ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ Μ.Β.Α

Μεταπτυχιακή Διπλωματική Εργασία

The Role of Big Data in drug discovery, development and commercialization: LDA and Sentiment Analysis case study

Συγγραφέας
Μαριάννα Παπαγεωργιάδη
ΑΜ: 19089

Επιβλέπων

Δημήτριος Παπακυριακόπουλος

Αθήνα, Ιούνιος 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ Μ.Β.Α

Μεταπτυχιακή Διπλωματική Εργασία

The Role of Big Data in drug discovery, development and commercialization: LDA and Sentiment Analysis case study

Μέλη Εξεταστικής Επιτροπής συμπεριλαμβανομένου και του Εισηγητή

Η μεταπτυχιακή διπλωματική εργασία εξετάστηκε επιτυχώς από την κάτωθι Εξεταστική Επιτροπή:

Α/α	ΟΝΟΜΑ ΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ/ΙΔΙΟΤΗΤΑ	ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ
	ΔΗΜΗΤΡΗΣ ΠΑΠΑΚΥΡΙΑΚΟΠΟΥΛΟΣ	ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ	Dimitrios Papakyriakopoulos Digitally signed by Dimitrios Papakyriakopoulos Date: 2023.12.07 18:16:34 +02'00'
	ΔΗΜΗΤΡΙΟΣ ΚΑΛΛΙΒΩΚΑΣ	ΛΕΚΤΟΡΑΣ	Dimitrios Kallivokas Digitally signed by Dimitrios Kallivokas Date: 2023.12.09 23:10:42 +02'00'
	ΦΑΙΔΩΝ ΚΟΜΙΣΟΠΟΥΛΟΣ	ΛΕΚΤΟΡΑΣ	Faidon Komisopoulos Digitally signed by Faidon Komisopoulos Date: 2023.12.09 20:04:49 +02'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένος/η Παπαγεωργιάδη Μαριάννα του Γεωργίου, με αριθμό μητρώου 19089 φοιτητής/τρια του Προγράμματος Μεταπτυχιακών Σπουδών Μ.Β.Α. του Τμήματος Διοίκησης Επιχειρήσεων της Σχολής Διοικητικών, Οικονομικών & Κοινωνικών Επιστημών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:

«Είμαι συγγραφέας αυτής της μεταπτυχιακής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο/Η Δηλών/ούσα

Πίνακας περιεχομένων

1 Εισαγωγή

Στην παρούσα εργασία θα εξεταστεί ο ρόλος των Big Data στην ανάπτυξη νέων φαρμάκων, θα αναλυθούν οι πιο σημαντικές αναλυτικές μέθοδοι επεξεργασίας Big Data, οι οποίες ήδη χρησιμοποιούνται για την ανάπτυξη νέων προϊόντων σε άλλους κλάδους εταιριών και τέλος θα μελετηθούν αναλυτικά μέσω διενέργειας πειραμάτων στο πρόγραμμα Rapid Miner η LDA analysis και η Sentiment analysis σχετικά με δύο κατηγορίες φαρμάκων, το birth control και την κατάθλιψη. Ο σκοπός είναι να διαπιστωθεί εάν θα μπορούσε μια εταιρία να εξάγει σαφή και αξιόπιστα συμπεράσματα αναλύοντας συλλογές δεδομένων από σχόλια χρηστών-ασθενών και να τα εκμεταλλευτεί στην ανάπτυξη νέων φαρμάκων.

1.1 Στόχος και κίνητρα

Κίνητρο για την παρούσα εργασία αποτελεί το γεγονός πως η φαρμακοβιομηχανία βρίσκεται ακόμη σε πρώιμο στάδιο σχετικά με την αξιοποίηση των Big Data στην ανάπτυξη νέων φαρμάκων. Ο τομέας των φαρμάκων χαρακτηρίζεται από πολύ αυστηρά πλαίσια και νομοθεσίες και γι' αυτό τον λόγο όλες οι διαδικασίες παραμένουν στα προκαθορισμένα και ελεγμένα πλαίσια. Απαιτείται μεγάλο χρονικό διάστημα έως ότου η φαρμακοβιομηχανία να ενσωματώσει επίσημα και να αξιοποιήσει τα Big Data για την ανάπτυξη νέων προϊόντων. Ωστόσο οι μελλοντικές τάσεις δείχνουν ότι οι φαρμακευτικές εταιρείες πρέπει να αναπτύξουν νέες μεθόδους επεξεργασίας δεδομένων και πληροφοριών για να ανταποκρίνονται ταχύτερα και με μεγαλύτερη ακρίβεια στις μεταβαλλόμενες αγορές και στις ανάγκες των ασθενών. Με την εκμετάλλευση των Big Data, τα οποία προσπαθούν να ενσωματώσουν δεδομένα από διαφορετικές πηγές δεδομένων, η βιομηχανία έχει εντοπίσει ένα νέο σύνορο που θα μπορούσε να παρέχει τις απαραίτητες γνώσεις για να βελτιώσει την ανάπτυξη νέων προϊόντων. Στόχος μας λοιπόν είναι να αποδείξουμε πως υπάρχουν συλλογές δεδομένων που περιγράφουν με σαφήνεια την εμπειρία του χρήστη-ασθενή σχετικά με την λήψη ενός φαρμακευτικού σκευάσματος και πως αυτά τα δεδομένα μπορούν να αξιοποιηθούν από τις φαρμακευτικές εταιρίες, ώστε να παράγονται πιο γρήγορα και οικονομικά φάρμακα με λιγότερες παρενέργειες, τα οποία θα ικανοποιούν τις ανάγκες των ασθενών σε μεγαλύτερο βαθμό.

1.2 Δομή της εργασίας

Στην παρούσα εργασία περιγράφεται αρχικά ο στόχος και τα κίνητρα για τα οποία επιλέξαμε να μελετήσουμε τον ρόλο των Big Data στην ανακάλυψη και ανάπτυξη φαρμάκων με την χρήση των μεθόδων LDA and Sentiment Analysis. Έπειτα ακολουθεί η βιβλιογραφική ανασκόπηση, όπου περιγράφεται η ανάπτυξη νέων προϊόντων και ο ρόλος της τεχνολογίας στον κύκλο ζωής του φαρμάκου. Πιο συγκεκριμένα παρατίθεται η θεωρία που αφορά τα Big Data και αναλύονται οι μέθοδοι Sentiment Analysis και Topic Modelling. Στην συνέχεια περιγράφεται η μεθοδολογία η οποία θα ακολουθηθεί στο εμπειρικό κομμάτι και η διαθεσιμότητα δεδομένων που χρησιμοποιήθηκαν στα πειράματα. Έπειτα ακολουθεί το εμπειρικό κομμάτι, στο οποίο χρησιμοποιήθηκε το πρόγραμμα Rapid Miner. Στα πειράματα που πραγματοποιήθηκαν, χρησιμοποιήθηκαν και αναλύθηκαν τα σχόλια χρηστών-ασθενών σχετικά με δύο κατηγορίες φαρμακευτικών σκευασμάτων: αντισυλληπτικά φάρμακα (birth control medicines) και αντικαταθλιπτικά φάρμακα (depression medicines). Τέλος παρατίθενται τα συμπεράσματα και ο σχολιασμός και η βιβλιογραφία.

2 Βιβλιογραφική ανασκόπηση

2.1 Ανάπτυξη νέων προϊόντων (NPD)

Στον επιχειρηματικό τομέα η ανάπτυξη νέων προϊόντων αποτελείται από την πλήρη διαδικασία εισαγωγής ενός νέου προϊόντος στην αγορά, είτε αυτό αφορά την ανανέωση ενός προ υπάρχοντος προϊόντος, είτε την εισαγωγή ενός προϊόντος σε μία νέα αγορά. Αυτό περιλαμβάνει την κατανόηση των αναγκών της αγοράς, τη σύλληψη της ιδέας του προϊόντος, τον σχεδιασμό του, την κυκλοφορία του στην αγορά και τέλος την συλλογή αξιολογήσεων από τους καταναλωτές. Ο σχεδιασμός του προϊόντος πάντα πραγματοποιείτε σε συνδυασμό με διάφορα επιχειρηματικά κριτήρια, καθώς αποτελεί την μετατροπή μιας ευκαιρίας που παρουσιάζεται στην αγορά, σε ένα έτοιμο προϊόν προς πώληση. Το προϊόν αυτό μπορεί να είναι υλικό ή άυλο, όπως μια υπηρεσία.
Για την ανάπτυξη νέων προϊόντων απαιτείται η επίγνωση των αναγκών και των επιθυμιών των πελατών, της δύναμης του ανταγωνισμού καθώς και της φύσης της αγοράς. Το κόστος, ο χρόνος και η ποιότητα είναι οι τρεις παράγοντες με βάση τους οποίους οι εταιρίες αναπτύσσουν συνεχείς πρακτικές ώστε να αυξήσουν την ικανοποίηση των πελατών και κατά συνέπεια το μερίδιο αγοράς τους. Η χρήση των βέλτιστων πρακτικών και η εξάλειψη των εμποδίων είναι οι βασικές προκλήσεις στην διαχείριση της ανάπτυξης νέων προϊόντων (Gurbuz, 2018).

Διαδικασία ανάπτυξης νέων προιόντων-8 στάδια
Η διαδικασία της ανάπτυξης νέων προϊόντων μπορεί να διαφέρει ανάλογα τον τύπο της βιομηχανίας, μπορεί όμως να χωριστεί σε οκτώ βασικά στάδια: δημιουργία της ιδέας, έλεγχος ιδεών, ανάπτυξη και δοκιμή ιδεών, ανάπτυξη στρατηγικής μάρκετινγκ, επιχειρηματική ανάλυση, ανάπτυξη προϊόντων, δοκιμή μάρκετινγκ και εμπορευματοποίηση.
Σε κάθε εταιρία η ανάπτυξη νέων προϊόντων, δεν είναι ποτέ ρόλος ενός συγκεκριμένου τμήματος, αλλά συνεργασία πολλών διαφορετικών τμημάτων όπως των τμημάτων σχεδίασης, μηχανολογίας, παραγωγής, μάρκετινγκ και πολλά άλλα (Gurbuz, 2018).
Πιο συγκεκριμένα θα αναφερθούν παρακάτω τα οκτώ στάδια ανάπτυξης νέων προϊόντων:

- Δημιουργία της ιδέας

Σε αυτό το στάδιο επεξεργάζονται πολλές διαφορετικές και μοναδικές ιδέες, οι οποίες μπορεί να προέρχονται είτε από εσωτερικές πηγές, όπως οι ομάδες έρευνας και ανάπτυξης της εταιρίας, είτε από εξωτερικές πηγές, όπως οι τάσεις της αγοράς, οι επιθυμίες των πελατών, οι καινοτομίες των ανταγωνιστών κ.α.

- Έλεγχος ιδεών

Σε αυτό το στάδιο ο μεγάλος όγκος ιδεών του προηγούμενου σταδίου φιλτράρεται με σκοπό να αποκλειστούν οι πιο επίφοβες ιδέες και να

παραμείνουν όσες συμπορεύονται με τις αξίες αλλά και τους οικονομικούς στόχους της εταιρίας.

- Ανάπτυξη και δοκιμή ιδεών

Στο τρίτο αυτό στάδιο, οι καλύτερες ιδέες θα πρέπει να εξελιχθούν σε λεπτομερείς ιδέες προϊόντων εκφρασμένες σε καταναλωτικούς όρους. Η ιδέα θα πρέπει να εκφράζει με σαφήνεια με ποιον τρόπο θα παρουσιαστεί στο καταναλωτικό κοινό και ποιο θα είναι το πιθανό καταναλωτικό κοινό στόχος. Έπειτα η ιδέα θα πρέπει να παρουσιαστεί στους καταναλωτές και να αναλυθεί η ανταπόκριση τους.

- Ανάπτυξη στρατηγικής μάρκετινγκ

Στο στάδιο αυτό η εταιρία θα προσπαθήσει να καταλήξει στην κατάλληλη στρατηγική για την εισαγωγή του νέου προϊόντος στην αγορά, την σωστή κοστολόγηση, τα κανάλια διανομής και τη διαφήμιση.

- Επιχειρηματική ανάλυση

Στη συνέχεια η ιδέα του προϊόντος θα υποβληθεί σε ενδελεχή επιχειρηματική ανάλυση, με σκοπό να προβλεφθούν τα έσοδα, να εκτιμηθεί ο κίνδυνος και το κατά πόσο η παραγωγή του προϊόντος συμφέρει οικονομικά την εταιρία. Εφόσον λοιπόν πληρούνται οι στόχοι της εταιρίας το προϊόν μπορεί να προχωρήσει στο επόμενο στάδιο.

- Ανάπτυξη προϊόντων

Στο στάδιο αυτό το τμήμα έρευνας και ανάπτυξης αφότου πάρει έγκριση από την διοίκηση της εταιρίας ξεκινά να εργάζεται πάνω ανάπτυξη του προϊόντος έως ότου σε ένα πλήρως λειτουργικό πρωτότυπο προϊόν. Το στάδιο αυτό μπορεί να διαρκέσει πολλούς μήνες έως και χρόνια.

- Δοκιμή μάρκετινγκ

Αφού ολοκληρωθεί η ανάπτυξη του προϊόντος, γίνεται δοκιμή του προϊόντος και του προτεινόμενου πλάνου μάρκετινγκ σε ρεαλιστικές συνθήκες αγοράς. Αυτό το στάδιο παρέχει μια εικόνα για τον τρόπο με τον οποίο το προϊόν θα εισαχθεί στην αγορά, πως θα παραχθεί, θα συσκευαστεί, θα διαφημιστεί, θα διανεμηθεί και θα πωληθεί στους πελάτες. Στόχος αυτού του σταδίου είναι να γίνουν βελτιστοποιήσεις εάν και εφόσον απαιτούνται.

- Εμπορευματοποίηση

Στο τελικό στάδιο της εμπορευματοποίησης, βασισμένη στην δοκιμή μάρκετινγκ που προηγήθηκε η εταιρία αποφασίζει να προχωρήσει στην κυκλοφορία του προϊόντος στην αγορά.

2.1.1 *Μηχανισμοί Ανάπτυξης νέων προϊόντων*

2.1.1.1 Reverse engineering

Το reverse engineering σαν ορισμός είναι η αποσυναρμολόγησης ενός αντικειμένου προκειμένου να δούμε πώς λειτουργεί. Γίνεται κυρίως για την ανάλυση και την απόκτηση γνώσεων σχετικά με τον τρόπο με τον οποίο λειτουργεί κάτι, αλλά συχνά χρησιμοποιείται για την αντιγραφή ή τη βελτίωση του αντικειμένου. Πολλά πράγματα μπορούν να δημιουργηθούν αντίστροφα, όπως λογισμικό, φυσικές μηχανές, στρατιωτική τεχνολογία και ακόμη και βιολογικές λειτουργίες που σχετίζονται με τον τρόπο λειτουργίας των γονιδίων (Saiga, 2021).

Ανάλογα με την τεχνολογία, η γνώση που αποκτάται από το reverse engineering μπορεί να χρησιμοποιηθεί για να επαναχρησιμοποιηθούν απαρχαιωμένα αντικείμενα, να γίνει ανάλυση ασφάλειας, να αποκτηθεί ανταγωνιστικό πλεονέκτημα ή απλώς να διδάξει κάποιον το πώς λειτουργεί κάτι. Ανεξάρτητα από το πώς χρησιμοποιείται η γνώση ή με τι σχετίζεται, η αντίστροφη μηχανική είναι η διαδικασία απόκτησης αυτής της γνώσης από ένα ολοκληρωμένο αντικείμενο. Συχνά ο στόχος του λογισμικού ή του υλικού αντίστροφης μηχανικής είναι να βρει έναν τρόπο να δημιουργήσει ένα παρόμοιο προϊόν πιο φθηνά ή επειδή το αρχικό προϊόν δεν είναι πλέον διαθέσιμο (Saiga, 2021).
Η διαδικασία αντίστροφης μηχανικής είναι συγκεκριμένη για το αντικείμενο για το οποίο εκτελείται. Ωστόσο, ανεξάρτητα από το πλαίσιο, υπάρχουν τρία γενικά βήματα κοινά για όλες τις προσπάθειες αντίστροφης μηχανικής.

Τα βήματα αυτά είναι τα παρακάτω:

- Εξαγωγή πληροφοριών
  Το αντικείμενο που αναστράφηκε μελετάται, εξάγονται πληροφορίες σχετικά με τη σχεδίασή του και εξετάζονται αυτές οι πληροφορίες για να προσδιοριστεί πώς τα κομμάτια ταιριάζουν μεταξύ τους. Στο reverse engineering, αυτό μπορεί να απαιτεί τη συλλογή πηγαίου κώδικα και σχετικών εγγράφων σχεδιασμού για μελέτη. Μπορεί επίσης να περιλαμβάνει τη χρήση εργαλείων, όπως αποσυναρμολογητή για να χωρίσει το πρόγραμμα στα συστατικά μέρη του.
- Modeling
  Οι πληροφορίες που συλλέγονται αφαιρούνται σε ένα εννοιολογικό μοντέλο, με κάθε κομμάτι του μοντέλου να εξηγεί τη λειτουργία του στη συνολική δομή. Ο σκοπός αυτού του βήματος είναι να ληφθούν συγκεκριμένες πληροφορίες για το πρωτότυπο και να αφαιρεθούν σε ένα γενικό μοντέλο που μπορεί να χρησιμοποιηθεί για να καθοδηγήσει το σχεδιασμό νέων αντικειμένων ή συστημάτων. Στο reverse engineering, αυτό μπορεί να λάβει τη μορφή ενός διαγράμματος ροής δεδομένων ή ενός διαγράμματος δομής.
- Ανασκόπηση
  Αυτό περιλαμβάνει την αναθεώρηση του μοντέλου και τη δοκιμή του σε διάφορα σενάρια για να διασφαλιστεί ότι είναι μια ρεαλιστική αφαίρεση του αρχικού αντικειμένου ή συστήματος. Στη μηχανική λογισμικού αυτό μπορεί να λάβει τη μορφή δοκιμής λογισμικού. Μόλις δοκιμαστεί, το μοντέλο μπορεί να εφαρμοστεί για να ανασχεδιάσει το αρχικό αντικείμενο.

2.1.1.2 Concurrent engineering

Το concurrent engineering σαν ορισμός είναι η συστηματική μέθοδος σχεδιασμού και ανάπτυξης προϊόντων όπου διαφορετικές δραστηριότητες

εκτελούνται ταυτόχρονα. Αυτό είναι επίσης γνωστό ως ταυτόχρονη μηχανική. Εκτελώντας διαφορετικές εργασίες ταυτόχρονα, η ταυτόχρονη μηχανική μειώνει τον χρόνο παραγωγής οδηγώντας σε μειωμένο κόστος. Το Concurrent Engineering είναι μια μακροπρόθεσμη επιχειρηματική στρατηγική που παρέχει μακροπρόθεσμα οφέλη σε οποιαδήποτε επιχείρηση ή διαδικασία παραγωγής. Σε αυτή τη μεθοδολογία, πολλές ομάδες μέσα σε έναν οργανισμό εργάζονται ταυτόχρονα για την ανάπτυξη προϊόντων και υπηρεσιών. Ο σχεδιασμός του προϊόντος, η ποιότητα, το μοναδιαίο κόστος και ο χρόνος κατασκευής είναι οι πιο σημαντικές παράμετροι που επηρεάζουν την κερδοφορία ενός οργανισμού. Η ταυτόχρονη μηχανική βοηθά στη βελτίωση αυτών των παραγόντων και παρέχει στις εταιρείες ένα εξαιρετικά ανταγωνιστικό πλεονέκτημα (Ali, 2023).

Τα βασικά πλεονεκτήματα του concurrent engineering είναι τα παρακάτω:

- Μειωμένος χρόνος παραγωγής
  Καθώς όλες οι απαιτούμενες δραστηριότητες εκτελούνται ταυτόχρονα, ο χρόνος ρελαντί για κάθε δραστηριότητα μειώνεται. Έτσι ο πραγματικός χρόνος που απαιτείται για οποιοδήποτε προϊόν μειώνεται και το προϊόν μπορεί εύκολα να εισέλθει στην αγορά σε λιγότερο χρόνο και με μειωμένο κόστος. Το Concurrent Engineering λειτουργεί με βάση την αρχή των παράλληλων κύκλων ζωής του έργου που επιταχύνουν την ανάπτυξη προϊόντων.
- Ιδιαίτερα καινοτόμες λύσεις
  Καθώς πολλές από τις πτυχές ανάπτυξης και τις δεξιότητες των εργαζομένων αλληλεπικαλύπτονται, εξαιρετικές καινοτόμες λύσεις επιτυγχάνονται με πολύτιμες πληροφορίες από όλα τα τμήματα. Ο καταιγισμός ιδεών των προβλημάτων μεταξύ όλων των κλάδων αποφεύγει μεγάλα λάθη κατά τη διάρκεια της ίδιας της φάσης του σχεδιασμού, γεγονός που με τη σειρά του εξοικονομεί χρόνο και χρήμα.
- Βελτιωμένη παραγωγικότητα και ανταγωνιστικό πλεονέκτημα
  Καθώς τα πιθανά προβλήματα διορθώνονται εύκολα από διάφορα τμήματα και μειώνεται ο χρόνος παραγωγής, όλα αυτά παρέχουν βελτιωμένη παραγωγικότητα και υψηλά ανταγωνιστικό πλεονέκτημα έναντι των ανταγωνιστών.

Υπάρχουν τέσσερα βασικά στοιχεία στην αρχή της ταυτόχρονης μηχανικής.

- Διαλειτουργικές ομάδες
  Οι ομάδες πολλαπλών λειτουργιών σχηματίζονται από άτομα από διαφορετικούς τομείς εργασίας που σχετίζονται με τη συγκεκριμένη διαδικασία ή την ανάπτυξη προϊόντος.
- Ταυτόχρονη υλοποίηση προϊόντος
  Η εκτέλεση πολλών εργασιών ταυτόχρονα είναι η βάση της ταυτόχρονης μηχανικής.
- Αυξητική κοινή χρήση πληροφοριών
  Οι πληροφορίες πρέπει να κοινοποιούνται αμέσως όταν είναι διαθέσιμες για να επιτύχετε στην ταυτόχρονη μηχανική. Ακόμη και οι

πληροφορίες μπορούν να κοινοποιηθούν με τη μορφή εκ των προτέρων πληροφοριών εάν η πραγματική διαδικασία έγκρισης πληροφοριών απαιτεί χρόνο.

- Ολοκληρωμένη διαχείριση έργου
  Διασφαλίζει την ευθύνη του έργου προς τους βασικούς επαγγελματίες.

Η αρχή της ταυτόχρονης μηχανικής χρησιμοποιεί το κατάλληλο ανθρώπινο δυναμικό την κατάλληλη στιγμή για να επιταχύνει την ανάπτυξη του προϊόντος, διατηρώντας παράλληλα την επανεπεξεργασία στο ελάχιστο. Η αρχή λειτουργίας της παράλληλης μηχανικής βασίζεται στους ακόλουθους παράγοντες:

- Ομαδική εργασία: Αυτή είναι η βασική προϋπόθεση για την ταυτόχρονη φιλοσοφία της μηχανικής. Η συνεργασία, η συνεργασία και οι διαπροσωπικές σχέσεις είναι αναπόσπαστα μέρη της ταυτόχρονης μηχανικής.
- Πολυεπιστημονική ομάδα: Η διεπιστημονική ομάδα για την ανάπτυξη προϊόντων/υπηρεσιών/διαδικασιών που περιλαμβάνει ειδικούς από κάθε κλάδο είναι σημαντική για την επιτυχία των ταυτόχρονων αρχών μηχανικής.
- Αποτελεσματική Επικοινωνία: Για να έχετε το μέγιστο όφελος από την ταυτόχρονη στρατηγική μηχανικής, η αποτελεσματική επικοινωνία είναι απαραίτητη προϋπόθεση. Η γρήγορη ανταλλαγή πληροφοριών μεταξύ μελών, προμηθευτών, πελατών και κατασκευαστών είναι σημαντική. Πρέπει να διασφαλιστεί ότι όλα τα μέλη γνωρίζουν τι κάνουν τα άλλα μέλη.
- Υποστήριξη διαχείρισης: Η σωστή υποστήριξη διαχείρισης βοηθά στην εφαρμογή των ταυτόχρονων αρχών μηχανικής.
- Συμμετοχή Πελατών και Προμηθευτών: Η επιτυχία της ανάπτυξης προϊόντων στην ταυτόχρονη μηχανική εξαρτάται από τη σωστή ενοποίηση μεταξύ του πελάτη, των προμηθευτών και του κατασκευαστή. Αυτό μειώνει το σφάλμα σχεδιασμού και επαναλειτουργεί σημαντικά.

Οι βασικές προκλήσεις που προκύπτουν στην ταυτόχρονη φιλοσοφία της μηχανικής είναι η εξάρτηση από την αποτελεσματική επικοινωνία μεταξύ των μελών της ομάδας, η υλοποίηση πρώιμων μελετών σχεδιασμού. συμβατότητα λογισμικού, η αποτελεσματική ανταλλαγή μοντέλων υπολογιστή/πληροφοριών σχεδιασμού, η αποτελεσματική οργάνωση και διαχείριση των ομάδων έργου και η εκπαίδευση ατόμων για το πώς να εκτελούν αποτελεσματικά ταυτόχρονο σχεδιασμό (Ali, 2023).

