογή της μεθόδου, η οποία περιγράφεται στην παραπάνω έρευνα, πάνω στην Κατηγοριοποίηση Κειμένου σε δύο διαφορετικές εργασίες, τις οποίες περιγράψαμε στο παραπάνω Κεφάλαιο, την αναγνώριση συγγραφέα από κείμενα, καθώς και στην αναγνώριση είδους ιστοσελίδας, με σκοπό να αποκτήσουμε αποτελέσματα για την εφαρμογή της τεχνικής του Fine Tuning σε διάφορες εργασίες.

5.1 Universal Language Model Fine Tuning (ULMFiT)

Όπως αναφέρεται στην έρευνα [9], η μέθοδος ULMFiT είναι μια αποτελεσματική μέθοδος εκμάθησης μεταφοράς (effective Transfer Learning method), η οποία μπορεί να εφαμοσεί σε κάθε εργασία Επεξεργασίας Φυσικής Γλώσσας (NLP), παρουσιάζοντας τεχνικές οι οποίες είναι το κλειδί για την εφαρμογή Fine Tuning σε ένα Γλωσσικό Μοντέλο (Language Model)[9].

Τα αποτελέσματα των πειραμάτων έδειξαν ότι ξεπερνούν σε σημαντικό βαθμό τις επιδόσεις του state-of-the-art σε έξι εργασίες ταξινόμησης κειμένου, με την μείωση των σφαλμάτων σε ποσοστό από 18% μέχρι 24%. Πιο συγκεκριμένα, η έρευνα δείχνει ότι με 100 παραδείγματα κειμένου τα οποία είναι ετικετοποιημένα (labeled) η απόδοση αντιστοιχεί με μια εκπαίδευση ενός μοντέλου με 100 φορές περισσότερα δεδομένα.

Τέλος, η ανάπτυξη των προ-εκπαιδευμένων μοντέλων και γενικότερα της μεθόδου, έγινε ανοιχτού κώδικα. Στην παρούσα ενότητα θα αποσαφηνίσουμε κάθε έννοια που διέπει την έρευνα αυτή, καθώς και τα αποτελέσματα των πειραμάτων της.

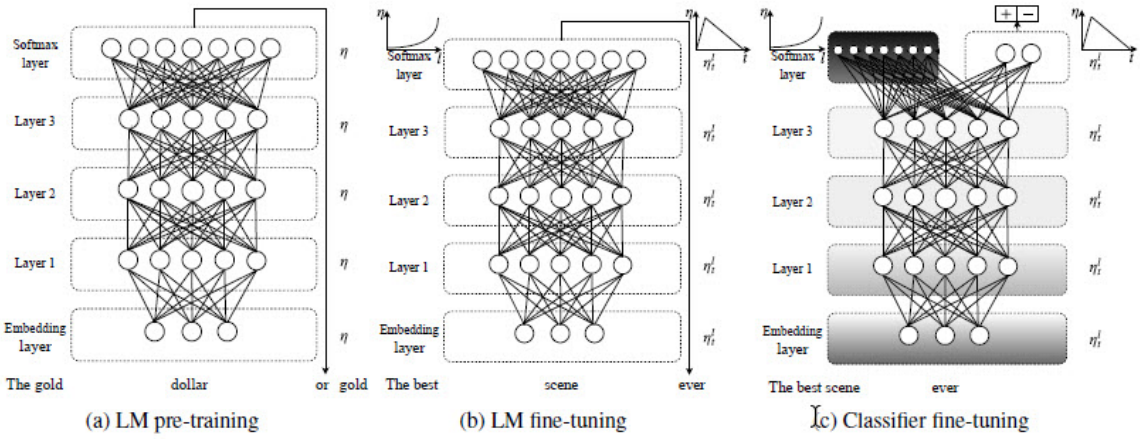
5.1.1 Περιγραφή Μεθόδου

Καθώς πολλές προσεγγίσεις εργασιών πάνω σε ΕΦΓ με την χρήση Deep Learning έχουν επιφέρει σημαντικά θετικά αποτελέσματα, η εκπαίδευση τους απαιτεί ένα μεγάλο σύνολο δεδομένων καθώς και πολλές ώρες εκπαίδευσης. Επιπρόσθετα, όλες οι προσεγγίσεις που βασίστηκαν στην εφαρμογή του Fine-Tuning συνεχίζουν να εκπαιδεύουν απο την αρχή το βασικό μοντέλο της εργασίας, κρατώντας τις παραμέτρους των προ-εκπαιδευμένων μοντέλων ως στατικές, περιορίζοντας την χρησιμότητα τους.

Η έρευνα διευθετεί αυτά του είδους τα προβλήματα, ενεργοποιώντας μια ισχυρή Επαγωγική Μεταφορική Μάθηση (inductive transfer learning) η οποία μπορεί να εφαρμοστεί σε κάθε πρόβλημα πάνω στον τομέα της ΕΦΓ, παρόμοια με την μέθοδο του Fine-Tuning πάνω στα μοντέλα του ImageNet, βασισμένα στην ίδια αρχιτεκτονική LSTM (Long Short Term Memory) με 3 επίπεδα (layers).

Συγκεκριμένα, έχει αρκετές θετικές συνεισφορές στον τομέα του NLP. Αρχικά προσφέρει το Universal Language Model Fine Tuning, το οποίο πετυχαίνει τις επιδόσεις του Transfer Learning στην υπολογιστική Όραση (CV), προβάλει τεχνικές όπως το discriminative fine-tuning, τα slanted triangular learning rates και gradual unfreezing οι οποίες έχουν ως στόχο να διατηρήσουν την προηγούμενη γνώση καθώς και να διορθώσουν το πρόβλημα της καταστροφικής αγνόησης της (Catastrophic Forgetting) και θα αναλυθούν σε αυτή την ενότητα. Επιπλέον, όσον αφορά την Κ.Κ, η οποία είναι και το βασικό θέμα της εργασίας μας, παρουσιάζει όπως αναφέραμε και παραπάνω μειωση των σφαλμάτων απο 18% μέχρι 24% μείωση σφαλμάτων σε έξι διαφορετικά σετ δεδομένων.

Η μέθοδος εφαρμογής του ULMFiT αποτελείται όπως βλέπουμε στο παρακάτω Σχήμα απο 3 στάδια. Στο πρώτο στάδιο το Γλωσσικό Μοντέλο (Language Model) εκπαιδεύεται σε ένα σύνολο δεδομένων γενικού περιεχομένου, με στόχο να αποκτήσει γενικά χαρακτηριστικά της γλώσσας στα διάφορα επίπεδα του. Στη συνέχεια, στο δεύτερο στάδιο, εφαρμόζεται η τεχνική του Fine-Tuning σε ολόκληρο το Γλωσσικό Μοντέλο χρησιμοποιώντας Discriminative Fine Tuning και Slanted Triangular Learning Rates, με σκοπό να μάθει τα συγκεκριμένα χαρακτηριστικά του συνόλου δεδομένου (dataset). Τέλος, εφαρμόζεται εξίσου Fine-Tuning και στον Ταξινομητή (Classifier) πάνω στην εργασία-στόχος χρησιμοποιώντας Gradual Unfreezing.



Σχήμα 5.1: Πίνακας επίλυσης κατηγοριοποίησης κειμένων με πολλές ετικέτες

5.1.2 Γενικού Τομέα Language Model Fine Tuning

Η γλωσσική μοντελοποίηση μπορεί να θεωρηθεί μια ιδανική πηγή και είναι αντίστοιχη του ImageNet για ΕΦΓ μιας και καλύπτει ένα μεγάλο φάσμα των πτυχών της γλώσσας, όπως είναι μακροπρόθεσμες εξαρτήσεις[19] και οι ιεραρχικές σχέσεις[20] μεταξύ των λέξεων, καθώς και συναισθήματα που προκύπτουν[21]. [9]

Όπως αναφέραμε παραπάνω, στο πρώτο στάδιο το Γλωσσικό Μοντέλο (Language Model) εκπαιδεύεται σε ένα σύνολο δεδομένων γενικού περιεχομένου. Συγκεκριμένα, ένα σύνολο δεδομένων ανάλογο του ImageNet θα πρέπει να είναι αρκετά μεγάλο και να καλύπτει ένα σύνολο από γενικές ιδιότητες της γλώσσας.

Με αυτό τον στόχο, το Γλωσσικό Μοντέλο στο πρώτο στάδιο εκπαιδεύεται στην συλλογή Wikitext-103 [30] η οποία αποτελείται απο 28.595 προ-επεξεργασμένα άρθρα της Wikipedia και από 103 εκατομμύρια λέξεις. Εκπαιδεύοντας το Γλωσσικό Μοντέλο σε αυτή την συλλογή δίνεται η δυνατότητα σε εργασίες με μικρές συλλογές κειμένων να επιτύχουν υψηλά ποσοστά ακρίβειας.

5.1.3 Language Model Fine Tuning

Παρόλης της μεγάλης ποικιλίας των δεδομένων των οποίων έχει προ-εκπαιδευτεί το Γλωσσικό Μοντέλο, το πιθανότερο είναι ότι τα δεδομένα της εργασίας της οποίας θέλουμε να εφαρμόσουμε, διαφέρουν. Για αυτό τον λόγο εφαρμόζουμε Fine-Tuning πάνω στο Γλωσσικό Μοντέλο με τα δεδομένα της εργασίας-στόχου. Η διαδικασία αυτή είναι αρκετά ταχύτερη απο την εκπαίδευση του Γλωσσικού Μοντέλου σε γενικού τομέα δεδομένα, καθώς το μόνο που χρειάζεται είναι το μοντέλο να υιοθετήσει τις ιδιαιτερότητες των δεδομένων δίνοντας την δυνατότητα της δημιουργίας ενός ισχυρού Γλωσσικού Μοντέλου ακόμα και από μικρές συλλογές δεδομένων. Παρακάτω γίνεται αποσαφήνιση των τεχνικών που χρησιμοποιούνται στην εκπαίδευση των Γλωσσικών Μοντέλων.

5.1.3.1 Discriminative Fine Tuning

Καθώς τα διαφορετικά επίπεδα του μοντέλου καταγράφουν διαφορετικούς τύπους πληροφορίας θα πρέπει να εφαρμοστεί και Fine Tuning με διαφορετικές προσεγγίσεις. Στην περίπτωση του Discriminative Fine Tuning αντί να χρησιμοποιηθεί ο ίδιος ρυθμός εκπαίδευσης σε όλα τα επίπεδα του μοντέλου, επιλέγεται να ρυθμιστεί το μοντέλο (Fine Tuning) με διαφορετικούς ρυθμούς εκπαίδευσης σε κάθε διαφορετικό επίπεδο. Για να κατανοήσουμε καλύτερα την έννοια, η κανονική περίπτωση της Στοχαστικής Κλίσης (Stochastic Gradient Descent - SGD) μεταβάλει τις παραμέτρους του μοντέλου σε χρόνο t με τον παρακάτω τρόπο:

$$\theta_t = \theta_{t-1} - \eta \cdot \nabla_\theta J(\theta)$$

όπου η ο ρυθμός αύξησης και όπου $\nabla_\theta J(\theta)$ είναι η βαθμίδα (Gradient). Στην περίπτωση του Discriminative Fine Tuning διαχωρίζουμε τις παραμέτρους θ σε $\left\{\theta^1, \ldots, \theta^L\right\}$ όπου θ^l περιέχονται οι παράμετροι του μοντέλου στο l-οστό επίπεδο και L είναι ο αριθμός των επιπέδων του μοντέλου. Με τον ίδιο ακριβώς τρόπο αποκτούμε τα $\left\{\eta^1, \ldots, \eta^L\right\}$ όπου η^l είναι ο ρυθμός αύξησης του l-οστού επιπέδου. Επομένως έχουμε την αναβαθμισμένη έκδοση της SGD με Discriminative Fine Tuning ώς εξής:

$$\theta_t^l = \theta_{t-1}^l - \eta^l \cdot \nabla_{\theta^l} J(\theta)$$

5.1.3.2 Slanted Triangular Learning Rates

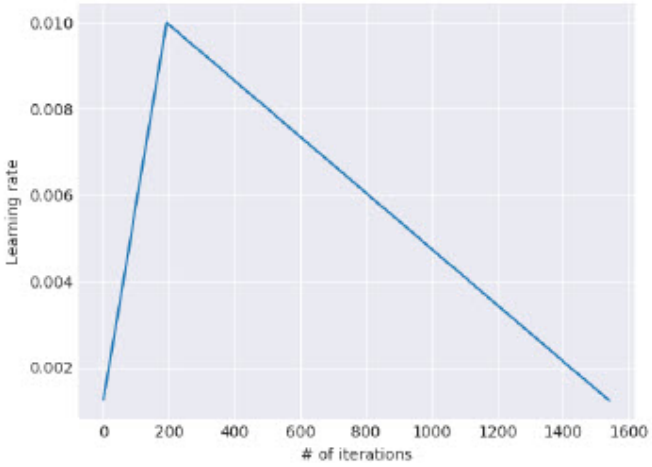
Προκειμένου να γίνει η προσαρμογή των παραμέτρων του μοντέλου στις συγκεκριμένες λειτουργίες της εργασίας που καλείται να λύσει (task), θα ήταν επιθυμητό το μοντέλο κατά την εκπαίδευση να συγκλίνει όσο το δυνατόν πιο άμεσα στην κατάλληλη περιοχή του χώρου των παραμέτρων και στη συνέχεια να τελειοποιεί την παραμετροποίηση του. Αυτός ο στόχος δεν μπορεί να επιτευχθεί θέτοντας ένα σταθερό ρυθμό εκπαίδευσης (Learning Rate) ή ένα σταθερά αυξητικό ρυθμό.

