Text mining project

Analysis of pain medicine sentiment based on customer rating via comments on online pharmaceutical encyclopaedia – drugs.com

Zofia Bentyn

Abstract

In main interest of this project lies extracting information about how sentiment of pain medicine changed throughout 14 years and which topic can be extracted from users reviews.

Keywords: tramadol, acetaminophen, oxycodone, analysis, text mining.

1 Thesis

Pharmaceutical market thrives due to multiple factors. Growing pace of life, consumerism, ecological state of our environment and political situation, abuse through advertising and overall safety impacts the need for medicine and affects people's sentiment towards it.

I will demonstrate how people's sentiment to three popular pain medication (tramadol, acetaminophen and oxycodone) changed over the years, from 2008 to 2021. Apart from that I will extract main topics based on the corpus, and analyse the similarities between mentioned medicine across the years.

2 Theoretical Background

Text mining, according to Wikipedia, „is a process of deriving high-quality information from text". Mentioned extraction is performed on data from different, written resources.

Opinion mining, also known as sentiment analysis, includes text mining and measures the overall sentiment on given topic.

3 Research Methodology

3.1 The Main Definitions Used in Project

Document Term Matrix is a matrix that includes frequency of term in a collection of documents.

Zip'f law, model of the distribution of term in a collection of documents.

Topic modelling, a type of statistical modelling useful for searching abstract topics.

Latent Dirichlet Allocation is a probabilistic model, which is useful for discovery of topics in a collection of documents.

Gibbs sampling is a method that draw an instance from the distribution of each document in a collection.

Multidimensional Scaling maps location of a document based on how the document differs from other documents.

K-means a method that allows for partitioning the documents into given clusters.

Sentiment an attitude, judgment prompted by feeling.

Sentiment lexicon is a collection of words associated with their sentiment orientation.

Tokens are building blocks of Natural Language.

3.2 Research Questions

3.2.1 Which topics can be distinguished from the comments?

3.2.2 How does the sentiment changes over 14 years?

3.2.3 Which medicine has the most positive overall reception?

3.3 Research Plan

3.3.1 Comments Scraping

Data that I am basing my analysis upon comes from online pharmaceutical encyclopaedia, precisely from comments written by users throughout 14 years.

The process of gathering data was completed by using Beautiful Soup, a Python package that was crated for enabling people parsing HTML documents and, furthermore, extracting the desired data.

Additionally I have used webbot from Browser package, which made possible to overcome difficulties that appeared whilst extracting mentioned information.

3.3.2 Corpus Pre-processing

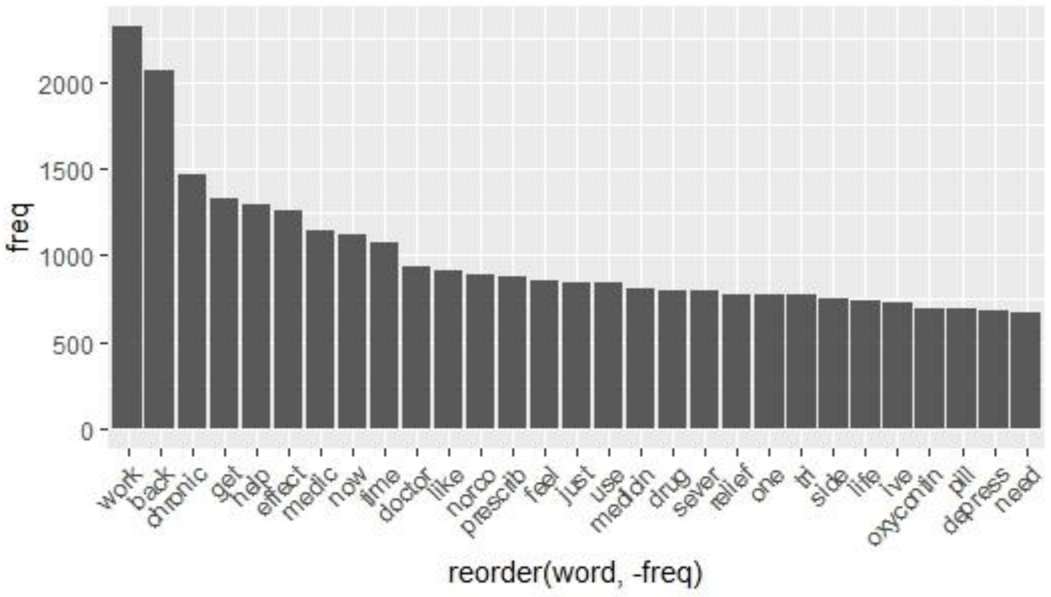
After extracting desired data into documents (based on what year and which medicine given comment was written) I began the pre-processing.