2.1.1.3 Crowdsourcing

Την τελευταία δεκαετία έχει σημειωθεί σημαντική αύξηση στην αξιοποίηση crowdsourcing και open innovation προσεγγίσεων, τα οποία εκμεταλλεύονται «επιστήμονες πολίτες» για να πραγματοποιήσουν μία νέα επιστημονική έρευνα. Οι προσεγγίσεις αυτές έχουν υιοθετηθεί από μεγάλες

φαρμακευτικές εταιρείες στον τομέα της ανακάλυψης φαρμάκων (Thompson, 2020).
Ο μηχανισμός της ανακάλυψης, ανάπτυξης, διανομής και εμπορίας ενός νέου φαρμάκου είναι ένα πολύπλοκο σύστημα. Τα τελευταία χρόνια σημειώθηκε αριθμός ρεκόρ νέων εγκρίσεων φαρμάκων από τις ρυθμιστικές αρχές των ΗΠΑ. Είναι πολύ νωρίς και ο τομέας ανάπτυξης νέων φαρμάκων είναι πολύ περίπλοκος για να διασαφηνιστούν με σιγουριά οι παράγοντες που συνέβαλαν σε αυτή τη σημαντική αύξηση, ωστόσο η αξιοποίηση της προσέγγισης του crowdsourcing φαίνεται να έχει διαδραματίσει σημαντικό ρόλο. Θεωρείται πολύ πιθανόν στο άμεσο μέλλον ένα μεγάλο μέρος της ανάπτυξης νέων φαρμάκων να σχετίζεται με crowdsource, καθώς το συγκεκριμένο μοντέλο είναι πολύ αποτελεσματικό στην εξερεύνηση του χώρου των ασθενειών και τον εντοπισμό ευκαιριών.
Το crowdsourcing σαν όρος διατυπώθηκε για πρώτη φορά σε ένα τεύχος του 2006 του δημοφιλούς τεχνολογικού περιοδικού Wired, όπου περιγράφηκε ως ως μια διαδικτυακή προσέγγιση, η οποία εκμεταλλεύεται την ικανότητα ατόμων εκτός ενός οργανισμού. Επίσης στενά συνδεδεμένες έννοιες με το crowdsourcing είναι οι πρακτικές της ανοιχτής καινοτομίας και της επιστήμης των πολιτών. Μια επίσημη περιγραφή της ανοιχτής καινοτομίας έγινε από τον Chesbrough το 2003 και περιγράφει ότι οι επιχειρήσεις πρέπει να χρησιμοποιούν εξωτερικές ιδέες καθώς και εσωτερικές ιδέες, και εσωτερικές και εξωτερικές οδούς προς την αγορά, καθώς οι επιχειρήσεις αναπτύσσουν την τεχνολογία τους. Η επιστήμη των πολιτών, η συμμετοχή δηλαδή του ευρύτερου κοινού στην έρευνα, είναι ένα συγκεκριμένο παράδειγμα crowdsourcing, που επικεντρώνεται αποκλειστικά στις επιστήμες.
Τον Ιανουάριο του 2017, ο Αμερικανικός Νόμος περί Καινοτομίας και Ανταγωνιστικότητας (AICA) οριστικοποίησε τον νόμο για το Crowdsourcing and Citizen Science, παρέχοντας την δυνατότητα στις επιχειρήσεις να χρησιμοποιούν το crowdsourcing, και πιο συγκεκριμένα την επιστήμη των πολιτών, για να προωθήσουν τις δράσεις των επιχειρήσεων και να ενισχύσουν την συμμετοχή των πολιτών (Thompson, 2020). Συγκεκριμένα, ο νόμος Crowdsourcing and Citizen Science Act ορίζει το crowdsourcing ως μια μέθοδο για την απόκτηση απαραίτητων υπηρεσιών, ιδεών ή περιεχομένου μέσω της προσέλκυσης εθελοντικών συνεισφορών από μια ομάδα ατόμων ή οργανισμών, ιδίως από μια διαδικτυακή κοινότητα.
Επιπλέον, το Government Accountability Office (GAO) ορίζει την ανοιχτή καινοτομία ως δραστηριότητες και τεχνολογίες για την αξιοποίηση των ιδεών, της τεχνογνωσίας και των πόρων εκείνων που βρίσκονται εκτός ενός οργανισμού για την αντιμετώπιση ενός ζητήματος ή την επίτευξη συγκεκριμένων στόχων (Thompson, 2020).
Το γεγονός ότι αυτά τα εργαλεία έχουν εγκριθεί επίσημα από την κυβέρνηση των ΗΠΑ υπογραμμίζει τόσο τη χρησιμότητά τους όσο και τον αυξανόμενο ρόλο τους στην πρακτική της «καινοτομίας» και στη διεύρυνση της συμμετοχής του κοινού στις επιχειρηματικές διαδικασίες. Υπάρχουν πολλές προφανείς ομοιότητες μεταξύ των ορισμών του crowdsourcing και της ανοιχτής καινοτομίας. Και οι δύο αντιπροσωπεύουν δομημένους τρόπους για την εστιασμένη δέσμευση να συμβεί πέρα από τα οργανωτικά όρια.
Ένας οργανισμός ορίζεται ως ένα σύνολο ημι-αυτόνομων «Actors», που ασχολούνται με ένα πρότζεκτ, το οποίο έχει ως αποτέλεσμα την παροχή μιας υπηρεσίας ή την παράδοση ενός προϊόντος στους πελάτες. Οι «actors» οι

οποίοι δεν αποτελούν επίσημα μέρος του Οργανισμού και δεν έχουν επίσημο ρόλο σε εργασία που σχετίζεται άμεσα με την παράδοση της υπηρεσίας ή των προϊόντων που καταναλώνονται από τους πελάτες του Οργανισμού, ορίζονται ως «Άλλοι».
Η ανοιχτή καινοτομία είναι η ευρύτερη, η πιο περιεκτική προσέγγιση για τη συμμετοχή ενός «Άλλου» και θα μπορούσε να χρησιμοποιηθεί για να παρέχει πιο εύκολα ένα πλαίσιο για την ευρύτερη στρατηγική πρόθεση ενός οργανισμού.
Υπάρχουν αρκετές πλατφόρμες που χρησιμοποιούνται συνήθως από οργανισμούς για δραστηριότητες crowdsourcing και ανοιχτής καινοτομίας. Οι τρεις μεγαλύτερες είναι οι παρακάτω: Kaggle, InnoCentive και DREAM (Thompson, 2020).

- Kaggle

Το Kaggle είναι μια διαδικτυακή κοινότητα που επικεντρώνεται στη συμμετοχή σε διαγωνισμούς μηχανικής μάθησης. Ο ιστότοπος χρησιμοποιεί πολλές πτυχές του gamification για να επιτρέψει και να επιδείξει την κυριαρχία της πρακτικής της επιστήμης δεδομένων. Το έτος 2019 σημειώθηκε μεγάλος αριθμός διαγωνισμών, που αφορούσαν κυρίως την ανίχνευση ασθενειών και την ανάλυση εικόνας. Μόνο λίγες φαρμακευτικές εταιρείες (Boehringer Ingelheim, Merck, Pfizer και Genentech) έχουν χρησιμοποιήσει ανοιχτά την πλατφόρμα Kaggle για δραστηριότητες crowdsourcing.
Υπάρχει ένα σταθερά υψηλό επίπεδο δέσμευσης μεταξύ της κοινότητας Kaggle και διαγωνισμών που σχετίζονται με την υγειονομική περίθαλψη, με πολλούς διαγωνισμούς να έχουν >1000 ενεργά συμμετέχουσες ομάδες. Δεδομένου του γενικού επιπέδου ενθουσιασμού για τη μηχανική μάθηση, την επιστήμη δεδομένων, την τεχνητή νοημοσύνη και την ικανότητα να μετασχηματίζονται εύκολα πολλά προβλήματα της υγειονομικής περίθαλψης σε μια μορφή που μπορεί να αναλυθεί από επιστήμονες δεδομένων, η συμμετοχή σε αυτούς τους αγώνες είναι υψηλή (Thompson, 2020).

- InnoCentive

Το InnoCentive είναι μια διαδικτυακή κοινότητα, που ιδρύθηκε το 2001 ως spinout από το Internet Incubator της Eli Lilly and Co. Είναι μια πλατφόρμα όπου δημοσιεύονται οι προκλήσεις και οι λύτες υποβάλλουν λύσεις.

- Dream

Σε αντίθεση με τις πλατφόρμες Kaggle και InnoCentive Crowdsourcing που φιλοξενούν προκλήσεις σε ένα ευρύ φάσμα βιομηχανιών και επιστημών, οι προκλήσεις του DREAM επικεντρώνονται στη βιολογία και την ιατρική. Αυτές οι προκλήσεις ξεκίνησαν το 2006 και υποστηρίζονται από την Sage Bionetworks. Οι προκλήσεις του DREAM οργανώνονται ως επί το πλείστων σε ακαδημαϊκό πλαίσιο και από μη κερδοσκοπικούς οργανισμούς και τα κίνητρα είναι ως επί το πλείστων μη χρηματικά με τη μορφή της συγγραφής λογοτεχνικών δημοσιεύσεων και παρακολούθησης συνεδρίων. Η AstraZeneca φαίνεται να είναι ο μόνος κερδοσκοπικός φαρμακευτικός οργανισμός που έχει χρησιμοποιήσει μια πρόκληση DREAM, υπογραμμίζοντας

περαιτέρω τη δέσμευση αυτής της εταιρείας για δέσμευση με εξωτερικούς «Άλλους» (Thompson, 2020).

2.1.2 Ο κύκλος ζωής του φαρμάκου

Οι φαρμακευτικές εταιρίες έρχονται αντιμέτωπες με πολύ μεγάλες εξωτερικές πιέσεις, όπως η παγκοσμιοποιημένη αγορά σε εξαιρετικά ανταγωνιστικό περιβάλλον, οι γρήγορες τεχνολογικές αλλαγές και οι σύντομοι κύκλοι ζωής των προϊόντων. Γι' αυτόν το λόγο η ανάπτυξη νέων προϊόντων είναι στρατηγικής σημασίας.
Οι δαπάνες μιας φαρμακευτικής εταιρίας για έρευνα και ανάπτυξη είναι κατά πέντε φορές μεγαλύτερη συγκριτικά με εταιρίες άλλων κλάδων.
Ωστόσο η έρευνα και ανάπτυξη νέων φαρμάκων έχει υψηλό κόστος και χαμηλή παραγωγικότητα. Επίσης οι αυστηρές ρυθμίσεις, οι χαμηλές πιθανότητες τεχνικής επιτυχίας, η αβέβαιη αγορά και το περιορισμένο εξειδικευμένο ανθρώπινο δυναμικό οδηγούν τις φαρμακευτικές εταιρίες πολύ μεγάλες προκλήσεις στην ανάπτυξη νέων προϊόντων.
Τα έσοδα από πολλά φάρμακα που κυκλοφορούν στο εμπόριο δεν υπερβαίνουν το μέσο κόστος έρευνας και ανάπτυξης. Γι' αυτό το λόγο η ανάπτυξη νέων προϊόντων απαιτεί πολύ μεγάλη προσοχή για την επίτευξη αποδεκτού επιπέδου οικονομικής απόδοσης (Yousefi, 2017).

Τα βασικά βήματα για την ανάπτυξης ενός νέου φαρμάκου είναι τα ακόλουθα:

- Προ κλινική έρευνα και ανακάλυψη καινοτόμων φαρμάκων
- Κλινικές δοκιμές συνταγογραφούμενων φαρμάκων
- Προετοιμασία και υποβολή αιτήσεων για έγκριση από τον Οργανισμό Τροφίμων και Φαρμάκων ή FDA
- Σχεδιασμός παραγωγικών διαδικασιών για το νέο προϊόν
- Κλινική δοκιμή ενός νέου φαρμάκου έναντι ενός υπάρχοντος φαρμάκου για να δείξει τελικά τα ανώτερα οφέλη του νέου προϊόντος
- Πρόσθετες κλινικές δοκιμές μετά την κυκλοφορία ενός νέου φαρμάκου στην αγορά, για την παρακολούθηση της ασφάλειας και τον εντοπισμό πρόσθετων παρενεργειών που μπορεί να μην έχουν παρατηρηθεί σε προηγούμενες δοκιμές κατά την ανάπτυξη
- Επεκτάσεις γραμμής ή καινοτομίες και βελτιώσεις για υπάρχοντα συνταγογραφούμενα φάρμακα, που περιλαμβάνουν την ανάπτυξη νέων δόσεων και συστημάτων χορήγησης και δοκιμές για πρόσθετες πιθανές χρήσεις

Μέχρι στιγμής η αξιοποίηση των Big Data στην ανάπτυξη νέων φαρμάκων δεν είναι διαδεδομένη.

2.1.2.1 Πρωτότυπα φάρμακα

Οι φαρμακευτικές εταιρίες, οι οποίες ασχολούνται με πρωτότυπα φάρμακα, έχουν αρκετές βασικές διαφορές στην ανάπτυξη νέων προϊόντων, όπως οι χρόνοι και το κόστος που απαιτούνται για την ανάπτυξη. Λόγω των

χρονοβόρων κλινικών δοκιμών και των μεγάλων περιόδων έγκρισης είναι λογικό να απαιτείται πολύ περισσότερος χρόνος μέχρι ένα προϊόν να πιστοποιηθεί ότι πληροί όλες τις απαιτούμενες προϋποθέσεις ώστε να κυκλοφορήσει στην αγορά, συγκριτικά με άλλου τύπου εταιρίες. Επίσης το κόστος κεφαλαίου και το κόστος ανάπτυξης νέων προϊόντων είναι πολύ μεγαλύτερα. Για να εισαχθούν επιτυχώς νέα προϊόντα, απαιτούνται τεχνολογικές γνώσεις οι οποίες μπορούν να μετατραπούν σε πολύτιμα νέα προϊόντα.
Στην περίπτωση των πρωτότυπων φαρμάκων για να μπορέσει η εταιρία να αποζημιωθεί για την πολύ δαπανηρή διαδικασία ανακάλυψης και ανάπτυξης του νέου φαρμάκου, το κατοχυρώνει με δίπλωμα ευρεσιτεχνίας. Επιπρόσθετα, το πρωτότυπο φαρμακευτικό προϊόν προστατεύεται από περίοδο προστασίας των δεδομένων του η οποία διαρκεί 10 χρόνια μετά την αδειοδότηση του και κατά την οποία κανένα αντίστοιχο γενόσημο φάρμακο δε μπορεί να κυκλοφορήσει στην αγορά.
Κατά την περίοδο που το πρωτότυπο φάρμακο έχει την αποκλειστικότητα στην αγορά, συνήθως έχει ψηλή τιμή. Μετά τη λήξη της περιόδου προστασίας των δεδομένων του πρωτότυπου φαρμάκου, μια άλλη φαρμακοβιομηχανία, μπορεί να εξασφαλίσει άδεια κυκλοφορίας και να κυκλοφορήσει νόμιμα γενόσημο φάρμακο.
Όμως λόγω του ότι η ανάπτυξη των προϊόντων μπορεί να διαρκέσει αρκετά χρόνια, η πατέντα μπορεί να έχει σχεδόν εξαντληθεί έως ότου το προϊόν πια κυκλοφορήσει στην αγορά. Τέλος ακόμη για τις εταιρίες που κατέχουν την πρωτότυπη ιδέα, η τιμή την οποία προτείνουν για το νέο φάρμακο είναι καθοριστικός παράγοντας για να υποδεχθεί το σύστημα υγείας το νέο προϊόν.
Παράγοντες οι οποίοι διαδραματίζουν σημαντικό ρόλο στην επιτυχημένη ανάπτυξη νέων προϊόντων είναι η αφοσίωση των υψηλόβαθμων στελεχών, οι σωστά κατηρτισμένες ομάδες, οι κατάλληλες εσωτερικές και εξωτερικές επικοινωνίες, η καινοτόμος κουλτούρα και η κατάλληλη υποστήριξη μάρκετινγκ. Συνεπώς οι βασικοί παράγοντες επιτυχίας είναι το ανθρώπινο δυναμικό, το πνευματικό κεφάλαιο, το οργανωτικό κεφάλαιο, το σχεσιακό κεφάλαιο και το οργανωτικό κεφάλαιο.
Όσον αφορά τους παράγοντες που σχετίζονται με το ίδιο το προϊόν, στην περίπτωση των φαρμακευτικών εταιριών, το νέο προϊόν δεν πηγάζει από την ζήτηση των πελατών αλλά από τα συμφέροντα και τις ανάγκες του συστήματος υγείας. Κύριος παράγοντας επιτυχίας λοιπόν δεν είναι μόνο η ποιότητα αλλά και η ταχύτητα ανάπτυξης καθώς ελλείψει ανταγωνιστών, το πρώτο προϊόν που είναι έτοιμο να διατεθεί αποτελεσματικά και άμεσα στην αγορά είναι και αυτό που θα επιτύχει. Συνεπώς η πρωτοπορία είναι πολύ σημαντικό πλεονέκτημα για τις φαρμακευτικές εταιρίες.

2.1.2.2 Γενόσημα φάρμακα

Οι εταιρίες που παράγουν γενόσημα φάρμακα στοχεύουν κατά κύριο λόγο στο να είναι οι πρώτες στην αγορά όταν λήξει η πατέντα για ένα πρωτότυπο φάρμακο. Σε αυτήν την περίπτωση ο χρόνος διάθεσης των νέων προϊόντων είναι τους προσδίδει συγκριτικό πλεονέκτημα.

Οι εταιρίες γενοσήμων εξαρτώνται σε μεγάλο βαθμό από τις συνθήκες της τοπικής αγοράς. Οι νομικοί και διοικητικοί περιορισμοί αποτελούν σημαντικούς φραγμούς για την είσοδο σε νέες αγορές.
Οι φάσεις της ανάπτυξης ενός γενοσήμου είναι οι παρακάτω:

1. Παραγωγή ιδεών
2. Προκαταρκτική αξιολόγηση
3. Εργαστηριακή ανάπτυξη
4. Ανάπτυξη της τεχνολογίας
5. Κατάθεση εγγράφων στις αρχές
6. Κυκλοφορία στην αγορά

Οι φάσεις 2,3,4 & 5 είναι αυτές με τον μεγαλύτερο αντίκτυπο στον χρόνο που απαιτείται για να κυκλοφορήσει το προϊόν στην αγορά (Prašnikar, 2006).

2.1.2.3 Ιχνηλασιμότητα στην φαρμακοβιομηχανία

Η υγεία και η ασφάλεια των καταναλωτών είναι η θεμελιώδης ευθύνη μιας φαρμακευτικής εταιρείας. Μια σημαντική απειλή για την αξιοπιστία του κλάδου είναι η ανεξέλεγκτη κυκλοφορία παραποιημένων προϊόντων. Για να αποφευχθεί αυτό, η φαρμακευτική αλυσίδα εφοδιασμού είναι μια από τις πιο αυστηρά ρυθμιζόμενες αλυσίδες εφοδιασμού στον κόσμο.
Κάθε πτυχή του κύκλου ζωής ανάπτυξης του προϊόντος, από την προμήθεια πρώτων υλών έως την παράδοση του προϊόντος στους πωλητές και τελικά στους ασθενείς, υπόκειται σε υψηλά επίπεδα ελέγχου μέσω κανονισμών και προτύπων ποιοτικού ελέγχου. Οι κατευθυντήριες γραμμές GMP στη φαρμακοβιομηχανία είναι ένας από τους πολλούς τέτοιους κανονισμούς. Η κατασκευή προϊόντων υψηλής ποιότητας με ταυτόχρονη τήρηση αυστηρών τοπικών και παγκόσμιων κανονισμών με χρήση συμβατικών μεθόδων είναι πέρα από τα όρια της δυνατότητας. Οι λύσεις ERP Life Science παρέχουν πλέον έναν τρόπο αυτοματοποίησης αυτής της εξαιρετικά περίπλοκης ροής εργασίας. Μία από τις βασικές πτυχές της διαχείρισης της φαρμακευτικής αλυσίδας εφοδιασμού είναι η ιχνηλασιμότητα (Gupta, 2023).
Οι 3 βασικές λειτουργίες για την ιχνηλασιμότητα στην φαρμακοβιομηχανία είναι οι παρακάτω:

- Serialization: Το πρώτο μέρος είναι το serialization που αναφέρεται στη δημιουργία μιας μοναδικής ταυτότητας για ένα φαρμακευτικό προϊόν στη μικρότερη εμπορεύσιμη μονάδα σε ένα ή περισσότερα επίπεδα συσκευασίας αυτού του προϊόντος.
- Η παρακολούθηση: Προσδιορίζει το πού βρίσκεται το προϊόν αυτήν τη στιγμή και καταγράφει τις πληροφορίες καθώς κινείται στην αλυσίδα εφοδιασμού. Η ανίχνευση δεν μας παρέχει απλά την πληροφορία για το πού ήταν το προϊόν ή σε ποιον ανήκει. Η ανίχνευση μας βοηθά να κατανοήσουμε τις αλλαγές ιδιοκτησίας και μας επιτρέπει να επιστρέψουμε σε ένα ορισμένο σημείο του κύκλου ζωής του προϊόντος.
- Επαλήθευση: Μετά τον καθορισμό μιας μοναδικής ταυτότητας σε ένα προϊόν και της ιχνηλασιμότητας των συναλλαγών που σχετίζονται με το προϊόν, είναι επιτακτική ανάγκη το σύστημα να επαληθεύει πληροφορίες σχετικά με τα προϊόντα και τις συναλλαγές σε ένα ή

περισσότερα σημεία της αλυσίδας εφοδιασμού. Η επαλήθευση αφορά περισσότερο τη ρύθμιση και οι νόμοι θα έχουν απαιτήσεις για την επαλήθευση του σειριακού αριθμού και του ιστορικού συναλλαγών των προϊόντων (Gupta, 2023).

Για τις προαναφερθείσες λειτουργίες, πέρα από το σύστημα του serialization των προϊόντων, αξιοποιείται και η συνταγογράφηση.

Είναι κρίσιμο να έχουμε ένα σύστημα που να μπορεί να καταγράφει δεδομένα σε κάθε στάδιο της παραγωγικής διαδικασίας. Εκτός από την παροχή πληροφοριών σε πραγματικό χρόνο σχετικά με τη διαδικασία παραγωγής, η ιχνηλασιμότητα προσφέρει επίσης άλλα οφέλη, όπως:

- Η εφαρμογή των διαδικασιών που απαιτούνται για τη διασφάλιση της συμμόρφωσης με τους κανονισμούς γίνεται πολύ απλούστερη με ένα σύστημα ιχνηλασιμότητας, καθώς βοηθά στη διενέργεια ελέγχων σε διαφορετικά στάδια.
- Επιτρέπει στον οργανισμό να ανακαλεί συγκεκριμένα προϊόντα σε περίπτωση που παρατηρηθεί ένα μη συμμορφούμενο γεγονός.
- Η ενσωμάτωση διαδικασιών ιχνηλασιμότητας σε μια ροή εργασιών μπορεί να βελτιώσει τον έλεγχο των αποθεμάτων και κατά συνέπεια, στη βελτίωση της κερδοφορίας.
- Η ύπαρξη ενός συστήματος ιχνηλασιμότητας λειτουργεί επίσης ως προστιθέμενη αξία που οδηγεί σε καλύτερες σχέσεις εργασίας με προμηθευτές και πελάτες.

2.1.2.3.1 Forward and Backward Traceability

Ο όρος forward traceability - ανίχνευση προς τα εμπρός σημαίνει χρήση των πληροφοριών για την παρακολούθηση της κίνησης των προϊόντων. Η ανίχνευση προς τα εμπρός λειτουργεί ως αποτελεσματικό μέτρο σε μια εποχή ανακλήσεων και εντοπισμού των ελαττωματικών προϊόντων. Έτσι, μόλις εντοπίσετε ελαττώματα στο συγκεκριμένο εξάρτημα, προϊόντα που περιέχουν αυτά τα μέρη, μπορείτε να τα αναγνωρίσετε και να τα ανακαλέσετε αμέσως.
Αντίθετα ο όρος backward traceability - ανίχνευση προς τα πίσω, σημαίνει παρακολούθηση των εγγραφών προς τα πίσω στη γραμμή χρόνου. Για παράδειγμα, εάν υπάρχει πρόβλημα με την αποστολή προϊόντων, μπορείτε να αναγνωρίσετε την παρτίδα και τα προϊόντα ανιχνεύοντας σωστά τις εγγραφές προς τα πίσω. Έτσι, η ανίχνευση πίσω βοηθά στη βελτίωση της ποιότητας της διαδικασίας, με αποτέλεσμα πιο σταθερά και κορυφαίας ποιότητας προϊόντα (Elqortobi, 2023).

2.2 Ο ρόλος της τεχνολογίας στον κύκλο ζωής του φαρμάκου

2.2.1 Big Data

Τα Big Data θα μπορούσαν να οριστούν ως πλούσιες σε πολυμέσα και διαδραστικές πληροφορίες χαμηλού κόστους που προκύπτουν από τη μαζική επικοινωνία.
Οι πληροφορίες αυτές μπορεί να δημιουργηθούν μέσω διαφορετικών πληροφοριακών συστημάτων και τεχνολογιών, συμπεριλαμβανομένων εφαρμογών smartphone, διαδικτυακών κοινοτήτων, δικτύων αισθητήρων, κλικ στο Διαδίκτυο και πλατφορμών μέσων κοινωνικής δικτύωσης (Tan, 2017).
Τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν για να επιτρέψουν στους πελάτες να εκφράσουν μη αναγνωρισμένες ανάγκες. Με την απόκτηση αυτών των πληροφοριών, οι διαχειριστές μπορούν να αποκτήσουν ευκαιρίες να αναπτύξουν προϊόντα με επίκεντρο τον πελάτη (Zhan, 2018).

Τα Big Data προσφέρουν στους πελάτες καλύτερη κατανόηση των νέων προϊόντων και παρέχουν νέους, απλοποιημένους τρόπους αλληλεπίδρασης μεγάλης κλίμακας μεταξύ πελατών και επιχειρήσεων.
Αν και προηγούμενες μελέτες έχουν επισημάνει ότι οι εταιρείες μπορούν να κατανοήσουν καλύτερα τις προτιμήσεις και τις ανάγκες των πελατών αξιοποιώντας διαφορετικούς τύπους διαθέσιμων δεδομένων, η κατάσταση εξελίσσεται, με αυξανόμενη εφαρμογή αναλυτικών δεδομένων Big Data για την ανάπτυξη προϊόντων, τις λειτουργίες και τη διαχείριση της εφοδιαστικής αλυσίδας (Zhan, 2018). Όσο πιο γρήγορα μια εταιρεία ολοκληρώσει τη διαδικασία ανάπτυξης προϊόντος, τόσο μεγαλύτερη είναι η πιθανότητα να ξεπεράσει τους ανταγωνιστές της στην αγορά (Tan, 2017).

Η διαθεσιμότητα των Big Data έχει τονώσει την ευρεία υιοθέτηση της εξόρυξης δεδομένων και της ανάλυσης δεδομένων στην έρευνα και στα επιχειρηματικά περιβάλλοντα. Με τα χρόνια, έχει προταθεί ένας συγκεκριμένος αριθμός μεθοδολογιών εξόρυξης δεδομένων, οι οποίες χρησιμοποιούνται εκτενώς στην πράξη και στην έρευνα. (Plotnikova, 2020)

Παρακάτω θα αναφερθούν παραδείγματα αξιοποίησης των Big Data στην ανάπτυξη νέων προϊόντων από διάφορους κλάδους επιχειρήσεων:

2.2.1.1 Αξιοποίηση των Big Data στην ανάπτυξη νέων προϊόντων στην βιομηχανία τροφίμων.

Μια αποτελεσματική ανάπτυξη νέων προϊόντων θεωρείται ευρέως ως ένας τρόπος για τις εταιρείες τροφίμων να αποκτήσουν ανταγωνιστικά πλεονεκτήματα, στο ακραίο επιχειρηματικό περιβάλλον. Σε αυτό το πλαίσιο, για να διασφαλιστεί η κερδοφορία και να κερδίσουν μερίδιο αγοράς, οι εταιρείες τροφίμων θα μπορούσαν να επιταχύνουν τη διαδικασία της ανάπτυξης προϊόντων και να πρωτοστατήσουν με το να είναι οι πρώτες στην κυκλοφορία του προϊόντος.

Η αξιοποίηση των Big Data έχει αρχίσει να διαδραματίζει ουσιαστικό ρόλο στην προώθηση της ανάπτυξης νέων προϊόντων σε διάφορους τομείς, συμβάλλοντας στη δημιουργία αξίας, τη δημιουργία ιδεών και το ανταγωνιστικό πλεονέκτημα.
Η χρήση των Big Data μπορεί να επιταχύνει τη διαδικασία της ανάπτυξης νέων προϊόντων σε μία εταιρία τροφίμων και να επιταχύνει την κυκλοφορία του προϊόντος σε σχετικά μικρότερο χρόνο. Επιτρέπει στα ενδιαφερόμενα μέρη να εντοπίσουν τυχόν ελλείψεις των προϊόντων διατροφής και να τις αντιμετωπίσουν νωρίτερα στη φάση ανάπτυξης του προϊόντος, οδηγώντας σε τεράστια εξοικονόμηση κόστους που συνεπάγεται η κυκλοφορία νέων προϊόντων. Επιπλέον, βοηθά τις επιχειρήσεις να αναπτύξουν προϊόντα διατροφής, τα οποία είναι επικεντρωμένα στον καταναλωτή και ικανοποιούν τις ανάγκες τους.
Τα Big Data παρέχουν τεράστιο όγκο επιχειρηματικών γνώσεων και πολύτιμων προοπτικών, οι οποίες μπορούν να ωφελήσουν τις επιχειρήσεις τροφίμων να πραγματοποιήσουν εις βάθος αναλύσεις.
Ωστόσο, τα Big Data έχουν επίσης ορισμένα μειονεκτήματα, όπως η ασφάλεια, το απόρρητο και η εμπιστευτικότητα των ευαίσθητων επιχειρηματικών δεδομένων. Επιπλέον, η αποθήκευση, η ανάλυση και η ενσωμάτωση των Big Data στη διαδικασία της ανάπτυξης νέων προϊόντων θα μπορούσε να είναι ένα δύσκολο έργο. Ωστόσο, οι νέες τεχνολογίες εξελίσσονται συνεχώς για την αντιμετώπιση αυτού του ζητήματος και τα οφέλη εξακολουθούν να υπερτερούν των περιορισμών τους. Οι επιχειρήσεις τροφίμων που υιοθετούν τα Big Data για να αξιοποιήσουν τον κύκλο της ανάπτυξης νέων προϊόντων παραμένουν μπροστά από τους ανταγωνιστές τους (Jagtap, 2019).