Η έννοια του Slanted Triangular Learning Rate επιλύει αυτού του είδους το πρόβλημα. Αρχικά, ο ρυθμός αυξάνεται γραμμικά και στη συνέχεια ξεκινά να μειώνεται πάλι γραμμικά, όπως παρατηρούμε στο Σχήμα. Παρακάτω βλέπουμε τον αλγόριθμο ο οποίος μας δείχνει την μαθηματική λογική που ακολουθείται στην εφαρμογή του STLR. Όπου T είναι ο αριθμός των επαναλήψεων εκπαίδευσης, όπου cut_frac το κλάσμα των επαναλήψεων το οποίο αυξάνουμε τον ρυθμό εκπαίδευσης, όπου cut είναι η επανάληψη η οποία αλλάζουμε τον ρυθμό εκπαίδευσης από αύξηση σε μείωση και όπου p είναι το κλάσμα των αριθμών των επαναλήψεων που έχουμε αυξήσει ή μειώσει τον ρυθμό εκπαίδευσης. Επιπλέον, το $ratio$ καθορίζει το πόσο μικρότερος είναι ο χαμηλότερος ρυθμός εκπαίδευσης (LR) από τον μεγαλύτερο n_{max} και τέλος το n_t είναι ο ρυθμός εκπαίδευσης στην επανάληψη t. Γενικότερα, χρησιμοποιούμε $cut\_frac = 0.1$, $ratio = 32$ και $\eta_{\max} = 0.01$. [9]

$$cut = \lfloor T \cdot cut\_frac \rfloor$$

$$p = \begin{cases} t/cut, & if t < cut \\ 1 - \frac{t-cut}{cut \cdot (ratio-1)}, & otherwise \end{cases}$$

$$\eta_t = \eta_{\max} \cdot \frac{1 + p \cdot (ratio - 1)}{ratio}$$



Σχήμα 5.2: Slanted Triangular Learning Rate

5.1.4 Classification Model Fine Tuning

Το Γλωσσικό Μοντέλο το οποίο εκπαιδεύεται αρχικά σε γενικού τομέα δεδομένα και στη συνέχεια σε μια συγκεκριμένη εργασία, χρησιμοποιείται στην εκπαίδευση του Ταξινομητή μέσω Fine-Tuning. Για την επίτευξη των καλύτερων αποτελεσμάτων επίδοσης,εφαρμόζεται η τεχνική του Gradual Unfreezing, η οποία θα περιγραφεί αναλυτικά παρακάτω. Ένας σημαντικός παράγοντας ο οποίος συμβάλει στην μείωση των σφαλμάτων είναι η χρήση του Discriminative Fine Tuning, της σταδιακής μείωσης του ρυθμού εκμάθησης δηλαδή, καθώς η εκπαίδευση προχωρά προς τα βαθύτερα επίπεδα του μοντέλου.

5.1.4.1 Gradual unfreezing

Εφαρμόζοντας Fine-Tuning σε όλα τα επίπεδα ταυτόχρονα υπάρχει μεγάλο ρίσκο καταστροφικής αγνόησης από το μοντέλο (catastrophic forgetting). επομένως η μέθοδος που προτείνεται ως βέλτιστη για να αποφύγουμε αυτό το φαινόμενο είναι να ξεπαγώνουμε σταδιακά το μοντέλο ξεκινώντας απο το τελευταίο επίπεδο καθώς περιέχει τις λιγότερες γενικές γνώσεις [22]. Αρχικά, ξεπαγώνουμε το τελευταίο επίπεδο και εφαρμόζουμε Fine-Tuning σε όλα τα παγωμένα για μια περίοδο εκπαίδευσης (epoch). Στη συνέχεια, ξεπαγώνουμε το επόμενο επίπεδο απο το τέλος και επαναλαμβάνουμε μέχρι να εφαρμόσουμε Fine-Tuning σε όλα τα επίπεδα. Παρόμοια τεχνική εφαρμόζεται και στην μέθοδο Chain-Thaw [25] με την διαφορά οτι προσθέτουμε ένα επίπεδο την φορά στο σύνολο των 'ξεπαγωμένων' επιπέδων αντί να εκπαιδεύουμε κάθε ένα επίπεδο την φορά. [9]

5.1.5 Σύνολα δεδομένων πειραμάτων

Η αξιολόγηση της μεθόδου έγινε με την εφαρμογή της πάνω σε έξι διαφορετικά σύνολα δεδομένων ευρείας μελέτης, τα οποία ποικίλουν σε αριθμό και σε μέγεθος κειμένων. Τα δεδομένα χρησιμοποιούνται σε τρία θεμελιώδη παραδείγματα για K.K: υφολογική ανάλυση, κατηγοριοποίηση ερωτήσεων και θεματική κατηγοριοποίηση. Στον παρακάτω πίνακα βλέπουμε αναλυτικά τα σύνολα δεδομένων, ως προς τον τύπο τους, τον αριθμό των ετικετών τους καθώς και τον αριθμό των δεδομένων-κειμένων.

Σύνολα Δεδομένων			
Dataset	Τύπος	Κλάσεις	Αριθμός Δεδομένων
TREC-6	Question	6	5.5k
IMDb-6	Sentiment	2	25k
Yelp-bi-6	Sentiment	2	560k
Yelp-full-6	Sentiment	5	650k
AG-6	Topic	4	120k
DBpedia-6	Topic	14	560k

Πίνακας 5.1: Σύνολα δεδομένων που χρησιμοποιήθηκαν για την αξιολόγηση του ULMFiT

5.1.6 Αξιολόγηση

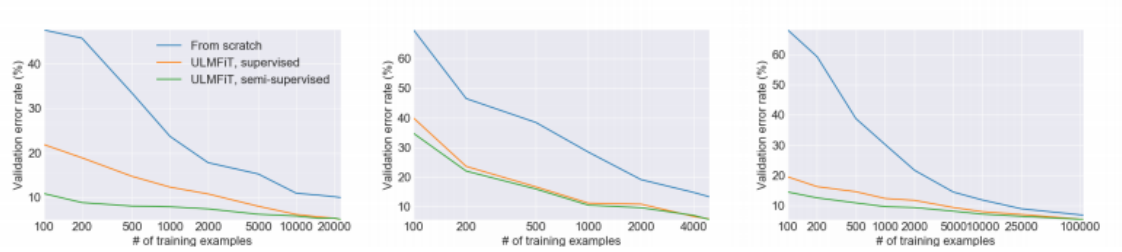
Στους παρακάτω πίνακες, παρατηρούμε τα ποσοστά σφαλμάτων σε κάθε διαφορετικό Σύνολο Δεδομένων για το ULMFiT, καθώς και για άλλες προσεγγίσεις ώστε να μπορούμε να συγκρίνουμε τα αποτελέσματα. Μπορούμε εύκολα να διαπιστώσουμε τις καλύτερες επιδόσεις του ULMFiT σε κάθε Σύνολο Δεδομένων σε σχέση με τις υπάρχουσες προσεγγίσεις.

	Model	Test		Model	Test
IMDb	CoVe (McCann et al., 2017)	8.2	TREC-6	CoVe (McCann et al., 2017)	4.2
	oh-LSTM (Johnson and Zhang, 2016)	5.9		TBCNN (Mou et al., 2015)	4.0
	Virtual (Miyato et al., 2016)	5.9		LSTM-CNN (Zhou et al., 2016)	3.9
	ULMFiT (ours)	**4.6**		ULMFiT (ours)	**3.6**

Σχήμα 5.3: Ποσοστά σφάλματος για τα Dataset IMDB & TREC-6

	AG	DBpedia	Yelp-bi	Yelp-full
Char-level CNN (Zhang et al., 2015)	9.51	1.55	4.88	37.95
CNN (Johnson and Zhang, 2016)	6.57	0.84	2.90	32.39
DPCNN (Johnson and Zhang, 2017)	6.87	0.88	2.64	30.58
ULMFiT (ours)	**5.01**	**0.80**	**2.16**	**29.98**

Σχήμα 5.4: Ποσοστά σφάλματος για 4 διαφορετικά datasets



Σχήμα 5.5: Ποσοστά σφαλμάτων κατά το Validation, Supervised, Semi-supervised, From-scratch

5.2 Άλλα Προ-Εκπαιδευμένα Γλωσσικά Μοντέλα

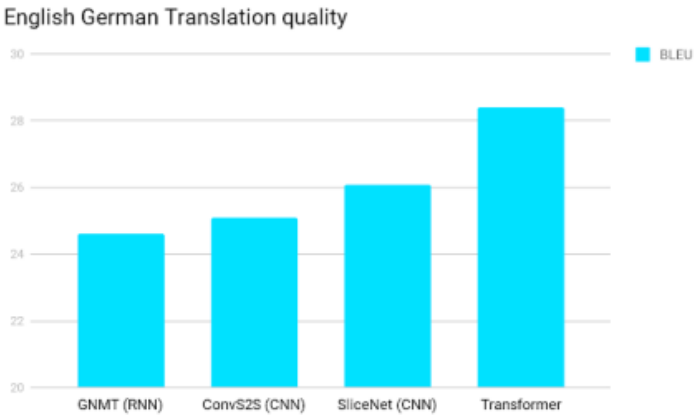
5.2.1 Transformer

Η αρχιτεκτονική του Transformer βρίσκεται στο κέντρο σχεδόν όλων των πρόσφατων σημαντικών εξελίξεων στο τομέα του NLP και επηρεάζει και άλλα γλωσσικά μοντέλα όπως θα αναφερθεί και παρακάτω. Εισήχθη πρόσφατα, το 2017 από την Google. Ως τότε, χρησιμοποιούνταν διάφορα επαναλαμβανόμενα νευρωνικά δίκτυα για γλωσσικές εργασίες, όπως η μηχανική μετάφραση και τα συστήματα απάντησης ερωτήσεων. Είχε λοιπόν καλύτερα αποτελέσματα από RNN και CNN νευρωνικά δίκτυα και επιπρόσθετα απαιτούσε λιγότερους πόρους. Τα RNN (αναδρομικά νευρωνικά δίκτυα) από τη μία είναι μία τάξη τεχνητών νευρωνικών δικτύων όπου οι συνδέσεις μεταξύ των κόμβων σχηματίζουν ένα κατευθυνόμενο γράφημα κατά μήκος μιας χρονικής αλληλουχίας και μπορούν να χρησιμοποιήσουν την εσωτερική τους κατάσταση (μνήμη) για να επεξεργαστούν ακολουθίες εισόδων.

