Top-Down Analysis on Medical Focus Group

Executive summary

The analysis on a document level using cosine similarity and summarization revealed an overview of the content and discussion. The results indicate that discussion shifted from alpha blocker, medical studies conducted, to cancer. In the in the second week the discussions are themed around cancer and consumers. The LDA analysis on the full corpus revealed four major topics of discussion and the overall sentiment of is leaning towards negative, apart from the positive sentiment in week1 Q1 and Q3.

# Introduction

This study was conducted to analyze the feedback of doctors during a workshop. Data pertaining to study is a collection of 7 text files, each pertaining to specific question discussed. A top down analysis is conducted on a document level, followed by analysis on the text overall and lastly the sentiment regarding the discussion (as shown in Figure 1).

Figure : Top-down analysis of corpus

# Data preprocessing

Data pre-processing makes the data easier to work with and ready for analytics and insight studies. Different pre-processing methods were applied such as tokenization, stop wording, lemmatization, abbreviations removal and punctuation removal.

The first step applied was removing the encoding characters and ‘\n’ values were replaced. The next step was tokenizing the data which breaks up the data into strings of words, keywords, symbols, and other elements called tokens. Many functions were created to normalize the data which includes removing punctuation which removes unrequired characters for the analysis.

To make the data easier to work with, stop wording was applied which helps removing the words that don’t add any meaning to the analysis like pronouns and other similar words which have less than 3 letters from the data.

Lemmatization was chosen over the Stemming method due to the fact that it takes the words context in consideration using POS(Part of speech) processing. Lemmatization helps in removing the modulation endings and returning the dictionary form of words or morphological words. The abbreviations present in the tokens were replaced with the actual terms through creating a dictionary of terms that needed replacement.

The results of the adjustments during the preprocessing step are presented in Table 1 below. After completing the preprocessing part, the length of the set of words was reduced from 1505 to 1153, while the total number of tokens were reduced from 7547 to 2965 tokens.

Table 1: Tokens after each step of preprocessing

|  |  |  |
| --- | --- | --- |
| Steps: | Number of tokens | Set of tokens |
| Tokenize | 7547 | 1505 |
| remove punctuations | 6940 | 1490 |
| stop wording | 2965 | 1235 |
| alter abbreviations | 2965 | 1235 |
| lemmatization | 2965 | 1153 |

# Analysis of Documents

## Summarization

From question 1’s collocations in the first week, the medical doctors (MDs) appear to be giving thought to the risk associated with tamsulosin hydrochloride, which is an alpha-blocker used to treat the symptoms of an enlarged prostrate. One can allude that this is the drug in consideration to become accessible over-the-counter. They also mention “population”, “propose”, “advisory committee” and “time period”, which may conclude that this question regarded medical management and clinical data.

The responses from question 2 in the same week included more technical collocations regarding medicine. “Benign prostatic” refers to benign prostatic hyperplasia which is the medical term for an enlarged prostrate. Two different kinds of cancer were also mentioned in the bladder and prostrate. These terms were also mentioned in week 2, questions 2 and 3, the latter of which highlighted more diseases, for example, liver, kidney and diabetes. Other questions in the two weeks provided collocations more associated with regulatory issues: week 1, questions 3 and 4, and week 2, question 1.

For more comprehensive insights of the responses, summaries were then processed using the raw corpus and collocations. These collocations differed from the ones provided in Table 2 as they were computed using bigram association measures. These summaries not only covered the essence of the responses but gave an idea of the response size through the number of sentences in the summary. For example, week 1, questions 1, 3 and 4 did not receive a lot of feedback from the MDs, however, there was much more discussion in week 1, question 2, and week 2, question 1. These specific questions and responses may be fruitful in further analysis especially in deriving sentiments. The sentence count and alternate collocations are shown in Table 3.

## Cosine similarity of documents

The cosine similarity of documents indicates that the conversation that took place in Week1\_Q1, Week1Q\_2 and Week2\_Q2 was highly similar, whilst the conversation in Week2\_Q3 was somewhat distinguished from the rest. The results of the cosine similarity calculated using tf-idf measure is presented in Table1.