2.2.1.2 Αξιοποίηση των Big Data στην ανάπτυξη νέων προϊόντων στην βιομηχανία ηλεκτρονικών.

Εξετάζοντας της αξιοποίηση των Big Data στην ανάπτυξη νέων προϊόντων τριών μεγάλων κινεζικών εταιριών (Xiaomi Inc.,Lenovo Group Ltd & Didiдache Inc.) ηλεκτρονικών
κάθε μία από τις περιπτώσεις μπορεί να θεωρηθεί ως ανεξάρτητη περίπτωση.
Οι τρεις περιπτώσεις δείχνουν μια ποικιλία προσεγγίσεων για τη χρήση των Big Data για την υποστήριξη της ανάπτυξης νέων προϊόντων. Ωστόσο και οι τρεις εταιρείες επικεντρώνονται στη δημιουργία ομάδων που μπορούν να εργαστούν τόσο αυτόνομα όσο και ταυτόχρονα, προκειμένου να επιταχύνουν την ανάπτυξη προϊόντων. Συνδέονται επίσης με το ευρύ φάσμα των πελατών τους στο πιο πρώιμο δυνατό στάδιο ανάπτυξης προϊόντων. Παρουσιάζουν τα νέα τους προϊόντα όσο το δυνατόν γρηγορότερα για να κερδίσουν την αναγνώριση της αγοράς καθώς και περαιτέρω σχόλια από τους πελάτες για να ενεργοποιήσουν περαιτέρω συνεχή καινοτομία. Προκειμένου να συλλέξουν Big Data για να ενημερώσουν την ανάπτυξη προϊόντων, εντοπίζουν και εξαλείφουν όσο το δυνατόν περισσότερο τις διαδικασίες σπατάλης χρόνου και κόστους κατά τη διάρκεια της ανάπτυξης προϊόντος

και είναι σε θέση να χρησιμοποιούν Big Data για να διευκολύνουν μια πιο ευέλικτη, λιτή, δυναμική διαδικασία ανάπτυξης προϊόντος που είναι πιο γρήγορη και προσαρμοστική.
Η ανάπτυξη νέων προϊόντων μπορεί να διευκολυνθεί με την εξέλιξη ιδεών όταν ακούγεται η φωνή των πελατών. Το προϊόν είναι καλύτερο όταν οι πιθανοί πελάτες μπορούν να εντοπιστούν και να ικανοποιηθούν οι ανάγκες τους. Για παράδειγμα, το σύστημα MIUI της Xiaomi και η πλατφόρμα μεγάλων δεδομένων Talend της Lenovo είναι και οι δύο καλοί τρόποι για τη δημιουργία στενής σύνδεσης με τους πελάτες. Η Dididache ξόδεψε πολύ χρόνο δημιουργώντας διάφορες πλατφόρμες για να συνδεθεί με τους χρήστες της καθώς και με την αγορά (συμπεριλαμβανομένης της χρήσης Big Data για να εντοπίσει τους κύριους ανταγωνιστές, το μέγεθος της αγοράς και τα προβλήματα και τις ανάγκες των πελατών της. Επιπλέον, αντί να γίνονται αλλαγές αργά στο έργο, η σύνδεση πελατών ενθαρρύνει τις αλλαγές να συμβαίνουν νωρίτερα, όταν είναι λιγότερο δαπανηρές. Η διαδικασία ανάπτυξης νέων προϊόντων μπορεί να επιταχυνθεί χρησιμοποιώντας Big Data με τη μορφή πληροφοριών χρήσης, οι οποίες είναι πολύ πιο γρήγορα διαθέσιμες από τα αποτελέσματα των ερευνών αγοράς. (Tan, 2017).

2.2.1.3 Αξιοποίηση των Big Data στην ανάπτυξη νέων προϊόντων στην φαρμακευτική βιομηχανία.

Πως όμως έχουν αξιοποιηθεί έως τώρα τα big data στην φαρμακοβιομηχανία και πως θα μπορούσαν να αξιοποιηθούν μελλοντικά;
Τα τελευταία 20 χρόνια, η παραγωγικότητα στη φαρμακευτική βιομηχανία μειώνεται λόγω του συνεχώς αυξανόμενου κόστους, ενώ η παραγωγή παραμένει στάσιμη. Ταυτόχρονα, οι πάροχοι υγειονομικής περίθαλψης απαιτούν καλύτερη σχέση ποιότητας/τιμής και σαφείς αποδείξεις ότι τα νέα φάρμακα είναι καλύτερα από τα ήδη υπάρχοντα και μπορούν να παρέχουν βελτιωμένες υπηρεσίες στους ασθενείς (Tormay, 2015). Συνάμα το r&d και η ανάπτυξη νέων προϊόντων σε μία φαρμακευτική εταιρία όπως αναφέρθηκε και παραπάνω είναι πολύ μεγάλες επενδύσεις, οι οποίες απαιτούν την εξερεύνηση και ανάλυση δεδομένων. (Seebode, 2013)
Γι’ αυτό το λόγο οι μελλοντικές τάσεις δείχνουν ότι οι φαρμακευτικές εταιρείες πρέπει να αναπτύξουν νέες μεθόδους επεξεργασίας δεδομένων και πληροφοριών για να ανταποκρίνονται ταχύτερα και με μεγαλύτερη ακρίβεια στις μεταβαλλόμενες αγορές και στις ανάγκες των ασθενών (Seebode, 2013).
Με την εκμετάλλευση των Big Data, τα οποία προσπαθούν να ενσωματώσουν δεδομένα από διαφορετικές πηγές δεδομένων και κλάδους που είναι διαθέσιμοι στη βιοεπιστήμη, η βιομηχανία έχει εντοπίσει ένα νέο σύνορο που θα μπορούσε να παρέχει τις απαραίτητες γνώσεις για να βελτιώσει την ανάπτυξη νέων προϊόντων και να επιτρέψει στη βιομηχανία να επιστρέψει στην βιώσιμη ανάπτυξη.(Tormay, 2015).
Μέχρι τώρα κατά κύριο λόγο οι φαρμακευτικές εταιρίες βασίζονται σε big data τα οποία σχετίζονται με αναλύσεις γονιδιωμάτων για συγκεκριμένες

ασθένειες, και σε big data που απορρέουν από πανεπιστημιακές αλλά και κλινικές μελέτες.
Ωστόσο σήμερα με την όλο και αυξανόμενη παρουσία της τεχνολογίας στην ζωή μας ο ασθενής βρίσκεται όλο και περισσότερο στο προσκήνιο. Δίνεται αυξανόμενη σημασία στα αποτελέσματα που αναφέρονται από τους ασθενείς, συμπεριλαμβανομένων αυτών που δημοσιεύονται στα μέσα κοινωνικής δικτύωσης όπως το Twitter, το Facebook και τα φόρουμ ασθενών. Με τις τεχνολογικές εξελίξεις, η χρήση αυτοματοποιημένων αισθητήρων και έξυπνων συσκευών γίνεται όλο και πιο διαδεδομένη. Συγκεκριμένα, τα smartphone γίνονται διαγνωστικά εργαλεία σημείου φροντίδας μέσω της ανάπτυξης νέων εφαρμογών που σχετίζονται με την υγειονομική περίθαλψη, καθώς και πρόσθετων διαγνωστικών αισθητήρων που χρησιμοποιούν το smartphone ως πλατφόρμα ενεργοποίησης (Tormay, 2015).
Με αυτόν τον τρόπο οι ασθενείς θα έχουν μεγαλύτερη επιρροή στην παροχή υγειονομικής περίθαλψης και κατά συνέπεια θα μπορούν να επηρεάζουν και να συμμετέχουν έμμεσα στην λήψη αποφάσεων (Seebode, 2013)
Συνεπώς μια πλατφόρμα η οποία θα μπορεί να εκμεταλλευτεί τέτοιου είδους big data θα ήταν σε θέση να ανοίξει το δρόμο προς τις απαραίτητες δυνατότητες διαχείρισης μεγάλων δεδομένων.

2.2.1.4 IBM Watson-NLP

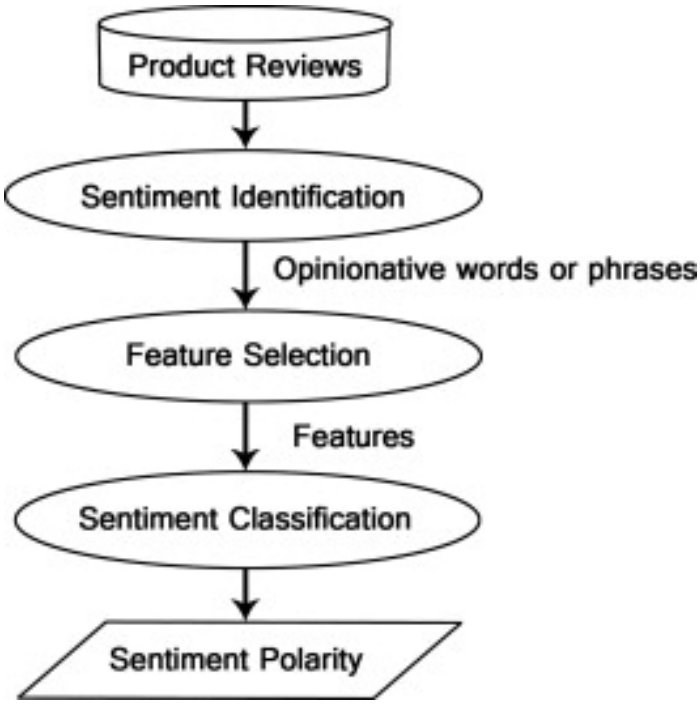
Οι εταιρείες φαρμάκων δεν έχουν κίνητρο για την ανάπτυξη νέων φαρμάκων όταν ο αριθμός των πιθανών ασθενών είναι πολύ μικρός. Μπορεί να χρειαστούν πολλά χρόνια για να ανακαλυφθεί, να αναπτυχθεί και να ερευνηθεί ένα φάρμακο. Επίσης στη συνέχεια ακολουθεί μια παρατεταμένη κλινική έρευνα και μια φάση όπου το φάρμακο πρέπει να δοκιμαστεί για φαρμακευτικές ανεπιθύμητες ενέργειες σε ανθρώπους και να επανεξεταστεί από κυβερνητικούς φορείς. Λόγω αυτών των περιορισμών, η ανάπτυξη φαρμάκων απαιτεί τεράστιες δαπάνες κεφαλαίου, και πολλές φορές αν και η ερευνητική διαδικασία είναι πολλά υποσχόμενη εν τέλη το νέο φάρμακο μπορεί να μην καταλήξει στην αγορά.
Ο Tushar Satav (Ph.D-Founder of Keystonemab) είχε την ιδέα να χρησιμοποιήσει την τεχνολογία για να εντοπίσει υπάρχοντα ή αποτυχημένα φάρμακα για νέες ενδείξεις. Ίδρυσε την Keystonemab μαζί με τον Tjerk Geersing και τον Dr. Roland Meisel, μια startup που χρησιμοποιεί την τεχνητή νοημοσύνη για να βρει κρυφούς συνδέσμους μεταξύ πληροφοριών που εξάγονται από εκατομμύρια επιστημονικές εργασίες που μπορούν να χρησιμοποιηθούν για την ανάπτυξη νέων θεραπειών. Συγκεκριμένα η Keystonemab χρησιμοποιεί το Watson Knowledge Studio και το Watson Natural Language Understanding (NLU).
Συνήθως, μπορεί να χρειαστούν δεκαετίες για τους επιστήμονες που ανακαλύπτουν φάρμακα για να επανεξετάσουν τη σχετική επιστημονική βιβλιογραφία σε ένα δεδομένο πεδίο. Ωστόσο, το λογισμικό της Keystonemab μπορεί να αποκαλύψει γρήγορα πρακτικές πληροφορίες σε πραγματικό χρόνο με τις ισχυρές δυνατότητες Επεξεργασίας Φυσικής Γλώσσας (NLP) της IBM Watson.
Ο λόγος που επιλέχθηκε η Watson είναι ότι αποτελεί μια δοκιμασμένη τεχνολογία και πολλές startup επιχειρήσεις που βρίσκονται στον

φαρμακευτικό τομέα την χρησιμοποιούν. Επίσης διαθέτει μεγάλα επίπεδα ασφάλειας, δεδομένου ότι οι πελάτες των φαρμακευτικών εταιριών απαιτούν υψηλό επίπεδο ασφαλείας για τα δεδομένα τους. Τέλος είναι πολύ φιλική προς τον χρήστη και μπορεί να χρησιμοποιηθεί και από χρήστες οι οποίοι δεν είναι ειδικοί στην τεχνητή νοημοσύνη.
Εκτός από την ανακάλυψη νέων φαρμάκων, οι ερευνητές μπορούν να χρησιμοποιήσουν την τεχνολογία για να βρουν δύο ή περισσότερα υπάρχοντα φάρμακα με συνεργατικά αποτελέσματα, επιτρέποντας στις φαρμακευτικές εταιρείες να πειραματιστούν με νέους συνδυασμούς φαρμάκων. Αυτή η προσέγγιση ονομάζεται «επαναπροσανατολισμός φαρμάκων» ή «επανατοποθέτηση φαρμάκων». Επιφέρει ουσιαστικές βελτιώσεις στην αποτελεσματικότητα της θεραπείας, εξοικονόμηση κόστους και χρόνου, κάτι που μπορεί να είναι σωτήριο σε σενάρια όπως η πανδημία COVID-19. Επιτρέπει επίσης στις φαρμακευτικές εταιρείες να στοχεύουν με πιο ασφαλή και οικονομικά αποδοτικό τρόπο σε σπάνιες παθήσεις. Οι συνδυασμοί φαρμάκων μπορεί επίσης να είναι λιγότερο τοξικοί και πιο αποτελεσματικοί από μια μεγάλη δόση ενός μόνο φαρμάκου.
Το χρονοδιάγραμμα ανάπτυξης για την επανατοποθέτηση φαρμάκων μπορεί να είναι 30 – 60% χαμηλότερο από αυτό για τις de novo ενώσεις και το συνολικό κόστος ανάπτυξης μπορεί να μειωθεί έως και 60%, καθώς τα υπάρχοντα φάρμακα έχουν ήδη υποβληθεί σε ακριβείς δοκιμές ασφάλειας.
Η Keystonemab είναι πεπεισμένη ότι υπάρχουν πολλές βιώσιμες θεραπείες για σπάνιες ασθένειες που περιμένουν να ανακαλυφθούν, αλλά ο μόνος τρόπος για να επιτευχθεί αυτό είναι να γίνουν συνδέσεις μεταξύ διαφορετικών φαρμάκων που συνήθως δεν θεωρούνται συμπληρωματικά.
Το επιχειρηματικό τους μοντέλο βασίζεται σε τεράστιο όγκο δεδομένων φαρμακευτικής έρευνας, πρόσφατα διαθέσιμα στο κοινό, που περιλαμβάνουν πληροφορίες για φάρμακα, ασθένειες, πρωτεΐνες-στόχους φαρμάκων, βιοδείκτες κ.α. Οι ερευνητές μπορούν να χρησιμοποιήσουν την τεχνητή νοημοσύνη (AI) για να ανακαλύψουν σημαντικές σημασιολογικές συνδέσεις μεταξύ δύο ή περισσότερων φαρμάκων που διαθέτουν συμπληρωματικά χαρακτηριστικά.
Το κοινό στο οποίο απευθύνεται η Keystonemab μπορεί να είναι επικεφαλείς επιστήμονες, επικεφαλείς ανάπτυξης και άλλα στελέχη σε μικρές και μεγάλες εταιρείες ανακάλυψης φαρμάκων, αλλά ο κύριος στόχος της είναι να παρέχει την πλατφόρμα σε εμπορικούς κατασκευαστές φαρμάκων, επιστήμονες εργαστηρίου, γιατρούς και άλλους χρήστες σε μικρότερη κλίμακα. Προς το παρόν, η ομάδα επικεντρώνεται στο να παρουσιάσει στους συνεργάτες ευρήματα που δεν θα μπορούσαν να ανακαλύψουν με το τρέχον εσωτερικό προσωπικό ερευνητών τους.
Η μεγαλύτερη πρόκληση είναι ότι η φαρμακευτική βιομηχανία δεν γνωρίζει σε βάθος πώς λειτουργεί η τεχνητή νοημοσύνη και είναι αρκετά συντηρητική στη χρήση μιας νέας τεχνολογίας στην οποία δεν είναι ειδικοί.

2.2.2 Sentiment Analysis

Η Sentiment Analysis είναι ένα πεδίο έρευνας στον τομέα του text mining. Η Sentiment Analysis (SA) είναι η υπολογιστική επεξεργασία απόψεων, συναισθημάτων και υποκειμενικότητας του κειμένου. Η Sentiment Analysis (SA) ή διαφορετικά το Opinion Mining (OP), είναι η υπολογιστική μελέτη των

απόψεων, των στάσεων και των συναισθημάτων των ανθρώπων απέναντι σε μια οντότητα. Η οντότητα μπορεί να αντιπροσωπεύει άτομα, γεγονότα ή θέματα. Αυτά τα θέματα είναι πολύ πιθανό να καλύπτονται από κριτικές. Οι δύο εκφράσεις SA ή OM εκφράζουν ένα αμοιβαίο νόημα. Ωστόσο, ορισμένοι ερευνητές θεωρούν ότι το OM και το SA έχουν ελαφρώς διαφορετικές έννοιες. Το Opinion Mining εξάγει και αναλύει τη γνώμη των ανθρώπων για μια οντότητα, ενώ η Ανάλυση Συναισθήματος προσδιορίζει το συναίσθημα που εκφράζεται σε ένα κείμενο και στη συνέχεια το αναλύει (Medhat, 2014). Ως εκ τούτου, ο στόχος της SA είναι να βρει απόψεις, να εντοπίσει τα συναισθήματα που εκφράζουν και στη συνέχεια να ταξινομήσει την πολικότητα τους όπως φαίνεται στην παρακάτω εικόνα:



Εικόνα 1. Sentiment analysis σε reviews προϊόντων.

Τα σύνολα δεδομένων που χρησιμοποιούνται στην SA είναι ένα σημαντικό ζήτημα. Οι κύριες πηγές δεδομένων προέρχονται από τις κριτικές προϊόντων. Αυτές οι κριτικές είναι σημαντικές για τους κατόχους επιχειρήσεων, καθώς μπορούν να λάβουν επιχειρηματικές αποφάσεις σύμφωνα με τα αποτελέσματα της ανάλυσης των απόψεων των χρηστών σχετικά με τα προϊόντα τους. Οι πηγές κριτικών είναι κυρίως ιστότοποι κριτικών.

2.2.2.1 Sentimental Analysis στην Φαρμακοβιομηχανία

Οι διαδικτυακοί ιστότοποι κριτικών και τα forums περιέχουν πληθώρα πληροφοριών σχετικά με τις προτιμήσεις και τις εμπειρίες των χρηστών. Αυτές οι πληροφορίες μπορούν να αξιοποιηθούν για την απόκτηση πολύτιμων πληροφοριών χρησιμοποιώντας προσεγγίσεις data mining, όπως η ανάλυση συναισθήματος. Στο φαρμακευτικό τομέα οι διαδικτυακές

κριτικές περιέχουν πληροφορίες που σχετίζονται με πολλαπλές πτυχές, όπως τα συναισθήματα που αφορούν τη συνολική ικανοποίηση, τις παρενέργειες και την αποτελεσματικότητα των φαρμάκων, που κάνουν την αυτόματη ανάλυση πολύ ενδιαφέρουσα αλλά και προκλητική. Ωστόσο, η ανάλυση των συναισθημάτων μπορεί να προσφέρει πολύτιμες γνώσεις, να βοηθήσει στη λήψη αποφάσεων και να βελτιώσει την παρακολούθηση της δημόσιας υγείας συνολικά (Gräßer, 2018).

2.2.2.2 Sentimental Analysis σε reviews φαρμάκων

Ένα πολύ σημαντικό κοινωνικό φαινόμενο που έχει προκύψει λόγω της προόδου της τεχνολογίας της πληροφορίας και των επικοινωνιών είναι η τεράστια ενίσχυση της δύναμης του word of mouth. Με τη βοήθεια του διαδικτύου, της ασύρματης δικτύωσης και της κινητής τηλεφωνίας, οι σημερινοί πολίτες και οι καταναλωτές ανταλλάσσουν με μεγάλη ευκολία απόψεις και εμπειρίες για εταιρείες, προϊόντα, υπηρεσίες, ακόμη και παγκόσμια γεγονότα.
Το word of mouth δεν αποτελεί ένα νέο φαινόμενο, ωστόσο, η έλευση του Διαδικτύου και η εξέλιξη της τεχνολογίας πρόσθεσε δύο σημαντικές νέες διαστάσεις σε αυτή τη διαχρονική ιδέα:

- Επεκτασιμότητα και ταχύτητα διάδοσης. Οι τεχνολογίες της πληροφορίας δίνουν τη δυνατότητα στις απόψεις ενός μεμονωμένου ατόμου να προσεγγίσουν αμέσως χιλιάδες ή και εκατομμύρια καταναλωτές. Αυτή η κλιμάκωση στο κοινό αλλάζει τη δυναμική πολλών βιομηχανιών στις οποίες το word of mouth παραδοσιακά διαδραματίζει σημαντικό ρόλο. Ένα τέτοιο παράδειγμα είναι η βιομηχανία του θεάματος.
- Επιμονή και δυνατότητα μέτρησης. Την περίοδο πριν το διαδίκτυο, με το word of mouth η πληροφορία χανόταν μόλις ολοκληρωνόταν η επικοινωνία. Σήμερα, με την ύπαρξη του διαδικτύου, οι πληροφορίες του word of mouth παραμένουν διαθέσιμες σε πολλά δημόσια διαθέσιμα φόρουμ, όπως σε αξιολογήσεις, ομάδες συζήτησης κ.α. Αυτά τα δημόσια δεδομένα παρέχουν στους οργανισμούς τη δυνατότητα να μετρούν γρήγορα και με ακρίβεια το word of mouth, όπως συμβαίνει με το mining πληροφοριών που είναι διαθέσιμες στα φόρουμ του Διαδικτύου (Dellarocas, 2004).

Η γρήγορη μέτρηση είναι η πρώτη προϋπόθεση για τις γρήγορες αντιδράσεις που χρειάζονται σε αυτόν τον νέο αγωνιστικό χώρο. Ωστόσο, η πληροφοριακή αξία των διαδικτυακών φόρουμ για οργανισμούς δεν είναι επί του παρόντος καλά κατανοητή. Υπάρχει διαμάχη σχετικά με την αξιοπιστία των διαδικτυακών κριτικών καθώς και με το κατά πόσο αυτές αντικατοπτρίζουν τις απόψεις του πληθυσμού των καταναλωτών. Κάποια στοιχεία υποδηλώνουν ότι πιθανώς ορισμένες από αυτές τις πληροφορίες μπορεί να είναι προκατειλημμένες και μερικές φορές παρέχονται ανώνυμα από τις ίδιες τις εταιρείες. Τέλος, παρόλο που ο αντίκτυπος των διαδικτυακών κριτικών στη συμπεριφορά των καταναλωτών έχει βρεθεί στο επίκεντρο προσφάτων ερευνών, υπάρχει πολύ λίγη μελέτη σχετικά με το πώς αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν από τις εταιρείες για να αποκτήσουν επιχειρηματικό πλεονέκτημα.

Οι Chrysanthos Dellarocas, Neveen Farag Awad και Xiaoquan (Michael) Zhang, μελέτησαν την εταιρική χρήση των δημοσίως διαθέσιμων διαδικτυακών κριτικών ταινιών στην πρόβλεψη εσόδων από κινηματογραφικές ταινίες. Επικεντρώθηκαν στη βιομηχανία του κινηματογράφου, καθώς στον συγκεκριμένο κλάδο το word of mouth διαδραματίζει σημαντικό ρόλο και επειδή οι online κριτικές ταινιών είναι άμεσα διαθέσιμες. Εμπνευσμένοι από το μοντέλο Bass, ανέπτυξαν ένα ιδιαίτερα ακριβές μοντέλο πρόβλεψης εσόδων που βασίζεται σε στατιστικά στοιχεία κριτικών ταινιών που δημοσιεύονται από χρήστες στο Yahoo στο διαδίκτυο την πρώτη εβδομάδα από την κυκλοφορία μιας νέας ταινίας. Η συγκεκριμένη εργασία δεν καταλήγει στο ότι οι διαδικτυακές κριτικές ταινιών επηρεάζουν τα μελλοντικά έσοδα, αλλά ότι αποτελούν μία μετρήσιμη ένδειξη για την αξιοποίηση του word of mouth. Εκτός από το να βοηθηθεί η εταιρεία να προβλέψει τη ζήτηση και να σχεδιάσει τις δικές της ενέργειες, οι διαδικτυακές κριτικές μπορούν να την βοηθήσουν να αναλύσει τον ανταγωνισμό. Σε πολλές κατηγορίες προϊόντων, οι προϋπολογισμοί πωλήσεων και μάρκετινγκ είναι μυστικές πληροφορίες και επομένως η ανάλυση του ανταγωνισμού ήταν δύσκολη υπόθεση. Οι τεράστιοι όγκοι αξιολογήσεων των καταναλωτών που είναι δημόσια διαθέσιμες στο Διαδίκτυο έχουν τη δυνατότητα να το αλλάξουν ριζικά (Dellarocas, 2004).

2.2.3 Topic Modeling

Η μοντελοποίηση θεμάτων (topic modelling) είναι ένας τύπος στατιστικής μοντελοποίησης που χρησιμοποιεί τη μη εποπτευόμενη Μηχανική Εκμάθηση για τον εντοπισμό ομάδων παρόμοιων λέξεων μέσα σε ένα σώμα κειμένου. Αυτή η μέθοδος εξόρυξης κειμένου χρησιμοποιεί σημασιολογικές δομές σε κείμενο για την κατανόηση μη δομημένων δεδομένων χωρίς προκαθορισμένες ταμπέλες ή δεδομένα εκπαίδευσης (Henderi, 2023).

