Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Αναγνώριση Επιχειρησιακών Διαδικασιών και Ταξινόμηση ως αυτοματοποιήσιμων ή μη με χρήση Μηχανικής Μάθησης

Μελέτη και υλοποίηση

Διπλωματικη Εργασια

της

ΜΕΓΑΛΟΟΙΚΟΝΟΜΟΥ ΕΙΡΗΝΗ

Επιβλέπων: Ανδρέας - Γεώργιος Σταφυλοπάτης
Καθηγητής ΕΜΠ

Αθήνα, Απρίλιος 2021

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Αναγνώριση Επιχειρησιακών Διαδικασιών και Ταξινόμηση ως αυτοματοποιήσιμων ή μη με χρήση Μηχανικής Μάθησης

Μελέτη και υλοποίηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΕΓΑΛΟΟΙΚΟΝΟΜΟΥ ΕΙΡΗΝΗ

Επιβλέπων: Ανδρέας - Γεώργιος Σταφυλοπάτης
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 22 Απριλίου 2021.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ανδρέας - Γεώργιος Σταφυλοπάτης	Στέφανος Κόλλιας	Γεώργιος Στάμου
Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ	Αναπληρωτής Καθηγητής ΕΜΠ

Αθήνα, Απρίλιος 2021

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Ειρήνη Μεγαλοοικονόμου

22 Απριλίου 2021

Περίληψη

Η συνεχής ψηφιοποίηση που λαμβάνει χώρα τα τελευταία χρόνια στον κόσμο των επιχειρήσεων, έχει ως αποτέλεσμα την ανάγκη για αυτοματοποίηση των διαδικασιών που τις αποτελούν. Αυτή η ανάγκη έχει οδηγήσει σε ένα αυξημένο ενδιαφέρον για τα ψηφιακά ρομπότ - RPA (Robotic Process Automation). Πρόκειται για λογισμικό το οποίο αυτόματα εκτελεί επαναλαμβανόμενες εργασίες. Το RPA βοήθησε τις επιχειρήσεις να μειώσουν δραστικά τα εργασιακά καθήκοντα των υπαλλήλων, ιδίως στη χειρονακτική εργασία, δίνοντας τους την ευκαιρία να εργαστούν σε καλύτερες συνθήκες εν αντιθέσει με το παρελθόν.

Τα οφέλη του RPA στην εξοικονόμηση του κόστους και σε άλλους σχετικούς δείκτες απόδοσης έχουν παρουσιαστεί εκτενώς. Παρ'όλα αυτά μια από τις μεγαλύτερες προκλήσεις που αφορούν το RPA είναι η αποτελεσματική αναγνώριση διαδικασιών και εργασιών που θεωρούνται κατάλληλες για αυτοματοποίηση. Συνήθως οι περιγραφές των διαδικασιών που ακολουθούνται σε έναν οργανισμό, όπως οι οδηγίες για μια αναφορά, υπάρχουν σε μορφή κειμένου και εκεί παρέχεται σημαντική γνώση για το ενδεχόμενο επίπεδο αυτοματοποίησης. Ωστόσο οι επιχειρήσεις συχνά διατηρούν εκατοντάδες ή ακόμα χιλιάδες από αυτές τις περιγραφές, το οποίο βεβαίως καθιστά την χειρονακτική ανάλυση μη εφικτή, ειδικά για μεγαλύτερου μεγέθους οργανισμούς.

Αναγνωρίζοντας την μεγάλη χειρονακτική προσπάθεια που απαιτείται για τον καθορισμό του τρέχοντα βαθμού αυτοματισμού σε μια επιχειρησιακή διαδικασία, προτείνουμε μια μέθοδο η οποία μπορεί αυτόματα να κάνει αυτήν την αναγνώριση.

Πιο συγκεκριμένα στην παρούσα διπλωματική εργασία έχοντας ως είσοδο έγγραφα διαδικασιών, χρησιμοποιούμε εργαλεία για την επεξεργασία φυσικής γλώσσας (NLP - Natural Language Process) με σκοπό την αναγνώριση του εκάστοτε κειμένου και αξιοποιούμε την επιβλεπόμενη μάθηση (supervised machine learning) για να ταξινομήσουμε μια εργασία που περιγράφεται σε αυτά τα έγγραφα ως πλήρως χειρονακτική (manual - 0), ως αλληλεπίδραση ανθρώπου με κάποιο σύστημα (interaction of human with information system - 1) ή ως αυτόματη (automated - 2).

Η αξιολόγηση έγινε με ένα σετ 507 δραστηριοτήτων από ένα σύνολο 47 κειμένων με περιγραφές διαδικασιών και επιδεικνύει πως η μέθοδος μας παράγει ικανοποιητικά αποτελέσματα.

Λέξεις Κλειδιά

Ρομπότ Λογισμικού, Επιχειρησιακές Διαδικασίες, Επεξεργασία Φυσικής Γλώσσας, Αλγόριθμοι Μηχανικής Μάθησης, Ταξινόμηση Δεδομένων, Επιβλεπόμενη Μάθηση, Υπερμοντελοποίηση

Abstract

The continuous digitization that takes place the last years in the world of business, resulted in the need of organizations to improve the automation of their business processes. This has led to an increased interest in digital robots - Robotic Process Automation (RPA). RPA solutions emerge in the form of software that automatically executes repetitive and routine tasks. RPA has helped businesses to drastically reduce the working assignments of the employees, mainly in the manual section, giving them the chance to work under better conditions in contrast to the past.

The benefits of RPA on cost savings and other relevant performance indicators have been demonstrated in different contexts a lot. Although one of the key challenges for RPA endeavors is to effectively identify processes and tasks that are suitable for automation. Textual process descriptions, such as work instructions, provide rich and important insights about this matter. However, organisations often maintain hundreds or even thousands of them, which makes a manual analysis unfeasible for larger organizations.

Recognizing the large manual effort required to determine the current degree of automation in an organization's business process, we use this paper to propose an approach that is able to automatically do so.

More specifically, in this thesis having as input documents of processes, we use the tools to process natural language (NLP - Natural Language Process) in order to identify each text and we leverage supervised machine learning to automatically identify whether a task described in a textual process description is manual (manual - 0), an interaction of a human with an information system (interaction of human with information system - 1) or automated (automated - 2).

An evaluation with a set of 507 activities from a total of 47 textual process descriptions demonstrates that our approach produces satisfactory results.

Keywords

Software Robot, Business Automation, Natural language Processing, Machine Learning Algorithms, Data Classification, Supervised Learning, Overfitting

στους γονείς μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Ανδρέα Σταφυλοπάτη για την ευκαιρία που μου έδωσε, να εκπονήσω τη διπλωματική μου εργασία στο Εργαστήριο Ευφυών Συστημάτων του τομέα Τεχνολογίας Πληροφορικής και Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου

Ιδιαίτερες ευχαριστίες θα ήθελα να δόσω στον κύριο Γεώργιο Σιόλα για την εξαιρετική συνεργασία και τη συνεχή καθοδήγηση κατά τη συγγραφή της εργασίας αυτής καθώς και την άμεση ανταπόκριση του σε όποια δυσκολία αντιμετώπιζα. Τον ευχαριστώ που με προέτρεψε να εργαστώ σε ένα θέμα τόσο ενδιαφέρον και δημιουργικό όπως ακριβώς το είχα στο μυαλό μου όταν του το πρότεινα.

Ένα ευχαριστώ οφείλω και στον Henrik Leopold καθηγητή του Kuehne Logistics University της Γερμανίας καθώς χάρη στη συμβολή του είχα το annotated data set για να μπορέσω να εργαστώ πάνω στο θέμα που διάλεξα.

Αυτήν την στιγμή τελειώνει ένα πολύ μεγάλο ταξίδι για εμένα. Τα χρόνια της ‹σχολής›, ήταν πολύ γεμάτα χρόνια. Πέρα από τις τεχνικές γνώσεις που κατάφερα να αποκτήσω, γέμισα με εμπειρίες, έκανα ταξίδια, συμμετείχα σε εθελοντικά προγράμματα, έκανα ένα απίστευτο Erasmus στο Βέλγιο, δούλεψα σε εφήμερες δουλειές αλλά έκανα και μια πολύ πετυχημένη πρακτική που μου έδωσε μια κατεύθυνση για το μέλλον. Το σημαντικότερο από όλα γνώρισα και διαμόρφωσα τον εαυτό μου.

Θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που ήταν μέρος αυτού του ταξιδιού. Πρώτα από όλα τους πιο κοντινούς μου φίλους την Μίνα και τον Λέντιο που είναι σαν αδέλφια μου, και που πάντα ήταν εκεί για να με βοηθήσουν. Την παρέα μου από την σχολή την Σοφία, τον Στέλιο, τον Νίκο, τον Κίμωνα, τον Κώστα, την Δήμητρα, τον Αλέξανδρο και τον Άγγελο με τους οποίους περάσαμε άπειρες στιγμές ξεγνοιασιάς, χαλάρωσης αλλά και πίεσης και άγχους. Ευχαριστώ όλους τους φίλους που είναι πολλοί για να ονοματίσω.

Τέλος, από καρδιάς θα ήθελα να ευχαριστήσω πολύ την οικογένεια μου, την μητέρα μου Βάσω, τον πατέρα μου Βαλέριο και τα αδέλφια μου Νίκο και Ηρώ για την στήριξη, την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια. Δεν θα είχα καταφέρει τίποτα χωρίς την βοήθεια τους.

Αθήνα, Απρίλιος 2021

Ειρήνη Μεγαλοοικονόμου

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κατάλογος Αλγορίθμων

Πρόλογος

Στο πρώτο κεφάλαιο γίνεται μια εισαγωγή της διπλωματικής εργασίας. Συζητείται το κίνητρο και το πρόβλημα που εντοπίσαμε και βασίστηκε όλη η εργασία. Επιπλέον, παρουσιάζεται ο στόχος και το περίγραμμα των διαφόρων κεφαλαίων του κειμένου αυτής της διπλωματικής εργασίας.

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρο

Πολλοί μεγάλοι οργανισμοί αυτή την στιγμη αντιμετωπίζουν την πρόκληση να συμβαδίσουν με την αυξημένη ψηφιοποίηση. Μεταξύ άλλων, απαιτείται να προσαρμόσουν τα υπάρχοντα επιχειρησιακά μοντέλα τους και αντίστοιχα να βελτιώσουν την αυτοματοποίηση των επιχειρηματικών διαδικασιών. Ενώ το πρώτο είναι κυρίως στρατηγικό καθήκον, το δεύτερο απαιτεί συγκεκριμένες λειτουργικές λύσεις. Μία από αυτές τις λύσεις για την αύξηση του αυτοματισμού είναι η πρόσφατη εξέλιξη του λεγόμενου Robotic Process Automation (RPA) - ή στα ελληνικά ψηφιακά ρομπότ. Συγκεκριμένα πρόκειται για λογισμικό που εκτελεί αυτόματα επαναλαμβανόμενες και συνήθεις εργασίες. Με αυτόν τον τρόπο οι εργαζόμενοι μπορούν να αφιερώσουν χρόνο και προσπάθεια σε πιο περίπλοκες και με μεγαλύτερη αξία εργασίες.

Ενώ τα οφέλη του RPA έχουν αποδειχθεί και σε θεωρητικό και πρακτικό πλαίσιο, μια από τις βασικότερες προκλήσεις είναι ο αποτελεσματικός προσδιορισμός διαδικασιών και εργασιών που είναι κατάλληλες για αυτοματοποίηση. Μέχρι στιγμής οι έρευνες έχουν επικεντρωθεί στην εδραίωση κριτηρίων και αναλυτικών οδηγιών για τον διαχωρισμό των κατάλληλων αυτών διαδικασιών. Ωστόσο, όλες αυτές οι μέθοδοι απαιτούν χειροκίνητη ανάλυση από κάποιον άνθρωπο, εξαρτώνται δηλαδή από την μη αυτόματη αναγνώριση των εργασιών και (υπό-)διεργασιών που αυτοματοποιούνται ή υποστηρίζονται από ένα σύστημα πληροφοριων. Αυτή η εργασία αναγνώρισης απαιτεί διεξοδική ανάλυση των εγγράφων της εκάστοτε διαδικασίας ή των μοντέλων, σχεδίων που μπορεί να υπάρχουν. Κυρίως τα έγγραφα, παρέχουν συνήθως πλούσιες και λεπτομερείς πληροφορίες σχετικά με το πιθανό επίπεδο αυτοματοποίησης. Οι οργανισμοί ωστόσο, τείνουν να διατηρούν εκατοντάδες ή ακόμα και χιλιάδες από αυτά, το οποίο καθιστά την αναγνώριση αυτή πολύ δύσκολη.

Εμείς, ως νέοι ηλεκτρολόγοι μηχανικοί και μηχανικοί υπολογιστών, είμαστε υποχρεωμένοι να παρακολουθούμε τις τεχνολογικές τάσεις που λαμβάνουν χώρα στον τομέα μας και να εκπαιδεύσουμε τους εαυτούς μας, ώστε να είμαστε σε θέση να συμβάλλουμε στην βελτίωση και στην συνεχή ανάπτυξη υπερσύγχρονων λύσεων.

Οι προαναφερθέντες λόγοι παρακίνησαν την έρευνα και τη σύνταξη αυτής της διπλωματικής εργασίας. Αναγνωρίζοντας τη μεγάλη χειροκίνητη προσπάθεια που απαιτείται για τον προσδιορισμό του βαθμού αυτοματισμού στις επιχειρηματικές διαδικασίες ενός οργανισμού, χρησιμοποιούμε αυτό το έγγραφο για να προτείνουμε ένα μηχανισμό που μπορεί να το κάνει

αυτόματα.

1.2 Το πρόβλημα της Αναγνώρισης Διαδικασιών

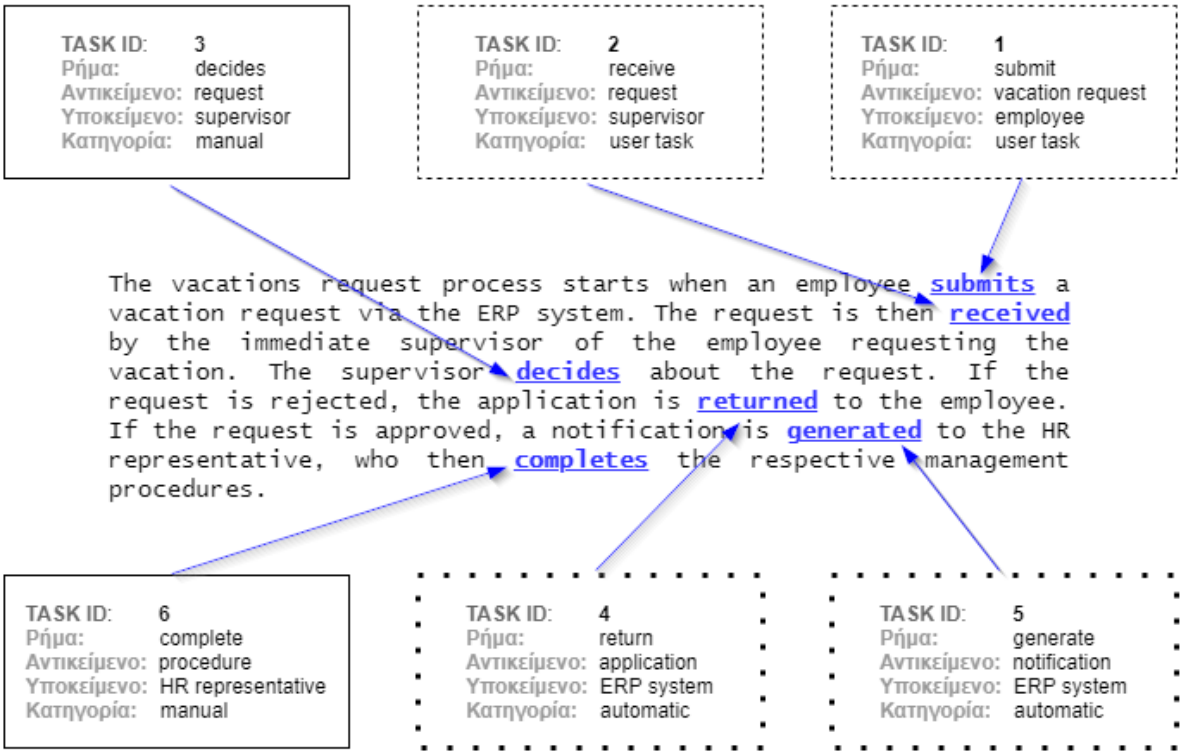
Ο προσδιορισμός του επιπέδου αυτοματισμού μιας διαδικασίας αποτελεί μια αρκετά δύσκολη δουλειά από την μεριά ενός αναλυτή. Ακόμα και αν καταφέρει να εντοπίσει διαδικασίες οι οποίες είναι κατάλληλες για αυτοματοποίηση με βάση τα κριτήρια που έχουν θεσπιστεί, δεν είναι σίγουρο πως θα είναι η καταλληλότερη διαδικασία από όλες όσες υπάρχουν στον εκάστοτε οργανισμό. Πέραν τούτου πρόκειται και για μια εξαιρετικά χρονοβόρα διαδικασία. Αν αναλογιστεί κανείς πόσες διευθύνσεις, τμήματα και υπό-ομάδες μπορεί να υπάρχουν σε έναν οργανισμό από τις οποίες η καθεμία έχει τις δικές τους διαδικασίες, μπορεί να καταλάβει ότι είναι πολύ δύσκολο να αναλυθεί όλος αυτός ο όγκος χειροκίνητα από οποιονδήποτε αριθμό ατόμων ώστε να επιλεχθεί η σωστή ή οι σωστές διαδικασίες. Αυτή η επιλογή η οποία δεν είναι η βέλτιστη έχει ως αποτέλεσμα να μην επιφέρει τα επιθυμητά αποτελέσματα όσον αφορά την εξοικονόμηση χρόνου και κόστους στις επιχειρήσεις.

1.3 Στόχος Διπλωματικής Εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι να σας παρουσιάσει μια πρόταση, που θα έλυνε το παραπάνω πρόβλημα. Δηλαδή έναν μηχανισμό ο οποίος θα μπορεί να αναγνωρίσει την δυνατότητα αυτοματοποίησης μιας εργασίας, χωρίς την εμπλοκή κάποιου ανθρώπου - αναλυτή. Πιο συγκεκριμένα στην δική μας προσέγγιση προσπαθήσαμε να συνδυάσουμε την επιβλεπόμενη μηχανική μάθηση (supervised machine learning) και τις τεχνικές επεξεργασίας φυσικών γλωσσών (natural language processing) με σκοπό τον διαχωρισμό κάθε εργασίας σε (1) χειροκίνητη εργασία (manual task) , (2) σε εργασία χρήστη, δηλαδή αυτή η οποία γίνεται με αλληλεπίδραση ενός ανθρώπου με ένα πληροφοριακό σύστημα (user task) ή (3) σε αυτοματοποιημένη εργασία (automated task). Στο Σχήμα 1 φαίνεται μια περιγραφή μιας διαδικασίας και αναφέρονται οι επιμέρους εργασίες με τον βαθμό αυτοματοποίησης της καθεμίας.

Το Σχήμα 1 δείχνει ότι αυτή η περιγραφή διαδικασίας περιλαμβάνει δύο χειροκίνητες εργασίες, δύο εργασίες χρήστη και δύο αυτοματοποιημένες διαδικασίες. Οι χειροκίνητες εργασίες περιλαμβάνουν την απόφαση του επόπτη σχετικά με το αίτημα άδειας (task 3) και την ολοκλήρωση των διαδικασιών διαχείρισης από τον εκπρόσωπο του HR (task 6). Οι εργασίες χρήστη είναι οι δύο εργασίες της διαδικασίας που εκτελούνται με τη βοήθεια ενός πληροφοριακού συστήματος. Στο συγκεκριμένο παράδειγμα δηλαδή η υποβολή και η λήψη του αιτήματος άδειας. (task 1 και task 2). Τέλος οι αυτοματοποιημένες εργασίες είναι αυτές που εκτελούνται από το σύστημα ERP. Αυτό περιλαμβάνει την επιστροφή της αίτησης στον υπάλληλο (task 4) καθώς και τη ειδοποίηση του εκπρόσωπου του HR (task 5).

Στο παρακάτω παράδειγμα φαίνεται ακριβώς ο στόχος της παρούσας διπλωματικής εργασίας, ο οποίος επιγραμματικά αποτελούταν από την εύρεση όλων των εργασιών μέσα από μια



Σχήμα 1.1: *Περιγραφή επιχειρησιακής διαδικασίας με υπογραμισμένες επιμέρους εργασίες και αναφορά των βαθμών αυτοματισμού τους*

περιγραφή μιας διαδικασίας, και η κατηγοριοποίηση τους με βάση τον βαθμό αυτοματισμού τους. Οι κύριες προκλήσεις για τα παραπάνω ήταν οι εξής:

- *Προσδιορισμός των εργασιών*: Πριν από την ταξινόμηση μιας εργασίας, μια αυτοματοποιημένη προσέγγιση πρέπει να είναι σε θέση να εντοπίζει τις εργασίες που περιγράφονται σε ένα κείμενο. Για παράδειγμα τα ρήματα από το παραπάνω κείμενο «ξεκινά» (starts), «απορρίφθηκε» (rejected) και «εγκρίθηκε» (approved) δεν σχετίζονται με εργασίες. Όσον αφορά το «ξεκινά» δεν σχετίζεται με την εργασία ταξινόμησης επειδή αντιπροσωπεύει ένα κομμάτι μετα-πληροφοριών (meta-data) της διαδικασίας. Τα δύο επόμενα ρήματα δεν είναι σχετικά καθώς σχετίζονται με όρους παρά με εργασίες, δηλαδή «εάν το αίτημα απορριφθεί» ή «εάν το αίτημα εγκριθεί» περιγράφουν μια κατάσταση και όχι μια δραστηριότητα που εκτελείται. Πέραν του προσδιορισμού του ρήματος, η αναγνώριση των εργασιών απαιτεί επίσης να βρεθεί σωστά το αντικείμενο στο οποίο αναφέρεται ένα ρήμα και ο πόρος που εκτελεί την εργασία.

- *Μελέτη των συμφραζόμενων*: Για να προβλεφθεί αξιόπιστα εάν μια συγκεκριμένη δραστηριότητα είναι εργασία χειροκίνητη, εργασία χρήστη ή αυτοματοποιημένη εργασία πρέπει να ληφθούν υπόψιν και τα συμφραζόμενα. Ας πάρουμε για παράδειγμα την εργασία 2 από το παραπάνω κείμενο. Ενώ σε αυτήν την περιγραφή διαδικασίας το αίτημα άδειας υποβάλλεται σε ένα σύστημα πληροφοριών, θα μπορούσε σε άλλες διαδικασίες

να υποβάλλεται γραπτά ή προφορικά. Το γεγονός ότι ένα σύστημα πληροφοριών αναφέρεται στην πρώτη πρόταση πρέπει αντίστοιχα να ληφθεί υπόψιν κατά την ταξινόμηση μιας εργασίας που περιγράφεται αργότερα στη διαδικασία.

Στις επόμενες ενότητες θα παρουσιαστούν εκτενώς οι λεπτομέρειες της προτεινόμενης λύσης μας.

1.4 Συνεισφορά Διπλωματικής Εργασίας

Όπως αναφέρθηκε παραπάνω, σκοπός της διπλωματικής είναι η ανάπτυξη ενός συστήματος το οποίο είναι ικανό να ανανωρίσει επιχειρηματικές διαδικασίες και να τις κατηγοριοποιήσει με βάση τον βαθμό αυτοματοποίησης τους. Χρησιμοποιώντας την μεθοδό μας ένας αναλυτής μπορεί να απαλλαχθεί από αυτό το χρονοβόρο έργο. Επιδιώκοντας αυτό το στόχο μελετήσαμε και αξιολογήσαμε σχετικά έργα, αναπτύξαμε κα δημιουργήσαμε μια δική μας προσέγγιση και την αξιολογήσαμε.

Επομένως αυτή η διπλωματική παρέχει τις ακόλουθες συνεισφορές:

- **Ανασκόπηση βιβλιογραφίας**: Συλλέξαμε διάφορα έργα που ασχολούνται με το πρόβλημα αυτοματοποιημένων διαδικασιών RPA και μη, τις μελετήσαμε και αξιολογήσαμε τα πλεονεκτήματα κα τις αδυναμίες τους.

- **Κατηγοριοποίηση ζητημάτων**: Αναπτύξαμε ενα θεωρητικό πλαισιο που κατηγοριοποιεί σημαντικά ζητήματα που σχετίζονται με το θέμα μας.

- **Νέα μέθοδος αναγνώρισης και κατηγοριοποίησης**: Μεταφέραμε το θεωρητικό μας πλαίσιο σε πράξη και αναπτύξαμε μια νέα προσέγγιση. Η μέθοδος μας βασίζεται σε συντακτική ανάλυση, γραμματικές σχέσεις και εξαρτήσεις. Συνδυάζουμε γνωστά εργαλεία NLP και στο τέλος συγκρίνουμε διάφορους αλγορίθμους μηχανικής μάθησης για να συγκρίνουμε αποτελέσματα.

- **Πλήρες σύνολο δεδομένων**: Για να αξιολογήσουμε την μέθοδό μας συλλέξαμε 47 περιγραφές διαδικασιών από διάφορους τομείς. Όλα τα κείμενα περιλαμβάνονται στο τέλος της διπλωματικής, το οποίο θα επιτρέψει και σε άλλους ερευνητές να κάνουν δοκιμές, καθώς είναι ένας τομέας που δεν είναι εύκολο να βρεθεί Data Set.

- **Προσέγγιση αξιολόγησης**: Για να εκτιμήσουμε την ακρίβεια του μοντέλου μας χρησιμοποιήσαμε ένα υποσημειωμένο data set 507 συνολικά εργασιών με ετικέτες ως προς το επίπεδο αυτοματοποίησης.

1.5 Δομή Διπλωματικής Εργασίας

Η εργασία αυτή είναι οργανωμένη σε 5 κεφάλαια εκτός της εισαγωγής όπου τοποθετήθηκε το πρόβλημα.

Στο Κεφάλαιο 2 καλύπτεται το θεωρητικό υπόβαθρο όσων κρίθηκαν απαραίτητα για την κατανόηση αυτής της διπλωματικής εργασίας. Αρχικά γίνεται αναφορά στο πρόβλημα της

αναγνώρισης επιχειρησιακών διαδικασιών αναλύοντας βασικές έννοιες, προβλήματα και ορολογίες. Παρουσιάζονται επίσης οι βασικές έννοιες του NLP και όλα τα εργαλεία που χρησιμοποιήσαμε. Στη συνέχεια γίνεται μια παρουσίαση του related work, δηλαδή των σημαντικότερων εργασιών που έχουν γίνει μέχρι σήμερα πάνω στο πρόβλημα που καλούμαστε να λύσουμε.

Στο Κεφάλαιο 3 αναλύονται οι σχεδιαστικές επιλογές που κάναμε κατά την ανάπτυξη των μοντέλων μας και τα βήματα υλοποίησης που ακολουθήσαμε. Παρουσιάζονται οι τρόποι με τους οποίους αντιμετωπίσαμε τα όποια προβλήματα συναντήσαμε και κάποιοι αλγόριθμοι για την περιγραφή των λύσεων.

Στο Κεφάλαιο 4 παρουσιάζονται τα δεδομένα με τα οποία εργαστήκαμε και τα χαρακτηριστικά τους. Παρουσιάζονται επίσης τα αποτελέσματα των πειραμάτων κι εξάγονται τα συμπεράσματα. Ξεκινάμε με μια συγκριτική μελέτη των διαφορετικών μοντέλων που δημιουργήσαμε, αναλύοντας ξεχωριστά τα διάφορα χαρακτηριστικά και τον τρόπο με τον οποίο αυτά επηρεάζουν την εκπαίδευση και τα αποτελέσματα του μοντέλου.

Στο Κεφάλαιο 5 γίνεται μια σύνοψη της διπλωματικής και στη συνέχεια εξάγονται και συγκεντρώνονται τα τελικά συμπεράσματα. Στη συνέχεια αναλύουμε προσωπικές παρατηρήσεις του συγγραφέα και κάνουμε μια ποιοτική αξιολόγηση των αποτελεσμάτων έχοντας απομακρυνθεί από τα αριθμητικά αποτελέσματα και κοιτώντας τη γενικότερη εικόνα. Τέλος αναφέρονται πιθανές μελλοντικές κατευθύνσεις της επιστημονικής μελέτης, περισσότερες από τις οποίες προκύπτουν ως λογική συνέχεια της δουλειάς που κάναμε για αυτή την εργασία.

Μέρος Ι

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζονται αναλυτικά οι τρεις βασικοί πυλώνες που έχουν σχέση με την εργασία αυτή, δηλαδή η τεχολογία RPA, η επεξεργασία φυσικης γλώσσας NLP, και κάποιοι αλγόριθμοι μηχανικής μάθησης με τους οποίους δουλέψαμε.

2.1 Robotic Process Automation

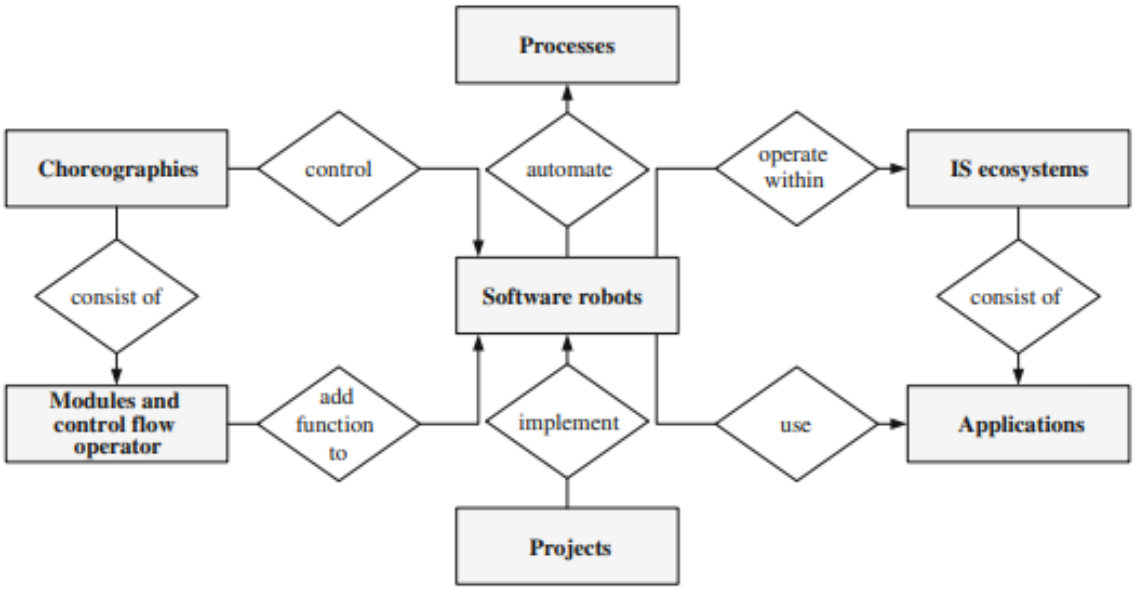
2.1.1 Η ανάγκη για ψηφιακά ρομπότ

Η ψηφιοποίηση δεν αποτελέι πλέον καινούργιο φαινόμενο. Ο τομέας του IT εξελίσσεται συνεχώς και δημιουργεί νέα προϊόντα και ευκαιρίες. Έτσι τα σημερινά επιχειρηματικά περιβάλλοντα αντιμετωπίζουν τον συνεχή ψηφιακό μετασχηματισμό, μέσα στον οποίο ούτε η αυτοματοποίηση, ούτε η ρομποτική είναι νέες εξελίξεις. Τα τελευταία χρόνια, το RPA έχει τραβήξει μεγάλη εταιρική προσοχή σχετικά με τις πρωτοβουλίες αυτοματισμού. Σύμφωνα με τον Όμιλο Υπηρεσιών Πληροφοριών (Information Services Group - ISG) (2021) οι πελάτες του RPA συνειδητοποιούν μείωση κατά 75% ή περισσότερο στην επιχειρησιακή προσπάθεια και αύξηση πάνω από 80% στην αποτελεσματικότητα των διαδικασιών.

2.1.2 Τι είναι τα ψηφιακά ρομπότ

Για να χαρακτηρίσουμε το RPA με δομημένο τρόπο, παρουσιάζουμε τα κύρια χαρακτηριστικά του στο παρακάτω σχήμα δίνοντας έμφαση σε τέσσερα μεγάλα χαρακτηριστικά. Οι IEEE Corporate Advisory Group ορίζουν το RPA ως τη χρήση μιας «προκαθορισμένης παρουσίας λογισμικού που χρησιμοποιεί επιχειρηματικούς κανόνες και προκαθορισμένη δραστηριότητα για την ολοκλήρωση της αυτόνομης εκτέλεσης ενός συνδυασμού διαδικασιών, δραστηριοτήτων, συναλλαγών και εργασιών σε ένα ή περισσότερα συστήματα λογισμικού για να αποδώσουν ένα αποτέλεσμα ή υπηρεσία που θα γινόταν με ανθρώπινη διαχείριση».

Το RPA λειτουργεί στη διεπαφή χρήστη των εργαλείων λογισμικού και αυτοματοποιεί τις αλληλεπιδράσεις με το ποντίκι και το πληκτρολόγιο για την αφαίρεση επαναλαμβανόμενων χειρονακτικών εργασιών. Αυτό ελαχιστοποιεί το ανθρώπινο λάθος που μπορεί να οφείλεται σε πλήξη ή σε εξάντληση. Σύμφωνα με ακαδημαϊκά κείμενα του τομέα διακρίνουμε τρεις προσεγγίσεις για την οικοδόμηση των RPA.



Σχήμα 2.1: *Κύριοι Τομείς του RPA*

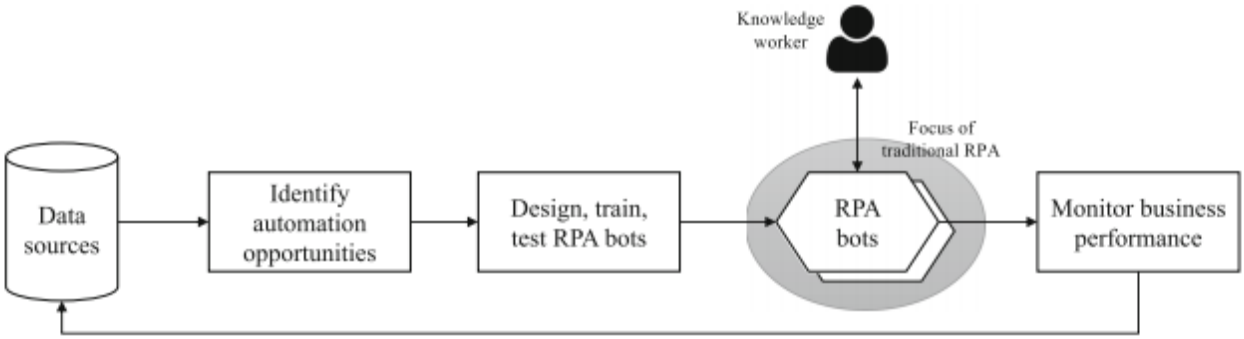
- *Πρώτη προσέγγιση*: Η πρώτη προσέγγιση μαθαίνει να αυτοματοποιεί εργασίες από παραδείγματα ή από επιδείξεις. Τα RPA είτε παρατηρούν όταν οι άνθρωποι εκτελούν τις εργασίες, είτε επεξεργάζονται τα αρχεία καταγραφής συμπεριφοράς (process logs) του λογισμικού. Ένα παράδειγμα αποτελεί μια εργασία που περιέχει αρχεία εισόδου και εξόδου από τα οποία το RPA μπορεί να εξαγάγει τον υποκείμενο κανόνα ή πρόγραμμα. Αυτή είναι μια από τις δημοφιλέστερες προσεγγίσεις για τα RPA. Ωστόσο δεν γενικεύεται καλά σε νέες εφαρμογές λόγω του ιδιαίτερα συγκεκριμένου σχεδιασμού αρχείων καταγραφής αλλά και των διαπεφών του χρήστη.

- *Δεύτερη προσέγγιση*: Η δεύτερη προσέγγιση μαθαίνει εργασίες από κείμενα περιγραφών διαδικασιών. Αυτή η μέθοδος έχει βασιστεί στους ανθρώπους κατά έμμεσο τρόπο αφού τα έγγραφα που περιγράφουν τις διαδικασίες έχουν γραφτεί από ανθρώπους. Δεδομένου ότι δεν απαιτεί την ύπαρξη ενσωματωμένης επιχειρηματικής διαδικασίας, είναι περισσότερο δύσκολο να μάθει τους κανόνες που πρέπει να αυτοματοποιηθούν. Στο κομμάτι αυτό περιλαμβάνεται και η επεξεργασία φυσικής γλώσσας, με την οποία και ασχοληθήκαμε.

- *Τρίτη προσέγγιση*: Η τρίτη προσέγγιση μαθαίνει από την ίδια τη διαδικασία όπως ορίζεται από ένα περιβάλλον με ορισμένα παραδείγματα εισόδου / εξόδου. Συχνά αναφέρεται ως RPA 2.0,καθώς αυτή η προσέγγιση αποσκοπεί στην εξάλειψη της εξαρτώμενης από τον άνθρωπο εκπαίδευσης. Βασίζεται στην υιοθέτηση αλγορίθμων εκμάθησης και την εκπαίδευσή τους για επίτευξη καλύτερης απόδοσης. Αυτή η προσέγγιση είναι η λιγότερο ώριμη μέχρι σήμερα, αλλά θα οδηγήσει σε γενικευμένες λύσεις RPA που προσεγγίζουν τον έξυπνο αυτοματισμό.

Η συνήθης ροή για την κατασκευή μιας διαδικασίας με RPA αποτελείται από αρκετά

βήματα. Πάντα ξεκινάει από την θέληση των εταιριών να εξελιχθούν και να κάνουν ένα βήμα παραπάνω όσον αφορά την ψηφιοποίηση. Τότε καλούν κάποιον αναλυτή ο οποίος θα αναγνωρίσει διαδικασίες οι οποίες θεωρούνται ως ευκαιρίες για αυτοματοποίηση. Τότε οι κατάλληλοι μηχανικοί θα φτιάξουν τα λεγόμενα RPA bots τα οποία είναι ουσιαστικά τα ψηφιακά ρομπότ που θα επαναλαμβάνουν την παραπάνω διαδικασία. Εννοείται ότι δεν πρόκειται για κάτι στατικό καθώς ανά πάσα στιγμή τα δεδομένα μπορεί να αλλάξουν και να πρέπει οι διαδικασίες να τροποποιηθούν ή να αλλάξουν εντελώς. Για αυτόν το λόγο γίνεται συνέχεια παρακολουθηση των έτοιμων διαδικασιών.



Σχήμα 2.2: *Κύκλος ζωής του RPA*

Ένα κρίσιμο στοιχείο για την επιτυχία του RPA είναι ο προσδιορισμός των ευκαιριών (δηλαδή διαδικασιών) που χρήζουν για αυτοματοποίηση. Τροποποιώντας τις σωστές διαδικασίες με RPA μπορεί να μεγιστοποιηθεί το κέρδος σε μια εταιρία, είτε είναι κέρδος χρόνου, χεριών ή και οικονομικό. Συνήθως αυτές οι διαδικασίες προσδιορίζονται από ειδικούς αναλυτές. Παρ'όλο που η έρευνα στο RPA δείχνει πολλά υποσχόμενες μεθόδους και οδηγίες για την αξιολόγηση του αυτοματισμού μια διαδικασίας, οι πληροφορίες που υπάρχουν για να γίνει αυτόματα όλη αυτή η αναζήτηση είναι ελάχιστες.

2.1.3 Οφέλη των ψηφιακών ρομπότ

Το RPA επέτρεψε την ενσωμάτωση συστημάτων που διαφορετικά δεν θα είχαν ενσωματωθεί και διευκόλυνε τον φόρτο εργασίας των εργαζόμενων αυτοματοποιώντας επεναλαμβανόμενες και συνήθεις εργασίες. Εύκολα παραδείγματα υπάρχουν πολλά. Κάποια απο αυτά είναι η αντιγραφή δεδομένων από το ένα σύστημα σε ένα άλλο ανά συγκεκριμένο χρονικό διάστημα, η έκδοση κάποιας αναφοράςμε συγκεκριμένα βήματα ανά συγκεκριμένο χρονικό διάστημα, η αποστολή αυτοματοποιημένων μηνυμάτων, κλήσεων ή email, κ.λπ. Όλα τα παραπάνω παραδείγματα έχουν το κοινό ότι ήταν διαδικασίες στις οποίες τα βήματα ήταν δεδομένα και δεν άλλαζαν πέρα από τους γενικούς κανόνες που μπορεί να υπηρχαν.

Το RPA συνηθίζεται κυρίως στα λογιστικά, οικονομικά τμήματα ή στα τμήματα παραγγελιών. Ο λόγος είναι πως εκεί συνήθως υπάρχουν εργασίες όπως κάποιες αναφορές ή η

δημιουργία πελάτη σε κάποιο σύστημα που είναι όπως λέγεται κατάλληλες για αυτοματοποίηση. Εννοείται υπάρχουν και περιπτώσεις όπως η εταιρία Telefónica Global της Ισπανίας που έχει αυτοματοποιήσει το μεγαλύτερο μέρος όλων των διαδικασιών γραφείου.

2.2 Natural Language Processing

Στόχος της επιστήμης της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing) και της Υπολογιστικής Γλωσσολογίας (Computatiotal Linguistics) είναι η εξαγωγή και η ανάλυση χρήσιμων πληροφοριών από κείμενα φυσικής γλώσσας ή από ομιλίες. Η πιο συνήθης εφαρμογή είναι η ανάλυση συναισθημάτων (sentiment analysis), όπου ο στόχος είναι να προσδιοριστεί αυτόματα η στάση ή η γνώμη για π.χ. ένα προϊόν ή ένα πρόσωπο ή ένα γεγονός, κ.λπ. Το πιο γνωστό παράδειγμα εφαρμογής της ανάλυσης συναισθημάτων είναι η ανάλυση των Tweets στο κοινωνικό μέσο δικτύωσης Twitter.

Όσον αφορά την εργασία μας, τρεις ήταν οι κύριοι τομείς - πυλώνες που απασχολούν για την Επεξεργασια Φυσικής Γλώσσας. Αυτοί οι τομείς είναι ζωτικής σημασίας και θα περιγραφούν εκτενέστερα παρακάτω.

- *Συντακτική Ανάλυση - Syntax Parsing*: Ο προσδιορισμός του συντακτικού δέντρου και των γραμματικών σχέσεων μεταξύ των τμημάτων της πρότασης.
- *Αναφορική Ανάλυση - Anaphora Resolution*: Ο προσδιορισμός των εννοιών που αναφέρονται με αντωνυμίες όπως προσωπικές («εμείς», «αυτός», «αυτό») ή δεικτικές («εκείνος», «αυτή»), κ.λπ.
- *Σημασιολογική Ανάλυση - Semantic Analysis*: Ο προσδιορισμός της έννοιας των λέξεων ή των φράσεων χρησιμοποιώντας τη λεξικολογική βάση δεδομένων WordNet.

2.2.1 Συντακτική Ανάλυση

Στον τομέα της επεξεργασίας κειμένου, ένας στόχος είναι ο αυτοματοποιημένος προσδιορισμός τμημάτων ομιλίας (POS - Part of Speech) και η αναγνώριση της συντακτικής δομής, δηλαδή ποιες λέξεις σχηματίζουν μία φράση και οι γραμματικές σχέσεις μεταξύ λέξεων μέσα σε μια πρόταση. Παραδείγματα τέτοιων συντακτικών αναλυτών είναι ο UC Berkley Parser, ο Stanford Parser ή τα ελεύθερα διαθέσιμα NLP εργαλεία NLTK, OpenNLP, GATE και RASP.

Ο συνταντικός αναλυτής Stanford Parser είναι σε θέση να προσδιορίσει ένα συντακτικό δέντρο στο οποίο περιλαμβάνονται και οι σχέσεις μεταξύ των λέξεων της πρότασης. Επιπλέον κάθε λέξη και φράση φέρουν την κατάλληλη ετικέτα POS. Οι ετικέτες που χρησιμοποιεί ο Stanford Parser είναι οι ίδιες που ορίζονται στο Penn Tree Bank.

Επίσης ο συνταντικός αναλυτής Stanford Parser παράγει και επιστρέφει εξαρτήσεις. Οι εξαρτήσεις αυτές αντικατοπτρίζουν τις γραμματικές σχέσεις μεταξύ των λέξεων.

Παρακάτω θα παρουσιάσουμε ένα παράδειγμα του συνταντικού αναλυτή Stanford Parser με την πρόταση «The vacation request process starts when an employee submits a vacation request via the ERP system.» η οποία είναι από το Σχήμα 1.1 .

Tagging

```
The/DT vacation/NN request/NN process/NN starts/VBZ when/WRB an/DT employee/NN submits/VBZ a/DT vacation/NN request/NN via/IN the/DT ERP/NNP system/NN ./.
```

Σχήμα 2.3: *Stanford Parser Tagging*

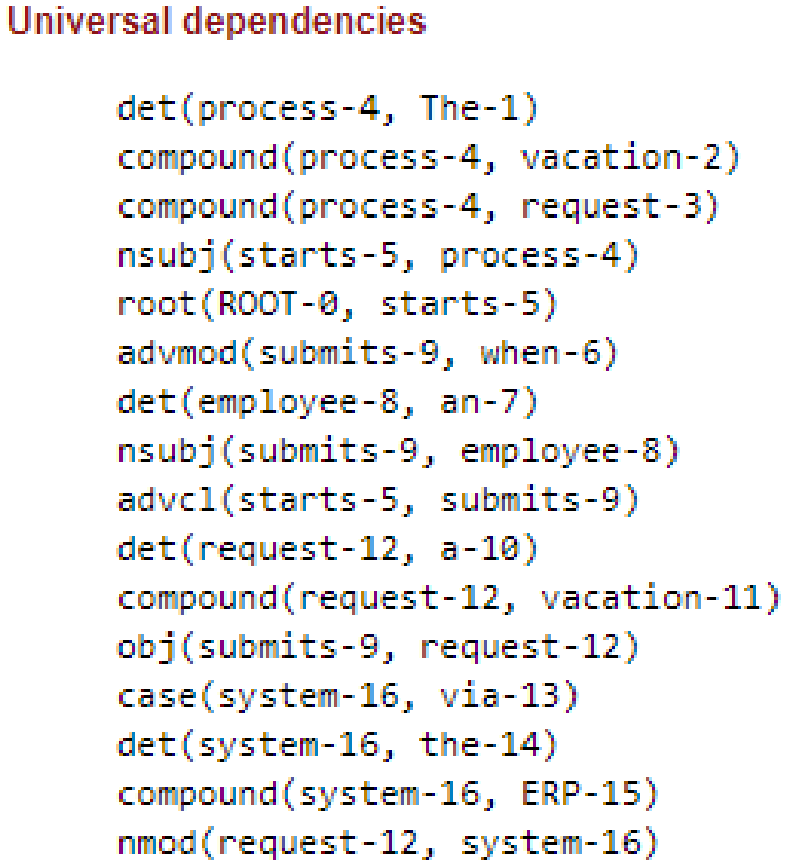
Στην παραπάνω εικόνα φαίνεται πως ο συνταντικός αναλυτής Stanford Parser μπορεί να αναγνωρίσει τι μέρος του λόγου είναι κάθε λέξη.