Firstly, I have downloaded libraries [*tm, Rcpp, topicmodels, lsa, ggplot2*] that would allow for pre-processing and analysis. Then I began cleaning my data, which consisted of getting rid of every symbol, accept letters, changing all words to lowercase, removing additional whitespace as well as stop words. Additionally I decided to get rid of the word “pain”, as this term is not giving us any valuable information.

After performing term association on name of each medicine I decided to recreate the corpus without names of given drugs. Since I performed Multidimensional Scaling I did not want any bias, which connection via keyword could introduce.

3.3.3 General Corpus Analysis

Creation of sparse Document Term Matrix allowed me to analyse the words of corpus. I created plot demonstrating Zipf's law, based on terms which occurred at least 1000 times throughout the documents. Presented below graph demonstrates that first most occurring word is "work". It either could suggest that the medicine *worked*, or it did not, or it is connected with work as in employment environment. Second, "back", could possibly mean the body part or position in time or space. Third one is the most direct, as it indicates that the topic of comments were mostly described as "chronic", as we know from main narrative of this project, the main topic is pain.



To show two main topics I used Latent Dirichlet Allocation. In the table below I put samples of words corelated with each topic. It is possible to name each topic. I distinguished the first group as words assimilated with motivation to start taking medicine, and second group as words used for describing the process of taking the drug and what comes with it, like "pharmacy".

Topic 1: Motivation	Topic 2: Aspects of medicine taking
Effect	Work
Feel	Brand
Depress	Pill
Use	Pharmaci
Try	Chronic
Help	Medic

Below I present how particular topic emerged throughout the years. We can observe, that the technical aspects connected with medicine have been addressed in reviews for just four years.

Year	Tramadol	Acetaminophen	Oxycodone
2008	1	1	1
2009	1	1	1
2010	1	1	1
2011	1	1	1
2012	1	1	1
2013	1	1	1
2014	1	1	1
2015	1	1	1
2016	1	1	1
2017	1	1	1
2018	1	2	2
2019	1	2	2
2020	1	2	2
2021	1	2	2