2.2.3.1 LDA

Η LDA (Latent Dirichlet Allocation) είναι μια μέθοδος που σας επιτρέπει να προσδιορίζετε θέματα σε έγγραφα. Η μοντελοποίηση θεμάτων (toping modeling) είναι μια μέθοδος ταξινόμησης εγγράφων χωρίς επίβλεψη, παρόμοια με τη ομαδοποίηση σε αριθμητικά δεδομένα, η οποία βρίσκει ορισμένες φυσικές ομάδες στοιχείων (θέματα) ακόμη και όταν δεν είμαστε σίγουροι για το τι ψάχνουμε. Ένα έγγραφο μπορεί να αποτελεί μέρος πολλών θεμάτων, όπως στην ασαφή ομαδοποίηση (soft clustering) στην οποία κάθε σημείο δεδομένων ανήκει σε περισσότερα από ένα cluster. Η μοντελοποίηση θεμάτων παρέχει μεθόδους για αυτόματη οργάνωση, κατανόηση, αναζήτηση και σύνοψη μεγάλων ηλεκτρονικών αρχείων. Μπορεί να βοηθήσει στην ανακάλυψη κρυμμένων θεμάτων, στην ταξινόμηση των εγγράφων στα θέματα που ανακαλύφθηκαν.
Η LDA είναι μια από τις πιο δημοφιλείς μεθόδους μοντελοποίησης θεμάτων. Κάθε έγγραφο αποτελείται από διάφορες λέξεις και κάθε θέμα έχει επίσης διάφορες λέξεις που ανήκουν σε αυτό. Ο στόχος του LDA είναι να βρει θέματα στα οποία ανήκει ένα έγγραφο, με βάση τις λέξεις σε αυτό.

Ο τρόπος λειτουργίας της LDA περιγράφεται συνοπτικά παρακάτω:
Οι λέξεις ταξινομούνται σε σχέση με τη βαθμολογία πιθανοτήτων τους.
Οι κορυφαίες x λέξεις επιλέγονται από κάθε θέμα για να αντιπροσωπεύουν το θέμα. Εάν x = 10, θα ταξινομηθούν όλες τις λέξεις στο θέμα1 με βάση τη βαθμολογία τους και θα ληφθούν οι 10 πρώτες λέξεις για να αντιπροσωπεύσουν το θέμα. Αυτό το βήμα μπορεί να μην είναι πάντα απαραίτητο γιατί αν το σώμα είναι μικρό μπορούν να αποθηκευτούν όλες οι λέξεις ταξινομημένες με βάση τη βαθμολογία τους.
Εναλλακτικά, μπορεί να οριστεί ένα όριο στο σκορ. Όλες οι λέξεις σε ένα θέμα που έχουν βαθμολογία πάνω από το όριο μπορούν να αποθηκευτούν ως αντιπροσωπευτικές του, κατά σειρά βαθμολογίας.

Κάθε έγγραφο είναι απλώς μια συλλογή λέξεων. Έτσι, η σειρά των λέξεων και ο γραμματικός ρόλος των λέξεων (υποκείμενο, αντικείμενο, ρήματα,…) δεν λαμβάνονται υπόψη στο μοντέλο. Λέξεις όπως am/is/are/of/a/the/but κ.ο.κ. δεν περιέχουν καμία πληροφορία σχετικά με τα «θέματα» και επομένως μπορούν να εξαλειφθούν από τα έγγραφα ως βήμα προεπεξεργασίας. Στην πραγματικότητα, μπορούν να εξαλειφθούν λέξεις που εμφανίζονται σε τουλάχιστον %80 ~ %90 των εγγράφων, χωρίς να χάθεί καμία πληροφορία. Για παράδειγμα, εάν το σώμα περιέχει μόνο ιατρικά έγγραφα, λέξεις όπως άνθρωπος, σώμα, υγεία κ.λπ. ενδέχεται να υπάρχουν στα περισσότερα έγγραφα και ως εκ τούτου μπορούν να αφαιρεθούν, καθώς δεν προσθέτουν συγκεκριμένες πληροφορίες που θα κάνουν το έγγραφο να ξεχωρίζει (Henderi, 2023).

2.2.3.2 LSA

Η LSA (Latent semantic analysis) είναι μια στατιστική τεχνική για την εξαγωγή και την αναπαράσταση των κύριων ιδεών σε ένα σώμα κειμένου. Το LSA βασίζεται στην αρχή ότι οι λέξεις που έχουν κοντινή σημασία τείνουν να χρησιμοποιούνται μαζί στο πλαίσιο. Το LSA συνδέει λέξεις σημασιολογικά με βάση το περιβάλλον και τη συχνότητα λέξης — πόσο συχνά εμφανίζεται μια λέξη σε ένα έγγραφο. Χρησιμοποιώντας μη εποπτευόμενη μηχανική εκμάθηση, το LSA δημιουργεί αυτόματα ξεχωριστά θέματα με βάση προηγούμενες εισόδους και εξόδους. Η LSA υποθέτει ότι όλα τα παρόμοια έγγραφα μοιράζονται τα ίδια μοτίβα όταν η συχνότητα και η σειρά των λέξεων τους είναι συνεπείς, κάτι που βοηθά στην ανάλυση όχι μόνο ενός αλλά μιας συλλογής εγγράφων. Αποτελεί μέρος της επεξεργασίας φυσικής γλώσσας, στενά συνδεδεμένο με την εκμάθηση και την κατανόηση της ανθρώπινης γλώσσας και κρίσης και έχει πολλές εφαρμογές, συμπεριλαμβανομένης της ταξινόμησης εγγράφων και εικόνων (Henderi, 2023).

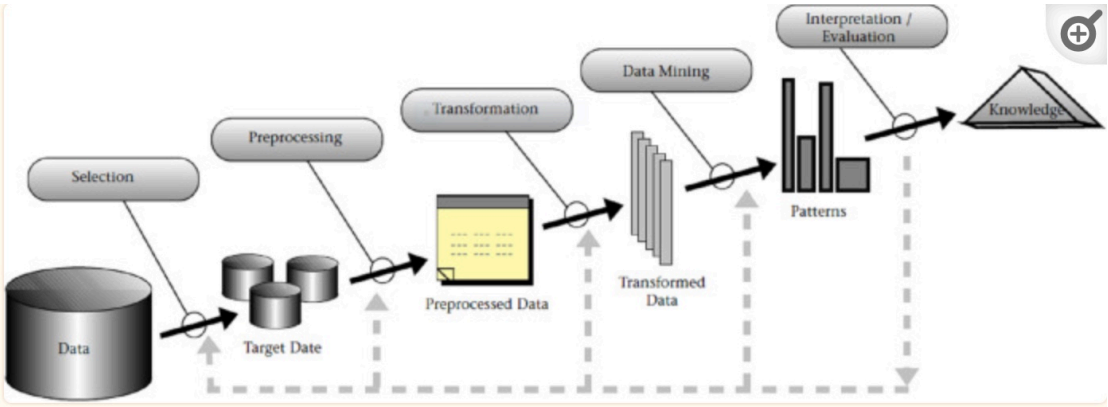
2.3 Αναγνώριση κενών για έρευνα

Παρότι όπως αναφέρθηκε και παραπάνω, τα τελευταία χρόνια έχουν γίνει τα πρώτα βήματα για την αξιοποίηση των Big Data και του Crowdsourncing στον τομέα των φαρμάκων, ο τομέας αυτός δεν είναι αρκετά ώριμος ώστε να βασίσει ένα μεγάλο μέρος την ανάπτυξης των φαρμάκων στις τεχνολογίες αυτές.
Οι νομοθεσίες είναι πολύ αυστηρές και οι διαδικασίες πολύ συγκεκριμένες, με αποτέλεσμα η μετάβαση σε μια νέα «τεχνολογική εποχή» να απαιτεί χρόνο για να αφομοιωθεί και να τεκμηριωθεί.
Κάθε νέα διαδικασία που πρόκειται να ακολουθηθεί θα πρέπει να περιγραφεί αναλυτικά στις γενικές διαδικασίες κάθε εταιρίας, να γίνει η απαραίτητη ανάλυση ρίσκου και φυσικά να ληφθούν οι απαραίτητες εγκρίσεις από τις αρμόδιες αρχές.
Ωστόσο, πολύ μεγάλες φαρμακευτικές εταιρείες όπως η AstraZeneca, η Lilly, η GSK, η Merck, η Janssen, η Daiichi-Sankyo, η Pfizer, η TransCelerate και άλλες έχουν ξεκινήσει πρωτοβουλίες για να επιτρέψουν στην επιστημονική καινοτομία να διασχίσει τα σύνορα μεταξύ εταιρειών, ακαδημαϊκών, κυβερνήσεων και μη κερδοσκοπικών οργανισμών.
Στην παρούσα εργασία θα επιχειρήσουμε να εξετάσουμε πως μια φαρμακευτική εταιρία, μπορεί να αξιοποιήσει τα αρνητικά σχόλια από μια πλατφόρμα με reviews φαρμάκων, ώστε να δημιουργήσει ένα νέο φάρμακο, πιο αποτελεσματικό και με λιγότερες παρενέργειες.

3 Μεθοδολογία

Η εξόρυξη δεδομένων (data mining) ορίζεται ως ένα σύνολο κανόνων, διαδικασιών, αλγορίθμων που έχουν σχεδιαστεί για τη δημιουργία πρακτικών πληροφοριών, την εξαγωγή μοτίβων και τον εντοπισμό σχέσεων από μεγάλα σύνολα δεδομένων. Η εξόρυξη δεδομένων ενσωματώνει την αυτοματοποιημένη εξαγωγή, επεξεργασία και μοντελοποίηση δεδομένων μέσω μιας σειράς μεθόδων και τεχνικών. Αντίθετα, η ανάλυση δεδομένων αναφέρεται σε τεχνικές που χρησιμοποιούνται για την ανάλυση και την απόκτηση νοημοσύνης από δεδομένα (συμπεριλαμβανομένων των «Big Data») και τοποθετείται ως ένα ευρύτερο πεδίο, που περιλαμβάνει ένα ευρύτερο φάσμα μεθόδων που περιλαμβάνει τόσο στατιστικά όσο και δεδομένα εξόρυξης. Ένας αριθμός αλγορίθμων έχει αναπτυχθεί σε τομείς στατιστικής, μηχανικής μάθησης και τεχνητής νοημοσύνης για την υποστήριξη και την ενεργοποίηση της εξόρυξης δεδομένων. Ενώ οι στατιστικές προσεγγίσεις προηγούνται, έχουν περιορισμούς, με τους πιο γνωστούς να είναι οι άκαμπτες συνθήκες διανομής δεδομένων. Οι τεχνικές μηχανικής μάθησης κέρδισαν δημοτικότητα καθώς επιβάλλουν λιγότερους περιορισμούς ενώ αντλούν κατανοητά πρότυπα από δεδομένα (Plotnikova 2020).

Τα θεμέλια των μεθοδολογιών δομημένης εξόρυξης δεδομένων προτάθηκαν για πρώτη φορά από τους Fayyad, Piatetsky-Shapiro & Smyth και αρχικά σχετίζονταν με την Ανακάλυψη Γνώσης στις Βάσεις Δεδομένων (KDD). Το KDD παρουσιάζει ένα εννοιολογικό μοντέλο διαδικασίας υπολογιστικών θεωριών και εργαλείων που υποστηρίζουν την εξαγωγή πληροφοριών (γνώση) με δεδομένα. Στο KDD, η συνολική προσέγγιση στην ανακάλυψη γνώσης περιλαμβάνει την εξόρυξη δεδομένων ως ένα συγκεκριμένο βήμα. Ως εκ τούτου, το KDD, με τα εννέα κύρια βήματα του (που παρουσιάζονται στην Εικόνα 2), έχει το πλεονέκτημα να εξετάζει την αποθήκευση και την πρόσβαση δεδομένων, την κλιμάκωση αλγορίθμων, την ερμηνεία και την οπτικοποίηση των αποτελεσμάτων και την αλληλεπίδραση με τον άνθρωπο. Η εισαγωγή του KDD επισημοποίησε επίσης σαφέστερη διάκριση μεταξύ εξόρυξης δεδομένων και ανάλυσης δεδομένων: Με την ανάλυση δεδομένων, εννοούμε ολόκληρη τη διαδικασία KDD, ενώ με την εξόρυξη δεδομένων, εννοούμε το τμήμα της ανάλυσης δεδομένων που στοχεύει στην εύρεση των κρυμμένων πληροφοριών στα δεδομένα, data mining (Plotnikova, 2020).



Εικόνα 2. Τα κύρια βήματα του KDD

Τα κύρια βήματα του KDD είναι τα εξής:

- Βήμα 1: Εκμάθηση πεδίου εφαρμογής: Στο πρώτο βήμα, χρειάζεται να αναπτυχθεί μια κατανόηση του τομέα εφαρμογής και η σχετική προηγούμενη γνώση, ακολουθούμενη από τον προσδιορισμό του στόχου της διαδικασίας KDD από την οπτική γωνία του πελάτη.
- Βήμα 2: Δημιουργία συνόλου δεδομένων: Το δεύτερο βήμα περιλαμβάνει την επιλογή ενός συνόλου δεδομένων, την εστίαση σε ένα υποσύνολο μεταβλητών ή δειγμάτων δεδομένων στα οποία πρόκειται να πραγματοποιηθεί η μελέτη.
- Βήμα 3: Καθαρισμός και επεξεργασία δεδομένων: Στο τρίτο βήμα, εκτελούνται βασικές λειτουργίες αφαίρεσης θορύβου ή ακραίων στοιχείων. Λαμβάνονται επίσης υπόψη η συλλογή των απαραίτητων πληροφοριών για τη μοντελοποίηση ή τον υπολογισμό του θορύβου, τη λήψη αποφάσεων για στρατηγικές χειρισμού πεδίων δεδομένων που λείπουν και τη λογιστική για τους τύπους δεδομένων, το σχήμα και την αντιστοίχιση τιμών που λείπουν.
- Βήμα 4: Μείωση και προβολή δεδομένων: Εδώ, διεξάγεται η εργασία εύρεσης χρήσιμων χαρακτηριστικών για την αναπαράσταση των δεδομένων, ανάλογα με τον στόχο της εργασίας, εφαρμογή μεθόδων μετασχηματισμού για την εύρεση βέλτιστων χαρακτηριστικών που έχουν οριστεί για τα δεδομένα.
- Βήμα 5: Επιλογή της συνάρτησης εξόρυξης δεδομένων: Στο πέμπτο βήμα, ορίζεται το αποτέλεσμα-στόχος (π.χ. σύνοψη, ταξινόμηση, παλινδρόμηση, ομαδοποίηση).
- Βήμα 6: Επιλογή αλγορίθμου εξόρυξης δεδομένων: Το έκτο βήμα αφορά την επιλογή μεθόδων για αναζήτηση μοτίβων στα δεδομένα, την απόφαση ποια μοντέλα και παραμέτρους είναι κατάλληλα και την αντιστοίχιση μιας συγκεκριμένης μεθόδου εξόρυξης δεδομένων με τα γενικά κριτήρια της διαδικασίας KDD.
- Βήμα 7: Εξόρυξη δεδομένων: Στο έβδομο βήμα, διεξάγεται η εργασία εξόρυξης των δεδομένων, δηλαδή η αναζήτηση προτύπων ενδιαφέροντος σε μια συγκεκριμένη αναπαραστατική μορφή ή ένα σύνολο τέτοιων αναπαραστάσεων: κανόνες ταξινόμησης ή δέντρα, παλινδρόμηση, ομαδοποίηση.

- Βήμα 8: Ερμηνεία: Σε αυτό το βήμα, τα περιττά και άσχετα μοτίβα φιλτράρονται, τα σχετικά μοτίβα ερμηνεύονται και οπτικοποιούνται με τέτοιο τρόπο ώστε να γίνεται κατανοητό το αποτέλεσμα στους χρήστες.
- Βήμα 9: Χρήση ανακαλυφθείσας γνώσης: Στο τελευταίο βήμα, τα αποτελέσματα ενσωματώνονται με το σύστημα απόδοσης, τεκμηριώνονται και αναφέρονται στους ενδιαφερόμενους φορείς και χρησιμοποιούνται ως βάση για αποφάσεις (Plotnikova, 2020).

Το 2000, ως απάντηση σε κοινά ζητήματα και ανάγκες της εποχής, μια μεθοδολογία καθοδηγούμενη από τη βιομηχανία που ονομάζεται Cross-Industry Standard Process for Data Mining (CRISP-DM) εισήχθη ως εναλλακτική λύση στο KDD. Επίσης, ενοποίησε το αρχικό μοντέλο KDD και τις διάφορες επεκτάσεις του. Ενώ το CRISP-DM βασίζεται στο KDD, αποτελείται από έξι φάσεις που εκτελούνται σε επαναλήψεις. Οι επαναληπτικές εκτελέσεις του CRISP-DM αποτελούν το πιο διακριτό χαρακτηριστικό σε σύγκριση με το αρχικό KDD που προϋποθέτει μια διαδοχική εκτέλεση των βημάτων του. Το CRISP-DM, όπως και το KDD, στοχεύει στο να παρέχει στους επαγγελματίες κατευθυντήριες γραμμές για την εκτέλεση εξόρυξης δεδομένων σε μεγάλα σύνολα δεδομένων. Ωστόσο, το CRISP-DM με τα έξι βασικά του βήματα με συνολικά 24 εργασίες και εξόδους, είναι πιο εκλεπτυσμένο σε σύγκριση με το KDD.
Τα κύρια βήματα του CRIPS-DM, όπως απεικονίζονται στην (Εικόνα 3) παρακάτω είναι τα εξής:

Phase	Short description
Business Understanding	The business situation should be assessed to get an overview of the available and required resources. The determination of the data mining goal is one of the most important aspect in this phase. First the data mining type should be explained (e. g. classification) and the data mining success criteria (like precision). A compulsory project plan should be created.
Data understanding	Collecting data from data sources, exploring and describing it and checking the data quality are essential tasks in this phase. To make it more concrete, the user guide describe the data description task with using statistical analysis and determining attributes and their collations.
Data preparation	Data selection should be conducted by defining inclusion and exclusion criteria. Bad data quality can be handled by cleaning data. Dependent on the used model (defined in the first phase) derived attributes have to be constructed. For all these steps different methods are possible and are model dependent.
Modeling	The data modelling phase consists of selecting the modeling technique, building the test case and the model. All data mining techniques can be used. In general, the choice is depending on the business problem and the data. More important is, how to explain the choice. For building the model, specific parameters have to be set. For assessing the model it is appropriate to evaluate the model against evaluation criteria and select the best ones.
Evaluation	In the evaluation phase the results are checked against the defined business objectives. Therefore, the results have to be interpreted and further actions have to be defined. Another point is, that the process should be reviewed in general.
Deployment	The deployment phase is described generally in the user guide. It could be a final report or a software component. The user guide describes that the deployment phase consists of planning the deployment, monitoring and maintenance.

Εικόνα 3. Οι έξι φάσεις του CRISP-DM

Οι έξι φάσεις αναλύονται παρακάτω:

- Φάση 1: Επιχειρησιακή κατανόηση: Το επίκεντρο του πρώτου βήματος είναι η κατανόηση των στόχων και των απαιτήσεων του έργου από επιχειρηματική σκοπιά, ακολουθούμενη από τη μετατροπή τους σε ορισμούς προβλημάτων εξόρυξης δεδομένων. Η παρουσίαση ενός προκαταρκτικού σχεδίου για την επίτευξη των στόχων περιλαμβάνεται επίσης σε αυτό το πρώτο βήμα.
- Φάση 2: Κατανόηση δεδομένων: Αυτό το βήμα ξεκινά με μια αρχική συλλογή δεδομένων και συνεχίζει με δραστηριότητες για την εξοικείωση με τα δεδομένα, τον εντοπισμό ζητημάτων ποιότητας δεδομένων, την πρώτη επίγνωση των δεδομένων και, ενδεχομένως, τον εντοπισμό και τη διαμόρφωση υποθέσεων.
- Φάση 3: Προετοιμασία δεδομένων: Το τρίτο βήμα καλύπτει δραστηριότητες που απαιτούνται για την κατασκευή του τελικού συνόλου δεδομένων από τα αρχικά ακατέργαστα δεδομένα. Οι εργασίες προετοιμασίας δεδομένων εκτελούνται επανειλημμένα.
- Φάση 4: Φάση μοντελοποίησης: Σε αυτό το βήμα, επιλέγονται και εφαρμόζονται διάφορες τεχνικές μοντελοποίησης ακολουθούμενες από βαθμονόμηση των παραμέτρων τους. Συνήθως, χρησιμοποιούνται διάφορες τεχνικές για το ίδιο πρόβλημα εξόρυξης δεδομένων.
- Φάση 5: Αξιολόγηση του μοντέλου: Το πέμπτο βήμα ξεκινά με την προοπτική ποιότητας και στη συνέχεια, πριν προχωρήσουμε στην τελική ανάπτυξη του μοντέλου, διαπιστώνεται ότι το μοντέλο επιτυγχάνει τους επιχειρηματικούς στόχους. Στο τέλος αυτής της φάσης, θα πρέπει να ληφθεί μια απόφαση σχετικά με τον τρόπο χρήσης των αποτελεσμάτων της εξόρυξης δεδομένων.
- Φάση 6: Φάση ανάπτυξης: Στο τελικό βήμα, τα μοντέλα αναπτύσσονται για να επιτρέψουν στους πελάτες να χρησιμοποιούν τα δεδομένα ως βάση για αποφάσειςF ή υποστήριξη στην επιχειρηματική διαδικασία. Ακόμα κι αν ο σκοπός του μοντέλου είναι να αυξήσει τη γνώση των δεδομένων, η γνώση που αποκτάται θα πρέπει να οργανωθεί, να παρουσιαστεί, να διανεμηθεί με τρόπο ώστε ο τελικός χρήστης να μπορεί να τη χρησιμοποιήσει. Ανάλογα με τις απαιτήσεις, η φάση ανάπτυξης μπορεί να είναι τόσο απλή όσο η δημιουργία μιας αναφοράς ή τόσο περίπλοκη όσο η εφαρμογή μιας επαναλαμβανόμενης διαδικασίας εξόρυξης δεδομένων (Plotnikova, 2020).

Η ανάπτυξη του CRISP-DM έγινε από την κοινοπραξία της βιομηχανίας και ως εκ τούτου, χρησιμοποιείται πλέον ευρέως από τη βιομηχανία και τις ερευνητικές κοινότητες. Αυτά τα διακριτικά χαρακτηριστικά έχουν κάνει το CRISP-DM να θεωρείται ως «de-facto» πρότυπο μεθοδολογίας εξόρυξης δεδομένων και ως πλαίσιο αναφοράς στο οποίο συγκρίνονται άλλες μεθοδολογίες (Plotnikova, 2020). Η φάση Κατανόησης Δεδομένων του CRISP-DM συνδυάζει τις φάσεις Επιλογής και Προεπεξεργασίας στο KDD. Το CRISP-DM είναι διαφορετικό από το KDD στο στάδιο της επιχειρηματικής κατανόησης επειδή δεν υπάρχει τέτοιο στάδιο στο KDD. Η φάση επιχειρηματικής κατανόησης καλύπτει τα βήματα του CRISP-DM για τη

δημιουργία ενός αξιόπιστου έργου επιστήμης δεδομένων, επειδή εστιάζει στην κατανόηση των στόχων και των απαιτήσεων. Και ως σημαντική διαφορά μεταξύ KDD και CRISP-DM, το CRISP-DM έχει αναστρέψιμα βήματα που παρέχουν μεγαλύτερη ευκολία στην εξόρυξη δεδομένων. Δεδομένου ότι τα στάδια μπορούν να αντιστραφούν, οποιοδήποτε λάθος μπορεί να διορθωθεί χωρίς να ολοκληρωθεί ολόκληρος ο κύκλος.

3.1 Διαθεσιμότητα δεδομένων

Στην παρούσα εργασία τα δεδομένα αντλήθηκαν από το πρόγραμμα Rapid Miner, το οποίο αποτελεί μια πλατφόρμα επιστήμης δεδομένων. Η πλατφόρμα αυτή προορίζεται να υποστηρίξει πολλούς χρήστες αναλυτικών στοιχείων σε έναν ευρύ κύκλο ζωής AI. Το RapidMiner παρέχει διαδικασίες εξόρυξης δεδομένων και μηχανικής μάθησης, όπως: φόρτωση και μετασχηματισμός δεδομένων, προεπεξεργασία και οπτικοποίηση δεδομένων, προγνωστική ανάλυση και στατιστική μοντελοποίηση, αξιολόγηση και ανάπτυξη. Το RapidMiner είναι γραμμένο στη γλώσσα προγραμματισμού Java. Πιο συγκεκριμένα αντλήθηκαν δεδομένα, σχόλια χρηστών, που αφορούσαν φάρμακα για την κατάθλιψης και την αντισύλληψη. Το κριτήριο με βάση το οποίο επιλέχθηκαν οι παθήσεις αυτές είναι ότι ήταν οι δύο με τον μεγαλύτερο αριθμό σχολίων.

Τα δεδομένα που αναλύθηκαν αντλήθηκαν από την βάση δεδομένων UCI Machine Learning Repository. Η συγκεκριμένη βάση δεδομένων παρέχει κριτικές ασθενών για συγκεκριμένα φάρμακα, όπως και για τις σχετικές παθήσεις και βαθμολογία ασθενών με 10 αστέρια που αντικατοπτρίζει τη συνολική ικανοποίηση των ασθενών. Τα δεδομένα ελήφθησαν με ανίχνευση διαδικτυακών ιστότοπων με reviews φαρμακευτικών προϊόντων. Kallumadi,Surya and Grer,Felix. (2018). Drug Review Dataset (Drugs.com). UCI Machine Learning Repository.
Ο σκοπός της συγκεκριμένης βάσης δεδομένων είναι να μελετήσει τα παρακάτω:

1. Ανάλυση συναισθήματος της εμπειρίας από τα φάρμακα σε πολλαπλές πτυχές, δηλαδή συναισθήματα που μάθαμε για συγκεκριμένες πτυχές, όπως η αποτελεσματικότητα και οι παρενέργειες,
2. Η δυνατότητα μεταφοράς μοντέλων μεταξύ τομέων, όπως για παράδειγμα οι παθήσεις
3. Η δυνατότητα μεταφοράς μοντέλων μεταξύ διαφορετικών πηγών δεδομένων

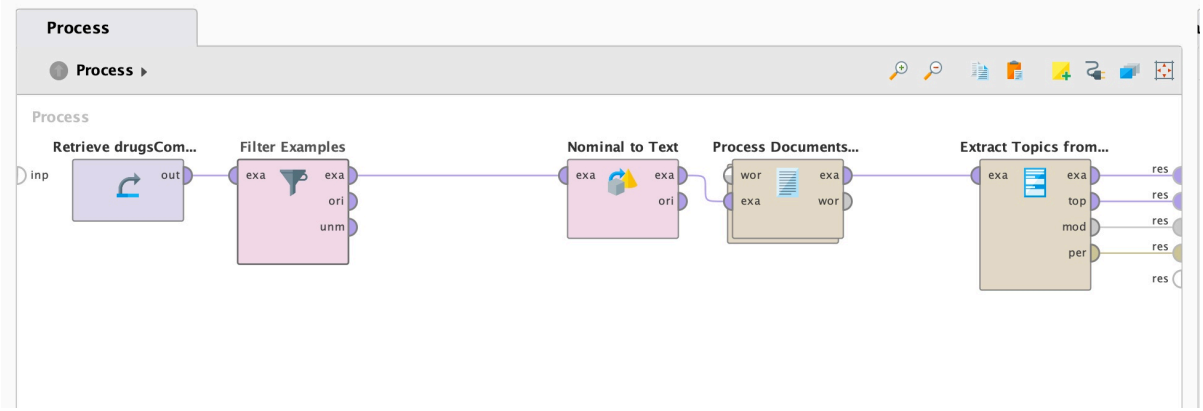
4 Εμπειρική έρευνα

Στην εμπειρική έρευνα της παρούσας εργασίας, χρησιμοποιήθηκε το πρόγραμμα Rapid Miner, όπως αναφέρθηκε και παραπάνω. Στα πειράματα που πραγματοποιήθηκαν, χρησιμοποιήθηκαν και αναλύθηκαν τα σχόλια χρηστών-ασθενών σχετικά με δύο κατηγορίες φαρμακευτικών σκευασμάτων: αντισυλληπτικά φάρμακα (birth control medicines) και αντικαταθλιπτικά φάρμακα (depression medicines).
Και για τις δύο κατηγορίες φαρμάκων πραγματοποιήθηκαν δύο ειδών αναλύσεις, Sentiment Analysis και LDA Analysis, με διαφορετικά φίλτρα σε κάθε πειραματισμό, τα οποία θα αναλυθούν παρακάτω.
Πιο συγκεκριμένα θα εξεταστούν τα παρακάτω:

Μέθοδος	**Παράμετροι**
LDA	• Birth Control, no filter in rating • Birth Control, filter in rating<5 • Depression, no filter in rating • Depression, filter in rating<5
Sentiment Analysis	• Birth Control, no filter in rating • Birth Control, filter in rating<5 • Depression, no filter in rating • Depression, filter in rating<5

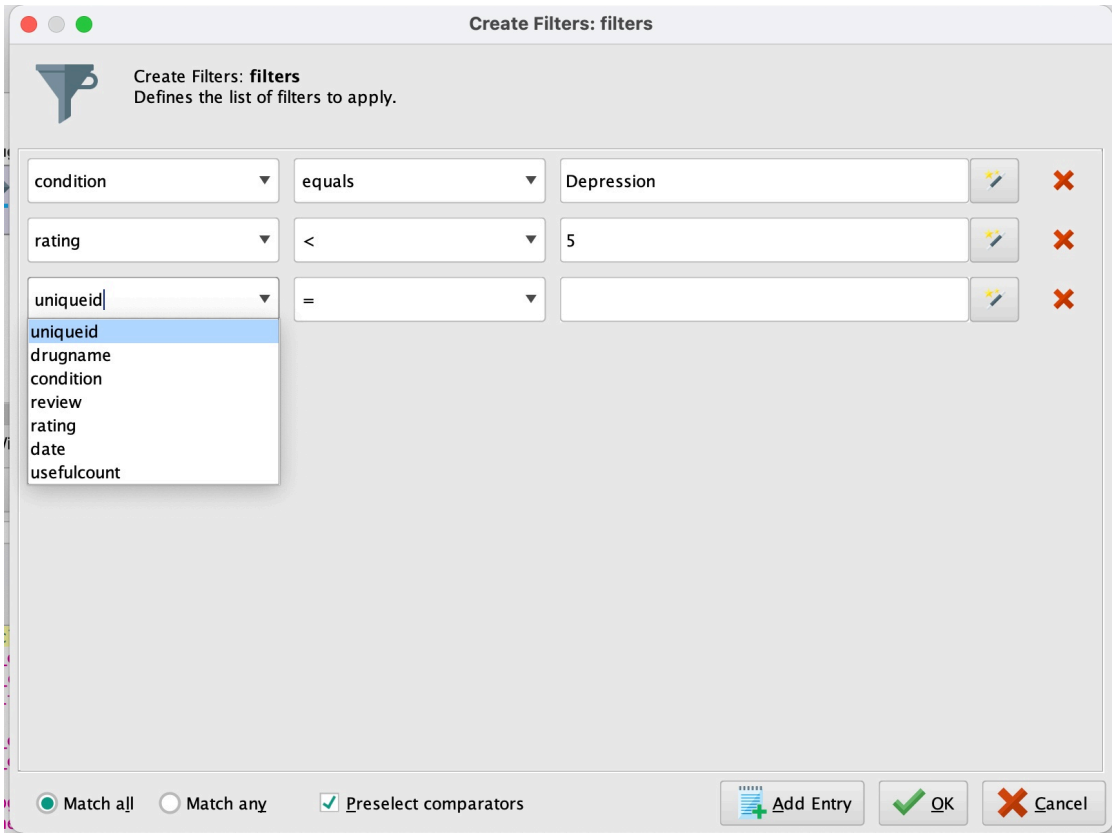
Πίνακας 1: Σύνοψη πειραμάτων

Παρακάτω παρουσιάζεται αναλυτικά η διαδικασία που ακολουθήθηκε για την LDA ανάλυση:
Αρχικά φορτώθηκαν στο Repository Entry του Rapid Miner σχόλια χρηστών (Example Set) που αφορούν έναν μεγάλο αριθμό παθήσεων και δραστικών φαρμακευτικών ουσιών σε μορφή αρχείου XML.



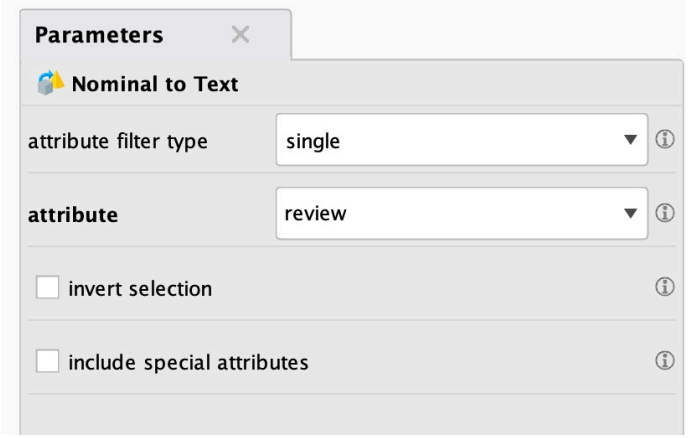
Εικόνα 4: Παράδειγμα LDA analysis process

Εν συνεχεία προστέθηκαν φίλτρα σε αυτά τα δεδομένα, στον operator Filter Examples. Ο συγκεκριμένος operator επιλέγει ποια παραδείγματα (σχόλια) από το σύνολο παραδειγμάτων θα κρατηθούν και ποια όχι. Ο operator επιστρέφει τα παραδείγματα που ταιριάζουν με την επιλεγμένη πάθηση. Οι παθήσεις επιλέγονται από τον χρήστη. Υπάρχουν επίσης αρκετές παθήσεις ως προεπιλογές στο πρόγραμμα. Παράλληλα μπορούν να χρησιμοποιηθούν φίλτρα και για άλλες παραμέτρους πέρα από το condition, όπως uniqueid, drugname, review, rating, date και usefulcount.



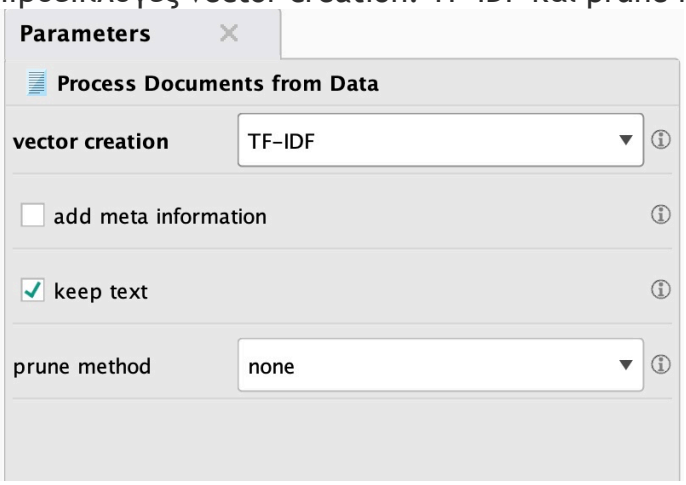
Εικόνα 5: Παράδειγμα επιλογή φίλτρων condition Depression και rating <5

Στην συνέχεια προστέθηκε ο operator Nominal to Text. Ο συγκεκριμένος operator αλλάζει τον τύπο των επιλεγμένων ονομαστικών χαρακτηριστικών σε κείμενο. Αντιστοιχίζει επίσης όλες τις τιμές αυτών των χαρακτηριστικών σε αντίστοιχες τιμές stringvalues. Στο attribute filter type χρησιμοποιήθηκε η επιλογή single. Αυτή η παράμετρος μας επιτρέπει να επιλέξουμε το φίλτρο επιλογής χαρακτηριστικών, την μέθοδο που θέλουμε να χρησιμοποιήσουμε για την επιλογή χαρακτηριστικών στα οποία θέλουμε να εφαρμόσετε τη μετατροπή ονομαστικής σε κείμενο. Η επιλογή single επιτρέπει την επιλογή ενός μόνο χαρακτηριστικού. Όταν επιλεγεί αυτή η επιλογή, μια άλλη παράμετρος (χαρακτηριστικό) γίνεται ορατή στον πίνακα Parameters. Άλλες διαθέσιμες επιλογές είναι: all, subset, regular_expression, value_type, block_type, no_missing_values and numeric value filter.
Επίσης στον operator επιλέχθηκε στο attribute: review.



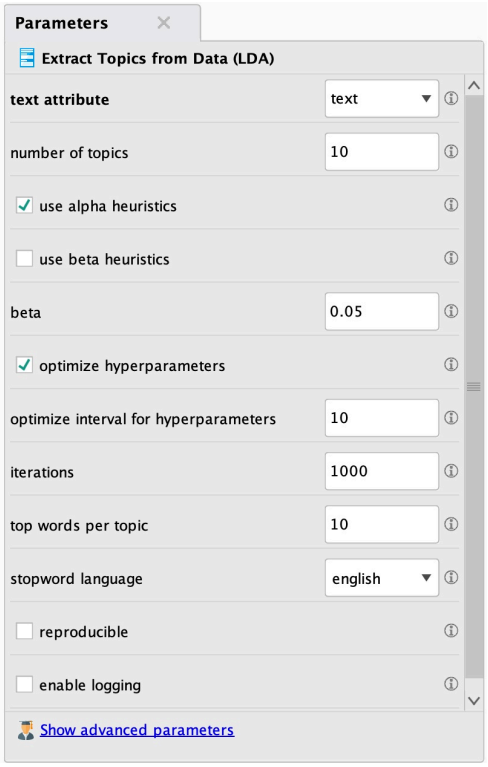
Εικόνα 6: Παράδειγμα Nominal to Text operator, επιλογή attribute filter type single and attribute review.

Έπειτα, στον operator Process Documents from Data κρατήθηκαν οι προεπιλογές vector creation: TF-IDF και prune method: none.



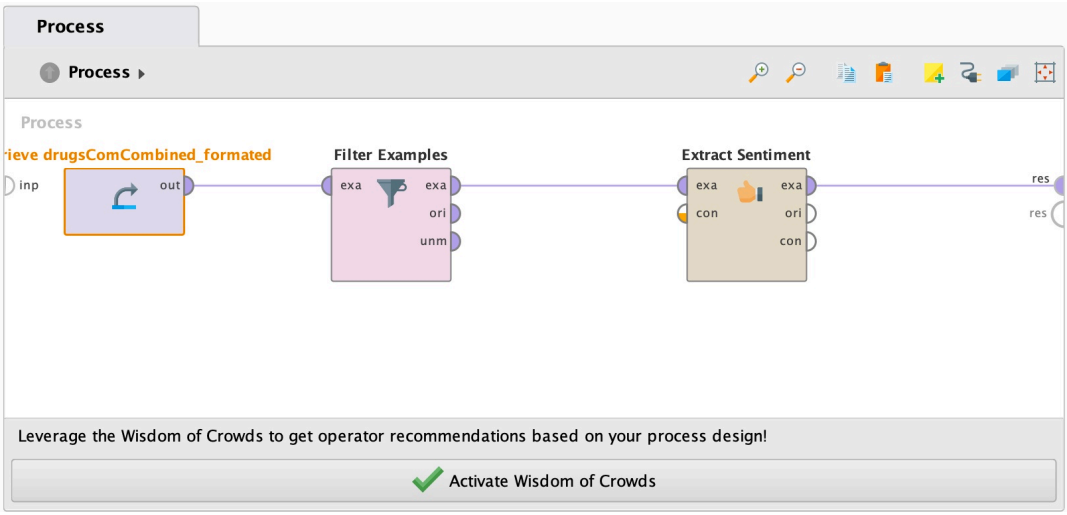
Εικόνα 7: Παράδειγμα Process Documents from Data operator

Τέλος ο operator Extract Topics from Data (LDA) βρίσκει topics χρησιμοποιώντας την μέθοδο LDA. Η μέθοδος LDA, όπως έχει αναφερθεί και παραπάνω είναι μια μέθοδος topic modeling, επιτρέποντας μας να προσδιορίζουμε θέματα σε έγγραφα. Αυτή η υλοποίηση του LDA χρησιμοποιεί το ParallelTopicModel της βιβλιοθήκης Mallet (πηγή: Newman, Asuncion, Smyth and Welling, Distributed Algorithms for Topic Models JMLR (2009)) με σχήμα δειγματοληψίας SparseLDA και δομή δεδομένων (πηγή: Yao, Mimno and McCallum, Efficient Methods for Topic Model Inference on Streaming Document Collections, KDD (2009)).
Επιλέχθηκαν οι επιλογές text attribute: text και number of topics: 10. Συνεπώς στα αποτελέσματα της ανάλυσης μας θα εμφανιστούν 10 topics. Επίσης το top words per topic επιλέχθηκε 10, συνεπώς σε κάθε topic θα εμφανιστούν οι top 10 λέξεις, με βάση την συχνότητα εμφάνισης τους στο topic. όλες οι υπόλοιπες επιλογές κρατήθηκαν by default και εμφανίζονται στην παρακάτω εικόνα.



Εικόνα 8: Παράδειγμα Extract Topics from Data (LDA)

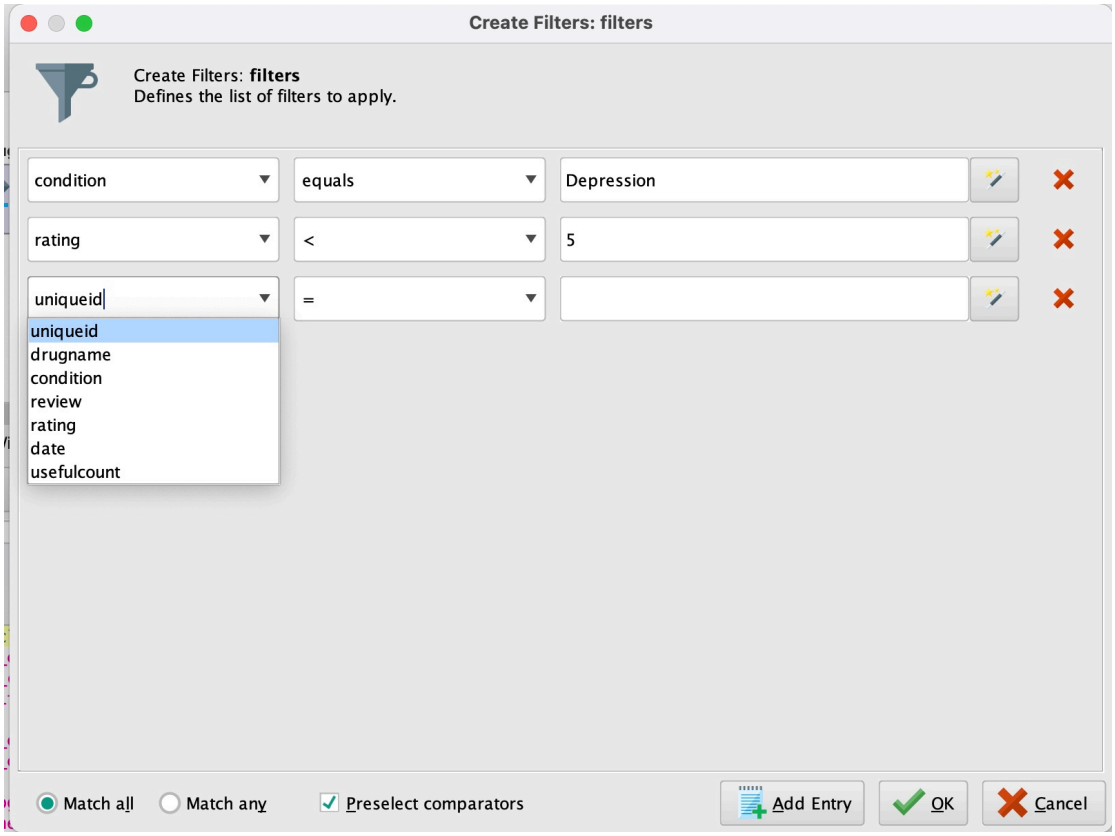
Παρακάτω παρουσιάζεται αναλυτικά η διαδικασία που ακολουθήθηκε για την Sentiment analysis:
Αρχικά φορτώθηκαν στο Repository Entry του Rapid Miner σχόλια χρηστών (Example Set) που αφορούν έναν μεγάλο αριθμό παθήσεων και δραστικών φαρμακευτικών ουσιών σε μορφή αρχείου XML.



Εικόνα 9: Παράδειγμα Sentiment analysis process

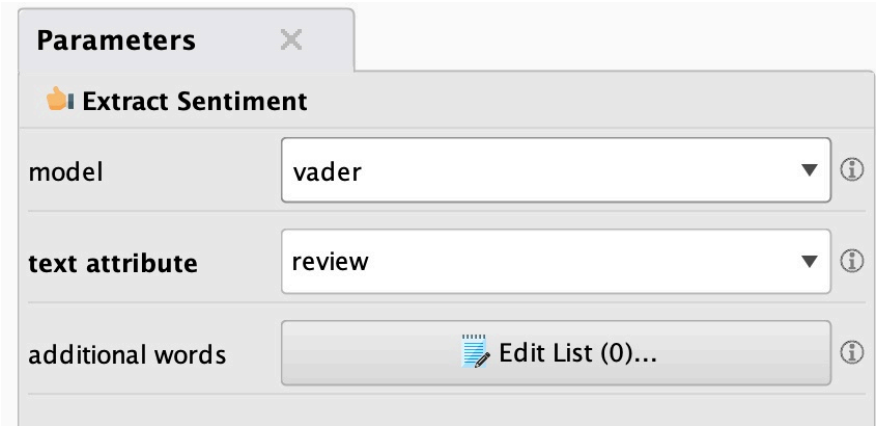
Εν συνεχεία προστέθηκαν φίλτρα σε αυτά τα δεδομένα, στον operator Filter Examples. Ο συγκεκριμένος operator επιλέγει ποια παραδείγματα (σχόλια) από το σύνολο παραδειγμάτων θα κρατηθούν και ποια όχι. Ο operator επιστρέφει τα παραδείγματα που ταιριάζουν με την επιλεγμένη πάθηση. Οι παθήσεις επιλέγονται από τον χρήστη. Υπάρχουν επίσης αρκετές παθήσεις

ως προεπιλογές στο πρόγραμμα. Παράλληλα μπορούν να χρησιμοποιηθούν φίλτρα και για άλλες παραμέτρους πέρα από το condition, όπως uniqueid, drugname, review, rating, date και usefulcount.



Εικόνα 10: Παράδειγμα επιλογής φίλτρων condition Depression και rating <5

Τέλος επιλέχθηκε ο operator Extract Sentiment.
Αυτός ο operator δημιουργεί μια βαθμολογία συναισθήματος εφαρμόζοντας είτε λεξικά συναισθήματος ανοιχτού κώδικα είτε ιδιόκτητες μεθόδους API σε ένα υπάρχον χαρακτηριστικό κειμένου. Υπάρχουν επιλογές για την έκθεση πρόσθετων αποτελεσμάτων ανάλογα με τη μέθοδο που έχει επιλεγεί.
Ως μοντέλο χρησιμοποιήθηκε το μοντέλο Vader και ως text attribute το review. Στο text attribute επιλέγεται η παράμετρος για την οποία επιθυμούμε να έχουμε score. Το μοντέλο VADER χρησιμοποιεί το λεξικό VADER (Valence Aware Dictionary and sEntiment Reasoner) και το συναίσθημα που βασίζεται σε κανόνες για να βαθμολογήσει το κείμενο. Το VADER είναι ειδικά προσαρμοσμένο στα συναισθήματα που εκφράζονται στα μέσα κοινωνικής δικτύωσης και παράγει αποτελέσματα με βάση ένα λεξικό λέξεων. Αυτός ο τελεστής υπολογίζει και στη συνέχεια εκθέτει το άθροισμα όλων των βαθμολογιών λέξεων συναισθήματος στο κείμενο. (Hutto, C.J. and Gilbert, E.E. (2014)).



Εικόνα 11: Παράδειγμα Extract Sentiment operator, model VADER

Παρακάτω παρατίθενται τα πειράματα που πραγματοποιήθηκαν με βάση τις μεθόδους και παραμέτρους που αναλύθηκαν.

- **Πείραμα 1: LDA Analysis for Birth Control (no filter for rating)- 10 topics**

Στο Πείραμα 1 πραγματοποιήθηκε LDA ανάλυση για σχόλια που αφορούν το birth control χωρίς να υπάρχει φίλτρο στο rating, συνεπώς είναι πιθανό να συναντήσουμε αρνητικά αλλά και θετικά σχόλια.
Παρακάτω θα παρουσιαστούν οι πίνακες με τα δεδομένα ανά topic. Η στήλη weight παρουσιάζει την συχνότητα με την οποία η κάθε λέξη εμφανίστηκε στο συγκεκριμένο topic.

Topic 0

topicId	word	weight
0,0	review	3283,0
0,0	read	2395,0
0,0	effect	1614,0
0,0	experi	1545,0
0,0	bodi	1495,0
0,0	side	1396,0
0,0	peopl	1042,0
0,0	take	1021,0
0,0	pill	916,0
0,0	neg	870,0

Αναλύοντας τις λέξεις από το Topic 0, συμπεραίνουμε ότι πιθανώς οι χρήστες σχολιάζουν πως έχουν διαβάσει reviews σύμφωνα με τα οποία οι άνθρωποι που έλαβαν αντισυλληπτικά φάρμακα αντιμετώπισαν ανεπιθύμητες παρενέργειες.

Topic 1

topicId	word	weight
1,0	gain	15470,0
1,0	weight	15432,0
1,0	month	4946,0
1,0	period	4934,0
1,0	year	4446,0
1,0	mood	4172,0
1,0	pound	3780,0
1,0	swing	3398,0
1,0	acn	3001,0
1,0	lost	2373,0

Αναλύοντας τις λέξεις από το Topic 1, συμπεραίνουμε ότι οι χρήστες οι οποίοι σχολίασαν έχουν χρησιμοποιήσει αντισυλληπτικά για συγκεκριμένα χρονικά διαστήματα (month/year) και έχουν αντιμετωπίσει παρενέργειες. Η υψηλότερη σε συχνότητα παρενέργεια είναι η αύξηση του σωματικού βάρους και έπειτα οι αλλαγές στην διάθεση. Τέλος πιθανώς αναφέρουν ότι εμφάνισαν μείωση και εξάλειψη της ακμής.

Topic 2

topicId	word	weight
2,0	period	18096,0
2,0	month	12190,0
2,0	dai	7673,0
2,0	week	7306,0
2,0	pill	6412,0
2,0	bleed	6313,0
2,0	cramp	6034,0
2,0	start	5993,0
2,0	spot	4539,0
2,0	take	4028,0

Αναλύοντας τις λέξεις από το Topic 2, συμπεραίνουμε ότι οι χρήστες πιθανώς σχολιάζουν πως παρουσίασαν αιμορραγία και κράμπες με την χρήση αντισυλληπτικών. Επίσης ίσως εμφάνισαν κηλίδες μεταξύ των περιόδων.

Topic 3

topicId	word	weight
3,0	control	15892,0
3,0	birth	15541,0
3,0	pill	7899,0
3,0	year	5108,0
3,0	take	5040,0
3,0	effect	4710,0
3,0	side	4142,0
3,0	switch	3975,0
3,0	period	3711,0
3,0	month	3446,0

Αναλύοντας τις λέξεις από το Topic 3, συμπεραίνουμε ότι οι χρήστες σχολιάζουν πως αντιμετώπισαν ως παρενέργεια αλλαγές στην εμφάνιση της περιόδου τους. Αυτό θα μπορούσε να σημαίνει ότι η περίοδος εμφανίστηκε αργότερα από το αναμενόμενο ή και να διακόπηκε για κάποιο διάστημα ή διαφορετικά μπορεί να είχαν άστατο κύκλο πριν την λήψη του φαρμάκου και αυτό τους βοήθησε στο να ομαλοποιηθεί ο κύκλος τους.

Topic 4

topicId	word	weight
4,0	pain	7890,0
4,0	insert	7889,0
4,0	cramp	7585,0
4,0	mirena	3035,0
4,0	skyla	2629,0
4,0	felt	2107,0
4,0	feel	1861,0
4,0	period	1629,0
4,0	spot	1544,0
4,0	took	1509,0

Αναλύοντας τις λέξεις από το Topic 4, συμπεραίνουμε ότι οι χρήστες σχολιάζουν πως με την χρήση ενδομήτριας αντισυλληπτικής συσκευής (IUD) σαν μέθοδο αντισύλληψης , όπως το Mirena και το Skyla τα οποία αναφέρονται, παρουσίασαν πόνο κατά την τοποθέτηση τους και εμφάνισαν κράμπες και κηλίδες.

Topic 5

topicId	word	weight
5,0	acn	5984,0
5,0	pill	5315,0
5,0	month	4070,0
5,0	breast	3327,0
5,0	skin	3124,0
5,0	take	2985,0
5,0	clear	2639,0
5,0	effect	2511,0
5,0	side	2493,0
5,0	start	2482,0

Αναλύοντας τις λέξεις από το Topic 5, συμπεραίνουμε ότι οι χρήστες σχολιάζουν πως είτε έκαναν χρήση αντισυλληπτικών φαρμάκων για να αντιμετωπίσουν την ακμή και αναφέρουν ότι καθάρισε το δέρμα τους, είτε παρουσίασαν ακμή από την χρήση του φαρμάκου. Πιθανώς αναφέρουν ότι παρουσίασαν κάποια ευαισθησία στο στήθος σαν παρενέργεια.

Topic 6

topicId	word	weight
6,0	pain	2602,0
6,0	blood	1270,0
6,0	start	1096,0
6,0	doctor	977,0
6,0	week	934,0
6,0	month	843,0
6,0	take	808,0
6,0	caus	806,0
6,0	cyst	780,0
6,0	hair	763,0

Αναλύοντας τις λέξεις από το Topic 6, υποθέτουμε ότι οι χρήστες σχολιάζουν πως είτε εμφάνισαν πόνο και αιμορραγία με την χρήση αντισυλληπτικών είτε ο γιατρός τους συνταγογράφησε τα συγκεκριμένα φάρμακα για την αντιμετώπιση των προαναφερθέντων συμπτωμάτων. Επίσης πιθανόν να εμφανίστηκαν παρενέργειες που σχετίζονται με κύστες και τριχοφυΐα.

Topic 7

topicId	word	weight
7,0	feel	4961,0
7,0	depress	4838,0
7,0	month	3343,0
7,0	mood	3272,0
7,0	pill	2914,0
7,0	anxieti	2682,0
7,0	take	2592,0
7,0	swing	2436,0
7,0	time	2391,0
7,0	start	2280,0

Αναλύοντας τις λέξεις από το Topic 7, συμπεραίνουμε ότι οι χρήστες σχολιάζουν πως εμφάνισαν αλλαγές στην διάθεση στους, με μεγαλύτερη συχνότητα εμφάνισης το αίσθημα της κατάθλιψης και έπειτα το άγχος.

Topic 8

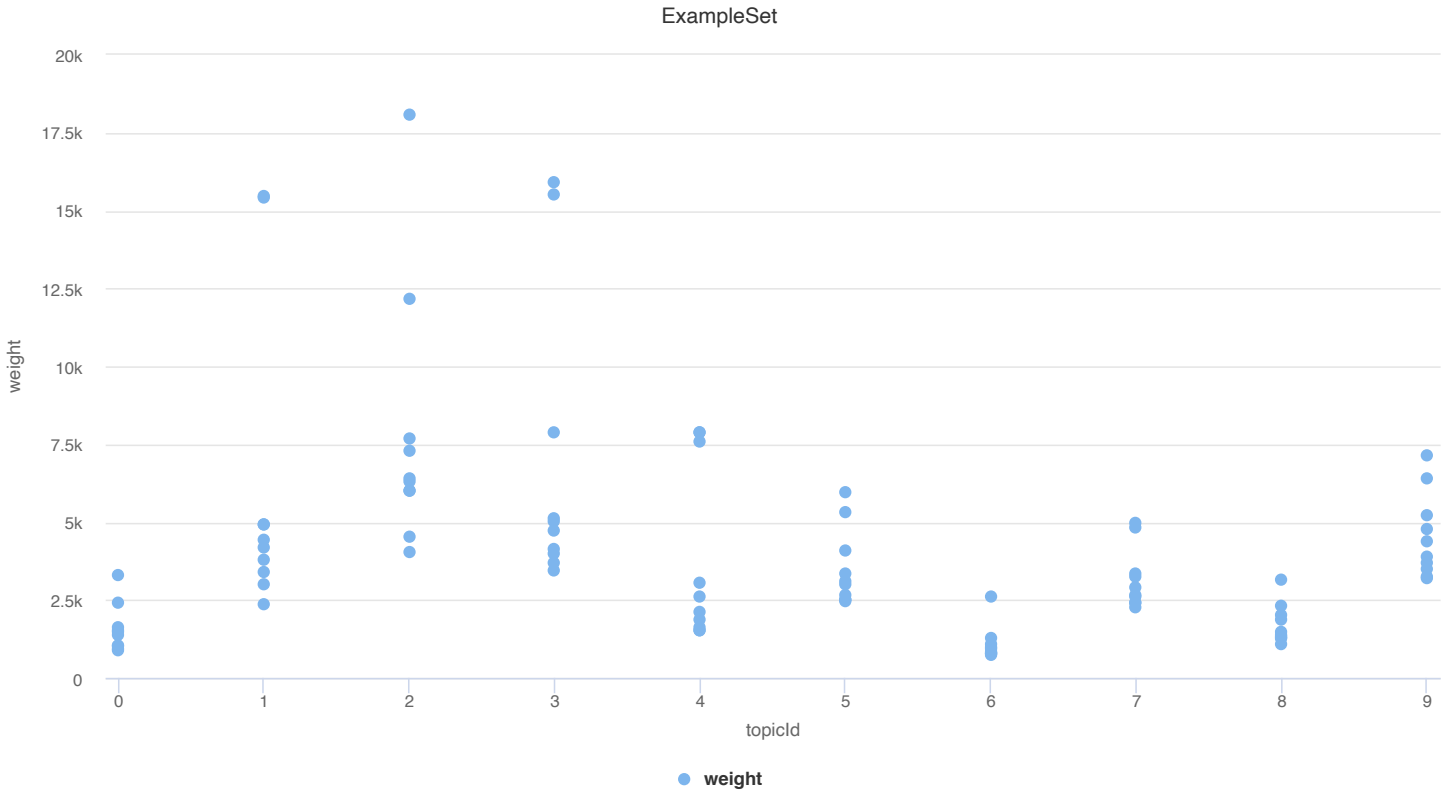
topicId	word	weight
8,0	shot	3171,0
8,0	patch	2338,0
8,0	depo	2010,0
8,0	nuvar	1873,0
8,0	year	1475,0
8,0	effect	1384,0
8,0	ring	1309,0
8,0	love	1288,0
8,0	side	1276,0
8,0	take	1066,0

Στο Topic 8 εμφανίζονται λέξεις που σχετίζονται με διάφορες μεθόδους αντισύλληψης όπως: the patch, the ring and the shot (depo-provera), ωστόσο δεν είναι ξεκάθαρη κάποια αναφορά σε θετική ή αρνητική εμπειρία, απλώς υπάρχει αναφορά σε παρενέργειες.

Topic 9

topicId	word	weight
9,0	period	7142,0
9,0	month	6383,0
9,0	bleed	5244,0
9,0	year	4791,0
9,0	nexplanon	4363,0
9,0	implant	3871,0
9,0	week	3702,0
9,0	remov	3520,0
9,0	insert	3250,0
9,0	implanon	3184,0

Στο Topic 9 εμφανίζονται λέξεις που σχετίζονται με εμφυτεύματα αντισύλληψης όπως το Implanon και το explanon και συμπεραίνουμε πως υπάρχουν σχόλια που αφορούν την εισαγωγή και αφαίρεση τους και πιθανώς το χρονικό διάστημα για το οποίο το κάθε ένα διατηρείται. Επίσης υπάρχει αναφορά σε αιμορραγία.
Παρακάτω παρουσιάζεται ένα διάγραμμα όπου εμφανίζονται τα topics στον άξονα x και η συχνότητα εμφάνισης των λέξεων για το κάθε topic στον άξονα y.



Γράφημα 1: Πείραμα 1-LDA Analysis for Birth Control (no filter for rating)-10 topics

Αναλύοντας το παραπάνω γράφημα μπορούμε να συμπεράνουμε ότι τα Topics 0,5,6,7,8,9 είναι πιο συνεκτικά, δηλαδή όλες οι λέξεις στα συγκεκριμένα έχουν παρόμοια συχνότητα εμφάνισης και συνεπώς τα συμπεράσματα που έχουμε εξάγει από αυτά είναι περισσότερο αξιόπιστα. Αντιθέτως στα topics 1,2,3,4 υπάρχουν σχόλια με πολύ μεγαλύτερη συχνότητα εμφάνισης από άλλα, οπότε είναι πιθανό να μην σχετίζονται και ως αποτέλεσμα τα συμπεράσματα που εξάγουμε είναι λιγότερο αξιόπιστα.

Δεδομένου ότι στο πείραμα 1 δεν χρησιμοποιήσαμε κάποιο φίλτρο στο rating των σχολίων, είναι δυσκολότερο να εξάγουμε σαφή συμπεράσματα καθώς δεν μπορούμε με σιγουριά να υποστηρίξουμε αν κάποιες λέξεις αναφέρονται με θετικό ή αρνητικό πρόσημο καθώς επίσης και αν αναφέρονται ως παρενέργειες οι οποίες δημιουργήθηκαν από την λήψη των φαρμάκων ή εάν τα φάρμακα λήφθηκαν για την αντιμετώπιση αυτών των συμπτωμάτων. Γι' αυτό το λόγο στο Πείραμα 2 θα κάνουμε την ίδια ανάλυση με χρήση φίλτρου στο rating, ώστε να εμφανίζονται μόνο τα αρνητικά σχόλια (<5) και θα αναλύσουμε ξανά τα δεδομένα που θα προκύψουν.

- **Πείραμα 2: LDA Analysis for Birth Control (filter for rating<5) - 10 topics**

Στο Πείραμα 2 πραγματοποιήθηκε LDA ανάλυση για σχόλια που αφορούν το birth control με φίλτρο στο rating <5 (κλίμακα 1-10), συνεπώς υποθέτουμε πως θα συναντήσουμε κατά κύριο λόγο αρνητικά σχόλια.

Topic 0

topicId	word	weight
0,0	pill	3163,0
0,0	month	1974,0
0,0	take	1636,0
0,0	period	1394,0
0,0	cramp	1235,0
0,0	effect	1221,0
0,0	side	1153,0
0,0	start	1084,0
0,0	breast	990,0
0,0	feel	943,0

Αναλύοντας τις λέξεις από το Topic 0, συμπεραίνουμε ότι οι χρήστες αναφέρουν ως ανεπιθύμητες ενέργειες τις κράμπες περιόδου και την ευαισθησία στο στήθος.

Topic 1

topicId	word	weight
1,0	period	5583,0
1,0	month	4244,0
1,0	bleed	3092,0
1,0	week	3071,0
1,0	dai	2338,0
1,0	start	2091,0
1,0	pill	1816,0
1,0	spot	1609,0
1,0	stop	1587,0
1,0	cramp	1536,0

Αναλύοντας τις λέξεις από το Topic 1, συμπεραίνουμε ότι οι χρήστες αναφέρουν ως ανεπιθύμητες ενέργειες την αιμορραγία, τις κηλίδες και τις κράμπες περιόδου.

Topic 2

topicId	word	weight
2,0	acn	1617,0
2,0	switch	1363,0
2,0	pill	1287,0
2,0	skin	987,0
2,0	month	730,0
2,0	face	672,0
2,0	take	592,0
2,0	year	522,0
2,0	clear	462,0
2,0	start	452,0

Αναλύοντας τις λέξεις από το Topic 2, συμπεραίνουμε ότι οι χρήστες αναφέρουν πως παρατήρησαν αλλαγές στην εμφάνιση του δέρματος και πιθανώς εμφάνισαν ακμή.

Topic 3

topicId	word	weight
3,0	depress	2627,0
3,0	feel	1888,0
3,0	mood	1637,0
3,0	anxieti	1508,0
3,0	pill	1457,0
3,0	month	1366,0
3,0	swing	1189,0
3,0	take	1055,0
3,0	start	942,0
3,0	thing	898,0

Αναλύοντας τις λέξεις από το Topic 3, συμπεραίνουμε ότι οι χρήστες αναφέρουν πως εμφάνισαν εναλλαγές στην διάθεση τους και βίωσαν αίσθημα κατάθλιψης και άγχους.

Topic 4

topicId	word	weight
4,0	shot	1170,0
4,0	effect	805,0
4,0	side	766,0
4,0	month	713,0
4,0	depo	560,0
4,0	infect	440,0
4,0	year	369,0
4,0	hair	349,0
4,0	ring	338,0
4,0	yeast	312,0

Αναλύοντας τις λέξεις από το Topic 4, συμπεραίνουμε ότι οι χρήστες που χρησιμοποίησαν αντισυλληπτικές μεθόδους (shot, depo και ring) εμφάνισαν μύκητες και προβλήματα που αφορούν την τριχοφυΐα τους.

Topic 5

topicId	word	weight
5,0	feel	832,0
5,0	start	653,0
5,0	take	636,0
5,0	pain	592,0
5,0	week	558,0
5,0	headach	471,0
5,0	night	446,0
5,0	felt	434,0
5,0	patch	403,0
5,0	dai	356,0

Αναλύοντας τις λέξεις από το Topic 5, συμπεραίνουμε ότι οι χρήστες που χρησιμοποίησαν patch ως μέσο αντισύλληψης εμφάνισαν πονοκέφαλο πιθανώς κατά την διάρκεια της νύχτας.

Topic 6

topicId	word	weight
6,0	gain	5160,0
6,0	weight	4448,0
6,0	month	1849,0
6,0	pound	1663,0
6,0	year	1375,0
6,0	period	1282,0
6,0	acn	990,0
6,0	mood	912,0
6,0	swing	802,0
6,0	lose	749,0

Αναλύοντας τις λέξεις από το Topic 6, συμπεραίνουμε ότι οι χρήστες αναφέρουν πως με την χρήση των φαρμάκων αυξήθηκε το σωματικό τους βάρος, εμφάνισαν ακμή, είχαν εναλλαγές στην διάθεση τους και πιθανώς διαταράχθηκε η περίοδος τους.

Topic 7

topicId	word	weight
7,0	control	6255,0
7,0	birth	6020,0
7,0	take	1107,0
7,0	pill	1037,0
7,0	effect	868,0
7,0	month	776,0
7,0	doctor	761,0
7,0	year	758,0
7,0	side	729,0
7,0	worst	588,0

Αναλύοντας τις λέξεις από το Topic 7, συμπεραίνουμε ότι οι χρήστες αναφέρουν πως αντιμετώπισαν σοβαρές παρενέργειες χρησιμοποιώντας φάρμακα για αντισύλληψη.

Topic 8

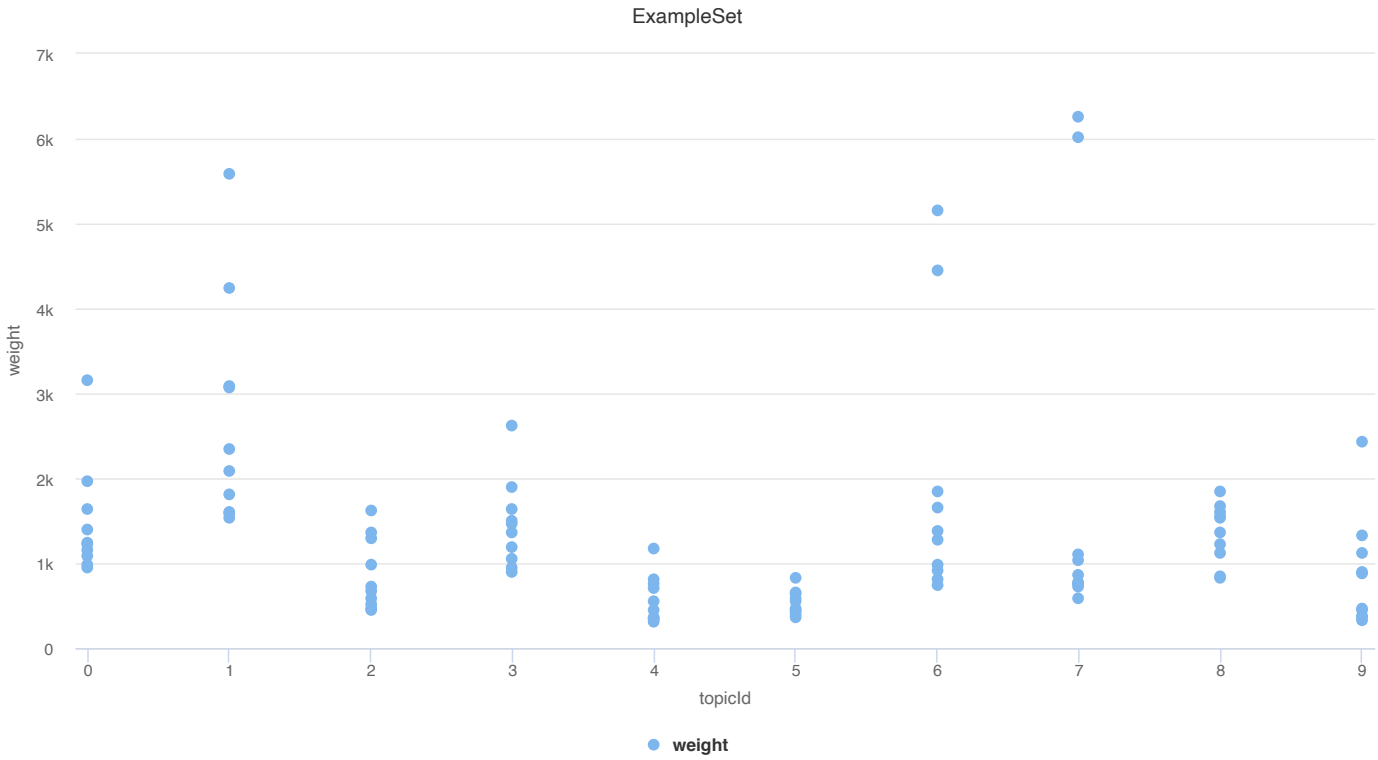
topicId	word	weight
8,0	bleed	1848,0
8,0	month	1669,0
8,0	remov	1599,0
8,0	period	1556,0
8,0	nexplanon	1535,0
8,0	implant	1369,0
8,0	year	1220,0
8,0	get	1115,0
8,0	week	847,0
8,0	time	834,0

Αναλύοντας τις λέξεις από το Topic 8, συμπεραίνουμε ότι οι χρήστες αναφέρουν πως με την χρήση του εμφυτεύματος Nexplanon αντιμετώπισαν αιμορραγία και αδυναμία.

Topic 9

topicId	word	weight
9,0	pain	2425,0
9,0	insert	1323,0
9,0	cramp	1116,0
9,0	remov	891,0
9,0	mirena	878,0
9,0	skyla	473,0
9,0	year	451,0
9,0	went	372,0
9,0	doctor	365,0
9,0	month	332,0

Αναλύοντας τις λέξεις από το Topic 9, συμπεραίνουμε ότι οι χρήστες σχολιάζουν πως με την χρήση ενδομήτριας αντισυλληπτικής συσκευής (IUD) σαν μέθοδο αντισύλληψης , όπως το Mirena και το Skyla τα οποία αναφέρονται, παρουσίασαν πόνο κατά την τοποθέτηση τους, εμφάνισαν κράμπες και χρειάστηκε να επισκεφθούν τον γιατρό τους.
Παρακάτω παρουσιάζεται ένα διάγραμμα όπου εμφανίζονται τα topics στον άξονα x και η συχνότητα εμφάνισης των λέξεων για το κάθε topic στον άξονα y.



- Γράφημα 2: Πείραμα 2- LDA Analysis for Birth Control (filter for rating<5) - 10 topics

Αναλύοντας το παραπάνω γράφημα μπορούμε να συμπεράνουμε ότι τα Topics 0,2,3,4,5,8,9 είναι πιο συνεκτικά, δηλαδή όλες οι λέξεις στα συγκεκριμένα έχουν παρόμοια συχνότητα εμφάνισης και συνεπώς τα συμπεράσματα που έχουμε εξάγει από αυτά είναι περισσότερο αξιόπιστα. Αντιθέτως στα topics 1,6,7 υπάρχουν σχόλια με πολύ μεγαλύτερη συχνότητα εμφάνισης από άλλα, οπότε είναι πιθανό να μην σχετίζονται και ως αποτέλεσμα τα συμπεράσματα που εξάγουμε είναι λιγότερο αξιόπιστα.

Παράλληλα παρατηρούμε ότι τα topics 1,6 και 7 έχουν πολύ μεγαλύτερη συχνότητα εμφάνισης στις λέξεις τους. Αυτό μας οδηγεί στο συμπέρασμα ότι οι αναφερόμενες περενέργειες: αιμορραγία, κηλίδες, κράμπες περιόδου, αύξηση σωματικού βάρους, ακμή, εναλλαγές στην διάθεση και διατάραξη περιόδου, είναι τα οι πιο κοινά και συχνά εμφανιζόμενες παρενέργειες των φαρμάκων αντισύλληψης.

Συμπεραίνουμε πως στο Πείραμα 2, όπου φιλτράραμε μόνο τα αρνητικά σχόλια, ήταν πολύ πιο εύκολο να αναλύσουμε τα αποτελέσματα και να εξάγουμε σαφή συμπεράσματα για τις παρενέργειες των φαρμάκων. Επίσης παρατηρήθηκε πως δεν υπήρχαν αρνητικά σχόλια που να αφορούν άλλους παράγοντες πέρα από τις παρενέργιες όπως για παράδειγμα η τιμή, η διαθεσιμότητα ή κάποιος ποιοτικό πρόβλημα των σκευασμάτων.

- **<u>Πείραμα 3: LDA Analysis for Depression (no filter for rating)- 10 topics</u>**

Στο Πείραμα 3 πραγματοποιήθηκε LDA ανάλυση για σχόλια που αφορούν τα αντικαταθλιπτικά φάρμακα χωρίς να υπάρχει φίλτρο στο rating, συνεπώς είναι πιθανό να συναντήσουμε αρνητικά αλλά και θετικά σχόλια.

Topic 0

topicId	word	weight
0,0	pain	812,0
0,0	nausea	462,0
0,0	week	438,0
0,0	headach	311,0
0,0	dai	307,0
0,0	heart	300,0
0,0	blood	294,0
0,0	sever	269,0
0,0	pressur	250,0
0,0	muscl	243,0

Αναλύοντας τις λέξεις από το Topic 0, συμπεραίνουμε ότι οι χρήστες σχολιάζουν πως με την χρήση αντικαταθλιπτικών βίωσαν πόνο, ναυτία, πονοκέφαλο, αυξημένη πίεση και προβλήματα με την καρδιά τους αλλά και μυϊκά θέματα.

Topic 1

topicId	word	weight
1,0	life	2532,0
1,0	feel	1162,0
1,0	year	1082,0
1,0	depress	934,0
1,0	chang	563,0
1,0	medic	524,0
1,0	thought	506,0
1,0	happi	464,0
1,0	thing	437,0
1,0	anxieti	428,0

Αναλύοντας τις λέξεις από το Topic 1, συμπεραίνουμε ότι οι χρήστες χρησιμοποιώντας αντικαταθλιπτικά πιθανώς αντιμετώπισαν αισθήματα κατάθλιψης και άγχους, ίσως χρειάστηκε να αλλάξουν φάρμακο. Μια διαφορετική ερμηνεία θα μπορούσε να είναι ότι τα αντικαταθλιπτικά τους βοήθησαν να ξεπεράσουν τα συναισθήματα κατάθλιψης και άγχους και τώρα αισθάνονται χαρούμενοι.

Topic 2

topicId	word	weight
2,0	feel	4664,0
2,0	take	3599,0
2,0	week	3099,0
2,0	start	2985,0
2,0	work	2412,0
2,0	felt	2034,0
2,0	depress	1706,0
2,0	month	1650,0
2,0	time	1542,0
2,0	doctor	1364,0

Αναλύοντας τις λέξεις από το Topic 2, συμπεραίνουμε ότι οι χρήστες χρησιμοποιώντας αντικαταθλιπτικά πιθανώς αντιμετώπισαν αισθήματα κατάθλιψης και αυτό ίσως χρειάστηκε να επισκεφθούνε τον γιατρό τους. Μια διαφορετική ερμηνεία είναι πως με την λήψη των φαρμάκων που συνταγογραφήθηκαν από τον γιατρό τους ξεπέρασαν τα συναισθήματα της κατάθλιψης.

Topic 3

topicId	word	weight
3,0	drug	766,0
3,0	withdraw	690,0
3,0	effexor	622,0
3,0	take	602,0
3,0	dose	578,0
3,0	brain	404,0
3,0	year	400,0
3,0	symptom	398,0
3,0	miss	369,0
3,0	stop	321,0

Αναλύοντας τις λέξεις από το Topic 3, δεν μπορούμε να διεξάγουμε κάποιο ξεκάθαρο συμπέρασμα για τα σχόλια των χρηστών. Αναφέρονται σε δόσεις φαρμάκου οι οποίες ίσως χάθηκαν ή διακόπηκαν.

Topic 4

topicId	word	weight
4,0	depress	4177,0
4,0	anxieti	2433,0
4,0	year	2044,0
4,0	help	1261,0
4,0	work	1179,0
4,0	take	1108,0
4,0	life	995,0
4,0	medic	949,0
4,0	panic	933,0
4,0	attack	929,0

Αναλύοντας τις λέξεις από το Topic 4, συμπεραίνουμε ότι οι χρήστες πιθανώς βοηθήθηκαν να ξεπεράσουν τα συναισθήματα κατάθλιψης, άγχους και κρίσεων πανικού.

Topic 5

topicId	word	weight
5,0	feel	1555,0
5,0	week	1055,0
5,0	mood	1045,0
5,0	notic	904,0
5,0	depress	797,0
5,0	improv	639,0
5,0	month	583,0
5,0	increas	579,0
5,0	anxieti	557,0
5,0	energi	547,0

Αναλύοντας τις λέξεις από το Topic 5, συμπεραίνουμε ότι οι χρήστες πιθανώς αισθάνθηκαν αδυναμία με την χρήση των φαρμάκων και ίσως βελτιώθηκε το αίσθημα της κατάθλιψης αλλά αυξήθηκε το συναίσθημα του άγχους.

Topic 6

topicId	word	weight
6,0	sleep	1811,0
6,0	night	1461,0
6,0	take	1265,0
6,0	morn	634,0
6,0	hour	605,0
6,0	wake	549,0
6,0	feel	543,0
6,0	time	414,0
6,0	help	390,0
6,0	fall	332,0

Αναλύοντας τις λέξεις από το Topic 6, συμπεραίνουμε ότι οι χρήστες αντιμετώπισαν ως παρενέργεια προβλήματα με τον ύπνο όπως αυπνία.

Topic 7

topicId	word	weight
7,0	work	323,0
7,0	gener	257,0
7,0	drug	230,0
7,0	insur	223,0
7,0	year	203,0
7,0	deplin	190,0
7,0	brand	179,0
7,0	name	158,0
7,0	depress	143,0
7,0	ssri	141,0

Αναλύοντας τις λέξεις από το Topic 7, δεν μπορούμε να διεξάγουμε κάποιο ξεκάθαρο συμπέρασμα για τα σχόλια των χρηστών. Αναφέρονται σε κάποια συγκεκριμένα αντικαταθλιπτικά φάρμακα.

Topic 8

topicId	word	weight
8,0	effect	6566,0
8,0	side	5557,0
8,0	depress	1842,0
8,0	work	1175,0
8,0	medic	884,0
8,0	week	855,0
8,0	anxieti	822,0
8,0	year	780,0
8,0	take	732,0
8,0	tri	728,0

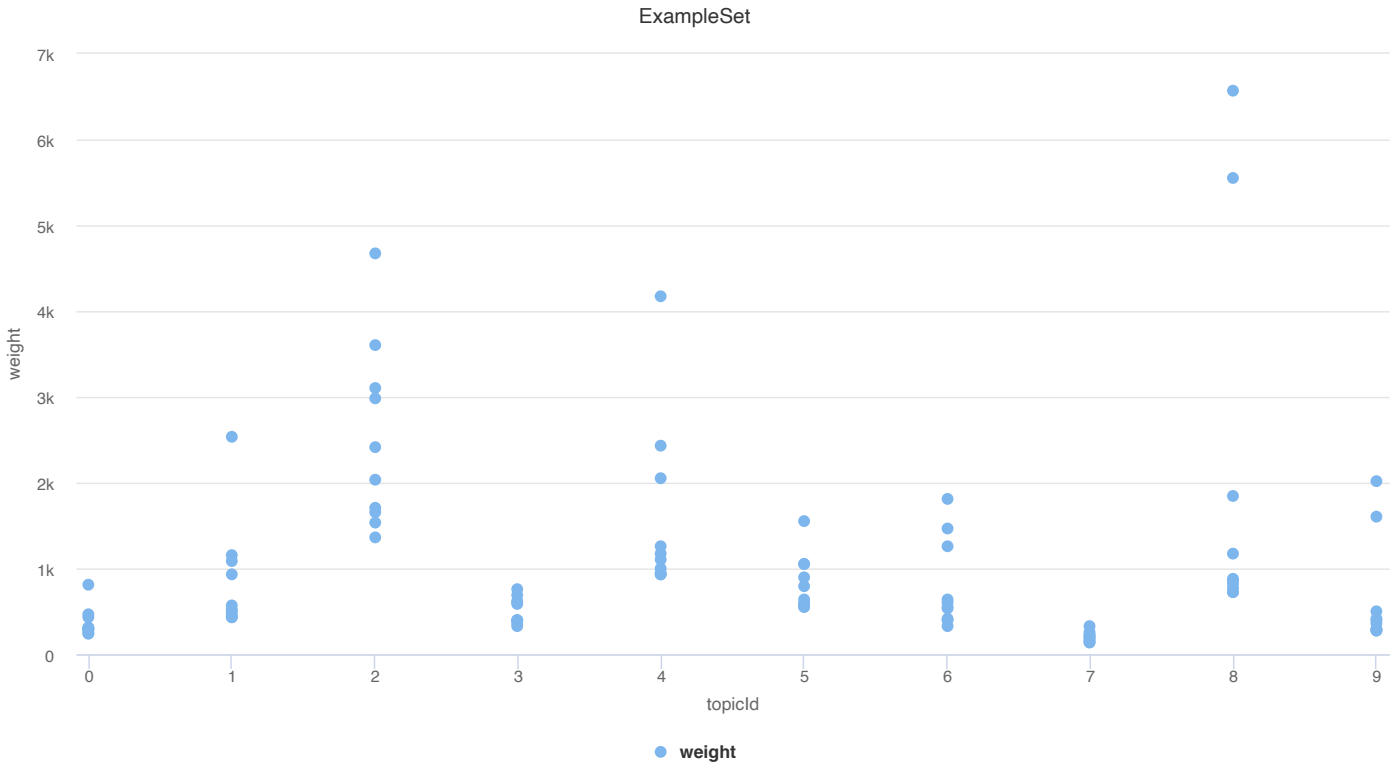
Αναλύοντας τις λέξεις από το Topic 8, συμπεραίνουμε ότι οι χρήστες χρησιμοποιώντας αντικαταθλιπτικά πιθανώς αντιμετώπισαν περενέργειες όπως κατάθλιψη και άγχος.

Topic 9

topicId	word	weight
9,0	weight	2020,0
9,0	gain	1606,0
9,0	lost	501,0
9,0	month	412,0
9,0	pound	392,0
9,0	depress	359,0
9,0	lose	301,0
9,0	wellbutrin	300,0
9,0	eat	279,0
9,0	year	271,0

Αναλύοντας τις λέξεις από το Topic 9, συμπεραίνουμε ότι οι χρήστες χρησιμοποιώντας αντικαταθλιπτικά πιθανώς αντιμετώπισαν διαταραχές στο σωματικό τους βάρος.

Παρακάτω παρουσιάζεται ένα διάγραμμα όπου εμφανίζονται τα topics στον άξονα x και η συχνότητα εμφάνισης των λέξεων για το κάθε topic στον άξονα y.



Γράφημα 3: Πείραμα 3-LDA Analysis for Depression (no filter for rating)- 10 topics

Αναλύοντας το παραπάνω γράφημα μπορούμε να συμπεράνουμε ότι τα Topics 0,3,5,6,7 είναι πιο συνεκτικά, δηλαδή όλες οι λέξεις στα συγκεκριμένα έχουν παρόμοια συχνότητα εμφάνισης και συνεπώς τα συμπεράσματα που έχουμε εξάγει από αυτά είναι περισσότερο αξιόπιστα. Αντιθέτως στα topics 1,2,4,8,9 υπάρχουν σχόλια με πολύ μεγαλύτερη συχνότητα εμφάνισης από άλλα, οπότε είναι πιθανό να μην σχετίζονται και ως αποτέλεσμα τα συμπεράσματα που εξάγουμε είναι λιγότερο αξιόπιστα.

Δεδομένου ότι στο πείραμα 3 δεν χρησιμοποιήσαμε κάποιο φίλτρο στο rating των σχολίων, είναι δυσκολότερο να εξάγουμε σαφή συμπεράσματα καθώς δεν μπορούμε με σιγουριά να υποστηρίξουμε αν κάποιες λέξεις αναφέρονται με θετικό ή αρνητικό πρόσημο καθώς επίσης και αν αναφέρονται ως παρενέργειες οι οποίες δημιουργήθηκαν από την λήψη των φαρμάκων ή εάν τα φάρμακα λήφθηκαν για την αντιμετώπιση αυτών των συμπτωμάτων. Γι' αυτό το λόγο στο Πείραμα 4 θα κάνουμε την ίδια ανάλυση με χρήση φίλτρου στο rating, ώστε να εμφανίζονται μόνο τα αρνητικά σχόλια (<5) και θα αναλύσουμε ξανά τα δεδομένα που θα προκύψουν.

- **<u>Πείραμα 4: LDA Analysis for Depression (filter for rating<5) - 10 topics</u>**

Στο Πείραμα 4 πραγματοποιήθηκε LDA ανάλυση για σχόλια που αφορούν την πάθηση της κατάθλιψης με φίλτρο στο rating <5 (κλίμακα 1-10), συνεπώς υποθέτουμε πως θα συναντήσουμε κατά κύριο λόγο αρνητικά σχόλια.

Topic 0

topicId	word	weight
0,0	medic	100,0
0,0	drug	93,0
0,0	thought	57,0
0,0	suicid	37,0
0,0	stop	36,0
0,0	psychiatrist	35,0
0,0	get	33,0
0,0	doctor	33,0
0,0	make	32,0
0,0	med	30,0

Αναλύοντας τις λέξεις από το Topic 0, συμπεραίνουμε ότι οι χρήστες χρησιμοποιώντας αντικαταθλιπτικά αντιμετώπισαν αυτοκτονικές τάσεις .

Topic 1

topicId	word	weight
1,0	gain	297,0
1,0	weight	279,0
1,0	month	146,0
1,0	take	95,0
1,0	depress	92,0
1,0	year	91,0
1,0	medic	66,0
1,0	week	62,0
1,0	gener	61,0
1,0	lexapro	58,0

Αναλύοντας τις λέξεις από το Topic 1, συμπεραίνουμε ότι οι ασθενείς αύξησαν το σωματικό τους βάρος και αντιμετώπισαν αισθήματα κατάθλιψης.

Topic 2

topicId	word	weight
2,0	effect	695,0
2,0	side	616,0
2,0	week	258,0
2,0	depress	197,0
2,0	nausea	163,0
2,0	month	159,0
2,0	experienc	120,0
2,0	start	117,0
2,0	headach	110,0
2,0	medic	104,0

Αναλύοντας τις λέξεις από το Topic 2, συμπεραίνουμε ότι οι ασθενείς αντιμετώπισαν παρενέργειες όπως κατάθλιψη, ναυτία και πονοκεφάλους.

Topic 3

topicId	word	weight
3,0	drug	372,0
3,0	withdraw	191,0
3,0	year	174,0
3,0	month	154,0
3,0	effexor	139,0
3,0	brain	130,0
3,0	life	112,0
3,0	zap	107,0
3,0	worst	107,0
3,0	medic	105,0

Αναλύοντας τις λέξεις από το Topic 3, συμπεραίνουμε ότι οι ασθενείς σχολιάζουν αρνητικά κάποιο συγκεκριμένο αντικαταθλιπτικό φάρμακο, το οποίο πιθανώς επηρέασε την εγκεφαλική τους λειτουργία.

Topic 4

topicId	word	weight
4,0	pain	171,0
4,0	muscl	73,0
4,0	start	55,0
4,0	sever	55,0
4,0	help	52,0
4,0	dai	49,0
4,0	cymbalta	39,0
4,0	depress	39,0
4,0	leg	36,0
4,0	take	36,0

Αναλύοντας τις λέξεις από το Topic 4, συμπεραίνουμε ότι οι ασθενείς αντιμετώπισαν σοβαρούς μυϊκούς πόνους και κατάθλιψη ως παρενέργειες.

Topic 5

topicId	word	weight
5,0	attack	192,0
5,0	panic	174,0
5,0	heart	132,0
5,0	anxieti	130,0
5,0	depress	103,0
5,0	took	100,0
5,0	blood	94,0
5,0	dai	93,0
5,0	pressur	91,0
5,0	week	81,0

Αναλύοντας τις λέξεις από το Topic 5, συμπεραίνουμε ότι οι ασθενείς αντιμετώπισαν κρίσεις πανικού, αύξηση την αρτηριακής τους πίεσης και άγχος και προβλήματα στην καρδιά τους ως παρενέργειες στα αντικαταθλιπτικά φάρμακα που έλαβαν.

Topic 6

topicId	word	weight
6,0	feel	95,0
6,0	medic	85,0
6,0	thing	57,0
6,0	becam	55,0
6,0	time	55,0
6,0	life	53,0
6,0	person	52,0
6,0	help	52,0
6,0	chang	50,0
6,0	think	50,0

Δεν μπορούμε να εξάγουμε κάποιο ξεκάθαρο συμπέρασμα αναλύοντας τις λέξεις από το Topic 6.

Topic 7

topicId	word	weight
7,0	night	288,0
7,0	sleep	259,0
7,0	feel	138,0
7,0	hour	128,0
7,0	take	105,0
7,0	took	97,0
7,0	time	91,0
7,0	felt	88,0
7,0	asleep	73,0
7,0	morn	71,0

Αναλύοντας τις λέξεις από το Topic 7, συμπεραίνουμε ότι οι ασθενείς αντιμετώπισαν προβλήματα αυπνίας με την λήψη αντικαταθλιπτικών φαρμάκων.

Topic 8

topicId	word	weight
8,0	take	626,0
8,0	feel	519,0
8,0	start	242,0
8,0	dai	233,0
8,0	doctor	219,0
8,0	nausea	194,0
8,0	felt	193,0
8,0	time	189,0
8,0	week	175,0
8,0	headach	158,0

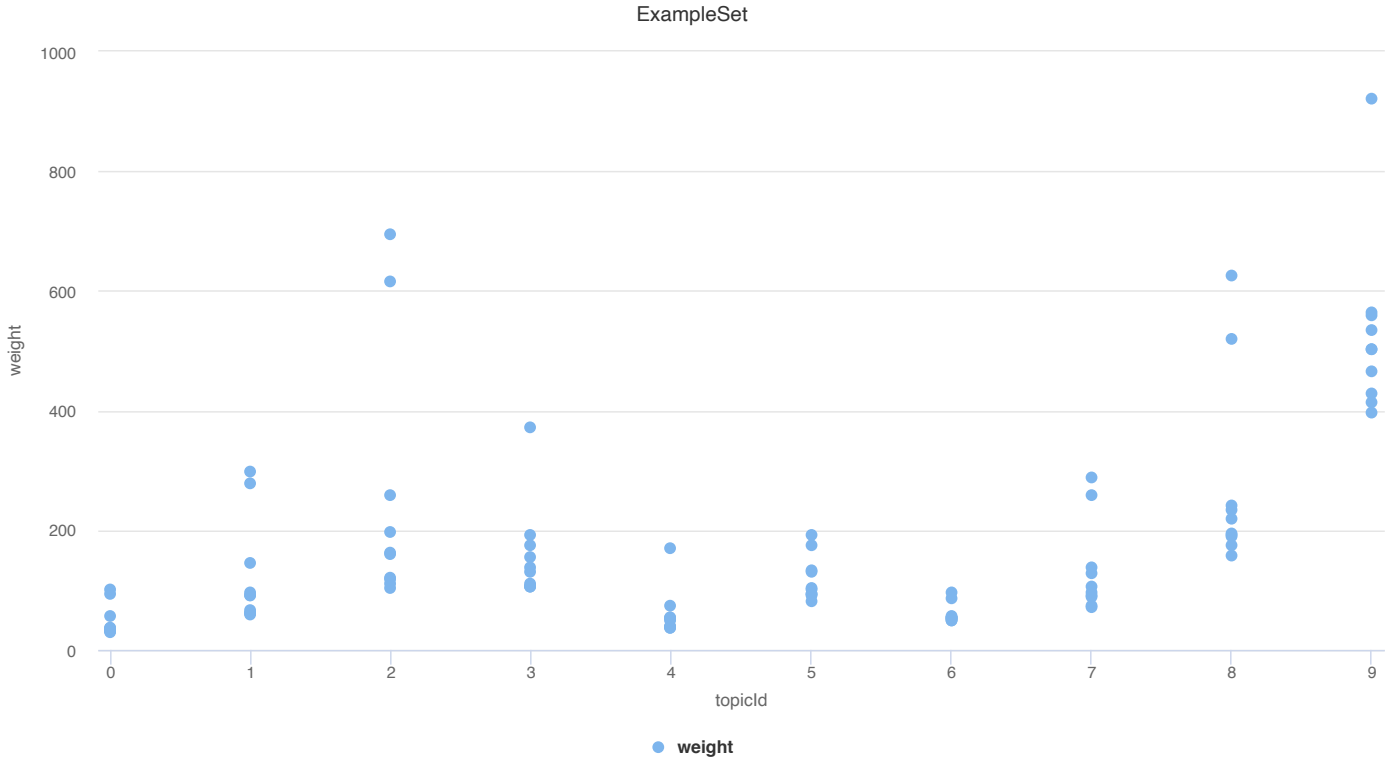
Αναλύοντας τις λέξεις από το Topic 8, συμπεραίνουμε ότι οι ασθενείς αντιμετώπισαν παρενέργειες όπως ναυτία και πονοκεφάλους.

Topic 9

topicId	word	weight
9,0	depress	922,0
9,0	work	564,0
9,0	take	558,0
9,0	feel	534,0
9,0	week	502,0
9,0	anxieti	502,0
9,0	start	466,0
9,0	month	428,0
9,0	help	413,0
9,0	felt	397,0

Αναλύοντας τις λέξεις από το Topic 9, συμπεραίνουμε ότι οι ασθενείς αντιμετώπισαν παρενέργειες όπως κατάθλιψη και άγχος.

Παρακάτω παρουσιάζεται ένα διάγραμμα όπου εμφανίζονται τα topics στον άξονα x και η συχνότητα εμφάνισης των λέξεων για το κάθε topic στον άξονα y.



Γράφημα 4: Πείραμα 4-LDA Analysis for Depression (filter for rating<5) - 10 topics

Αναλύοντας το παραπάνω γράφημα μπορούμε να συμπεράνουμε ότι τα Topics 0,4,5,6 είναι πιο συνεκτικά, δηλαδή όλες οι λέξεις στα συγκεκριμένα έχουν παρόμοια συχνότητα εμφάνισης και συνεπώς τα συμπεράσματα που έχουμε εξάγει από αυτά είναι περισσότερο αξιόπιστα. Αντιθέτως στα topics 1,2,3,7,8,9 υπάρχουν σχόλια με πολύ μεγαλύτερη συχνότητα εμφάνισης από άλλα, οπότε είναι πιθανό να μην σχετίζονται και ως αποτέλεσμα τα συμπεράσματα που εξάγουμε είναι λιγότερο αξιόπιστα.

Παράλληλα παρατηρούμε ότι τα topics 2,8 και ακόμη περισσότερο το topic 9 έχουν πολύ μεγαλύτερη συχνότητα εμφάνισης στις λέξεις τους. Αυτό μας οδηγεί στο συμπέρασμα ότι οι αναφερόμενες περενέργειες: κατάθλιψη, άγχος, ναυτία και πονοκέφαλος, κάποιες από τις οποίες είναι επίσης κοινές σε αυτά τα topics, είναι τα οι πιο κοινά και συχνά εμφανιζόμενες παρενέργειες των αντικαταθλιπτικών φαρμάκων.

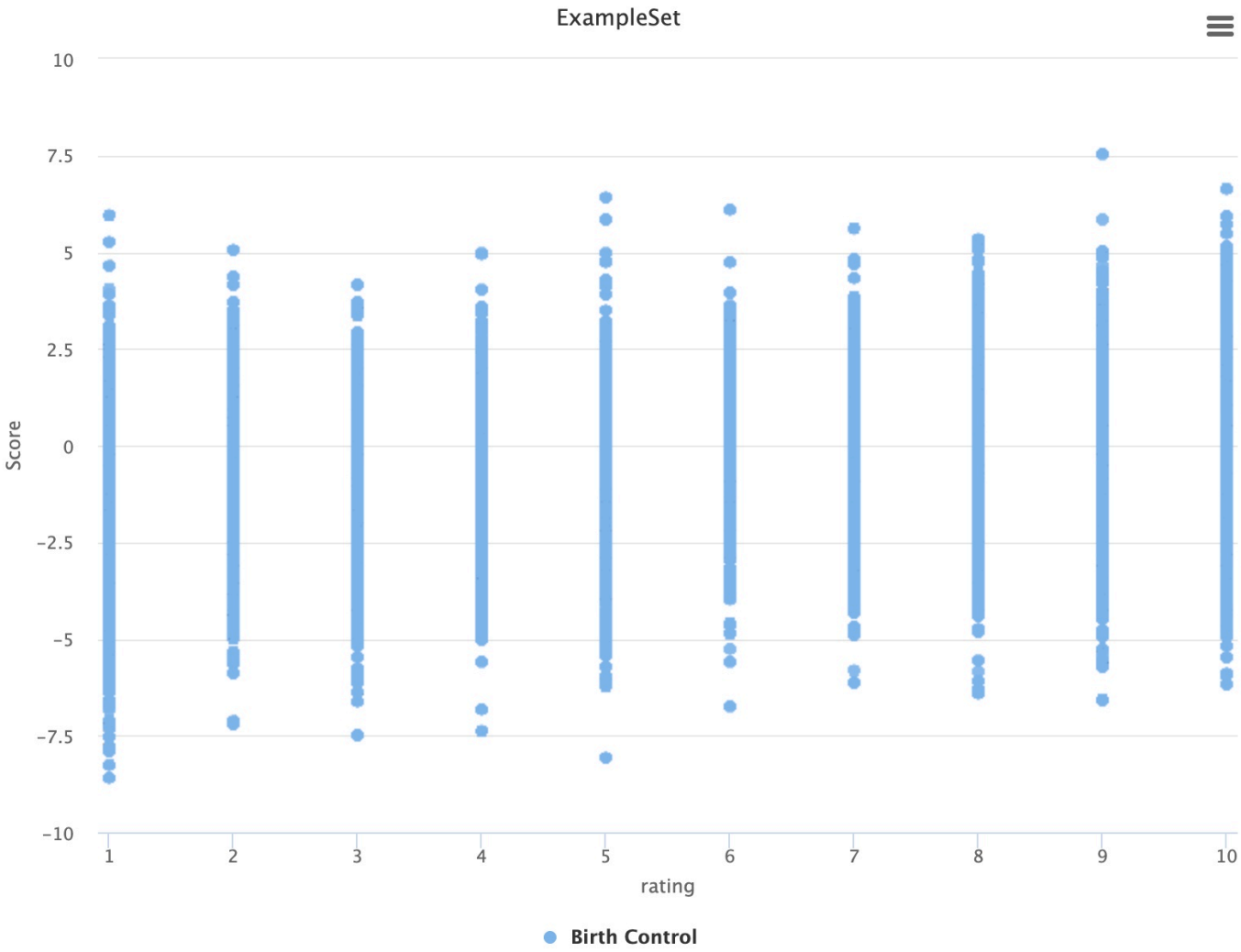
Συμπεραίνουμε πως στο Πείραμα 4, όπου φιλτράραμε μόνο τα αρνητικά σχόλια, ήταν πολύ πιο εύκολο να αναλύσουμε τα αποτελέσματα και να εξάγουμε σαφή συμπεράσματα για τις παρενέργειες των φαρμάκων. Επίσης παρατηρήθηκε πως δεν υπήρχαν αρνητικά σχόλια που να αφορούν άλλους

παράγοντες πέρα από τις παρενέργειες όπως για παράδειγμα η τιμή, η διαθεσιμότητα ή κάποιος ποιοτικό πρόβλημα των σκευασμάτων.

Συνολικά μετά την ολοκλήρωση των Πειραμάτων 1-4 παρατηρούμε ότι συγκρίνοντας τα weight των σχολίων μεταξύ των 2 παθήσεων που εξετάσαμε υπάρχει πολύ μεγάλη απόκλιση τόσο στα πειράματα χωρίς φίλτρο στο rating, όσο και στα πειράματα που έχουν φίλτρο στο rating. Τα σχόλια που αφορούν το birth control είναι έως και 10 φορές περισσότερα από τα σχόλια για το depression και στις 2 περιπτώσεις που αναφέρουμε. Αυτό θα μπορούσε να δικαιολογηθεί εάν αναλογιστεί κανείς την διαφορά στην ψυχολογική κατάσταση μεταξύ των δύο κατηγοριών ασθενών. Πιθανώς οι ασθενείς που λαμβάνουν φάρμακα για την πάθηση της κατάθλιψης να μην έχουν την διάθεση ή την ψυχολογική δύναμη για να μπούνε σε μια διαδικασία να γράψουν κριτική για τα φάρμακα που λαμβάνουν. Αντίθετα ένας ασθενής που λαμβάνει κάποιο αντισυλληπτικό φάρμακο πιθανώς έχει καλύτερη διάθεση και περισσότερη όρεξη για να ασχοληθεί αφήνοντας μια κριτική για τα φάρμακο που λαμβάνει.

- **Πείραμα 5: Sentiment Analysis for Birth Control (no filter for rating)**

Στο συγκεκριμένο πείραμα πραγματοποιήθηκε Sentiment Analysis με την χρήση Rapid Miner. Επιλέχθηκε η πάθηση του Birth Control, χωρίς να χρησιμοποιηθεί φίλτρο στο rating. Αυτός ο operator δημιουργεί μια βαθμολογία συναισθήματος εφαρμόζοντας είτε λεξικά συναισθήματος ανοιχτού κώδικα είτε ιδιόκτητες μεθόδους API σε ένα υπάρχον χαρακτηριστικό κειμένου.
Στις παραμέτρους επιλέχθηκε το model: vader και text attribute: review.
Το μοντέλο vader χρησιμοποιεί το λεξικό VADER (Valence Aware Dictionary and sEntiment Reasoner) και το συναίσθημα που βασίζεται σε κανόνες για να βαθμολογήσει το κείμενο. Το VADER είναι ειδικά προσαρμοσμένο στα συναισθήματα που εκφράζονται στα μέσα κοινωνικής δικτύωσης και παράγει αποτελέσματα με βάση ένα λεξικό λέξεων. Αυτός ο τελεστής υπολογίζει και στη συνέχεια εκθέτει το άθροισμα όλων των βαθμολογιών λέξεων συναισθήματος στο κείμενο.
Παρακάτω παρουσιάζεται ο πίνακας που προέκυψε από την παραπάνω διαδικασία. Στον άξονα x εμφανίζεται το rating και στον άξονα y εμφανίζεται το score:

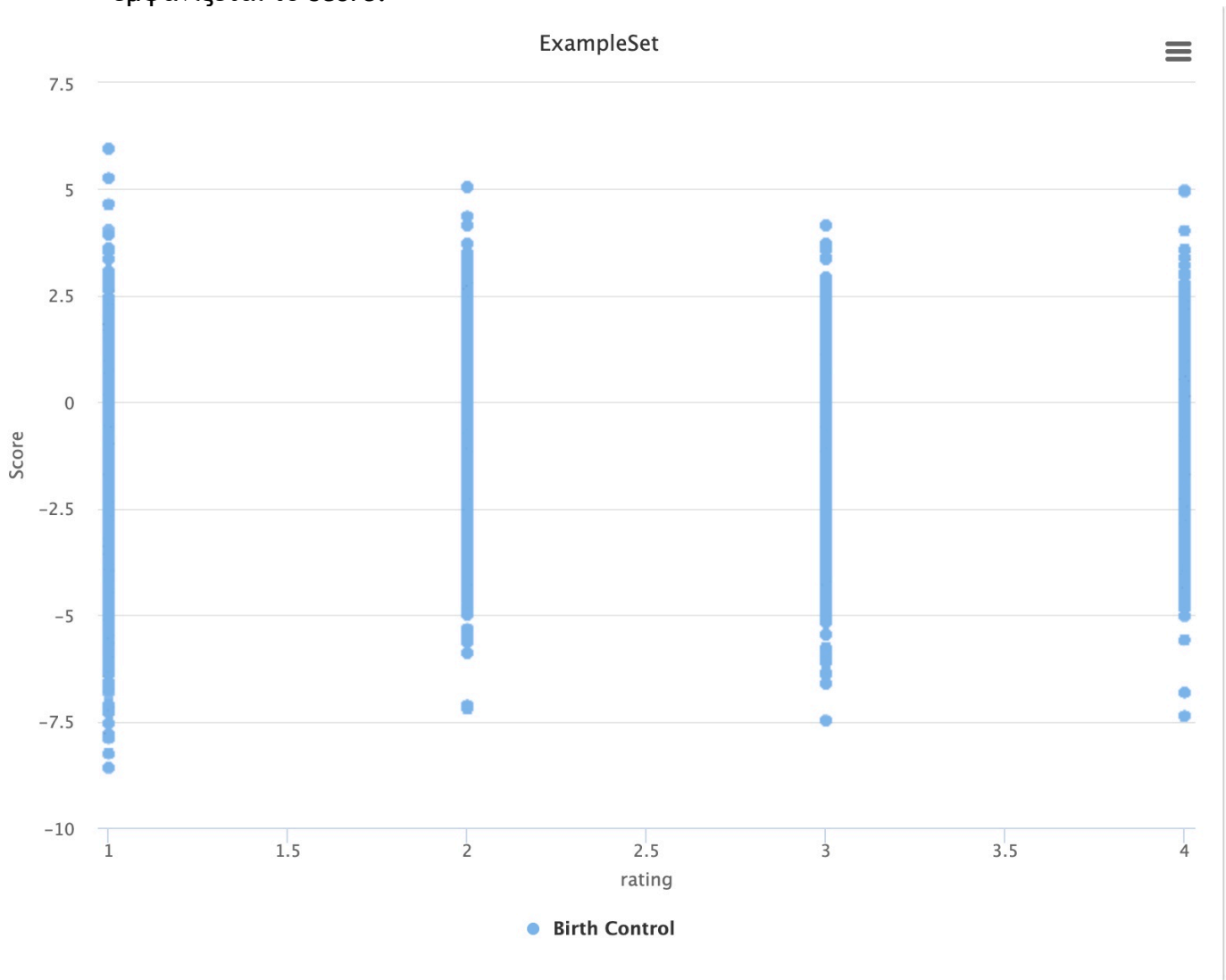


Γράφημα 5: Πείραμα 5-Sentiment Analysis for Birth Control (no filter for rating)

Αναλύοντας το παραπάνω διάγραμμα διαπιστώνουμε πως τόσο σε αρνητικά όσο και σε θετικά ratings παρατηρούμε χαμηλά και υψηλά sentiment scores. Αυτό πιθανώς οφείλετε στο ότι μπορεί ένα σχόλιο να περιγράφει πως ο ασθενής με την χρήση του φαρμάκου απαλλάχθηκε από τα αρνητικά συμπτώματα που παρουσίαζε, γι αυτό και δίνει ένα θετικό rating, όμως η sentiment analysis λόγω της ύπαρξης των λέξεων με αρνητικό πρόσημο στο σχόλιο αποδίδουν ένα χαμηλό score. Συνεπώς δεν μας βοηθάει μια τέτοια ανάλυση στο να εξάγουμε έγκυρα συμπεράσματα.