Parse

```
(ROOT
  (S
    (NP (DT The) (NN vacation) (NN request) (NN process))
    (VP (VBZ starts)
      (SBAR
        (WHADVP (WRB when))
        (S
          (NP (DT an) (NN employee))
          (VP (VBZ submits)
            (NP
              (NP (DT a) (NN vacation) (NN request))
              (PP (IN via)
                (NP (DT the) (NNP ERP) (NN system))))))))
    (. .)))
```

Σχήμα 2.4: *Stanford Parser Parser*

Στην παραπάνω εικόνα φαίνεται πως ο συνταντικός αναλυτής Stanford Parser μπορεί να αναγνωρίσει το συντακτικό δέντρο της πρότασης στο οποίο φαίνονται και οι ετικέτες των λέξεων.



Σχήμα 2.5: *Stanford Parser Dependencies*

Στην παραπάνω εικόνα φαίνεται πως ο συνταντικός αναλυτής Stanford Parser επιστρέφει και τις εξαρτήσεις μεταξύ των λέξεων της πρότασης. Αυτός είναι και ο κυριότερος λόγος που επιλέξαμε να δουλέψουμε με το συγκεκριμένο εργαλείο.

Η εξάρτηση *det* δηλώνει ότι η πρώτη λέξη της πρότασης «The» ορίζει την τέταρτη λέξη «process». Η σχέση *nsubj* με τη σειρά της δείχνει ότι η λέξη «process» είναι το υποκείμενο του ρήματος «starts» που είναι η πέμπτη λέξη. Συνολικά ο συνταντικός αναλυτής Stanford Parser περιλαμβάνει 55 σχέσεις εξαρτήσεων οι οποίες υπάρχουν αναλυτικά στο εγχειρίδιο του. Το πρώτο στοιχείο αυτών των εξαρτήσεων ονομάζεται κυβερνήτης (governor) ενώ το δεύτερο στοιχείο ονομάζεται εξαρτώμενος (dependent).

Οι παραπάνω εικόνες είναι από την online πλατφόρμα του συνταντικού αναλυτή Stanford Parser όπου μπορεί κανείς να κάνει δοκιμές με διάφορες προτάσεις. Το link είναι το εξής *http://nlp.stanford.edu:8080/parser/index.jsp*.

2.2.2 Αναφορική Ανάλυση

Το επόμενο που μας απασχόλησε για την παραγωγή εννοιολογικών μοντέλων από κείμενο ήταν η επίλυση αναφορικών πληροφοριών. Αναφορικές λέξες θεωρούνται οι κτητικές αντωνυμίες (π.χ. «μου», «της», «του»), προσωπικές αντωνυμίες (π.χ. «εγώ», «εσύ», «εκείνος»),

δεικτικές αντωνυμίες («αυτό», «εκείνο»), σχετικές αντωνυμίες («ποιος», «οποίος») ή φράσεις που περιγράφουν το αντικείμενο με διαφορετικό τρόπο (π.χ. ο «Steve Jobs», «ο CEO της Apple»).

Έχουν προταθεί πολλοί τρόποι για την επίλυση αυτού το θέματος. Ο επικρατέστερος πρόκειται για έναν αλγόριθμο ο οποίος εντοπίζοντας την αντωνυμία και αναλύοντας την, σαρώνει το κείμενο για υποψήφιες λέξεις που θα μπορούσαν να είναι αυτές στις οποίες αναφέρεται η αντωνυμία. Για κάθε υποψήφια λέξη υπολογίζεται ένα σκορ το οποίο αποτελείται από:

- Την απόσταση μεταξύ της αναφορικής αντονυμίας και της υποψήφιας λέξης (οι κοντινότερες λέξεις προτιμώνται).
- Το φύλο, τον αριθμό και το αν είναι πρόσωπο ή όχι η υποψήφια λέξη.
- Ο συντακτικός ρόλος της υποψήφιας λέξης στην πρόταση που υπάγεται (τα υποκείμενα προτιμώνται).
- Ο αριθμός των προηγούμενων εμφανίσεων του υποψηφίου (μεγαλύτερος αριθμός προτιμάται).

Αυτή η τεχνική έχει ποσοστό επιτυχίας 84,2% ενώ σε επόμενες έρευνες η μέθοδος έχει εξελιχθεί ακόμα παραπάνω. Παραδείγματα υλοποίησης αναφορικών αλγορίθμων ανάλυσης είναι το GuiTAR Framework7, το BART8 ή το Reconcile framework9.

2.2.3 Σημασιολογική Ανάλυση

Το τρίτο και τελευταίο μέρος που μας απασχόλησε για τη δημιουργία του συστήματος μας, είναι η σημασιολογική ανάλυση. Εκτός από συντακτικό ρόλο, κάθε λέξη σε μια πρόταση έχει και μια συγκεκριμένη έννοια η οποία ονομάζεται σημασιολογία. Μια λέξη μπορεί να έχει πολλαπλές ερμηνείες ανάλογα την πρόταση, το ύφος και την χρήση που της γίνεται. Συνεπώς και ο τομέας αυτός είναι εξίσου σημαντικός με τους προηγούμενους για την καλύτερη ανάλυση ενός κειμένου. Υπάρχουν διάφορα συστήματα που προσπαθούν να συλλάβουν σημασιολογικές σχέσεις. Εμείς χρησιμοποιήσαμε τη λεξικολογική βάση δεδομενων WordNet για την ανάπτυξη της μεθόδου μας.

Το WordNet είναι μια σημασιολογική βάση δεδομένων που αναπτύχθηκε το 1985 στο Πανεπιστήμιο του Princeton. Από τότε, αυξάνεται σταθερά και σήμερα περιέχει περισσότερες από 155.000 μοναδικές λέξεις και περισσότερα από 200.000 ζεύγη λέξεων. Αυτές οι λέξεις οργανώνονται στα λεγόμενα SynSets (σύνολα συνωνύμων). Ένα SynSet περιέχει πολλές λέξεις που έχουν την ίδια έννοια. Επιπλέον, τα SynSets στο WordNet συνδέονται μεταξύ τους μέσω δεικτών διαφόρων τύπων. Επομένως, ένα πρόγραμμα είναι σε θέση να εξαγάγει διαφορετικές σημασιολογικές σχέσεις για μια δεδομένη λέξη. Συγκεκριμένα, οι πιο συνηθισμένες κατηγορίες συνόλων είναι οι παρακάτω:

- Συνώνυμα - Synonyms : Λέξεις οι οποίες έχουν το ίδιο νόημα.
- Ομώνυμα - Homonyms : Λέξεις που γράφονται πανομοιότυπα, αλλά έχουν διαφορετική σημασία.

- Υπερνύμια - Hypernyms : Ονόματα που υπερισχύουν του δεδομένου ουσιαστικού.
- Μερωνύμια - Meronyms : Δομές με σχέση «μερών» (π.χ. αυτοκίνητο - τροχός).
- Αντώνυμα - Antonyms : Χρησιμοποιείται κυρίως για επίθετα και επιρρήματα και περιγράφει το αντίθετο (π.χ. υγρό - ξηρό, ζεστό - κρύο).
- Εμπλοκή - Entailment : Μια επίπτωση που μπορεί να εξαχθεί λογικά (π.χ. διαζύγιο - γάμος).
- Ονομασίες - Nominalizations : Ένα ρήμα, επίθετο ή επίρρημα που χρησιμοποιείται ως ουσιαστικό (π.χ. δημιουργία - δημιουργώ).

Έτσι με το WordNet θα μπορούσαμε πολύ εύκολα για παράδειγμα να διακρίνουμε ένα σύστημα από ένα αντικείμενο ή από έναν ενεργό άνθρωπο. Παραπάνω λεπτομέρειες σχετικά με αυτό υπάρχουν στην ενότητα της υλοποίησης. Από την έκδοση 3.0 το WordNet επίσης ενσωματώνει την τεχνική του word stemming. Ο σκοπός του word stemming είναι η μείωση μιας λέξης στη βασική της μορφή χωρίς προθέματα (prefixes) ή/και καταλήξεις (suffixes). Επομένως μπορούμε να το χρησιμοποιήσουμε για να ομαλοποιήσουμε διαφορετικές αναπαραστάσεις των λέξεων του κειμένου κατά τη διαδικασία της μεθόδου μας.

2.3 Machine Learning

2.3.1 Ορισμός Μηχανικής Μάθησης

Η μηχανική εκμάθηση (ML) είναι το σύνολο των μεθόδων που μπορούν να ανιχνεύουν αυτόματα τα μοτίβα στα δεδομένα, και στη συνέχεια χρησιμοποιούν αυτά τα μοτίβα για να προβλέψουν μελλοντικά δεδομένα, ή να εκτελέσουν άλλα είδη αποφάσεων υπό αβεβαιότητα.

Η μηχανική μάθηση αποτελεί μέρος του κλάδου της πληροφορικής που συνήθως αναφέρεται με τον όρο «Τεχνητή Νοημοσύνη». Πολύ συχνά οι δύο όροι λανθασμένα συγχέονται. Η τεχνητή νοημοσύνη, είναι η επιστήμη και η τεχνολογία κατασκευής ευφυών μηχανών. Πρόκειται λοιπόν για τον γενικότερο κλάδο της πληροφορικής που ασχολείται με την ανάπτυξη ευφυών μηχανών, μιας προσπάθειας δηλαδή να κάνουμε τους υπολογιστές να μιμηθούν την ευφυή συμπεριφορά του ανθρώπου ώστε να εκτελέσουμε εργασίες οι οποίες απαιτούν ανθρωπινή ευφυΐα όπως αναγνώριση λόγου, οπτική αντίληψη και λήψη αποφάσεων. Η βασική διαφορά της μηχανικής μάθησης συγκριτικά με τους γράφους γνώσης και άλλα πεδία της τεχνητής νοημοσύνης, είναι η ικανότητα της να τροποποιείται και να προσαρμόζεται στα νέα δεδομένα που δέχεται. Μέσω αυτής της ικανότητας της μηχανικής μάθησης, οι επιστήμονες μπορούν πλέον τροφοδοτώντας με δεδομένα να οδηγήσουν τους υπολογιστές σε νέα γνώση την οποία ούτε οι ίδιοι κατέχουν. Στις μέρες μας όπου πληθώρα δεδομένων είναι διαθέσιμη και υπάρχει ολοένα και μεγαλύτερη ψηφιοποίηση αυτών των δεδομένων, η μηχανική μάθηση αποτελεί ένα πολύ ισχυρό εργαλείο για την επεξεργασία και αξιοποίηση αυτού του όγκου πληροφορίας, καθώς και για τη δημιουργία νέας γνώσης στην οποία δεν είχαμε πρόσβαση πρωτύτερα.

Οι αλγόριθμοι μηχανικής μάθησης δημιουργούν ένα μοντέλο βασισμένο σε δείγματα δεδομένων, γνωστά ως «train data», προκειμένου να δημιουργούν προβλέψεις ή να παίρνουν αποφάσεις χωρίς να έχουν προγραμματιστεί ρητά να το κάνουν.

Οι προσεγγίσεις μηχανικής εκμάθησης χωρίζονται σε τρεις ευρείες κατηγορίες, ανάλογα με τη φύση του «σήματος» ή «feedback» που είναι διαθέσιμο στο σύστημα μάθησης:

- Supervised learning : Στην κατηγορία αυτή υπάρχουν παραδείγματα εισόδων μαζί με τις επιθυμητές εξόδους τους, που δίνονται από έναν «δάσκαλο», και ο στόχος είναι ο αλγόριθμος να μάθει έναν γενικό κανόνα που χαρτογραφεί τις εισόδους στις εξόδους.

- Unsupervised learning : Στην κατηγορία αυτή δεν δίνονται ετικέτες στον αλγόριθμο εκμάθησης, αφήνοντάς τον από μόνος του να βρει δομή ή μοτίβα στην είσοδό του.

- Reinforcement learning : Στην κατηγορία αυτή ένα πρόγραμμα υπολογιστή αλληλεπιδρά με ένα δυναμικό περιβάλλον στο οποίο πρέπει να εκτελεί έναν συγκεκριμένο στόχο (όπως οδήγηση οχήματος ή παιχνίδι εναντίον αντιπάλου). Καθώς πλοηγεί στον προβληματικό του χώρο, το πρόγραμμα παρέχει ανατροφοδότηση ανάλογη με τις ανταμοιβές, την οποία προσπαθεί να μεγιστοποιήσει.

2.3.2 Ταξινόμηση

Ταξινόμηση (Classification) είναι η διαδικασία πρόβλεψης της κατηγορίας συγκεκριμένων δεδομένων. Κλάσεις ονομάζονται οι στόχοι / ετικέτες ή κατηγορίες. Η ταξινόμηση της προβλεπτικής μοντελοποίησης είναι η προσέγγιση της συνάρτησης χαρτογράφησης (f) από τις μεταβλητές εισόδου (X) στις διακριτές μεταβλητές εξόδου (y).

Για παράδειγμα, η ανίχνευση ανεπιθύμητων μηνυμάτων (spam detection) σε παρόχους υπηρεσιών email μπορεί να αναγνωριστεί ως πρόβλημα ταξινόμησης. Αυτή είναι η δυαδική ταξινόμηση, καθώς υπάρχουν μόνο 2 τάξεις, τα ανεπιθύμητα και τα όχι ανεπιθύμητα email. Ένας ταξινομητής χρησιμοποιεί ορισμένα δεδομένα εκπαίδευσης για να κατανοήσει πώς οι δεδομένες μεταβλητές εισόδου σχετίζονται με την κλάση. Σε αυτήν την περίπτωση, τα γνωστά ανεπιθύμητα μηνύματα και τα μηνύματα spam δεν πρέπει να χρησιμοποιούνται ως εκπαιδευτικά δεδομένα. Όταν ο ταξινομητής εκπαιδευτεί με ακρίβεια, μπορεί να χρησιμοποιηθεί για τον εντοπισμό ενός άγνωστου email.

Η ταξινόμηση ανήκει στην κατηγορία της εποπτευόμενης - επιβλεπόμενης μάθησης (supervised learning όπου τα δεδομένα εισαγωγής παρείχαν επίσης τους στόχους. Υπάρχουν πολλές εφαρμογές κατάταξης σε πολλούς τομείς, όπως στην έγκριση πίστωσης, στην ιατρική διάγνωση, στοχευόμενο μάρκετινγκ κ.λπ.

2.3.3 Αλγόριθμοι Ταξινόμησης

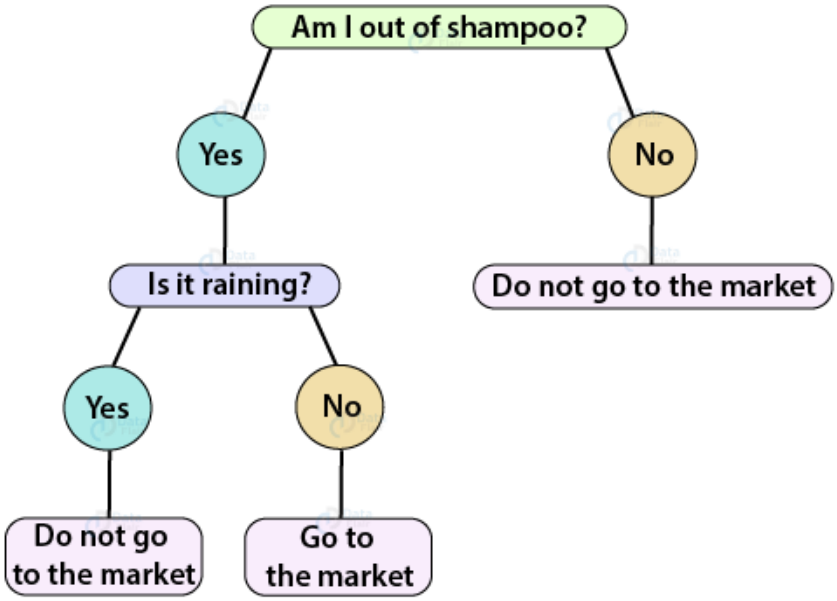
Υπάρχουν πολλοί αλγόριθμοι ταξινόμησης, αλλά δεν είναι δυνατόν να εξαχθεί το συμπέρασμα ποιος είναι ανώτερος από άλλον. Εξαρτάται από την εφαρμογή και τη φύση των διαθέσιμων συνόλων δεδομένων. Παρακάτω θα παρουσιαστούν κάποιοι από τους πιο γνωστούς, από τους οποίους κάποιους τους δοκιμάσαμε στο δικό μας σύστημα.

Decision Tree

Το δέντρο αποφάσεων δημιουργεί μοντέλα ταξινόμησης ή παλινδρόμησης με τη μορφή μιας δομής δέντρου. Χρησιμοποιεί ένα σύνολο κανόνων if-then το οποίο είναι αμοιβαία αποκλειστικό και εξαντλητικό για ταξινόμηση. Οι κανόνες μαθαίνονται διαδοχικά χρησιμοποιώντας τα δεδομένα εκπαίδευσης ένα ένα κάθε φορά. Κάθε φορά που μαθαίνεται ένας κανόνας, οι πλειάδες που καλύπτονται από τους κανόνες αφαιρούνται. Αυτή η διαδικασία συνεχίζεται στο σετ προπόνησης έως ότου ικανοποιηθεί ένας όρος τερματισμού.

Το δέντρο είναι κατασκευασμένο από πάνω προς τα κάτω με αναδρομικό διαίρει και βασίλευε τρόπο. Όλα τα χαρακτηριστικά πρέπει να είναι κατηγορηματικά. Διαφορετικά, θα πρέπει να διακριθούν εκ των προτέρων. Τα χαρακτηριστικά στην κορυφή του δέντρου έχουν μεγαλύτερη επίδραση στην ταξινόμηση και ταυτοποιούνται χρησιμοποιώντας την έννοια της απόκτησης πληροφοριών.

Ένα δέντρο αποφάσεων μπορεί εύκολα να φτάσει στην υπερμοντελοποίηση (overfitting) δημιουργώντας πάρα πολλά κλαδιά και μπορεί να φέρει ανωμαλίες λόγω θορύβου ή ακραίων τιμών. Ένα τέτοιο μοντέλο έχει πολύ κακή απόδοση στα καινούργια δεδομένα, παρόλο που δίνει εντυπωσιακή απόδοση στα δεδομένα εκπαίδευσης. Αυτό μπορεί να αποφευχθεί με προ-κλάδεμα που σταματά την κατασκευή δέντρων νωρίς ή μετά το κλάδεμα που αφαιρεί κλαδιά από το πλήρως αναπτυγμένο δέντρο.



Σχήμα 2.6: *Decision Tree Example*

Random Forest

Ο αλγόριθμος τυχαίων δασών είναι μια επέκταση του δέντρου αποφάσεων, με την έννοια ότι, πρώτα δημιουργεί δέντρα αποφάσεων σε πραγματικό κόσμο μερικών αξόνων με δεδομένα

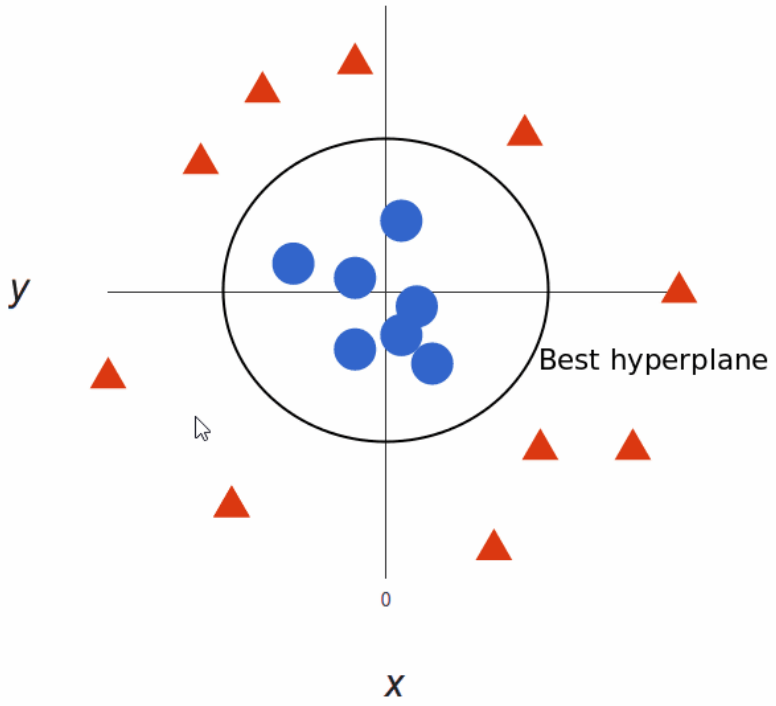
εκπαίδευσης και, στη συνέχεια, ταιριάζει τα νέα δεδομένα σε ένα από τα δέντρα ως «τυχαίο δάσος».

Ουσιαστικά, ο μέσος όρος των δεδομένων συνδέεται με το πλησιέστερο δέντρο στην κλίμακα δεδομένων. Τα τυχαία δασικά μοντέλα είναι χρήσιμα καθώς διορθώνουν το πρόβλημα του «εξαναγκασμού» σημείων δεδομένων των decision trees σε μια κατηγορία χωρίς λόγο.

Support Vector Machines

Ένα μηχάνημα φορέα υποστήριξης (SVM) χρησιμοποιεί αλγόριθμους για να εκπαιδεύσει και να ταξινομήσει δεδομένα εντός βαθμών πολικότητας, μεταφέροντάς τα σε βαθμό πέρα από την πρόβλεψη X / Y.

Το Support Vector Machine δημιουργεί ένα hyperplane που διαχωρίζει καλύτερα τις ετικέτες. Σε δύο διαστάσεις αυτό είναι απλά μια γραμμή. Σε δυαδική ταξινόμηση, οτιδήποτε από τη μία πλευρά της γραμμής είναι η μια κατηγορία και οτιδήποτε από την άλλη πλευρά είναι η άλλη κατηγορία. Στην ανάλυση συναισθημάτων, για παράδειγμα, η μία κατηγορία θα ήταν το θετικό και η άλλη το αρνητικό. Χρησιμοποιώντας το SVM, όσο πιο περίπλοκα είναι τα δεδομένα, τόσο ακριβέστερη θα γίνει η πρόβλεψη.



Σχήμα 2.7: *Support Vector Machine Example*

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 3

Ανάλυση και υλοποίηση

Στο κεφάλαιο αυτό παρουσιάζεται η μελέτη που έγινε για την υλοποίηση του συστήματος. Έτσι, αρχικά προηγήθηκε έρευνα σχετικά με τα ζητήματα που υπήρχαν για την ανάλυση φυσικής γλώσσας (natural language processing) όσον αφορά ένα γραπτό κείμενο. Με βάση τα ευρήματα σχηματίστηκαν κάποιες υποθέσεις στις οποίες βασίστηκε όλη η στρατηγική μας η οποία και επεξηγείται παρακάτω. Τέλος, κάθε στάδιο της μεθόδου μας και η λειτουργία του θα περιγράφονται λεπτομερώς χρησιμοποιώντας και ψευδοκώδικα.

3.1 Κατηγοριοποίηση Ζητημάτων

Το πιο σημαντικό ζήτημα που αντιμετωπίζουμε όταν προσπαθούμε να δημιουργήσουμε ένα σύστημα για την κατανόηση της φυσικής γλώσσας, είναι η πολυπλοκότητά της. Η φυσική γλώσσα θεωρείται πολύπλοκη υπό πολλές έννοιες, ειδικά όταν πρόκειται για την κατανόηση της από έναν υπολογιστή. Σύμφωνα με τη βιβλιογραφία για την επιτυχή ανάλυση της φυσικής γλώσσας εντοπίζονται τρεις κύριες κατηγορίες ζητημάτων προς επίλυση. Οι κατηγορίες αυτές είναι οι εξής:

- **Σημασιολογία - Σύνταξη**: Η αναντιστοιχία μεταξύ της σημασιολογικής και της συντακτικής ανάλυσης του κειμένου.
- **Συνεκτικότητα**: Ο τρόπος αναζήτησης κατάλληλων φράσεων από μια πρόταση.
- **Συνάφεια**: Τμήματα του κειμένου ή της πρότασης που μπορεί να μην έχουν σχέση με το γενικό πλαίσιο του κειμένου.

Κάθε μία από τις παραπάνω κατηγορίες θα την απλοποιήσουμε σε επιμέρους κομμάτια - ζητήματα και θα παραθέσουμε παραδείγματα για την πλήρη κατανόηση τους. Τα παραδείγματα θα είναι από το dataset με το οποίο δουλέψαμε για να δείξουμε τον τρόπο που εμφανίζονται.

3.1.1 Σημασιολογία - Σύνταξη

Πίνακας 3.1: *Πίνακας με Ζητήματα Σημασιολογίας - Σύνταξης της Φυσικής Γλώσσας*

Σημασιολογία - Σύνταξη
1. Ενεργητική - Παθητική Φωνή
2. Αναδιατύπωση Φράσης
3. Υπονοούμενη Φράση

Η έκφραση μίας σημασιολογικής έννοιας σε μια γλώσσα μπορεί να γίνει με διαφορετικά συντακτικά πρότυπα, άρα η συντακτική δομή ενός όρου δεν συνδέεται απαραίτητα άμεσα με την σημασιολογία του. Δηλαδή η σχέση σύνταξης και εννοιολογίας δεν είναι 1-1. Επομένως είναι σημαντικό από κάθε πρόταση να μπορούν να καθοριστούν οι ρόλοι των λέξεων, δηλαδή ποιο είναι το Υποκείμενο - Ρήμα - Αντικείμενο.

Για παράδειγμα, ακολουθούν δύο προτάσεις από τις οποίες η πρώτη είναι σε ενεργητική και η δεύτερη σε παθητική φωνή:

•Ο Γιώργος τραυμάτισε το χέρι του κάνοντας skateboard.
•Το χέρι του Γιώργου τραυματίστηκε κάνοντας skateboard.

Στο πρώτο παράδειγμα «ο Γιώργος» είναι το κύριο θέμα της πρότασης και το υποκείμενο. Στο δεύτερο παράδειγμα το υποκείμενο είναι «το χέρι» και «ο Γιώργος» αναφέρεται μόνο σε μια εμπρόθετη φράση, η οποία θα μπορούσε επίσης να παραλειφθεί, μολονότι το σημασιολογικό μοτίβο των δύο προτάσεων είναι το ίδιο.

Ένα άλλο ζήτημα είναι η δυνατότητα ανταλλαγής ορισμένων λέξεων ή φράσεων χωρίς αλλαγή της έννοια μιας πρότασης, όπως σε αυτό το παράδειγμα:

•Ο Γιώργος πρέπει να πάει στο νοσοκομείο αν το χέρι του είναι σπασμένο.
•Σε περίπωση που το χέρι του Γιώργου είναι σπασμένο, πρέπει να πάει στο νοσοκομείο.

Τέλος, ένα ακόμη ζήτημα είναι η αναγνώριση της ρητορικής χροιάς του κειμένου και ο διαχωρισμός των φράσεων αυτών. Προβλήματα προκύπτουν όταν το κείμενο περιέχει έμμεσες σχέσεις λόγου, όπως σε αυτήν την πρόταση:

•Για νέους ασθενείς, δημιουργείται ένα αρχείο ασθενούς.

Προφανώς, υπάρχει ένας όρος εδώ. Το αρχείο ασθενούς πρέπει να δημιουργηθεί μόνο όταν ασχολούμαστε με έναν νέο ασθενή. Σε περίπτωση γνωστού ασθενούς, μπορούμε να παραλείψουμε αυτό το βήμα. Όμως, για να αναγνωρίσουμε αυτήν την κατάσταση χρειάζεται εκτενής ανάλυση. Με μια εμπεριστατωμένη σημασιολογική ανάλυση δεν είμαστε σε θέση να εντοπίσουμε αυτές τις συνθήκες χρησιμοποιώντας μόνο τα συντακτικά μέσα που παρέχονται από τα εργαλεία που έχουν προαναφερθεί και που χρησιμοποιήσαμε.

3.1.2 Συνεκτικότητα

Πίνακας 3.2: *Πίνακας με Ζητήματα Συνεκτικότητας της Φυσικής Γλώσσας*

Συνεκτικότητα
1. Σύνθετες Προτάσεις
2. Αναφορικές Προτάσεις

Η συνεκτικότητα ασχολείται με το ερώτημα του ποια μέρη της πρότασης αντιστοιχούν σε δραστηριότητα - εργασία. Είναι πιθανό πως ορισμένες προτάσεις περιλαμβάνουν μια ξεκάθαρη και απλή περιγραφή δραστηριότητας. Ωστόσο στο σύνολο των δεδομένων συναντάμε κυρίως σύνθετες προτάσεις. Τα παρακάτω δύο παραδείγματα είναι από το dataset μας και θα βοηθήσουν στην κατανόηση των σύνθετων προτάσεων.

• Sometimes, we buy details for cold calls, sometimes, our marketing staff participates in exhibitions and sometimes, you just happen to know somebody, who is interested in the product.
• The GO (Grid operator) or the MPON (Metering point operator new) confirms the invoice with payment advice to the MPOO (Metering point operator old) or the MSPO (Metering service provider old), or the GO (Grid operator) or the MPON (Metering point operator new) rejects the invoice of the MPOO (Metering point operator old) or the MSPO (Metering service provider old).

Αναλύοντας τις προτάσεις προκείπτει πως η πρώτη αντιστοιχεί σε τρεις διαφορετικές δραστηριότητες οι οποίες είναι η αγορά λεπτομερειών (buy details for cold calls), η συμμετοχή σε εκθέσεις (participates in exhibitions) και το να τύχει να γνωρίζουν κάποιον (happen to know somebody), ενώ η δεύτερη πρόταση αντιστοιχεί σε τέσσερις δραστηριότητες οι οποίες είναι η επιβεβαίωση τιμολογίου από τον GO και τον MPON (confirms the invoice) και η απόρριψη τιμολογίου από τον GO και τον MPON (rejects the invoice).

Για να καταφέρουμε εμείς να αναγνωρίσουμε τις υποπροτάσεις χρησιμοποιήσαμε την μέθοδο **ClauseIE** η οποία εξάγει υποπροτάσεις σε μορφή Υποκείμενο - Ρήμα - Αντικείμενο. Παραπάνω πληροφορίες θα δωθούν πιο κάτω στην ενότητα αυτή.

Ένα άλλο ζήτημα είναι ότι μια πρόταση μπορεί να περιλαμβάνει αρκετές υποπροτάσεις οι οποίες όμως να μην αντιπροσωπεύουν δραστηριότητες ή ενέργειες, να μην παίζουν ρόλο δηλαδή στον βαθμό αυτοματοποίησης όλης της διαδικασίας. Για παράδειγμα:

•If the treasurer accepts the expenses for processing, the report moves to an automatic activity that links to a payment system.

Η πρώτη υποπρόταση («If the treasurer accepts the expenses for processing») αποτελεί μια προυπόθεση για να γίνει μια ενέργεια ή να μην γίνει, δηλαδή εάν ο ταμίας αποδεκτεί τα έξοδα τότε θα γίνει το επόμενο βήμα αλλιώς όχι. Αντίστοιχα η τελευταία υποπρόταση («that

links to a payment system») αποτελεί μια περαιτέρω επεξήγηση της «αυτόματης δραστηριότητας». Κάποιες από τις πληροφορίες λοιπόν μιας πρότασης μπορεί να είναι σημαντικές κάποιες άλλες όχι. Στην δική μας περίπτωση έχοντας χωρίσει ήδη κάθε πρόταση σε υποπροτάσεις μπορέσαμε να βρούμε ποιες ήταν απαραίτητες με βάση το dataset που είχαμε. Περισσότερες λεπτομέρειες θα δωθούν παρακάτω.

3.1.3 Συνάφεια

Πίνακας 3.3: *Πίνακας με Ζητήματα Συνάφειας της Φυσικής Γλώσσας*

Συνάφεια
1. Σχετικότητα Πρότασης
2. Μετα-Προτάσεις

Παραπάνω θέσαμε το πρόβλημα για την εύρεση υποπροτάσεων που αποτελούν μέρος κάποιας ενέργειας ή δραστηριότητας. Παρόμοιο και εξίσου σημαντικό είναι το ζήτημα αν η σχετική πρόταση που αναλύεται είναι σχετική και συναφής με το συνολικό θέμα. Για παράδειγμα θα μπορούσε μια πρόταση να παρουσιάζει ένα παράδειγμα για παραπάνω κατανόηση, και άρα οι δρασηριότητες αυτές που υπάρχουν μέσα στο παράδειγμα να μην αποτελούν μέρος της συνολικής διαδικασίας.

Ένα άλλο ζήτημα συνάφειας που μπορεί να παρατηρηθεί είναι ότι μέσα στις περιγραφές διαδικασιών υπάρχουν πληροφορίες σχετικά για το επίπεδο της περιγραφής. Δηλαδή εδώ συμπεριλαμβάνονται τα βήματα που πρέπει να διεξαχθούν, η αλλαγή σταδίου, η εκκίνηση ή η τερματισμός ενός βήματος, κ.λπ. Παραθέτουμε κάποια παραδείγματα από το dataset :

•Η διαδικασία ξεκινά όταν ένας πελάτης υποβάλλει [...].
•Εάν ο σχεδιασμός αποτύχει στη δοκιμή, τότε επιστρέφει στην πρώτη δραστηριότητα.
•Μετά την επιβεβαίωση της πληρωμής, η διαδικασία τελειώνει.

Από τις παραπάνω προτάσεις είναι φανερό πως υποπροτάσεις του τύπου «η διαδικασία ξεκινά» ή «η διαδικασία τελειώνει» δεν αποτελούν επιθυμητό υλικό για επεξεργασία. Για το φιλτράρισμα των υποπροτάσεων αυτών χρησιμοποιήσαμε το dataset για το οποίο είχαμε δεδομένα σχετικά με την αυτοματοποίηση του. Περισσότερες λεπτομέρειες θα δωθούν παρακάτω.

3.1.4 Στρατηγική Λύσης

Είναι φανερό από όλα τα παραπάνω ζητήματα που υπάρχουν, πως η αυτόματη επεξεργασία κειμένου με στόχο την έυρεση βαθμού αυτοματισμού δεν μπορεί να κατασκευαστεί εύκολα. Αφού έχουμε θίξει τη σημασία του προβλήματος που υπάρχει όσον αφορά την χειροκίνητη ανάλυση και έρευση επιχειρησιακών διαδικασιων για αυτοματοποίηση, στόχος της προτεινόμενης προσέγγισης είναι να δώσει μια αρχική ιδέα για το πως μπορεί να αντιμετωπιστεί αυτό το ζήτημα. Σαφώς ότι βαθμό ακρίβειας και να έχει η λύση, σκοπός δεν είναι να

αντικαταστήσουμε τον άνθρωπο αναλυτή, αλλά να γίνει μια αρχική προσέγγιση, και να αποκλειστεί ένας μεγάλος όγκος διαδικασιών, ώστε να μπορεί μετά να κάνει πιο αποτελεσματικά τη δουλειά του.

Η ανάλυση του κειμένου που προτείνουμε και έχουμε υλοποιήσει έχει βασιστεί σε κάποια από τα ζητήματα που έχουν αναφερθεί παραπάνω και στην δομή του dataset που είχαμε στα χέρια μας. Συγκεκριμένα στο dataset αναφέρονται μεμονομένες δραστηριότητες από κάθε κείμενο οι οποίες έχουν ταξινομηθεί με βάση τον βαθμό αυτοματοποίησης τους σε τρεις κατηγορίες, σε (1) χειροκίνητη εργασία (manual task) , (2) σε εργασία χρήστη, δηλαδή αυτή η οποία γίνεται με αλληλεπίδραση ενός ανθρώπου με ένα πληροφοριακό σύστημα(user task) ή (3) σε αυτοματοποιημένη εργασία (automated task). Έπειτα από αυτόν τον διαχωρισμό μπορεί να βγει και ένα συνολικό συμπέρασμα για κάθε κείμενο. Για τον λόγο αυτό στην προσέγγιση μας οι προτάσεις έχουν αναλυθεί μεμονομένα ώστε να βγει ένα συμπέρασμα ανά δραστηριότητα που αναφέρεται.

Για να το επιτύχουμε αυτό έχουμε χρησιμοποιήσει εργαλεία όπως ο Stanford Syntax Parser, το WordNet, το ClausIE και το Fuzzy. Η ανάλυση θα διεξαχθεί σε διαφορετικά στάδια, που είναι ανεξάρτητα το ένα από το άλλο.

Είναι πολύ σημαντικό να αναφέρουμε ότι η προσέγγιση μας βασίζεται σε τρεις βασικές παραδοχές που κάναμε με σκοπό να περιορίσουμε τις απαιτούμενες αναλύσεις. Αυτές είναι:

- Το κείμενο περιγράφει πραγματικά μια διαδικασία και όχι κάτι άλλο.
- Το κείμενο δεν περιέχει ερωτήσεις.
- Οι μετα-πληροφορίες («η επόμενη εργασία είναι ...», «τότε η διαδικασία ...») αν υπάρχουν θα αγνοούνται.

Ενώ η πρώτη υπόθεση είναι αυτονόητη, η δεύτερη χρησιμοποιείται για να περιορίσει το πεδίο εφαρμογής εξαρτήσεων που πρέπει να ληφθούν υπόψη, καθώς υπάρχουν ειδικές ετικέτες για ερωτήσεις. Επιπλέον, οι ερωτήσεις αναμένεται να είναι πολύ σπάνιες γενικά σε επιχειρηματικές διαδκασίες. Στο dataset μας δεν περιλαμβάνονται ερωτήσεις και, επομένως, αυτός ο περιορισμός φαίνεται να ήταν πρακτικός. Ενώ θα παρουσιάσουμε μηχανισμούς για τον εντοπισμό και τη μείωση των μετα-πληροφοριών, γενικά προτιμάται ένα άμεσο στυλ γραφής.

Ωστόσο, η παραβίαση μιας υπόθεσης δεν καθιστά την ανάλυση του κειμένου αδύνατη, αλλά απλώς θα μειώσει την ποιότητα του αποτέλεσματος. Ωστόσο θεωρούμε πως ικανοποιητικά αποτελέσματα μπορούν να επιτευχθούν παρά τους περιορισμούς.

3.2 Ανάλυση Προτάσεων

Σε αυτήν την ενότητα θα περιγράψουμε την ανάλυση που έγινε σε επίπεδο προτάσεων χρησιμοποιώντας ψευδοκώδικα. Η διαδικασία εξαγωγής αποτελείται από διάφορα στάδια.

3.2.1 Αποσύνθεση προτάσεων από το κείμενο

Το πρώτο στάδιο επεξεργασίας κειμένου ξεκινά με την αφαίρεση συμβόλων, πολλαπλών κενών ή αλλαγές γραμμών και κομμάτων. Ουσιαστικά κρατάμε όλους τους χαρακτήρες και

τα νούμερα, τις τελείες και το μοναδικό σύμβολο που χρειάζεται, την πάυλα («-»). Όπως γνωρίζουμε, οι υπολογιστές δεν είναι τόσο καλοί στην κατανόηση της φυσικής γλώσσας. Καθαρίζοντας το κείμενο καταφέρνουμε μια πιο αποτελεσματική ανάλυση χωρίς να πρέπει να λάβουμε υπόψιν όλες τις διαφορετικές περιπτώσεις συμβόλων που μπορεί να υπάρχουν, καθώς με κάθε σύμβολο μπορεί να αλλάζει και η ερμηνεία της πρότασης.

Το επόμενο στάδιο είναι η διάσπαση του κειμένου σε μεμονωμένες προτάσεις. Η πρόκληση εδώ είναι να διακρίνουμε μια τελεία που χρησιμοποιείται για μια συντομογραφία (π.χ. M.Sc.) από μια τελεία που σηματοδοτεί το τέλος μιας πρότασης. Αυτή η δυσκολία μπορεί να αντιμετωπιστεί από τον προεπεξεργαστή εγγράφων που περιλαμβάνεται στη Βιβλιοθήκη NLTK. Αυτό το εργαλείο (tokeniser) διαιρεί ένα κείμενο σε μια λίστα προτάσεων χρησιμοποιώντας έναν αλγόριθμο χωρίς επίβλεψη για να δημιουργήσει ένα μοντέλο για λέξεις συντομογραφίας, εκφράσεων και λέξεις που ξεκινούν προτάσεις. Κανονικά πρέπει να εκπαιδευτεί σε μια μεγάλη συλλογή απλού κειμένου στη γλώσσα προορισμού πριν μπορέσει να χρησιμοποιηθεί. Το πακέτο δεδομένων NLTK περιλαμβάνει όμως ένα προ-εκπαιδευμένο Punkt tokenizer για τα Αγγλικά. Έτσι μπορούμε για κάθε κείμενο να βρούμε τις προτάσεις από τις οποίες αποτελείται. Προφανώς, σε περιπτώσεις που μπορεί να λείπει κάποια τελεία, ο επεξεργαστής αυτός δεν θα καταφέρει να αναγνωρίσει κάποιο λάθος και θα επιστρέψει λάθος εκτίμηση προτάσεων.

Μία ακόμη ενέργεια που έγινε πριν ακόμη και από την αφαίρεση συμβόλων αφορά τα κείμενα από το set Federal Network Agency Enatcment. Σε αυτά τα κείμενα υπήρχαν πολλές συντομογραφίες ειδικα στα υποκείμενα που αναφέρονταν σε κάποιον ρόλο εργαζομένου, που θα δυσκόλευαν την ανάλυση του κειμένου. Για αυτόν τον λόγο αντικαταστήθηκαν αυτές οι συντομογραφίες με ολόκληρο τον τίτλο του εργαζομένου. Η λίστα με τις λέξεις αυτές βρίσκετα στο παράρτημα στο τέλος της διπλωματικής.

Τέλος, έπειτα από τα παραπάνω βήματα, έγινε μετατροπή κάθε κειμένου σε πεζούς χαρακτήρες και αφαιρέθηκαν όλοι ο κεφαλαίοι. Η μετατροπή όλων των γραμμάτων σε πεζά βοηθάει την διαδικασία του preprocessing που χρειάζεται ειδικότερα σε μετέπειτα στάδιο όταν θα γίνει το parsing των προτάσεων.

Αλγοριθμος 3.1: *Αποσύνθεση προτάσεων*

Require One or more texts
Result List of sentences
if NameofDocument is from Federal Network Agency Enatcment Dataset **then**
    ReplaceActors(Abbriviations, CompleteNames)
end if
for all the texts **do**
    Text = Remove [^A-Za-z0-9 ,-] and ReplaceWith " "
    Text = SplitWords
    Text = JoinAllWords ▷ Για να μην υπάρχει παραπάνω από ένα κενό
    Text.toLower
end for
return List of Sentences

3.2.2 Απλοποίηση προτάσεων

Έχοντας χωρίσει το κείμενο σε προτάσεις, επόμενο βήμα είναι η διάσπαση των προτάσεων σε υποπροτάσεις, εάν και εφόσον χειάζεται. Σκοπός μας είναι η εύρεση δραστηριοτήτων και ενεργειών για τις οποίες θα βγάλουμε ένα συμπέρασμα σχετικά με το επίπεδο αυτοματοποίησης τους. Όπως αναφέρθηκε και παραπάνω, είναι πολύ σύνηθες μια πρόταση να είναι σύνθετη, δηλαδή να περιλαμβάνει παραπάνω από μία διαφορετικές δραστηριότητες. Για την λύση του προβήματος αυτού χρησιμοποιήσαμε το εργαλείο Stanford Parser και πιο συγκεκριμένα το ClausIE το οποίο εκκμεταλλεύεται το Stanford Parser.

Το ClausIE είναι μία προσέγγιση που βασίζεται σε όρους (clauses) για την εξαγωγή πληροφοριών. Ο λόγος που διαφέρει ουσιαστικά από τις υπόλοιπες προσεγγίσεις είναι ότι διαχωρίζει την ανίχνευση «χρήσιμων» πληροφοριών που εκφράζονται σε μια πρόταση. Πιο αναλυτικά, το ClausIE εκκμεταλλεύεται τις γραμματικές σχέσεις που υπάρχουν σε μια πρόταση και έπειτα τις συντακτικές για να προσδιορίσει τον τύπο κάθε λέξης σε συνδυασμό με ένα μικρό σετ λεξικολογικής βάσης. Με βάση τις πληροφορίες αυτές το ClausIE είναι σε θέση να παράγει υποπροτάσεις υψηλής ακρίβειας. Οι υποπροτάσεις αυτές επιστρέφονται σε μορφή Υποκείμενου - Ρήμα - Αντικειμένου (Subject - Verb - Object / SVO). Το ClausIE λειτουργεί ανά πρόταση χωρίς καμία παραπάνω επεξεργασία και δεν απαιτεί δεδομένα εκπαίδευσης.

Για παράδειγμα ας πάρουμε την πρόταση:

•A. Einstein, who was born in Ulm, has won the Nobel Prize.