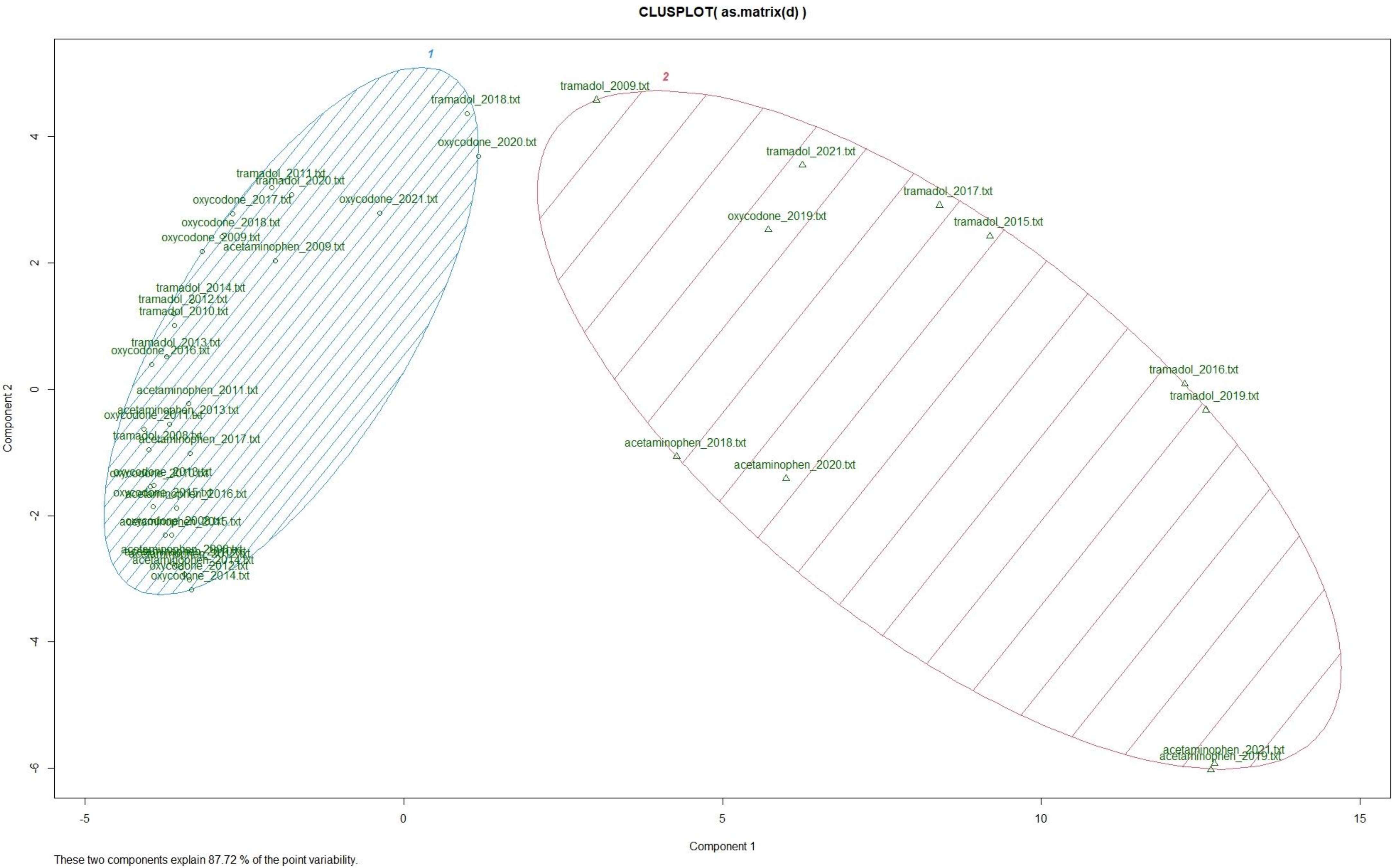
I decided to perform term association to see which terms correspond best with given medicine. Next to each word there is a frequency at which the word was displayed across the whole corpus. For clearance, I would suggest looking at word “leg” as a stemmed version of “legal”, “side” could be “side-effect”.

Under Acetaminophen, “Norco”, precisely “Norco 5/325” is another name for this medicine, the same case is with “Vicodin”, “Lortab” and “Lorcet”. To clarify the case of word “climb”, it has been scientifically proven that Acetaminophen is (next to Ibuprofen) effective prophylactic treatment of Acute Mountain Sickness.