Δεν δίνει μια συγκεκριμένη σειρά για τις λέξεις και το πως θα τοποθετηθούν, πέρα από το να δοθεί σε κάθε λέξη η ακριβής της θέση. Η Google κυκλοφόρησε πέρυσι μια βελτιωμένη έκδοση του Transformer, γνωστή ως Universal Transformer. Υπάρχει μια ακόμη νεότερη και πιο έξυπνη έκδοση, που ονομάζεται Transformer-XL.

Το Transformer βασίζεται αποκλειστικά σε self attention θεωρία (η πρώτη μεθόδους που την εφάρμοσε επιτυχώς). Η συγκεκριμένη θεωρία έχει ως στόχο να μάθουμε τη σχέση των λέξεων σε 1 φράση. Σε αυτό μπορεί να βοηθήσει και η επόμενη πρόταση ή η προηγούμενη. Για παράδειγμα αν η μία πρόταση "Οι Transformers" είναι μια ιαπωνική μπάντα. Το συγκρότημα δημιουργήθηκε το 1968". Εδώ μπορούμε να καταλάβουμε ότι η λέξη συγκρότημα της δεύτερης πρότασης αναφέρεται στους "Transformers" από τη πρώτη πρόταση. Άρα η προηγούμενη φράση βοηθάει στο να μάθουμε την αναφορά και έννοια της λέξης "συγκρότημα" της δεύτερης φράσης. Και γενικά το να εξετάζουμε έναν όρο μέσω των γειτόνων του είναι ένα μεγάλο πλεονέκτημα.

Άρα εν γένει, εξετάζεται μία πρόταση μέσω πολλών επαναλήψεων, στις οποίες επαναλήψεις εξετάζονται οι σχέσεις που έχει ο κάθε όρος με τους γείτονές του.



Σχήμα 5.6: Σύγκριση Pre Trained Models για μετάφραση από αγγλικά σε γερμανικά ΠJournal of the american medical informatics associationηγή:www.analyticsvidhya.com/blog/2019/03/pretrained-models-get-started-nlp

5.2.2 BERT

Το BERT είναι μια μέθοδος προ-κατάρτισης γλωσσικών παραστάσεων, που σημαίνει ότι εκπαιδεύουμε ένα μοντέλο γενικής χρήσης «γλωσσικής κατανόησης» σε ένα μεγάλο corpus κειμένου (όπως το Wikipedia) και στη συνέχεια χρησιμοποιούμε αυτό το μοντέλο για τα downstream NLP καθήκοντα που μας ενδιαφέρουν (όπως η απάντηση σε ερώτημα). Είναι ένα πρόσφατο μοντέλο φυσικής επεξεργασίας γλώσσας που έχει δείξει πρωτοποριακά αποτελέσματα σε πολλές εργασίες όπως η απάντηση σε ερωτήσεις, η εξαγωγή φυσικής γλώσσας και η ανίχνευση παραφράσεων. Δεδομένου ότι είναι ανοιχτό προς το κοινό, έχει γίνει δημοφιλής στην ερευνητική κοινότητα γενικά. Έχει δημιουργηθεί από τη Google. Το B.E.R.T. (Bidirectional Encoder Representations) όπως παρατηρούμε το τελευταίο γράμμα είναι το T για τη λέξη Transformer, που μας δείχνει ότι βασίζεται στον στον Transformer, ο οποίος αναλύθηκε στο προηγούμενο κεφάλαιο

Επίσης όπως προδίδει η λέξη το πρώτο γράμμα έχει να κάνει με αμφίδρομες κωδικοποιήσεις (Bidirectional). Ουσιαστικά επιθεωρεί το εξεταζόμενο πλαίσιο ή φράση και από τις δύο πλευρές του, δηλαδή και αριστερά και δεξιά. Οι προηγούμενες προσπάθειες εξέταζαν από τη μια πλευρά μιας φράσης κάθε φορά - είτε αριστερά είτε δεξιά [1]. Η αμφίδρομη κατεύθυνση βοηθά το μοντέλο να αποκτήσει μια καλύτερη κατανόηση του πλαισίου μέσα στο οποίο χρησιμοποιήθηκαν οι λέξεις. Επιπλέον, το BERT έχει σχεδιαστεί για να κάνει μάθηση πολλαπλών εργασιών, δηλαδή μπορεί να εκτελεί ταυτόχρονα διαφορετικά καθήκοντα NLP. Όταν βγήκε στην αγορά, το BERT παρήγαγε πολύ καλά αποτελέσματα σε 11 NLP tasks.

5.2.3 OpenAI's GPT-2

Η ομάδα του OpenAI δημιούργησε το μοντέλο GPT2, το οποίο εκπαιδεύτηκε στο να βρίσκει την επόμενη λέξη, δεδομένου όλων των προηγούμενων λέξεων μέσα σε κάποιο κείμενο. Και αυτό δεδομένα τύπου internet text με όγκο μέχρι και 40GB! Το dataset το οποίο εξέτασε περιείχε 8 εκατομμύρια ιστοσελίδες. Βασίζεται και αυτό στο Transformer που αναλύθηκε προηγουμένως. Το μοντέλο αυτό μπορεί μέσω λίγων φράσεων σαν input να μας δώσει ένα συμπαγές και ξεκάθαρο αποτέλεσμα. Χρειάζονται λίγες προσπάθειες για να αποκτηθεί ένα καλό δείγμα, με τον αριθμό των προσπαθειών να εξαρτάται βέβαια και από το πόσο οικείο είναι το μοντέλο με το εξεταζόμενο περιβάλλον.

Το αρχικό μοντέλο είχε 1,5 δισεκατομμύρια παραμέτρους ενώ το open source μοντέλο δειγμάτων είχε τελικά 117 εκατομμύρια. Το GPT-2 είναι μια βελτίωση της προηγούμενης έκδοσης της ομάδας OpenAI, τη GPT. Η πρώτη είχε 10 φορές λιγότερες παραμέτρους και εκπαιδευόταν σε δεδομένα των οποίων η ποσότητα ήταν 10 φορές λιγότερη από την ποσότητα των δεδομένων της GPT2 έκδοσης [5]. Το GPT-2 παράγει στην αρχή κάποια σύνθετα δείγματα κειμένου για το μοντέλο που προετοιμάζεται να εξεταστεί (με αυθαίρετη είσοδο). Το GPT2 μπορεί και προσαρμόζεται στο στυλ και το περιεχόμενο του κειμένου κάθε φορά. Αυτό επιτρέπει στον χρήστη να παράγει ρεαλιστικές και συνεπείς συνέχειες για ένα θέμα.

Το εκπαιδευμένο μοντέλο δε δίνεται στο κοινό όμως ολοκληρωμένο, διότι οι δημιουργοί του GPT2 ανησυχούν σχετικά με κακόβουλες εφαρμογές της τεχνολογίας που μπορεί να έχουν αρνητικό αντίκτυπο και στο GPT2. Όμως μπορούν να προσφέρουν σε διάφορους επίδοξους ερευνητές και μελετητές ένα μικρό δείγμα από το μοντέλο προς εξέταση μαζί με ένα documentation. Το GPT-2 ξεπερνά συχνά τα υπόλοιπα γλωσσικά μοντέλα

[1] https://www.analyticsvidhya.com/blog/2019/03/pretrained-models-get-started-nlp/

που έχουν εκπαιδευτεί σε συγκεκριμένους τομείς, όπως η Wikipedia ή πάνω σε ειδήσεις ή βιβλία. Βέβαια και στο συγκεκριμένο είδος προ-εκπαιδευμένου μοντέλου έχουν παρατηρηθεί λάθη ή αδυναμίες, όπως διάφορες λειτουργίες αστοχίας, όπως επαναλαμβανόμενο κείμενο ή αφύσικη αλλαγή του θέματος.

Γενικά το GPT-2 έχει πετύχει κορυφαίες βαθμολογίες σε διάφορες εργασίες μοντελοποίησης γλωσσών αλλά για συγκεκριμένους τομείς, ενώ αξιολογείται μόνο ως τελικό τεστ. Αλλά και σε άλλες περιπτώσεις, όπως η απάντηση σε ερωτήσεις, η κατανόηση της ανάγνωσης, η συνοπτική παρουσίαση και η μετάφραση, παίρνουμε πολύ καλά αποτελέσματα, χωρίς κάποια τελειοποίηση των μοντέλων.

Κεφάλαιο 6

Πειράματα

Εφόσον έχουμε ολοκληρώσει την διεξοδική έρευνα η οποία αφορά τα προ-εκπαιδευμένα Γλωσσικά Μοντέλα, στο Κεφάλαιο αυτό θα παρουσιαστούν αναλυτικά την εφαρμογή της μεθόδου του ULMFiT πάνω στα δύο πειράματα και θα παρουσιάσουμε συγκρίσεις με αποτελέσματα από άλλες μελέτες στα ίδια δεδομένα.

6.1 Αναγνώριση Συγγραφέα - Simple Domain

Σε αυτή την ενότητα θα αναπτύξουμε αναλυτικά την χρήστη της μεθόδου του ULMFiT πάνω στον τομέα της Αναγνώρισης Συγγραφέα σε Simple Domain πρόβλημα. Σε πρώτο στάδιο θα αναλύσουμε τα δεδομένα τα οποία θα χρησιμοποιήσουμε για την αξιολόγηση του μοντέλου, καθώς και τον τρόπο με τον οποίο θα τα επεξεργαστούμε. Στη συνέχεια, θα αναλυθούν τα στάδια της εκπαίδευσης του Γλωσσικού Μοντέλου και παρακάτω θα δούμε την εφαρμογή των πληροφοριών που έχουν εξαχθεί απο το Γλωσσικό Μοντέλο, στον Ταξινομητή (Classifier). Τέλος θα παρουσιάσουμε τα αποτελέσματα τα οποία δόθηκαν από την διεξαγωγή του πειράματος καθώς και συγκρίσεις με άλλες μεθόδους πάνω στο ίδιο Σύνολο Δεδομένων.

6.1.1 Σύνολο Δεδομένων (Dataset-Corpus)

Το dataset το οποίο θα χρησιμοποιήσουμε για την αξιολόγηση του Μοντέλου είναι το C10 το οποίο αναλύσαμε στο Κεφάλαιο 4. Το C10, όπως αναφέραμε θα μας βοηθήσει να αξιολογήσουμε το μοντέλο σε Simple Domain πρόβλημα.

Αρχικά, τα δεδομένα του dataset είναι δομημένα σε φακέλους ανά συγγραφέα, επομένως θα χρησιμοποιήσουμε την βιβλιοθήκη του fastai, ώστε να δημιουργήσουμε το dataset σε μορφή την οποία θα αναγνωρίζει το μοντέλο, μέσω της μεθόδου $from\_folder$. Η μέθοδος αυτή, επιστρέφει τα δεδομένα έτοιμα για προ-επεξεργασία, επομένως, θα χρησιμοποιήσουμε την μέθοδο του $DataBunch$ για να μας επιστρέψει επεξεργασμένα τα δεδομένα.

6.1.1.1 Προ-επεξεργασία

Η βιβλιοθήκη του fastai παρέχει ένα σύνολο μεθόδων για την αυτόματη προ-επεξεργασία των δεδομένων υλοποιώντας ένα ελάχιστο τμήμα κώδικα.

Συγκεκριμένα, με είσοδο τα δεδομένα (τα οποία δέχεται σε διάφορες μορφές), αρχικά εφαρμόζει Tokenization χωρίζοντας τα δεδομένα του κάθε κειμένου σε tokens. Στη συνέ-

χεια, τα tokenized δεδομένα αποστέλονται στην διαδικασία του Numericalization μετατρέποντας τις λέξεις σε αριθμούς, με στόχο να μπορούν να δοθούν ως είσοδος στο μοντέλο του Νευρωνικού Δικτύου. Καθώς μας ενδιαφέρει να εξάγουμε χαρακτηριστικά σχετικά με το ύφος, τον τρόπο και τις μεθόδους τις οποίες γράφει ο συγγραφέας, δεν εφαρμόζουμε τεχνικές όπως stemming και lemmatization, διότι θα αφαιρέσουν σημαντικές πληροφορίες απο τα κείμενα τις οποίες χρειαζόμαστε.