Σκοπός του ClausIE είναι να εξάγει τριπλέτες (triples) από μια πρόταση, δηλαδή εδώ θα επιστρέψει ("A. Einstein", "has won", "the Nobel Prize") και ("A. Einstein", "was born", "in Ulm"). Καμία περαιτέρω ανάλυση σημασιολογική δεν γίνεται. Κάθε τριπλέτα όπως αναφέραμε παραπάνω θα αποτελείται από το κύριο θέμα ή υποκείμενο ("A. Einstein"), μια σχεσιακή φράση ("has won") και κανένα, ένα ή παραπάνω επιχειρήματα ("the Nobel Prize").

Ως αποτέλεσμα, έπειτα από την χρήση του επεξεργαστή αυτού, έχουμε για κάθε πρόταση ένα σύνολο υποπροτάσεων. Αν μια πρόταση είναι απλή εξ αρχής τότε το ClausIE θα επιστρέψει μια μόνο τριπλέτα. Προφανώς οι υποπροτάσεις θα έχουν ότι πληροφορία υπάρχει μέσα στην αρχική πρόταση ασχέτως άμα πρόκειται για δραστηριότητα ή όχι και γενικά για επιθυμητή εννοιολογικά έκφραση ή όχι. Ο διαχωρισμός θα γίνει αργότερα.

Παρακάτω θα εστιάσουμε στον τρόπο λειτουργίας του ClausIE και έπειτα θα δώσουμε κάποια παραδείγματα αποτελεσμάτων του.

Το ClausIE πήρε το όνομα του από τον αγγλικό όρο clause. Το clause είναι το μέρος μιας πρότασης που εκφράζει κάποια συνεκτική πληροφορία. Αποτελειται συνήθως από ένα υποκείμενο (Subject - S), ένα ρήμα (Verb - V), ένα έμμεσο ή άμεσο αντικείμενο (Object - O), ένα κατηγορούμενο (Complement - C) και έναν ή περισσότερους προσδιορισμούς (Adverbial). Στην αγγλική γλώσσα δεν εμφανίζονται όλοι οι συνδυασμοί αυτών των συστατικών μερών μιας πρότασης μαζί. Συγκεκριμένα όταν ταξινομούμε clauses σύμφωνα με τη γρμματική λειτουργία των μερών τους, βρίσκουμε εφτά διαφορετικά είδη, όπως παρουσιάζονται και στον παρακάνω πίνακα.

Ο όρος clause βασίζεται στο ότι είναι η ελάχιστη μονάδα πληροφορίας η οποία όμως είναι

συνεκτική. Διαισθητικά αυτό σημαίνει ότι αν αφαιρεθεί ένα συστατικό ενός clause που είναι μέρος του, τότε αυτό που θα απομείνει δεν θα έχει σημασιολογικό νόημα (ή αν έχει τότε έχει αλλάξει η έννοια του ρήματος).

Οι εφτά τύποι clauses, που ουσιαστικά εξαρτώνται από το ρήμα είναι οι εξής:

- Intransitive (Αμετάβατο): Εάν το ρήμα δεν προσδιορίζεται από κάτι.
- Extended-copular (Εκτεταμένο κυκλικό): Το ρήμα συνδέει το υποκείμενο με κάποιο προσδιορισμό.
- Copular (Κυκλικό): Το ρήμα συνδέει το υποκείμενο με κάποιο κατηγορούμενο.
- Monotransitive (Μονομεταβατικό): Άν το ρήμα παίρνει άμεσο αντικείμενο.
- Ditransitive (Δι-μεταβατικό): Άν το ρήμα έχει και άμεσο και έμμεσο αντικείμενο.
- Complex-transitive (Σύνθετο Μεταβατικό): Άν το ρήμα έχει άμεσο αντικείμενο και κατηγορούμενο.

Πίνακας 3.4: *Πίνακας με είδη προτάσεων στο ClausIE*

Sentence	**Pattern**	**Verb Type**	**Example**	**Derived Clause**
S1:	SV_i	Intransitive	AE died	(AE, died)
S2:	SV_eA	Extended-copular	AE remained in Princeton	(AE, remained, in Princeton)
S3:	SV_cC	Copular	AE is smart	(AE, is, smart)
S4:	$SV_{mt}O$	Monotransitive	AE has won the Nobel Prize	(AE, has won, the Nobel Prize)
S5:	$SV_{dt}O_iO$	Ditransitive	RSAS gave AE the Nobel Prize	(RSAS, gave, AE, the Nobel Prize)
S7:	$SV_{ct}OC$	Complex-transitive	AE declared the meeting open	(AE, declared, the meeting, open)

S: Υποκείμενο, V: Ρήμα, C: Κατηγορούμενο, O: Αντικείμενο, A: Προσδιορισμός
Intransitive: Αμετάβατο, Extended-copular: Εκτεταμένο κυκλικό, Copular: Κυκλικό, Monotransitive: Μονομεταβατικό, Ditransitive: Δι-μεταβατικό, Complex-transitive: Σύνθετο Μεταβατικό

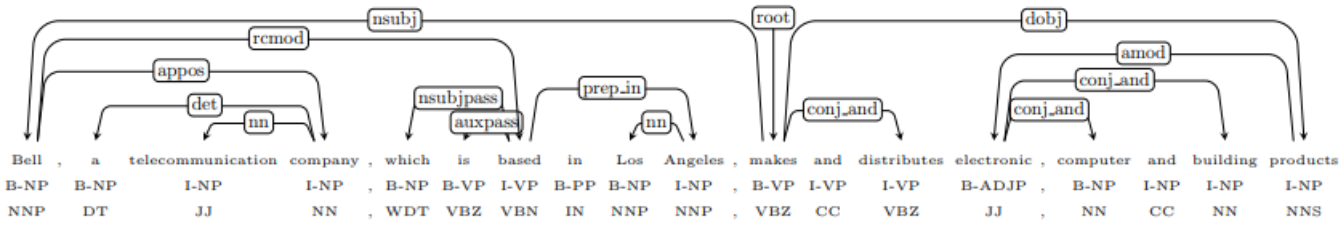
•AE remained in Princeton. •AE died in Princeton.

Για παράδειγμα η παραπάνω πρώτη πρόταση αποτελείται από ένα υποκείμενο, ένα ρήμα και ένα προσδιορισμό. Αφαιρώντας τον προσδιορισμό η υπόλοιπη πληροφορία που μένει AE remained.» δεν βγάζει σημασιολογικό νοήμα. Αντίθετα στη δεύτερη πρόταση που αποτελείται επίσης από ένα υποκείμενο, ένα ρήμα και ένα προσδιορισμό, αν αφαιρέσουμε τον προσδιορισμό είναι και πάλι συνεκτική «AE died.» και έτσι επιλέγεται αυτή ως clause.

Dependency Parsing

Με το ClausIE χρησιμοποιούμε τον Stanford parser για να ανακαλύψουμε τη συντακτική δομή μιας πρότασης που δίνεται ως είσοδος. Το συντακτικό δέντρο (dependency parsing) αποτελείται από ένα σύνολο κατευθυνόμενων συντακτικών σχέσεων μεταξύ των λέξεων της πρότασης. Η ρίζα είναι συνήθως ένα κύριο ρήμα. Για παράδειγμα στη φράση « ο Γιώργος παίζει ποδόσφαιρο», η λέξη «παίζει» αποτελεί ρίζα του συντακτικού δέντρου. Πέρα από

αυτό συνδέει το υποκείμενο που είναι «ο Γιώργος» με το άμεσο αντικείμενο που είναι «το ποδόσφαιρο». Μια πιο σύνθετη πρόταση αναλύεται παρακάτω.



Σχήμα 3.1: *Dependency Parse σύνθετης πρότασης (Συντακτικό δέντρο)*

From Dependency to Clauses

Για την αναγνώριση των clauses γίνεται χαρτογράφηση των σχέσεων εξάρτησης που δημιουργήθηκαν από το Dependency Parsing. Πρώτα δημιουργείται ένα clause για κάθε εξάρτηση υποκειμένου (π.χ. nsubj). Ο εξαρτώμενος αποτελεί το υποκείμενο (S), και ο κυβερνήτης το ρήμα (V). Τα υπόλοιπα συστατικά θα εξαρτώνται από το ρήμα, δηλαδή αντικείμενα (O) και κατηγορούμενα (C) μέσω dobj, iobj, xcomp ή ccomp, και προσδιορισμοί (A) μέσω σχέσεων εξάρτησης όπως ως advmod, advcl ή prep$_i$n.

Για την βελτίωση των αποτελεσμάτων το ClausIE δημιουργεί κάποια clauses τα οποία δεν εμφανίζονται άμεσα μέσα στην πρόταση. Αυτό συμβαίνει όταν υπάρχει μια αναφορική λέξη στο συντακτικό δέντρο. Συγκεκριμένα, αντικαθιστούμε την αναφορική αντωνυμία (π.χ. ποιος ή που) ενός clause από το προηγούμενο, το οποίο λαμβάνεται μέσω της συντακτικής εξάρτησης rcmod από τον κυβερνήτη της σχετικής λέξης. Η αντικατάσταση των αναφορικών αντωνυμιών στοχεύει στην αύξηση της πληροφορίας από τα αποσπάσματα.

Αλγοριθμος 3.2: *Απλοποίηση προτάσεων*

```
   Require One or more sentences
   Result List of clauses for each sentence
 1: ParsedSentence = parseSentence(Sentence)
 2: dependencies = extractElements(ParsedSentence)
 3: for ParsedSentence do
 4:     for dependencies = nsubj or dependencies = nsubjpass do
 5:         subject ← dependant
 6:         verb ← governor
 7:         if dobj or iobj or xcomp or ccomp that depends on verb then
 8:             object ← dependant
 9:         end if
10:         CreateClause(subject, verb, object)
11:     end for
12: end for
13: for each clause do
14:     DefineTypeofClause(verb)
15:     if TypeofClause not SV then
16:         while Clause = coherent do
17:             ReduseClause(clause)
18:         end while
19:     end if
20: end for
21: return List of Clauses
```

3.2.3 Διαχωρισμός Επιθυμητών Προτάσεων

Έχοντας εντοπίσει όλες τις υποπροτάσεις των κειμένων, επόμενο βήμα είναι ο διαχωρισμός αυτών των οποίων μας ενδιαφέρουν και πρόκειται να αναλυθούν ώστε να βρεθεί ο βαθμός αυτοματοποίησης τους. Όπως έχουμε αναφέρει παραπάνω, η διαδικασία εύρεσης προτάσεων και υποπροτάσεων δεν περιλάμβανει καμία σημασιολογική ανάλυση, έτσι μέσα στο σύνολο υποπροτάσεων μπορεί να υπάρχουν δραστηριότητες που δεν μας ενδιαφέρουν εννοιολογικά, που αφορούν π.χ. παραδείγματα ή μετα-δεδομένα.

Σκοπός είναι να βρεθούν υποπροτάσεις που υπάρχουν μέσα στο dataset ώστε να υπάρχει και ο αναφορικός βαθμός αυτοματοποίησης τους για να μπορέσουμε να κάνουμε σύγκριση αποτελεσμάτων στο τέλος. Το dataset είναι της μορφής ενός .csv αρχείου όπου για κάθε δραστηριότητα περιλαμβάνει πληροφορίες για το κείμενο και το σετ κειμένων από το οποίο προέρχεται, αριθμό πρότασης μέσα στο κείμενο κα αριθμό δραστηριότητας μέσα στο κείμενο και τέλος τον βαθμό - κατηγορία αυτοματοποίησης που ανήκει.

```
datasetID,textID,sentenceID,activityID,activity,GS
HU Berlin,Bicycle manufacturing,0,0,manufacture customized bicycles,0
HU Berlin,Bicycle manufacturing,1,1,receive an order,1
HU Berlin,Bicycle manufacturing,1,2,create a new process instance,1
HU Berlin,Bicycle manufacturing,2,3,reject the order,1
HU Berlin,Bicycle manufacturing,2,4,accept the order,1
HU Berlin,Bicycle manufacturing,4,5,the storehouse are informed,1
HU Berlin,Bicycle manufacturing,4,6,the engineering department are informed,1
HU Berlin,Bicycle manufacturing,5,7,process the part list,1
HU Berlin,Bicycle manufacturing,5,8,check the required quantity,1
HU Berlin,Bicycle manufacturing,6,9,part is reserved,1
HU Berlin,Bicycle manufacturing,7,10,part is back-ordered,1
HU Berlin,Bicycle manufacturing,9,11,prepare everything for the assembling,0
HU Berlin,Bicycle manufacturing,10,12,assemble the bicycle,0
HU Berlin,Bicycle manufacturing,11,13,ship the bicycle,0
```

Σχήμα 3.2: *Παράδειγμα Dataset για ένα κείμενο*

Για την αντιστοίχιση των υποπροτάσεων που έχουμε βρει από την ανάλυση των κειμένων με τις δραστηριότητες από το dataset έχουμε χρησιμοποιήσει την βιβλιοθήκη Fuzzy.

Η βιβλιοθήκη Fuzzy χρησιμοποιείται σε περιπτώσεις ομοιότητας strings. Υπολογίζει την απόσταση μεταξύ δύο φράσεων, όσο πολύπλοκων και αν είναι και επιστρέφει ένα ποσοστό το οποίο έχει προκύψει από την συνάρτηση Levenshtein. Το ενδιαφέρον της βιβλιοθήκης αυτής είναι ότι προσφέρει κάποιες συναρτήσεις με επιπλέον δυνατότητες από αυτή της απλής απόστασης που είνα χρήσιμες σε πολύπλοκες περιπτώσεις.

```
Str1 = "Los Angeles Lakers"

Str2 = "Lakers"

Ratio = fuzz.ratio(Str1.lower(),Str2.lower())

Partial_Ratio = fuzz.partial_ratio(Str1.lower(),Str2.lower())
```

Σχήμα 3.3: *Παράδειγμα 1 για Fuzzy*

Στο παραπάνω παράδειγμα συγκρίνουμε δυο φράσεις το str1 (Los Angeles Lakers), και το str2 (Lakers). Οι δυο αυτές φράσεις αναφέρονται στην ίδια ακριβώς ομάδα. Ωστόσο η συνάρτηση fuzz.ratio που υπολογίζει την απόσταση Levenshtein επιστρέφει ποσοστό 50% καθώς συγκρίνει απλώς λέξεις. Αντίθετα η συνάρτηση fuzz.partial_ratio είναι ικανή να εντοπίσει ότι και οι δύο φράσεις αναφέρονται στην ομάδα Lakers και έτσι επιστρέφει ποσοστό 100%. Ο τρόπος που λειτουργεί είναι χρησιμοποιώντας μια «βέλτιστη μερική» λογική. Με άλλα λόγια, εάν η συντομότερη συμβολοσειρά έχει μήκος k και η μεγαλύτερη συμβολοσειρά έχει το μήκος m, τότε ο αλγόριθμος αναζητά τη βαθμολογία Levenshtein της καλύτερης αντιστοίχισης υποφράσης μήκους-k.

```
Str1 = "united states v. nixon"
Str2 = "Nixon v. United States"
Ratio = fuzz.ratio(Str1.lower(),Str2.lower())
Partial_Ratio = fuzz.partial_ratio(Str1.lower(),Str2.lower())
Token_Sort_Ratio = fuzz.token_sort_ratio(Str1,Str2)
```

Σχήμα 3.4: *Παράδειγμα 2 για Fuzzy*

Στο παραπάνω παράδειγμα βλέπουμε την ίδια φράση με διαφορετική σειρά. Οι παραπάνω συναρτήσεις που έχουμε αναφέρει δεν καταφέρνουν να αναγνωρίσουν πως πρόκειται για την ίδια πρόταση καθώς η fuzz.ratio επιστρέφει ποσοστό 59% ενώ η fuzz.partial_ratio επιστρέφει ποσοστό 74%. Για τον λόγο αυτό υπάρχει ένα ακόμα σύνολο συναρτήσεων οι οποίες μπορούν να προσπεράσουν το πρόβλημα αυτό. Οι συναρτήσεις fuzz.token χωρίζουν τις φράσεις σε λέξεις και τις προεπεξεργάζονται μετατρέποντας τις σε πεζά και αφαιρώντας τα σημεία στίξης. Στην περίπτωση του fuzz.token_sort_ratio, οι συμβολοσειρές ταξινομούνται αλφαβητικά και στη συνέχεια ενώνονται. Μετά από αυτό, εφαρμόζεται ένα απλό fuzz.ratio για να ληφθεί το ποσοστό ομοιότητας. Έτσι στο συγκεκριμένο παράδειγμα επιστρέφει ποσοστό 100%. Αυτήν την συνάρτηση έχουμε αξιοποιήσει και εμείς στον αλγόριθμό μας. Ουσιαστικά προσπαθούμε να εντοπίσουμε τις ομοιότερες προτάσεις μεταξύ dataset και clauses ώστε για κάθε clause να υπάρχει μια ετικέτα (label) για τον βαθμό αυτοματοποίησης του. Την ετικέτα την χρειαζόμαστε για να μπορέσουμε αργότερα να εφαρμόσουμε τους αλγορίθμους μηχανικής μάθησης.

Αλγοριθμος 3.3: *Διαχωρισμός επιθυμητών προτάσεων*

Require Clauses and Input Dataset
Result List of desired clauses

1: **for** each phrase from Dataset **do**
2: **for** every clause of the same sentence as phrase **do**
3: ratio = fuzz.token_sort_ratio(clause, phrase)
4: Add.list(ratio)
5: **end for**
6: Max(listofratio)
7: Find index of clause with max ratio
8: **if** max ratio > 50% **then**
9: Add.list(ratio, index)
10: **end if**
11: **end for**
12: Sort.list(descending)
13: KeepUniqueIndexes(list) ▷ Αφαιρούνται αυτά με το μικρότερο ποσοστό ratio
14: **return** List of desired Clauses

3.2.4 Εξαγωγή Χαρακτηριστικών ανά Clause

Στο σημείο αυτό για κάθε clause που έχουμε βρει, του έχουμε αντιστοιχίσει και ετικέτα ως προς τον βαθμό αυτοματοποίησης του. Σκοπός είναι από το clause αυτό σαν δεδομένο, να καταλήξουμε στο συμπέρασμα που θα είναι η ετικέτα. Για να γίνει αυτό θα πρέπει να αντλήσουμε κάποια δεδομένα από κάθε clause ώστε να δημιουργηθεί ένας πίνακας ο οποίος θα είναι η είσοδος στους αλγορίθμους μηχανικής μάθησης όπως θα περιγράψουμε παρακάτω. Η επιλογή και ο υπολογισμός των κατάλληλων χαρακτηριστικών είναι το βασικότερο βήμα κατά την κατασκευή μιας λύσης που βασίζεται στη μηχανική μάθηση.

Επομένως, αναλύσαμε χειροκίνητα ποια χαρακτηριστικά στο σύνολο δεδομένων μας επηρεάζουν τον βαθμό αυτοματοποίησης μιας εργασίας. Σαν αποτέλεσμα, επιλέξαμε και εφαρμόσαμε τέσσερα χαρακτηριστικά:

- Κατηγορία Υποκειμένου
- Ρήμα
- Αντικείμενο
- Τομέας Τεχνολογίας

Στις επόμενες παραγράφους εξηγούμε τον ορισμό και το σκεπτικό του κάθε χαρακτηριστικού καθώς και τον υπολογισμό του.

Κατηγορία Υποκειμένου

Συνήθως ο διαχωρισμός των κατηγοριών ενός υποκειμένου γίνεται αναφορικά με ένα δυαδικό χαρακτηριστικό που χαρακτηρίζει τον πόρο που εκτελεί μια εργασία ως «ανθρώπινο» είτε «μη ανθρώπινο». Ωστόσο στο δικό μας dataset ανάλογα με τον τομέα της εξεταζόμενης διαδικασίας, οι πόροι κυμένονταν, μεταξύ άλλων, να σχετίζονται με συγκεκριμένους ρόλους (π.χ. «διευθυντής» ή «λογιστής»), τμήματα (π.χ. «τμήμα Ανθρώπινου Δυναμικού» ή «λογιστικό τμήμα») και επίσης συστήματα («Σύστημα ERP» ή «σύστημα πληροφοριών»). Για αυτόν τον λόγο αποφασίσαμε να χωρίσουμε τα υποκείμενα σε αυτές τις τρεις κατηγορίες.

Από το Clausie για κάθε clause έχουμε πληροφορίες σχετικά με το Υποκείμενο - Ρήμα - Αντικείμενο (αντικείμενο όχι πάντα). Αυτά μπορεί να είναι λέξεις ή φράσεις ανάλογα την πρόταση.

Ξεκινάμε με το Υποκείμενο. Με την βοήθεια της βιβλιοθήκης NLTK δημιουργούμε ένα συντακτικό δέντρο (Dependency Parse) για την φράση του υποκείμενου από το οποίο μπορούμε να βρούμε ποια είναι η ρίζα του. Στόχος είναι να καταλήξουμε σε μια λέξη για το υποκείμενο.

- His son John won the game.

Για παράδειγμα στην παραπάνω πρόταση το Clausie θα επιστρέψει ως υποκείμενο την φράση «His son John». Εμείς με την συντακτική ανάλυση θα βρούμε ότι η ρίζα της φράσης είναι το όνομα «John» άρα και το κύριο υποκείμενο μας. Η έυρεση της ρίζας είναι εύκολη καθώς απο το dependency parsing η λέξη αυτή θα έχει την ετικέτα «ROOT».

Έχοντας βρει το κύριο υποκείμενο, μπορούμε να βρούμε και σε τι κατηγορία ανήκει. Για να προσδιορίσουμε τον τύπο του υποκειμένου, χρησιμοποιούμε τη λεξική βάση δεδομένων WordNet. Το WordNet ομαδοποιεί τις αγγλικές λέξεις σε σύνολα, τα λεγόμενα synsets. Για κάθε ένα από τα 117.000 σύνολα που περιέχει το WordNet, παρέχει σύντομους ορισμούς, παραδείγματα και μια σειρά σημασιολογικών σχέσεων με άλλα σύνολα. Για να υπολογίσουμε αυτήν τη δυνατότητα, αξιοποιούμε τη σχέση υπερνύμου hypernym από το WordNet. Σε γενικές γραμμές, ένα hypernym είναι ένας πιο γενικός όρος για μια δεδομένη λέξη. Για παράδειγμα, η λέξη «όχημα» είναι το hypernym του «αυτοκινήτου» και η λέξη «πουλί» είναι το hypernym του «αετού». Βάσει αυτής της έννοιας της υπερνυμίας και της ιεραρχικής οργάνωσης του WordNet, είμαστε σε θέση να συμπεράνουμε για έναν δεδομένο πόρο, αν το hypernym του είναι «person», «computer_system» ή «social_group». Με βάση το hypernym μπορούμε τότε να κατηγοριοποιήσουμε αυτόματα εάν ένας πόρος είναι άνθρωπος, τμήμα ή σύστημα.

Ρήμα

Η κύρια ιδέα πίσω από αυτό το χαρακτηριστικό είναι ότι ορισμένα ρήματα είναι πιθανότερο να σχετίζονται με αυτοματοποιημένες εργασίες από άλλα. Για παράδειγμα, τα ρήματα «δημιουργώ» ή «μεταδίδω», πιθανώς σχετίζονται με αυτοματοποιημένες εργασίες. Τα ρήματα «αναλύω» και «αποφασίζω», αντιθέτως, είναι πιο πιθανό να σχετίζονται με χειροκίνητες εργασίες. Το πλεονέκτημα της εισαγωγής ενός χαρακτηριστικού ρήματος έναντι της χρήσης προκαθορισμένων κλάσεων ρήματος (όπως οι κλάσεις ρήματος Levin) είναι ότι ένα ρήμα δεν είναι προκαθορισμένο σε ποια κατηγορία θα ανήκει. Το ρήμα «δημιουργώ», για παράδειγμα, μπορεί επίσης να χρησιμοποιηθεί στο πλαίσιο «δημιουργία ιδεών» και, επομένως, να αναφέρεται σε μια χειροκίνητη εργασία. Μια τέτοια χρήση που σχετίζεται με το περιβάλλον μπορεί να ληφθεί υπόψην όταν το ρήμα θεωρείται μέρος ενός συνόλου χαρακτηριστικών.

Η εύρεση του ρήματος είναι εύκολη καθώς από το Clausie έχουμε το δεδομένο αυτό. Όπως αναφέρθηκε παραπάνω μπορεί να μην έχουμε μια λέξη ως ρήμα αλλά μια φράση. Για παράδειγμα στην παρακάτω πρόταση που υπάρχει ένα βοηθητικό ρήμα (modal verb), σαν ρήμα έχει επιστραφεί όλη η φράση «have to do».

- I have to do the laundry.

Για αυτόν τον λόγο γίνεται πρώτα μια συντακτική ανάλυση όπως παραπάνω, αλλα ταυτόχρονα και ένα έλεγχος σχετικά με τα βοηθητικά ρήματα. Συγκεκριμένα για αυτά ελέγχουμε την ύπαρξη της πρόθεσης «το» που τα συνοδεύει. Έχοντας βρει το βασικό ρήμα της πρότασης το επόμενο βήμα είναι να βρούμε την ρίζα του. Με αυτόν τον τρόπο θα βοηθήσουμε τους αλγορίθμους μηχανικής μάθησης να βρουν παραπάνω ομοιότητες για να γίνει σωστά η ταξινόμηση. Για την εύρεση της ρίζας χρησιμοποιήσαμε το stemmer εργαλείο SnowballStemmer("english") το οποίο ανήκει στην βιβλιοθήκη NLTK.

Αντικείμενο

Το ρήμα μιας δραστηριότητας συνήθως να αναφέρεται σε ένα αντικείμενο. Το σκεπτικό πίσω από αυτό το χαρακτηριστικό είναι, παρόμοια με το ρήμα, ότι ορισμένα αντικείμενα είναι πιο πιθανό να σχετίζονται με αυτοματοποιημένες εργασίες από άλλα. Για παράδειγμα, ας δούμε τους δύο συνδυασμούς ρήματος-αντικειμένου «στέλνω επιστολή» και «στέλνω e-mail». Παρόλο που και τα δύο περιέχουν το ρήμα «στέλνω», το αντικείμενο αποκαλύπτει ότι το πρώτο σχετίζεται με μια χειροκίνητη διαδικασία και το δεύτερο σχετίζεται με μια εργασία χρήστη (η αποστολή ενός e-mail απαιτεί σίγουρα την αλληλεπίδραση με έναν υπολογιστή). Για αυτόν τον λόγο το αντικειμένο ως χαρακτηριστικό μπορεί να βοηθήσει στην ταξινόμηση διαφορετικών βαθμών αυτοματοποίησης εργασιών.

Παρόμοια με τη λειτουργία ρήματος, η εύρεση του αντικειμένου αποτελεί μέρος της γλωσσικής προεπεξεργασίας, και συγκεκριμένα είναι αποτέλεσμα του Clausie. Το διαφορετικό με το αντικείμενο είνα πως δεν είναι υποχρεωτικό να υπάρχει πάντα. Έτσι κάποιες φορές μπορεί το πεδίο Αντικείμενο από το Clausie να είναι κενό. Σε αυτή την περίπτωση κάνουμε έναν επιπλέον έλεγχο για να δούμε αν η πρόταση είναι σε παθητική φωνή. Αν είναι τότε ως αντικείμενο χρησιμοποιούμε το Υποκείμενο για να μην χαθεί καθόλου πληροφορία.

Όπως αναφέρθηκε παραπάνω μπορεί να μην έχουμε μια λέξη ως αντικείμενο αλλά μια φράση ολόκληρη. Για αυτόν τον λόγο γίνεται πρώτα μια συντακτική ανάλυση για να βρούμε ποιο είναι το βασικό αντικείμενο, ή αλλιώς η ρίζα του συντακτικού δέντρου. Έχοντας βρει το βασικό αντικείμενο της πρότασης το επόμενο βήμα είναι να βρούμε την ρίζα του. Με αυτόν τον τρόπο θα βοηθήσουμε όπως αναφέραμε τους αλγορίθμους μηχανικής μάθησης να βρουν παραπάνω ομοιότητες για να γίνει σωστά η ταξινόμηση. Για την εύρεση της ρίζας χρησιμοποιήσαμε το stemmer εργαλείο SnowballStemmer("english") το οποίο ανήκει στην βιβλιοθήκη NLTK.

Τομέας Τεχνολογίας

Το χαρακτηριστικό του τομέα τεχνολογίας (IT) είναι ένα δυαδικό χαρακτηριστικό που αποκαλύπτει εάν μια εργασία σχετίζεται με τον τομέα IT ή όχι. Το σκεπτικό πίσω από αυτήν τη δυνατότητα είναι ότι μια εργασία που σχετίζεται με τον τομέα IT είναι πιθανό να είναι μια εργασία χρήστη ή ακόμη και μια αυτοματοποιημένη εργασία.

•Ο πελάτης υποβάλλει παράπονο μέσω του συστήματος διαχείρισης παραπόνων.

Για παράδειγμα, ας δούμε την παραπάνω πρόταση. Περιλαμβάνει τον ανθρώπινο όρο «πελάτη», το ρήμα «υποβάλλει» και το αντικείμενο «παράπονο». Κανένα από αυτά τα στοιχεία δεν δείχνει σαφώς έναν βαθμό αυτοματοποίησης. Ωστόσο, η φράση αναφέρει επίσης ένα «σύστημα διαχείρισης παραπόνων». Ο στόχος του χαρακτηριστικού τομέα τεχνολογίας είναι να λαμβάνει υπόψην λέξεις ή φράσεις που έχουν σχέση με την τεχνολογία.

Για να υπολογίσουμε αυτό το χαρακτηριστικό, αξιοποιούμε το γλωσσάριο των όρων υπολογιστών που αναπτύχθηκε από το Πανεπιστήμιο της Utah. Εκτός από μια ολοκληρωμένη κάλυψη τεχνικών όρων, αυτή η λίστα περιέχει επίσης ρήματα και επίθετα που χρησιμοποιούνται σε ένα περιβάλλον πληροφορικής. Εάν μια εξεταζόμενη πρόταση, περιέχει έναν

ή περισσότερους όρους από αυτήν τη λίστα, τότε το χαρακτηριστικο αυτό λαμβάνει την τιμή «ναι» για οποιαδήποτε εργασία που αποτελεί μέρος αυτής της πρότασης. Αλλιώς έχε τιμή «όχι». Τονίζουμε πως το χαρακτηριστικό αυτό αφορά ολόκληρη την πρόταση και όχι κάποια υποπρόταση. Έτσι αν μια πρόταση περιλαμβάνει μια λέξη που έχει σχέση με την τεχνολογία, τότε κάθε δραστηριότητα της πρότασης αυτής θα έχει θετική τιμή στο πεδίο αυτό. Η λίστα με τις λέξεις παρουσιάζεται στο τέλος της διπλωματικής εργασίας.

Αλγοριθμος 3.4: *Εξαγωγή Χαρακτηριστικών ανά Clause*

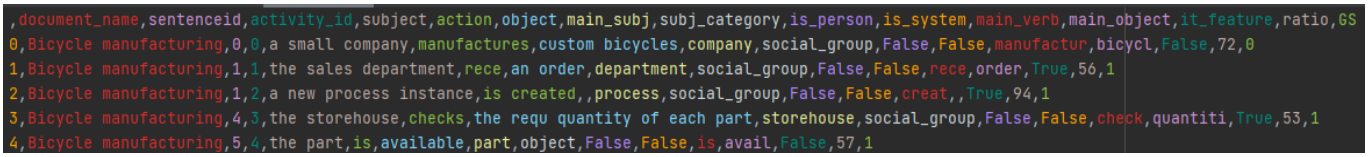
Require Clauses, it_word_list
Result Data for ML algorithms

for each Clause **do**
  it_word_flag ← *Exists*(*it_word_list*, *clause*)
  subject ← *DefineSubject*(*Triple*[0])
  ishuman ← *is_human*(*subject*)
  issystem ← *is_system*(*subject*)
  issocialgroup ← *is_social_group*(*subject*)
  verb ← *DefineVerb*(*Triple*[1])
  if verb = ModalVerb **then**
    verb ← *CorrectVerb*
  end if
  verb ← *Root*(*verb*)
  object ← *DefineObject*(*Triple*[2])
  if object = " " **then**
    if Clause in Passive voice **then**
      object ← *subject*
      object ← *Root*(*object*)
    end if
  else
    object ← *Root*(*object*)
  end if
end for
return [it_word_list, ishuman, issystem, issocialgroup, verb, object] for each Clause

3.3 Προετοιμασία για Μηχανική Μάθηση

Έχοντας ολοκληρώσει όλο το παραπάνω pre-processing των κειμένων καταλήγουμε σε ένα σύνολο δεδομένων - χαρακτηριστικών για κάθε δραστηριότητα που έχουμε εντοπίσει. Τα δεδομένα αυτά τα έχουμε ομαδοποιήσει στη μορφή ενός Dataframe, που είναι μια δομή δεδομένων με αριθμημένες γραμμές και στήλες. Έτσι έχουμε δημιουργήσει ουσιαστικά ένα dataset που για κάθε δραστηριότητα που έχει εντοπιστεί από τα κείμενα περιλαμβάνει τις παρακάτω πληροφορίες:

1. Ρήμα φράσης (σε μορφή ρίζας)
2. Αντικείμενο φράσης (σε μορφή ρίζας)
3. Αν το υποκείμενο είναι πρόσωπο
4. Αν το υποκείμενο είναι σύστημα
5. Αν το υποκείμενο είναι τμήμα
6. Αν στη φράση υπάρχει λέξη σχετική με την τεχνολογία



Σχήμα 3.5: *Απόσπασμα δεδομένων μετά το preprocesing*

Από το τρίτο χαρακτηριστικό και μετά πρόκειται για δυαδικές πληροφορίες δηλαδή που οι τιμές κυμαίνονται μεταξύ του TRUE και του FALSE. Ωστόσο στις κατηγορίες ένα και δύο υπάρχουν κατηγορηματικά δεδομένα. Αυτό σημαίνει ότι δεν είναι προκαθορισμένο το σύνολο τιμών καθώς πρόκειται για τα ρήματα και τα αντικείμενα κάθε φράσης.

Στην επεξεργασία κειμένων είναι αναμενόμενο οι λέξεις του κειμένου να αντιπροσωπεύονται με διακριτά, κατηγορηματικά χαρακτηριστικά. Αυτό αποτελεί πρόβλημα για κάποιους αλγόριθμους μηχανικής μάθησης καθώς χρειάζονται σαν είσοδο δεδομένα αριθμών ή λογικά (boolean) που επίσης μεταφράζονται σε 0 και 1.

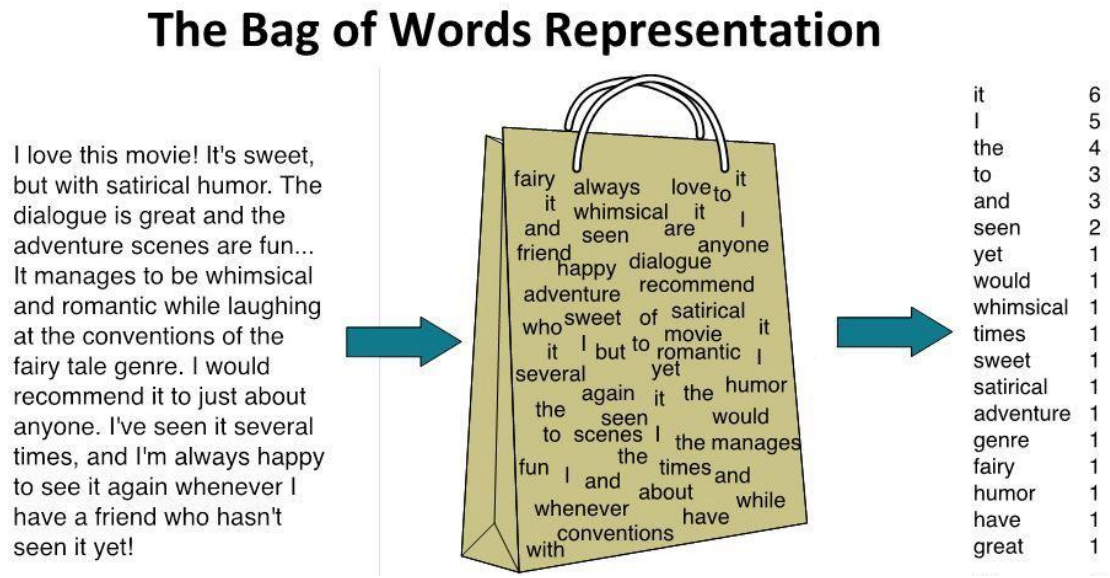
Ορισμένοι αλγόριθμοι μπορούν να λειτουργήσουν με κατηγορηματικά δεδομένα απευθείας. Για παράδειγμα, ένα δέντρο αποφάσεων (decision tree) μπορεί να μάθει απευθείας από κατηγορηματικά δεδομένα χωρίς να απαιτείται μετασχηματισμός δεδομένων. Πολλοί αλγόριθμοι μηχανικής μάθησης δεν μπορούν να λειτουργήσουν απευθείας σε δεδομένα κατηγορηματικά. Απαιτούν όλες οι μεταβλητές εισόδου και μεταβλητές εξόδου να είναι αριθμητικές. Αυτό σημαίνει ότι τα κατηγορηματικά δεδομένα πρέπει να μετατραπούν σε αριθμητική μορφή.

Πώς κωδικοποιούμε τέτοια δεδομένα με τρόπο που να είναι αποδεκτός ώστε να χρησιμοποιηθεί από τους αλγόριθμους· Η χαρτογράφηση από δεδομένα κειμένου σε πραγματικούς αριθμούς ονομάζεται εξαγωγή χαρακτηριστικών (Feature Extraction). Μία από τις απλούστερες τεχνικές για την αριθμητική αναπαράσταση κειμένου είναι το Bag of Words.

3.3.1 Bag of Words

Με τον όρο Bag of Words εννοούμε μια λίστα των μοναδικών λέξεων στο κείμενο που ονομάζεται λεξιλόγιο. Στη συνέχεια μπορούμε να αντιπροσωπεύσουμε κάθε λέξη με τους

αριθμούς 0 και 1 για το αν αντιστοιστιχούν σε μία λέξη από το λεξιλόγιο ή όχι. Ένας ακόμη τρόπος αναπαράστασης είναι η απαρίθμηση των φορών που επαναλαμβάνεται μια λέξη μέσα στο κείμενο, αλλά δεν μπορούσε να εφαρμοστεί στην δική μας περίπτωση.



Σχήμα 3.6: *Bag of Words*

Στο παραπάνω παράδειγμα μπορούμε να δούμε πως έχει σχηματιστεί ένα λεξιλόγιο με όλες τις λέξεις του κειμένου (movie, secenes, have, κ.λπ.) και αποτυπώνεται με αριθμό δίπλα η επανάληψη εμφανίσεων για κάθε δεδομένο.

Στον αλγόριθμό μας έχουμε χρησιμοποιήσει την μέθοδο του One Hot Encoding. Είναι της κατηγορίας με την δυαδική αναπαρασταση δεδομένων. Έτσι έχουμε τιμές 0 και 1 ανάλογα για το αν η εκάστοτε λέξη αντιστοιχεί σε κάποια λέξη από το λεξιλόγιο ή αλλιώς bag of words. Έχουμε προτιμήσει την τεχνική αυτή καθώς η συχνότητα επανάληψης λέξεων, δημιουργεί μια ταξινόμηση ή μια προτεραιότητα που μπορεί να επηρεάσει τα αποτελέσματα των αλγορίθμων.

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Σχήμα 3.7: *One Hot Encoding*

Στο παραπάνω παράδειγμα μπορούμε να δούμε πως έχει σχηματιστεί ένα λεξιλόγιο με τις λέξεις (Apple, Chicken, Broccoli, κ.λπ.) και αποτυπώνεται με 0 ή 1 η εμφάνιση της λέξης για κάθε διακριτή λέξη (ανά σειρά).

Ένα από τα κύρια μειονεκτήματα της χρήσης του Bag of Words είναι ότι απορρίπτει τη σειρά εμφάνισης των λέξεων αγνοώντας και την έννοια τους. Για την επεξεργασία φυσικής γλώσσας (NLP) είναι εξαιρετικά σημαντική η διατήρηση του πλαισίου των λέξεων. Για να λύσουμε αυτό το πρόβλημα χρησιμοποιήσαμε και μια άλλη προσέγγιση που ονομάζεται Word Embedding.

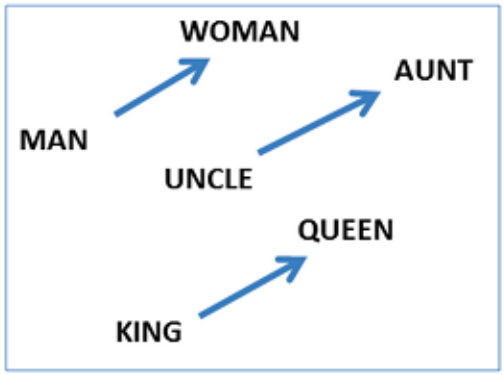
3.3.2 Word Embeddings

Το word embedding είναι η αναπαράσταση του κειμένου όπου λέξεις που έχουν την ίδια έννοια έχουν και παρόμοια αναπαράσταση. Αυτή η προσέγγιση για την αναπαράσταση λέξεων και εγγράφων μπορεί να θεωρηθεί ως ένα από τα πιο βασικά επιτεύγματα της βαθιάς μάθησης (deep learning) σχετικά με τα δύσκολα προβλήματα επεξεργασίας φυσικών γλωσσών.

Τα word embeddings λέξεων είναι στην πραγματικότητα μια κατηγορία τεχνικών όπου μεμονωμένες λέξεις αντιπροσωπεύονται ως διανύσματα πραγματικών τιμών σε έναν προκαθορισμένο διανυσματικό χώρο. Κάθε λέξη χαρτογραφείται σε έναν φορέα και οι τιμές του διανύσματος μαθαίνονται με τρόπο που μοιάζει με ένα νευρωνικό δίκτυο

Κάθε λέξη αντιπροσωπεύεται από ένα πραγματικό διάνυσμα, συχνά δεκάδων ή εκατοντάδων διαστάσεων. Αυτό έρχεται σε αντίθεση με τις χιλιάδες ή εκατομμύρια διαστάσεις που απαιτούνται για αναπαράσταση αραίων λέξεων, όπως με το One Hot Encoding. Ένα από τα πιο γνωστά μοντέλα των word embeddings και αυτό το οποίο χρησιμοποιήσαμε είναι το Word2Vec.

Το Word2vec παίρνει ως είσοδο του ένα μεγάλο σώμα κειμένου και παράγει ένα διανυσματικό χώρο με κάθε μοναδική λέξη να αντιστοιχεί σε ένα αντίστοιχο διάνυσμα στο χώρο. Τα διανύσματα λέξεων τοποθετούνται στο χώρο του διανύσματος έτσι ώστε οι λέξεις που μοιάζουν νοηματικά να βρίσκονται σε κοντινή απόσταση μεταξύ τους στο χώρο.



Σχήμα 3.8: *Διανυσματικός Χώρος με το Word2Vec*

Στο παραπάνω παράδειγμα βλέπουμε μια αναπαράσταση κάποιων λέξεων στον χώρο. Αυτό το είδος διανυσματικής αναπαράσταση μας επιτρέπει επίσης να απαντήσουμε στο ερώτημα «King - Man + Woman =;» και να φτάσουμε στο αποτέλεσμα «Queen»! Όλα αυτά είναι πραγματικά αξιοσημείωτα αφού όλες αυτές οι γνώσεις προέρχονται απλώς από την εξέταση πολλών λέξεων στο κείμενο χωρίς άλλες πληροφορίες σχετικά με τη σημασιολογία τους.

Όσον αφορά την περίπτωσή μας κωδικοποιήσαμε τα δύο χαρακτηριστικά (ρήμα και αντικείμενο) με τις παραπάνω μεθόδους και χρησιμοποιήσαμε τους αλγορίθμους μηχανικής μάθησης. Η σύγκριση των αποτελεσμάτων θα γίνει στο επόμενο κεφάλαιο.

Κεφάλαιο 4

Αξιολόγηση Μοντέλων

Για την αξιολόγηση της διαδικασίας που περιγράψαμε στην προηγούμενη ενότητα, συλλέχθηκε ένα σύνολο δεδομένων δοκιμής (test data). Στις επόμενες ενότητες θα περιγράψουμε λεπτομερέστερα τη σύνθεση και το περιεχόμενο αυτού του συνόλου δεδομένων δοκιμής. Στη συνέχεια, θα συγκρίνουμε τα μοντέλα που δημιουργήθηκαν από το πρόγραμμά μας με τα δεδομένα και θα αναλύσουμε τα αποτελέσματα.

4.1 Σύνολο Δεδομένων - Data Set

Η επιλογή των κατάλληλων δεδομένων είναι από τους σημαντικότερους παράγοντες, αν όχι ο πιο σημαντικός, που καθορίζουν την επιτυχία ενός προγράμματος επεξεργασίας φυσικής γλώσσας και μηχανικής μάθησης. Η εύρεση όμως των κατάλληλων δεδομένων αποτελεί ένα πολύ μεγάλο πρόβλημα στο χώρο αυτό και ειδικότερα όταν πρόκειται για επιχειρησιακά δεδομένα. Για ένα πρόγραμμα επιτηρούμενης μάθησης, όπως αυτό που αναπτύξαμε, τα δεδομένα μας θέλουμε να είναι πρωτίστως έγκυρα και ορθά κατηγοριοποιημένα. Ένα αποτέλεσμα δεν έχει καμία αξία εάν έχει βασιστεί σε ψευδή ή λάθος δεδομένα. Σημαντικό για το πρόγραμμά μας είναι και το πλήθος των δεδομένων, μιας και όσο μεγαλύτερο είναι αυτό, τόσο πιο έμπιστο και στιβαρό γίνεται το μοντέλο μας, και τόσο καλύτερα μπορούμε να γενικεύσουμε τα αποτελέσματά μας σε νέα δεδομένα, χωρίς βέβαια αυτό να αποτελεί το μόνο παράγοντα. Είναι χρήσιμο ένα σύνολο δεδομένων να είναι ισορροπημένο, να υπάρχει δηλαδή ισοκατανομή των στοιχείων στις κλάσεις του.

Ένα στοιχείο αυτού του συνόλου δεδομένων δοκιμής αποτελείται από μια περιγραφή διαδικασίας κειμένου. Συνολικά καταφέραμε να συλλέξουμε 44 από αυτά τα περιργραφικά κείμενα. Δεν περιορίσαμε τη συλλογή σε συγκεκριμένο τομέα ή τύπο λόγω της έλλειψης δεδομένων. Έτσι, διαφορετικές πηγές ενσωματώθηκαν στο σύνολο δεδομένων δοκιμών μας. Ταξινομήσαμε το καθένα στοιχείο σε μία από τις τέσσερις κατηγορίες:

- **Ακαδημαϊκά**: Στοιχεία που παρέχονται από πανεπιστήμια ή υπαλλήλους πανεπιστημίου.
- **Βιομηχανία**: Στοιχεία που παρέχονται από εταιρείες ή υπαλλήλους τους.
- **Κείμενο βιβλίου**: Στοιχεία που λαμβάνονται από βιβλία.

- **Δημόσιος τομέας**: Στοιχεία που λαμβάνονται από εθνικούς φορείς του δημόσιου τομέα.

Η κατανομή των δεδομένων δοκιμής σύμφωνα με αυτές τις κατηγορίες φαίνεται στο παρακάτω πινακάκι. Ενώ τα ακαδημαϊκά μοντέλα και τα μοντέλα του δημόσιου τομέα αντιπροσωπεύουν το μεγαλύτερο μερίδιο, όλες οι άλλες κατηγορίες είναι επίσης επαρκώς εκπροσωπημένες.

Πίνακας 4.1: *Πηγές για Test Data*

Type	**Amount**	**Frequency**
Academic	14	31,81%
Industry	8	18,18%
Textbook	8	18,18%
Public Sector	14	31,81%
Total	44	100%

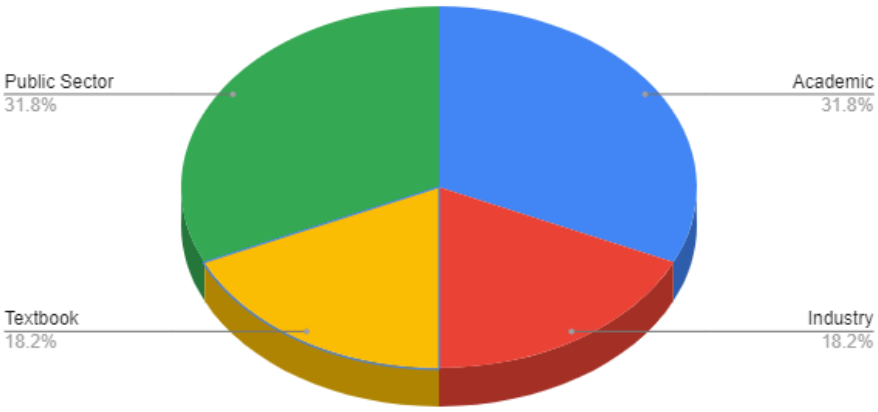
Τα ακαδημαϊκά κείμενα του συνόλου δεδομένων δοκιμών προέρχονται από το Humboldt Universit at zu Berlin, το Technische Universit at Berlin, το Queensland University of Technology και το Technische Universiteit Eindhoven. Στο σύνολο αυτό τα κείμενα είναι πιο μεγάλα και πιο πολύπλοκα από ότι τα υπόλοιπα. Περιλαμβάνουν δηλαδή μεγαλύτερο αριθμό προτάσεων και λέξεων σε σχέση με τον μέσο αριθμό των υπόλοιπων κειμένων.

Τα βιομηχανικά μοντέλα προέρχονται από δύο κύριες πηγές. Αρχικά τα μισά προέρχονται από τρεις προμηθευτές εργαλείων BPM, συγκεκριμένα τους Active VOS, Oracle και BizAgi. Τα υπόλοιπα είναι από την inubit AG, στην οποία τα χρησιμοποιούν για την εκπαίδευση πελατών και εργαζομένων.

Επιπλέον ένα μικρό ποσοστό κειμένων προέρχεται από δύο εγχειρίδια για μοντελοποίηση BPMN, το BPMN Modeling and Reference Guide και το Praxishandbuch BPMN.

Τέλος, υπάρχει ένα σύνολο κειμένων από την Ομοσπονδιακής Υπηρεσίας Δικτύου της Γερμανίας σχετικά με κάποιες διαδικασίες. Εκεί οι περιγραφές κειμένου δεν ήταν ξεκάθαρες, αλλά ημι-δομημένες σε μορφή πίνακα και για αυτό έγινε η προεπεξεργασία που έχουμε αναφέρει στο προηγούμενο κεφάλαιο.

Είναι πολύ σημαντικό εδώ να αναφέρουμε πως τα κείμενα αυτά μας τα παρείχε ο Fabian Friedrich από το paper του Automated Generation of Business Process Models from Natural Language Input.



Σχήμα 4.1: *Διακύμανση πηγών data set*

Οι διαφορετικές περιγραφές καλύπτουν διάφορους τομείς, όπως π.χ. ασφάλιση, ιατρική, τραπεζική, μάρκετινγκ, πωλήσεις και παροχή ηλεκτρικού ρεύματος. Το γεγονός ότι είμαστε σε θέση να ανλύσουμε κείμενα όλων αυτών των τομέων επισημαίνει την ανεξαρτησία της προτεινόμενης λύσης από τον τομέα.

Παρακάτω θα δούμε κάποια χαρακτηριστικά των κειμένων.

Πίνακας 4.2: *Χαρακτηριστικά του Test Data ανά πηγή*

Source	**Type**	**Texts**	**Sents**	**Words**	**MT**	**UT**	**AT**
HU Berlin	Academic	4	10.0	18.1	32	21	0
TU Berlin	Academic	2	34.0	21.2	20	39	25
QUT	Academic	7	6.1	18.3	37	20	5
TU Eindhoven	Academic	1	40.0	18.5	23	15	4
Vendor Tutorials	Industry	4	9.0	18.2	4	34	16
inubit AG	Industry	4	11.5	18.4	37	21	2
BPM Practisioners	Industry	1	7	9.7	6	1	0
BPMN practical handbook	Textbook	3	4.7	17.0	8	14	0
BPMN guide	Textbook	4	7.0	20.8	26	7	9
FNA	Public Sector	14	6.43	13.95	33	48	0
Total		**44**	**8.06**	**15.37**	**226**	**220**	**61**

Sents: Μέσος αριθμός προτάσεων ανά κείμενο, Words: Μέσος αριθμός λέξεων ανά πρόταση, MT: Χειρονακτικές δραστηριότητες, UT: Δραστηριότητες χρήστη, AT: Αυτοματοποιημένες δραστηριότητες

Βλέπουμε πως οι περιγραφές διαδικασιών με τις οποίες δουλέψαμε διαφέρουν αρκετά στις διαστάσεις. Μεγαλύτερη απόκλιση παρουσιάζουν στο μέγεθος καθώς οι μέσες τιμές κυμαίνονται από 4.7 μέχρι 40.0. Διαφορές υπάρχουν όμως και στο μήκος των προτάσεων αφού και εκεί υπάρχει διακύμανση από 9.7 σε 21.2 λέξεις ανά πρόταση. Ακόμη μία σημαντική διαφορά η οποία δεν φαίνεται στον παραπάνω πίνακα αλλά γίνεται αντιληπτή μέσα από τα κείμενα, είναι στον τρόπο γραφής τους. Δηλαδή στο πόσο κατανοητά και ξεκάθαρα περιγράφονται οι διαδικασίες. Όλες αυτές οι διαφορές εννοείται θα επηρεάσουν τα αποτελέσματα της προτεινόμενης λύσης.

Είναι φανερό από το παραπάνω πινακάκι πως τα δεδομένα για τις αυτοματοποιημένες

διαδικασίες ήταν σαφώς λιγότερα σε σχέση με τις άλλες δύο κατηγορίες. Αυτό είναι ένα μειονέκτημα του dataset που είχαμε αλλά δυστυχώς όπως περιγράψαμε κI παραπάνω ήταν αρκετά δύσκολο να βρεθούν επιχειρησιακά δεδομένα. Για αυτόν τον λόγο δουλέψαμε με τα συγκεκριμένα τα οποία παρουσιάζονται στο Παράρτημα.

4.2 Μέθοδος

Πριν από την ανάλυση των αποτελεσμάτων θα εμβαθύνουμε στις μεθόδους που ακολουθήσαμε για να παραχθούν τα αποτελέσματα. Ο στόχος των πειραμάτων αξιολόγησης είναι να δειχθεί ότι η προσέγγιση μας σε αυτό το έγγραφο μπορεί να καθορίσει αξιόπιστα τον βαθμό αυτοματοποίησης των κειμένων διαδικασιών. Για να το πετύχουμε αυτό, έπειτα από όλη την επεξεργασία κειμένου που ακολουθήσαμε όπως έχει περιγραφεί παραπάνω για την συλλογή χαρακτηριστικών, εφαρμόσαμε κάποιους αλγορίθμους μηχανικής μάθησης.

Τα δεδομένα χωρίστηκαν τυχαία σε train και test set ώστε τα αποτελέσματα να είναι έγκυρα. Όσον αφορά τα χαρακτηριστικά των δεδομένων τα οποία ήταν 6 διαφορετικά, δημιουργήσαμε διαφορετικούς συνδυασμούς μεταξύ τους, για να δούμε ποιοι μας επιστρέφουν τα καλύτερα αποτελέσματα. Έτσι θα δούμε αποτελέσματα για κάθε διαφορετικό σύνολο χαρακτηριστικών, ανά αλγόριθμο που χρησιμοποιήθηκε και ανά κωδικοποίηση των δεδομένων. Κάθε τρόπος θα συνοδεύεται με ποσοτικά αποτελέσματα και γραφικές παραστάσεις.

Για να απεικονίσουμε την απόδοση κάθε μεθόδου χρησιμοποιούμε τις μετρικές ακρίβειας precision, recall, και F1-measure. Οι παραπάνω μετρικές υπολογίζονται για κάθε τάξη ξεχωριστά.

Με την μετρική precision υπολογίζουμε για μια συγκεκριμένη τάξη τον αριθμό των εργασιών που έχουν ανατεθεί σωστά σε αυτήν, διαιρούμενο με τον συνολικό αριθμό εργασιών που έχουν ανατεθεί σε αυτήν την τάξη.

$$\text{Precision} = TP/TP+FP$$

Με το recall τον αριθμό των εργασιών που έχουν ανατεθεί σωστά σε αυτήν την τάξη διαιρούμενο με τον συνολικό αριθμό εργασιών που ανήκουν σε αυτήν την τάξη.

$$\text{Recall} = TP/TP+FN$$

Το F1-measure είναι ο αρμονικός μέσος όρος των δύο προηγούμενων.

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

Στις παραπάνω σχέσεις που αφορούν 3 τάξεις έχουμε:

- True Positives (TP): Οι σωστά προβλεπόμενες τιμές που σημαίνει ότι η τιμή της πραγματικής τάξης είναι ίδια με την τιμή της προβλεπόμενης τάξης.
- True Negatives (TN): Πρόκειται για τις πραγματικές τιμές που δεν ανήκουν στην τάξη που μελετάμε, και σωστά προβλέφθηκε ότι δεν ανήκουν σε αυτή την κλάση, ασχέτως αν ταξινομήθηκαν σωστά ή όχι.

- False Positives (FP): Όταν η πραγματική τάξη είναι διαφορετική από την προβλεπόμενη τάξη που είναι η τάξη που μελετάμε.
- False Negatives (FN): Όταν η πραγματική τάξη είναι η τάξη που μελετάμε αλλά η πρόβλεψη είναι λάθος.

4.3 Αποτελέματα

Στα παρακάτω αποτελέσματα που θα παρουσιάσουμε φαίνεται ότι υπάρχουν 8 διαφορετικά data sets. Αυτά τα 8 διαφορετικά σύνολα είναι οι διαφορετικοί συνδυασμοί χαρακτηριστικών των κειμένων. Παρακάτω φαίνεται τι περιλαμβάνει κάθε σύνολο.

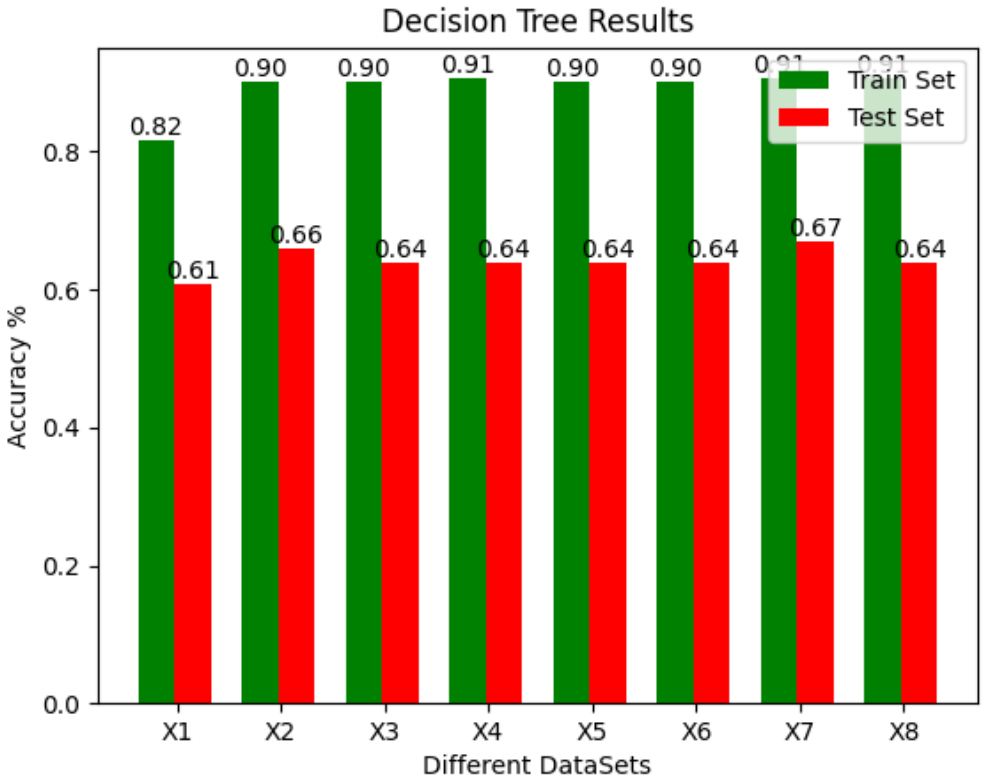
Πίνακας 4.3: *Σύνολα Χαρακτηριστικών*

Set	**Characteristics**
X1	[Ρήμα]
X2	[Ρήμα, Αντικείμενο]
X3	[Ρήμα, Αντικείμενο, Είναι_Πρόσωπο]
X4	[Ρήμα, Αντικείμενο, Είναι_Πρόσωπο, Είναι_Σύστημα]
X5	[Ρήμα, Αντικείμενο, Είναι_Πρόσωπο, Είναι_Σύστημα, Είναι_Τμήμα]
X6	[Ρήμα, Αντικείμενο, Είναι_Πρόσωπο, Λέξη_IT]
X7	[Ρήμα, Αντικείμενο, Είναι_Πρόσωπο, Είναι_Σύστημα, Λέξη_IT]
X8	[Ρήμα, Αντικείμενο, Είναι_Πρόσωπο, Είναι_Σύστημα, Είναι_Τμήμα, Λέξη_IT]

4.3.1 Decision Tree Results

Ο πρώτος αλγόριθμος μηχανικής μάθησης που χρησιμοποιήσαμε είνα το Δέντρο Αποφάσεων - Decision Tree.

Πρώτα θα δούμε τα αποτελέσματα για την κωδικοποίηση δεδομένων με One Hot Encoder.



Σχήμα 4.2: *Decision Tree - One Hot Encoder*

Από το παραπάνω διάγραμμα βλέπουμε πως ενώ στο train set έχουμε πολύ καλά αποτελέσματα της τάξεως του 90% και πάνω, η ακρίβεια στο test set μειώνεται κατά 25% περίπου ανάλογα το σύνολο δεδομένων. Τα αποτελέσματα αυτά παρατηρούνται έπειτα από την προσπάθεια για μείωση του overfitting.

Το **overfitting** είναι ένα σφάλμα μοντελοποίησης που παρουσιάζεται όταν μια συνάρτηση είναι πολύ ακριβής για ένα περιορισμένο σύνολο δεδομένων (στη περιπτωσή μας το train set). Το μοντέλο λαμβάνει γενικά τη μορφή δημιουργίας ενός υπερβολικά περίπλοκου μοντέλου για την εξήγηση των ιδιοσυγκρασιών στα δεδομένα που μελετώνται, χάνοντας όμως την ακρίβεια στη γενική περίπτωση των δεδομένων.

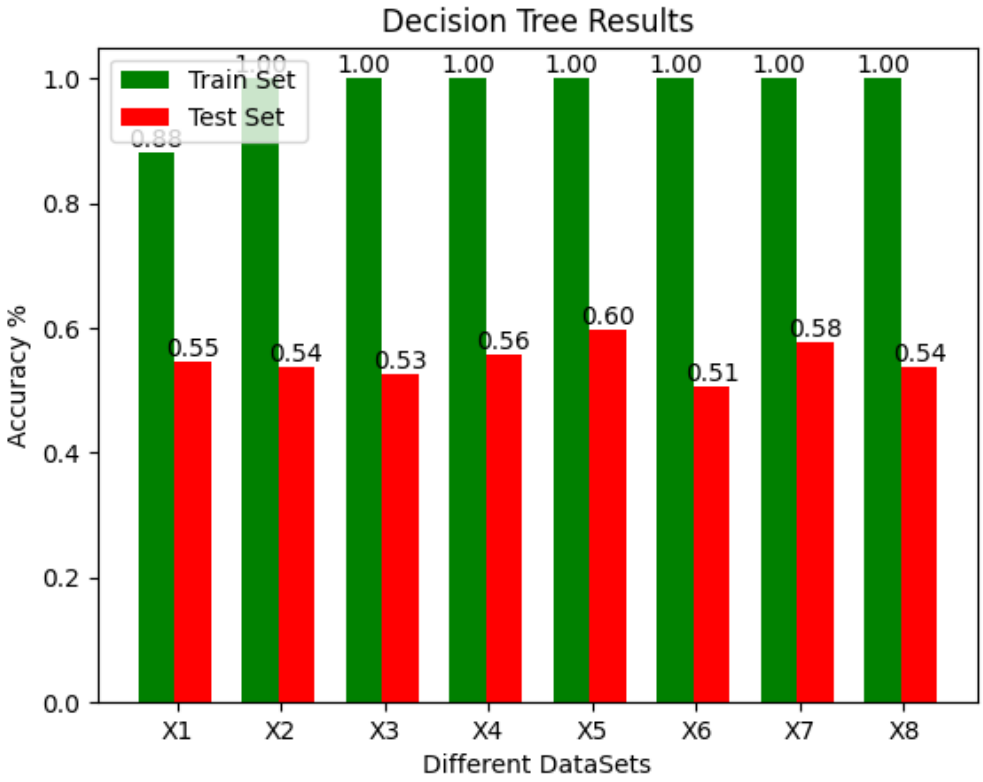
Για να μειώσουμε το φαινόμενο αυτό στη συγκεκριμένη περίπτωση με τα Δέντρα Αποφάσεων, μείωσαμε το βάθος του δέντρου κατά 25% περίπου του αρχικού. Στο παρακάτω πινακάκι φαίνονται οι τιμές για τα βάθη των δέντρων:

Πίνακας 4.4: *Depth of Decision Trees*

Set	Initial Depth	Reduced Depth
X1	76	57
X2	73	55
X3	73	55
X4	74	56
X5	73	55
X6	73	55
X7	74	56
X8	74	56

Έτσι καταφέραμε να έχουμε τα παραπάνω βαλτιωμένα αποτελεσματα στο test set.

Παρακάτω θα δούμε τα αποτελέσματα των Δέντρων Αποφάσεων έχοντας χρησιμοποιήσει τα Word Embeddings για κωδικοποίηση.



Σχήμα 4.3: *Decision Tree - Word Embeddings*

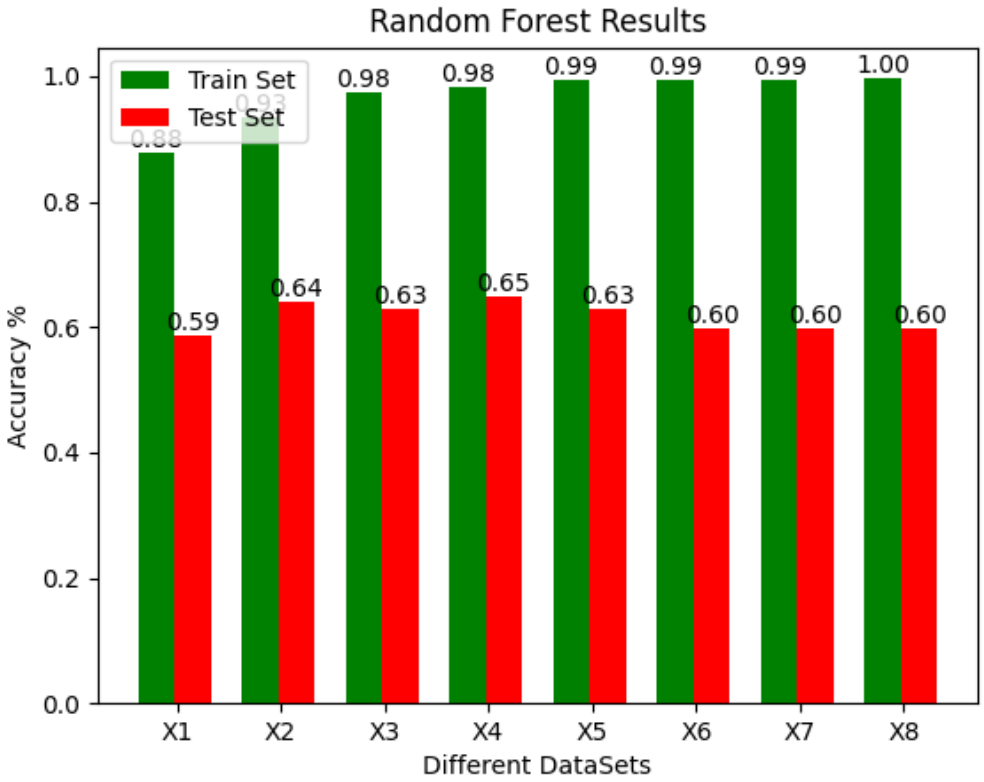
Παραδόξως παρατηρούμε ότι τα αποτελέσματα είναι χειρότερα με την χρήση των Word Embeddings σε σχέση με την χρήση του One Hot Encoder. Και στην συγκεκριμένη περίπτωση έχουμε χρησιμοποιήσει την μέθοδο μείωσης του βάθους των δέντρων για την αποφυγή του overfitting, καθώς η ακρίβεια στο test set ήταν ακόμη μικρότερη.

Ωστόσο, αυτό που είναι κοινό και στις δύο περίπτωσεις είναι πως το σύνολο X1 που αποτελείται μόνο από το ρήμα σαν χαρακτηριστικό, λειτουργεί ως βάση των αποτελεσμάτων, καθώς η προσθήκη κι άλλων χαρακτηριστικών μεταβάλλει ελάχιστα τα αποτελέσματα. Από αυτό προκύπτει πως το ρήμα ·ίσως να αποτελεί το πιο σημαντικο χαρακτηριστικό για την ταξινόμηση των εργασιών.

4.3.2 Random Forest Results

Επόμενος αλγόριθμος που δοκιμάσαμε ήταν το Τυχαίο Δάσος - Random Forest. Όπως έχουμε αναφέρει παραπάνω είναι μια επέκταση του δέντρου αποφάσεων, με την έννοια ότι, πρώτα δημιουργεί δέντρα αποφάσεων σε πραγματικό κόσμο μερικών αξόνων με δεδομένα εκπαίδευσης και, στη συνέχεια, ταιριάζει τα νέα δεδομένα σε ένα από τα δέντρα ως «τυχαίο δάσος».

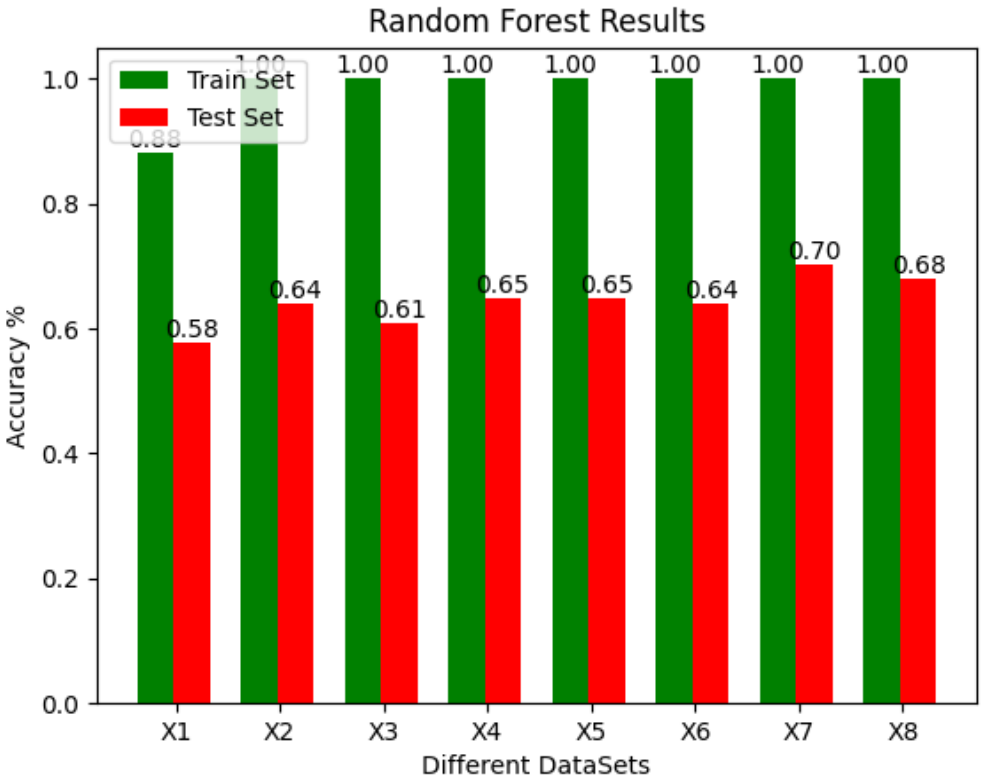
Θα δούμε πρώτα τα αποτελέσματα για κωδικοποίηση με One Hot Encoder.



Σχήμα 4.4: *Random Forest - One Hot Encoder*

Και στη συγκεκριμένη περίπτωση για να μειώσουμε το φαινόμενο του overfitting μείωσαμε το βάθος τών δέντρων κατά 25% περίπου του αρχικού. Είναι φανερό πως τα αποτελέσματα είναι κατά ένα μικρό ποσοστό χειρότερα από τα Δέντρα αποφάσεων επίσης με την χρήση του One Hot Encoder.

Παρακάτω θα δούμε τα αποτελέσματα των Τυχαίων Δασών έχοντας χρησιμοποιήσει τα Word Embeddings για κωδικοποίηση.



Σχήμα 4.5: *Random Forest - Word Embeddings*

Στη συγκεκριμένη περίπτωση βλέπουμε μια ελάχιστη μείωση της ακρίβειας σε σχέση με τη χρήση του One Hot Encoder μόνο στο σύνολο X3. Στα υπόλοιπα σύνολα το ποσοστό ακρίβειας διατηρείται σταθερό ή και αυξάνεται ελάχιστα. Συνολικά έως τώρα παρατηρούμε πως η καλύτερη απόδοση ακρίβειας βρίσκεται σε αυτήν την περίπτωση με ποσοστό 70% στο σύνολο X7.

4.3.3 Support Vector Machine Results

Επόμενη κατηγορία αλγόριθμου μηχανικής μάθησης που υλοποιήσαμε είναι ο Support Vector Machine - SVM ή στα ελληνικά μηχάνημα φορέα υποστήριξης.

Η βασική στρατηγική ενός SVM είναι να βρει το λεγόμενο hyperplane που διαιρεί καλύτερα ένα σύνολο δεδομένων. Ενώ σε δισδιάστατο χώρο, μια απλή γραμμή θα ήταν αρκετή, ένα SVM χαρτογραφεί τα δεδομένα σε υψηλότερες και υψηλότερες διαστάσεις μέχρι ένα hyperplane να μπορεί να σχηματιστεί που διαχωρίζει σαφώς τα δεδομένα. Δεδομένου ότι ένα SVM είναι αλγόριθμος εποπτευόμενης μηχανικής μάθησης, πρέπει να εκπαιδευτεί σε ένα σύνολο δεδομένων με ετικέτα με μη αυτόματο τρόπο.

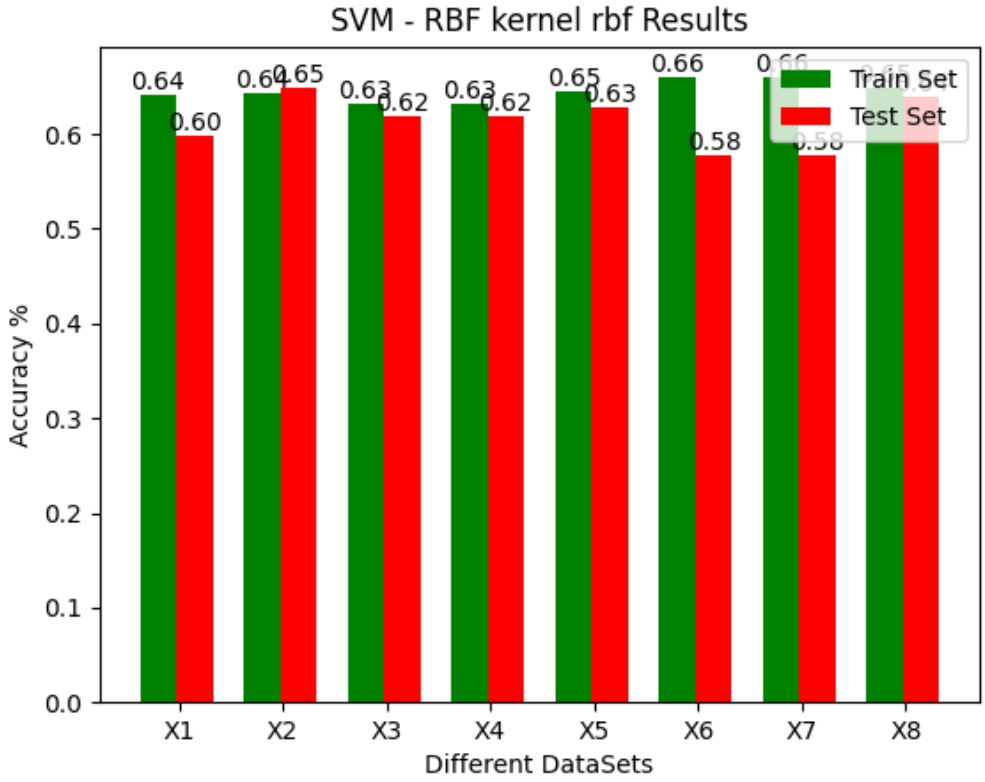
Στον αλγόριθμο αυτό πρέπει να οριστούν διάφορες παράμετροι οι οποίες είναι οι kernel, c, gamma, cv.

- kernel: Καθορίζει τον τύπο του πυρήνα που θα χρησιμοποιηθεί στον αλγόριθμο. Πρέπει να είναι ένα από τα «linear», «poly», «rbf», «sigmoid», «precomputed». Εάν δεν δοθεί κανένα, θα χρησιμοποιηθεί το «rbf». Στην δική μας περίπτωση έγιναν δοκιμές με τα «poly», «rbf», «sigmoid».
- c: Πρόκειται για παράμετρο κανονικοποίησης. Η ισχύς της κανονικοποίησης είναι αντιστρόφως ανάλογη με το c. Πρέπει να είναι αυστηρά θετικό. Εμείς κάναμε δοκιμές για τιμές $\log_2 C$ ε {-5,...,15}.
- gamma: Συντελεστής πυρήνα για «poly», «rbf», «sigmoid». Εμείς κάναμε δοκιμές για τιμές $\log_2 \gamma$ ε {-15,...,3}.
- cv: Καθορίζει τη στρατηγική του cross validation. Η ιδέα πίσω από αυτήν την προσέγγιση επικύρωσης είναι να χωριστεί τυχαία το σύνολο δεδομένων σε X αμοιβαία αποκλειστικά υποσύνολα (τα λεγόμενα folds) ίσου μεγέθους. Εμείς κάναμε δοκιμές με τιμές cv=5 και cv=10.

CV = 5 ||*OneHotEncoding*

Θα παρουσιάσουμε πρώτα τα αποτελέσματα για τον πυρήνα RBF.

Στο παρακάτω διάγραμμα βλέπουμε τα ποσοστά ακρίβειας για κάθε σύνολο δεδομένων. Σαν πρώτη παρατήρηση βλέπουμε παρόμοια ποσοστά ακρίβειας με τους προηγούμενους αλγορίθμους. Φαίνεται ωστόσο πως έχει εξαληφθεί το φαινόμενο του overfitting καθώς τα νούμερα των train και test set είναι πολύ κοντινά.

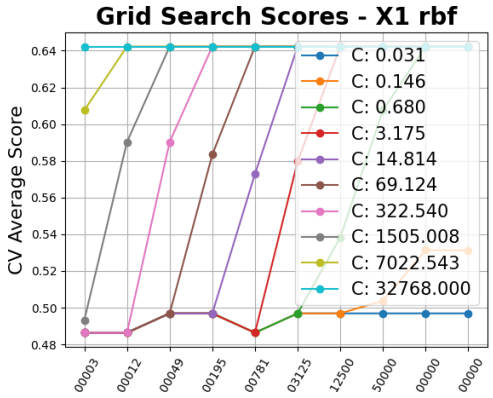


Σχήμα 4.6: *SVM RBF - CV 5 - One Hot Encoder*

Για κάθε σύνολο δεδομένων αναφερόμαστε σε διαφορετικές παραμέτρους c και gamma ανάλογα τα καλύτερα αποτελέσματα που έχουν επιστρέψει. Για την εύρεση του καλύτερου ζεύγους παραμέτρων έχουμε χρησιμοποιήσει την μέθοδο του Grid Search.

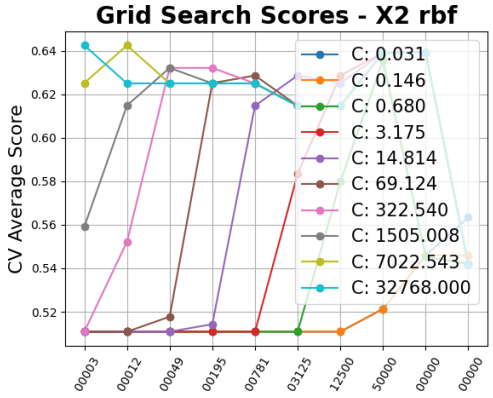
Πίνακας 4.5: *Best Parameters from Grid Search - RBF kernel - CV 5 - One Hot Encoding*

Set	**Parameter C**	**Parameter Gamma**
X1	0.6803950000871886	2.0
X2	7022.542707532377	0.0001220703125
X3	14.81399539659665	0.125
X4	14.81399539659665	0.125
X5	69.12382328910758	0.0078125
X6	14.81399539659665	0.125
X7	14.81399539659665	0.125
X8	14.81399539659665	0.03125



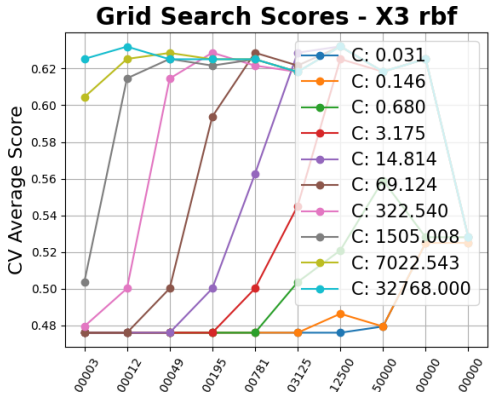
Detailed classification report:

	precision	recall	f1-score	support
0	0.56	0.82	0.67	45
1	0.68	0.50	0.58	38
2	0.67	0.14	0.24	14
accuracy			0.60	97
macro avg	0.64	0.49	0.49	97
weighted avg	0.62	0.60	0.57	97



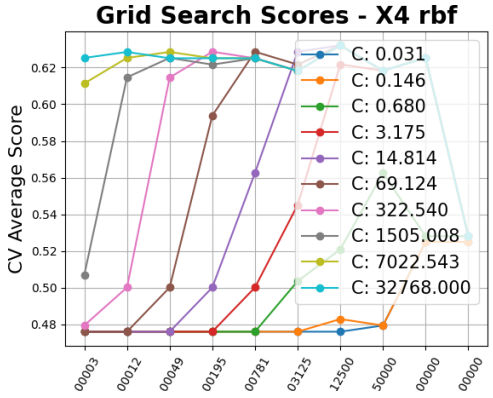
Detailed classification report:

	precision	recall	f1-score	support
0	0.67	0.76	0.71	45
1	0.65	0.58	0.61	38
2	0.58	0.50	0.54	14
accuracy			0.65	97
macro avg	0.63	0.61	0.62	97
weighted avg	0.65	0.65	0.65	97



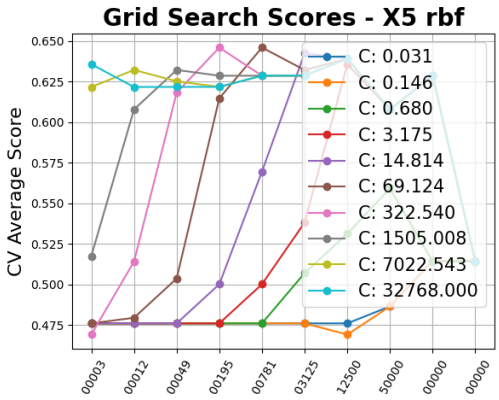
Detailed classification report:

	precision	recall	f1-score	support
0	0.65	0.69	0.67	45
1	0.59	0.58	0.59	38
2	0.58	0.50	0.54	14
accuracy			0.62	97
macro avg	0.61	0.59	0.60	97
weighted avg	0.62	0.62	0.62	97



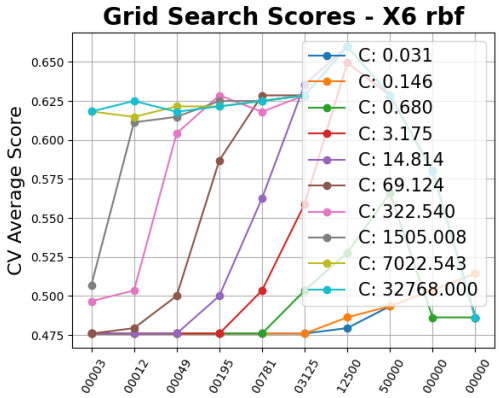
Detailed classification report:

	precision	recall	f1-score	support
0	0.65	0.69	0.67	45
1	0.59	0.58	0.59	38
2	0.58	0.50	0.54	14
accuracy			0.62	97
macro avg	0.61	0.59	0.60	97
weighted avg	0.62	0.62	0.62	97



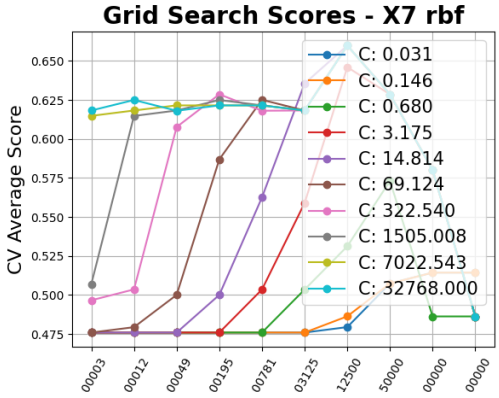
Detailed classification report:

	precision	recall	f1-score	support
0	0.65	0.71	0.68	45
1	0.61	0.58	0.59	38
2	0.58	0.50	0.54	14
accuracy			0.63	97
macro avg	0.62	0.60	0.60	97
weighted avg	0.63	0.63	0.63	97



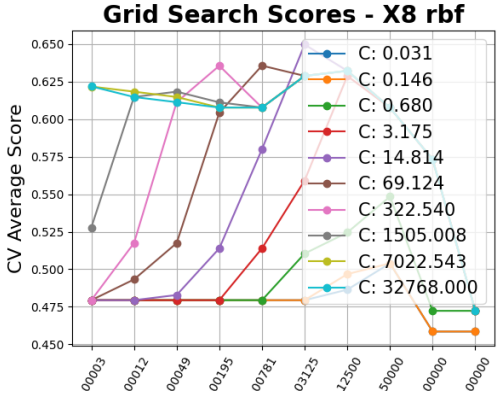
Detailed classification report:

	precision	recall	f1-score	support
0	0.62	0.58	0.60	45
1	0.53	0.55	0.54	38
2	0.60	0.64	0.62	14
accuracy			0.58	97
macro avg	0.58	0.59	0.59	97
weighted avg	0.58	0.58	0.58	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.62	0.58	0.60	45
1	0.53	0.55	0.54	38
2	0.60	0.64	0.62	14
accuracy			0.58	97
macro avg	0.58	0.59	0.59	97
weighted avg	0.58	0.58	0.58	97

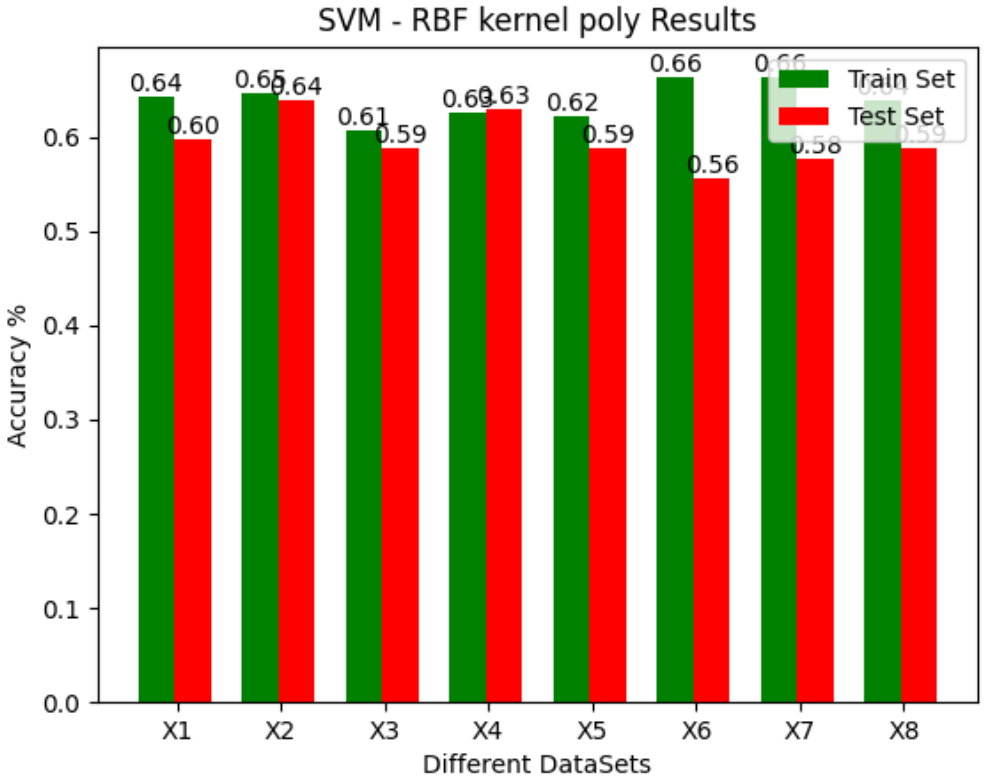


Detailed classification report:

	precision	recall	f1-score	support
0	0.66	0.73	0.69	45
1	0.62	0.55	0.58	38
2	0.62	0.57	0.59	14
accuracy			0.64	97
macro avg	0.63	0.62	0.62	97
weighted avg	0.64	0.64	0.64	97

Σχήμα 4.7: *Grid Search for RBF kernel - CV 5 - One Hot Encoder*

Συνεχίζουμε με τα αποτελέσματα για τον πυρήνα POLY. Το φαινόμενο του overfitting είναι ελάχιστο και εδώ καθώς τα νούμερα των train και test set είναι πολύ κοντά. Γενικά βλέπουμε μια πολύ μικρή μείωση στην ακρίβεια σε σχέση με τον πυρήνα RBF.

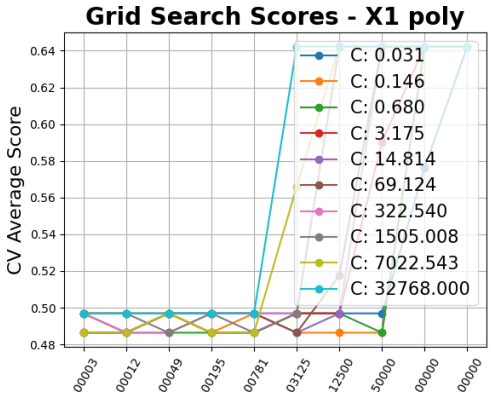


Σχήμα 4.8: *SVM POLY - CV 5 - One Hot Encoder*

Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει επίσης την μέθοδο του Grid Search.