Table : Cosine similarity

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Docs per week | W1\_Q1 | W1\_Q2 | W1\_Q3 | W1\_Q4 | W2\_Q1 | W2\_Q2 | W2\_Q3 |
| W1\_Q1 | 1 | 0.844\* | 0.765 | 0.793 | 0.746 | 0.846\* | 0.692 |
| W1\_Q2 | 0.844\* | 1 | 0.752 | 0.774 | 0.776 | 0.858\* | 0.666 |
| W1\_Q3 | 0.765 | 0.752 | 1 | 0.707 | 0.668 | 0.733 | 0.567 |
| W1\_Q4 | 0.793 | 0.774 | 0.707 | 1 | 0.697 | 0.773 | 0.569 |
| W2\_Q1 | 0.746 | 0.776 | 0.668 | 0.697 | 1 | 0.787 | 0.613 |
| W2\_Q2 | 0.846\* | 0.858\* | 0.733 | 0.773 | 0.787 | 1 | 0.681 |
| W2\_Q3 | 0.692 | 0.666 | 0.567 | 0.569 | 0.613 | 0.681 | 1 |

Table : Summary of collocations

| Week | Question | No. of Sentences | Collocations |
| --- | --- | --- | --- |
| 1 | 1 | 5 | ('alpha', 'blocker'), ('actual', 'study') |
| 1 | 2 | 29 | ('tamsulosin', 'hydrochloride'), ('bladder', 'cancer'), ('prostate', 'cancer'), ('cancer', 'diagnosis') |
| 1 | 3 | 4 | ('believe', 'dipstick'), ('urine', 'dipstick') |
| 1 | 4 | 6 | ('exclusion', 'criterion'), ('clinical', 'trial') |
| 2 | 1 | 24 | ('tamsulosin', 'hydrochloride'), ('prostatic', 'hyperplasia'), ('benign', 'prostatic'), ('enlarged', 'prostate'), ('prostate', 'urinary'), ('symptom', 'reliever'), ('urinary', 'symptom') |
| 2 | 2 | 18 | ('tamsulosin', 'hydrochloride'), ('prostate', 'cancer'), ('self-selection', 'decision'), ('underlying', 'condition'), ('self-selection', 'error') |
| 2 | 3 | 10 | ('liver', 'kidney'), ('kidney', 'disease'), ('urinary', 'retention'), ('think', 'might'), ('urinary', 'symptom'), ('diabetes', 'prostate'), ('bladder', 'cancer'), ('bladder', 'prostate'), ('prostate', 'bladder') |

From the analysis of collocation in Table 2, some of the important themes of the documents can be extracted. Week1\_Q2 stands out in this regards as the discussion in this week seemed to cover a lager area of topic as seen in the dispersion plot, week1\_Q1 appear to relate to the trial of the drug, while week2 Q1 and Q2 was ‘consumer’ orientated.

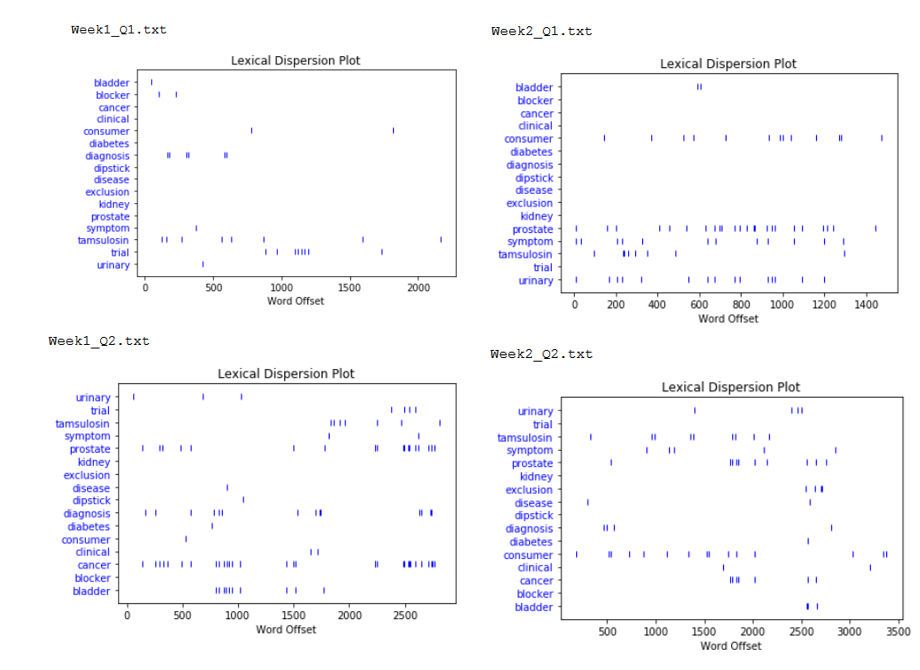


Figure : Lexical Dispersion Plot of Key words

# Analysis of corpus

## LDA

To discover the main topics, the focus group is talking about, LDA (Latent Dirichlet Allocation) is used which gives the best results on the focus group data using four clusters of words. The main topic in the corpus is a study about prostate cancer symptoms and the second biggest topic is consumer product labels that are related to prostate cancer. Some of the important topics are listed in Table 4 in order of importance.