Τέλος, εφόσον τα μοντέλα τα οποία μπορούμε πλέον να δημιουργήσουμε μπορούν να είναι ισχυρότερα και πιο σύνθετα, οι έρευνες πάνω στην ΕΦΓ τείνουν να απομακρύνονται από αυτές τις τεχνικές.

6.1.2 Fine Tuning Γλωσσικού Μοντέλου (Language Model)

Μετά την προ-επεξεργασία των δεδομένων μέσω των εργαλείων της βιβλιοθήκης του FastAI, θα εισάγουμε τα επεξεργασμένα δεδομένα στο Γλωσσικό Μοντέλο. Πριν εκπαιδεύσουμε το Μοντέλο όμως, θα χρησιμοποιήσουμε την lr_find μέθοδο της βιβλιοθήκης ώστε να αναζητήσουμε τον πιο αποδοτικό ρυθμό εκμάθησης του μοντέλου, και εφόσον βρούμε το εύρος θα ξεκινήσουμε το Fine Tuning, το οποίο βλέπουμε στο παρακάτω στιγμιότυπο. Τέλος, αποθηκεύουμε την χρήσιμη πληροφορία του Γλωσσικου Μοντέλου στον encoder.

epoch	train_loss	valid_loss	accuracy	time
0	3.858167	3.383929	0.356236	00:31
1	3.265446	3.219716	0.380552	00:30
2	2.697619	3.158464	0.386593	00:30
3	2.256337	3.167700	0.388505	00:30

Σχήμα 6.1: Fine Tuning του Language Model

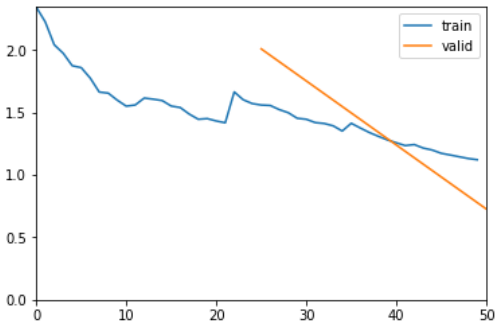
6.1.3 Fine Tuning Ταξινομητή (Classifier)

Έχοντας εκπαιδεύσει το Γλωσσικό Μοντέλο, είμαστε σε θέση να χρησιμοποιήσουμε τα δεδομένα τα οποία άντλησε το Μοντέλο, στον Ταξινομητή. Εισάγουμε επομένως τον encoder από το Γλωσσικό Μοντέλο στον Ταξινομητή, ώστε να ξεκινήσουμε το Fine Tuning. Όπως θα παρατηρήσουμε στα παρακάτω στιγμιότυπα, ακολουθούμε τις τεχνικές οι οποίες παρουσιάστηκαν στην έρευνα του ULMFiT, δηλαδή Discriminative Fine Tuning και Gradual Unfreezing.

```
learn = text_classifier_learner(data_clas, AWD_LSTM, drop_mult=0.5,
                                callback_fns=ShowGraph)
learn.load_encoder('enc')
```

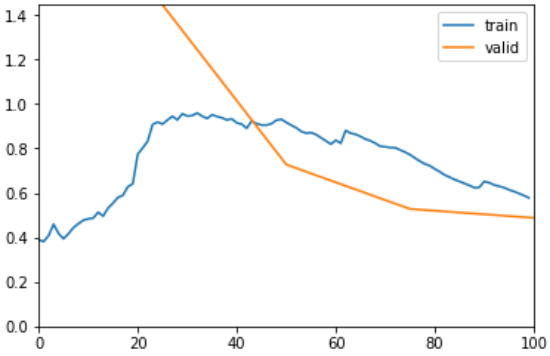
Σχήμα 6.2: Fine Tuning του Language Model

epoch	train_loss	valid_loss	accuracy	time
0	1.572525	2.010800	0.270000	00:32
1	1.122614	0.727663	0.750000	00:32



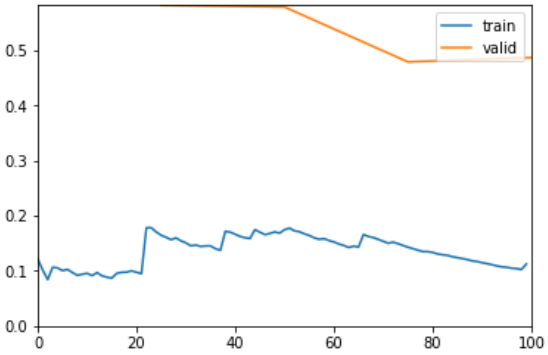
Σχήμα 6.3: Fine Tuning του Classification Model

epoch	train_loss	valid_loss	accuracy	time
0	0.918806	1.445314	0.630000	00:36
1	0.931306	0.727428	0.750000	00:36
2	0.783048	0.527657	0.790000	00:36
3	0.577876	0.487857	0.860000	00:36



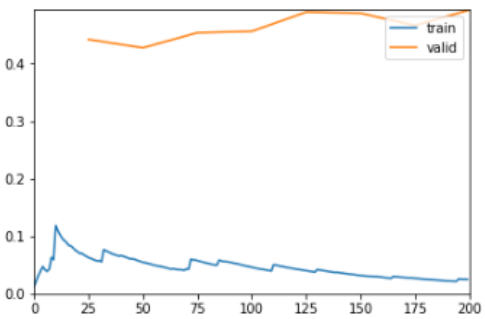
Σχήμα 6.4: Fine Tuning του Classification Model

epoch	train_loss	valid_loss	accuracy	time
0	0.170480	0.581785	0.780000	00:51
1	0.168535	0.578771	0.810000	00:51
2	0.146091	0.479222	0.830000	00:51
3	0.112315	0.487047	0.840000	00:51



Σχήμα 6.5: Fine Tuning του Classification Model

epoch	train_loss	valid_loss	accuracy	time
0	0.065000	0.441952	0.840000	01:04
1	0.055688	0.427342	0.870000	01:04
2	0.058434	0.453630	0.870000	01:04
3	0.046605	0.456133	0.850000	01:04
4	0.040239	0.489485	0.870000	01:04
5	0.031757	0.487548	0.870000	01:04
6	0.026875	0.466230	0.860000	01:04
7	0.024476	0.493291	0.850000	01:04



Σχήμα 6.6: Fine Tuning του Classification Model

6.1.4 Αποτελέσματα

Έχοντας εφαρμόσει Fine Tuning και στον Ταξινομητή, θα τον χρησιμοποιήσουμε για την αξιολόγηση του Μοντέλου, στο σύνολο δεδομένων C10, το οποίο είχε ήδη ξεχωριστό Test set 500 κειμένων απο τους 10 συγγραφείς. Στον παρακάτω πίνακα βλέπουμε τα αποτελέσματα απο τις προβλέψεις του μοντέλου, ενώ στον επόμενο παρατηρούμε τις συγκρίσεις με άλλες προσεγγίσεις οι οποίες έγιναν στο ίδιο σύνολο δεδομένων.

Συνολικά Αποτελέσματα	
Validation Accuracy	Test Accuracy
78.8	70.7

Πίνακας 6.1: Αποτελέσματα ακρίβειας για το dataset C10

Συνολικά Αποτελέσματα	
Μέθοδος	Ακρίβεια
Character n-grams – Baseline	80.6%
Character n-grams – DV-MA	78.2%
Character n-grams – DV-SA	77.4%
Token n-grams – Baseline	80.0%
Token n-grams – DV-MA	79.2%
Token n-grams – DV-SA	79.4%
Fine Tuning Pretrained Models (ULMFiT)	**70.7**%

Πίνακας 6.2: Συγκρίσεις επιδόσεων για το δύο σύνολο δεδομένων KI-04

6.2 Αναγνώριση Συγγραφέα - Cross Domain

Εφόσον αναπτύξαμε το πείραμα της Αναγνώρισης Συγγραφέα πάνω σε Simple Domain πρόβλημα, σε αυτή την ενότητα θα δούμε τα αποτελέσματα του πειράματος πάνω σε Cross Domain δεδομένα. Αρχικά, θα παρουσιάσουμε την προετοιμασία που έγινε στα δεδομένα μας και την προ-επεξεργασία ώστε να μπορούν να χρησιμοποιηθούν από το Γλωσσικό Μοντέλο και στην συνέχεια από τον Ταξινομητή (Classifier). Έπειτα, θα περιγράψουμε αναλυτικά τον τρόπο με τον οποίο εκπαιδεύουμε το Γλωσσικό Μοντέλο, τεχνικές που εφαρμόζουμε για καλύτερα αποτελέσματα, ενώ στη συνέχεια θα περιγράψουμε και την αντίστοιχη διαδικασία στην εκπαίδευση του Ταξινομητή.

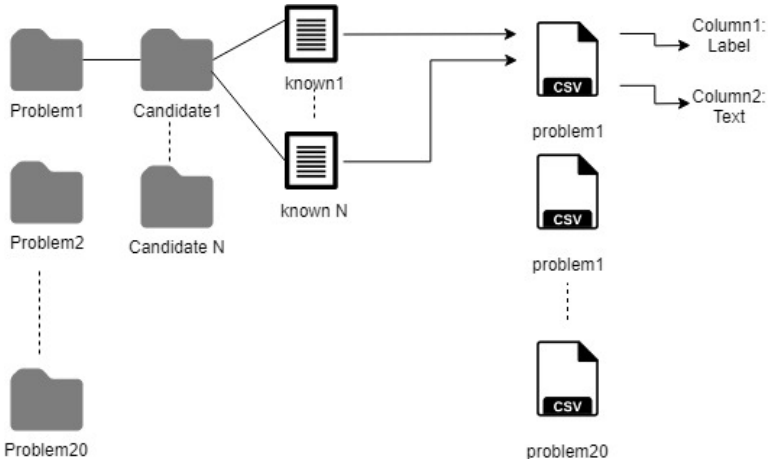
Τέλος, θα παρουσιάσουμε τα αποτελέσματα τα οποία προέκυψαν απο την εφαρμογή της μεθόδου στο πείραμα, καθώς και συγκρίσεις με άλλες προσεγγίσεις.

6.2.1 Σύνολο Δεδομένων (Dataset-Corpus)

Όπως αναπτύξαμε στο Κεφάλαιο 4, η έρευνα της παρούσας εργασίας πάνω στον τομέα της Αναγνώρισης συγγραφέα εστιάζεται στο Fanfiction. Επομένως, το σύνολο δεδομένων (dataset) το οποίο θα χρησιμοποιήσουμε για να εφαρμόσουμε την μέθοδο ULMFiT είναι αυτό το οποίο περιγράψαμε στην Ενότητα 4.1.6, το Fanfiction Dataset το οποίο δόθηκε για τον διαγωνισμό του PAN18.

6.2.1.1 Προετοιμασία

Πριν φτάσουμε στο στάδιο της προ-επεξεργασίας των δεδομένων μετατρέψαμε τα δεδομένα τα οποία ήταν δομημένα σε φακέλους ανά πρόβλημα και με την σειρά τους ανά συγγραφέα περιέχοντας κείμενα σε μορφή txt, σε αρχεία csv ανά πρόβλημα δίνοντας ετικέτα σε κάθε κείμενο ανάλογα με τον συγγραφέα τους. Στόχος αυτής της μετατροπής των δεδομένων ήταν η πιο εύκολη κατανόηση και επεξεργασία τους, σε αντίθεση με την δομή των φακέλων, παρόλο του οτι η βιβλιοθήκη του fastai παρέχει την δυνατότητα της εισόδου των δεδομένων σε φακέλους. Ο κώδικας ο οποίος μετατρέψαμε τα δεδομένα είναι ανοιχτός και μπορεί να βρεθεί στην Ενότητα του Παραρτήματος.