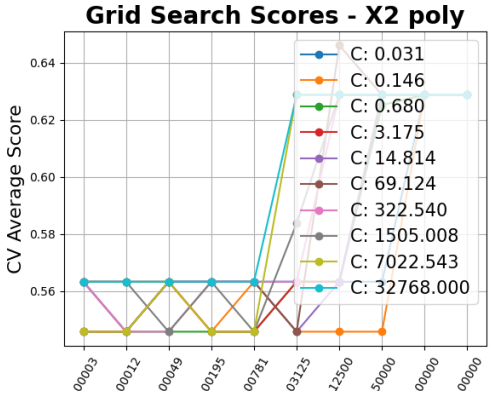
Πίνακας 4.6: *Best Parameters from Grid Search - POLY kernel - CV 5 - One Hot Encoding*

Set	**Parameter C**	**Parameter Gamma**
X1	322.53978877308765	0.125
X2	69.12382328910758	0.125
X3	14.81399539659665	0.125
X4	0.03125	0.125
X5	0.03125	2.0
X6	1505.0081200902148	0.03125
X7	0.03125	2.0
X8	322.53978877308765	0.03125



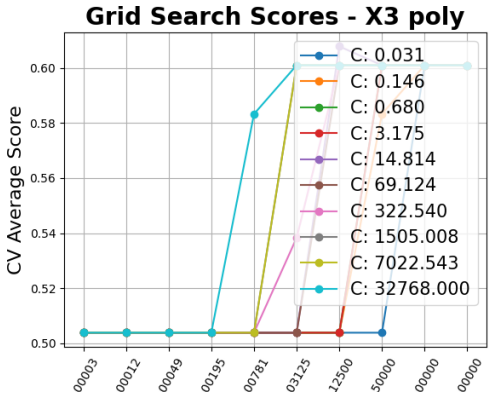
Detailed classification report:

	precision	recall	f1-score	support
0	0.56	0.82	0.67	45
1	0.68	0.50	0.58	38
2	0.67	0.14	0.24	14
accuracy			0.60	97
macro avg	0.64	0.49	0.49	97
weighted avg	0.62	0.60	0.57	97



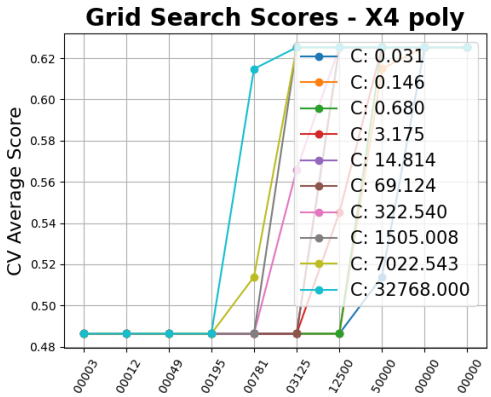
Detailed classification report:

	precision	recall	f1-score	support
0	0.59	0.89	0.71	45
1	0.73	0.50	0.59	38
2	1.00	0.21	0.35	14
accuracy			0.64	97
macro avg	0.77	0.53	0.55	97
weighted avg	0.70	0.64	0.61	97



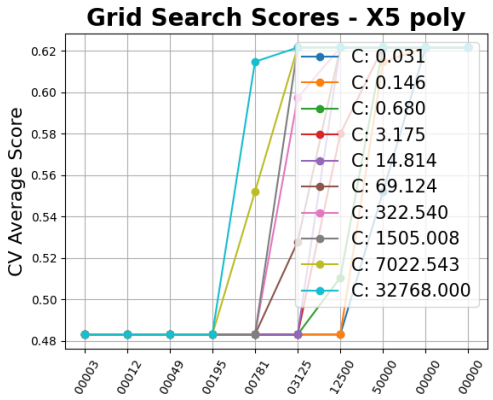
Detailed classification report:

	precision	recall	f1-score	support
0	0.55	0.84	0.67	45
1	0.64	0.42	0.51	38
2	1.00	0.21	0.35	14
accuracy			0.59	97
macro avg	0.73	0.49	0.51	97
weighted avg	0.65	0.59	0.56	97

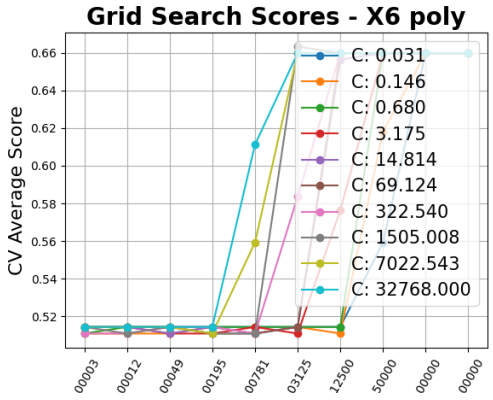


Detailed classification report:

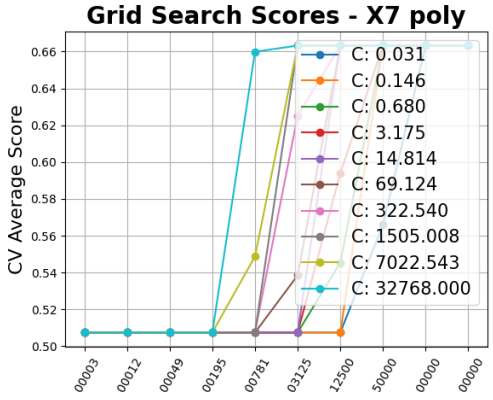
	precision	recall	f1-score	support
0	0.62	0.76	0.68	45
1	0.64	0.55	0.59	38
2	0.67	0.43	0.52	14
accuracy			0.63	97
macro avg	0.64	0.58	0.60	97
weighted avg	0.63	0.63	0.62	97



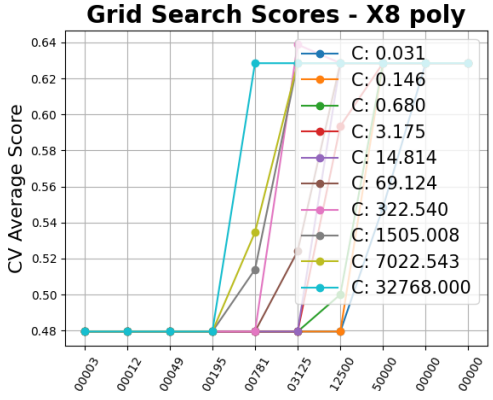

```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.58      0.71      0.64        45
           1       0.55      0.47      0.51        38
           2       0.78      0.50      0.61        14

    accuracy                           0.59        97
   macro avg       0.64      0.56      0.59        97
weighted avg       0.60      0.59      0.58        97
```




```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.59      0.60      0.59        45
           1       0.51      0.53      0.52        38
           2       0.58      0.50      0.54        14

    accuracy                           0.56        97
   macro avg       0.56      0.54      0.55        97
weighted avg       0.56      0.56      0.56        97
```



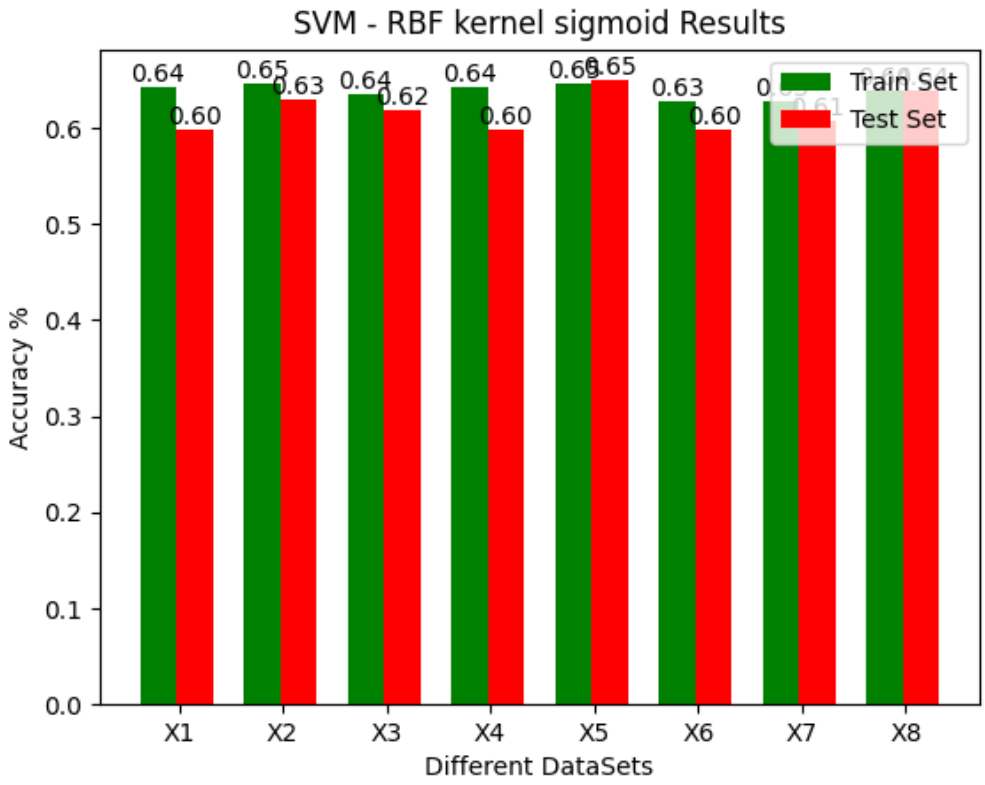

```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.61      0.60      0.61        45
           1       0.54      0.55      0.55        38
           2       0.57      0.57      0.57        14

    accuracy                           0.58        97
   macro avg       0.57      0.57      0.57        97
weighted avg       0.58      0.58      0.58        97
```




```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.60      0.69      0.64        45
           1       0.56      0.47      0.51        38
           2       0.62      0.57      0.59        14

    accuracy                           0.59        97
   macro avg       0.59      0.58      0.58        97
weighted avg       0.59      0.59      0.58        97
```

Σχήμα 4.9: *Grid Search for POLY kernel - CV 5 - One Hot Encoder*

Τέλος παρουσιάζουμε τα αποτελέσματα για τον πυρήνα SIGMOID. Το φαινόμενο του overfitting είναι ελάχιστο και εδώ καθώς τα νούμερα των train και test set είναι πολύ κοντινά. Παρατηρούμε μια μικρή άνοδο ακρίβειας σε σχέση με τον πυρήνα POLY.

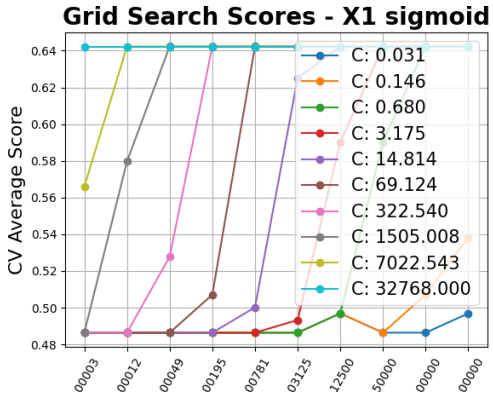


Σχήμα 4.10: *SVM SIGMOID - CV 5 - One Hot Encoder*

Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει επίσης την μέθοδο του Grid Search.

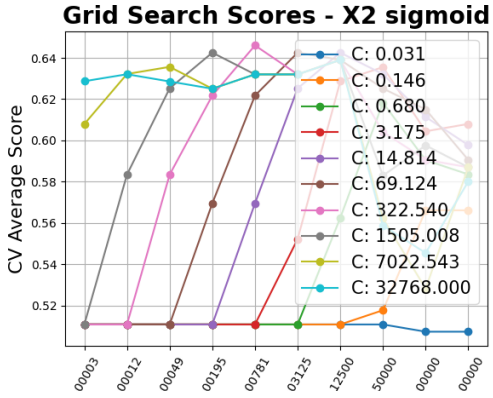
Πίνακας 4.7: *Best Parameters from Grid Search - SIGMOID kernel - CV 5 - One Hot Encoding*

Set	**Parameter C**	**Parameter Gamma**
X1	0.6803950000871886	2.0
X2	322.53978877308765	0.0078125
X3	7022.542707532377	0.00048828125
X4	69.12382328910758	0.125
X5	7022.542707532377	0.0001220703125
X6	1505.0081200902148	0.03125
X7	1505.0081200902148	0.03125
X8	32768.0	3.0517578125ε-05



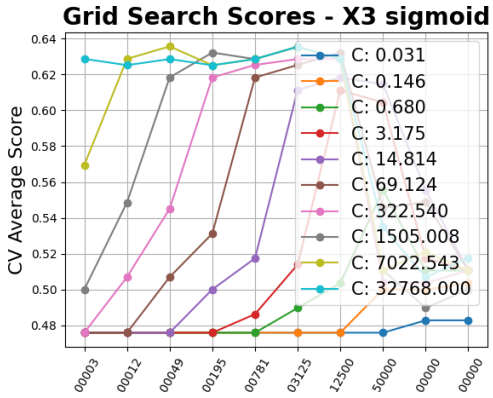
Detailed classification report:

	precision	recall	f1-score	support
0	0.56	0.82	0.67	45
1	0.68	0.50	0.58	38
2	0.67	0.14	0.24	14
accuracy			0.60	97
macro avg	0.64	0.49	0.49	97
weighted avg	0.62	0.60	0.57	97



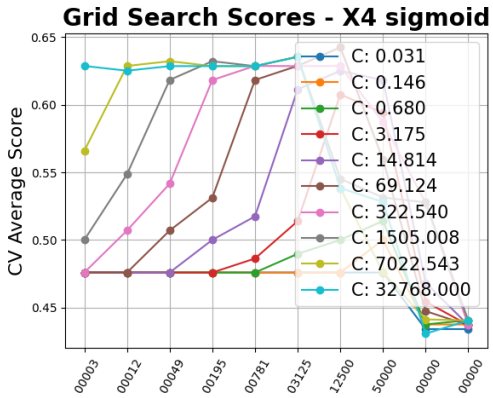
Detailed classification report:

	precision	recall	f1-score	support
0	0.65	0.71	0.68	45
1	0.61	0.58	0.59	38
2	0.58	0.50	0.54	14
accuracy			0.63	97
macro avg	0.62	0.60	0.60	97
weighted avg	0.63	0.63	0.63	97



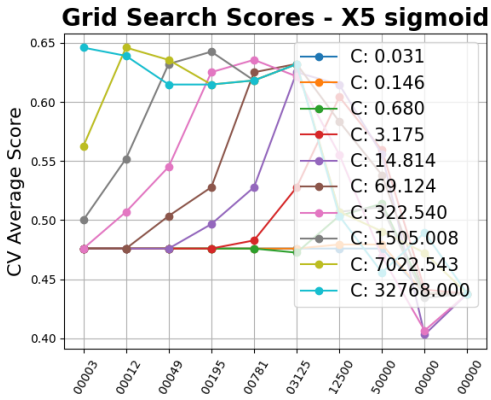
Detailed classification report:

	precision	recall	f1-score	support
0	0.65	0.69	0.67	45
1	0.59	0.58	0.59	38
2	0.58	0.50	0.54	14
accuracy			0.62	97
macro avg	0.61	0.59	0.60	97
weighted avg	0.62	0.62	0.62	97



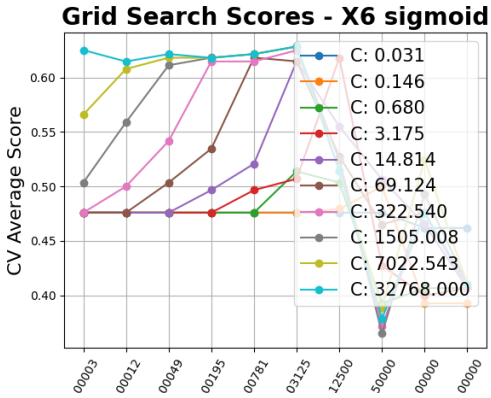
Detailed classification report:

	precision	recall	f1-score	support
0	0.63	0.64	0.64	45
1	0.56	0.58	0.57	38
2	0.58	0.50	0.54	14
accuracy			0.60	97
macro avg	0.59	0.57	0.58	97
weighted avg	0.60	0.60	0.60	97



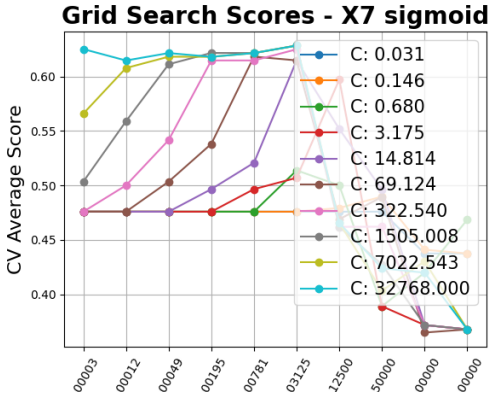
Detailed classification report:

	precision	recall	f1-score	support
0	0.67	0.73	0.70	45
1	0.63	0.58	0.60	38
2	0.62	0.57	0.59	14
accuracy			0.65	97
macro avg	0.64	0.63	0.63	97
weighted avg	0.65	0.65	0.65	97



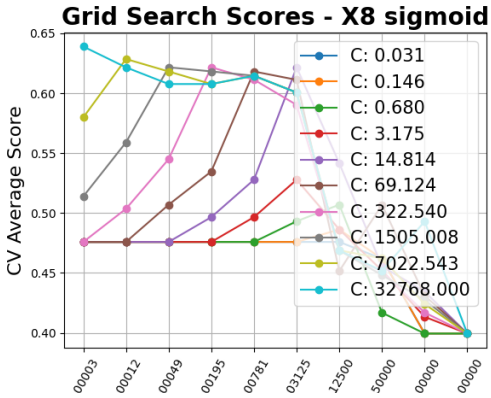
Detailed classification report:

	precision	recall	f1-score	support
0	0.62	0.69	0.65	45
1	0.57	0.53	0.55	38
2	0.58	0.50	0.54	14
accuracy			0.60	97
macro avg	0.59	0.57	0.58	97
weighted avg	0.60	0.60	0.60	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.63	0.69	0.66	45
1	0.58	0.55	0.57	38
2	0.58	0.50	0.54	14
accuracy			0.61	97
macro avg	0.60	0.58	0.59	97
weighted avg	0.61	0.61	0.61	97



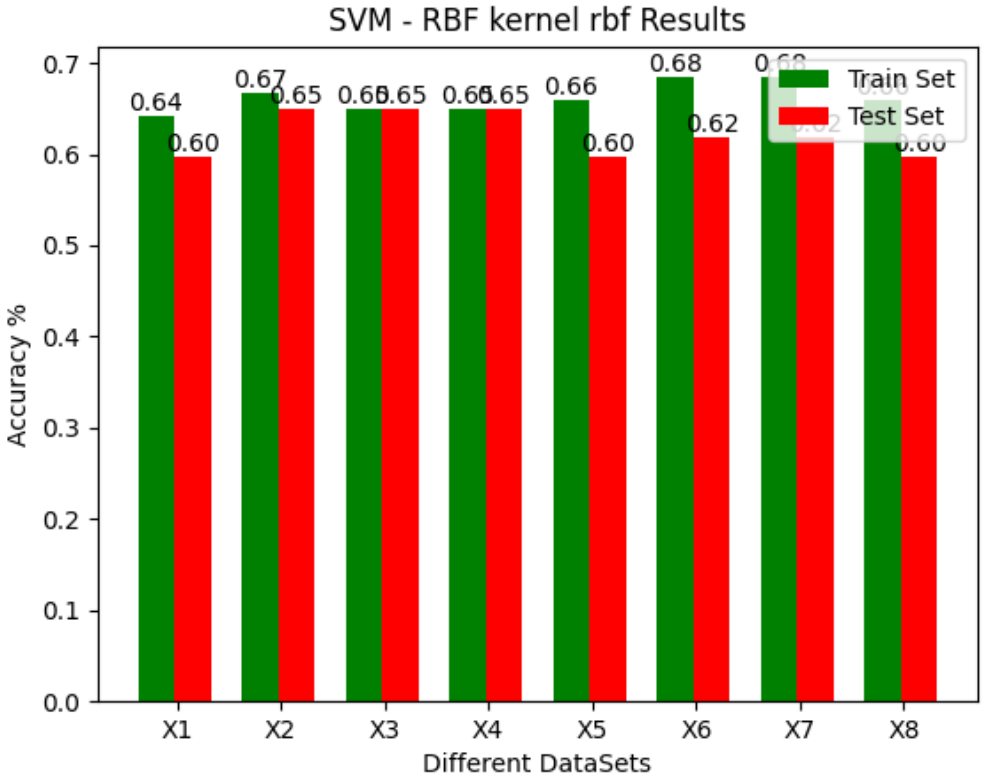
Detailed classification report:

	precision	recall	f1-score	support
0	0.67	0.71	0.69	45
1	0.61	0.58	0.59	38
2	0.62	0.57	0.59	14
accuracy			0.64	97
macro avg	0.63	0.62	0.63	97
weighted avg	0.64	0.64	0.64	97

Σχήμα 4.11: *Grid Search for SIGMOID kernel - CV 5 - One Hot Encoder*

CV = 10 ||*OneHotEncoding*

Συνεχίζουμε με τις δοκιμές για cv=10. Παρουσιάζουμε πρώτα τα αποτελέσματα για τον πυρήνα RBF. Σαν πρώτη παρατήρηση βλέπουμε παρόμοια ποσοστά ακρίβειας με την προηγούμεη τιμή του cv.

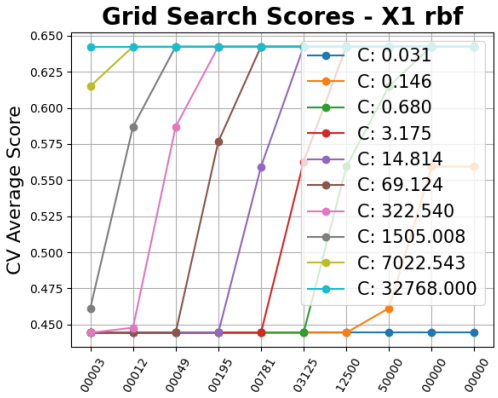


Σχήμα 4.12: *SVM RBF - CV 10 - One Hot Encoder*

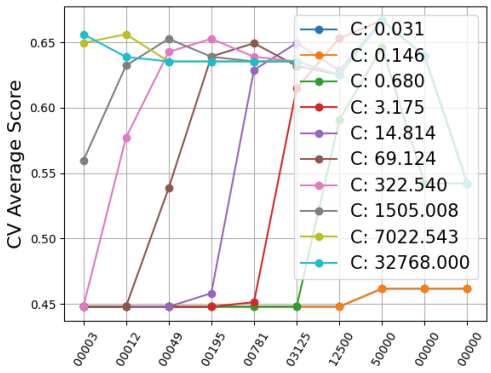
Για κάθε σύνολο δεδομένων αναφερόμαστε σε διαφορετικές παραμέτρους c και gamma ανάλογα τα καλύτερα αποτελέσματα που έχουν επιστρέψει. Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει την μέθοδο του Grid Search.

Πίνακας 4.8: *Best Parameters from Grid Search - RBF kernel - CV 10 - One Hot Encoding*

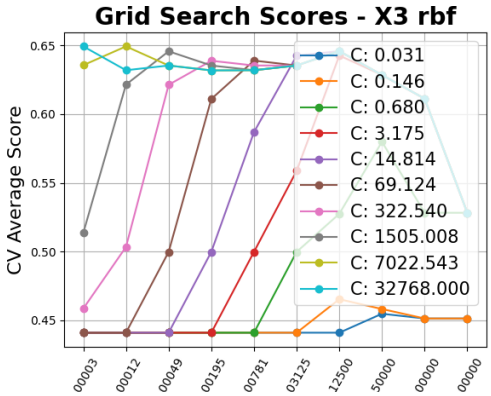
Set	**Parameter C**	**Parameter Gamma**
X1	0.6803950000871886	2.0
X2	3.1748021039363996	0.5
X3	7022.542707532377	0.0001220703125
X4	1505.0081200902148	0.00048828125
X5	14.81399539659665	0.125
X6	3.1748021039363996	0.125
X7	3.1748021039363996	0.125
X8	3.1748021039363996	0.125



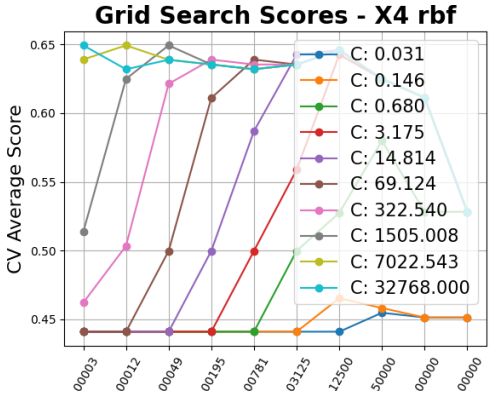

```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.56      0.82      0.67        45
           1       0.68      0.50      0.58        38
           2       0.67      0.14      0.24        14

    accuracy                           0.60        97
   macro avg       0.64      0.49      0.49        97
weighted avg       0.62      0.60      0.57        97
```



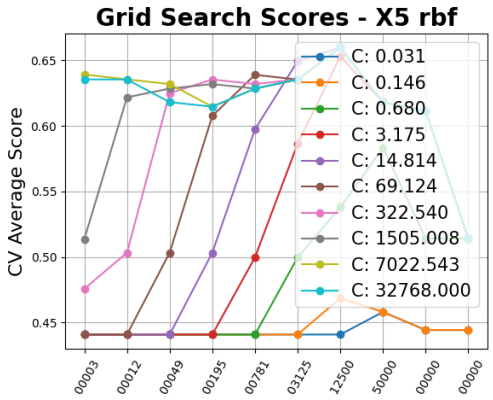

```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.62      0.80      0.70        45
           1       0.69      0.58      0.63        38
           2       0.71      0.36      0.48        14

    accuracy                           0.65        97
   macro avg       0.67      0.58      0.60        97
weighted avg       0.66      0.65      0.64        97
```




```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.67      0.76      0.71        45
           1       0.65      0.58      0.61        38
           2       0.58      0.50      0.54        14

    accuracy                           0.65        97
   macro avg       0.63      0.61      0.62        97
weighted avg       0.65      0.65      0.65        97
```

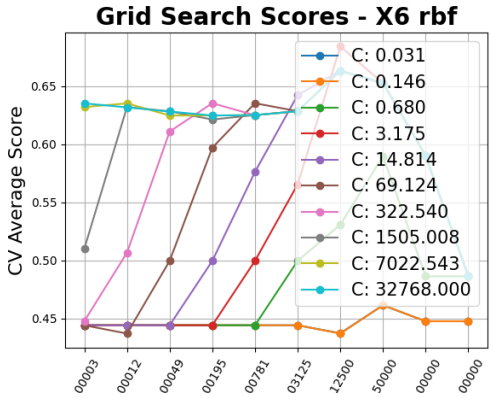



```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.67      0.76      0.71        45
           1       0.65      0.58      0.61        38
           2       0.58      0.50      0.54        14

    accuracy                           0.65        97
   macro avg       0.63      0.61      0.62        97
weighted avg       0.65      0.65      0.65        97
```



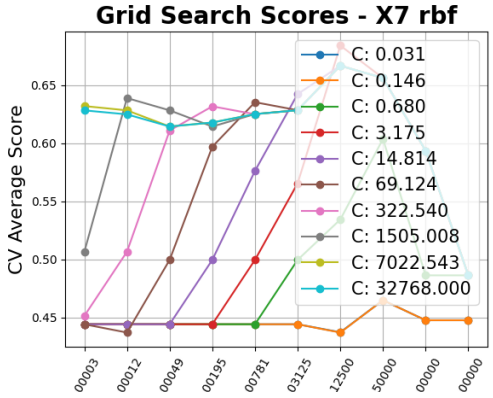
Detailed classification report:

	precision	recall	f1-score	support
0	0.60	0.71	0.65	45
1	0.58	0.50	0.54	38
2	0.64	0.50	0.56	14
accuracy			0.60	97
macro avg	0.61	0.57	0.58	97
weighted avg	0.60	0.60	0.59	97



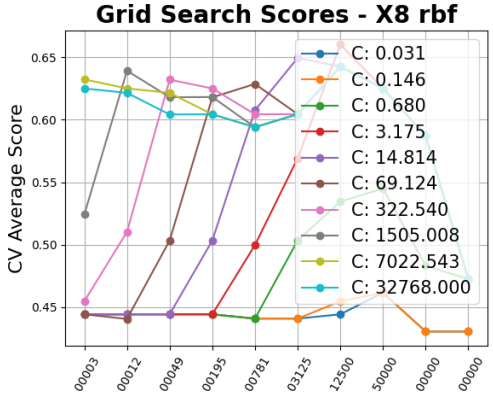
Detailed classification report:

	precision	recall	f1-score	support
0	0.64	0.71	0.67	45
1	0.60	0.55	0.58	38
2	0.58	0.50	0.54	14
accuracy			0.62	97
macro avg	0.61	0.59	0.60	97
weighted avg	0.62	0.62	0.62	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.64	0.71	0.67	45
1	0.60	0.55	0.58	38
2	0.58	0.50	0.54	14
accuracy			0.62	97
macro avg	0.61	0.59	0.60	97
weighted avg	0.62	0.62	0.62	97

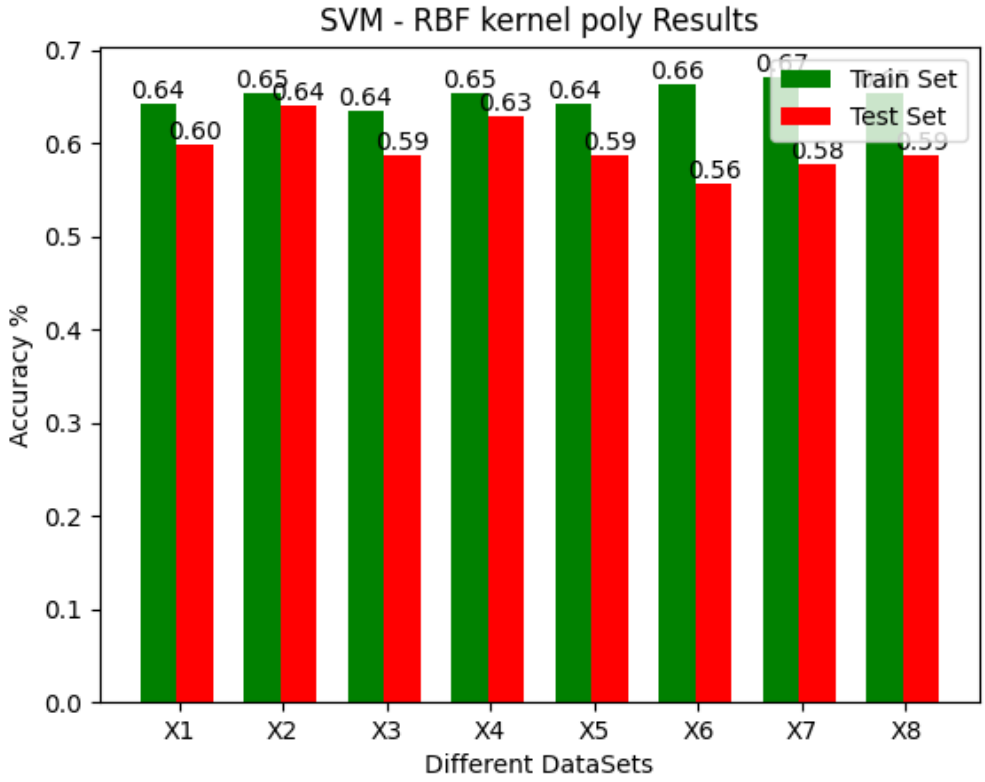


Detailed classification report:

	precision	recall	f1-score	support
0	0.58	0.73	0.65	45
1	0.59	0.50	0.54	38
2	0.75	0.43	0.55	14
accuracy			0.60	97
macro avg	0.64	0.55	0.58	97
weighted avg	0.61	0.60	0.59	97

Σχήμα 4.13: *Grid Search for RBF kernel - CV 10 - One Hot Encoder*

Συνεχίζουμε με τα αποτελέσματα για τον πυρήνα POLY. Στο παρακάτω διάγραμμα βλέπουμε τα ποσοστά ακρίβειας για κάθε σύνολο δεδομένων. Το φαινόμενο του overfitting είναι ελάχιστο και εδώ καθώς τα νούμερα των train και test set είναι πολύ κοντινά.

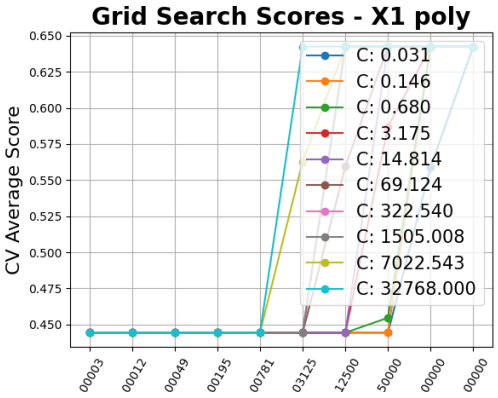


Σχήμα 4.14: *SVM POLY - CV 10 - One Hot Encoder*

Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει επίσης την μέθοδο του Grid Search.

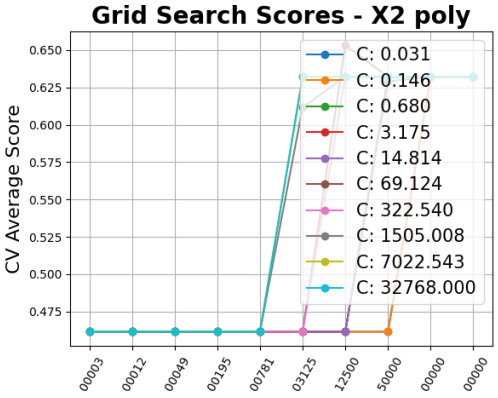
Πίνακας 4.9: *Best Parameters from Grid Search - POLY kernel - CV 10 - One Hot Encoding*

Set	**Parameter C**	**Parameter Gamma**
X1	322.53978877308765	0.125
X2	69.12382328910758	0.125
X3	14.81399539659665	0.125
X4	0.03125	2.0
X5	0.03125	2.0
X6	0.03125	2.0
X7	0.03125	2.0
X8	322.53978877308765	0.03125



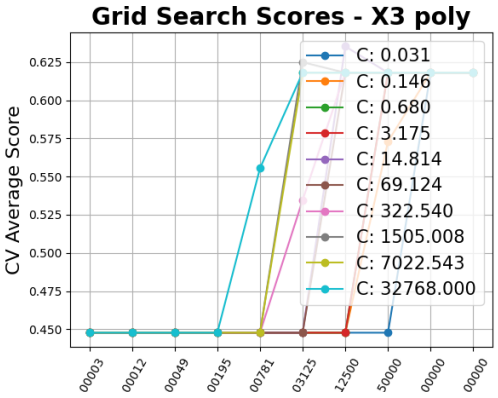
Detailed classification report:

	precision	recall	f1-score	support
0	0.56	0.82	0.67	45
1	0.68	0.50	0.58	38
2	0.67	0.14	0.24	14
accuracy			0.60	97
macro avg	0.64	0.49	0.49	97
weighted avg	0.62	0.60	0.57	97



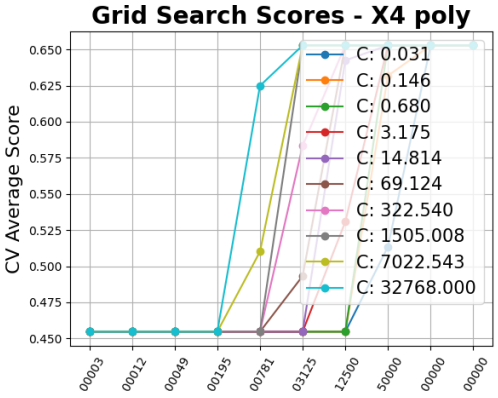
Detailed classification report:

	precision	recall	f1-score	support
0	0.59	0.89	0.71	45
1	0.73	0.50	0.59	38
2	1.00	0.21	0.35	14
accuracy			0.64	97
macro avg	0.77	0.53	0.55	97
weighted avg	0.70	0.64	0.61	97



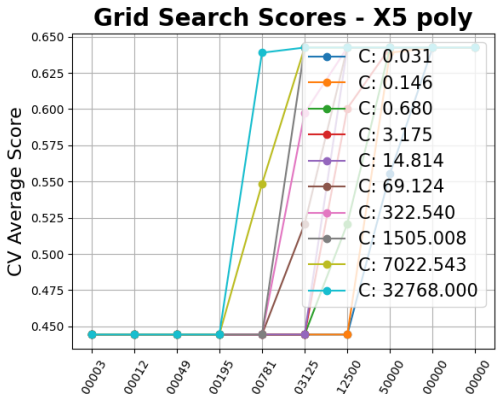
Detailed classification report:

	precision	recall	f1-score	support
0	0.55	0.84	0.67	45
1	0.64	0.42	0.51	38
2	1.00	0.21	0.35	14
accuracy			0.59	97
macro avg	0.73	0.49	0.51	97
weighted avg	0.65	0.59	0.56	97



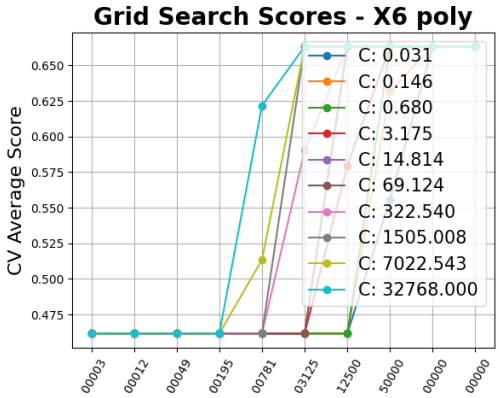
Detailed classification report:

	precision	recall	f1-score	support
0	0.62	0.76	0.68	45
1	0.64	0.55	0.59	38
2	0.67	0.43	0.52	14
accuracy			0.63	97
macro avg	0.64	0.58	0.60	97
weighted avg	0.63	0.63	0.62	97



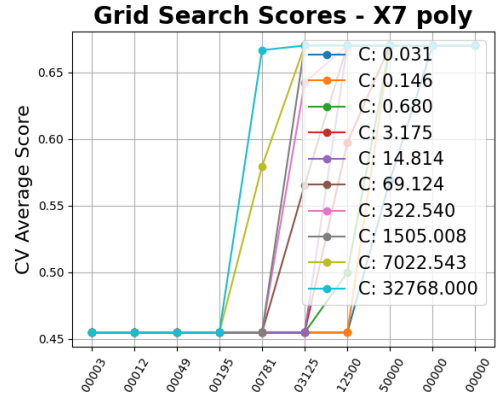
Detailed classification report:

	precision	recall	f1-score	support
0	0.58	0.71	0.64	45
1	0.55	0.47	0.51	38
2	0.78	0.50	0.61	14
accuracy			0.59	97
macro avg	0.64	0.56	0.59	97
weighted avg	0.60	0.59	0.58	97



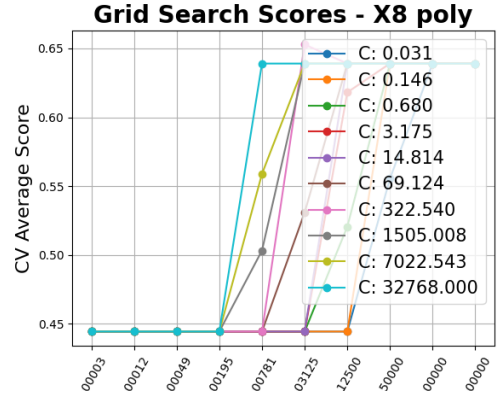
Detailed classification report:

	precision	recall	f1-score	support
0	0.59	0.60	0.59	45
1	0.51	0.53	0.52	38
2	0.58	0.50	0.54	14
accuracy			0.56	97
macro avg	0.56	0.54	0.55	97
weighted avg	0.56	0.56	0.56	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.61	0.60	0.61	45
1	0.54	0.55	0.55	38
2	0.57	0.57	0.57	14
accuracy			0.58	97
macro avg	0.57	0.57	0.57	97
weighted avg	0.58	0.58	0.58	97

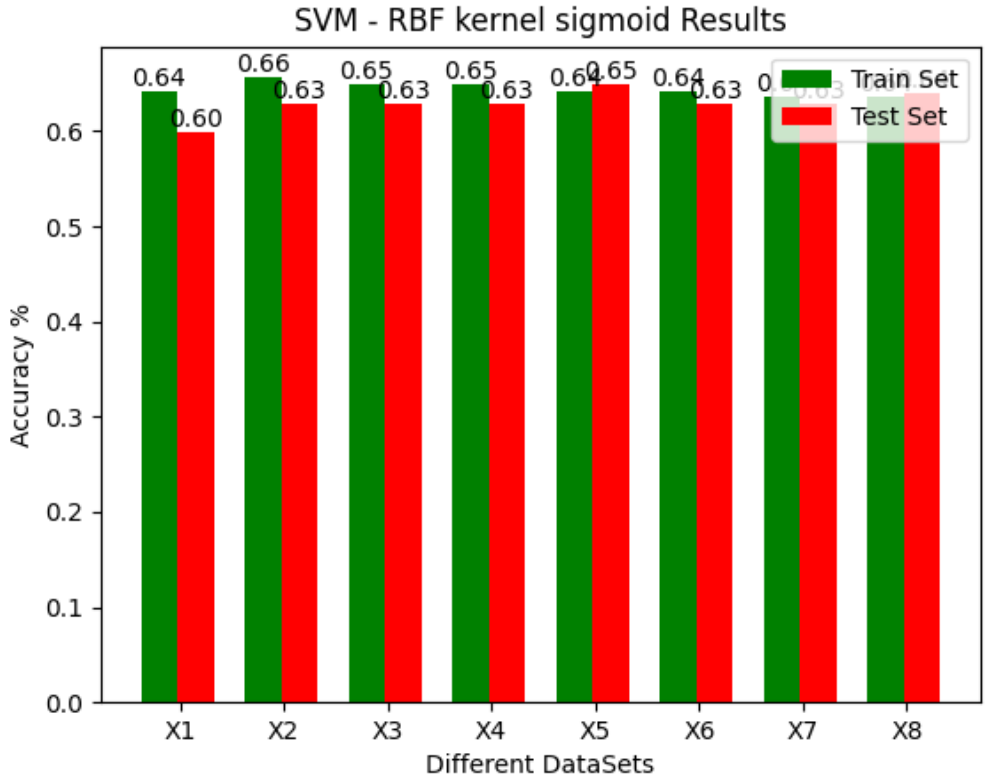


Detailed classification report:

	precision	recall	f1-score	support
0	0.60	0.69	0.64	45
1	0.56	0.47	0.51	38
2	0.62	0.57	0.59	14
accuracy			0.59	97
macro avg	0.59	0.58	0.58	97
weighted avg	0.59	0.59	0.58	97

Σχήμα 4.15: *Grid Search for POLY kernel - CV 10 - One Hot Encoder*

Τέλος παρουσιάζουμε τα αποτελέσματα για τον πυρήνα SIGMOID.

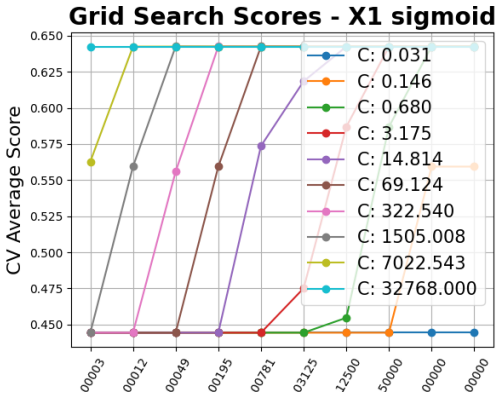


Σχήμα 4.16: *SVM SIGMOID - CV 10 - One Hot Encoder*

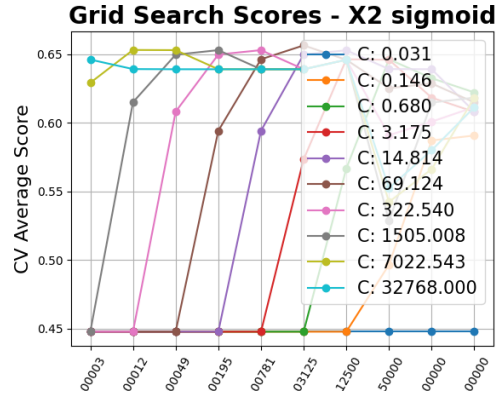
Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει επίσης την μέθοδο του Grid Search.

Πίνακας 4.10: *Best Parameters from Grid Search - SIGMOID kernel - CV 10 - One Hot Encoding*

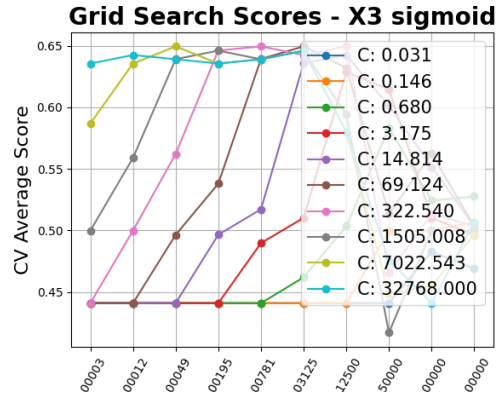
Set	**Parameter C**	**Parameter Gamma**
X1	0.6803950000871886	2.0
X2	69.12382328910758	0.03125
X3	69.12382328910758	0.03125
X4	69.12382328910758	0.03125
X5	7022.542707532377	0.0001220703125
X6	322.53978877308765	0.0078125
X7	322.53978877308765	0.0078125
X8	1505.0081200902148	0.00048828125



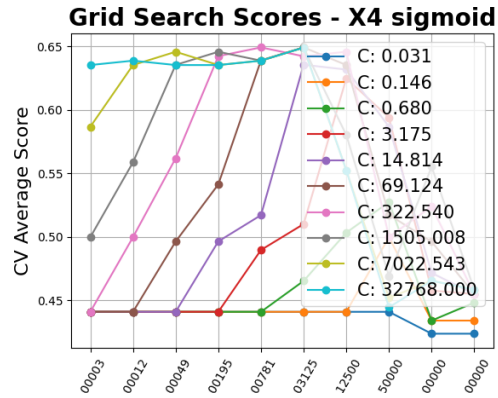

```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.56      0.82      0.67        45
           1       0.68      0.50      0.58        38
           2       0.67      0.14      0.24        14

    accuracy                           0.60        97
   macro avg       0.64      0.49      0.49        97
weighted avg       0.62      0.60      0.57        97
```



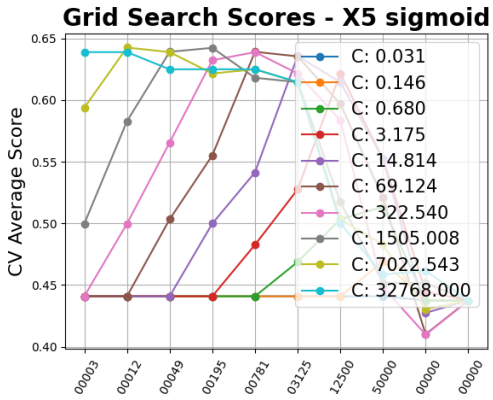

```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.65      0.71      0.68        45
           1       0.61      0.58      0.59        38
           2       0.58      0.50      0.54        14

    accuracy                           0.63        97
   macro avg       0.62      0.60      0.60        97
weighted avg       0.63      0.63      0.63        97
```




```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.65      0.71      0.68        45
           1       0.61      0.58      0.59        38
           2       0.58      0.50      0.54        14

    accuracy                           0.63        97
   macro avg       0.62      0.60      0.60        97
weighted avg       0.63      0.63      0.63        97
```

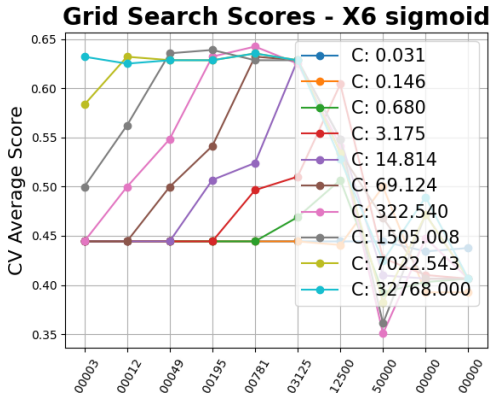



```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.65      0.71      0.68        45
           1       0.61      0.58      0.59        38
           2       0.58      0.50      0.54        14

    accuracy                           0.63        97
   macro avg       0.62      0.60      0.60        97
weighted avg       0.63      0.63      0.63        97
```



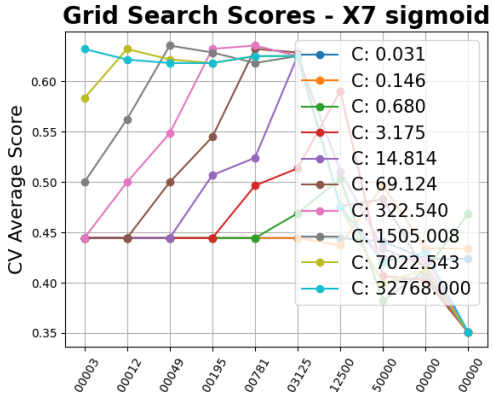
Detailed classification report:

	precision	recall	f1-score	support
0	0.67	0.73	0.70	45
1	0.63	0.58	0.60	38
2	0.62	0.57	0.59	14
accuracy			0.65	97
macro avg	0.64	0.63	0.63	97
weighted avg	0.65	0.65	0.65	97



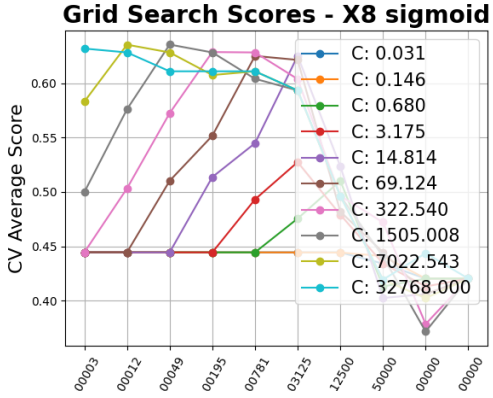
Detailed classification report:

	precision	recall	f1-score	support
0	0.65	0.71	0.68	45
1	0.61	0.58	0.59	38
2	0.58	0.50	0.54	14
accuracy			0.63	97
macro avg	0.62	0.60	0.60	97
weighted avg	0.63	0.63	0.63	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.65	0.71	0.68	45
1	0.61	0.58	0.59	38
2	0.58	0.50	0.54	14
accuracy			0.63	97
macro avg	0.62	0.60	0.60	97
weighted avg	0.63	0.63	0.63	97



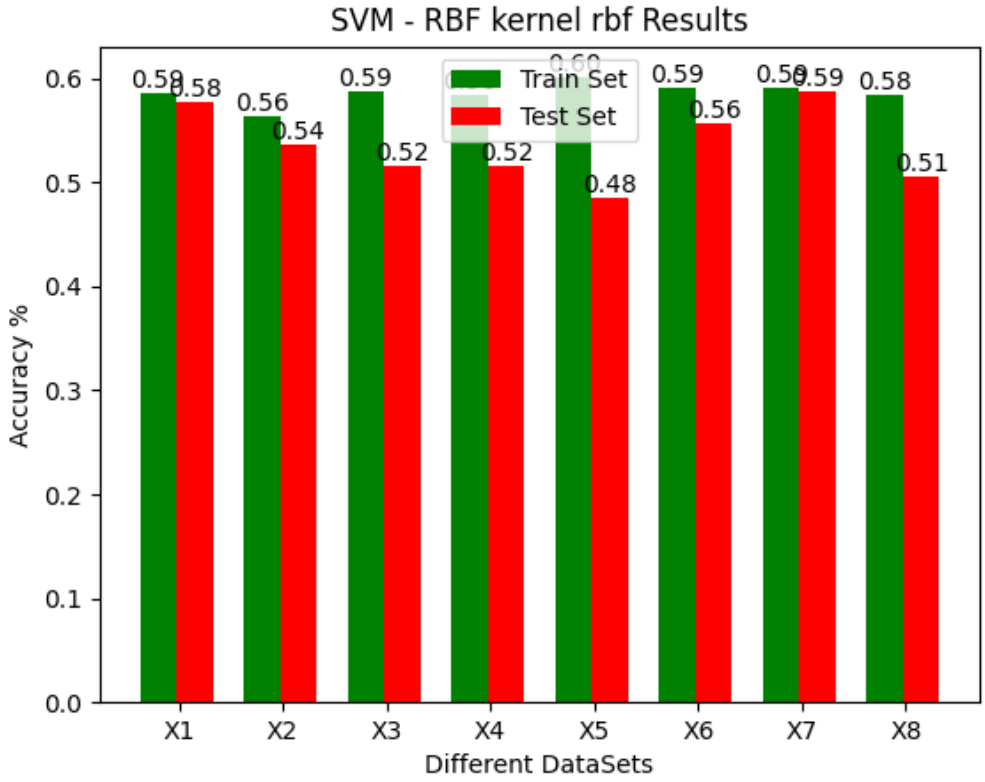
Detailed classification report:

	precision	recall	f1-score	support
0	0.66	0.73	0.69	45
1	0.63	0.58	0.60	38
2	0.58	0.50	0.54	14
accuracy			0.64	97
macro avg	0.62	0.60	0.61	97
weighted avg	0.64	0.64	0.64	97

Σχήμα 4.17: *Grid Search for SIGMOID kernel - CV 10 - One Hot Encoder*

CV = 5 ||*WordEmbeddings*

Παρακάτω θα δούμε τα αποτελέσματα για κωδικοποίηση δεδομένων με Word Embeddings. Θα ξεκινήσουμε πρώτα για τον πυρήνα RBF. Όπως και στους δυο πρώτους αλγορίθμους παρατηρούμε μείωση σε σχέση με το One Hot Encoding.

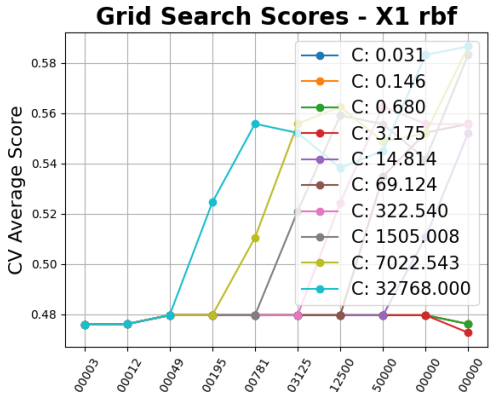


Σχήμα 4.18: *SVM RBF - CV 5 - Word Embeddings*

Για κάθε σύνολο δεδομένων αναφερόμαστε σε διαφορετικές παραμέτρους c και gamma ανάλογα τα καλύτερα αποτελέσματα που έχουν επιστρέψει. Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει την μέθοδο του Grid Search.

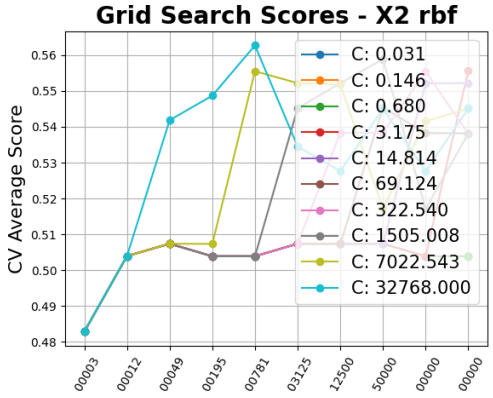
Πίνακας 4.11: *Best Parameters from Grid Search - RBF kernel - CV 5 - Word Embeddings*

Set	**Parameter C**	**Parameter Gamma**
X1	7022.542707532377	8.0
X2	32768.0	0.0078125
X3	322.53978877308765	0.5
X4	322.53978877308765	0.5
X5	1505.0081200902148	0.125
X6	1505.0081200902148	0.125
X7	322.53978877308765	0.5
X8	32768.0	0.03125



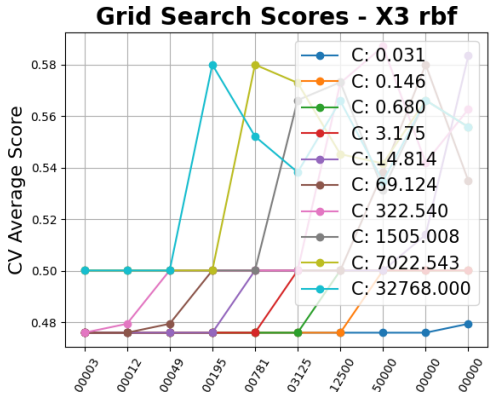
Detailed classification report:

	precision	recall	f1-score	support
0	0.68	0.56	0.61	45
1	0.56	0.63	0.59	38
2	0.41	0.50	0.45	14
accuracy			0.58	97
macro avg	0.55	0.56	0.55	97
weighted avg	0.59	0.58	0.58	97



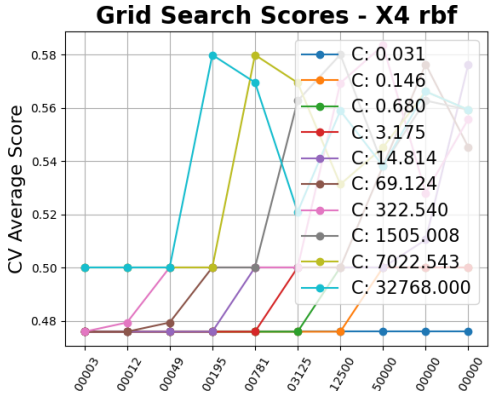
Detailed classification report:

	precision	recall	f1-score	support
0	0.54	0.58	0.56	45
1	0.48	0.53	0.50	38
2	0.86	0.43	0.57	14
accuracy			0.54	97
macro avg	0.62	0.51	0.54	97
weighted avg	0.56	0.54	0.54	97



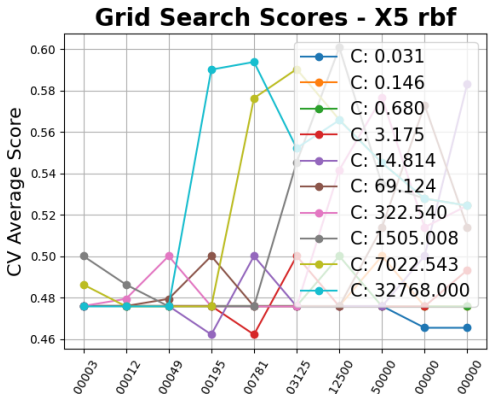
Detailed classification report:

	precision	recall	f1-score	support
0	0.53	0.60	0.56	45
1	0.47	0.53	0.49	38
2	1.00	0.21	0.35	14
accuracy			0.52	97
macro avg	0.66	0.45	0.47	97
weighted avg	0.57	0.52	0.51	97



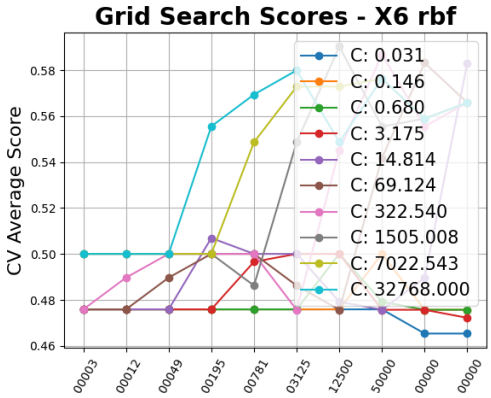
Detailed classification report:

	precision	recall	f1-score	support
0	0.53	0.60	0.56	45
1	0.47	0.53	0.49	38
2	1.00	0.21	0.35	14
accuracy			0.52	97
macro avg	0.66	0.45	0.47	97
weighted avg	0.57	0.52	0.51	97



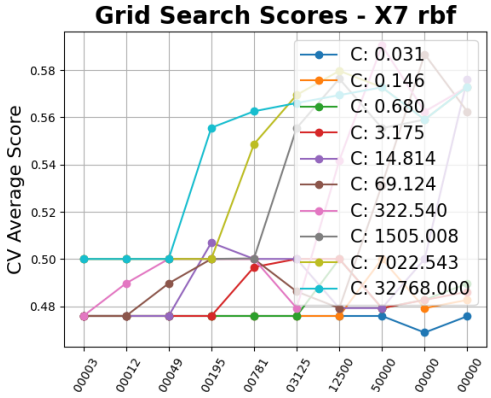
Detailed classification report:

	precision	recall	f1-score	support
0	0.49	0.58	0.53	45
1	0.44	0.45	0.44	38
2	0.80	0.29	0.42	14
accuracy			0.48	97
macro avg	0.58	0.44	0.46	97
weighted avg	0.51	0.48	0.48	97



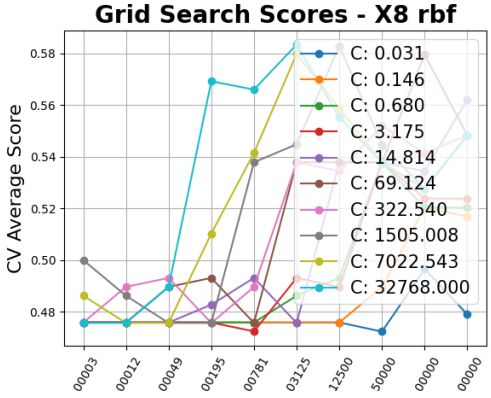
Detailed classification report:

	precision	recall	f1-score	support
0	0.55	0.58	0.57	45
1	0.53	0.63	0.58	38
2	0.80	0.29	0.42	14
accuracy			0.56	97
macro avg	0.63	0.50	0.52	97
weighted avg	0.58	0.56	0.55	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.58	0.58	0.58	45
1	0.57	0.68	0.62	38
2	0.83	0.36	0.50	14
accuracy			0.59	97
macro avg	0.66	0.54	0.57	97
weighted avg	0.61	0.59	0.58	97

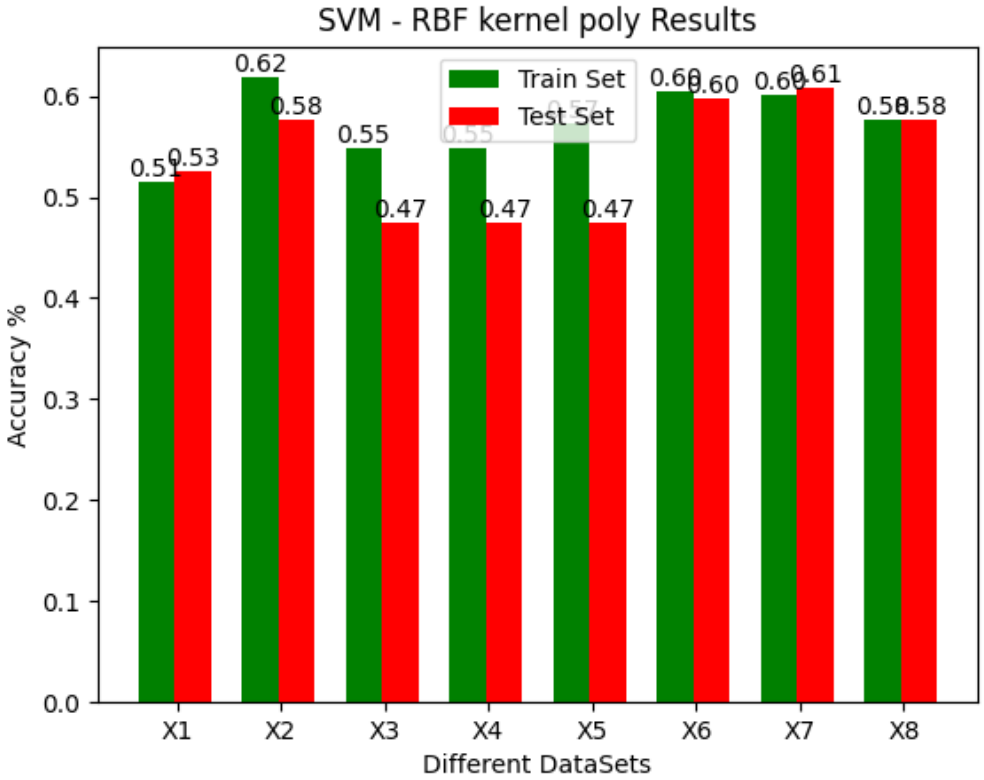


Detailed classification report:

	precision	recall	f1-score	support
0	0.57	0.58	0.57	45
1	0.50	0.47	0.49	38
2	0.33	0.36	0.34	14
accuracy			0.51	97
macro avg	0.47	0.47	0.47	97
weighted avg	0.51	0.51	0.51	97

Σχήμα 4.19: *Grid Search for RBF kernel - CV 5 - Word Embeddings*

Συνεχίζουμε με τα αποτελέσματα για τον πυρήνα POLY. Στο παρακάτω διάγραμμα βλέπουμε αρκετά χαμηλά ποσοστά ακρίβειας για κάθε σύνολο δεδομένων.

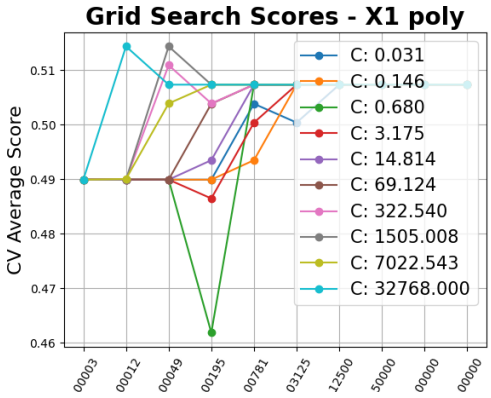


Σχήμα 4.20: *SVM POLY - CV 5 - Word Embeddings*

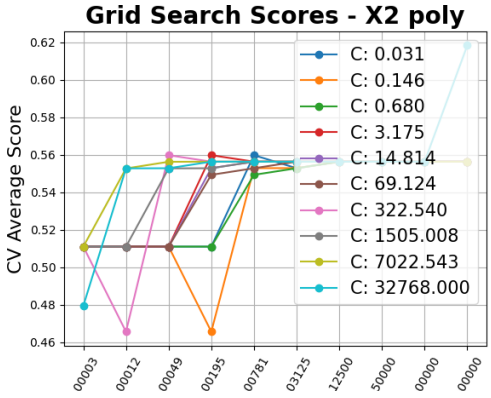
Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει επίσης την μέθοδο του Grid Search.

Πίνακας 4.12: *Best Parameters from Grid Search - POLY kernel - CV 5 - Word Embeddings*

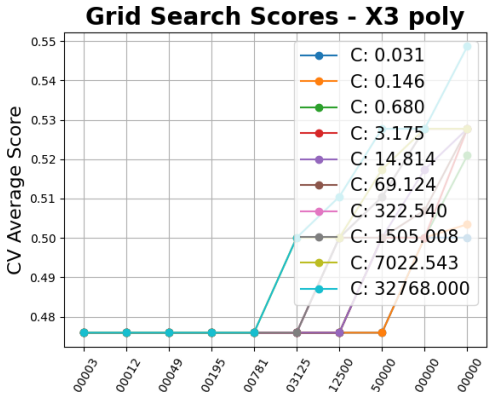
Set	**Parameter C**	**Parameter Gamma**
X1	1505.0081200902148	0.00048828125
X2	32768.0	8.0
X3	32768.0	8.0
X4	32768.0	8.0
X5	32768.0	0.125
X6	14.81399539659665	2.0
X7	14.81399539659665	2.0
X8	3.1748021039363996	2.0



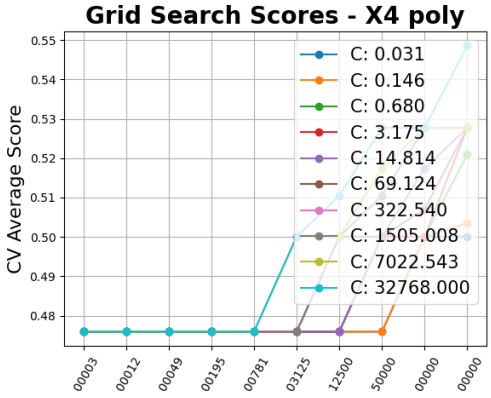

```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.49      0.96      0.65        45
           1       0.80      0.21      0.33        38

   micro avg       0.53      0.61      0.57        83
   macro avg       0.65      0.58      0.49        83
weighted avg       0.63      0.61      0.51        83
```



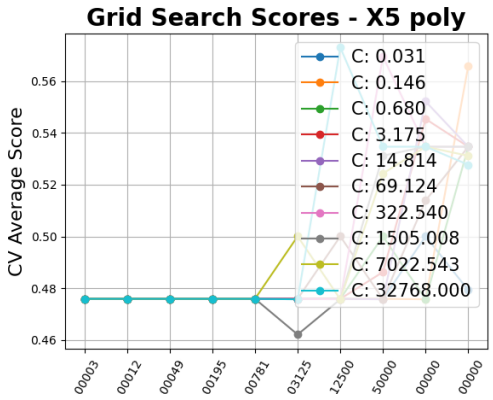

```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.53      0.87      0.66        45
           1       0.70      0.42      0.52        38
           2       1.00      0.07      0.13        14

    accuracy                           0.58        97
   macro avg       0.74      0.45      0.44        97
weighted avg       0.66      0.58      0.53        97
```




```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.47      0.84      0.61        45
           1       0.47      0.18      0.26        38
           2       0.50      0.07      0.12        14

    accuracy                           0.47        97
   macro avg       0.48      0.37      0.33        97
weighted avg       0.48      0.47      0.40        97
```

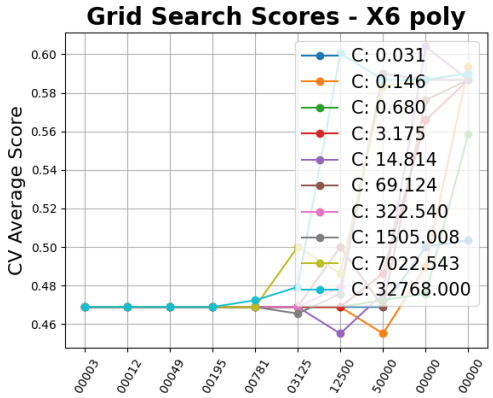



```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.47      0.84      0.61        45
           1       0.47      0.18      0.26        38
           2       0.50      0.07      0.12        14

    accuracy                           0.47        97
   macro avg       0.48      0.37      0.33        97
weighted avg       0.48      0.47      0.40        97
```



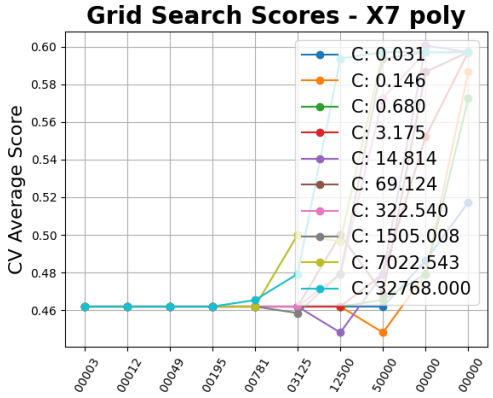
Detailed classification report:

	precision	recall	f1-score	support
0	0.48	0.60	0.53	45
1	0.45	0.47	0.46	38
2	1.00	0.07	0.13	14
accuracy			0.47	97
macro avg	0.64	0.38	0.38	97
weighted avg	0.54	0.47	0.45	97



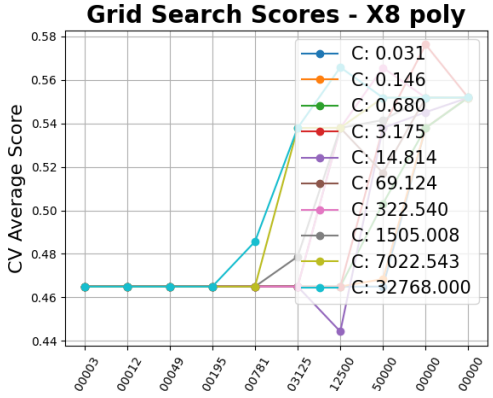
Detailed classification report:

	precision	recall	f1-score	support
0	0.60	0.56	0.57	45
1	0.57	0.68	0.62	38
2	0.78	0.50	0.61	14
accuracy			0.60	97
macro avg	0.65	0.58	0.60	97
weighted avg	0.61	0.60	0.60	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.62	0.53	0.57	45
1	0.58	0.74	0.65	38
2	0.70	0.50	0.58	14
accuracy			0.61	97
macro avg	0.63	0.59	0.60	97
weighted avg	0.62	0.61	0.60	97

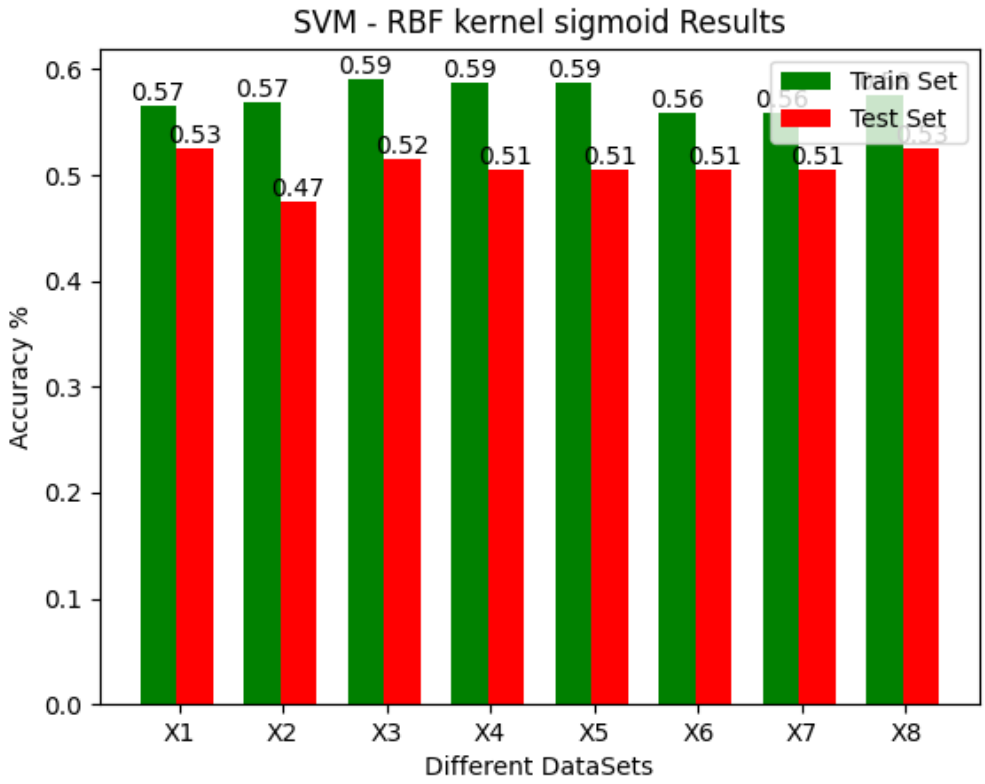


Detailed classification report:

	precision	recall	f1-score	support
0	0.55	0.64	0.59	45
1	0.59	0.63	0.61	38
2	1.00	0.21	0.35	14
accuracy			0.58	97
macro avg	0.71	0.50	0.52	97
weighted avg	0.63	0.58	0.56	97

Σχήμα 4.21: *Grid Search for POLY kernel - CV 5 - Word Embeddings*

Τέλος παρουσιάζουμε τα αποτελέσματα για τον πυρήνα SIGMOID. Παρατηρούμε καλύτερα αποτελέσματα απο τον πυρήνα POLY.

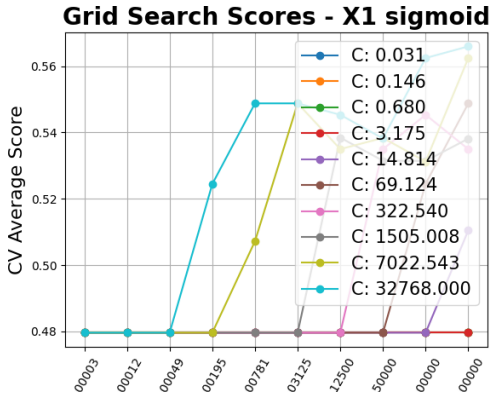


Σχήμα 4.22: *SVM SIGMOID - CV 5 - Word Embeddings*

Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει επίσης την μέθοδο του Grid Search.

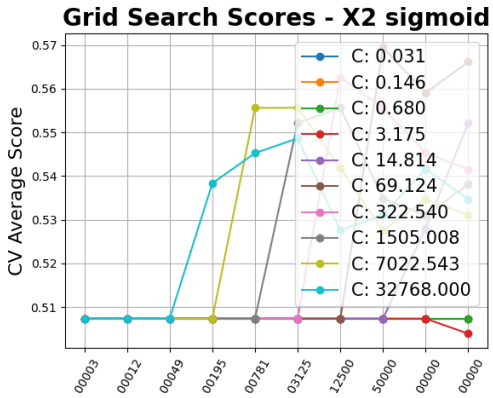
Πίνακας 4.13: *Best Parameters from Grid Search - SIGMOID kernel - CV 5 - Word Embeddings*

Set	**Parameter C**	**Parameter Gamma**
X1	32768.0	8.0
X2	69.12382328910758	0.5
X3	7022.542707532377	0.03125
X4	7022.542707532377	0.03125
X5	32768.0	0.0078125
X6	32768.0	0.0078125
X7	32768.0	0.0078125
X8	32768.0	0.0078125



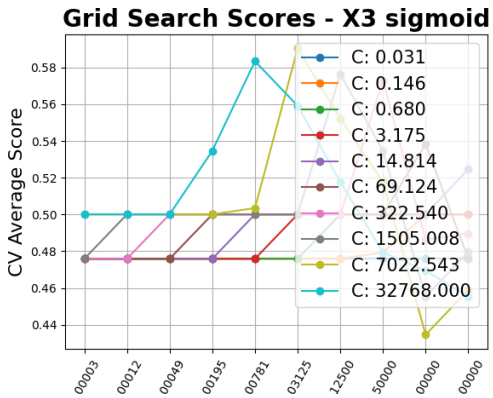
Detailed classification report:

	precision	recall	f1-score	support
0	0.64	0.56	0.60	45
1	0.50	0.53	0.51	38
2	0.33	0.43	0.38	14
accuracy			0.53	97
macro avg	0.49	0.50	0.49	97
weighted avg	0.54	0.53	0.53	97



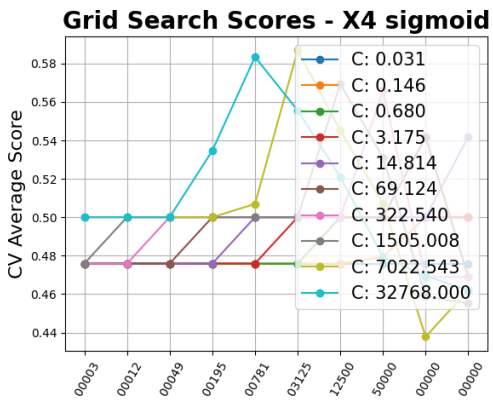
Detailed classification report:

	precision	recall	f1-score	support
0	0.48	0.60	0.53	45
1	0.46	0.50	0.48	38
micro avg	0.47	0.55	0.51	83
macro avg	0.47	0.55	0.51	83
weighted avg	0.47	0.55	0.51	83



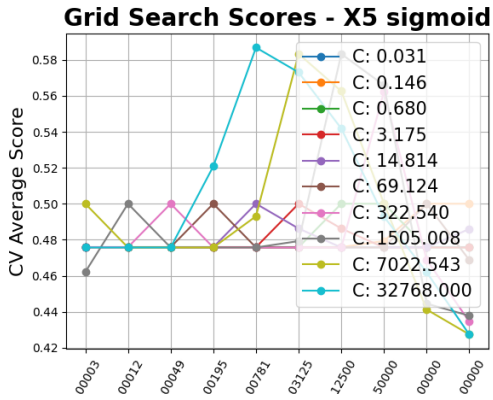
Detailed classification report:

	precision	recall	f1-score	support
0	0.53	0.62	0.57	45
1	0.49	0.55	0.52	38
2	1.00	0.07	0.13	14
accuracy			0.52	97
macro avg	0.67	0.42	0.41	97
weighted avg	0.58	0.52	0.49	97



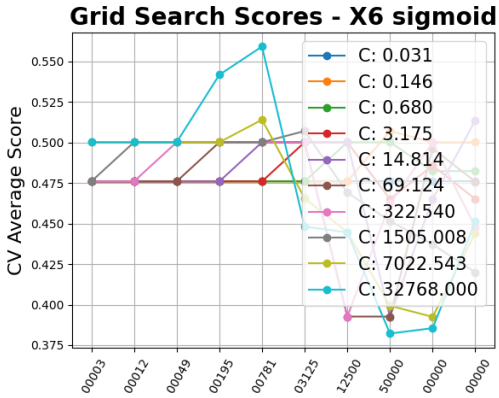
Detailed classification report:

	precision	recall	f1-score	support
0	0.52	0.62	0.57	45
1	0.48	0.53	0.50	38
2	1.00	0.07	0.13	14
accuracy			0.51	97
macro avg	0.66	0.41	0.40	97
weighted avg	0.57	0.51	0.48	97



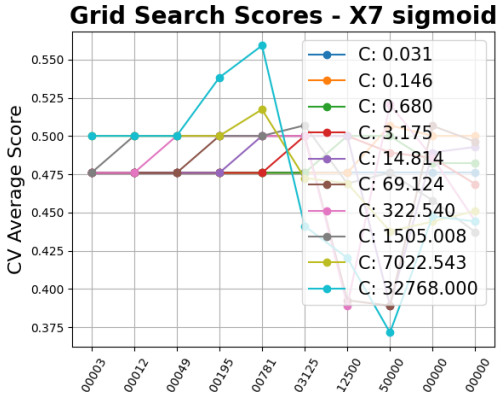
Detailed classification report:

	precision	recall	f1-score	support
0	0.51	0.60	0.55	45
1	0.49	0.53	0.51	38
2	0.67	0.14	0.24	14
accuracy			0.51	97
macro avg	0.55	0.42	0.43	97
weighted avg	0.52	0.51	0.49	97



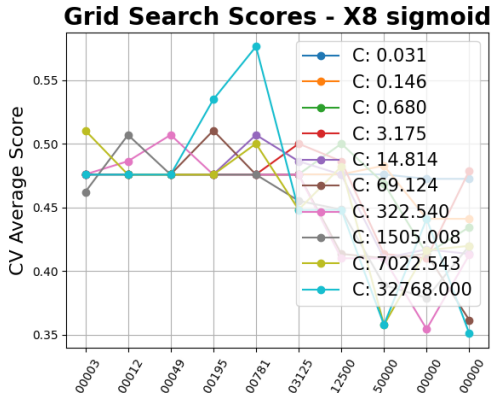
Detailed classification report:

	precision	recall	f1-score	support
0	0.52	0.58	0.55	45
1	0.47	0.55	0.51	38
2	1.00	0.14	0.25	14
accuracy			0.51	97
macro avg	0.66	0.42	0.43	97
weighted avg	0.57	0.51	0.49	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.52	0.58	0.55	45
1	0.47	0.55	0.51	38
2	1.00	0.14	0.25	14
accuracy			0.51	97
macro avg	0.66	0.42	0.43	97
weighted avg	0.57	0.51	0.49	97



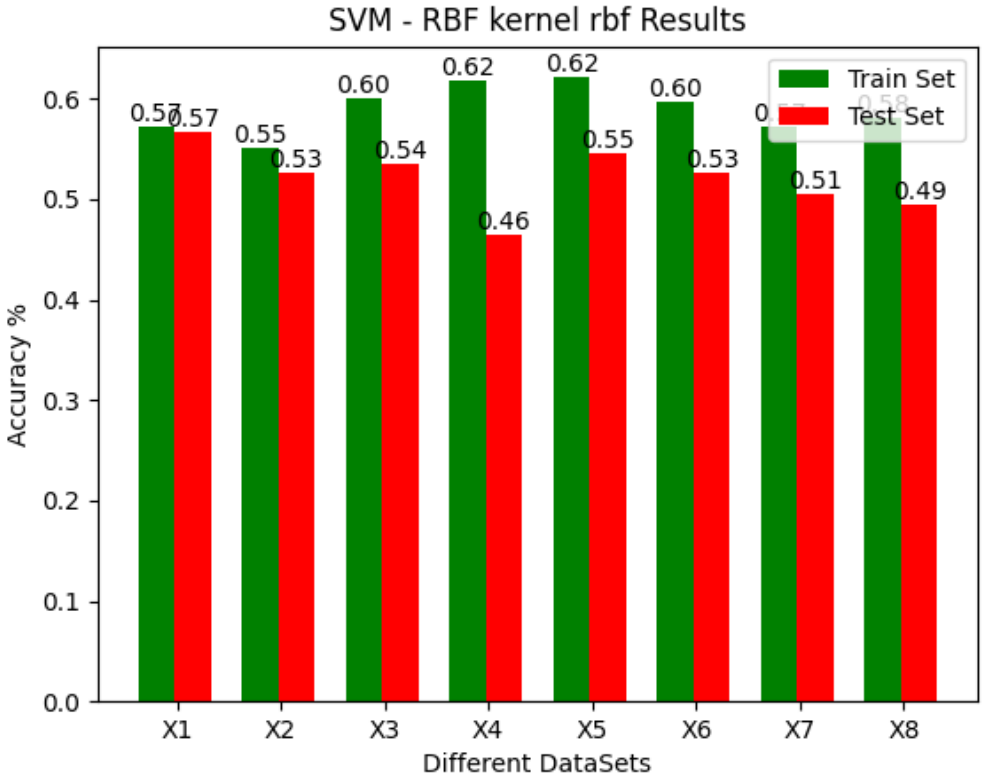
Detailed classification report:

	precision	recall	f1-score	support
0	0.52	0.60	0.56	45
1	0.49	0.53	0.51	38
2	1.00	0.29	0.44	14
accuracy			0.53	97
macro avg	0.67	0.47	0.50	97
weighted avg	0.58	0.53	0.52	97

Σχήμα 4.23: *Grid Search for SIGMOID kernel - CV 5 - Word Embeddings*

CV = 10 ||*WordEmbeddings*

Συνεχίζουμε για cv=10

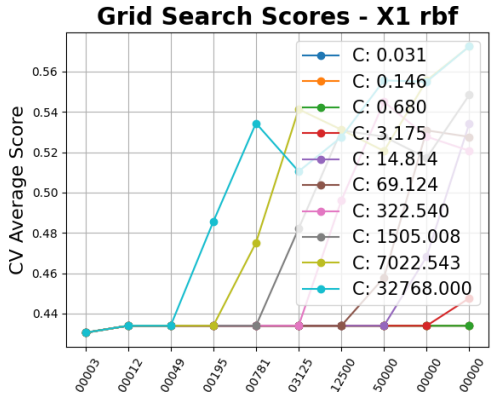


Σχήμα 4.24: *SVM RBF - CV 10 - Word Embeddings*

Για κάθε σύνολο δεδομένων αναφερόμαστε σε διαφορετικές παραμέτρους c και gamma ανάλογα τα καλύτερα αποτελέσματα που έχουν επιστρέψει. Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει την μέθοδο του Grid Search.

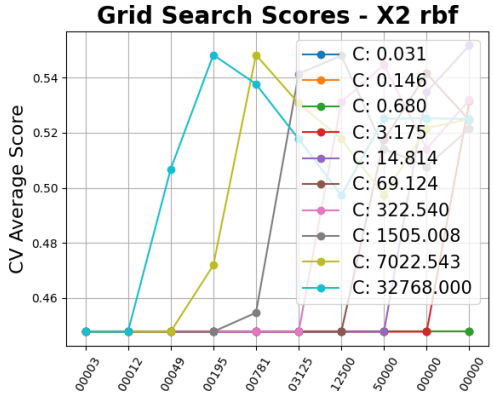
Πίνακας 4.14: *Best Parameters from Grid Search - RBF kernel - CV 10 - Word Embeddings*

Set	**Parameter C**	**Parameter Gamma**
X1	7022.542707532377	8.0
X2	14.81399539659665	8.0
X3	14.81399539659665	8.0
X4	7022.542707532377	0.5
X5	322.53978877308765	2.0
X6	7022.542707532377	0.5
X7	1505.0081200902148	0.125
X8	32768.0	0.125



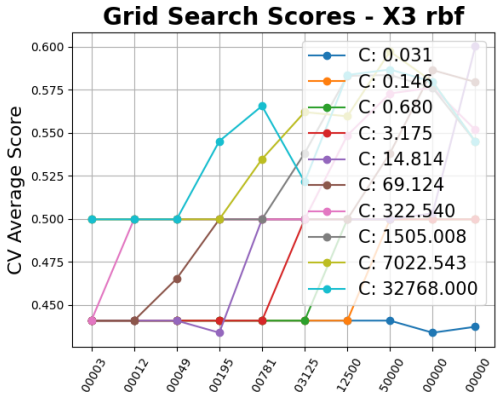
Detailed classification report:

	precision	recall	f1-score	support
0	0.65	0.58	0.61	45
1	0.52	0.61	0.56	38
2	0.46	0.43	0.44	14
accuracy			0.57	97
macro avg	0.54	0.54	0.54	97
weighted avg	0.57	0.57	0.57	97



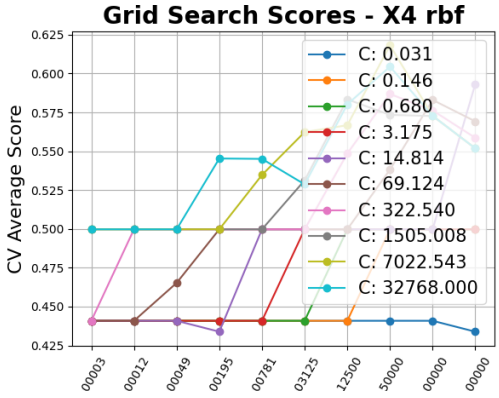
Detailed classification report:

	precision	recall	f1-score	support
0	0.55	0.51	0.53	45
1	0.48	0.66	0.56	38
2	1.00	0.21	0.35	14
accuracy			0.53	97
macro avg	0.68	0.46	0.48	97
weighted avg	0.59	0.53	0.51	97



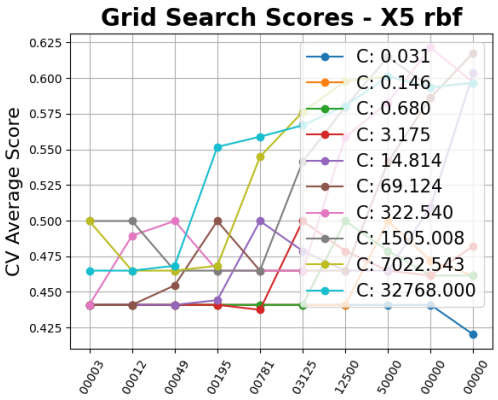
Detailed classification report:

	precision	recall	f1-score	support
0	0.55	0.67	0.60	45
1	0.49	0.50	0.49	38
2	1.00	0.21	0.35	14
accuracy			0.54	97
macro avg	0.68	0.46	0.48	97
weighted avg	0.59	0.54	0.52	97



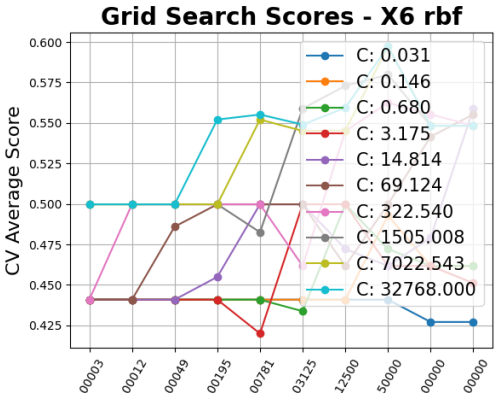
Detailed classification report:

	precision	recall	f1-score	support
0	0.53	0.51	0.52	45
1	0.42	0.45	0.44	38
2	0.36	0.36	0.36	14
accuracy			0.46	97
macro avg	0.44	0.44	0.44	97
weighted avg	0.47	0.46	0.46	97



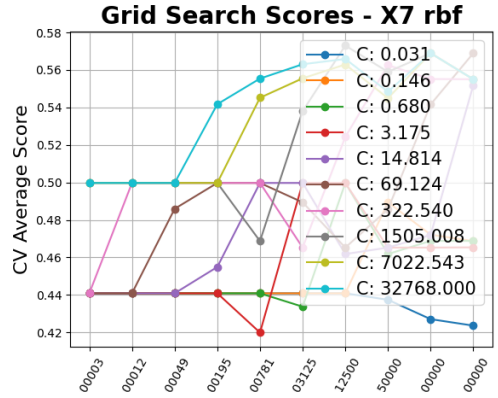
Detailed classification report:

	precision	recall	f1-score	support
0	0.56	0.53	0.55	45
1	0.50	0.58	0.54	38
2	0.70	0.50	0.58	14
accuracy			0.55	97
macro avg	0.59	0.54	0.56	97
weighted avg	0.56	0.55	0.55	97



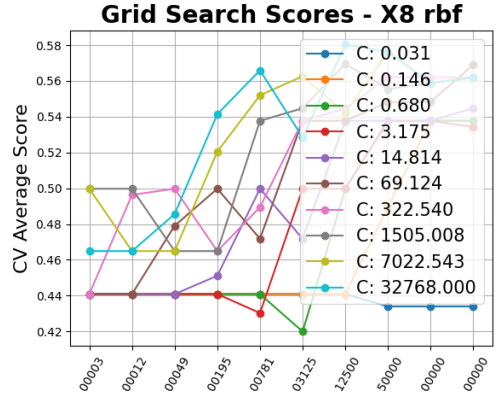
Detailed classification report:

	precision	recall	f1-score	support
0	0.57	0.56	0.56	45
1	0.50	0.53	0.51	38
2	0.46	0.43	0.44	14
accuracy			0.53	97
macro avg	0.51	0.50	0.51	97
weighted avg	0.53	0.53	0.53	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.52	0.53	0.53	45
1	0.45	0.55	0.49	38
2	1.00	0.29	0.44	14
accuracy			0.51	97
macro avg	0.66	0.46	0.49	97
weighted avg	0.56	0.51	0.50	97

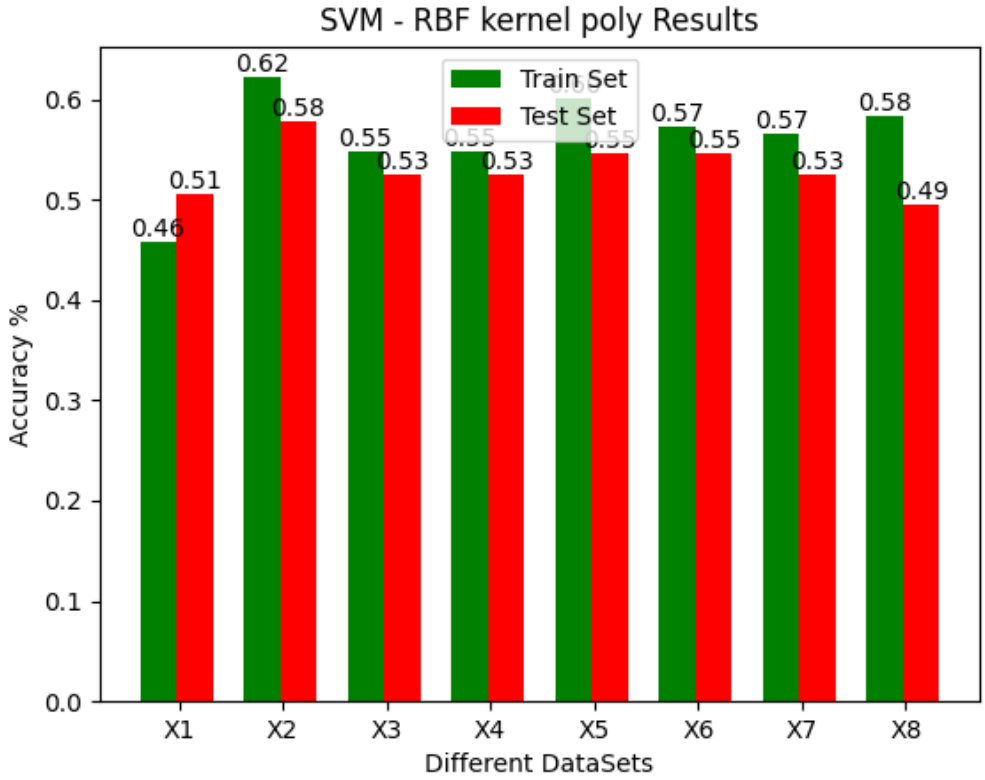


Detailed classification report:

	precision	recall	f1-score	support
0	0.53	0.53	0.53	45
1	0.44	0.50	0.47	38
2	0.56	0.36	0.43	14
accuracy			0.49	97
macro avg	0.51	0.46	0.48	97
weighted avg	0.50	0.49	0.49	97

Σχήμα 4.25: *Grid Search for RBF kernel - CV 10 - Word Embeddings*

Συνεχίζουμε με τα αποτελέσματα για τον πυρήνα POLY. Στο παρακάτω διάγραμμα παρατηρούμε εν αντιθέση με πριν μια μικρή αύξηση σε σχέση με τα ποσοστά του RBF πυρήνα.

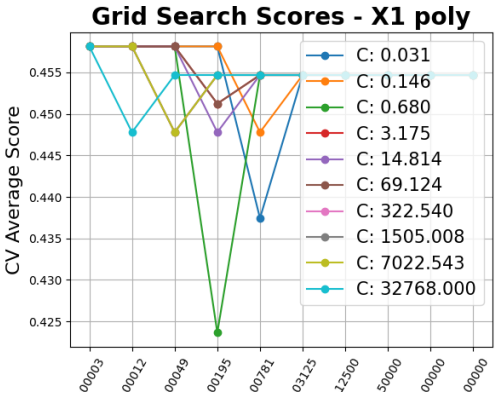


Σχήμα 4.26: *SVM POLY - CV 10 - Word Embeddings*

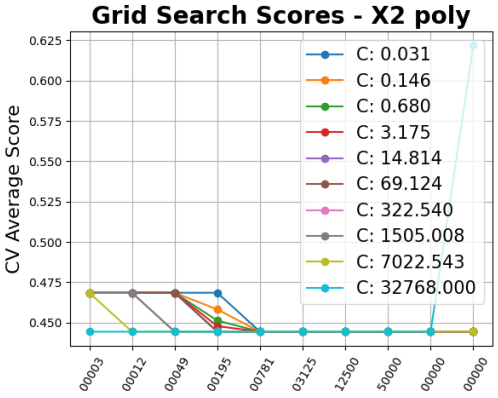
Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει επίσης την μέθοδο του Grid Search.

Πίνακας 4.15: *Best Parameters from Grid Search - POLY kernel - CV 10 - Word Embeddings*

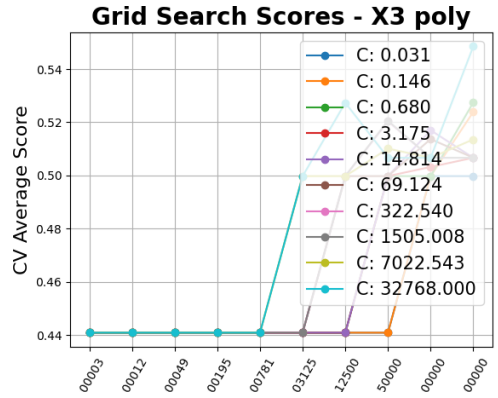
Set	**Parameter C**	**Parameter Gamma**
X1	0.03125	3.0517578125ε-05
X2	32768.0	8.0
X3	32768.0	8.0
X4	32768.0	8.0
X5	0.6803950000871886	8.0
X6	32768.0	8.0
X7	3.1748021039363996	8.0
X8	1505.0081200902148	0.5



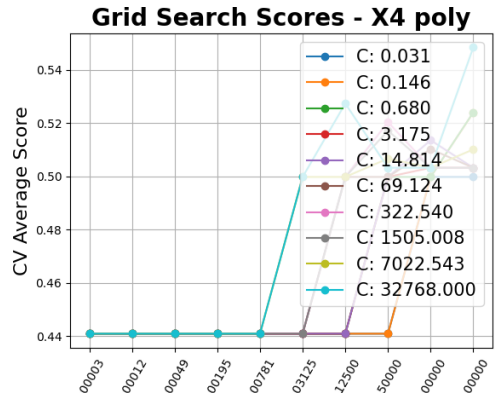

```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.52      0.51      0.52        45
           1       0.49      0.68      0.57        38

   micro avg       0.51      0.59      0.54        83
   macro avg       0.51      0.60      0.54        83
weighted avg       0.51      0.59      0.54        83
```



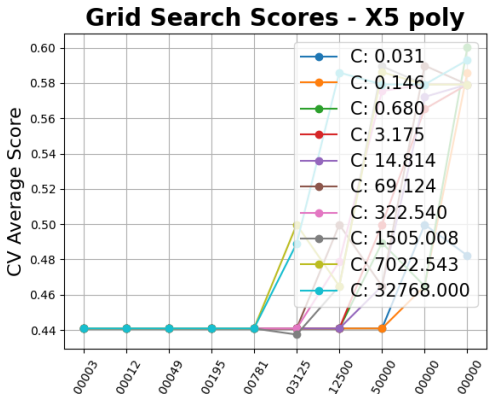

```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.54      0.82      0.65        45
           1       0.65      0.45      0.53        38
           2       1.00      0.14      0.25        14

    accuracy                           0.58        97
   macro avg       0.73      0.47      0.48        97
weighted avg       0.65      0.58      0.55        97
```




```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.50      0.87      0.63        45
           1       0.57      0.21      0.31        38
           2       0.80      0.29      0.42        14

    accuracy                           0.53        97
   macro avg       0.62      0.45      0.45        97
weighted avg       0.57      0.53      0.48        97
```

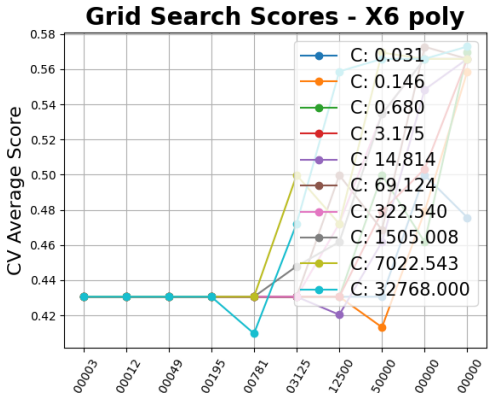



```
Detailed classification report:

              precision    recall  f1-score   support

           0       0.50      0.87      0.63        45
           1       0.57      0.21      0.31        38
           2       0.80      0.29      0.42        14

    accuracy                           0.53        97
   macro avg       0.62      0.45      0.45        97
weighted avg       0.57      0.53      0.48        97
```



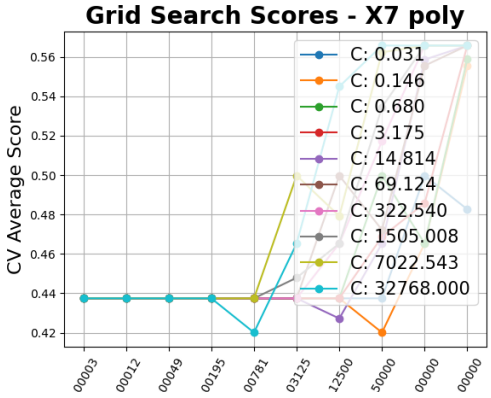
Detailed classification report:

	precision	recall	f1-score	support
0	0.54	0.58	0.56	45
1	0.54	0.58	0.56	38
2	0.62	0.36	0.45	14
accuracy			0.55	97
macro avg	0.57	0.50	0.52	97
weighted avg	0.55	0.55	0.54	97



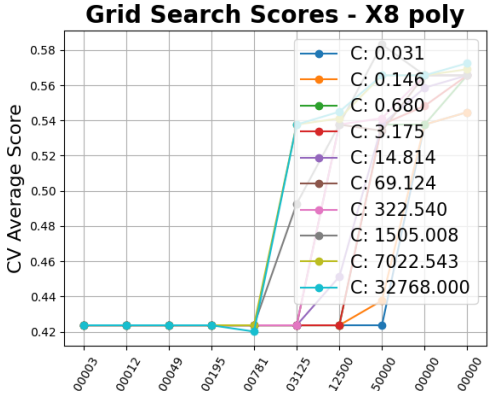
Detailed classification report:

	precision	recall	f1-score	support
0	0.58	0.58	0.58	45
1	0.49	0.55	0.52	38
2	0.67	0.43	0.52	14
accuracy			0.55	97
macro avg	0.58	0.52	0.54	97
weighted avg	0.56	0.55	0.55	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.55	0.60	0.57	45
1	0.46	0.45	0.45	38
2	0.64	0.50	0.56	14
accuracy			0.53	97
macro avg	0.55	0.52	0.53	97
weighted avg	0.53	0.53	0.52	97

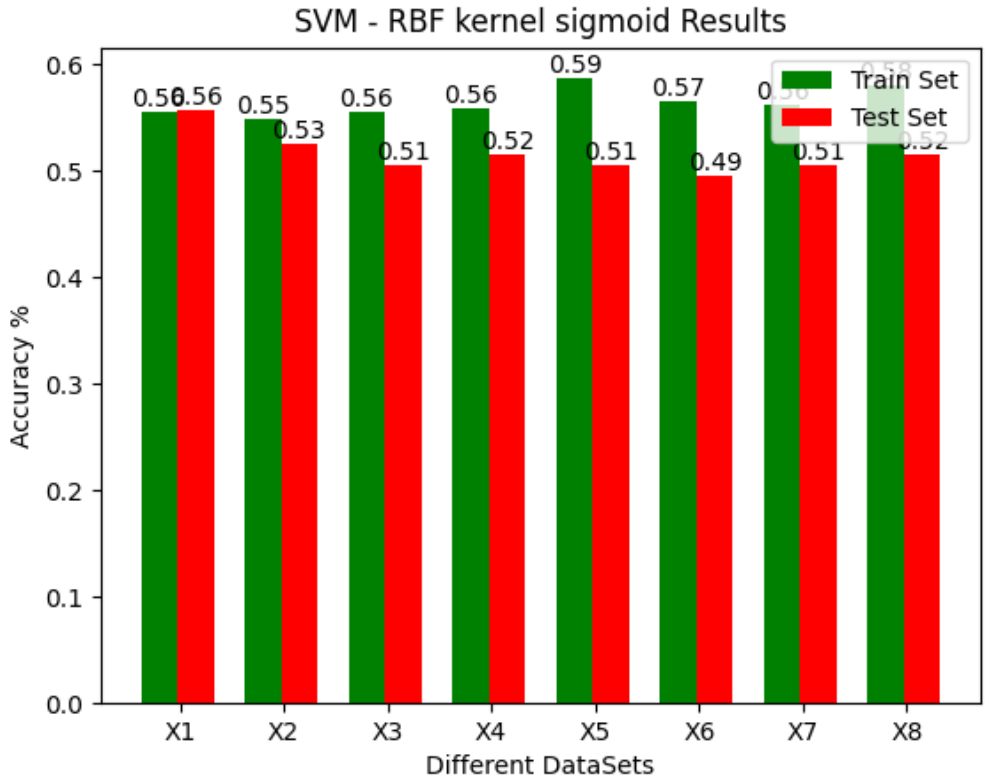


Detailed classification report:

	precision	recall	f1-score	support
0	0.51	0.53	0.52	45
1	0.45	0.47	0.46	38
2	0.60	0.43	0.50	14
accuracy			0.49	97
macro avg	0.52	0.48	0.49	97
weighted avg	0.50	0.49	0.50	97

Σχήμα 4.27: *Grid Search for POLY kernel - CV 10 - Word Embeddings*

Τέλος παρουσιάζουμε τα αποτελέσματα για τον πυρήνα SIGMOID.

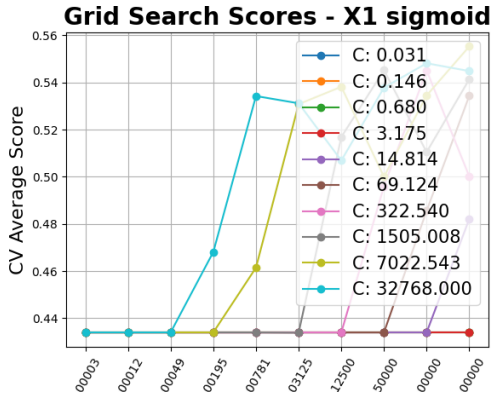


Σχήμα 4.28: *SVM SIGMOID - CV 10 - Word Embeddings*

Για την εύρεση των καλύτερων παραμέτρων έχουμε χρησιμοποιήσει επίσης την μέθοδο του Grid Search.

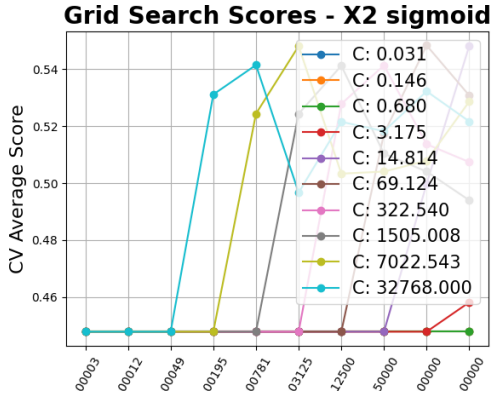
Πίνακας 4.16: *Best Parameters from Grid Search - SIGMOID kernel - CV 10 - Word Embeddings*

Set	**Parameter C**	**Parameter Gamma**
X1	7022.542707532377	8.0
X2	69.12382328910758	2.0
X3	32768.0	0.0078125
X4	7022.542707532377	0.03125
X5	32768.0	0.0078125
X6	32768.0	0.0078125
X7	32768.0	0.0078125
X8	32768.0	0.0078125



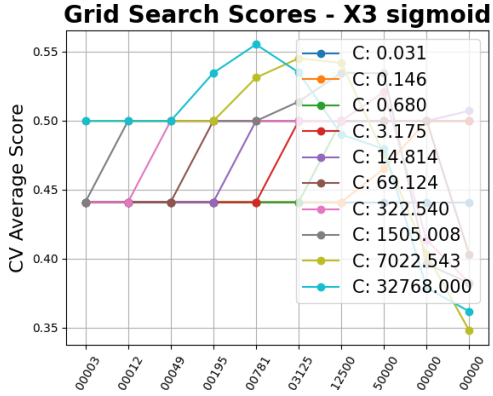
Detailed classification report:

	precision	recall	f1-score	support
0	0.59	0.60	0.59	45
1	0.53	0.61	0.57	38
2	0.50	0.29	0.36	14
accuracy			0.56	97
macro avg	0.54	0.50	0.51	97
weighted avg	0.55	0.56	0.55	97



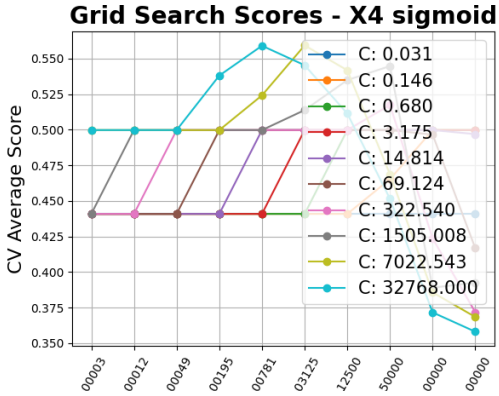
Detailed classification report:

	precision	recall	f1-score	support
0	0.54	0.60	0.57	45
1	0.51	0.63	0.56	38
micro avg	0.53	0.61	0.57	83
macro avg	0.53	0.62	0.57	83
weighted avg	0.53	0.61	0.57	83



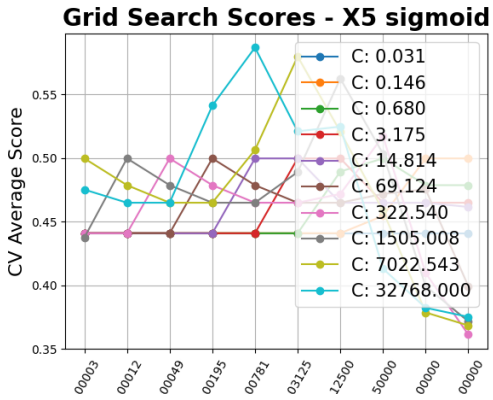
Detailed classification report:

	precision	recall	f1-score	support
0	0.52	0.53	0.53	45
1	0.46	0.58	0.51	38
2	1.00	0.21	0.35	14
accuracy			0.51	97
macro avg	0.66	0.44	0.46	97
weighted avg	0.57	0.51	0.50	97



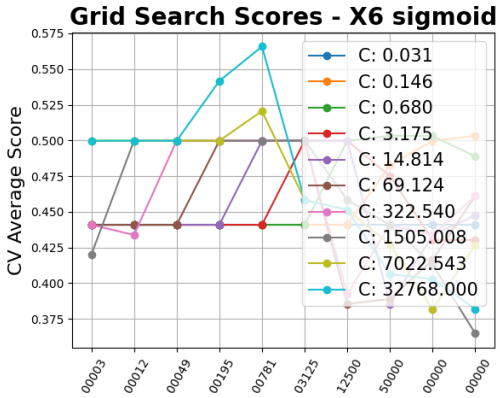
Detailed classification report:

	precision	recall	f1-score	support
0	0.52	0.56	0.54	45
1	0.48	0.58	0.52	38
2	1.00	0.21	0.35	14
accuracy			0.52	97
macro avg	0.67	0.45	0.47	97
weighted avg	0.57	0.52	0.51	97



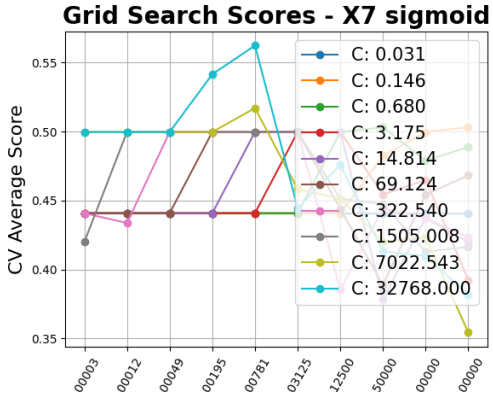
Detailed classification report:

	precision	recall	f1-score	support
0	0.51	0.53	0.52	45
1	0.47	0.58	0.52	38
2	1.00	0.21	0.35	14
accuracy			0.51	97
macro avg	0.66	0.44	0.46	97
weighted avg	0.56	0.51	0.50	97



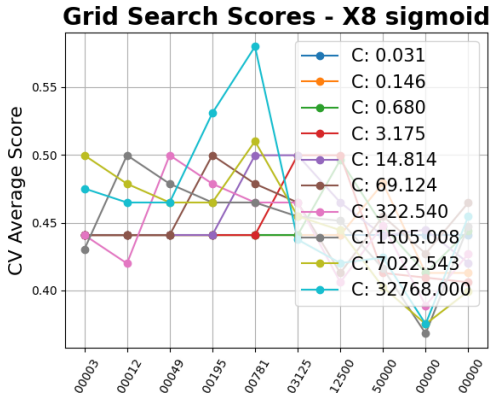
Detailed classification report:

	precision	recall	f1-score	support
0	0.51	0.49	0.50	45
1	0.44	0.58	0.50	38
2	1.00	0.29	0.44	14
accuracy			0.49	97
macro avg	0.65	0.45	0.48	97
weighted avg	0.55	0.49	0.49	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.52	0.51	0.52	45
1	0.45	0.58	0.51	38
2	1.00	0.29	0.44	14
accuracy			0.51	97
macro avg	0.66	0.46	0.49	97
weighted avg	0.56	0.51	0.50	97



Detailed classification report:

	precision	recall	f1-score	support
0	0.52	0.49	0.51	45
1	0.47	0.63	0.54	38
2	1.00	0.29	0.44	14
accuracy			0.52	97
macro avg	0.66	0.47	0.50	97
weighted avg	0.57	0.52	0.51	97

Σχήμα 4.29: *Grid Search for SIGMOID kernel - CV 10 - Word Embeddings*

Μέρος III

Επίλογος

Κεφάλαιο 5

Επίλογος

Στο τελευταίο κεφάλαιο περιγράφεται εν συντομία η προτεινόμενη μέθοδος. Συζητείται η μεθοδολογία και τα κυριότερα χαρακτηριστικά της και υπογραμίζονται τα πλεονεκτήματα και οι αδυναμίες της. Στη συνέχεια, τα αποτελέσματα από τις περιπτώσεις που μελετήσαμε (case studies) συζητούνται μαζί με τον βαθμό στον οποίο η εφαρμογή της μεθόδου θεωρείται επιτυχής. Τέλος, αναφέρονται οι περιορισμοί που υπήρχαν αλλά και οι πιθανές μελλοντικές εργασίες επέκτασης της διπλωματικής εργασίας.

5.1 Σύνοψη

Στην παρούσα διπλωματική εργασία πραγματευθήκαμε το πρόβλημα της αναγνώρισης διαδικασιών που είναι κατάλληλες προς αυτοματοποίηση. Πρόκειται για ένα δύσκολο κι απαιτητικό πρόβλημα ακόμη και για τον άνθρωπο. Προς το παρόν αυτή η διαδικασία γίνεται από κάποιον αναλυτή χειρονακτικά, και πέρα από το γεγονός ότι είναι χρονοβόρα δεν υπάρχει τρόπος επιβεβαίωσης ότι επιλέχτηκαν οι σωστές διαδικασίες. Θεωρήθηκε ως δεδομένο πως οι αναλύσεις διαδικαδιών είναι σε έγγραφη μορφή. Σκοπός μας ήταν η μελέτη κι ανάλυση των έγγραφων κειμένων περιγραφής διαδικασιών, και του τρόπου με το οποίο μπορεί αυτόματα να γίνει η ταξινόμηση τους ως προς την δυνατότητα αυτοματοποίση τους.

Αναλυτικότερα χρησιμοποιήσαμε μια πληθώρα βιβλιοθηκών επεξεργασίας φυσικής γλώσσας (NLP) για την ανάλυση των κειμένων και την εξαγωγή δραστηριοτήτων που θα έπρεπε να κατηγοριοποιηθούν. Συνθέσαμε ένα λεπτομερές dataset με τα χαρακτηριστικά που καταέραμε να εξάγουμε από την ανάλυση των κειμένων και των προτάσεων, και κάναμε δοκιμές με διάφορους αλγορίθμους μηχανικής μάθησης όσον αφορά την ταξνόμηση τους.

Εργαστήκαμε με ένα σύνολο κειμένων με αρκετές διαφορές μεταξύ τους. Χωρίς απαραίτητα να είναι όλα από τον επιχειρισιακό τομέα όπως στην υπόθεση μας, αναλύσαμε όλα τα κείμενα και βγάλαμε αποτελέσματα για όλα. Εντοπίσαμε ένα ποσοστό επιτυχίας που μας επιτρέπει να θεωρούμε ότι η μέθοδος μας κινείται προς την σωστή κατεύθυνση.

Συνοπτικά τα βήματα που ακολουθήσαμε ήταν τα παρακάτω:

- Συλλέξαμε ένα σύνολο κειμένων που περιγράφουν διαδικασίες.
- Αναλύσαμε τα κείμενα και τις προτάσεις αυτών.
- Εντοπίσαμε τις δραστηριότητες μέσα από στα κείμενα.

- Ερευνήσαμε και καταλήξαμε σε τέσσερα χαρακτηριστικά των φράσεων δραστηριοτήτων που θα μας φαινόντουσαν χρήσιμα.
- Εξάγαμε τα χαρακτηριστικά αυτά από κάθε δραστηριότητα.
- Χρησιμοποιήσαμε αλγορίθμους μηχανικής μάθησης για να ταξινομήσουμε τις δραστηριότητες αυτές.

5.2 Σχολιασμός Αποτελεσμάτων

Τα ακριβή ποσοστά επιτυχίας για κάθε αλγόριθμο που χρησιμοποιήσαμε και δοκιμάσαμε παρουσιάζονται εκτενώς στο Κεφάλαιο 4. Στην παρούσα εννότητα θα γίνει σχλιασμός αυτών και θα βγουν κάποια γενικά συμπεράσματα.

Τα αποτελέσματα των αλγορίθμων κινούνται κατά μέσο όρο σε ποσοστά από 60 έως 70 τοις εκατό. Υπάρχουν και κάποιες περιπτώσεις της τάξεως του 45 έως 60 τοις εκατό τις οποίες και θα σχολιάσουμε παρακάτω.

Το γεγονός ότι τα ποσοστά κειμένονται σε αυτά τα πλαίσια είναι θετικό καθώς δείχνει μια συσχέτιση των δεδομένων με τις ετικέτες. Ενώ στη αρχή είχαμε κάποια κείμενα χωρίς οποιαδήποτε ταξινόμηση, ή κάποιο διαχωρισμό χαρακτηριστικών, καταφέραμε να βρούμε έναν τρόπο να τα ταξινομήσουμε επιλέγοντας με βάση την έρευνα μας κάποια συγκεκριμένα χαρακτηριστικά. Σε γενικές γραμμές φαίνεται πως η μέθοδος μας λειτουργεί καλά από πλευράς λογικής καθώς προσεγγίζουμε επιτυχία σχεδόν 70%.

Σχετικά με την συνεισφορά των χαρακτηριστικών όπως παρατηρήσαμε και στο προηγούμενο στάδιο, βλέπουμε πως το ρήμα είναι ίσως το πιο σημαντικό χαρακτηριστικό. Στους περισσότερους αλγορίθμους παρατηρούμε πως το σύνολο X1 του test set που αποτελείται μόνο από το ρήμα σαν χαρακτηριστικό, έχει ένα ενδεικτικό ποσοστό ακρίβειας το οποίο μεταβάλλεται ελάχιστα με την προσθήκη των υπολοίπων χαρακτηριστικών. Ακόμα και όταν υπάρχει αύξηση του ποσοστού ακρίβειας βλέπουμε πως τα νούμερα δεν μεταβάλλονται πολύ, αλλά παραμένουν γύρω από το ποσοστό του X1 συνόλου. Άρα συμπεραίνουμε πως ήδη από την αναγνώριση του ρήματος, μπορούμε κατά ένα πολύ μεγάλο ποσοστο να κατηγοριοποιήσουμε την δραστηριότητα πο έχουμε εντοπίσει χωρίς να λάβουμε υπόψιν μας άλλο χαρακτηριστικό.

Παρατηρούμε ακόμα πως τόσο το χαρακτηριστικό του αντικειμένου όσο και το χαρακτηριστικό της κατηγορίας του υποκειμένου, προσθέτει κάποιες φορές περισσότερη ακρίβεια στα αποτελέσματα μας, χωρίς όμως να αποκλείνουν πολύ από τα αρχικά. Για παράδεγμα μπορούμε να δούμε ενδεικτικά τα αποτελέσματα του αλγορίθμου Decision Tree με χρήση One Hot Encoder. Εκεί το σύνολο X1 με χαρακτηριστικό μόνο το ρήμα είχε ποσοστό ακρίβειας 61% ενώ η προσθήκη των χαρακτηριστικών του αντικειμένου και της κατηγορίας του υποκειμένου το ανέβασαν σε ποσοστό απο 64% έως 66%.

Το χαρακτηριστικό που αποδείχθηκε ότι παίζει τον μικρότερο ρόλο είναι το αν η πρόταση περιλαμβάνει κάποια λέξη του τομέα του IT. Σε όλους τους αλγορίθμους παρατηρήσαμε μηδενική μεταβολή των αποτελεσμάτων προσθέτοντας αυτό το χαρακτηριστικό στο σύνολο δεδομένων, και ακόμα σε συγκεκριμένες περιπτώσεις στους αλγορίθμους Random Forest και SVM είδαμε και μείωση της ακρίβειας σε σχέση με το ποσοστό που είχε το σύνολο X1.

Αναφορικά με τους τρόπους κωδικοποίησης των δεδομένων παρατηρήσαμε και εκεί κάποιες αποκλισεις στα αποτελέσματα. Επιγραμματικά χρησιμοποιήσαμε δύο τρόπους για την κωδικοποίηση των string δεδομένων οι οποίοι ήταν το One Hot Encoding που ανήκει στην κατηγορορία του Bag of Words και το Word2Vec που ανήκει στην κατηγορία του Word Embeddings. Προς εκκπληξή μας τα αποτελέσματα ήταν καλύτερα με την χρήση του One Hot Encoding παρά με το Word2Vec. Συγκεκριμένα φάνηκε με την χρήση όλων των αλγορίθμων μια μείωση της ακρίβειας αλλάζοντας τον τρόπο κωδικοποίησης. Η μείωση αυτή ήταν πιο έντονη στον αλγόριθμο SVM.

Αξίζει να τονίσουμε ωστόσο πως για την χρήση των embeddings απαιτούνται μεγάλες ποσότητες τόσο στα συνολικά δεδομένων όσο και σε επαναλαμβανόμενες εμφανίσεις μεμονωμένων παραδειγμάτων, καθώς και μακράς διάρκειας εκπαίδευση. Στη δική μας περίπτωση το Data Set ήταν μικρού μεγέθους χωρίς πολλά επαναλαμβανόμενα δεδομένα και για αυτό ίσως να μην μας απέδωσε καλά αποτελέσματα αυτή η μέθοδος.

Τέλος θα κάνουμε κάποια σχόλια για τους αλγόριθμους που χρησιμοποιήσαμε. Βλέπουμε πως οι αλγόριθμοι Decision Tree και Random Forest επιστρέφουν κοντινά αποτελέσματα γεγονός που είναι λογικό καθώς πρόκειται για δύο παρόμοιες μεθόδους όπως αναφέραμε και παραπάνω, αφού το Random Forest χρησιμοποιεί Decision Trees. Δοκιμάζοντας την μέθοδο SVM είδαμε και εκεί παρόμοια ποσοστά ακρίβειας με μια απειρο-ελάχιστη μείωση της τάξης 1 έως 3 %, έχοντας εξαφανίσει όμως το φαινόμενο του overfitting που παρατηρήσαμε στις προηγούμενες δύο μεθόδους. Ο αλγόριθμος SVM παρόλο που είναι πιο περίπλοκος δεν εγγυάται καλύτερα αποτελέσματα από κάποιον άλλον. Αυτό εξαρτάται από το πρόβλημα και τα δεδομένα και συνήθως χρειάζεται μεγάλο όγκο δεδομένων για να επιστρέψει πιο έγκυρα αποτελέσματα. Στην δική μας περίπτωση και με τα δεδομένα που καταφέραμε να συλλέξουμε πρατηρούμε πως όλοι οι αλγόριθμοι επιστρέφουν παρεμφερή αποτελέσματα με μια μικρή προτεραιότητα στον αλγόριθμο Decision Tree που φαίνεται συνολικά να φέρει τα καλύτερα αποτελέσματα για το test set.

5.3 Περιορισμοί

Πραγματοποιήσαμε ανάλυση σφαλμάτων για να λάβουμε πληροφορίες σχετικά με τα όρια που είχε η προσέγγιση μας και για ποιον λόγο δεν καταφέραμε να λάβουμε υψηλότερα ποσοστά ακριβείας.

Πιο συγκεκριμένα, διερευνήσαμε ποιες περιπτώσεις εργασιών ταξινομήθηκαν λανθασμένα και γιατί. Στην ουσία, παρατηρήσαμε τρεις βασικούς τύπους εσφαλμένων ταξινομήσεων: εσφαλμένες ταξινομήσεις λόγω λάθους αναγνώρισης δραστηριότητας ή λόγω λάθους εξαγωγής χαρακτηριστικών, εσφαλμένες ταξινομήσεις λόγω παρεκκλίνουσας χρήσης χαρακτηριστικών και εσφαλμένες ταξινομήσεις λόγω ανεπαρκών δεδομένων εκπαίδευσης.

Η πρώτη κατηγορία σχετίζεται με έλλειψη ακρίβειας στον τρόπο αναγνώρισης προτάσεων, δραστηριοτήτων ή και εξαγωγής χαρακτηριστικών. Σε αυτό μπορεί να οφείλεται κάποια αδυναμία των βιβλιοθηκών που χρησιμοποιήθηκαν ή και του τρόπου ανάλυσης.

Η επόμενη κατηγορία σχετίζεται με δραστηριότητες που ταξινομήθηκαν λανθασμένα επειδή τα χαρακτηριστικά γνωρίσματα συνήθως σχετίζονταν με άλλη τάξη. Για παράδειγμα, ας δούμε την χειροκίνητη εργασία «επισυνάψαν το νέο έγγραφο sct». Για αυτήν την εργασία,

η προσέγγισή μας δεν γνωρίζει το γεγονός ότι το «έγγραφο sct» είναι στην πραγματικότητα ένα φυσικό έγγραφο. Το ταξινομεί ως εργασία χρήστη επειδή το ρήμα «επισυνάπτω» συσχετίζεται συχνά με εργασίες χρήστη στο σύνολο δεδομένων μας (π.χ. «επισύναψη e-mail»).

Η τρίτη κατηγορία εσφαλμένων ταξινομήσεων αφορά περιπτώσεις στις οποίες η προσέγγισή μας ταξινόμησε εσφαλμένα μια εργασία επειδή δεν έχει δει αρκετά εκπαιδευτικά δεδομένα. Για παράδειγμα, στην εργασία χρήστη «μεταδώστε ένα σχόλιο απάντησης», η προσέγγισή μας την ταξινόμησε ως μη αυτόματη εργασία. Εδώ το πρόβλημα είναι ότι η προσέγγισή μας δεν έχει παρατηρήσει επαρκή αριθμό δεδομένων χρησιμοποιώντας το ρήμα «μεταδίδω», το οποίο σχετίζεται σαφώς με τη χρήση ενός συστήματος πληροφοριών.

Παρά αυτές τις εσφαλμένες ταξινομήσεις, μπορούμε να δηλώσουμε ότι η παρουσιαζόμενη προσέγγιση αντιπροσωπεύει μια πολλά υποσχόμενη λύση για τον αυτόματο προσδιορισμό του βαθμού αυτοματοποίησης εργασιών.

5.4 Συμπεράσματα

Από όλα τα παραπάνω που παρουσιάστηκαν στην παρούσα διπλωματική μπορούμε να βγάλουμε τα εξής συμπεράσματα. Η μέθοδος που δημιουργήσαμε και ακολουθήσαμε είναι μια καλή βάση για το ξεκίνημα της αυτόματης αναγνώρισης διαδικασιών προς αυτοματοποίηση. Φαίνεται πως υπάρχει συσχέτιση των χαρακτηριστικών που διαλέξαμε και του βαθμού αυτοματοποιήσης τους και αυτό αποδεικνύεται από τα ποσοστά ακρίβειας που για μια πρώτη προσέγγιση είναι πολύ ενθαρυντικά. Ο βασικός σκοπός πίσω από την ιδέα που αναπτύχθηκε ήταν η μείωση του φόρτου εργασίας του εργαζόμενου ως αναλυτή και ακόμα και με επιτυχία 70% θα μπορούσε να γίνει μια αρχική ταξινόμηση των δραστηριοτήτων πριν ξεκινήσει ο άνθρωπος την χειρονακτική ανάλυση. Εννοείται πρέπει να ληφθούν υπόψιν όλοι οι περιορισμοί που αναφέραμε παραπάνω καθώς και ότι το data set δεν είναι ούτε ενδεικτικό ούτε ιδανικό για τις περιγραφές των διαδικασιών. Στην πραγματικότητα, οι περιγραφές των διαδικασιών κειμένου ενδέχεται να αποκλίνουν από αυτές στο σύνολο δεδομένων μας με διάφορους τρόπους. Ωστόσο, προσπαθήσαμε να μεγιστοποιήσουμε την εγκυρότητα της αξιολόγησής μας επιλέγοντας ένα σύνολο δεδομένων που συνδυάζει διαφορετικές πηγές. Επιπλέον, η προσέγγισή μας θα μπορούσε εύκολα να επανεκπαιδευτεί σε άλλα σύνολα δεδομένων για να αυξήσει περαιτέρω την απόδοσή της.

5.5 Μελλοντικές Επεκτάσεις

Ενώ η αξιολόγηση που πραγματοποιήθηκε σε αυτή τη διπλωματική απέδειξε ενθαρρυντικά αποτελέσματα, διάφορες βελτιώσεις θα μπορούσαν να ακολουθηθούν προκειμένου να ενισχυθεί η ποιότητα της διαδικασίας αναγνώρισης διαδικασιών προς αυτοματοποίηση.

Το πιο σημαντικό κομμάτι που ήταν και το κύριο μειονέκτημα αυτής της λύσης που παρουσιάζουμε στη διπλωματική είναι το data set. Θα μπορούσε να γίνει μια καλύτερη έρευνα για εύρεση ενός πιο αντιπροσωπευτικού δείγματος κειμένων ή ίσως κάποια παραχώρηση από κάποια επιχείρηση, ώστε τα αποτελέσματα τα οποία θα έβγαιναν να έδειχναν καλύτερα την επιτυχία ή αποτυχία της διαδικασίας. Με αυτόν τον τρόπο θα βελτιώνονταν τα αποτελέσματα σε ένα πολύ μεγάλο ποσοστό. Πέρα από τη σημασιολογική πτυχή, θα βοηθούσε

να υπήρχαν κείμενα και σε μεγαλύτερη ποσότητα αλλά και σε καλύτερη ποιότητα. Όπως είδαμε παραπάνω ένα ποσοστό αποτυχίας οφειλόταν στην αδυναμία της μεθόδου να αναγνωρίσει προτάσεις και δραστηριότητες. Σε αυτό μπορεί να οφείλεται όχι μόνο η αδυναμία της μεθόδου και της βιβλιοθήκης αλλά και του αδύναμου συντακτικού που υπήρχε σε κάποια κείμενα.

Επίσης θα μπορούσε να γίνει παραιτέρω έρευνα σχετικά με τα χαρακτηριστικά που πιθανότατα να σχετίζονται με την ταξινόμηση ως προς την αυτοματοποίηση μιας διαδικασίας. Εμείς ορίσαμε 4 κατηγορίες και εξάγαμε και δουλέψαμε με αυτά τα χαρακτηριστικά. Θα μπορούσε να γίνει μια προσθήκη σε αυτό το σύνολο των χαρακτηριστικών και να γίνει ξανά εκπαίδευση των αλγορίθμων ώστε να αυθηθούν τα ποσοστά εγκυρότητας.

Προαιρετικά θα μπορούσε να γίνει κάποια εναλλαγή στην επιλογή βιβλιοθηκών που επιλέχθηκαν και χρησιμοποήθηκαν προκειμένου να υπάρχει μεγαλύτερο ποσοστό ακρίβειας στην αναγνώριση των δρστηριοτήτων μέσα από τα κείμενα.

Τέλος, η μεταφορά του συστήματος σε άλλες γλώσσες είναι ένα σημαντικό έργο. Μια ενσωμάτωση θα πρέπει να είναι εύκολα εφικτή καθώς κάποια συστατικά που χρησιμοποιήθηκαν έχουν ήδη εκπαιδευτεί σε διαφορετικές γλώσσες. Δυστυχώς, η εκπροσώπηση του Stanford Dependency είναι προς το παρόν διαθέσιμη μόνο για Αγγλικά και Κινέζικα. Καθώς η ανάλυση των Κινέζικων θέτει διαφορετικές προκλήσεις, π.χ. δεδομένου ότι οι λέξεις συνήθως δεν χωρίζονται χρησιμοποιώντας κενό διάστημα, δεν εμπίπτει στο πεδίο αυτής της διατριβής, αλλά παρουσιάζει μια ενδιαφέρουσα κατεύθυνση για περαιτέρω έρευνα. Εννοείται θα ήταν πολύ ενδιαφέρον να γίνει και μια υλοποίηση για τν ελληνική γλώσσα, ώστε να μπορέσει η μέθοδος να εφαρμοστεί σε ελληνικές επιχειρήσεις.

Παραρτήματα

Παράρτημα Α΄

Παραδείγματα Βιβλιογραφικών Αναφορών

Τύπος βιβλιογραφικής πηγής	**Αριθμός αναφοράς**
Βιβλίο ξενόγλωσσο	[1]
	[2]
	[3]
	[4]
	[5]
	[6]
Άρθρο σε επιστημονικό περιοδικό	[7] , [8]
	[9] , [10]
	[11] , [12]
	[13] , [14]
	[15] , [16]
	[17] , [18]
	[19] ,[20]
	[21] , [22]
	[23] , [24]
	[25] , [26]
	[27] , [28]
	[29] , [30]
	[31] , [32]
	[33] , [34]
	[35] , [36]
Ιστοσελίδα	[37]
	[38]
	[39]
	[40]
	[41]
	[42]
	[43]
	[44]
	[45]
	[46]
	[47]

Παράρτημα Β΄

Person Corrector List

Αυτή η ενότητα παραθέτει την λίστα λέξεων προσώπων που χρησιμοποιήθηκε κατά τη διαδικασία ως περιγράφεται στην ενότητα 3. Οι λέξεις λήφθηκαν βάσει διερευνητικής ανάλυσης και σύμφωνα με το βιβλίο «Practical English Usage».

1. resource provisioning
2. customer service
3. support
4. support office
5. support officer
6. client service back office
7. master
8. masters
9. assembler ag
10. acme ag
11. acme financial accounting
12. secretarial office
13. office
14. registry
15. head
16. storehouse
17. atm
18. crs
19. company
20. garage
21. kitchen
22. department
23. ec
24. sp
25. mpo
26. mpoo
27. mpon
28. msp
29. mspo
30. mspn
31. go
32. pu
33. ip
34. inq
35. sp/pu/go
36. fault detector

Παράρτημα Γ΄

IT Word List of Utah University

Αυτή η ενότητα παραθέτει την λίστα λέξεων τεχνολογίας του πανεπιστημίου Υταη που χρησιμοποιήθηκε κατά τη διαδικασία ως περιγράφεται στην ενότητα 3.