Table : Topics derived from LDA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Terms | Topic 1 | Topic 2 | Topic 3 | Topic 4 |
| 1 | Study | symptom | population | symptom |
| 2 | Consumer | prostate | information | conditions |
| 3 | Think | cancer | study | concern |
| 4 | Might | tamsulosin | label | diagnosis |
| 5 | Cancer | hydrochloride |  | product |
| 6 | Bladder | condition |  |  |
| 7 | patient |  |  |  |

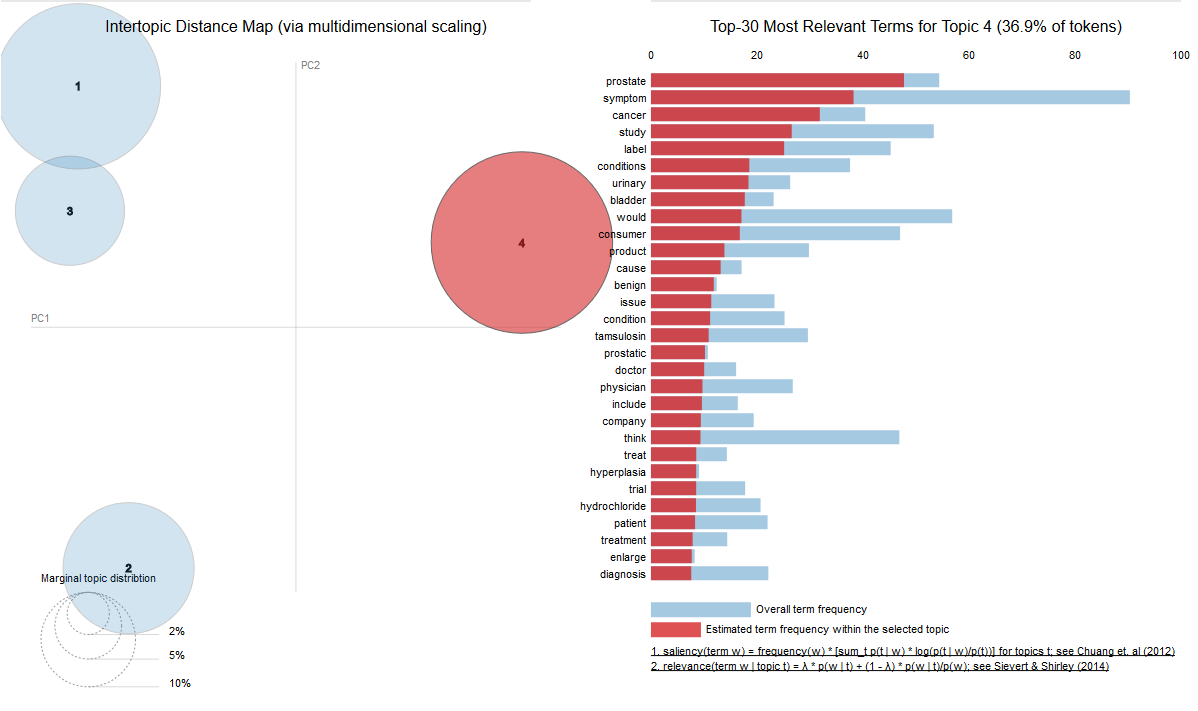


Figure : LDA Visualization

## Word2vec

Word2vec comes in action to discover words in the corpus that share common context which produces vector space of several hundreds of dimensions, to reduce the dimensionality TSNE is used to word vectors to transform the 64-dimensional space into a 2D space and by using the bokeh.iobokeh.io library we can visualize the output.

Taking a closer look at the output in Figure 4 below, we can notice that “company” “consumer” “drug” usually come in the same context, “physician” and “doctor” come in the same context and “patient” and “symptom” come in the same context as well.