Σχήμα 6.7: Προετοιμασία δεδομένων

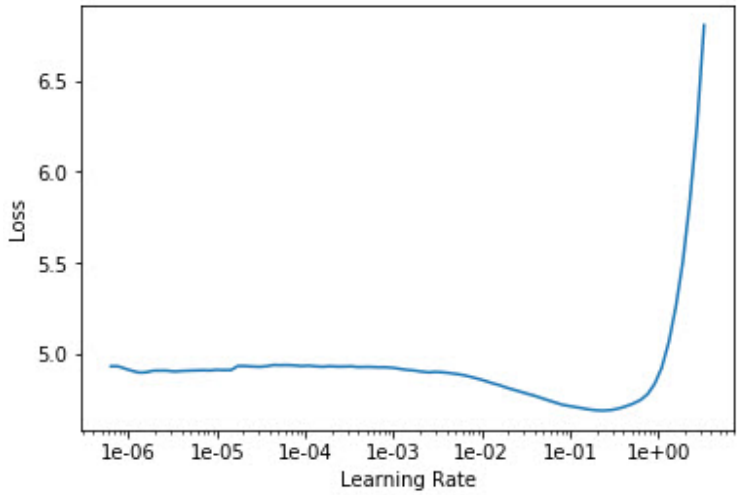
6.2.1.2 Προ-επεξεργασία

Όπως και στη προηγούμενη ενότητα, τα δεδομένα τα οποία εισάγαμε στο μοντέλο πέρασαν απο προ-επεξεργασία, με την χρήση μεθόδων απο την βιβλιοθήκη του fastai και συγκεκριμένα μέσω της μεθόδου $DataBunch$.

6.2.2 Fine-Tuning Γλωσσικού Μοντέλου (Language Model)

Τα δεδομένα τα οποία υπέστησαν προ-επεξεργασία, θα χρησιμοποιηθούν στο Fine Tuning του προ-εκπαιδευμένου Γλωσσικού Μοντέλου (Pretrained Language Model), το οποίο παρέχεται από την βιβλιοθήκη του fastai με την αρχιτεκτονική AWD-LSTM Νευρωνικού Δικτύου.

Αρχικά, για την αποδοτικότερη εφαρμογή του Fine-Tuning πάνω στο Μοντέλο, χρησιμοποιούμε την μέθοδο εύρεσης του καταλληλότερου ρυθμού εκπαίδευσης (Learning Rate) απο την βιβλιοθήκη του fastai. Στο παρακάτω διάγραμμα μπορούμε να παρατηρήσουμε τα αποτελέσματα τα οποία έγιναν εξαγωγή απο την μέθοδο lr_find της βιβλιοθήκης.



Σχήμα 6.8: Εύρεση καταλληλότερου ρυθμού εκπαίδευσης για το Γλωσσικό Μοντέλο

Εφόσον η μέθοδος μας έδωσε τις κατάλληλες πληροφορίες για τον ρυθμό εκπαίδευσης, θα τις χρησιμοποιήσουμε ανάλογα στην εκπαίδευση του Γλωσσικού Μοντέλου. Παρατηρώντας το γράφημα διαπιστώνουμε ότι από τις τιμές $1e-02$ μέχρι $1e-1$ ρυθμού εκπαίδευσης, αποδίδει καλύτερα το μοντέλο, επομένως θα χρησιμοποιήσουμε αυτό το εύρος για να εφαρμόσουμε Fine-Tuning.

Ξεκινώντας, εκπαιδεύσαμε το προ-εκπαιδευμένο Γλωσσικό Μοντέλο για 2 epochs, ενώ στην συνέχεια ξεπαγώσαμε το μοντέλο και συνεχίσαμε το Fine-Tuning για άλλα 4 epochs. Στα παρακάτω Σχήματα βλέπουμε την εφαρμογή του Fine-Tuning στο Γλωσσικό Μοντέλο για το πρώτο πρόβλημα του Fanfiction Dataset το οποίο είναι σε Αγγλική Γλώσσα.

```
learn.fit_one_cycle(2,slice(1e-3,1e-2),moms=moms)
```

epoch	train_loss	valid_loss	accuracy	time
0	4.656541	3.878199	0.261682	00:10
1	4.281924	3.790302	0.263914	00:10

Σχήμα 6.9: 1ο Στάδιο Fine Tuning του Language Model

```
learn.unfreeze()
learn.fit_one_cycle(4, slice(1e-3,1e-2), moms=moms)
```

epoch	train_loss	valid_loss	accuracy	time
0	3.901910	3.697851	0.263690	00:13
1	3.748298	3.624519	0.278869	00:13
2	3.471506	3.606552	0.278646	00:13
3	3.197897	3.627117	0.281324	00:13

Σχήμα 6.10: 2ο Στάδιο Fine Tuning του Language Model

Το τμήμα του Γλωσσικού Μοντέλου το οποίο περιέχει την πληροφορία της κατανόησης μιας πρότασης καλείται encoder, οπότε αποθηκεύουμε την κωδικοποίηση του Γλωσσικού Μοντέλου (encoder) ώστε να τη χρησιμοποιήσουμε παρακάτω για Ταξινόμηση (Classification).

```
learn.save_encoder('enc')
```

Σχήμα 6.11: Αποθήκευση encoder Γλωσσικού Μοντέλου

6.2.3 Εκπαίδευση Μοντέλου Ταξινόμησης (Classification Model)

Εφόσον εφαρμόσαμε Fine-Tuning στο Γλωσσικό Μοντέλο, θα το χρησιμοποιήσουμε για Fine-Tuning στο Μοντέλο Ταξινομητή (Classifier) μέσω του encoder που αποθηκεύσαμε.

```
learn = text_classifier_learner(data_clas, AWD_LSTM, drop_mult=0.6,
                                callback_fns=ShowGraph)
learn.load_encoder('enc')
```

Σχήμα 6.12: Αρχικοποίηση του Ταξινομητή

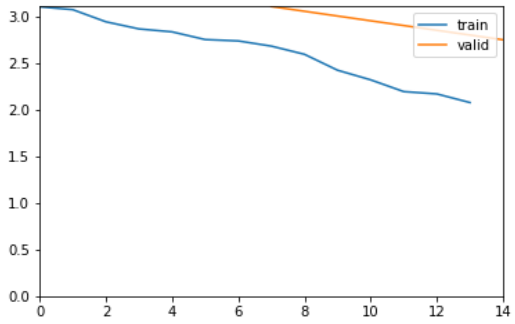
Στη συνέχεια, ξεκινάμε την διαδικασία του Fine Tuning εκπαιδεύοντας για 2 epochs τον Ταξονομητή, ενώ παρακάτω θα εφαρμόσουμε την διαδικασία του Gradual Unfreezing και Discriminative Fine Tuning τις οποίες περιγράψαμε παραπάνω, ξεπαγώνοντας σταδιακά το μοντέλο, ξεκινώντας από το τελευταίο επίπεδο και ταυτόχρονα χρησιμοποιώντας διαφορετικούς ρυθμούς εκμάθησης. Συγκεκριμένα, όπως βλέπουμε στα παρακάτω Σχήματα, ξεπαγώνουμε τα τελευταία δύο επίπεδα (layers) και εφαρμόζουμε Fine-Tuning. Έπειτα, ξεπαγώνουμε τα 3 τελευταία επίπεδα και συνεχίζουμε το Fine-Tuning. Τέλος, ξεπαγώνουμε όλο το μοντέλο και εφαρμόζουμε την ίδια λογική. Σε κάθε διαφορετικό 'ξεπάγωμα' μειώνουμε τον ρυθμό εκμάθησης εφόσον πλησιάζουμε στα βαθύτερα επίπεδα.

Τα αποτελέσματα των προβλέψεων των οποίων προκείπτουν για κάθε πρόβλημα του συνόλου δεδομένων, παρουσιάζονται στο επόμενο κεφάλαιο αναλυτικά, καθώς και συγκρίσεις με άλλες έρευνες πάνω στα ίδια δεδομένα.

```
lr = 1e-1
```

```
learn.fit_one_cycle(2,lr,moms=moms)
```

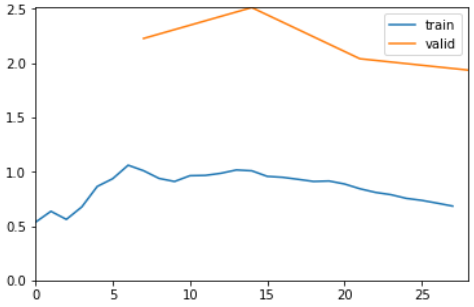
epoch	train_loss	valid_loss	accuracy	time
0	2.736776	3.105718	0.117647	00:13
1	2.075985	2.748805	0.058824	00:12



Σχήμα 6.13: 1ο Στάδιο Fine Tuning του Language Model

```
learn.freeze_to(-2)
lr /= 2
learn.fit_one_cycle(4, slice(lr/(2.6**4),lr), moms=moms, wd=0.1)
```

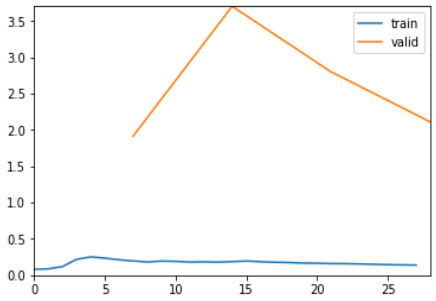
epoch	train_loss	valid_loss	accuracy	time
0	1.060409	2.227448	0.294118	00:15
1	1.016433	2.511911	0.294118	00:15
2	0.887522	2.040399	0.470588	00:15
3	0.683879	1.936123	0.470588	00:15



Σχήμα 6.14: 2ο Στάδιο Fine Tuning του Language Model

```
learn.freeze_to(-3)
lr /= 2
learn.fit_one_cycle(4, slice(lr/(2.6**4),lr), moms=(0.8,0.7), wd=0.1)
```

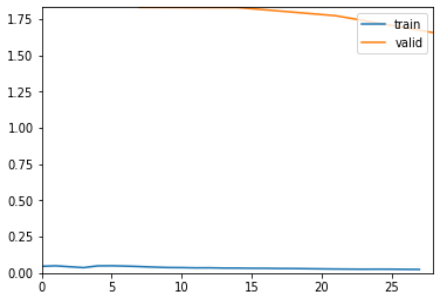
epoch	train_loss	valid_loss	accuracy	time
0	0.212993	1.915191	0.470588	00:21
1	0.180711	3.703254	0.235294	00:21
2	0.164889	2.799371	0.411765	00:21
3	0.139816	2.111270	0.529412	00:21



Σχήμα 6.15: 3ο Στάδιο Fine Tuning του Language Model

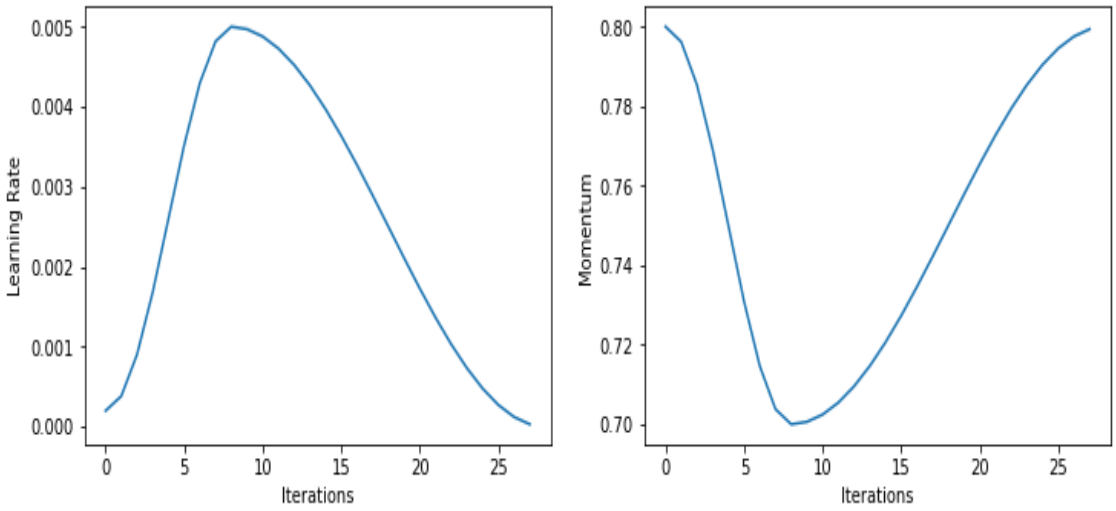
```
learn.unfreeze()
lr /= 5
learn.fit_one_cycle(8, slice(lr/(2.6**4),lr),
                    moms=moms, wd=0.1)
```

epoch	train_loss	valid_loss	accuracy	time
0	0.047569	1.831402	0.470588	00:26
1	0.033970	1.830043	0.411765	00:26
2	0.028355	1.773709	0.470588	00:26
3	0.024589	1.656777	0.470588	00:26



Σχήμα 6.16: 4ο Στάδιο Fine Tuning του Language Model

Παρακάτω παρατηρούμε στο αριστερό Σχήμα ότι ακολουθείτε η τεχνική Slanted Triangular Learning Rates που περιγράφεται στην έρευνα του ULMFiT. Ο ρυθμός εκπαίδευσης ξεκινάει να αυξάνεται γραμμικά και στη συνέχεια μειώνεται γραμμικά. Στο δεξιά Σχήμα παρατηρούμε τα στάδια των τιμών momentum τα οποία διαμορφώνονται αντιστρόφως ανάλογα του ρυθμού εκμάθησης.