1. system
2. send email
3. automate
4. access
5. Access Control List
6. access time
7. account
8. account name
9. address
10. aggregate
11. aggregate data
12. algorithm
13. alias
14. analog
15. Application Layer
16. application
17. application program
18. application software
19. archive,
20. argument
21. ASCII
22. assembler
23. ATM
24. ATM Forum
25. audit
26. authentication
27. authorization
28. autonomous system
29. backbone
30. background processing
31. backspace
32. backup
33. bandwidth
34. baseband
35. BASIC
36. batch processing
37. batch query
38. binary
39. binary number
40. bit
41. bitmapped terminal
42. BITNET
43. bits per second (bps)
44. block
45. bold
46. booting
47. break
48. bridge
49. broadband
50. broadcast
51. browser
52. buffer
53. bug
54. bulletin board (BBS)
55. BUS topology
56. byte
57. cable
58. carriage return
59. CD-ROM
60. cell relay
61. channel
62. character
63. character set
64. chip
65. client
66. client/server

67. Client-Server Interface
68. COBOL
69. code
70. collision
71. column
72. command,
73. communications line
74. communications program
75. compiler
76. computer
77. concentrator
78. conference
79. configuration
80. connect time
81. control character
82. control key
83. copy
84. CPU
85. crash
86. cursor
87. cursor control
88. Cyberspace
89. Data Link Layer
90. data
91. data communications
92. data entry
93. data processing
94. Dataset
95. database
96. database management system
97. DBMS
98. debug
99. default
100. delete key
101. DHCP
102. dial-up
103. dictionary file
104. digital
105. direct access
106. directory
107. disk
108. diskette
109. display
110. distributed
111. distributed application
112. distributed database
113. distributed file system
114. document
115. documentation
116. DOS
117. dot-matrix printer
118. down
119. download
120. downtime
121. drag and drop
122. drive
123. dump
124. edit
125. editor,email
126. e-mail
127. e-mail address
128. e-mail server
129. e-mail service
130. e-mail system
131. encapsulation
132. enter key
133. environment
134. erase
135. error message
136. error checking
137. Ethernet
138. execute
139. fiber optics
140. field
141. file
142. file format
143. file server
144. folder
145. font
146. foreground
147. form
148. form feed
149. format
150. FORTRAN
151. fragment
152. frame
153. freeware
154. frequency
155. FAQ
156. FTP
157. FUD
158. function key
159. garbage
160. gateway
161. GIF

162. gopher
163. graphic
164. Groupware
165. GUI
166. handshaking
167. hang
168. hard copy
169. hard disk
170. hardware
171. hardwired
172. header
173. help
174. hierarchical file
175. hierarchical file structure
176. host
177. host computer
178. HTML
179. HTTP
180. hub
181. hyperlink
182. hypermedia
183. hypertext
184. icons
185. I/O
186. IEEE
187. inbox
188. index
189. information hiding
190. information server
191. information super-highway
192. inheritance
193. input
194. instance
195. instantiation
196. instruction
197. interactive
198. INTERNET
199. IP
200. IP Address
201. interrupt
202. IRC
203. ISDN
204. ISO
205. job
206. JPEG
207. justify
208. Kermit
209. keykeyboard
210. kilobyte(K)
211. LAN
212. LAN e-mail system
213. laserdisc
214. laser printer
215. Layer
216. line
217. line editor
218. line printer
219. link
220. LISTSERV
221. load
222. logical record
223. login or logon
224. login ID
225. logoff
226. Longitudinal Study
227. LPR
228. lynx
229. machine language
230. macro
231. magnetic disk
232. magnetic tape
233. MAIL
234. mailbox
235. MAILER
236. main memory
237. mainframe
238. mainframe
239. minicomputer
240. micro-computer
241. MB
242. medium
243. memory
244. menu
245. message
246. method
247. methodology
248. microcomputer
249. microprocessor
250. Microwave
251. mission
252. modem
253. modem setup
254. module
255. monitor
256. Mosaic
257. mouse
258. multimedia
259. multimedia mail
260. multiplexer
261. multiuser

262. nesting
263. NetScape
264. Network Layer
265. network
266. nickname
267. node
268. noise
269. object
270. object-based
271. object code
272. object-oriented
273. object-oriented technology
274. OLE
275. off-line
276. on-line
277. Online Service
278. open
279. open platform
280. open system
281. OSI
282. OpenWindows
283. operating system
284. output
285. packet
286. parameter
287. parity
288. password
289. peripheral
290. PC
291. Physical Layer
292. ping
293. pixel
294. platform
295. plotter
296. polymorphism
297. port
298. portable
299. post
300. PostScript
301. Power PC
302. Presentation layer
303. printer
304. printout
305. procedure
306. process
307. program
308. programmer
309. programming
310. prompt
311. protocol
312. public domain
313. quality
314. query
315. queue
316. quit
317. RAID
318. RAM
319. random access
320. Re-engineering
321. read
322. read/write
323. realtime
324. record
325. record length
326. record type
327. recovery
328. rectangular file
329. reel tape
330. relational database
331. relational structure
332. remote
333. remote access
334. resource
335. response
336. retiming
337. return key
338. reuse and reuse-ability
339. reverse engineering
340. ROM
341. root directory
342. router
343. routine
344. routing
345. run
346. scanner
347. scheduling
348. screen
349. screen editor
350. scroll
351. segment
352. sequential
353. server
354. service
355. service provider
356. session
357. Session Layer
358. shareware
359. shell
360. simulation
361. smiley

362. soft copy
363. software
364. software tool
365. sort
366. source code
367. SPARC
368. SPARCstation
369. sponge
370. spool
371. spreadsheet
372. SQL
373. storage
374. strategy
375. string
376. striping
377. Sun Microsystems
378. SunOS
379. surfing
380. tape density
381. task
382. TCP/IP
383. TEAM
384. telecommunication
385. telecomputing
386. TELNET
387. terminal
388. terminal emulation
389. terminal server
390. terabyte
391. text
392. time out
393. time series
394. TN3270
395. toggle
396. token ring
397. topic
398. transfer
399. Transport Layer
400. tree
401. UNIX
402. upload
403. URL
404. Usenet
405. user
406. user-friendly
407. userid
408. username
409. utility
410. variable
411. vision
412. virtual
413. virtual terminal
414. VMS
415. virus
416. volume
417. wavelength
418. whois
419. window
420. Windows
421. word processor
422. wordwrap
423. work space
424. workstation
425. write
426. WWW
427. X window system
428. X-term

Παράρτημα Δ΄

Αναλυτικό Data Set

Δ΄.1 Humboldt-Universit at zu Berlin

A small company manufactures customized bicycles. Whenever the sales department receives an order, a new process instance is created. A member of the sales department an then reject or accept the order for a customized bike. In the former case, the process instance is finished. In the latter case, the storehouse and the engineering department are informed. The storehouse immediately processes the part list of the order and checks the required quantity of each part. If the part is available in-house, it is reserved. If it is not available, it is back-ordered. This procedure is repeated for each item on the part list. In the meantime, the engineering department prepares everything for the assembling of the ordered bicycle. If the storehouse has successfully reserved or back-ordered every item of the part list and the preparation activity has finished, the engineering department assembles the bicycle. Afterwards, the sales department ships the bicycle to the customer and finishes the process instance.

Text 1: Process Description 1-1: Bicycle manufacturing.

A customer brings in a defective computer and the CRS checks the defect and hands out a repair cost calculation back. If the customer decides that the costs are acceptable, the process continues, otherwise she takes her computer home unrepaired. The ongoing repair consists of two activities, which are executed, in an arbitrary order. The first activity is to check and repair the hardware, whereas the second activity checks and configures the software. After each of these activities, the proper system functionality is tested. If an error is detected another arbitrary repair activity is executed, otherwise the repair is finished.

Text 2: Process Description 1-2: Computer repair.

The Evanstonian is an upscale independent hotel. When a guest calls room service at The Evanstonian, the room-service manager takes down the order. She then submits an order ticket to the kitchen to begin preparing the food. She also gives an order to the sommelier (i.e., the wine waiter) to fetch wine from the cellar and to prepare any other alcoholic beverages. Eighty percent of room-service orders include wine or some other alcoholic beverage. Finally, she assigns the order to the waiter. While the kitchen and

the sommelier are doing their tasks, the waiter readies a cart (i.e., puts a tablecloth on the cart and gathers silverware). The waiter is also responsible for nonalcoholic drinks. Once the food, wine, and cart are ready, the waiter delivers it to the guests room. After returning to the room-service station, the waiter debits the guests account. The waiter may wait to do the billing if he has another order to prepare or deliver.

Text 3: Process Description 1-3: Hotel Service.

Whenever a company makes the decision to go public, its first task is to select the underwriters. Underwriters act as financial midwives to a new issue. Usually they play a triple role: First they provide the company with procedural and financial advice, then they buy the issue, and finally they resell it to the public. Established underwriters are careful of their reputation and will not handle a new issue unless they believe the facts have been presented fairly. Thus, in addition to handling the sale of a company's issue, the underwriters in effect give their seal of approval to it. They prepare a registration statement for the approval of the Securities and Exchange Commission (SEC). In addition to registering the issue with the SEC, they need to check that the issue complies with the so-called blue-sky laws of each state that regulate sales of securities within the state. While the registration statement is awaiting approval, underwriters begin to firm up the issue price. They arrange a road show to talk to potential investors. Immediately after they receive clearance from the SEC, underwriters fix the issue price. After that they enter into a firm commitment to buy the stock and then offer it to the public, when they haven't still found any reason not to do it.

Text 4: Process Description 1-4: Underwriters.

Δ΄.2 Technische Universit at Berlin

At the beginning the customer perceives that her subscribed service has degraded. A list with all the problem parameters is then sent to the Customer Service department of TELECO. At the customer service an employee enters (based on the received data) a problem report into system T.. Then the problem report is compared to the customer SLA to identify what the extent and the details of the service degradation are. Based on this, the necessary counter measures are determined including their respective priorities. An electronic service then determines the significance of the customer based on information that has been collected during the history of the contractual relationship. In case the customer is premium, the process will link to an extra problem fix process (this process will not be detailed here). In case the customer is of certain significance which would affect the counter measures previously decided upon, the process goes back to re-prioritize these measures otherwise the process continues. Taking together the information (i.e. contract commitment data + prioritized actions) a detailed problem report is created. The detailed problem report is then sent to Service Management. Service Management deals on a first level with violations of quality in services that are provided to customers. After receiving the detailed problem report, Service management investigates whether the problem is analyzable at the level of their department or whether the

problem may be located at Resource Provisioning. In case Service Management assesses the problem to be not analyzable by themselves, the detailed problem report is sent out to Resource Provisioning. If Service Management is sure they can analyze it, they perform the analysis and based on the outcome they create a trouble report that indicates the type of problem. After Resource Provisioning receives the detailed problem report, it is checked whether there are any possible problems. If no problems are detected, a notification about the normal service execution is created. If a problem is detected this will be analyzed by Resource Provisioning and a trouble report is created. Either trouble report or the normal execution notification will be included in a status report and sent back to Service Management. Service Management then prepares the final status report based on the received information. Subsequently it has to be determined what counter measures should be taken depending on the information in the final status report. Three alternative process paths may be taken. For the case that no problem was detected at all, the actual service performance is sent back to the Customer Service. For the case that minor corrective actions are required, Service Management will undertake corrective actions by themselves. Subsequently, the problem resolution report is created and then sent out to Customer Service. After sending, this process path of Service Management ends. For the case that automatic resource restoration from Resource Provisioning is required, Service Management must create a request for automatic resource restoration. This message is then sent to Resource Provisioning. Resource Provisioning has been on-hold and waiting for a restoration request but this must happen within 2 days after the status report was sent out, otherwise Resource Provisioning terminates the process. After the restoration request is received, all possible errors are tracked. Based on the tracked errors, all necessary corrective actions are undertaken by Resource Provisioning. Then a trouble-shooting report is created. This report is sent out to Service Management; then the process ends. The trouble-shooting report is received by Service Management and this information goes then into the creation of the problem resolution report just as described for ii). Customer Service either receives the actual service performance (if there was no problem) or the problem resolution report. Then, two concurrent activities are triggered, i.e. i) a report is created for the customer which details the current service performance and the resolution of the problem, and ii) an SLA violation rebate is reported to Billing Collections who will adjust the billing. The report for the customer is sent out to her. After all three activities are completed the process ends within Customer Service. After the customer then receives the report about service performance and problem resolution from Customer Service, the process flow at the customer also ends.

Text 5: Process Description 2-1: SLA Violation.

The process is initiated by a switch-over request. In doing so, the customer transmits his data to the customer service department of the company. Customer service is a shared service center between the departments Sales and Distribution. The customer data is received by customer service and based on this data a customer data object is entered into the CRM system. After customer data has been entered it should then be compared with the internal customer data base and checked for completeness and

plausibility. In case of any errors these should be corrected on the basis of a simple error list. The comparison of data is done to prevent individual customer data being stored multiple times. In case the customer does not exist in the customer data base, a new customer object is being created which will remain the data object of interest during the rest of the process flow. This object consists of data elements such as the customers name and address and the assigned power gauge. The generated customer object is then used, in combination with other customer data to prepare the contract documents for the power supplier switch (including data such as bank connection, information on the selected rate, requested date of switch-over). In the following an automated check of the contract documents is carried out within the CIS (customer information system) in order to confirm their successful generation. In case of a negative response, i.e. the contract documents are not (or incorrectly) generated, the causing issues are being analyzed and resolved. Subsequently the contract documents are generated once again. In case of a positive response a confirmation document is sent out to the customer stating that the switch-over to the new supplier can be executed. A request to the grid operator is automatically sent out by the CIS. It puts the question whether the customer may be supplied by the selected supplier in the future. The switch-over request is checked by the grid operator for supplier concurrence, and the grid operator transmits a response comment. In the case of supplier concurrence the grid operator would inform all involved suppliers and demand the resolution of the conflict. The grid operator communicates with the old supplier and carries out the termination of the sales agreement between the customer and the old supplier (i.e. the customer service (of the new supplier) does not have to interact with the old supplier regarding termination). If there are not any objections by the grid operator (i.e. no supplier concurrence), customer service creates a CIS contract. The customer then has the chance to check the contract details and based on this check may decide to either withdraw from the switch contract or confirm it. Depending on the customers acceptance/rejection the process flow at customer service either ends (in case of withdrawal) or continues (in case of a confirmation). An additional constraint is that the customer can only withdraw from the offered contract within 7 days after the 7th day the contract will be regarded as accepted and the process continues. The confirmation message by the customer is therefore not absolutely necessary (as it will count as accepted after 7 days in any way) but it can speed up the switch process. On the switch-date, but no later than 10 days after power supply has begun, the grid operator transmits the power meter data to the customer service and the old supplier via messages containing a services consumption report. At the same time, the grid operator computes the final billing based on the meter data and sends it to the old supplier. Likewise the old supplier creates and sends the final billing to the customer. For the customer as well as the grid operator the process ends then. After receiving the meter data customer service imports the meter data to systems that require the information. The process of winning a new customer ends here.

Text 6: Process Description 2-2: Supplier Switch.

Δ΄.3 Queensland University of Technology

The party sends a warrant possession request asking a warrant to be released. The Client Service Back Office as part of the Small Claims Registry Operations receives the request and retrieves the SCT file. Then, the SCT Warrant Possession is forwarded to Queensland Police. The SCT physical file is stored by the Back Office awaiting a report to be sent by the Police. When the report is received, the respective SCT file is retrieved. Then, Back Office attaches the new SCT document, and stores the expanded SCT physical file. After that, some other MC internal staff receives the physical SCT file (out of scope).

Text 7: Process Description 3-1: 2009-1 MC Finalise SCT Warrant Posession.

Each morning, the files which have yet to be processed need to be checked, to make sure they are in order for the court hearing that day. If some files are missing, a search is initiated, otherwise the files can be physically tracked to the intended location. Once all the files are ready, these are handed to the Associate, and meantime the Judges Lawlist is distributed to the relevant people. Afterwards, the directions hearings are conducted.

Text 8: Process Description 3-2: 2009-2 Conduct Directions Hearing.

After a claim is registered, it is examined by a claims officer. The claims officer then writes a ”settlement recommendation”. This recommendation is then checked by a senior claims officer who may mark the claim as ”OK” or ”Not OK”. If the claim is marked as ”Not OK”, it is sent back to the claims officer and the recommendation is repeated. If the claim is OK, the claim handling process proceeds.

Text 9: Process Description 3-3: 2009-3 Repetition - Cycles.

In the context of a claim handling process, it is sometimes necessary to send a questionnaire to the claimant to gather additional information. The claimant is expected to return the questionnaire within five days. If no response is received after five days, a reminder is sent to the claimant. If after another five days there is still no response, another reminder is sent and so on until the completed questionnaire is received.

Text 10: Process Description 3-4: 2009-4 Event-based Gateways.

Mail from the party is collected on a daily basis by the Mail Processing Unit. Within this unit, the Mail Clerk sorts the unopened mail into the various business areas. The mail is then distributed. When the mail is received by the Registry, it is opened and sorted into groups for distribution, and thus registered in a manual incoming Mail Register. Afterwards, the Assistant Registry Manager within the Registry performs a quality check. If the mail is not compliant, a list of requisition explaining the reason for rejection is compiled and sent back to the party. Otherwise, the matter details (types of action) are captured and provided to the Cashier, who takes the applicable fees attached to the mail. At this point, the Assistant Registry Manager puts the receipt and copied documents into an envelope and posts it to the party. Meantime, the Cashier captures the Party Details and prints the Physical Court File.

Text 11: Process Description 3-5: 2009-5 PE - Lodge Originating Document by Post.

When a claim is received, it is first checked whether the claimant is insured by the organization. If not, the claimant is informed that the claim must be rejected. Otherwise, the severity of the claim is evaluated. Based on the outcome (simple or complex claims), relevant forms are sent to the claimant. Once the forms are returned, they are checked for completeness. If the forms provide all relevant details, the claim is registered in the Claims Management system, which ends the Claims Notification process. Otherwise, the claimant is informed to update the forms. Upon reception of the updated forms, they are checked again.

Text 12: Process Description 3-6: 2010-1 Claims Notification.

The Police Report related to the car accident is searched within the Police Report database and put in a file together with the Claim Documentation. This file serves as input to a claims handler who calculates an initial claim estimate. Then, the claims handler creates an Action Plan based on an Action Plan Checklist available in the Document Management system. Based on the Action Plan, a claims manager tries to negotiate a settlement on the claim estimate. The claimant is informed of the outcome, which ends the process.

Text 13: Process Description 3-7: 2010-2 Claims Creation.

Δ΄.4 Technische Universiteit Eindhoven

The intake workflow starts with a notice by telephone at the secretarial office of the mental health care institute. This notice is done by the family doctor of somebody who is in need of mental treatment. The secretarial worker inquires after the name and residence of the patient. On the basis of this information, the doctor is put through to the nursing officer responsible for the part of the region that the patient lives in. The nursing officer makes a full inquiry into the mental, health, and social state of the patient in question. This information is recorded on a registration form. At the end of the conversation, this form is handed in at the secretarial office of the institute. Here, the information on the form is stored in the information system and subsequently printed. For new patients, a patient file is created. The registration form as well as the print from the information system are stored in the patient file. Patient files are kept at the secretarial office and may not leave the building. At the secretarial office, two registration cards are produced for respectively the future first and second intaker of the patient. The registration card contains a set of basic patient data. The new patient is added on the list of new notices. Halfway the week, at Wednesday, a staff meeting of the entire medical team takes place. The medical team consists of social-medical workers, physicians, and a psychiatrist. At this meeting, the team-leader assigns all new patients on the list of new notices to members of the team. Each patient will be assigned to a social-medical worker, who will act as the first intaker of the patient. One of the physicians will act as the second intaker. In assigning intakers, the team-leader takes into account their expertise, the

region they are responsible for, earlier contacts they might have had with the patient, and their case load. The assignments are recorded on an assignment list which is handed to the secretarial office. For each new assignment, it is also determined whether the medical file of the patient is required. This information is added to the assignment list. The secretarial office stores the assignment of each patient of the assignment list in the information system. It passes the produced registration cards to the first and second intaker of each newly assigned patient. An intaker keeps this registration with him at times when visiting the patient and in his close proximity when he is at the office. For each patient for which the medical file is required, the secretarial office prepares and sends a letter to the family doctor of the patient, requesting for a copy of the medical file. As soon as this copy is received, the secretarial office will inform the second intaker and add the copy to the patient file. The first intaker plans a meeting with the patient as soon as this is possible. During the first meeting, the patient is examined using a standard checklist which is filled out. Additional observations are registered in a personal notebook. After a visit, the first intaker puts a copy of these notes in the file of a patient. The standard checklist is also added to the patient's file. The second intaker plans the first meeting only after the medical information of the physician if required has been received. Physicians use dictaphones to record their observations made during meetings with patients. The secretarial office types out these tapes, after which the information is added to the patient file. As soon as the meetings of the first and second intaker with the patient have taken place, the secretarial office puts the patient on the list of patients that reach this status. For the staff meeting on Wednesday, they provide the team-leader with a list of these patients. For each of these patients, the first and second intaker together with the team-leader and the attending psychiatrist formulate a treatment plan. This treatment plan formally ends the intake procedure.

Text 14: Process Description 4-1: Intaker Workflow.

Δ΄.5 BPM Vendor Tutorials

The loan approval process starts by receiving a customer request for a loan amount. The risk assessment Web service is invoked to assess the request. If the loan is small and the customer is low risk, the loan is approved. If the customer is high risk, the loan is denied. If the customer needs further review or the loan amount is for $10,000 or more, the request is sent to the approver Web service. The customer receives feedback from the assessor or approver.

Text 15: Process Description 5-1: Active VOS Tutorial.

The process of Vacations Request starts when any employee of the organization submits a vacation request. Once the requirement is registered, the request is received by the immediate supervisor of the employee requesting the vacation. The supervisor must approve or reject the request. If the request is rejected, the application is returned to the applicant/employee who can review the rejection reasons. If the request is approved a

notification is generated to the Human Resources Representative, who must complete the respective management procedures.

Text 16: Process Description 5-2: BizAgi Tutorial 1.

The process of an Office Supply Request starts when any employee of the organization submits an office supply request. Once the requirement is registered, the request is received by the immediate supervisor of the employee requesting the office supplies. The supervisor must approve or ask for a change, or reject the request. If the request is rejected the process will end. If the request is asked to make a change then it is returned to the petitioner/employee who can review the comments for the change request. If the request is approved it will go to the purchase department that will be in charge of making quotations (using a sub-process) and select a vendor. If the vendor is not valid in the system the purchase department will have to choose a different vendor. After a vendor is selected and confirmed, the system will generate and send a purchase order and wait for the product to be delivered and the invoice received. In any case the system will send a notification to let the user know what the result was. In any of the cases, approval, rejection or change required the system will send the user a notification.

Text 17: Process Description 5-3: BizAgi Tutorial 2.

An employee purchases a product or service he requires. For instance, a sales person on a trip rents a car. The employee submits an expense report with a list of items, along with the receipts for each item. A supervisor reviews the expense report and approves or rejects the report. Since the company has expense rules, there are circumstances where the supervisor can accept or reject the report upon first inspection. These rules could be automated, to reduce the workload on the supervisor. If the supervisor rejects the report, the employee, who submitted it, is given a chance to edit it, for example to correct errors or better describe an expense. If the supervisor approves the report, it goes to the treasurer. The treasurer checks that all the receipts have been submitted and match the items on the list. If all is in order, the treasurer accepts the expenses for processing (including, e.g. , payment or refund, and accounting). If receipts are missing or do not match the report, he sends it back to the employee. If a report returns to the employee for corrections, it must again go to a supervisor, even if the supervisor previously approved the report. If the treasurer accepts the expenses for processing, the report moves to an automatic activity that links to a payment system. The process waits for the payment confirmation. After the payment is confirmed, the process ends.

Text 18: Process Description 5-4: Oracle Tutorial.

Δ΄.6 inubit AG

As a basic principle, ACME AG receives invoices on paper or fax. These are received by the Secretariat in the central inbox and forwarded after a short visual inspection to an accounting employee. In ”ACME Financial Accounting”, a software specially developed for

the ACME AG, she identifies the charging suppliers and creates a new instance (invoice). She then checks the invoice items and notes the corresponding cost center at the ACME AG and the related cost center managers for each position on a separate form (”docket”).The docket and the copy of the invoice go to the internal mail together and are sent to the first cost center manager to the list. He reviews the content for accuracy after receiving the copy of the invoice. Should everything be in order, he notes his code one on the docket (”accurate position - AP”) and returns the copy of the invoice to the internal mail. From it, the copy of the invoice is passed on to the next cost center manager, based on the docket, or if all items are marked correct, sent back to accounting. Therefore, the copy of invoice and the docket gradually move through the hands of all cost center managers until all positions are marked as completely accurate. However, if inconsistencies exist, e.g. because the ordered product is not of the expected quantity or quality, the cost center manager rejects the AP with a note and explanatory statement on the docket, and the copy of the invoice is sent back to accounting directly. Based on the statements of the cost center managers, she will proceed with the clarification with the vendor, but, if necessary, she consults the cost center managers by telephone or e-mail again. When all inconsistencies are resolved, the copy of the invoice is sent to the cost center managers again, and the process continues. After all invoice items are AP, the accounting employee forwards the copy of the invoice to the commercial manager. He makes the commercial audit and issues the approval for payment. If the bill amount exceeds EUR 20,000, the Board wants to check it again (4-eyes-principle). The copy of the invoice including the docket moves back to the accounting employee in the appropriate signature file. Should there be a complaint during the commercial audit, it will be resolved by the accounting employee with the supplier. After the commercial audit is successfully completed, the accounting employee gives payment instructions and closes the instance in ”ACME financial accounting”.

Text 19: Process Description 6-1: ACME.

The process starts periodically on the first of each month, when Assembler AG places an order with the supplier in order to request more product parts. a) Assembler AG sends the order to the supplier. b) The supplier processes the order. c) The supplier sends an invoice to Assembler AG. d) Assembler AG receives the invoice.

Text 20: Process Description 6-2: inubit AG Tutorial.

Every time we get a new order from the sales department, first, one of my masters determines the necessary parts and quantities as well as the delivery date. Once that information is present, it has to be entered into our production planning system (PPS). It optimizes our production processes and creates possibly uniform work packages so that the setup times are minimized. Besides, it creates a list of parts to be procured. Unfortunately it is not coupled correctly to our Enterprise Resource Planning system (ERP), so the data must be transferred manually. By the way, that is the second step. Once all the data is present, we need to decide whether any parts are missing and must be procured or if this is not necessary. Once production is scheduled to start, we receive

a notice from the system and an employee takes care of the implementation.Finally, the order will be checked again for its quality.

Text 21: Process Description 6-3: New Order.

The first step is to determine contact details of potential customers. This can be achieved in several ways. Sometimes, we buy details for cold calls, sometimes, our marketing staff participates in exhibitions and sometimes, you just happen to know somebody, who is interested in the product. Then we start calling the customer. That is done by the call center staff. They are determining the contact person and the budget which would be available for the project. Of course, asking the customer whether he is generally interested is also important. If this is not the case, we leave him alone, except if the potential project budget is huge. Then the head of development personally tries to acquire the customer. If the customer is interested in the end, the next step is a detailed online presentation. It is given either by a sales representative or by a pre-sales employee in case of a more technical presentation. Afterwards we are waiting for the customer to come back to us. If we are not contacted within 2 weeks, a sales representative is calling the customer. The last phase is the creation of a quotation.

Text 22: Process Description 6-4: Turbopixel.

Δ΄.7 BPM Practitioners

First, the Manager checks the open leads. Afterwards, he selects the top five ones. He then tells his Sales Assistant to call the contact person of the leads. The Sales Assistant calls each customer. If someone is interested, he sends a note to the Manager. The Manager then processes the lead. Otherwise, he calls the next customer.

Text 23: Process Description 7-1: Calling Leads.

Δ΄.8 BPMN Practical Handbook

The process is triggered by the demand of a functional department to fill a post. The post is advertised, applicants apply, the applications are checked and the post is filled. The process finishes when the post was filled, precisely through the conclusion of a contract of employment.

Text 24: Process Description 8-1: HR Process - Simple.

When a vacancy is reported to me, I create a job description from the information. Sometimes there is still confusion in the message, then I must ask the Department again. I am submitting the job description for consideration and waiting for the approval. But, it can also happen that the department does not approve it, but rejects it, and requests a correction. Then I correct the description and submit it again for consideration. If the description is finally approved, I post the job.

Text 25: Process Description 8-2: HR Process - HR Department.

When I have detected a number of personnel requirements, I report the vacancy to the Personnel Department. Then I wait to get the job description for review before it is advertised. Under certain circumstances, I must ask for corrections again, otherwise I approve the job description. Sometimes it also happens that the colleague from the HR department still has questions about the tasks and requirements before he can describe the job. Then I am available for clarifications, of course.

Text 26: Process Description 8-3: HR Process - Functional Department.

Δ΄.9 BPMN Modeling an Reference Guide

The Customer Service Representative sends a Mortgage offer to the customer and waits for a reply. If the customer calls or writes back declining the mortgage, the case details are updated and the work is then archived prior to cancellation. If the customer sends back the completed offer documents and attaches all prerequisite documents then the case is moved to administration for completion. If all pre-requisite documents are not provided a message is generated to the customer requesting outstanding documents. If no answer is received after 2 weeks, the case details are updated prior to archive and cancellation.

Text 27: Process Description 9-1: Exercise 2.

In November of each year, the Coordination Unit at the Town Planning Authority drafts a schedule of meetings for the next calendar year and adds draft dates to all calendars. The Support Officer then checks the dates and suggests modifications. The Coordination Unit then rechecks all dates and looks for potential conflicts. The final schedule of meeting dates is sent to all the independent Committee Members by email, who then check their diaries and advise the Coordination Unit of any conflicts.

Text 28: Process Description 9-2: Exercise 3a.

Once the dates are finalized (by the Coordination Unit), the Support Officer updates all group calendars and creates meeting folders for each meeting and ensures all appropriate documents are uploaded to system. Committee Members are advised a week before each meeting to read all related documents. The Committee Members hold their meeting, and the Support Office then produces minutes including any Action Points for each Committee Member. Within 5 working days, the Coordination Unit must conduct a QA check on the minutes, which are then sent to all Committee Members. The Support Officer then updates all departmental records.

Text 29: Process Description 9-3: Exercise 3b.

After the Process starts, a Task is performed to locate and distribute any relevant existing designs, both electrical and physical. Next, the design of the electrical and physical systems starts in parallel. Any existing or previous Electrical and Physical

Designs are inputs to both Activities. Development of either design is interrupted by a successful update of the other design. If interrupted, then all current work is stopped and that design must restart. In each department (Electrical Design and Physical Design), any existing designs are reviewed, resulting in an Update Plan for their respective designs (i.e. one in Electrical and another in Physical). Using the Update Plan and the existing Draft of the Electrical/Physical Design, a revised design is created. Once completed the revised design is tested. If the design fails the test, then it is sent back to the first Activity (in the department) to review and create a new Update Plan. If the design passes the test, then it tells the other department that they need to restart their work. When both of the designs have been revised, they are combined and tested. If the combined design fails the test, then they are both sent back to the beginning to initiate another design cycle. If the designs pass the test, then they are deemed complete and are then sent to the manufacturing Process [a separate Process].

Text 30: Process Description 9-4: Exercise 4.

Δ΄.10 Federal Network Agency Enactment

Abbr. Ger.	Amplification Ger.	Amplification Eng.	Abbr. Eng.
L	Letztverbraucher	End consumer	EC
LF	Lieferant	Supplier	SP
MSB	Messstellenbetreiber	Metering point operator	MPO
MSBA	Messstellenbetreiber alt	Metering point operator old	MPOO
MSBN	Messstellenbetreiber neu	Metering point operator new	MPO
MDL	Messdienstleister	Metering service provider	MSP
MDLA	Messdienstleister alt	Metering service provider old	MSPO
MDLN	Messdienstleister neu	Metering service provider new	MSPN
NB	Netzbetreiber	Grid operator	GO
AN	Anschlussnutzer	Power supply user	PU
AG	Angefragter	Inquired person	IP
AF	Anfragender	Inquirer	INQ

Σχήμα Δ΄.1: *List of abbreviations and translations used in the Test Data Set.*

The MPON sents the dismissal to the MPOO. The MPOO reviews the dismissal. The MPOO opposes the dismissal of MPON or the MPOO confirmes the dismissal of the MPON.

Text 31: Process Description 10-1: Process B2.

The MPON reports the meter operation to the GO. The GO examines the application of the MPON. The GO rejects the application of the MPON or the GO confirmes the application of the MPON. The GO informs the MPOO about the registration confirmation of the MPON. The GO informs the MSPO about the registration confirmation of the MPON. The MPON and the MPOO perform the equipment acquisition and/or equipment changes. The MPON

informs the GO about the failure of the entire process or the MPON informs the GO about the successful completion of the entire process. The GO informs the MPON about the failure of the overall transaction by deadline if after a maximum time limit no message of the MPON is present at the GO. If the MPON informs the GO about the failure of the entire process, the GO confirms the failure of the assignment to the MPON. If the MPON informs the GO about the successful completion of the overall process, the GO assigns the MPON. The GO confirms the assignment to the MPON. The GO informs the MPOO about the failure of the assignment of the MPON or the GO informs the MPOO about the assignment of the MPON. The GO informs the MSPO about the failure of the assignment of the MPON or the GO informs the MSPO about the assignment of the MPON. The GO informs the SP about the assignment of the MPON.

Text 32: Process Description 10-2: Process B3.

The MPOO deregisters at the GO. The GO verifies the deregistration. The GO rejects the deregistration of the MPOO or the GO preliminarily confirms the deregistration of the MPOO. The GO prepares the readmission of the measuring point. Optionally, the GO may oblige the MPOO to continue the operations. If the GO binds the MPOO to continue the operation, the MPOO confirms the continuation to the MPOO. The GO performs the equipment acquisition and/or equipment changes. The GO assigns the GO as MPO. The GO informs the MPOO about the end of the assignment of the MPOO and the beginning of the assignment of the GO. The GO informs the MSPO about the assignment of the GO. The GO informs the SP about the assignment of the GO.

Text 33: Process Description 10-3: Process B4.

The MPON notifies the MPOO about equipment change intentions. The MPOO announces self dismounting to the MPON or the MPOO shall notify the MPON about no self-dismounting of the MPOO. The MPON or the MPOO perform the final reading. The MPON or the MPOO dismount the old equipment. The MPON mounts the new device. The MPON reads the meter count from the installed meter. The MPON sents the values of the final reading to the GO. The MPON tells the GO about the device changes, the master data and the meter count at installation. The GO shall notify the MSP about the device changes, the master data, the meter count at dismounting, and the meter count at installation.

Text 34: Process Description 10-4: Process B5.1.

The MPON requests a device takeover bid of the MPOO. The MPOO sends a tender for the equipment takeover to the MPON. The MPON places an order at the MPOO. The MPOO confirms the order of the MPON and sends the master data.

Text 35: Process Description 10-5: Process B5.2.

The MSPN sends a dismissal to the MSPO. The MSPO reviews the dismissal. The MSPO rejects the dismissal of the MSPN or The MSPO confirms the dismissal of the MSPN.

Text 36: Process Description 10-6: Process B6.

The MSPN registers the measurement at the GO. The GO examines the application of the MSPN. The GO rejects the application of the MSPN or the GO confirms the application of the MSPN. The GO assigns the MSPN. The GO informs the MSPO about the assignment of MSPN. The GO informs the MPO about the assignment of the MSPN. The GO informs the SP about the assignment of MSPN.

Text 37: Process Description 10-7: Process B7.

The MSPO deregisters at the GO. The GO verifies the deregistration. The GO rejects the deregistration of the MSPO or the GO preliminarily confirms the deregistration of the MSPO. The GO assigns himself as MSP. The GO informs the MSPO about the end of the assignment and the beginning of the assignment of the GO. The GO informs the MPO about the assignment of the GO. The GO informs the SP about the assignment of the GO.

Text 38: Process Description 10-8: Process B8.

The SP/PU/GO request changes of the MPO or the MPO himself causes a change. The MPO reviews the change request. The MPO rejects the change of the measuring point by the SP/PU/GO or the MPO confirms the request of the SP/PU/GO. The MPO performs the measuring point change. The MPO reports the implementation to the SP/PU/GO or notifies the SP/PU/GO about the failure of the changes.

Text 39: Process Description 10-9: Process C1.

The fault detector reports a failure to the MPO or MPO has a suspicion of their own fault. The MPO shall examine the failure. The MPO rejects the failure of the fault detector or the MPO confirms the failure of the fault detector. If the MPO confirms the failure of the fault detector, he informs the GO and the MSP. The MPO fixes the fault at the measuring device. The MPO shares the results of the repairs carried out with the fault detector. The MPO will inform the GO about the resolution of the interference. The MPO will inform the MSP about the resolution of the interference.

Text 40: Process Description 10-10: Process C2.

The GO requests the measurements of the MSP. The MSP checks the received request. The MSP denies the request of the GO or the MSP performs the measurement. The MSP informs the GO about the failure of the reading or the MSP transmits the measured values to the GO. The GO processes the measured values. The GO sents the changed values to the MSP. The GO transmit the readings to the SP.

Text 41: Process Description 10-11: Process C3.

The EC tells the INQ about the change of his master data. The INQ notifies the IP of the change. The IP checks whether the master data can be changed at the desired time. The IP confirms the changes of the INQ or the IP rejects the changes of the INQ.

Text 42: Process Description 10-12: Process D1.

The INQ transmits the transaction data request to the IP. The IP checks the request of the INQ. The IP answers the question of the INQ depending on the outcome of the examination, i.e. Transmission of data or rejection.

Text 43: Process Description 10-13: Process D2.

If the MPOO sends the bill for the temporary continuation of the metering point operations to the GO, the GO examines the bill. If the MSPO sends the bill for the temporary continuation of the measurement to the GO, the GO examines the bill. If the MSPO sends the bill for additional readings to the GO, the GO examines the bill. If the MPOO sends the bill for the equipment acquisition to the MPON or the GO, the MPON or the GO examines the bill. The GO or the MPON confirms the invoice with payment advice to the MPOO or the MSPO, or the GO or the MPON rejects the invoice of the MPOO or the MSPO.

Text 44: Process Description 10-14: Process D3.

Βιβλιογραφία

[1] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0262018020, 9780262018029.

[2] John McCarthy. *What is artificial intelligence?* 01, 2004.

[3] Domingos P. *A few useful things to know about machine learning*. Commun. ACM 55(10), 78–87, 2012.

[4] Levin B. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, 1993.

[5] Oliver Holschke. *Impact of granularity on adjustment behavior in adaptive reuse of business process models*. Springer,, 2010. In Richard Hull, Jan Mendling, and Stefan Tai, editors, Business Process Management, volume 6336 of Lecture Notes in Computer Science, pages 112–127.

[6] Jakob Freund, Bernd Rcker και Thomas Henninger. *Practical Handbook BPMN*. Hanser, 2010.

[7] Friedrich F., Mendling J. και F. Puhlmann. *Automated Generation of Business Process Models from Natural Language Input*. *School of Business and Economics of the Humboldt-Universität zu Berli*, 2010. http://dx.doi.org/10.1002/andp.19053221004.

[8] Leopold H., van der Aa H. και Reijers H.A. *Identifying Candidate Tasks for Robotic Process Automation in Textual Process Descriptions*. *Gulden J., Reinhartz-Berger I., Schmidt R., Guerreiro S., Guédria W., Bera P. (eds) Enterprise, Business-Process and Information Systems Modeling. BPMDS 2018, EMMSAD 2018. Lecture Notes in Business Information Processing, vol 318. Springer, Cham.*, 2018. https://doi.org/10.1007/978-3-319-91704-7_5.

[9] Dr Nishad Nawaz. *Robotic Process Automation For Recruitment Process*. *International Journal of Advanced Research in Engineering and Technology, 10(2) pp. 608-611*, 2019. https://ssrn.com/abstract=3536295.

[10] Tathagata Chakraborti, Vatche Isahagian, Rania Khalaf, Yasaman Khazaeni, Vinod Muthusamy, Yara Rizk και Merve Unuvar. *From Robotic Process Automation to Intelligent Process Automation*. *Asatiani A. et al. (eds) Business Process*

Management: Blockchain and Robotic Process Automation Forum. BPM 2020. Lecture Notes in Business Information Processing, vol 393. Springer, Cham., 2020. https://doi.org/10.1007/978-3-030-58779-6$_1$5.

[11] Luciano Del Corro και Rainer Gemulla. *ClausIE: clause-based open information extraction Share on. WWW '13: Proceedings of the 22nd international conference on World Wide Web Pages 355-366*, 2013. https://doi.org/10.1145/2488388.2488420.

[12] I.A. Melcuk. *Dependency syntax: theory and practice. State University of New York Press*, 1988.

[13] M.P. Marcus, M.A. Marcinkiewicz και B. Santorini. *Building a Large Annotated Corpus of English: The Penn Treebank. University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-93-87.*, 1993. https://repository.upenn.edu/cis$_r$*eports*/237/.

[14] M.C. De Marneffe, B. MacCartney και C.D. Manning. *Generating typed dependency parses from phrase structure parses. LREC*, 2006.

[15] M.C. De Marneffe, B. MacCartney και C.D. Manning. *Stanford typed dependencies manual. Technical report. Stanford University*, 2008.

[16] M.C. De Marneffe, B. MacCartney και C.D. Manning. *The Stanford typed dependencies representation. Coling 2008: Proceedings of the workshop on Cross-Framework and CrossDomain Parser Evaluation, pages 1-8. Association for Computational Linguistics*, 2008.

[17] N. Ge, J. Hale και E. Charniak. *A statistical approach to anaphora resolution. Proceedings of the Sixth Workshop on Very Large Corpora, pages 161-170*, 1998.

[18] Miller G. και Fellbaum C. *WordNet: An Electronic Lexical Database. MIT Press, Cambridge*, 1998.

[19] H.A. Reijers. *Design and control of workflow processes: business process management for the service industry. Eindhoven University Press,*, 2003.

[20] S.A. White και D. Miers. *BPMN Modeling and Reference Guide: Understanding and Using BPMN. Future Strategies Inc.*, 2008.

[21] Robert Blumberg και Shaku Atre. *The Problem with Unstructured Data. Dm Review*, 2003.

[22] Somayya Madakam, Rajesh M. Holmukhe και Durgesh Kumar Jaiswal. *THE FUTURE DIGITAL WORK FORCE: ROBOTIC PROCESS AUTOMATION (RPA). Journal of Information Systems and Technology Management - Jistem USP*, 2018. DOI: 10.4301/S1807-1775201916001.

[23] Fabian Friedrich1, Jan Mendling2 και Frank Puhlmann1. *Process Model Generation from Natural Language Text. International Conference on Advanced Information Systems Engineering. Springer, Berlin, Heidelberg*, 2011.

[24] Aditya K. Ghose, George Koliadis και Arthur Cheung. *Process discovery from model and text ar om model and text artefacts*. *University of Wollongong*, 2007.

[25] Fabian Friedrich1, Jan Mendling2 και Frank Puhlmann1. *Generating Natural Language Texts from Business Process Models*. *International Conference on Advanced Information Systems Engineering. Springer, Berlin, Heidelberg*, 2012.

[26] Tathagata Chakraborti, Vatche Isahagian, Rania Khalaf, Yasaman Khazaeni, Vinod Muthusamy, Yara Rizk και Merve Unuvar. *From Robotic Process Automation to Intelligent Process Automation*. *IBM Research AI, Cambridge, MA, USA*, 2020. https://doi.org/10.1007/978-3-030-58779-6$_1$5.

[27] Martin Devillers. *Business process modeling as a means to bridge the business-IT divide*. *Radbourd University Nijmegen*, 2011.

[28] Koliadis G., Vranesevic A., Bhuiyan M., Krishna A. και Ghose A. *Combining i* and BPMN for business process model lifecycle management*. *International Conference on Business Process Management. Springer, Berlin, Heidelberg*, 2006.

[29] Vander Aa H., Carmona Vargas J, Leopold H., Mendling J. και Padró L. *Challenges and Opportunities of Applying Natural Language Processing in Business Process Management*. *The 27th International Conference on Computational Linguistics: Proceedings of the Conference: August 20-26, 2018 Santa Fe, New Mexico, USA*, 2018.

[30] Stanley M, Sutton Jr., Avik Sinha και Amit Paradkar. *Text2Test: Automated Inspection of Natural Language Use Cases*. *Third International Conference on Software Testing, Verification and Validation. IEEE*, 2010.

[31] Avik Sinha, Amit Paradkar, Palani Kumanan και Branimir Boguraev. *A Linguistic Analysis Engine for Natural Language Use Case Description and Its Application to Dependability Analysis in Industrial Use Cases*. *2009 IEEE/IFIP International Conference on Dependable Systems Networks. IEEE*, 2009.

[32] Willcocks L. και Lacity M.C. *Service Automation: Robots and the Future of Work. Steve Brookes Publishing, Warwickshire*, 2016.

[33] Tong S. και Koller D. *Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2*, 2001.

[34] Riefer M., Ternis S.F. και Thaler T. *Mining process models from natural language text: A state-of-the-art analysis. Multikonferenz Wirtschaftsinformatik (MKWI-16). Universit at Illmenau , Illmenau, Germany, 9-11 March 2016*, 2016.

[35] B. Levin. *English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago6*, 1993.

[36] A.K. Ghose, G. Koliadis και A. Chueng. *Process discovery from model and text artefacts. Proceedings of the IEEE Congress on Services, pp. 167-174. IEEE Computer Society*, 2007.

[37] *Stanford NLP*. https://nlp.stanford.edu/. Stanford University.

[38] *Stanford Parser online editor*. http://nlp.stanford.edu:8080/parser/index.jsp. Stanford University.

[39] *Machine Learning*. https://en.wikipedia.org/wiki/Machine_learning. Ημερομηνία δημοσίευσης: 21-3-2021.

[40] *Machine Learning Classification - 8 Algorithms for Data Science Aspirants*. https://data-flair.training/blogs/machine-learning-classification-algorithms/.

[41] *Machine Learning Classifiers*. https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623. Ημερομηνία δημοσίευσης: 11-7-2018.

[42] *Classification Algorithms in Machine Learning: How They Work*. https://monkeylearn.com/blog/classification-algorithms/. Ημερομηνία δημοσίευσης: 26-8-2020.

[43] *Fuzzy String Matching in Python*. https://www.datacamp.com/community/tutorials/fuzzy-string-python?utm_source=adwords_ppc&utm_campaignid=898687156&utm_adgroupid=48947256715&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=229765585183&utm_targetid=aud-299261629574:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=9061574&gclid=CjwKCAjwjbCDBhAwEiwAiudBy1qjY0ukhmw1PJ1ONVjDQN--MqHFlkUfoJnt3zD5t6JRKUoXH3OVeRoC7UMQAvD_BwE. Ημερομηνία δημοσίευσης: 6-2-2019.

[44] *Glossary of Computer Related Terms*. https://web.archive.org/web/20200304041552/www.math.utah.edu/~wisnia/glossary.html. Ημερομηνία δημοσίευσης: 6-7-1997.

[45] *Machine Learning — Text Processing*. https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958. Ημερομηνία δημοσίευσης: 13-9-2018.

[46] *Why One-Hot Encode Data in Machine Learning?* https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/. Ημερομηνία δημοσίευσης: 28-7-2017.

[47] *What Are Word Embeddings for Text?* https://machinelearningmastery.com/what-are-word-embeddings/. Ημερομηνία δημοσίευσης: 11-10-2017.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

βλπ	βλέπε
π.χ.	παραδείγματος χάρην
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
ΕΜΠ	Εθνιό Μετσόβειο Πολυτεχνείο
RPA	Robotic Process Automation
ML	Machine Learning
NLP	Natural Language Processing
BPMN	Business Process Modeling Notation
BPM	Business Process Modeling