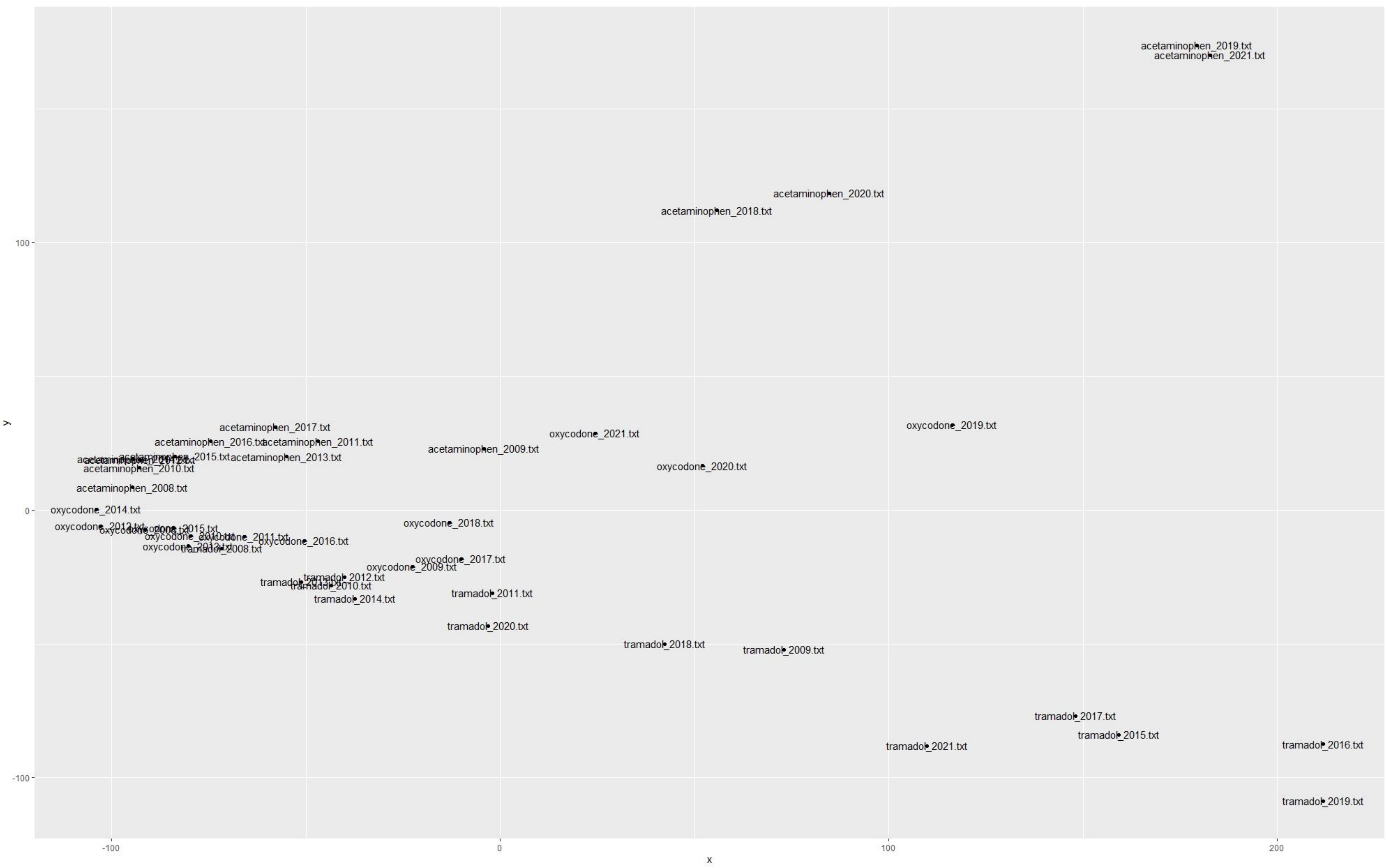
With Oxycodone we have similar case of discovering other names for this drug (“Xtampza”, “Dazidox”). Quite interestingly “breakthrough” is one of the most frequent words, which might suggest positive sentiment.

Tramadol		Acetaminophen		Oxycodone	
Mood	**0.93**	Norco	**0.87**	Xtampza	**0.77**
Symptom	**0.89**	Vicodin	**0.81**	Breakthrough	**0.76**
Leg	**0.88**	Lortab	**0.78**	Equival	**0.76**
Prescrib	**0.88**	Lorcet	**0.73**	Dazidox	**0.75**
Side	**0.87**	Climb	**0.70**	Chronic	**0.73**

At last, I wanted to perform Multidimensional Scaling to see how these documents look on a x and y axis. I performed k-means algorithm to make the distinguish more direct. We can also see that with time, comments are becoming less similar, accept for “oxycodone_2021.txt”, which is the closest to groups of comments from earlier years. Other than that, acetaminophen behaves similarly to tramadol - there are more differences between comments on these drugs with time. Comments from oxycodone users from 2016-2021 are the closest to comments on tramadol from 2008-2015.