- **Πείραμα 6: Sentiment Analysis for Birth Control (filter for rating<5)**

Πραγματοποιήσαμε ακριβώς το ίδιο πείραμα με το Πείραμα 5 με την προσθήκη φίλτρου στο rating<5, ώστε να φιλτράρουμε μόνο τα αρνητικά σχόλια. Παρακάτω παρουσιάζεται ο πίνακας που προέκυψε από την παραπάνω διαδικασία. Στον άξονα x εμφανίζεται το rating και στον άξονα y εμφανίζεται το score:

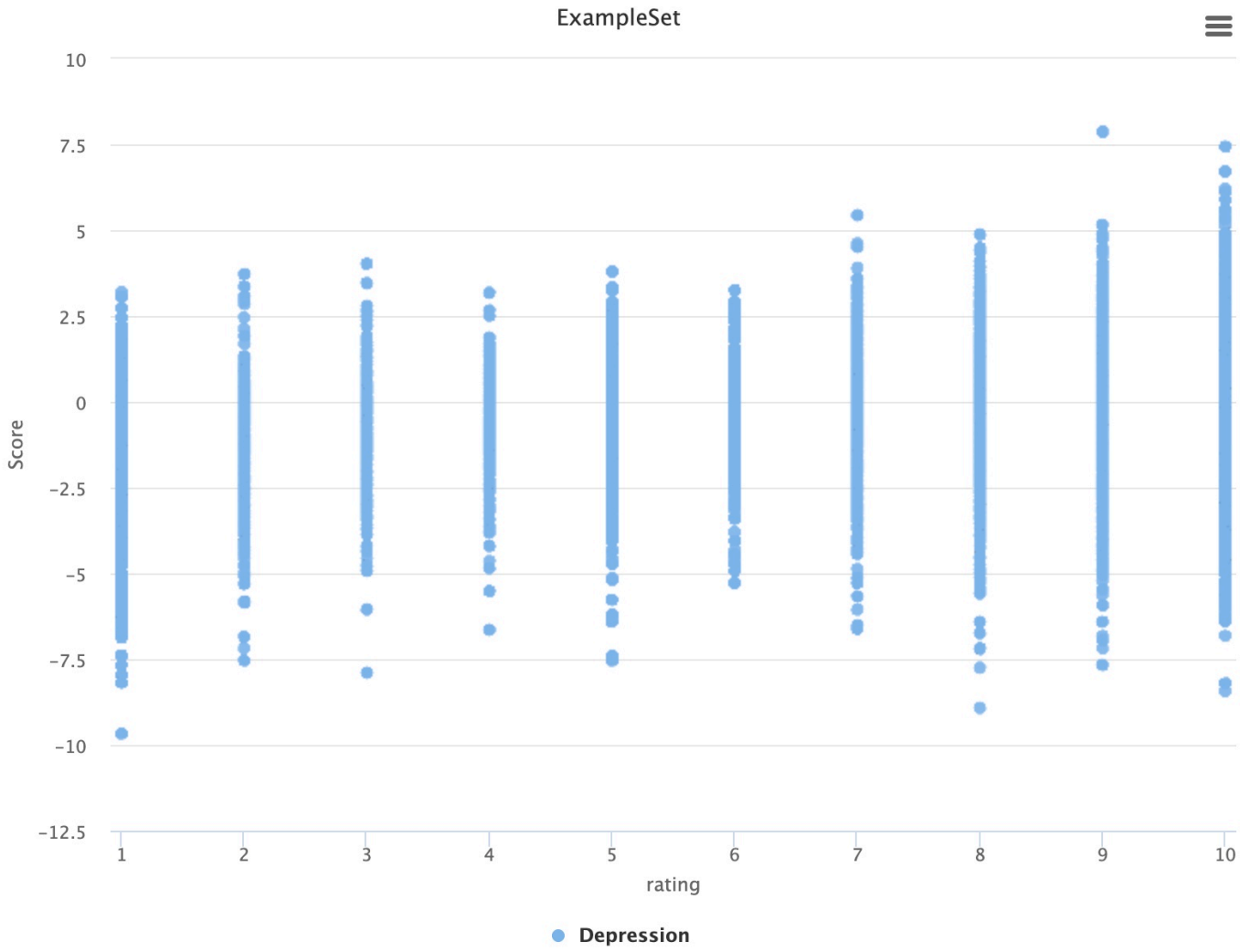


Γράφημα 6: Πείραμα 6-Sentiment Analysis for Birth Control (filter for rating<5)

Αναλύοντας το παραπάνω διάγραμμα διαπιστώνουμε πως παρότι μελετάμε μόνο αρνητικά ratings παρατηρούμε χαμηλά αλλά και υψηλότερα sentiment scores. Συγκριτικά με το Πείραμα 5 τα scores είναι χαμηλότερα, ωστόσο και σε αυτήν την περίπτωση δεν μπορούμε να έχουμε μια σαφή εικόνα και να εξάγουμε ουσιαστικά συμπεράσματα.

- **Πείραμα 7: Sentiment Analysis for Depression (no filter for rating)**

Στο συγκεκριμένο πείραμα πραγματοποιήθηκε Sentiment Analysis με την χρήση Rapid Miner. Επιλέχθηκε η πάθηση του Depression, χωρίς να χρησιμοποιηθεί φίλτρο στο rating. Στις παραμέτρους επιλέχθηκε το model: vader και text attribute: review. Παρακάτω παρουσιάζεται ο πίνακας που προέκυψε από την παραπάνω διαδικασία. Στον άξονα x εμφανίζεται το rating και στον άξονα y εμφανίζεται το score:



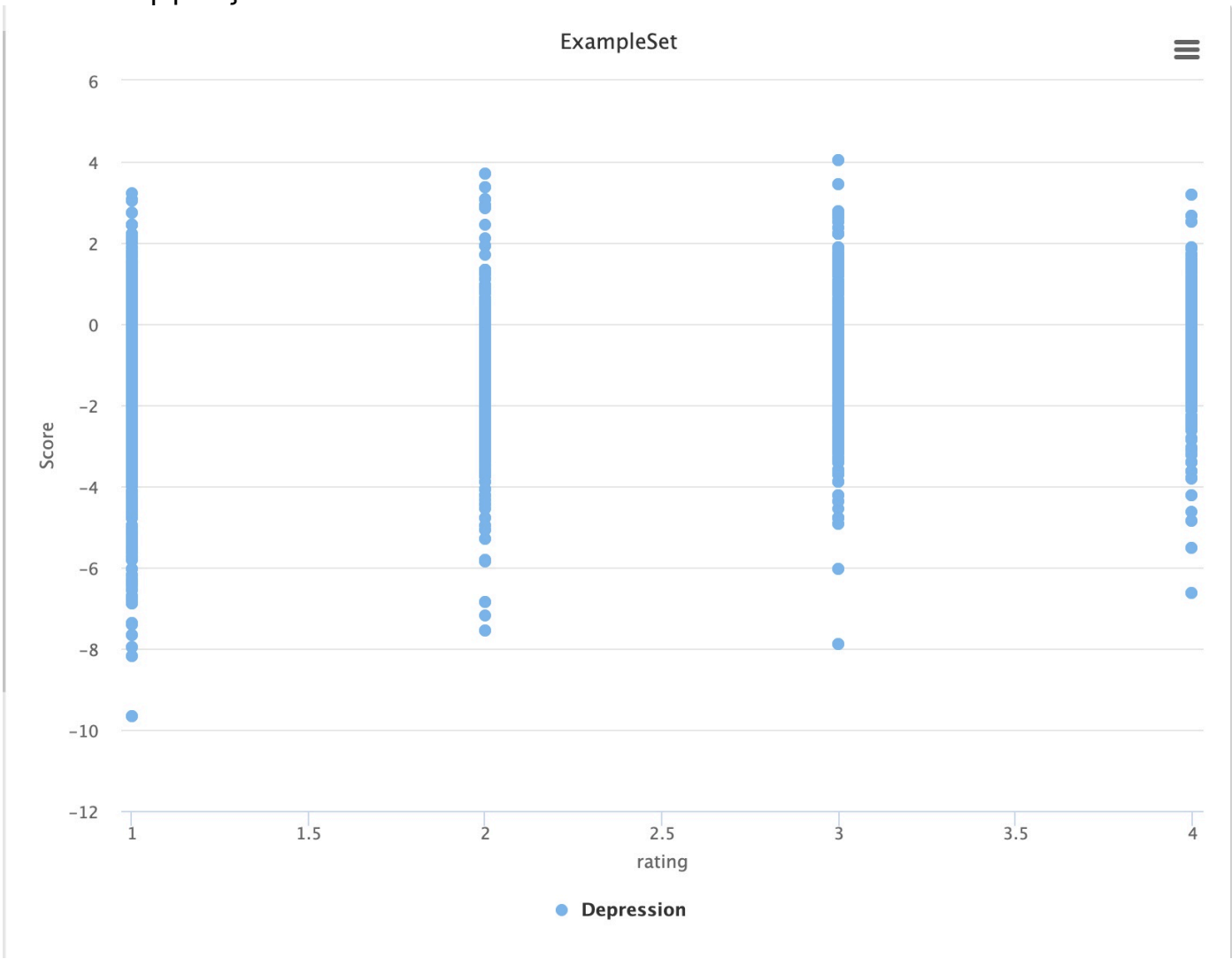
Γράφημα 7: Πείραμα 7-Sentiment Analysis for Depression (no filter for rating)

Αναλύοντας το παραπάνω διάγραμμα διαπιστώνουμε πως τόσο σε αρνητικά όσο και σε θετικά ratings παρατηρούμε χαμηλά και υψηλά sentiment scores. Αυτό πιθανώς οφείλετε στο ότι μπορεί ένα σχόλιο να περιγράφει πως ο ασθενής με την χρήση του φαρμάκου απαλλάχθηκε από τα αρνητικά συμπτώματα που παρουσίαζε, γι' αυτό και δίνει ένα θετικό rating, όμως η sentiment analysis λόγω της ύπαρξης των λέξεων με αρνητικό πρόσημο στο

σχόλιο αποδίδουν ένα χαμηλό score. Συνεπώς δεν μας βοηθάει μια τέτοια ανάλυση στο να εξάγουμε έγκυρα συμπεράσματα.

- **Πείραμα 8: Sentiment Analysis for Depression (filter for rating<5)**

Πραγματοποιήσαμε ακριβώς το ίδιο πείραμα με το Πείραμα 7 με την προσθήκη φίλτρου στο rating<5, ώστε να φιλτράρουμε μόνο τα αρνητικά σχόλια. Παρακάτω παρουσιάζεται ο πίνακας που προέκυψε από την παραπάνω διαδικασία. Στον άξονα x εμφανίζεται το rating και στον άξονα y εμφανίζεται το score:



Γράφημα 8: Πείραμα 8-Sentiment Analysis for Depression (filter for rating<5)

Αναλύοντας το παραπάνω διάγραμμα διαπιστώνουμε πως παρότι μελετάμε μόνο αρνητικά ratings παρατηρούμε χαμηλά αλλά και υψηλότερα sentiment scores. Συγκριτικά με το Πείραμα 7 τα scores είναι χαμηλότερα, ωστόσο και σε αυτήν την περίπτωση δεν μπορούμε να έχουμε μια σαφή εικόνα και να εξάγουμε ουσιαστικά συμπεράσματα.

5 Συμπεράσματα-Σχολιασμός

Με την ολοκλήρωση των πειραμάτων μας μπορούμε με σιγουριά να υποστηρίξουμε ότι υπάρχει πληθώρα από διαθέσιμες συλλογές δεδομένων που αφορούν reviews φαρμάκων. Τα reviews αυτά μπορούν να αναλυθούν από πλατφόρμες επιστήμης δεδομένων, όπως είναι το RapidMiner το οποίο χρησιμοποιήσαμε εμείς. Στα reviews αυτά μπορούν να χρησιμοποιηθούν διάφορα φίλτρα για να εξεταστούν συγκεκριμένες παθήσεις ή και διαφορετικές δραστικές ουσίες φαρμάκων. Μπορούν επίσης να χρησιμοποιηθούν φίλτρα στο rating των σχολίων κ.α. και τέλος να αναλυθούν με διαφορετικές μεθόδους και μοντέλα. Συνεπώς συμπεραίνουμε ότι η διαθεσιμότητα των δεδομένων είναι τεράστια και οι παράμετροι που μπορούν να εξεταστούν εξίσου πολλές.
Αναλύοντας τα αποτελέσματα των πειραμάτων μας, τα οποία βασίστηκαν σε κριτικές ασθενών σχετικά με την λήψη αντισυλληπτικών και αντικαταθλιπτικών φαρμάκων, καταλήγουμε σε κάποια γενικά συμπεράσματα.
Η sentiment analysis δεν μπορεί να μας παρέχει σαφή και αξιόπιστα αποτελέσματα κατά την ανάλυση των σχολίων των ασθενών. Αυτό πιθανώς οφείλεται εν μέρη στις συγκεκριμένες παθήσεις που επιλέξαμε, καθώς θεωρούνται και οι δύο παθήσεις που προκαλούν συναισθηματική φόρτιση, με αποτέλεσμα η sentiment analysis σε αυτά τα δεδομένα να είναι αρκετά δύσκολη.
Αντιθέτως η LDA, όπως διαπιστώθηκε στα πειράματα μας, είναι μια τεχνική η οποία μας δίνει μια πολύ ξεκάθαρη εικόνα σχετικά με τις παρενέργειες των φαρμάκων. Αυτό οφείλεται στο γεγονός πως η LDA κάνει γλωσσική ανάλυση και όχι συναισθηματική, με αποτέλεσμα τα δεδομένα να είναι πιο αντικειμενικά. Συνεπώς θα μπορούσε να αξιοποιηθεί από τις φαρμακευτικές εταιρίες στην διαδικασία της ανάπτυξης νέων φαρμάκων. Κατά την ανάπτυξη νέων φαρμάκων, η φαρμακευτική εταιρία θα μπορεί να μελετήσει όλες τις παρενέργειες των υπαρχόντων φαρμάκων και να τις ελαχιστοποιήσει, αυξάνοντας με αυτόν τον τρόπο την ικανοποίηση των ασθενών και μειώνοντας επίσης τον χρόνο και το κόστος.
Η φαρμακοβιομηχανία έχει πολύ αυστηρές και προκαθορισμένες διαδικασίες σε κάθε βήμα του κύκλου ζωής ενός φαρμάκου. Η αξιοποίηση της τεχνολογίας και του τεράστιου όγκου των big data που βρίσκονται στο διαδίκτυο για την ανάπτυξη νέων φαρμάκων δεν είναι ακόμη μια δεδομένη πρακτική και βρίσκεται σε πολύ πρώιμο στάδιο.
Ωστόσο έχουν ήδη γίνει τα πρώτα βήματα προς αυτήν την κατεύθυνση από τις μεγαλύτερες φαρμακευτικές εταιρίες παγκοσμίως. Είναι πολύ πιθανό η αξιοποίηση των big data στην φαρμακοβιομηχανία να αποτελέσει μελλοντικά έναν πολύ βασικό παράγοντα άντλησης δεδομένων. Εκτιμούμε πως οι εταιρίες που θα «αγκαλιάσουν» από νωρίς την τεχνολογία και θα αξιοποιήσουν όλη την δύναμη της θα βρεθούν με σημαντικό πλεονέκτημα έναντι των ανταγωνιστών, καθώς θα μπορούν να έχουν άμεση επαφή με τις ανάγκες των ασθενών, να μειώσουν σημαντικά τους χρόνους της ανάπτυξης των φαρμάκων και εν τέλη να παράξουν φάρμακα με ελαχιστοποιημένες παρενέργειες και καλύτερη απόδοση.

6 References

1. Ali, N. (2023). Concurrent Design Strategy in Product Design and Development. *International Design Journal*, 13(5), 473-488.
2. Dellarocas, Chrysanthos; Awad, Neveen; and Zhang, Xiaoquan, "Exploring the Value of Online Reviews to Organizations: Implications for Revenue Forecasting and Planning" (2004). ICIS 2004 Proceedings.
3. Elqortobi, M., Rahj, A., & Bentahar, J. (2023, August). Granular Traceability Between Requirements and Test Cases for Safety-Critical Software Systems. In International Conference on Mobile Web and Intelligent Information Systems (pp. 202-220). Cham: Springer Nature Switzerland.
4. Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018, April). Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In Proceedings of the 2018 International Conference on Digital Health (pp. 121-125).
5. Gupta, A. (2023). Pharmaceutical Drug Serialization: A Comprehensive Review. Universal Journal of Pharmacy and Pharmacology, 26-33.
6. Gurbuz, E. (2018). Theory of new product development and its applications. Marketing, 57-75.
7. Henderi, H., Hayadi, B. H., Sofiana, S., Padeli, P., & Setiyadi, D. (2023). Unsupervised Learning Methods for Topic Extraction and Modeling in Large-scale Text Corpora using LSA and LDA. Journal of Applied Data Sciences, 4(3), 103-118.
8. Jagtap, S., & Duong, L. N. K. (2019). Improving the new product development using big data: A case study of a food company. British Food Journal.
9. Kallumadi,Surya and Grer,Felix. (2018). Drug Review Dataset (Drugs.com). UCI Machine Learning Repository. https://doi.org/10.24432/C5SK5S.
10. Literature Review on Applying CRISP-DM Process Model Christoph Schröerab*, Felix Kruseb, Jorge Marx Gómezb Christoph Schröerab*, Felix Kruseb, Jorge Marx Gómezb
11. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113.
12. Plotnikova V, Dumas M, Milani F. Adaptations of data mining methodologies: a systematic literature review. PeerJ Comput Sci. 2020 May 25;6:e267. doi: 10.7717/peerj-cs.267. PMID: 33816918; PMCID: PMC7924527.
13. Prašnikar, J., & Škerlj, T. (2006). New product development process and time-to-market in the generic pharmaceutical industry. Industrial Marketing Management, 35(6), 690-702.
14. Procedia Computer Science 181 (2021) 526–534. A Systematic Literature Review on Applying CRISP-DM Process Model.
15. Saiga, K., Ullah, A. S., & Kubo, A. (2021). A Sustainable Reverse Engineering Process. Procedia CIRP, 98, 517-522.

16. Seebode, C., Ort, M., Regenbrecht, C., & Peuker, M. (2013, October). BIG DATA infrastructures for pharmaceutical research. In 2013 IEEE International Conference on Big Data (pp. 59-63). IEEE.
17. Tan, K. H., & Zhan, Y. (2017). Improving new product development using big data: a case study of an electronics company. R&D Management, 47(4), 570-582.
18. Thompson, D. C., & Bentzien, J. (2020). Crowdsourcing and open innovation in drug discovery: recent contributions and future directions. Drug discovery today, 25(12), 2284–2293.
19. Tormay, P. (2015). Big data in pharmaceutical R&D: Creating a sustainable R&D engine. Pharmaceutical medicine, 29(2), 87-92.
20. Yousefi, N., Mehralian, G., Rasekh, H. R., & Yousefi, M. (2017). New Product Development in the Pharmaceutical Industry: Evidence from a generic market. Iranian journal of pharmaceutical research : IJPR, 16(2), 834–846.
21. Zhan, Y., Tan, K. H., Li, Y., & Tse, Y. K. (2018). Unlocking the power of big data in new product development. Annals of Operations Research, 270(1), 577-595.
22. https://www.techtarget.com/searchsoftwarequality/definition/reverse-engineering
23. https://whatispiping.com/concurrent-engineering/
24. https://blog.nbs-us.com/why-is-traceability-important-in-the-pharmaceutical-industry
25. https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2
26. https://www.ibm.com/blog/keystonemab-drug-discovery/

7 Παράρτημα

Παρακάτω παρατίθενται τα XML αρχεία, τα οποία έτρεξαν στο Rapid Miner για την Sentiment Analysis και LDA Analysis αντίστοιχα:

Sentiment Analysis:

```
<?xml version="1.0" encoding="UTF-8"?><process version="9.10.008">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.10.008"
expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.10.008"
expanded="true" height="68" name="Retrieve drugsComCombined_formated"
width="90" x="45" y="34">
        <parameter key="repository_entry"
value="drugsComCombined_formated"/>
      </operator>
      <operator activated="true" class="filter_examples"
compatibility="9.10.008" expanded="true" height="103" name="Filter
Examples" width="90" x="246" y="34">
        <parameter key="parameter_expression" value=""/>
        <parameter key="condition_class" value="custom_filters"/>
        <parameter key="invert_filter" value="false"/>
        <list key="filters_list">
          <parameter key="filters_entry_key" value="condition.equals.Birth
Control"/>
        </list>
        <parameter key="filters_logic_and" value="true"/>
        <parameter key="filters_check_metadata" value="true"/>
      </operator>
      <operator activated="true" class="operator_toolbox:extract_sentiment"
compatibility="2.14.000" expanded="true" height="103" name="Extract
Sentiment" width="90" x="514" y="34">
        <parameter key="model" value="vader"/>
        <parameter key="text_attribute" value="review"/>
        <parameter key="show_advanced_output" value="true"/>
        <parameter key="use_default_tokenization_regex" value="true"/>
        <list key="additional_words"/>
```

```
    </operator>
    <connect from_op="Retrieve drugsComCombined_formated"
from_port="output" to_op="Filter Examples" to_port="example set input"/>
    <connect from_op="Filter Examples" from_port="example set output"
to_op="Extract Sentiment" to_port="exa"/>
    <connect from_op="Extract Sentiment" from_port="exa" to_port="result
1"/>
    <portSpacing port="source_input 1" spacing="0"/>
    <portSpacing port="sink_result 1" spacing="0"/>
    <portSpacing port="sink_result 2" spacing="0"/>
    <portSpacing port="sink_result 3" spacing="0"/>
   </process>
  </operator>
</process>
```

LDA Analysis:

```
<?xml version="1.0" encoding="UTF-8"?><process version="9.10.008">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.10.008"
expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.10.008"
expanded="true" height="68" name="Retrieve drugsComCombined_formated"
width="90" x="45" y="34">
        <parameter key="repository_entry"
value="drugsComCombined_formated"/>
      </operator>
      <operator activated="true" class="filter_examples"
compatibility="9.10.008" expanded="true" height="103" name="Filter
Examples" width="90" x="179" y="34">
        <parameter key="parameter_expression" value=""/>
        <parameter key="condition_class" value="custom_filters"/>
        <parameter key="invert_filter" value="false"/>
        <list key="filters_list">
          <parameter key="filters_entry_key" value="condition.equals.Birth
Control"/>
```

```
        <parameter key="filters_entry_key"
value="condition.equals.Depression"/>
      </list>
      <parameter key="filters_logic_and" value="true"/>
      <parameter key="filters_check_metadata" value="true"/>
    </operator>
    <operator activated="true" class="nominal_to_text"
compatibility="9.10.008" expanded="true" height="82" name="Nominal to
Text" width="90" x="447" y="34">
      <parameter key="attribute_filter_type" value="single"/>
      <parameter key="attribute" value="review"/>
      <parameter key="attributes"
value="condition|date|drugname|review"/>
      <parameter key="use_except_expression" value="false"/>
      <parameter key="value_type" value="nominal"/>
      <parameter key="use_value_type_exception" value="false"/>
      <parameter key="except_value_type" value="file_path"/>
      <parameter key="block_type" value="single_value"/>
      <parameter key="use_block_type_exception" value="false"/>
      <parameter key="except_block_type" value="single_value"/>
      <parameter key="invert_selection" value="false"/>
      <parameter key="include_special_attributes" value="false"/>
    </operator>
    <operator activated="true" class="text:process_document_from_data"
compatibility="9.4.000" expanded="true" height="82" name="Process
Documents from Data" width="90" x="581" y="34">
      <parameter key="create_word_vector" value="true"/>
      <parameter key="vector_creation" value="TF-IDF"/>
      <parameter key="add_meta_information" value="false"/>
      <parameter key="keep_text" value="true"/>
      <parameter key="prune_method" value="none"/>
      <parameter key="prune_below_percent" value="3.0"/>
      <parameter key="prune_above_percent" value="30.0"/>
      <parameter key="prune_below_rank" value="0.05"/>
      <parameter key="prune_above_rank" value="0.95"/>
      <parameter key="datamanagement" value="double_sparse_array"/>
      <parameter key="data_management" value="memory-optimized"/>
      <parameter key="select_attributes_and_weights" value="false"/>
      <list key="specify_weights"/>
      <process expanded="true">
        <operator activated="true" class="text:tokenize"
compatibility="9.4.000" expanded="true" height="68" name="Tokenize"
width="90" x="112" y="34">
          <parameter key="mode" value="non letters"/>
          <parameter key="characters" value=".:"/>
          <parameter key="language" value="English"/>
          <parameter key="max_token_length" value="3"/>
        </operator>
```

```
        <operator activated="true" class="text:filter_stopwords_english"
compatibility="9.4.000" expanded="true" height="68" name="Filter Stopwords
(English)" width="90" x="246" y="34"/>
        <operator activated="true" class="text:filter_by_length"
compatibility="9.4.000" expanded="true" height="68" name="Filter Tokens
(by Length)" width="90" x="380" y="34">
          <parameter key="min_chars" value="4"/>
          <parameter key="max_chars" value="30"/>
        </operator>
        <operator activated="true" class="text:stem_porter"
compatibility="9.4.000" expanded="true" height="68" name="Stem (Porter)"
width="90" x="514" y="34"/>
        <connect from_port="document" to_op="Tokenize"
to_port="document"/>
        <connect from_op="Tokenize" from_port="document" to_op="Filter
Stopwords (English)" to_port="document"/>
        <connect from_op="Filter Stopwords (English)" from_port="document"
to_op="Filter Tokens (by Length)" to_port="document"/>
        <connect from_op="Filter Tokens (by Length)" from_port="document"
to_op="Stem (Porter)" to_port="document"/>
        <connect from_op="Stem (Porter)" from_port="document"
to_port="document 1"/>
        <portSpacing port="source_document" spacing="0"/>
        <portSpacing port="sink_document 1" spacing="0"/>
        <portSpacing port="sink_document 2" spacing="0"/>
      </process>
    </operator>
    <operator activated="true" class="operator_toolbox:lda_exampleset"
compatibility="2.14.000" expanded="true" height="124" name="Extract Topics
from Data (LDA)" width="90" x="782" y="34">
      <parameter key="text_attribute" value="text"/>
      <parameter key="number_of_topics" value="10"/>
      <parameter key="show_optimization_settings" value="true"/>
      <parameter key="use_alpha_heuristics" value="true"/>
      <parameter key="alpha_sum" value="0.1"/>
      <parameter key="use_beta_heuristics" value="false"/>
      <parameter key="beta" value="0.05"/>
      <parameter key="optimize_hyperparameters" value="true"/>
      <parameter key="optimize_interval_for_hyperparameters" value="10"/>
      <parameter key="iterations" value="1000"/>
      <parameter key="top_words_per_topic" value="10"/>
      <parameter key="stopword language" value="english"/>
      <parameter key="reproducible" value="false"/>
      <parameter key="enable_logging" value="false"/>
      <parameter key="use_local_random_seed" value="false"/>
      <parameter key="local_random_seed" value="1992"/>
    </operator>
    <connect from_op="Retrieve drugsComCombined_formated"
from_port="output" to_op="Filter Examples" to_port="example set input"/>
```

```
      <connect from_op="Filter Examples" from_port="example set output"
to_op="Nominal to Text" to_port="example set input"/>
      <connect from_op="Nominal to Text" from_port="example set output"
to_op="Process Documents from Data" to_port="example set"/>
      <connect from_op="Process Documents from Data" from_port="example
set" to_op="Extract Topics from Data (LDA)" to_port="exa"/>
      <connect from_op="Extract Topics from Data (LDA)" from_port="exa"
to_port="result 1"/>
      <connect from_op="Extract Topics from Data (LDA)" from_port="top"
to_port="result 2"/>
      <connect from_op="Extract Topics from Data (LDA)" from_port="mod"
to_port="result 3"/>
      <connect from_op="Extract Topics from Data (LDA)" from_port="per"
to_port="result 4"/>
      <portSpacing port="source_input 1" spacing="0"/>
      <portSpacing port="sink_result 1" spacing="0"/>
      <portSpacing port="sink_result 2" spacing="0"/>
      <portSpacing port="sink_result 3" spacing="0"/>
      <portSpacing port="sink_result 4" spacing="0"/>
      <portSpacing port="sink_result 5" spacing="0"/>
    </process>
  </operator>
</process>
```