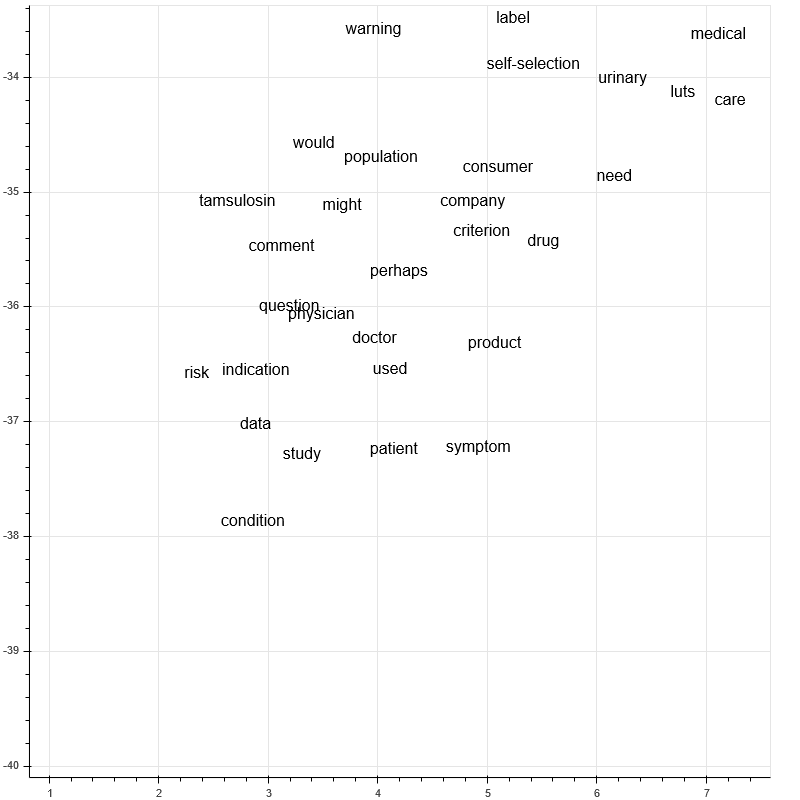


Figure : Word2Vec closeup

# Sentiment Analysis

Sentiment analysis was conducted to identify the affective opinions and attitudes towards switching tamsulosin from a prescription to Over the Counter medication expressed by Medical Doctors participating in the focus group.

The analysis was conducted using the SentiWordNet3.0 Lexicon Dictionary for each week and question. The original lexicon was manually enhanced to better fit the specificity of the medical dataset by manipulating certain words’ scores. For instance, ‘risk’ and ‘cancer’ in all contexts were changed to negative 0.9, while words that may suggest benefits of the treatment were assigned more positive scores. The final score is a weighted average for all words in each question; with score above 1.5 sentiment is positive, score below 1.5 is negative, and scores in between indicate a neutral attitude.

Subjectivity of the discussion topics for each question was generated through TextBlob. Subjectivity refers to being influenced by private opinions and beliefs, while objectivity refers to measurable facts generally agreed upon societies; 0 is objective and 1 is subjective. Product. Table below shows the key concepts for each question and their corresponding sentiment scores.

Discussion in week 1, question 2 is the most negative and suggests that the treatment containing Tamsulosin Hydrochloride might cause bladder cancer and prostate cancer. The following weeks suggest that the medication may also lead to liver and kidney disfunctions, prostatic hyperplasia, as well as diabetes. The neutral responses touch mainly upon the methodology of the study including clinical trials, exclusion criteria, etc. The positive answers include possibly optimistic outcomes of the treatment with an emphasis on the benefits of alpha blockers that are good symptom relievers. In week 1, question 3, the group discusses the results of a urine test strip that seems to be encouraging and supportive of the switch to OTC. Discussion topics within each question score between .43 and .50 meaning they are closer to being objective.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Week | Key concepts | Sentiment | Score | Subjectivity |
| Week 1 Q1 | alpha blocker, actual study | Positive | 9.725 | 0.45 |
| Week 1 Q2 | tamsulosin, hydrochloride, bladder, cancer, prostate, cancer, diagnosis | Negative | -18.18 | 0.43 |
| Week 1 Q3 | urine dipstick | Positive | 6.2 | 0.46 |
| Week 1 Q4 | Exclusion criteria, clinical trial | Neutral | 0.825 | 0.47 |
| Week 2 Q1 | tamsulosin, hydrochloride, prostatic, hyperplasia, prostatic, enlarged, prostate, prostate, urinary, symptom reliever, urinary | Negative | -8.4 | 0.46 |
| Week 2 Q2 | surgery, tamsulosin, hydrochloride, exclusion, criterion, practice, standard, prostate, cancer, study, underlying, condition, self-selection, symptom | Neutral | 0.05 | 0.47 |
| Week 2 Q3 | liver, kidney, disease, urinary, retention, symptom, diabetes, prostate, bladder, cancer | Negative | -3.65 | 0.5 |