Σχήμα 6.17: Άποψη του ρυθμού εκπαίδευσης και του momentum

6.2.4 Αποτελέσματα

Το Μοντέλο του Ταξινομητή αντιμετώπιζε σε κάθε Πρόβλημα του συνόλου δεδομένων, πρόβλημα overfitting όπως παρατηρείται στον παρακάτω πίνακα στα πρώτα τέσσερα προβλήματα, τα οποία αφορούσαν την Αγγλική Γλώσσα. Στην περίπτωση των άλλων 15 Προβλημάτων το μοντέλο Ταξινόμησης αδυνατούσε να ανταποκριθεί.

Αποτελέσματα					
Πρόβλημα	Γλώσσα	Επιτυχημένα Κείμενα	Σύνολο Κειμένων	Ακρίβεια	Αριθμός Συγγραφέων
1	Αγγλικά	9	79	0.113	20
2	Αγγλικά	12	74	0.162	15
3	Αγγλικά	97	40	0.175	10
4	Αγγλικά	5	16	0.312	5

Πίνακας 6.3: Αποτελέσματα προσπαθειών ανά πρόβλημα

6.2.5 Συγκρίσεις με άλλες μεθόδους του διαγωνισμού PAN18

Αποτελέσματα							
Ομάδα	Ταξινομητής	Συνολική	Αγγλικά	Γαλλικά	Ιταλικά	Πολωνικά	Ισπανικά
Custódio and Paraboni	ensemble	0.685	0.744	0.668	0.676	0.482	0.856
Murauer et al.	SVM	0.643	0.762	0.607	0.663	0.450	0.734
Halvani and Graner	similarity	0.629	0.679	0.536	0.752	0.426	0.751
Yigal et al.	SVM	0.598	0.672	0.609	0.642	0.431	0.636
Martín dCR et al.	SVM	0.588	0.601	0.510	0.571	0.556	0.705
PAN18-BASELINE	SVM	0.584	0.697	0.585	0.605	0.419	0.615
Miller et al.	SVM	0.582	0.573	0.611	0.670	0.421	0.637
Schaetti	ESN	0.387	0.538	0.332	0.337	0.388	0.343
Gagala	NN	0.267	0.376	0.215	0.248	0.216	0.280
López-Anguita et al.	SVM	0.139	0.190	0.065	0.161	0.128	0.153
Εμείς	**ULMFiT**	**-**	**0.376**	**-**	**-**	**-**	**-**

Πίνακας 6.4: Σύγκριση των αποτελεσμάτων χρήσης ULMFiT με άλλες μεθόδους

6.3 Αναγνώριση Είδους Ιστοσελίδας (Webpage Genre Recognition)

Όπως και στο πείραμα της Αναγνώρισης Συγγραφέα, έτσι και στο πείραμα Αναγνώριση είδους ιστοσελίδας θα ακολουθήσουμε την ίδια συγκεκριμένη διαδικασία για να εξάγουμε αποτελέσματα πρόβλεψης.

Πιο αναλυτικά, θα περιγράψουμε την προετοιμασία του συνόλου δεδομένων, την προεπεξεργασία που ακολουθείται ώστε να είναι κατάλληλα για είσοδο στο Γλωσσικό Μοντέλο, ενώ στη συνέχεια θα παρουσιάσουμε τον τρόπο με τον οποίο εκπαιδεύτηκαν τα Μοντέλα, καθώς και την εφαρμογή του 10 Fold Cross Validation ώστε να γίνουν συγκρίσιμα τα αποτελέσματα με άλλες μελέτες πάνω στα ίδια δεδομένα.

Τέλος, θα αναλυθούν τα αποτελέσματα των πειραμάτων στα δύο σύνολα δεδομένων (7Genre και KI-04) σε πίνακες για κάθε Fold, καθώς και συγκρίσεις με άλλες μεθόδους-προσεγγίσεις πάνω στα ίδια δεδομένα.

6.3.1 Σύνολο Δεδομένων (Dataset-Corpus)

Στην περίπτωση της Αναγνώρισης Ιστοσελίδας, τα πειράματα τα οποία διεξαγάγαμε αφορούσαν δύο σύνολα δεδομένων, τα οποία περιγράφηκαν στο προηγούμενο Κεφάλαιο στην ομώνυμη Ενότητα. Το ένα σύνολο δεδομένων είναι το 7Genre (SANTINIS) ενώ το δεύτερο είναι το KI-04. Αναλυτικές πληροφορίες με το περιεχόμενο και των δύο datasets μπορούν να βρεθούν στο προηγούμενο Κεφάλαιο.

6.3.1.1 Προετοιμασία

Όπως και στο προηγούμενο πείραμα, μετατρέψαμε τα δεδομένα τα οποία ήταν δομημένα σε κείμενα ανά φακέλους στο 7Genre Dataset με τις ετικέτες να φέρονται από τα ονόματα των φακέλων και σε κείμενα με ετικέτες τα ονόματα των αρχείων, σε δόμηση μορφής csv, με σκοπό την ευκολότερη διαχείριση τους.

6.3.1.2 Προ-επεξεργασία

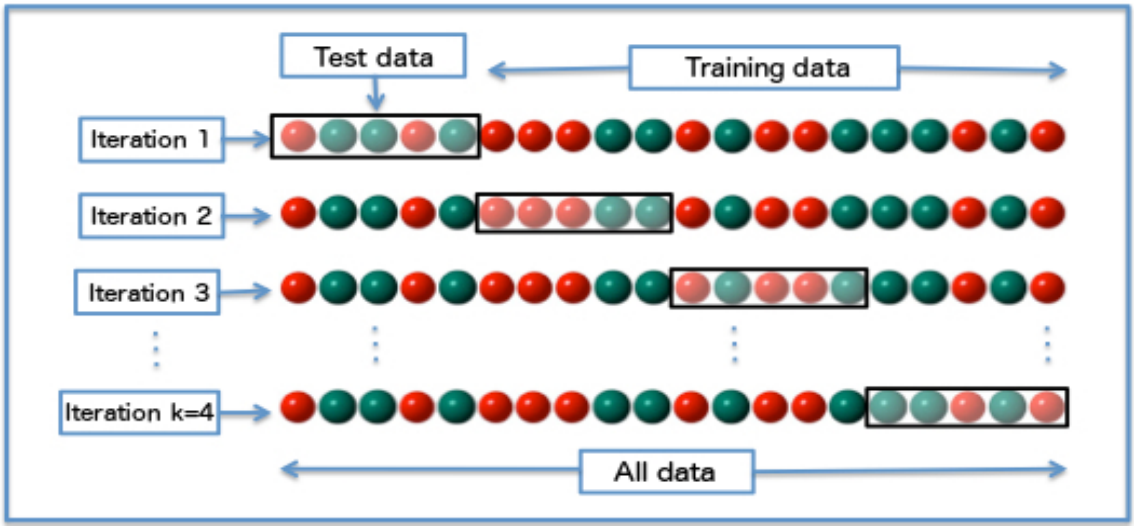
Η προ-επεξεργασία η οποία εφαρμόστηκε στο πείραμα ήταν ακριβώς η ίδια, περνώντας αρχικά από την μέθοδο του Tokenization, ώστε να μετατραπούν τα κείμενα σε Tokens, ενώ στην συνέχεια πέρασαν από την μέθοδο του Numericalization ωστε να μετατραπούν σε αριθμούς, με σκοπό την είσοδο τους στο Γλωσσικό Μοντέλο και στη συνέχεια στον Ταξινομητή.

6.3.2 10 Fold Cross Validation

Λόγω του ότι οι έρευνες πάνω σε αυτά τα δύο σύνολα δεδομένων εφαρμόζουν την τεχνική 10-Fold Cross Validation, θα υλοποιήσουμε ακριβώς την ίδια τεχνική ώστε να μπορούμε να συγκρίνουμε κατάλληλα τα δεδομένα.

Γενικότερα, η μέθοδος K-Fold Cross Validation ή αλλιώς: Διασταυρωμένη επικύρωση σε Κ-μέρη, είναι μια στατιστική τεχνική κατά την οποία το πείραμα διεξάγεται Κ φορές, κάθε φορά χρησιμοποιώντας διαφορετικό τμήμα του συνόλου δεδομένων για training και test, βγάζοντας ως αποτέλεσμα τον μέσο όρο τον προβλέψεων απο κάθε επανάληψη. Με αυτό τον τρόπο επικυρώνεται κατά πόσο τα αποτελέσματα γενικεύονται σε ανεξάρτητα σύνολα δεδομένων. Καθώς στο Κεφάλαιο αυτό θα δείξουμε την εφαρμογή του πειράματος και την τεχνική του K-Fold Cross Validation για 10 επαναλήψεις, στο επόμενο Κεφάλαιο

θα παραουσιαστούν αναλυτικά τα αποτελέσματα σε κάθε επανάληψη αλλά και το συνολικό αποτέλεσμα των προβλέψεων.



Σχήμα 6.18: Η εφαρμογή του K-Fold Cross Validation σε ένα σύνολο δεδομένων
Πηγή: https://en.wikipedia.org/wiki/Cross-validation$_{(statistics)}$

Για την υλοποίηση του K-Fold Validation, η βιβλιοθήκη SKLearn παρέχει μεθόδους κατάλληλα υλοποιημένες για τον ορθό διαχωρισμό του συνόλου δεδομένων. Με τον παρακάτω τρόπο ορίζουμε των αριθμό των Folds των οποίων θα χωρίσουμε τα σύνολα δεδομένων.

```
1 from sklearn.model_selection import KFold
```

```
1 kf = KFold(n_splits=10)
```

Σχήμα 6.19: Αρχικοποίηση των 10 Folds

Για τα δεδομένα χρησιμοποιούμε DataFrame από την βιβλιοθήκη Pandas, μιας και οι βιβλιοθήκες του fastai παρέχουν μεθόδους που το υποστηρίζουν. Στη συνέχεια λοιπόν, χωρίζεται το DataFrame του dataset στα 10 διαφορετικά Folds τα οποία ορίσαμε. Εφόσον

```
folds = kf.split(df)
```

Σχήμα 6.20: Διαχωρισμός συνόλου δεδομένου στα 10 Μέρη

χωρίσαμε επιτυχώς τα δεδομένα, εφαρμόζουμε το πείραμα για κάθε Fold μέσω επαναληπτικής διαδικασίας. Την διαδικασία η οποία ακολουθείτε μέσα στις επαναλήψεις, η οποία βασίζεται στις μεθόδους του ULMFiT, θα την αναλύσουμε διεξοδικά σε κάθε παρακάτω υπό-ενότητα.

```
for train, test in folds:
```