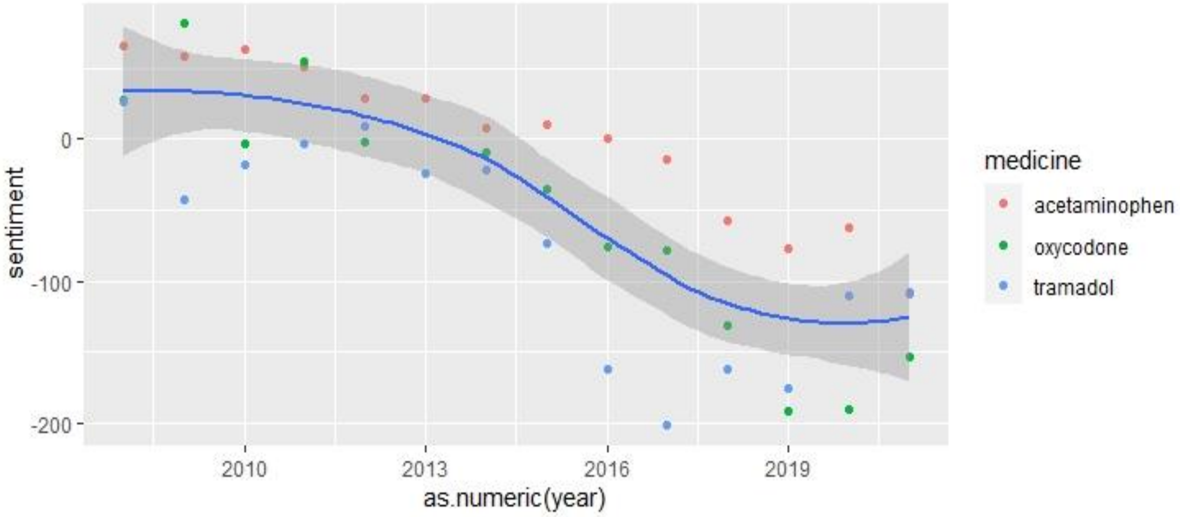
These two components explain 87.72 % of the point variability.


acetaminophen_2019.txt
acetaminophen_2021.txt
acetaminophen_2020.txt
acetaminophen_2018.txt
100
acetaminophen_2017.txt
acetaminophen_2016.txt
acetaminophen_2011.txt
acetaminophen_2015.txt
acetaminophen_2013.txt
acetaminophen_2012.txt
acetaminophen_2010.txt
acetaminophen_2008.txt
acetaminophen_2009.txt
oxycodone_2021.txt
oxycodone_2019.txt
oxycodone_2020.txt
y
0
oxycodone_2014.txt
oxycodone_2012.txt
oxycodone_2008.txt
oxycodone_2015.txt
oxycodone_2010.txt
oxycodone_2011.txt
oxycodone_2016.txt
oxycodone_2013.txt
tramadol_2008.txt
oxycodone_2018.txt
oxycodone_2017.txt
oxycodone_2009.txt
tramadol_2012.txt
tramadol_2013.txt
tramadol_2010.txt
tramadol_2014.txt
tramadol_2011.txt
tramadol_2020.txt
tramadol_2018.txt
tramadol_2009.txt
tramadol_2017.txt
tramadol_2015.txt
tramadol_2021.txt
tramadol_2016.txt
-100
tramadol_2019.txt
-100
0
100
200
x