Σχήμα 6.21: Επαναληπτική διαδικασία των Folds

Σε κάθε μέρος (Fold) υπάρχουν ένα Train τμήμα και ένα Test. Επομένως, για κάθε επανάληψη θα ορίζουμε κατάλληλα τα δεδομένα του Γλωσσικού Μοντέλου και του Ταξινομητή. Το τμήμα του train dataset θα χωριστεί επίσης σε 90% για εκπαίδευση (train) και 10% για επικύρωση (validation) του μοντέλου, επομένως θα έχουμε τα 3 διαφορετικά τμήματα δεδομένων όπως παρατηρούμε στο παρακάτω Σχήμα.

```
for train, test in folds:
  print(pos)
  pos += 1
  msk =  np.random.rand(len(train)) < 0.9
  train_df = df.iloc[msk]
  valid_df = df.iloc[~msk]
  test_df = df.iloc[test]
  data_lm = TextLMDataBunch.from_df(path,train_df=train_df,valid_df=valid_df)
  data_clas = TextClasDataBunch.from_df(path,train_df=train_df,valid_df=valid_df
                                        , vocab=data_lm.train_ds.vocab, bs=32)
```

Σχήμα 6.22: Διαχωρισμός train,test,validation

6.3.3 Εκπαίδευση Γλωσσικού Μοντέλου (Language Model)

Εφόσον ορίσαμε κατάλληλα τα δεδομένα του Γλωσσικού Μοντέλου αλλά και του Ταξινομητή, θα αρχικοποιήσουμε το Γλωσσικό Μοντέλο, ώστε να εφαρμόσουμε Fine-Tuning όπως και στο προηγούμενο πείραμα. Ξεκινώντας θα εκπαιδευτεί για 2 epoch, ενώ στη συνέχεια θα ξεπαγώσουμε το Μοντέλο ώστε να το εκπαιδεύσουμε συνολικά. Στο παρακάτω Σχήμα βλέπουμε τα ποσοστά σφαλμάτων κατά την εκπαίδευση του Γλωσσικού Μοντέλου, καθώς και την ακρίβεια του, για το πρώτο μέρος (Fold).

epoch	train_loss	valid_loss	accuracy	time
0	3.812201	3.519989	0.381519	04:03
1	3.339729	3.362379	0.400820	04:03

Σχήμα 6.23: Εκπαίδευση Γλωσσικού Μοντέλου

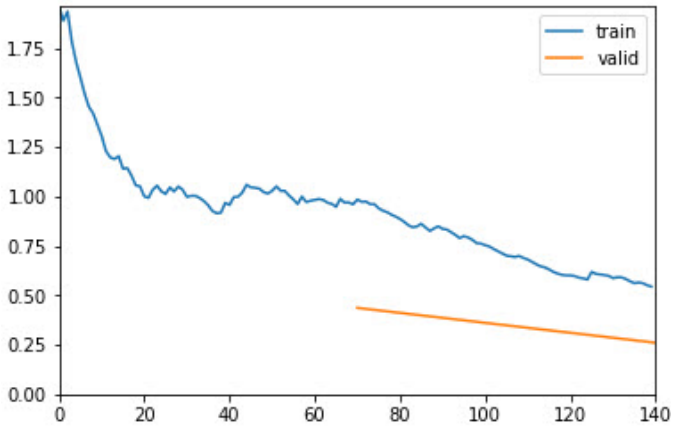
epoch	train_loss	valid_loss	accuracy	time
0	3.013442	3.285369	0.419108	04:48
1	2.843863	3.177294	0.437123	04:48
2	2.665537	3.112240	0.446736	04:48
3	2.584531	3.124093	0.447968	04:48

Σχήμα 6.24: Εκπαίδευση Γλωσσικού Μοντέλου

6.3.4 Εκπαίδευση μοντέλου ταξινόμησης (Classification Model)

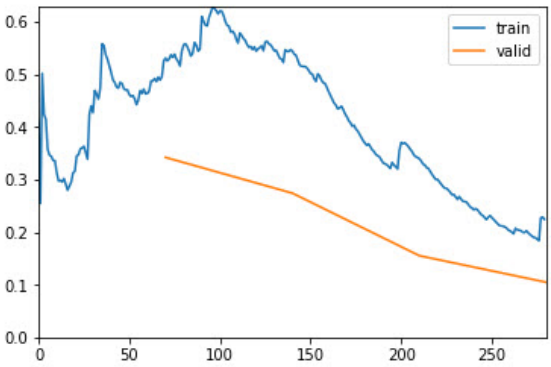
Χρησιμοποιώντας όλη την απαραίτητη πληροφορία η οποία αποκτήθηκε από το Γλωσσικό Μοντέλο στα δεδομένα, θα εκπαιδεύσουμε τον Ταξινομητή (Classifier) με τον ίδιο ακριβώς τρόπο όπου δράσαμε στο προηγούμενο πείραμα. Συγκεκριμένα θα εφαρμόσουμε Discriminative Fine Tuning και Gradual Unfreezing, ξεπαγώνοντας και εκπαιδεύοντας σταδιακά το Μοντέλο Ταξινομητή και χρησιμοποιώντας διαφορετικούς ρυθμούς εκμάθησης, χαμηλότερους όσο προχωράμε προς τα βαθύτερα επίπεδα (layers) του μοντέλου, όπως θα δούμε στα παρακάτω Σχήματα.

epoch	train_loss	valid_loss	accuracy	time
0	0.959708	0.436407	0.857143	05:33
1	0.544086	0.260142	0.912698	06:24



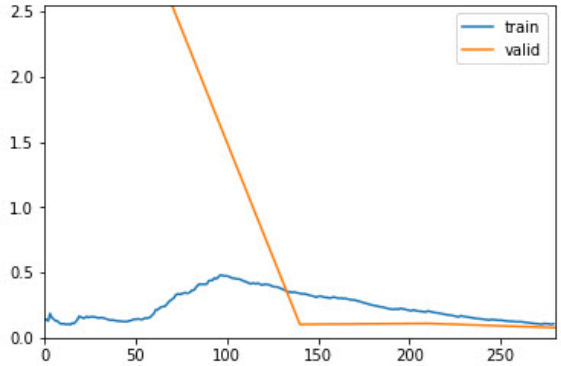
Σχήμα 6.25: Αναγνώριση είδους Ιστοσελίδας - 1ο Στάδιο Εκπαίδευσης Γλωσσικού Μοντέλου

epoch	train_loss	valid_loss	accuracy	time
0	0.525311	0.342006	0.857143	05:53
1	0.546807	0.273997	0.912698	06:33
2	0.341778	0.155170	0.960317	06:14
3	0.224251	0.104620	0.968254	06:31



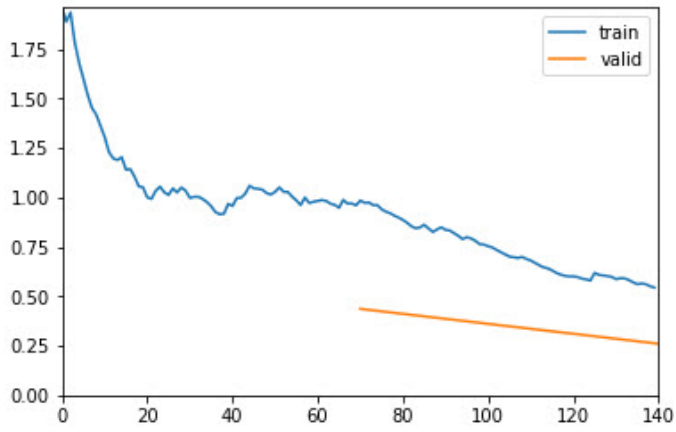
Σχήμα 6.26: Αναγνώριση είδους Ιστοσελίδας - Εκπαίδευση Γλωσσικού Μοντέλου

epoch	train_loss	valid_loss	accuracy	time
0	0.284292	2.544368	0.579365	08:29
1	0.343033	0.101743	0.960317	08:31
2	0.196846	0.109016	0.968254	07:24
3	0.106058	0.076994	0.976190	07:34

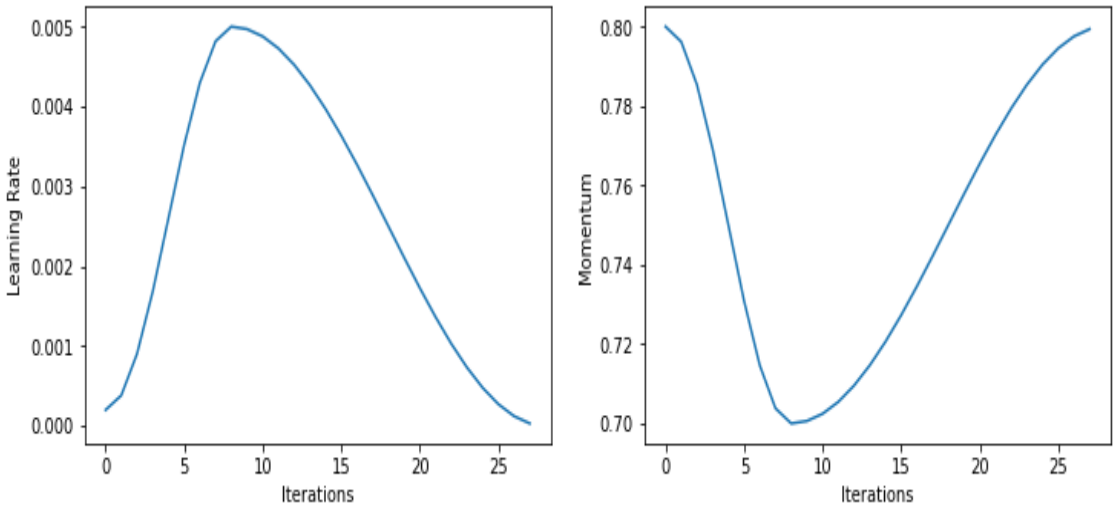


Σχήμα 6.27: Αναγνώριση είδους Ιστοσελίδας - Εκπαίδευση Γλωσσικού Μοντέλου

epoch	train_loss	valid_loss	accuracy	time
0	0.959708	0.436407	0.857143	05:33
1	0.544086	0.260142	0.912698	06:24



Σχήμα 6.28: Αναγνώριση είδους Ιστοσελίδας - Εκπαίδευση Γλωσσικού Μοντέλου



Σχήμα 6.29: Άποψη του ρυθμού εκπαίδευσης και του momentum

6.3.5 Αποτελέσματα - 7Genre-SANTINIS

Παρακάτω παρουσιάζεται σε αναλυτικό πίνακα η επίδοση του μοντέλου σε Validation και σε Test, για το σύνολο δεδομένων 7Genre-SANTINIS, ανά μέρος (Fold), καθώς και η συνολική του επίδοση (Μέσος Όρος) στον επόμενο πίνακα.

Συνολικά Αποτελέσματα	
Validation Accuracy	Test Accuracy
95.2	99.1

Πίνακας 6.5: Συνολικά Αποτελέσματα ακρίβειας για το dataset 7Genre-Santinis

Χρησιμοποιώντας έναν μικρό αριθμό κύκλων εκπαίδευσης των μοντέλων, μπορούμε να παρατηρήσουμε ότι με μικρό υπολογιστικό κόστος μπορούμε να έχουμε ένα πολύ ικανοποιητικό αποτέλεσμα πάνω στο σύνολο των δεδομένων (dataset) 7Genre, χρησιμοποιώντας την τεχνική Fine-Tuning πάνω σε προ-εκπαιδευμένα Μοντέλα. Στην επόμενη ενότητα, θα δώσουμε αναλυτικές συγκρίσεις σε πίνακες, με άλλες έρευνες πάνω στο ίδιο σύνολο δεδομένων, οι οποίες παρουσιάστηκαν επίσης στο προηγούμενο Κεφάλαιο.

6.3.5.1 Συγκρίσεις με άλλες μεθόδους

Συνολικά Αποτελέσματα	
Μέθοδος	Ακρίβεια
(Santini, 2007)	90.6%
(Kim & Ross, 2007)	92.7%
(Mason, et al., 2009)	94.6%
Character n-grams – Binary	96.2%
Character n-grams – TF	92.5%
Words – Binary	95.5%
Words – TF	95.1%
Textual + structural	96.5%
Fine Tuning Pretrained Models (ULMFiT)	**99.1**%

Πίνακας 6.6: Συγκρίσεις επιδόσεων για το σύνολο δεδομένων 7Genre

6.3.6 Αποτελέσματα - KI-04

Όπως και στην προηγούμενη υπό-ενότητα, θα παρουσιάσουμε στους παρακάτω πίνακες τα αποτελέσματα της επίδοσης του μοντέλου στο σύνολο δεδομενων του KI-04 ως προς την ακρίβεια στο Validation και στο Test, για κάθε διαφορετικό Fold, καθώς και την συνολική επίδοση μέσω του υπολογισμού του Μέσου Όρου των επιδόσεων σε κάθε επανάληψη.

Συνολικά Αποτελέσματα	
Validation Accuracy	Test Accuracy
78.6	92.6

Πίνακας 6.7: Συνολικά Αποτελέσματα ακρίβειας για το dataset KI-04

Παρατηρώντας τους πίνακες αποτελεσμάτων και σε αυτό το σύνολο δεδομένων, βλέπουμε μια ικανοποιητική εικόνα επιδόσεων μέσα σε ένα μικρό αριθμό κύκλων εκμάθησης (epochs). Στην επόμενη ενότητα θα δώσουμε μια ολοκληρωμένη εικόνα των επιδόσεων, καθώς θα συγκρίνουμε αναλυτικά τα αποτελέσματα των προβλέψεων πάνω στο σύνολο δεδομένων KI-04 με άλλες προσεγγίσεις.

6.3.6.1 Συγκρίσεις με άλλες μεθόδους

Αποτελέσματα	
Μέθοδος	Ακρίβεια
(Meyer zu Eissen & Stein, 2004)	70.0%
(Boese & Howe, 2005)	74.8%
(Santini, 2007)	68.9%
Character n-grams – Binary	82.8%
Character n-grams – TF	79.6%
Words – Binary	82.0%
Words – TF	81.8%
Textual + structural	84.1%
Fine Tuning Pretrained Models (ULMFiT)	**92.6**%

Πίνακας 6.8: Συγκρίσεις επιδόσεων για το C10 dataset

6.3.7 Συγκρίσεις συνολικά για τα δύο σύνολα δεδομένων με άλλες μεθόδους

Συνολικά Αποτελέσματα		
Μέθοδος	7Genre	KI-04
(Meyer zu Eissen & Stein, 2004)	-	70.0%
(Boese & Howe, 2005)	-	74.8%
(Santini, 2007)	90.6%	68.9%
(Kim & Ross, 2007)	92.7%	-
(Mason, et al., 2009)	-	-
Character n-grams – Binary	96.2%	82.8%
Character n-grams – TF	92.5%	79.6%
Words – Binary	95.5%	82.0%
Words – TF	95.1%	81.8%
Textual + structural	96.5%	84.1%
Fine Tuning Pretrained Models (ULMFiT)	**99.1**%	**92.6**%

Πίνακας 6.9: Συγκρίσεις επιδόσεων για τα δύο σύνολα δεδομένων

Κεφάλαιο 7

Επίλογος

7.1 Συμπεράσματα

Μέσω της παρούσας Διπλωματικής Εργασίας, μας δόθηκε η ευκαιρία να εμβαθύνουμε στην χρήση των προ-εκπαιδευμένων Γλωσσικών Μοντέλων στην Κατηγοριοποίηση Κειμένων καθώς και να εξάγουμε σημαντικά αποτελέσματα από τα πειράματα τα οποία διεξήχθησαν. Διαπιστώνουμε λοιπόν ότι, τα προ-εκπαιδευμένα Γλωσσικά Μοντέλα δείχνουν πολύ καλές επιδόσεις ανα περίπτωση. Επιπλέον, παρατηρούμε ότι οι επιδόσεις ποικίλουν από θέμα σε θέμα, καθώς και από επίπεδο σε επίπεδο προβλήματος.

Όσον αφορά την Αναγνώριση Συγγραφέα, μπορούμε εύκολα να διαπιστώσουμε ότι σε Simple Domain πρόβλημα το Μοντέλο είχε ικανοποιητικά αποτελέσματα, και συγκεκριμένα πάνω στο Σύνολο Δεδομένων C10. Αντίθετα, στο Cross Domain πρόβλημα, με το dataset του Fanfiction, το οποίο ήταν από τα βασικά θέματα της Διπλωματικής Εργασίας, παρατηρούμε ότι τα αποτελέσματα δεν ήταν καθόλου ενθαρρυντικά μιας και οι επιδόσεις του ήταν αρκετά χαμηλές, όπως και οι υπόλοιπες προσεγγίσεις με Νευρωνικά Δίκτυα, πάνω στο ίδιο Σύνολο Δεδομένων του Fanfiction.

Στην περίπτωση Κατηγοριοποίησης Ιστοσελίδων με βάση το ύφος/δομή τους, τα αποτελέσματα ήταν αρκετά ενθαρρυντικά δίνοντας επιδόσεις πάνω από κάθε άλλη προσέγγιση και στα δύο Σύνολα Δεδομένων στα οποία διεξαγάγαμε το πείραμα. Γενικότερα, οι επιδόσεις ποικίλουν από πρόβλημα σε πρόβλημα και φαίνεται πως τα προ-εκπαιδευμένα Γλωσσικά Μοντέλα, χαράζουν ένα καινούργιο δρόμο στην Κατηγοριοποίηση Κειμένων δίνοντας σημαντικά θετικά αποτελέσματα.

7.2 Παράρτημα

Τον πηγαίο κώδικα των πειραμάτων μπορείτε να τον βρείτε στο ανοιχτό Repository στο GitHub στον παρακάτω σύνδεσμο: https://github.com/AthanCB/Thesis

Βιβλιογραφία

[1] Collobert - Weston - Bottou - Karlen - Kavukcuoglu και Kuksa. *Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research 12, σελίδες: 2493 - 2537, 2011*.

[2] Ioannis Nasikas. *Text Mining: Μια νεα προτεινόμενη μέθοδος με χρήση κανόνων συσχέτισης, σελίδες: 131 , 2006*. URL: `http://nemertes.lis.upatras.gr/jspui/handle/10889/518`.

[3] Heikki Mannila. *Data mining: machine learning, statistics, and databases*. (Επίσκεψη 10/04/2019).

[4] Steve Mutuvi. *PreTrainedModels*. URL: `https://heartbeat.fritz.ai/using-transfer-learning-and-pre-trained-language-models-to-classify-spam-549fc0f56c20`.

[5] OpenAi Team. *Better Language Models and Their Implications, 2019*. URL: `https://openai.com/blog/better-language-models/`.

[6] M. E. Maron. *Automatic Indexing: An Experimental Inquiry, Journal of the ACM, σελίδες: 404-417, 1961*. URL: `https://dl.acm.org/citation.cfm?doid=321075.321084`.

[7] Upendra Sapkota - Steven Bethard - Solorio - Manuel MontesyGomez. *Not All Character N-grams Are Created Equal: A Study in Authorship Attribution, σελίδες: 93 - 100, 2015*. URL: `https://wing.comp.nus.edu.sg/~antho/N/N15/N15-1010.pdf`.

[8] Alexandra Chronopoulou. *"Τεχνικές Μεταφοράς Μάθησης σε Βαθιά Νευρωνικά Δίκτυα για Ανάλυση Συναισθήματος και Σημασιολογική Μοντελοποίηση", σελίδες: 21-22, 2019*.

[9] Jeremy Howard και Sebastian Ruder. *Universal Language Model Fine-tuning for Text Classification, Association for Computational Linguistics, σελίδες: 328–339, 2018*. URL: `arxiv.org/pdf/1801.06146.pdf/`.

[10] Efstathios Stamatatos. *A Survey of Modern Authorship Attribution Methods, Journal of the American Society for Information Science and Technology - σελίδες: 28 - 2008*.

[11] Mike Kestemont - Michael Tschuggnall - Efstathios Stamatatos - Walter Daelemans - Günther Specht - Benno Stein και Martin Potthast. *Overview of the Author Identification Task at PAN-2018 Cross-domain Authorship Attribution and Style Change Detection. CLEF conference - 2018*.

[12] Charu C. Aggarwal - ChengXiang Zhai. *A Survey of Text Classification Algorithms. Mining Text Data - σελίδες: 163 - 222 - 2012*.

[13] Conrad Sanderson - Simon Guenter. *Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation - σελίδες 482-491 -2006*.

[14] Conrad Sanderson - Simon Guenter. *Author verification by linguistic profiling: An exploration of the parameter, Journal ACM Transactions on Speech and Language Processing (TSLP) , 2007*.

[15] Grieve. *Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, Oxford Journal - σελίδες: 251–270 - 2007*.

[16] Ioannis kanaris Efstathios Stamatatos. *Learning to Recognize Webpage Genres, Journal Information Processing and Management - σελίδες: 499-512 - 2009*.

[17] Jorg Hakenberg - Conrad Plake - Ulf Leser. *Genic Interaction Extraction - Identification of Language Patterns Based on Alignment and Finite State Automata - 2005*.

[18] Efstathios Stamatatos. *Authorship Attribution Using Text Distortion, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017*.

[19] Tal Linzen - Emmanuel Dupoux και Yoav Goldberg. *Assessing the ability of lstms to learn syntax-sensitive dependencies, Transactions of the Association for Computational Linguistics - 2016*.

[20] Kristina Gulordava - Piotr Bojanowski - Edouard Grave - Tal Linzen και Marco Baroni. *Colorless green recurrent networks dream hierarchically. 2018*.

[21] Alec Radford - Rafal Jozefowicz και Ilya Sutskever. *Learning to generate reviews and discovering sentiment, 2017*.

[22] Jason Yosinski - Jeff Clune - Yoshua Bengio και Hod Lipson. *How transferable are features in deep neural networks - In Advances in neural information processing systems - 2014*.

[23] Anders Tolver. *An introduction to Markov Chains, σελίδες :1 - 159, 2016*. URL: `http://web.math.ku.dk/noter/filer/stoknoter.pdf`.

[24] Ξενη Μαρια. *"Λογιστική Παλινδρόμηση Διαχωριστική Ανάλυση", σελίδες: 5 - 15*. URL: `https://nemertes.lis.upatras.gr/jspui/bitstream/10889/5174/1/Diplwmatiki.pdf`.

[25] Bjarke Felbo - Alan Mislove - Anders Søgaard - Iyad Rahwan και Sune Lehmann. *Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.*

[26] Joaquim Ferreira da Silva1 -Gaël Dias1 - Sylvie Guilloré και José Gabriel Pereira Lopes. *Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units, 1999*. URL: `https://www.researchgate.net/publication/220773699_Using_LocalMaxs_Algorithm_for_the_Extraction_of_Contiguous_and_Non-contiguous_Multiword_Lexical_Units`.

[27] Pedro Barahona - Jose J. Alferes. *Progress in Artificial Intelligence, 9th Portuguese Conference on Artificial Intelligence, 1999*. URL: https://books.google.gr/books?id=5spuCQAAQBAJ&pg=PA118&lpg=PA118&dq=Symmetrical+Conditional+Probability+silva&source=bl&ots=LZXgIyV-SK&sig=ACfU3U1-QiAVVMPDQFxV77JcR3GlaptI8w&hl=el&sa=X&ved=2ahUKEwi0ofTWssbkAhUC6aQKHWx
v=onepage&q=Symmetrical.

[28] Fokianos. *"Εισαγωγή στην R", page: 109, 2010*. URL: http://www.mas.ucy.ac.cy/~fokianos/GreekRbook/dialeksi10.pdf.

[29] Robnik-Sikonja - Kononenko. *Theoretical and Empirical Analysis of ReliefF and RReliefF, 2003*.

[30] Stephen Merity - Caiming Xiong - James Bradbury και Richard Socher. 2017b. *Pointer Sentinel Mixture Models, In Proceedings of the International Conference on Learning Representations, 2017*.

[31] Konstantinou George. *"Μελέτη κατηγοριοποίησης δεδομένων με Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) και υλοποίηση εφαρμογής", 2012*. URL: http://apothesis.teicm.gr/xmlui/handle/123456789/859.

[32] Serge Sharoff - Zhili Wu - Katja Markert. *The Web Library of Babel: evaluating genre collections,LREC, 2010*.

[33] Helmut Schmid. *TreeTagger - a part-of-speech tagger for many languages, TC project at the Institute for Computational Linguistics of the University of Stuttgart, 1995*.