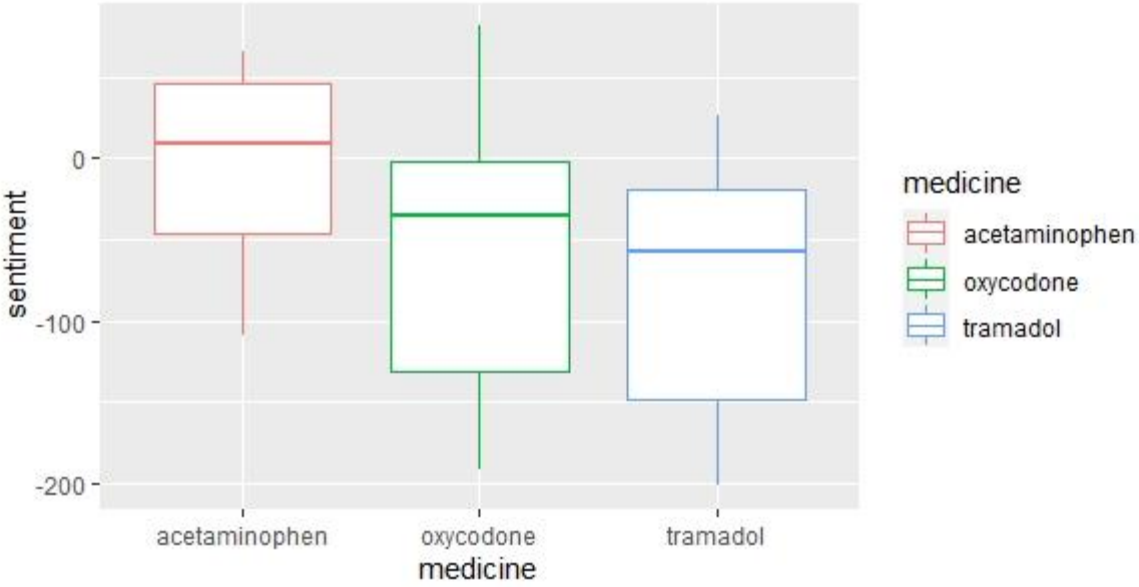
3.3.4 Sentiment Analysis throughout years

To calculate the sentiment I used different libraries [*tidyverse, tidytext, glue, stringr*]. Using “tidytext” enabled me access to “bing” sentiment lexicon, which was created mainly by Bing Liu. After pre-processing and tokenization I created two graphs. This time I also deleted the term “pain”. I tried to perform the sentiment analysis with or without it, to understand what impact has deleting this word. The word “pain” might be associated with negative sentiment, but what really is important, is how the pain is described.

The first graph displays how sentiment changed throughout 14 years. As we can see there is a fall in sentiment around years 2012 until 2021. Another thing that we can observe is how many peaks and falls tramadol sentiment has, while oxycodone and acetaminophen are falling at a moreover steady pace.



Second graph presents overall sentiment. As we can see, the most positive reviews has Acetaminophen. The mean of sentiment in case of Oxycodone and Tramadol are similar, as well as the range of it, while the range of sentiment in Acetaminophen is visibly smaller.



4 Conclusions

Analysis has shown, that we can distinguish two topic from the corpus. First of which is motivation, which can describe mainly why person has decided to try medicine, second topic relates to exposing aspects of taking given drug.

Throughout the years there has been steady fall in sentiment. This might be due to epidemic of Covid-19, but also better knowledge and data on the topic of prescription drug abuse. Possibly, even technological progress might have change the users review style.

Overall the best reviews are of Acetaminophen medicine. Tramadol is at a growth rate, but judging by it's overall sentiment, it might not be trustworthy.

References

https://www.elsevier.com/about/press-releases/research-and-journals/acetaminophen-a-viable-alternative-for-preventing-acute-mountain-sickness

https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

https://en.wikipedia.org/wiki/Zipf%27s_law

https://en.wikipedia.org/wiki/Document-term_matrix

https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

https://towardsdatascience.com/gibbs-sampling-8e4844560ae5

https://www.sciencedirect.com/topics/psychology/multidimensional-scaling

https://en.wikipedia.org/wiki/K-means_clustering

https://www.merriam-webster.com/dictionary/sentiment

https://link.springer.com/article/10.1007/s10115-020-01497-6

https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp