Erasmus University Rotterdam

MSc in Economics and Business

with specialisation in Data Science and Marketing Analytics

Master Thesis

Aspect-Level Sentiment Analysis using Patient-Authored Online Reviews for Oral Contraceptives

Author
Alina Heimgartner

Student Number
530903

Supervisor
Dr. Anastasija Tetereva

Second Assessor
Prof. Dr. Bas Donkers

July 2020

Abstract

Online reviews have become an increasingly important source of information for women when making their contraceptive choice, and thus for pharmaceutical companies, to understand their customers. However, it remains a challenge to obtain insightful information from the vast amount of textual data on the Internet. This problem is increasingly approached by innovative methods in the field of text mining. Sentiment analysis is a sub-field of text mining which particularly aims at extracting and classifying opinions found in textual data. This thesis contributes to the field of sentiment analysis by presenting a novel approach, which extracts more meaningful aspects, yields best-in-class accuracy of 81% and visualises the insights in a Perception Map, showing the relevance and sentiment of the extracted aspects. The thesis found that the most substantial benefits of using oral contraceptives are about not gaining weight and having light(er) cramps, while the most pressing concerns are about depression, anxiety, nausea and not bleeding. Additionally, aspects less commonly known, such as the loss of sex drive and tender breasts, have emerged and aspects expected to be more relevant, such as effectiveness, have turned out to be not as relevant. These insights can be leveraged in many marketing-related use-cases, for example, to define gaps in external communication or identify new concerns of customers not yet addressed appropriately.

Table of Contents

List of Figures

List of Tables

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Anastasija Tetereva. With her expertise and patience, she guided me from the start, encouraged me to think further and shared her knowledge and time generously. I would also like to thank my second assessor, Prof. Dr. Bas Donkers, for his lessons in text analytics, which sparked my interest in the field and gave me a solid theoretical foundation. Third, I would like to thank my friends for giving me the inspiration to look into reviews on contraceptive methods and for their unconditional friendship. Lastly, I would like to thank my parents for their endless love and support. My mother, for encouraging me to pursue a master. You are my hero and role model. My father, for always taking care of me, for being my teacher and making sure I stay healthy.

1 Introduction

In 2019, out of 1.9 billion women at a reproductive age, 842 million used modern methods of contraception, such as the pill (United Nations, 2019). The Sustainable Development Goals (SDGs) have declared it as one of their targets to ensure universal access to contraceptive methods, including the availability of information and education until 2030 (UN DESA, 2019; target 3.7). When women need to get informed about contraceptives, they traditionally consult their doctor, friends or family members (Melo et al., 2015). However, with the increasing availability and use of the Internet, information on contraceptives has become more accessible. Women can share their experience and opinion with a broader audience, discuss questions anonymously and share unexpected effects of a contraceptive method. Therefore, the Internet offers a new place to share and acquire information for patients. At the same time, it is an added source for companies to understand their customer's opinion better. Besides, research has shown that online reviews increasingly influence the decision-making process and are thus becoming critical for sales. As a result, pharmaceutical companies are increasingly aware of the significant impact of online reviews as a source to understand the opinions, needs and questions of their customers and ultimately to make their product more successful.

However, synthesizing a large amount of textual data remains a challenge. For example, a search for reviews on a contraceptive pill returned over 1000 results in a medical forum. To gain a summarized understanding of those reviews, one would need to read each review, which would be overwhelming and time consuming (Liu, 2012). Text mining proposes automated solutions to this problem by transforming unstructured textual data into meaningful information. A sub-field of text mining is sentiment analysis, which focuses mainly on the extraction, classification and interpretation of opinions or emotions in textual data, most frequently using online reviews as a data source (Liu, 2012).

This thesis focuses on a specific area of research in sentiment analysis, referred to as aspect-level sentiment analysis (Liu, 2012). Aspect-level sentiment analysis is applied when a more detailed view of the level of sentiment is needed. Other fields of sentiment analysis focus primarily on identifying the general sentiment of a sentence or review. In aspect-level sentiment analysis, the focus lies on identifying the sentiment of several aspects of a product or service. In the case of a medical drug, this could be its effectiveness, price or side effect. This is especially useful for companies, because it allows them to understand the needs and opinions of customers in more

detail and adds an additional source for market research, beyond the traditional methods (e.g. clinical studies, interviews, focus groups).

Up to this point, it has remained difficult to identify meaningful aspects, categorise them and present them in an intuitive way to support decision-making in business. In addition, previous studies on aspect-level sentiment analysis have focused mainly on restaurant, hotel or product reviews. Only a few studies used health-related reviews, and none have been applied to contraceptive reviews. Therefore, the work in this thesis investigates how to improve the extraction, classification and interpretation of meaningful and relevant aspects of oral contraceptive reviews. In particular, this thesis answers the research question: *What are the relevant and meaningful positive and negative aspects that are expressed in patient-authored online reviews for oral contraceptives?*

The methodological approach to answer this research question follows four steps.

1) For aspect extraction, different extensions of topic modelling methods are tested to create meaningful aspects.
2) For aspect classification, the performance of the aspects from step (1) is compared against commonly applied features for sentiment classification.
3) For aspect classification, the performance and interpretability of several different classifiers are compared.
4) For aspect interpretation, a Perception Map is created, which illustrates, in one dimension, how relevant an aspect is, and in another dimension, the sentiment towards that aspect.

Following the above steps, the thesis contributes to aspect-level sentiment analysis by presenting a novel approach that improves the extraction and interpretation of meaningful aspects and illustrates their relevance (i.e. how often do reviewers talk about this aspect) and sentiment (i.e. how positive or negative is this aspect perceived). The approach uses an extended topic modelling method, which produces best-in-class results. A Logistic Regression model is recommended for the classification task, as it achieves the best interpretability while performing similar to other, more complex classifiers.

From a marketing perspective, this thesis contributes by adding a simple, yet powerful visualisation called the Perception Map, which plots the aspects along two dimensions, their relevance and sentiment. This map is specifically useful to non-experts, such as marketeers, to quickly identify USPs and concerns found in contraceptive reviews, define gaps in external communication, such as websites or brochures or identify new concerns not yet addressed appropriately.

2 Theoretical Background

The first part of this chapter focuses on the theories which surround sentiment analysis from a marketing perspective. The second part of this chapter focuses on sentiment analysis from a technical point of view. The third part includes a short literature review on the most relevant health-related literature for sentiment analysis.

2.1 Marketing Perspective

Word-of-mouth (WOM) is defined by Arndt (1967) as a person to person communication, where the person receiving the message perceives the communicator as non-commercial regarding a product, brand or service. With the increasing use of the Internet, the WOM has transformed from a one-to-one communication between two people towards a one-to-many communication (Kumabam et al., 2017). This led to the increasing relevance of online WOM, in which customers across the world can share their experiences through blogs, forums or reviews. In this thesis, the focus lies on texts from online reviews. Research has shown that reviews have an important impact on the purchasing decisions of customers, and thus a company's sales, as well as other factors such as brand awareness and brand perception (Duan, Gu & Whunston, 2008).

The availability of reviews has changed the customer decision-making journey significantly (López & Sicilia, 2011; Varkaris & Neuhofer, 2017). From having only offline touchpoints, towards many cross-channel interaction points. In general, the customer journey consists of a pre-purchase phase, in which a customer receives the information and considers their options; a purchase phase, in which the customer interacts with the seller and purchases the product; and a post-purchase phase, in which the customer uses the product (Neslin et al. 2006). However, depending on the product or service, the journey may vary. This is specifically the case for prescriptive drugs. Strict regulations are in place on how a company is allowed to market and sell its products. For example, direct-to-consumer marketing is forbidden in many European countries. Purchasing decisions must be discussed and approved by a doctor, and costs are covered mostly by insurances (Kornfield et al., 2013).

For contraceptive methods, consultations from the immediate social network, mainly friends and family, are extremely important in addition to the consultation of a doctor. Melo et al. (2015) developed a framework (Figure 1.), which shows the influence of peers on the decision journey for contraceptives. Considering that more women are using the web as their source for peer

influence, the relevance of online reviews for the decision-making process is crucial, and thus their content valuable, both for patients and for pharmaceutical companies.

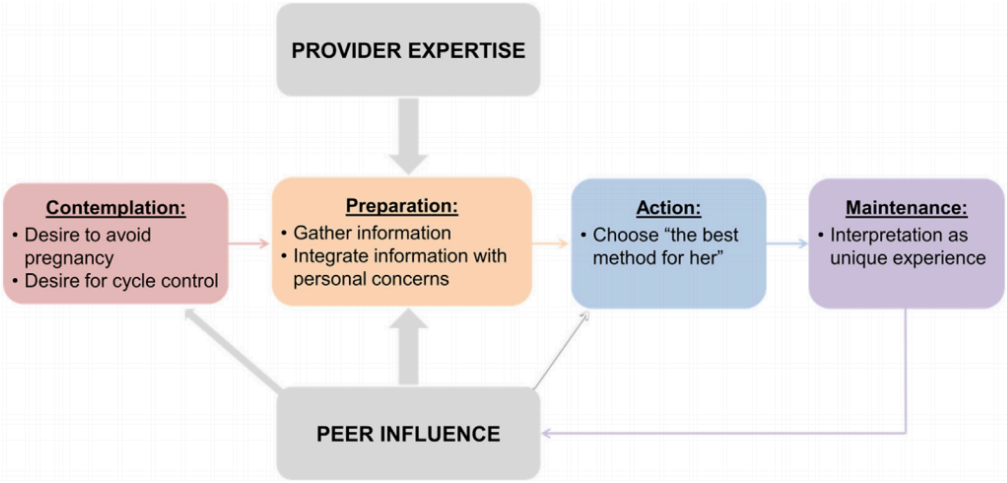


Figure 1. Conceptual framework on patient journey for contraceptive methods (Melo et al., 2015)

Pharmaceutical companies are just starting to utilise the large amount of textual data from patients, which can be leveraged, for example, to develop better drugs, find new opportunities, screen for Adverse Drug Reactions (ADR) or improve the patient journey. Within the marketing research field, gaining access to the opinions of customers, especially for pharmaceutical companies, has long been expensive, difficult and highly regulated. For most, it includes purchasing market research reports or hiring a market research agency that performs surveys and collects data. With patients expressing their experiences online, a new source to gain customer insights has appeared, which can potentially be more authentic in comparison to responses of customers to qualitative or quantitative surveys, as the motivation is intrinsic (sharing and caring) rather than extrinsic (getting paid) (Berger et al., 2020). However, there may be a biased in textual review data due to self-selection amongst people more inclined to write a review (Schoenmüller, Netzer and Stahl, 2019).

2.2 Technical Perspective

Sentiment analysis is a research area that explores the sentiments or emotions found in textual data from several people towards an entity. The entity is mostly a product or service, but can also be an individual or topic and their characteristics (Liu & Zhang, 2012). In general, the main purpose of sentiment analysis is to find out if a text indicates a positive, neutral or negative sentiment. The terms sentiment and opinion are used interchangeably throughout this thesis. The term "opinion" is defined as "judgment or belief not founded on certainty or proof" (Collins English Dictionary, 2015, as cited in Schouten & Fransincar, 2016). The main characteristics of sentiments are that

they are subjective, in contrast to factual information. As an opinion is subjective, a single sentiment from one person to an entity provides only one perspective and cannot be used to create a summary or to generate insights. However, by utilizing sentiment analysis and applying it to many reviews, the opinions of many people can be summarized and used for interpretation (Liu, 2012).

2.2.1 Level of Analysis

Sentiment analysis has been studied on different levels, most often on document and sentence level. On a document-level, sentiment analysis aims to identify whether the review expresses a positive or negative sentiment. On the other hand, sentence-level sentiment analysis focuses not on the whole review but on each individual sentence. Both approaches lack the ability to identify which aspects of a review or sentence are positive or negative. The problem with this level of analysis is that a review or sentence likely includes both positive and negative aspects for a product or service. Therefore, identifying a single sentiment for a review or sentence can lead to generalized conclusions and does not provide specific information on what a customer likes or dislikes. As a result, document- or sentence-level classification lacks the level of detail that is needed to fully understand the customer's opinion and is thus less useful in many business-related use-cases.

Aspect-level sentiment analysis, first introduced by Hu and Liu (2004), solves this problem by extracting aspects from the reviews and categorizing them into positive and negative ones. The results are thus more granular and can be leveraged in a wider range of applications, specifically in the field of marketing (Liu, 2012).

2.2.2 Aspect-Level Sentiment Analysis

Aspect-level sentiment analysis is also known as feature-level analysis or feature-based opinion mining (Hu & Liu, 2004; Pang & Lee, 2009; Liu, 2012). The core idea of aspect-level sentiment analysis is that an opinion is not just negative or positive, but includes different aspects which can be positive or negative, depending on the aspect. For example, in the reviews for a contraceptive drug, the drug is the entity and characteristics related to the drug (e.g., side effects, usage, dosage, etc.) are its aspects. The aim of the aspect-level analysis is to find the sentiment towards the aspects. The example "The pill is expensive, but I have fewer stomach cramps" contains two aspects: the price and a side effect of the pill. The sentiment expressed towards the target aspect price is negative, whilst the sentiment expressed towards the side effect is positive.

The two main tasks in aspect-level sentiment analysis are the extraction and the classification of the aspect. Aspect extraction is often achieved by using one of the following four approaches: frequency-based, syntax-based, supervised and unsupervised machine learning (Schouten & Frasincar, 2015). This thesis focuses on unsupervised machine learning. The rationales behind

using an unsupervised approach are as follows: extracting aspects using word frequency is the simplest and most common approach. It assumes that words which frequently occur are most likely to be important aspects. However, the main disadvantage of the frequency-based method is that some nouns occur many times and are thus wrongly used as aspects. In addition, words which describe a similar aspect are not recognised as one aspect and are counted several times. Therefore, to gain meaningful aspects, the frequency-based method has some limitations, and its performance is relatively poor in comparison to other approaches for aspect extraction (Hu & Liu, 2004; Long et al., 2010; Hai et al., 2011; Liu et al., 2005). Syntax-based extraction uses the syntactical relation between words to identify aspects. This approach is useful if the dataset is small as it relies on a limited set of data to work. The approach requires a significant amount of linguistic expertise, which lie outside of the scope of this thesis. The use of supervised machine learning methods for aspect extraction is part of the general information extraction research. Supervised methods need a labelled dataset to train, which would require to manually label a dataset into words that are considered an aspect and words not considered an aspect. As this is time-consuming and depends on the subjective perception of what an aspect might be, it is rather unpractical to apply in business. Thus, it is not further pursued in this thesis. Unsupervised machine learning methods, in particular topic modelling, are potentially a more suitable method towards aspect extraction, as they generate topics that can be used directly as aspects (Schouten & Frasincar, 2015). The most common method to extract sentiment using an unsupervised approach is the Latent Dirichlet Allocation (LDA), which is described in more detail in the next chapter.

The second task in aspect-level sentiment analysis is the aspect classification, in which a polarity (e.g. positive or negative) or a score is assigned to an aspect (Schouten & Frasincar, 2015). The three main approaches applied to aspect classification are dictionary-based (linguistic approach), supervised and unsupervised machine learning approaches (Liu, 2012). In this thesis, a supervised approach is presented. The rationale behind using this approach is the following. Unsupervised methods do not rely on a labelled dataset to predict the sentiment. As this thesis focuses on reviews which include their overall rating, the scores may subsequently be used to represent labels in a supervised approach. The dictionary approach includes the construction of a dictionary containing words that express a sentiment. Each document is assigned a score depending on the frequency of the positive and negative words it contains (Taboada et al., 2011). Such methods may handle complex grammatical relations. However, this approach is static and must be maintained and updated regularly.

2.2.3 Latent Dirichlet Allocation (LDA) for Aspect Extraction

Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2003) is a flexible and strong topic modelling algorithm. It is most commonly applied to textual data to generate several topics. LDA has been applied in various ways to extract sentiment from textual data (e.g. Titov & McDonald, 2008; Lu, Ott & Cardie, 2011; Lakkaraju et al., 2011; Hai et al., 2014). Lu et al. (2011) demonstrated 80.3% accuracy when applying a local LDA approach to classify the sentiment of restaurant reviews. Moghadam and Ester (2011) proposed a probabilistic graphical model based on LDA, the Factorised LDA, to address the cold start problem, demonstrating 79%-86% accuracy. LDA has gained attention for aspect extraction because it can, at the same time, identify expressions that describe aspects of an entity as well as cluster semantically related expressions (Wang et al., 2014). However, existing approaches using LDA have some limitations and drawbacks. For example, a topic generated from an LDA may not be deductible into a meaningful topic. As LDA applies the bag-of-words assumption[1], the words that appear to relate to one topic may not be semantically related to each other, which makes it difficult to characterize and interpret the topic (Schouten & Frasincar, 2015). Another important drawback is that LDA produces high-level topics called global topics. This is a problem as most reviews focus on the same aspects, which makes them relatively consistent. This makes it easy to find general and frequent aspects but hard to identify smaller, less frequent aspects that are potentially more relevant for aspect-level sentiment analysis and ultimately more interesting for business (Titov & McDonald, 2008). Furthermore, LDA relies on a large dataset and needs a significant amount of tuning in order to achieve good results (Schouten & Frasincar, 2015).

2.3 Health-Related Perspective

Sentiment analysis has mainly been applied to reviews for a range of different services, such as hotels or restaurants, entertainments such as movies and books, or for consumer goods products. The use of health-related reviews for text mining in general, and specifically for sentiment analysis, is a new area of research (Carrillo-de-Albornoz et al., 2018). This has mainly to do with issues in the quality and trustworthiness of patient-authored reviews (Na et al., 2012). However, over the past years, both quality and trustworthiness have improved and are expected to rise further, as addressed in more detail in Chapter 4.4. This section provides a short literature overview on the current approaches in the field of sentiment analysis using patient-authored texts.

[1] The bag-of-words assumption refers to a frequently used premise for textual data. It assumes that the order of words does not matter. All that matters is whether or not a word occurs.

In 2009, Denecke and Nejdl published a content analysis on health-related information, specifically for medical blogs, reviews and wikis. They used both patient and doctor written posts and distinguished between informative and affective comments. They used simple, frequency-based features, but also more complex features, such as lexical features based on the Unified Medical Language System (UMLS), achieving 78.59% in *F*-measure[2]. As a result, they were able to show the content differences from several types of sources, e.g. blogs or wiki, and for the type of person writing the review, e.g. patient, nurse or doctor. They suggested that their findings could be used to improve the rankings of search engines in order to direct users to the best knowledge source.

In 2012, Goeuriot et al. (2012) investigated the use of opinion mining for user-generated content on drugs and medications. They report that opinion mining, at that time, was not well explored because the trustworthiness of reviews from patients was rather low. They created a medical lexicon based on drug reviews, which included general and medical opinion words from existing lexicons. Na et al. (2012) tested a purely linguistic technique which considers the grammatical dependency of words to classify the sentiment of reviews on drugs. They made use of MetaMap, a tool to review medical text to increase domain expertise. Their linguistic approach exceeded the accuracy of their base-line model by five percentage points, resulting in 78% accuracy.

Bobicev, Sokolova, Jafer and Schramm (2012) classified Tweets on personal health information using sentiment analysis. In a first step, they labelled tweets as positive, negative or neutral. In a second step, they applied different machine learning algorithms on multi-class and binary classifications. By using correlated words as features, they achieved 74.5% in *F*-measure.

A more recent study by Carrillo-de-Albornoz et al. (2018) evaluates different sets of features, including semantic, sentiment-based, network-based and word-embeddings, to predict the sentiment of patient-authored reviews. They applied different learning algorithms in which it was possible to predict polarity with a 67.2% accuracy, using word embeddings, which significantly outperformed the more traditional features based on the bag-of-words assumption. Most similar to this study, Salas-Zárate et al. (2017) performed an aspect-level sentiment analysis for diabetes twitter data. The sentiment of the aspects was calculated by using n-grams and word-frequency measures, achieving an *F*-measure of 81.24%.

To summarize, both machine learning and linguistic techniques have been applied to perform sentiment analysis on health-related texts. However, aspect-level sentiment analysis has been underrepresented in the health domain, and no studies have investigated sentiment analysis for patient-authored texts on contraceptives.

---

[2] *F-measure* is the performance metric that calculates the natural mean between precision and recall.

3 Methodology

This chapter introduces in more depth the methods used during the analysis and the reasoning behind the approach. The aim is to explain the concepts applied in Chapter 5, which presents the overall results achieved.

3.1 LDA

In Chapter 2, LDA was introduced as the preferred topic modelling method for aspect-level sentiment analysis. The LDA is an unsupervised machine learning method, which aims to convert a selection of texts to a set of topics. It was first introduced by Blei et al. (2003) and is built on the premise that every document consists of topics and that topics consist of the unique words found in the documents. Patterns are formed when words co-occur frequently. In this thesis, LDA is used to extract aspects from the corpus, which are later used as features in the classification task. Second, the average probability of a topic occurring in the reviews is used to give an indication of how relevant an aspect is.

LDA is a soft-clustering method, in contrast to hard-clustering methods, such as k-means (Airoldi et al., 2015). The main difference is that for LDA, each document belongs to all the topics, but with different probabilities. In hard-clustering methods, one document can only belong to one topic exclusively. As a review often does not belong to just one topic, LDA is preferred over hard-clustering methods. The outcome of the LDA is represented by "a vector of continuous non-negative latent variables that add up to 1" (Airoldi et al., 2015, p. 4). LDA thus observes the word frequency distribution among all documents to find k, the number of topics that emerge. Each document belongs to a topic with a different probability. Most words of a topic have a probability close to zero. Only a few words receive a larger probability and are therefore more indicative of what the topic is about. Usually, only a few words with the largest probability are selected and used to interpret and name the topic. In order to understand how likely a topic occurs, the average topic probability is calculated, taking the mean probability for each topic over all reviews (Kwartler, 2017).

More technically, LDA is a probabilistic generative process that calculates a joint probability distribution over the words of the documents, called observed variables, and the topic structure called the hidden variables. Figure 2. shows the observed variable, $W_{d,n}$, as a group of words in document N and all the documents D. There are several dependencies. First, the specific word

$w_{d,n}$ depends on all topics $\beta_{1:K}$, in which β_K represents the distribution for all words on a topic k, and the hidden topic measure for the n^{th} word in document d, named $z_{d,n}$. In return, $z_{d,n}$ depends on θ_d, which is the distribution per document (Reisenbichler & Reutterer, 2018). The Dirichlet distribution is applied beforehand on the topic word distribution η and the document topic distribution α.

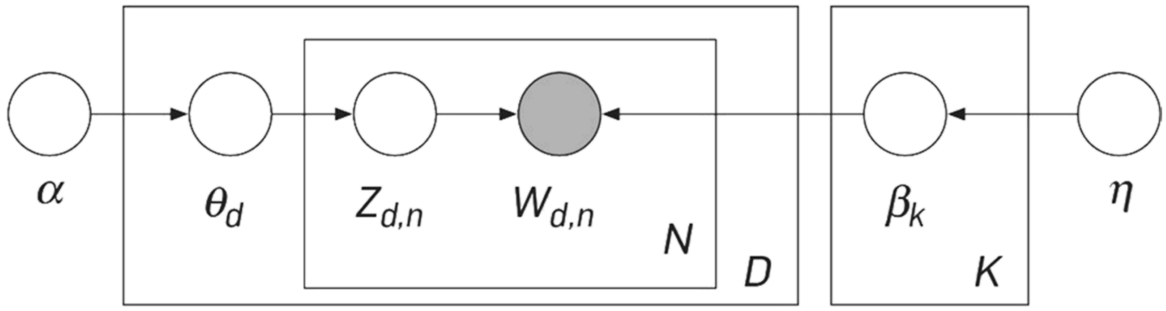


Figure 2. Graphical figure for LDA dependencies (Blei, 2012).

Mathematically, this is solved by calculating the conditional distribution of the topic structure based on the observed documents, the posterior (Blei, 2012). The posterior is defined by

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

in which the denominator is the likelihood of assigning the observed documents a random topic model and the numerator is the joint distribution of all the variables, which can be calculated for any set of hidden variables (Blei et al., 2003).

This thesis uses a built-in function in the R programming language to calculate the LDA, which requires the tuning of two parameters. The first parameter to tune is the number of topics, *k*, that are extracted from the document. This is done by splitting the dataset into a training and validation set and iterating over different numbers of *k*, for example, 5, 10, 15 and 20 topics, to find the smallest perplexity score for the validation set. A validation set is needed in order to avoid overfitting. The perplexity score indicates the accuracy of how well the probability distribution predicts a sample (Blei et al., 2003). It calculates the modeling power of LDA with the given parameters by using the inverse probability of unobserved documents, as shown in the formula below:

$$\text{Perplexity of set of documents} = \exp\left(\frac{-\log(\Pr[\text{all words in docs}])}{\text{Total number of words}}\right)$$

The lower the perplexity becomes, the better is the model, which means that there is less uncertainty about the unobserved documents. Minimizing the perplexity is thus similar to maximizing the probability. Of course, there are other factors that can be taken into consideration, such as interpretability. For some applications, having 50 or more topics leads to uninterpretable results, although it is where the perplexity is lowest (Kwartler, 2017).

The second parameter to tune, α, is responsible for controlling the document-topic density, also called the sparseness of the topic distribution. A high α increases the probability that a document is composed of a mix of relatively equally dominant topics. For example, with a high α, the topic distribution, (0.01, 0.39, 0.6), would more likely be (0.3, 0.3, 0.4). The variance is in this case low. Contrary, for a very low α, the probability is higher that each document is composed of only one or a few topics, in which case the variance is high. This concept can be understood further when looking at the Dirichlet distribution, which shows how different α can control the sparseness, for $p \sim$ Dirichlet ($\alpha 1, \alpha 2, .. \alpha k$)

$$\text{Expectation } pk = \mathrm{E}[pk] = \frac{\alpha k}{\sum_j \alpha j}$$

and

$$\text{Variance } pk = \mathrm{Var}[pk] = \frac{\mathrm{E}[pk](1 - \mathrm{E}[pk])}{1 + \sum_j \alpha j} \quad (1)$$

in which Here $\sum_j \alpha j$ is controlling the sparseness. In (1), a lower value of α leads to larger variance, and thus more sparseness. This increases the probability that a document will be just about one or two topics. In return, a higher α decreases the variance, which decreases sparseness, and thus it is more likely that a document is about a mixture of equally dominant topics (Kwartler, 2017).

The LDA is extendable to many use-cases due to its open framework. The only requirement is a large dataset that includes discrete units that are distributed unequally. The flexibility of LDA has led to a lot of extensions in research and is the basis for various types of new topic models, such as Dynamic LDA (Tam & Schulz, 2005), Correlated Topic Model (CTM) (Blei & Lafferty, 2006) or the Supervised Topic Model (Ramage, 2009). In this thesis, the LDA model is extended by not only using one word as an input for the LDA but two or three words that appear next to or near each other. This relaxes the bag-of-words assumption as the order of the words occurring becomes slightly more relevant (Stojanovski et al., 2018). As neighbouring words are taken into account, the order matters slightly more, and thus this assumption is relaxed. Using not just a single word but two or more words is called an n-gram or skip-gram. In the next section, they are explained in more detail.

3.2 N-grams and Skip-grams

N-grams are made up of single words, called unigrams, or a group of words. If a sequence of two or three words is generated from text, they are called bi-, respectively tri-, grams. An N-gram representation of text follows the same assumptions as the bag-of-words approach. Each n-gram is a unique feature of the text. As with most approaches that rely on the bag-of-words assumption, using n-grams is simple, can be done quickly and is computationally inexpensive. This is due to the simple organization of the words and documents in a matrix format, called the Document Term Matrix (DTM). Each row of the DTM is represented by a document or review, and each column is represented by a single word or n-gram (Kwartler, 2017).

The simplest statistical approach to find n-gram features is to use the most frequently used n-grams in the corpus as polarity indicators. This approach has proven to be effective and is fully automatic. The main disadvantage is that some texts which contain a lot of sentiment might be ignored and that the individual words might not be very meaningful for subsequent analysis (Duric & Song, 2012). Pang and Lee (2009) found that most reviews have many different words to describe sentiment, and thus, they might not appear frequently. Therefore, the following approach can lead to words that appear often but are not the most subjective words. Usually, using a sequence of 2 to 3 words improves the performance of a classifier compared to using unigrams (Fürnkranz, 1998).

A simple extension of n-grams is the skip-gram approach, or more precisely the k-skip-n-gram technique (Guthrie et al., 2006), which allows the skipping of a maximum of k words to overcome the data sparsity problem of traditional n-grams. Skip-grams are more general than the traditional n-grams as words do not need to be consecutive. They have been used mainly in supervised machine learning and language modelling tasks, usually in combination with other approaches, with the overall aim being to improve performance (Fernández et al., 2014; HaCohen-Kerner et al., 2017; Sidorov et al., 2014). An example of a 2-skip-bi-gram of ABCDE may be AC, BD or CE but also AB. Guthrie et al. (2006) formulates the k-skip-n-grams for a review or document $w_1 ... w_m$ as

$$\left\{ w_{i_1}, w_{i_2}, \dots w_{i_n} \mid \sum_{j=1}^{n} i_j - i_{j-1} < k \right\}$$

in which *n* is the number of words and *k* the number of maximum skips. One drawback is that the number of possible combinations of skip-grams is relativity large, depending on *k* and *n*.

3.3 Dictionary

Dictionaries are the most common and simple approach to extract sentiment from text. Words that express an opinion are the key indicators of sentiment in text and are called the opinion words, for example, *good* or *bad*. A dictionary is in its core a collection of words with a score assigned to them (Stojanovski et al., 2018). For example, for *good*, *great* or *happy* +1 and for *bad*, *terrible*, or *unhappy* -1. By counting the number of positive and negative opinion words, one may decide whether a review is negative or positive if the list of words is extensive enough. Over the years, researchers have investigated numerous ways to compile general and domain-specific dictionaries as well as improved existing ones by adding new rules, for example by using calculations to include the effect of negations and amplifications.

However, the dictionaries have several drawbacks. Firstly, they are static and thus need to be checked carefully before being applied in an unknown or complex domain (Berger et al., 2018). For example, the opinion word *gained*, in combination with *points gained* may express positive sentiment in the gaming industry, whereas *weight gained* is likely to indicate a negative sentiment in the health domain. Additionally, some opinion words may be used in context and not indicate a sentiment, for example, *"What makes you happy"*. Such dictionaries must also be curated, updated and carefully crafted, which requires time and a lot of expertise. Lastly, such dictionaries have difficulties handling sarcasm or irony. Nevertheless, despite their weaknesses, dictionaries are still frequently applied to extract sentiment from text, particularly as a baseline.

3.4 Word embeddings

The aforementioned methods use the bag-of-words approach, in which the order of words is irrelevant. Word embeddings overcome the bag-of-words assumption as they use the words that surround a word to describe it. Thus, the order of the word matters, as it is the key indicator of what the word means based on the company it keeps.

In 2013, Mikolov et al. developed Word2vec. Instead of using the bag-of-words assumptions, where a word is represented by zero if it is not relevant and one if it is relevant, each word is represented by a vector of continuous values, called the continuous bag of words (CBOW). These word vectors are usually high-dimensional, dense and not interpretable, for example: [-2.51, -0.42, …, -1.27, +4.22] for the word *weight*.

Several parameters need to be defined when using Word2vec, or, as applied in this thesis, the similar method, Glove2vec (Pennington et al., 2014), which is available as a function in R. First, the window size needs to be selected. The window size is the number of context words that should be considered to predict the focal word. If the window size is equal to two and symmetric, the

word pill (focal word) in the sentence *"this **pill** works i took it days after **unprotected** sex and on ovulation day and got my period days late"* would have the independent variables {*this, works, i*} while the word *unprotected* would use {*days, after, sex, and*} as context words. Second, the number of latent dimensions needs to be pre-specified. If 50 latent dimensions are chosen, a word vector will consist of 50 values. These values represent the location of the word in a high-dimensional word embedding space. Using a similarity measure, such as cosine similarity, the similarity between the words that are closer to one another in this space can be calculated.

Word embeddings provide a more detailed representation of the words in a corpus. In order to use the word embeddings to represent a review, summary statistics, such as the mean of the word vectors, need to be calculated (Prakash & Rao, 2017). The words of the review thus define where, in the latent space, the review is placed. Using word embeddings in supervised sentiment classification has yielded state-of-the-art results in recent years (Pennington et al., 2014; Tang et al., 2014; Yu et al., 2017). For example, in the paper "Feature engineering for sentiment analysis in e-health forums", Carrillo-de-Albornoz, Vidal and Plaza (2018) test a broad range of traditional features and compare the accuracy against word embeddings. They found that word embeddings perform considerably higher in comparison to more traditional features that are based on the bag-of-words assumption. Thus, word-embeddings set the aspiration level for accuracy in this thesis.

3.5 Logistic Regression

The Logistic Regression describes the association between a categorical dependet variable and one or more categorical or continuous independent variables. It is considered the model with the highest interpretability out of several classification methods due to its ability to measure how relevant a feature is through its coefficient size and the direction of the association (positive or negative) directly in the model output (Kwatler, 2017). Logistic Regression generally performs well when the data is linearly separable. It is a simple method and considered to be fast (Kuhn & Johnson, 2016). However, in order for the Logistic Regression to work well, several assumptions need to be fulfilled, such as independence of error terms, no multicollinearity or a large enough sample size compared to the number of features (Hoffmann, 2004). Overall, the Logistic Regression is a great baseline algorithm for classification tasks and interpretable, both on a global and local level.

From a technical perspective, the Logistic Regression models the probability that the dependent variable, *Y*, belongs to a particular category, expressed as $p(X)$. For example, if the probability of a review to be positive is 0.4, the review would be classified as being negative, assuming that the threshold for being positive is $p > 0.5$. The threshold can be changed when the data is imbalanced,

but if the data is balanced[3], as it is the case in this thesis, the threshold is normally set to 0.5 (Kwatler, 2017).

The mathematical function that expresses the above example is,

$$p(\mathrm{X}) = \frac{\mathrm{e}^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} ,$$

which provides outputs between 0 and 1 for all values of *X*. In this equation, β_0 is the *Y* intercept and β_1 is the coefficient size. With several manipulations and by taking the logarithm on both sides, the formula changes to

$$Logit(Y) = \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \qquad (2),$$

called the logit, which is linear in *X*. This means that the Logistic Regression in (2) has a linear *X*. Similar to a Linear Regression, β_1 explains the average change in *Y* with a one-unit increase in *X*. However, for the Logistic Regression, a change in *X* by one unit changes the log-odds by β_1. It is important to note that the relationship between *p(X)* and *X* in (2) is not linear, other than in the Linear Regression. Therefore, the coefficient size β_1 is not the actual change in *P(X)* associated with a one-unit increase in *X*. The actual change depends on the value of *X*. Nevertheless, the magnitude and coefficient direction can be associated with *P(X),* meaning that a positive β_1 is associated with an increase in *p(X)* while a negative β_1 is associated with a decrease in *p(X)*. To estimate the coefficients, β_0 and β_1, the maximum likelihood method is applied. In the method, the coefficients are fitted to minimize the difference between the actual and predicted outcome of an observation. In a formulized way, this is represented with the likelihood function

$$l(\beta_0, \beta_1) = \prod_{i:x_i=1} p(x_i) \prod_{i':yi'=0} \big(1 - p(x_i)\big) ,$$

which the coefficients aim to maximize. Thus, an optimal model returns probabilities close to one for positive reviews and close to zero for negative reviews.

3.6 Other Classification Techniques

This chapter compares the Logistic Regression to other well-known supervised classifiers which have proven to be effective for the text classification task (Devika, Sunitha & Ganesh, 2016). The dimensions discussed are mainly the interpretability of a method, but also its simplicity, accuracy and efficiency. The methods considered are SVM, Naïve Bayes, Decision Tree and Random Forest.

[3] A dataset is called balanced when the number of positive and negative classes is equally distributed.

3.6.1 Support Vector Machines (SVM)

SVM (Cortes & Vapnik et al., 1995) is known to perform well in text classification tasks due to the properties of text, which are, for example, that there are only a few irrelevant features, that there is a high dimensional input space, that the document vector is sparse or that text is linearly separable. SVM is well suited to handle those properties as its aim is to find a linear separator, and it can learn independently from the dimensionality of the input space. It is therefore protected towards overfitting, and there is no need for feature selection (Joachims, 1998).

In Figure 3., several loss functions are compared to each other. The figure shows that the hinge-loss function applied in SVM and the cross-entropy loss function applied in the Logistic Regression is relatively similar, while the exponential loss function is quite different. This means that when an observation is incorrectly classified, a comparable penalty size is given to that mistake in the SVM and the Logistic Regression. Thus, both methods usually behave relatively similarly. However, there are differences. The cross-entropy loss function is non-linear, steeper and never fully reaches zero. Thus, the Logistic Regression gives a greater penalty or weight to observations that are outliers. The Logistic Regression even puts a penalty on observations that are, with high certainty, correctly classified, while the SVM does not put any penalty on an observation once it reaches a certain point (Chorowski et al., 2014). This has the effect that the SVM is more robust to outliers than the Logistic Regression.

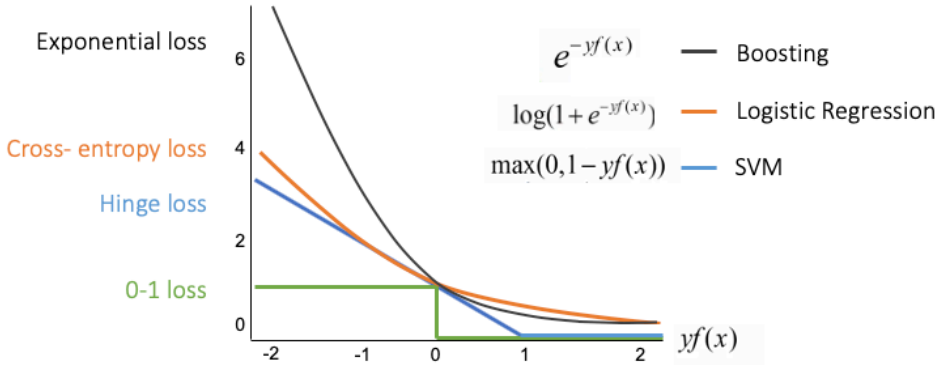


Figure 3. Comparison of loss functions (own illustration adapted from Roscher, 2013).

A limitation of the SVM is its lesser interpretability, as there is no probabilistic explanation for the classification. By calculating the variable importance, through permuting the values of each feature and checking how it changes the performance of the model, it is possible to find the feature importance. However, the direction of the impact is unknown and reveals nothing about the outcome on a local level. In addition, it is not suitable for large datasets as it is computationally expensive. To summarize, due to its lower interpretability and similar performance, the Logistic Regression is preferred over SVM.

3.6.2 Naïve Bayes

The Naïve Bayes (Friedmann et al., 1997) is a linear classification method based on the Bayes Theorem. It is a generative model, which means that it learns the combined probability, $p(x,y)$ of the input features, x and the outcome y. The predictions are based on the Bayes rules, calculating $p(y,x)$ and selecting the most probable class for y. In contrast, the Logistic Regression is a discriminative model that calculates probability $p(x,y)$ directly. Vapnik (1998) noted that "one should solve the [classification] problem directly and never solve a more general problem as an intermediate step.]. Ng and Jordan (2002) compared the performance of the Logistic Regression with the Naïve Bayes classifier using 15 different datasets. They found that when the training data is sufficiently large, the discriminative model (Logistic Regression) tends to outperform the Naïve Bayes. However, for fewer training sets, the Naïve Bayes reaches the asymptotic solution faster than the Logistic Regression. Based on the assumption that the dataset used in this thesis is sufficiently large and the lesser interpretability of the Naïve Bayes (similar to SVM), the Logistic Regression is preferred over the Naïve Bayes.

3.6.3 Decision Tree

The main difference between the Logistic Regression and the Decision Tree is in the way they generate the decision boundaries. While the Logistic Regression fits a single line to divide the space into two, the Decision Tree divides the space into smaller regions. The linear boundary used by the Logistic Regression can be restrictive when the data is non-linear, and this is where the Decision Tree can generate better results. Modelling a Decision Tree requires less effort during the pre-processing stage, as the data does not need to be scaled, correlated and there is no need to apply any regularization technique (Sharma & Dey, 2012). Most importantly, Decision Trees are an easy way to visualize and interpret the output as it is simple to understand and can be shown to non-experts. However, only a small tree can be examined, as for example, a tree with a depth of 10 can have hundreds of nodes, making it difficult to interpret. Another weakness of Decision Trees is that they are unstable, meaning that slight changes in the dataset can result in a completely different tree. Further, when classes are not well separated, trees tend to overfit. Most importantly, the predictive accuracy of a single Decision Tree is relatively low compared to other models. To summarize, due to the significant drawbacks in terms of stability and accuracy, the Logistic Regression is preferred over the Decision Tree.

3.6.4 Random Forrest

Random Forest cures the major drawbacks of the Decision Tree, such as stability and accuracy, but loses its biggest advantages, its high interpretability and visualization. Random Forest consists of a large number of trees that are each trained separately on a random subset of the data and a random selection of features per node. Thus, gaining a full understanding of the decision process is not feasible, and therefore, it is a black-box model[4], like SVM or Naïve Bayes, which is less interpretable than the Logistic Regression. Similar to SVM, one way of gathering insights into Random Forests is to compute feature importance. However, this gives little insight. Random Forest is expected to perform better if the data contains a lot of noise or if the data is not linearly separable (Kirasich, Smith & Sadler, 2018). However, from a technical point of view, the Logistic Regression is preferred over Random Forrest due to its interpretability.

3.7 Evaluation Metrics

In this chapter, the evaluation metrics used to evaluate the performance of the analysis in Chapter 5 are introduced. The most common performance metric is accuracy, which measures the number of correct classifications in relation to the total number of all observations, which is calculated as follows:

$$Accuracy = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative}$$

and provides the percentage of results correctly classified. In some cases, accuracy is not an ideal metric as it provides a false sense of performance. For example, if a model to detect cancer has an accuracy of 99%, one would assume that this is a good model. But in a scenario where 99 people do not have cancer and one person has cancer, the model can simply classify all people as not having cancer, reaching an accuracy of 99%. Of course, this model is completely useless. This problem is also referred to as the accuracy paradox and occurs in datasets that are imbalanced (Zhu & Davidson, 2007), which is not the case in this thesis. Nevertheless, an additional metric is calculated, called the *F*-measure, which overcomes the accuracy paradox and is often used in other papers, thus needed for comparison. The *F*-measure consists of two metrics, precision and recall. Precision calculates the number of observations that are relevant by dividing the correctly classified true observations through the total number of actual results:

[4] A black-box model is an algorithm in which the inputs and outputs are observable, but the internal workings of how predictions are made cannot be understood by humans.

$$precision = \frac{true\ positive}{true\ positive + false\ positive}.$$

Recall, on the other, hand refers to the number of correctly classified true observations divided by the total number of predicted results:

$$recall = \frac{true\ positive}{true\ positive + false\ negatives}.$$

The *F*-measure, (*F*1-score) calculates a weighted mean of recall and precision, as shown in the following calculation:

$$F - measure = \frac{2 \times recall \times precision}{recall + precision}.$$

Lastly, the Akaike Information Criterion (AIC) provides a single score for each model and estimates the performance of the model compared to other models. It is helpful specifically for model selection and comparison. The lower the AIC score, the better. AIC evaluates the fit of a model on the training data and adds a penalty, similar to regularization, for the complexity of the model. The lowest AIC score is the model that best balances fit with generalizability, which in return maximizes the fit in the out-of-sample data.

4 Data

This chapter describes first the retrieval process of the data used to support the analysis as well as the associated metadata. Second, the pre-processing steps undertaken to transform the original text to its outcome variable are described. Lastly, the trustworthiness of the reviews is discussed.

4.1 Description of Data

The data used in this thesis was scraped from the website www.drugs.com using the online tool ParseHub on the 21st of March 2020. According to the website, it is the largest, independent, and most frequently visited medical information website. The website provides online reviews for various drugs and medical devices and includes the manufacturers' information for dosages, side effects, medical articles and more. The main target audience is consumers and health professionals in the U.S.

The raw dataset includes 35,190 patient-authored reviews for 206 different contraceptive drugs or devices between February 2008 and March 2020. It contains the following variables: *brand_name*, *users_name, review_text, URL*, *date*, *product_rating* and *helpfulness*. The 10 most frequently reviewed brand names received 50% of all reviews, while around 70% of all other brand names got less than 100 reviews. The website includes a steady stream of reviews collected from 2008 until March 2020. It is therefore assumed that the website is in active use. The thesis focuses on the reviews for oral contraceptives (pills). The five most frequently reviewed brands for oral contraceptives were selected for subsequent analysis. This resulted in a total of 3,447 reviews. Subsequently, the variables *review_text* and *product_rating* were selected for further analysis.

4.2 Data Pre-Processing

The variable *review_text* contains the raw text written by reviewers. This text is referred to as the 'corpus' throughout the thesis.

First, in order to achieve a standardised corpus, it was converted to lowercase and punctuations, emoticons (pictorial representations of facial expressions, e.g. :D), numbers and excess of whitespace were removed. Additionally, contractions, such as "isn't" into "is not", were expanded to capture negations which are used at a later stage of the analysis.

Stop words, such as "the" or "a" were also removed, using the pre-defined list available in the programming language *R*. Additional domain-specific stop words were manually included, such as brand names, or words or letters that had no meaning on their own. After removing the stop words, it was calculated that there were 68% fewer words in the corpus. In order to further standardise the corpus, stemming was applied to reduce words to their root form. For example, the word *day*, *daily* or *days* were reduced to their stem *dai*. After stemming, it was calculated that the corpus had 35% less unique words.

Next, in order to reduce spelling errors, the most frequent and infrequent words were identified and removed from the text. Words occurring less than 1% were assumed as being infrequent and meaningless. Frequent words are removed because they tend to be general words used in almost every review and thus do not provide any insightful information to understand and differentiate between positive and negative reviews. In this thesis, the four most frequent words, *contraception, month, pill* and *dai* were excluded. Lastly, the variable *word_count* was added to count the number of words per review. If a review contained less than three words, the review was removed, which guaranteed that only reviews with enough words are considered in the analysis.

4.3 Transformation of Outcome Variable

The *product_rating* variable was used as the Logistic Regression's outcome variable. The rating has an ordinal scale from 1 to 10, in which 1 is the lowest and 10 the highest rating value for a review. The number of very bad (rating 1,2,3) and very good (rating 8,9,10) reviews was almost equally distributed, with 1,387 number of reviews receiving very bad ratings, and 1,359 number of reviews achieving a good rating.

The variable was transformed into a binary variable for positive or negative ratings by giving all the reviews with a *product_rating* of 1, 2 or 3 the label "not happy" and with 8, 9 or 10 the label "happy". Observations with a *product_rating* of 4, 5, 6 or 7 were excluded from the analysis, as the main goal was to classify features describing positive and negative sentiment and not those that may be reviewed as expressing a neutral or no sentiment. Excluding the reviews with rating scores between 4 and 7 reduced the dataset to contain 2,746 reviews.

4.4 Trustworthiness of Reviews

The trustworthiness of patient-authored reviews is important to establish internal and external validity of the dataset. The process of identifying fake reviews and assessing the reviews' trustworthiness is a large and active research field (Barbado, Araque & Iglesias, 2019). There are different types of fake reviews, and some are easier to identify than others. Frequently found and easy to identify, untrustworthy reviews are often commercials or duplicated reviews (Jindal, Liu, 2007). On the other hand, unauthentic, carefully crafted reviews are much harder to identify. To detect this type of fake reviews, more advanced methods are currently being explored such as psycholinguistic, style, deep syntax features or deeper details such as understandability, level of detail, writing style and cognition indicators (Barbado, Araque & Iglesias, 2019). However, these lie outside the scope of this study.

This thesis assumes that the reviews used are credible for three reasons. First, the website assesses each review before publishing it on the website. Therefore, unethical content, bad quality content and duplicates are not shared. Indeed, no duplicates were found in the dataset. When trying to post a commercial or a review which is not related to the topic, the review was not released. Second, the website has a strong interest in maintaining high quality and authentic set of reviews, which makes it likely that additional mechanisms are in place to prevent fake ones from being published. Lastly, the pharmaceutical industry is very regulated and must adhere to the highest ethical standards. Faking patient reviews may result in ann immense reputational, financial and legal loss for the individual and the company.

5 Results

In this chapter, the results of the analysis are presented. In the first section, the process in which seven different feature-sets were created is presented. The second and third sections focus on achieving the highest accuracy by finding the best performing feature-set and machine learning classifier. In the fourth section, the focus lies on the interpretation of the results.

5.1 Feature Engineering

5.1.1 LDA

When applying the LDA model, the first step is to calculate the DTM. As discussed in Chapter 3, five different DTM are created and are individually applied to the LDA model. The first DTM includes unigrams, and the second includes bigrams and the third includes trigrams. In Figure 4., an example of one topic per LDA outcome is shown. Whilst using unigrams, the LDA does not include negations. The word with the highest probability of occurring in topic one is "acne". However, it is not clear if the acne is bad or if the acne is improving, as other words such as "bad" or "clear" are following the first word. The second LDA topic example includes bigrams. The negations are included in these corpora and appear twice in this topic. However, the topic consists of bigrams with many different meanings, such as *anxiety, depress* or *doctor no* which cannot be described as one topic area. The third example of a topic which uses trigrams shows that the negations are almost always present. While two trigrams have the same meaning about not gaining weight, the others have a different meaning, and thus, no clear topic has emerged. Nevertheless, the meaning of the trigrams indicates a positive sentiment, as *no gaining weight*, *no mood swing* and *period no cramp* are all positive experiences when using the contraceptive pill.

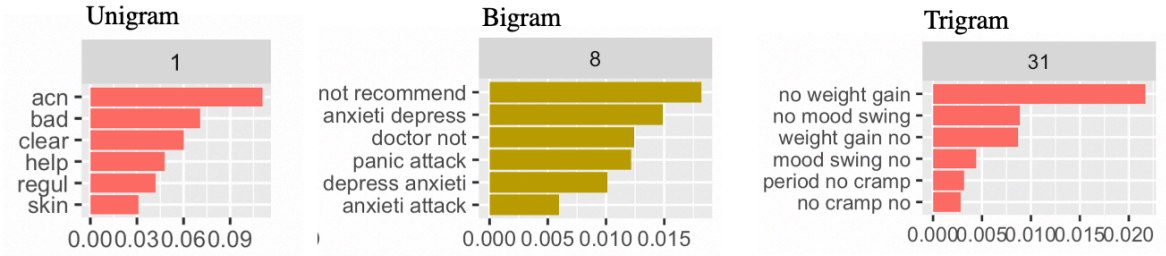


Figure 4. Example topic from LDA model for uni-, bi- and trigrams.

Next, the DTM with two different types of skip-grams was created. Both skip-grams use two as the number of words that can be skipped. However, the first skip-gram DTM uses unigrams and bigrams while the second uses unigrams, bigrams and also trigrams. In Figure 5., the first 2-skip-1-2-gram shows that most words with a high probability of occurring in the topic are unigrams. The first topic is mainly about *spot*, which is a common term describing the experience of irregular or no period after starting oral contraception. In this topic, the terms *lighter* and *no_period* support this meaning of the term.

In the last skip-gram, where also trigrams are used, the topic clearly describes not gaining weight when taking the oral contraceptive. Although the additional trigram does not add a lot of value to the interpretation, the importance of the negations is evident. Without their presence, this topic may be interpreted differently. The difference in gaining weight and not gaining weight will later be a discriminating factor when calculating the sentiment of a topic.

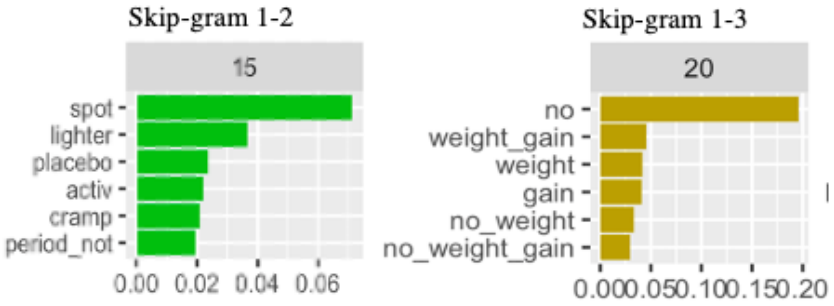


Figure 5. Example topic from LDA model for skip-grams.

In order to find the number of k topics and α with the lowest perplexity score, all five models were tuned. For example, in Figure 6., k is iterated between 5 to 50 topics, with a 5-step difference between each k. The rationale behind using a training and validation set is evident. While the perplexity of the training set decreases, the perplexity of the validation set increases when there are less than 20 topics. This indicates that for less than 20 topics, the model starts to overfit; thus, $k=20$ was selected in this example.

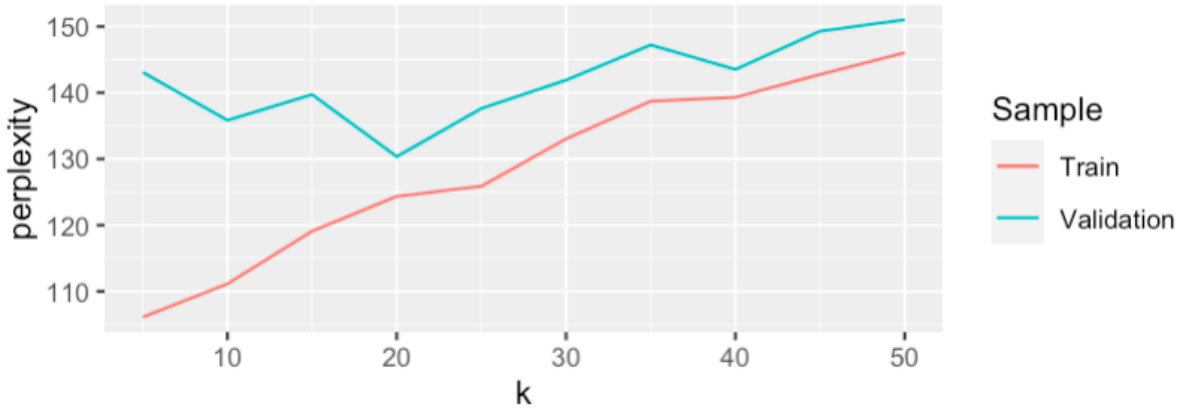


Figure 6. Example of perplexity scores as a result of integrating over k.

In a similar manner as k, α, the document topic density, was tuned, as is shown in Figure 7. The iteration for α was selected for a range of 0.1 to 3, with the lowest α being at 0.5 for the validation set.

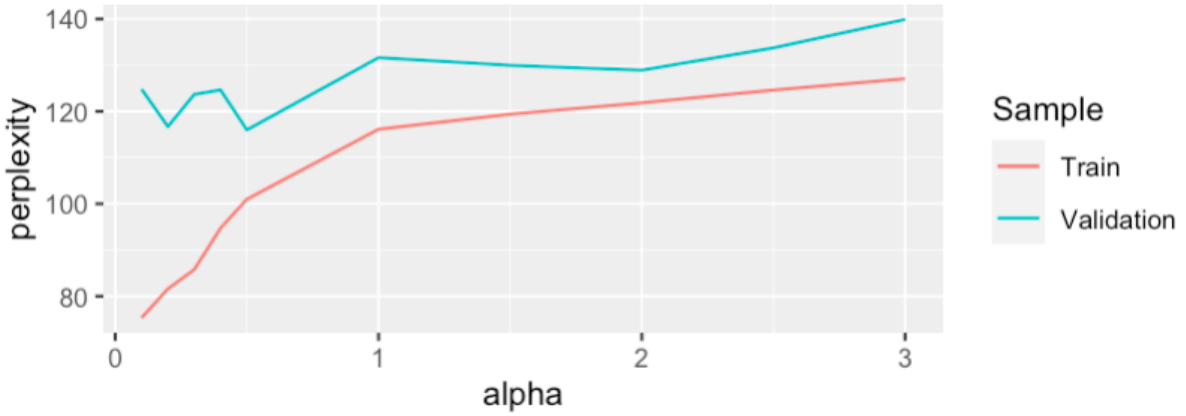


Figure 7. Example of perplexity scores as a result of iterating over α.

To summarize, five different LDA-based feature sets are created and tuned in order to be used for the classification task. As shown in Table 1, each review receives a topic distribution for the number of k topics which are used as features in the classification task.

Table 1. Example of a topic distribution for one review if k=25 topics.

ID	**Topic_1**	**Topic_2**	**Topic_3**	**[..]**	**Topic_22**	**Topic_23**	**Topic_24**	**Topic_25**
644	0.023	0.001	0.012	..	0.052	0.001	0.041	0.072

5.1.2 Dictionary

The NRC dictionary (2010, Mohammad & Turney) was used in order to create dictionary-based features for the classification task as a baseline comparison to the LDA features. The NRC lexicon

provides a score of 1 if a sentiment word is positive or negative, but also a score of 1 if it belongs to one or more of eight emotions; joy, anticipation, surprise, trust, fear, disgust, sadness and anger. For example, the word *hate* is referenced in the dictionary as a negative sentiment word but also as an expression of anger and disgust. Similarly, the sentiment words *loved* is positive and expresses joy and trust. Thus, the pre-processed and cleaned review with ID number 143 and the text "*I hate after first month I loved acne*", would receive the scores as shown in Table 2. The scores are given to each review, and the two sentiments and eight emotions are used as input features for the Logistic Regression to perform the classification task.

Table 2. Example of the scores for emotions and sentiments for the NRC sentiment dictionary.

ID	**joy**	**anticipation**	**surprise**	**trust**	**fear**	**disgust**	**sadness**	**anger**	**negative**	**positive**
143	1	0	0	1	0	1	0	1	1	1

5.1.3 Word embeddings

Similar to the dictionary-based features, word embedding features are created and used for the classification tasks as a baseline comparison to the LDA features. They were calculated using the *Glove* function available in R. Before running the model, the number of latent dimensions was specified to 50, and the window size was set to 5. As a result, the word vector obtained consisted of 50 features. In order to use the word embeddings for the reviews, the word vectors need to be summarized using a summary statistic. In this thesis, the mean was used. Subsequently, each review consisted of a 50-dimensional vector, as shown in Table 3. These 50 dimensions per review were used as input features in the classification task.

Table 3. Example of the 50-dimensional vector for one review.

ID	**Dim_1**	**Dim_2**	**Dim_3**	**[..]**	**Dim_47**	**Dim_48**	**Dim_49**	**Dim_50**
1545	-1.433	0.331	2.437	..	-0.538	3.852	0.345	1.455

5.2 Feature Comparison

After creating the feature-sets, the Logistic Regression is used as a baseline classifier to perform the classification task. Table 4. summarizes the findings, showing the name of the model on the left, in which *glm* stands for the Generalized Linear Model, which covers the Logistic Regression. The number of features indicates the number of inputs for the Logistic Regression. AIC, Accuracy, *F*-measure, and Kappa are all performance metrics to be able to compare the feature sets along different dimensions.

The dictionary features have demonstrated to perform relatively poorly compared to word embeddings and LDA features. This was expected, as the dictionary used is not domain-specific. However, the results are comparable to other sentiment predictions using dictionary features, where the accuracy achieved was around 70% (Hu & Liu, 2004; Moghaddam & Ester, 2011; Zhu et al., 2009).

When using word embeddings, the model performs well with an accuracy of 79.38%. As word-embeddings are currently regarded as state-of-the-art features when the aim is to achieve high accuracy, this was expected. For the LDA features using unigrams, achieving 78.19% is thus already a very good result in terms of accuracy. However, as discussed in Chapter 5.1, the LDA consisting of only unigrams is less interpretable, and thus less relevant for aspect extraction, than when including negations, bi- and trigrams. Therefore, the analysis included several extensions of the LDA. When using only bigrams and only trigrams, the LDA performs poorly, with bigrams achieving an accuracy of 72.34%, similar to the dictionary-features, and trigrams with an accuracy of 59.27%. However, the LDA models using skip-grams outperform the LDA using unigrams in terms of AIC and *F*-measure. Overall, the k-skip-1-3 grams performs best, achieving an accuracy of 81.42% and an *F*-measure of 81.38%, outperforming even the word embeddings. The flexibility of using one, two or three words and skipping over certain unimportant words as well as including negations in the corpus improves the accuracy of the model, which indicates that the topics are more distinct and differentiating.

Table 4. Results from Logistic Regression using different feature sets.

Model	**Nr. of Features**	**AIC**	**Accuracy** (Std)	***F*-measure**	**Kappa**
glm.dictionary	10	2726.6	0.72 (0.027)	0.72	0.44
glm.word_embeddings	50	2299.5	0.79 (0.023)	0.79	0.58
glm.LDA.uni	34	2311.4	0.78 (0.024)	0.77	0.55
glm.LDA.bi	19	2600.8	0.72 (0.023)	0.71	0.44
glm.LDA.tri	34	3243.6	0.59 (0.027)	0.53	0.18
glm.LDA.skip_1_2	24	2229.0	0.78 (0.026)	0.78	0.56
glm.LDA.skip_1_3	24	**2058.5**	**0.81** (0.024)	**0.81**	**0.62**
glm.LDA.all	135	1970.7	0.84 (0.025)	0.84	0.68

Note. glm stands for Generalized Linear Model. In bold is the best performing model using only one set of LDA features.

In order to increase accuracy further, the model is run with all LDA features (total 135), instead of only using one set at a time. This yields a higher accuracy (84.28%), which is expected, as more

features generally lead to better performance. However, having 135 aspects to interpret is not feasible. Thus, a feature selection method, such as Lasso, can be applied. In this thesis, the use of Lasso yielded an accuracy of 83.21% (out-of-sample) and selected 55 features (lambda 1se, Appendix 1.). However, the interpretation of the aspects becomes more difficult and time-intensive. One needs to define and assess topics from different LDA models, and there are still a lot of aspects compared to using only the features from one model. Further, the aspects extracted are repetitive, and some aspects are not selected, which are considered to be interesting. To summarize, the improved accuracy does not outweigh the drawbacks of a more complex approach, which yields less insightful aspects and thus, the LDA features using 2 skip 1-3 gram is selected.

5.3 Comparison of Classifiers

In this section, the performance of the Logistic Regression baseline is compared against more complex classifiers, namely the SVM, Random Forrest, Decision Tree and Naïve Bayes. The classifiers are compared by using the best performing features from the previous section, the LDA skip 1-3 features. The models were trained using 10-fold-cross-validation and repeated five times, as suggested in literature (Kuhn & Johnson, 2016). When applicable to the method, the model was scaled and tuned using grid-search. The Decision Tree was tuned to a length of 50.

The results in Figure 8. show that the Logistic Regression performs similarly in comparison to more complex classifiers. The accuracy for the Logistic Regression (80.85%) is comparable to the accuracy achieved by the Random Forrest (80.31%) and the SVM (80.75%). The Naïve Bayes (78.46%) and the Decision Tree (74.77%) achieved a lower classification performance. Although the variance of the SVM is lower than in the Logistic Regression, the Logistic Regression is preferred due to reasons presented in the theoretical comparison in Chapter 3.

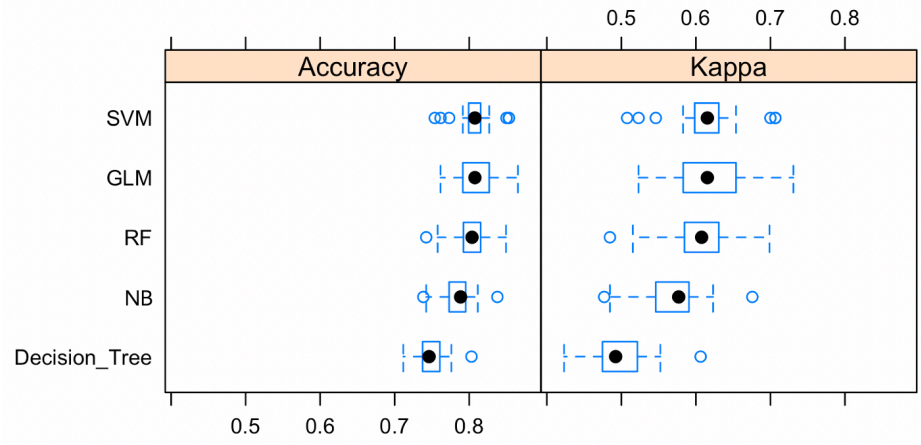


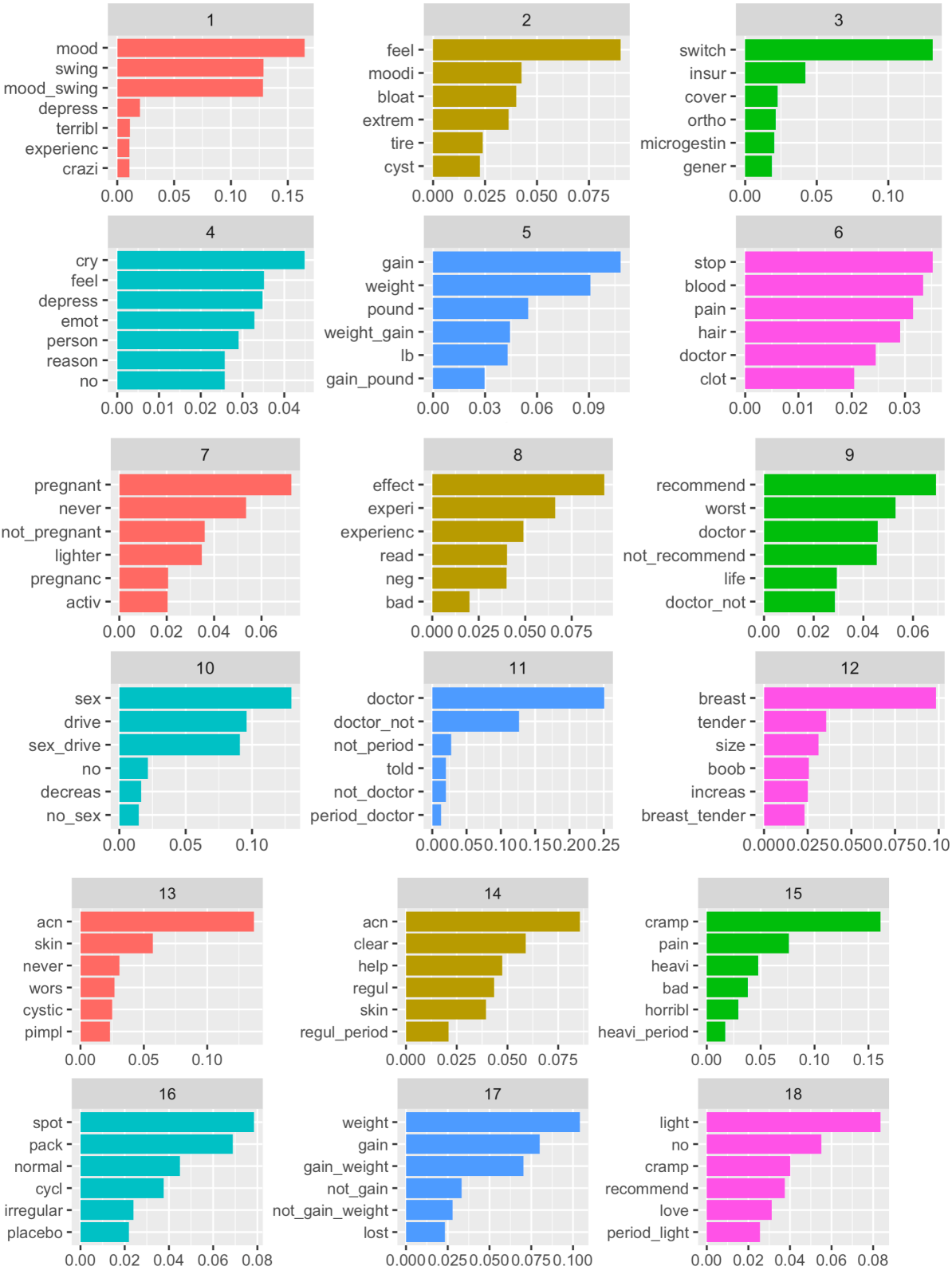
Figure 8. Accuracy and Kappa for SVM, Logistic Regression (GLM), Random Forrest (RF), Naïve Bayes (NB) and Decision Tree.

The results achieved in this thesis are comparable to studies which use health-related datasets in the field of text mining and sentiment analysis. Denecke & Nejdl (2009) achieved an F-measure of 78.59% with a large set of different features. Na et al. (2012) achieved 78% accuracy using SVM as a classifier and a domain-specific dictionary. Bobicev, Sokolova, Jafer and Schramm (2012) achieved an F-measure of 74.51% when using Naïve Bayes and correlated words as features. Carrillo-de-Albornoz et al. (2018) achieved 67.21% accuracy using word embeddings. The best performing results found in literature are from Lu et al. (2011), with an accuracy of 80.32%, from Moghadam and Ester (2011), with accuracy between 79%-86% and lastly, from Salas-Zárate et al. (2017), achieving an *F*-measure of 81.24%. Overall, the results obtained herein demonstrate the efficacy of the proposed approach.

5.4 Interpretation of Results

The LDA outcome using the 2-skip-1-3 gram features performed best and was thus used for the interpretation in this chapter. Figure 9. shows the 25 topics with the six most likely words per topic produced by the tuned LDA. While most words are unigrams, there are several bigrams and one trigram that is among the top six words of the topic.

A well-performing LDA should produce topics that are interpretable, unique and specific. When regarding interpretability, each topic can be clearly summarized into one meaningful topic, except for one, topic 9, which contains irrelevant topic words such as *recommend, worst, doctor, not_recommend*. This topic is not useful for interpretation in subsequent analysis. In terms of uniqueness, several topics are similar, such as topic 1 and 2, with both discussing being moody and having mood swings. Topic 6 and 21 are also similar, with both discussing the stopping of bleeding during their period. However, their difference is that topic 6 discusses *pain*, while topic 21 includes words such as *heavy* and *breakthrough*. Topic 4 and 22 are both partly about depression. However, while topic 4 discusses crying, topic 22 includes the words *anxiety* and *attack*. Similarly, topic 13 and 14 are both about acne, but in 13, the topic sentiment is more negative, including words such as *worst* or *never*, while for topic 14, the sentiment is more positive, including words such as *help* and *clear*.


1
mood
swing
mood_swing
depress
terribl
experienc
crazi
0.00 0.05 0.10 0.15
2
feel
moodi
bloat
extrem
tire
cyst
0.000 0.025 0.050 0.075
3
switch
insur
cover
ortho
microgestin
gener
0.00 0.05 0.10
4
cry
feel
depress
emot
person
reason
no
0.00 0.01 0.02 0.03 0.04
5
gain
weight
pound
weight_gain
lb
gain_pound
0.00 0.03 0.06 0.09
6
stop
blood
pain
hair
doctor
clot
0.00 0.01 0.02 0.03
7
pregnant
never
not_pregnant
lighter
pregnanc
activ
0.00 0.02 0.04 0.06
8
effect
experi
experienc
read
neg
bad
0.000 0.025 0.050 0.075
9
recommend
worst
doctor
not_recommend
life
doctor_not
0.00 0.02 0.04 0.06
10
sex
drive
sex_drive
no
decreas
no_sex
0.00 0.05 0.10
11
doctor
doctor_not
not_period
told
not_doctor
period_doctor
0.000.050.100.150.200.25
12
breast
tender
size
boob
increas
breast_tender
0.0000.0250.0500.0750.10
13
acn
skin
never
wors
cystic
pimpl
0.00 0.05 0.10
14
acn
clear
help
regul
skin
regul_period
0.000 0.025 0.050 0.075
15
cramp
pain
heavi
bad
horribl
heavi_period
0.00 0.05 0.10 0.15
16
spot
pack
normal
cycl
irregular
placebo
0.00 0.02 0.04 0.06 0.08
17
weight
gain
gain_weight
not_gain
not_gain_weight
lost
0.0000.0250.0500.0750.100
18
light
no
cramp
recommend
love
period_light
0.00 0.02 0.04 0.06 0.08

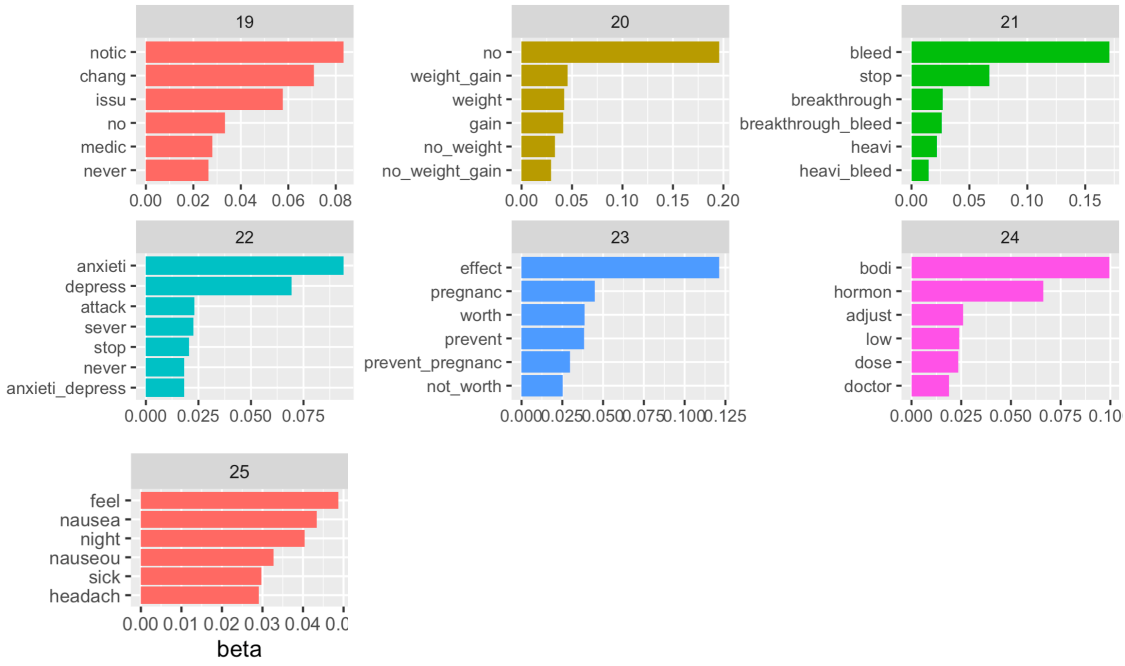


Figure 9. 25 LDA topics for the 2-skip-1-3 gram.

The 25 topics presented above were named using the extracted topic words, logical interpretation and domain understanding. In a real-life use case, the output of the LDA and the topic names would need to be confirmed and interpreted by a domain expert, such as a doctor, and checked against other research (e.g. clinical studies, pharmacovigilance[5]). For the purpose of this study, the topic names were manually selected based on the first one or two words. In Table 5., the topic names are listed together with the average probability of a topic occurring. The average probability of a topic occurring is calculated by taking the mean probability for each topic over all reviews. The most frequent topics are *No_weigh_gain, Worst_acne*, *Feeling_nausea*, and *Anxiety_depression*. Less frequent topics are about *Effectiveness*, *Feeling_moody* and *Notice_change*.

Table 5. Topic list for LDA results using the 2-skip 1-3 gram and the average topic probability.

Topic Name	**Average topic probability**
1_Mood_swings	0.042
2_Feeling_moody	0.036
3_Switching_insurance	0.039
4_Crying_depressed	0.042
5_Gaining_weight	0.041

[5] Pharmacovigilance monitors the unexpected conditions and effects of drugs, in order to record, measure and evaluate reactions to drugs not yet documented, also referred to as Adverse Drug Reactions (ADR).

6_Stop_bleeding	0.040
7_Not_pregnant	0.038
8_Experience	0.040
9_Not_recommend	0.039
10_No_sex_drive	0.037
11_Doctor	0.040
12_Tender_breasts	0.040
13_Worst_acne	0.046
14_Clear_acne	0.037
15_Heavy_pain_cramps	0.043
16_Spot_pack	0.039
17_Medium_weight_gain	0.038
18_Light_no_cramps	0.041
19_Notice_change	0.037
20_No_weight_gain	0.046
21_Stop_bleeding	0.040
22_Anxiety_depression	0.044
23_Effectivness	0.034
24_Hormons	0.037
25_Feeling_nausea	0.044

Note. Topic names are manually created based on the most frequent words of a topic. Average topic probability is calculated by taking the mean probability for each topic.

The Logistic Regression describes how a dependent, categorical feature and one or several independent features relate to each other. In this thesis, the dependent feature has two levels, positive or negative, and 24 continuous independent features. By definition, the 25 topics from the LDA and their probability distributions, scores between 0 and 1, add up to one, which means they are perfectly collinear, resulting in the problem of perfect multicollinearity (Allen, 1997). To overcome this problem, one variable was excluded from the original set of 25 independent variables, as suggested in literature. As topic 6 and topic 21 are almost identical, topic 21 is removed. Thus, the Logistic Regression is performed with 24 features instead of 25.

Table 7. presents the output of the Logistic Regression sorted by coefficient size. The outcome feature is labelled 1 for positive (happy) reviews and 0 for negative (unhappy) reviews. A negative coefficient tends to decrease the likelihood of a review being happy while a positive coefficient tends to increase the likelihood of a review being happy. For example, Table 6. shows the probability distribution for reviews with ID 644, which scores high on Topic_22 about anxiety

and depression, and on Topic_25, which is about feeling nausea, both topics that have a strongly negative coefficient in Table 7., thus this review is likely to be negative.

Table 6. Example of a topic distribution for one review if *k*=25 topics.

ID	**Topic_1**	**Topic_2**	**Topic_3**	**[..]**	**Topic_22**	**Topic_23**	**Topic_24**	**Topic_25**
644	0.023	0.001	0.012	..	0.052	0.001	0.002	0.072

Overall, there are more positive than negative coefficients. Out of the 25 coefficients, five are not significant, as indicated by the z-value in Table 7., in which having no stars means the coefficient is not significant. The most positive coefficients were: *No_weight_gain, Light_no_cramps, Experience* and *Clear_acne*. The most negative reviews were: *Not_recommend, Anxiety_depression, Crying_depressed, Stop_bleeding* and *Worst_acne*.

One interesting insight is that the topics that are similar are represented in the Logistic Regression in the expected order of the coefficient size. For example, *No_weight_gain* is very positive, while the topic *Gaining_weight* is negative, and the *Medium_weigth_gain* is less sentiment-baring, as expected from the topic interpretation.

Table 7 Coefficients for 24 LDA topics using skip-gram 1-3.

Coefficients	**Estimate**	**Std.**	**Error**	**z-value**	**Pr(>|z|)**
(Intercept)	-3.67	0.92	-4	6.25E-05	***
20_No_weight_gain	16.13	1.66	9.71	<2.00E-16	***
18_Light_no_cramps	14.34	1.57	9.11	<2.00E-16	***
8_Experience	13.64	1.66	8.24	<2.00E-16	***
14_Clear_acne	11.16	1.49	7.51	6.14E-14	***
17_Medium_weight_gain	10.25	1.39	7.36	1.85E-13	***
12_Tender_breasts	8.6	1.36	6.34	2.33E-10	***
19_Notice_change	8.15	1.5	5.44	5.31E-08	***
7_Not_pregnant	7.45	1.38	5.42	6.07E-08	***
24_Hormons	6.37	1.42	4.48	7.65E-06	***
11_Doctor	5.53	1.43	3.88	0.000104	***
23_Effectivness	4.55	1.43	3.17	0.001511	**
16_Spot_pack	3.63	1.42	2.56	0.010587	*
15_Heavy_pain_cramps	3.6	1.34	2.69	0.007096	**
10_Sex_drive	3.23	1.24	2.61	0.009078	**
1_Mood_swings	3.05	1.28	2.38	0.017406	*

5_Gaining_weight	1.68	1.21	1.4	0.163097	
3_Switching_insurance	1.29	1.36	0.95	0.342747	
2_Feeling_moody	-0.67	1.44	-0.47	0.641125	
25_Feeling_nausea	-0.7	1.27	-0.56	0.579047	
13_Worst_acne	-1.59	1.17	-1.36	0.173684	
6_Stop_bleeding	-3.78	1.45	-2.61	0.009123	**
4_Crying_depressed	-6.03	1.5	-4.01	6.12E-05	***
22_Anxiety_depression	-8.43	1.63	-5.19	2.13E-07	***
9_Not_recommend	-9.22	1.76	-5.25	1.53E-07	***

Note. The dependent variable has two classes, 0 (unhappy) and 1 (happy). The independent variable has values between 0 and 1. Topic 21 is excluded to overcome the problem of perfect multicollinearity. *** p*<0.001, ** p<0.01, * p<0.05.*

In the last step of the analysis, the average probability of a topic occurring, as shown in Table 5. and the coefficient size, as shown in Table 7., were plotted against each other to illustrate the findings for the aspects in a Perception Map (Figure 10.). Although the position of the aspects does not represent the exact linear relation of the aspects to each other, it shows how they relate to each other in terms of sentiment (y-axis) and relevance (x-axis). The aspects located in the top right corner are perceived as being more positive and frequently discussed in the reviews, such as *No_weight_gain* and *Light_no_cramps*. These can be interpreted as the unique-selling-position (USP) for oral contraceptives. Aspects in the bottom right corner are perceived as being more negative and frequently discussed. For example, *anxiety_depressed*, *crying_depressed*, or *worst_acne,* which is not as negative as *anxiety_depressed* but extremely often discussed in the reviews. These may be the most pressing concerns from patients, which companies must be aware of and address. The Perception Map also shows that some aspects, such as *effectiveness* and *not_pregnant*, are much less important to women than one would expect based on literature (Johnson et al., 2013). Currently, the focus of promotional activities and the main messages focus on the effectiveness (Kornfield et al., 2013; Tyrawski & DeAndrea, 2015; Wu et al., 2016), however, this thesis suggests that other topics might be more relevant for women and need to be addressed in the marketing and information booklets provided by the producer. In Figure 11., circles are manually added to Figure 10. to support the quick interpretation of the Perception Map and convey the key message to end-users, for example, marketeers. The size of the circles depends on the context and can be adapted depending on the individual graph.

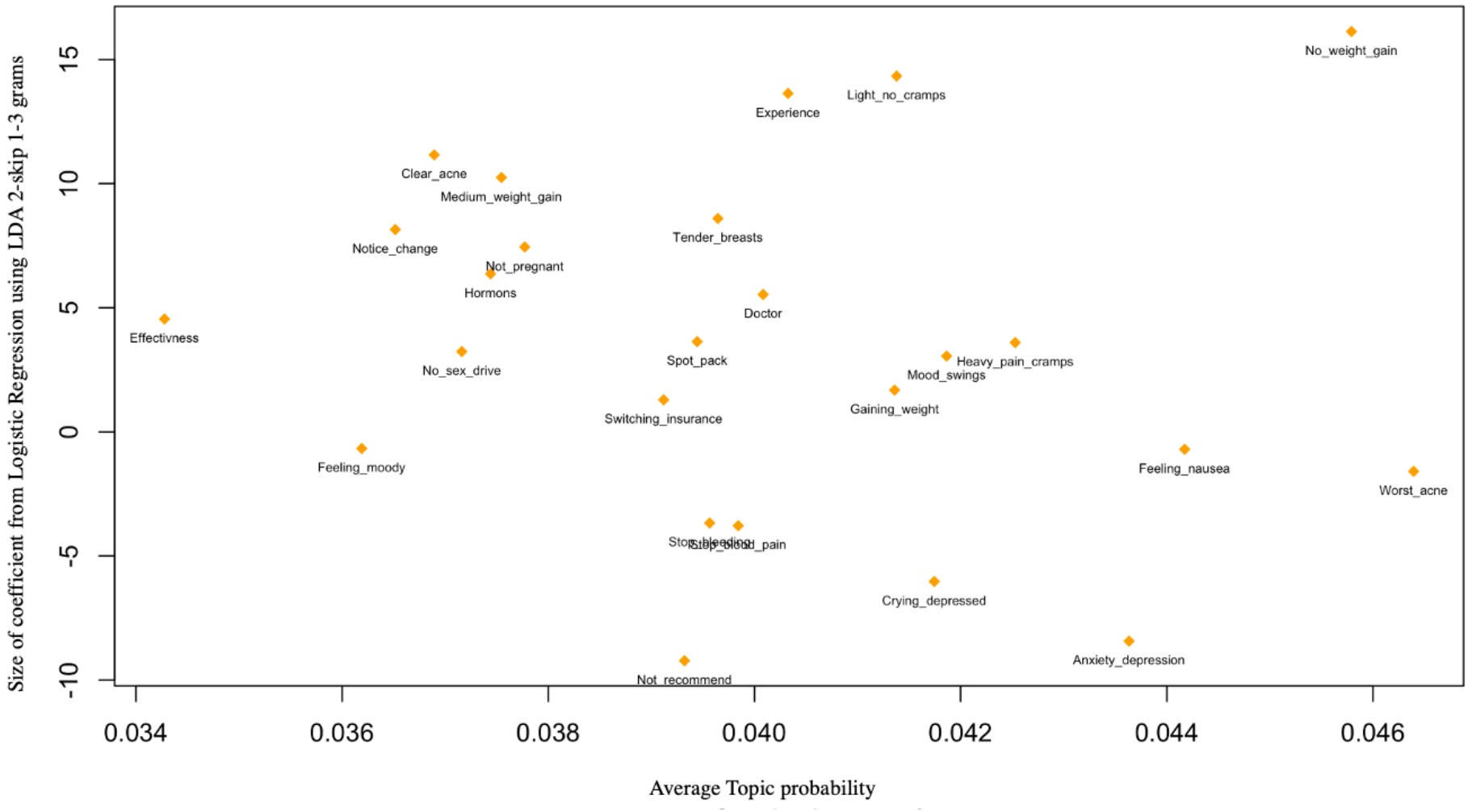


Figure 10. Perception Map for Oral Contraceptives

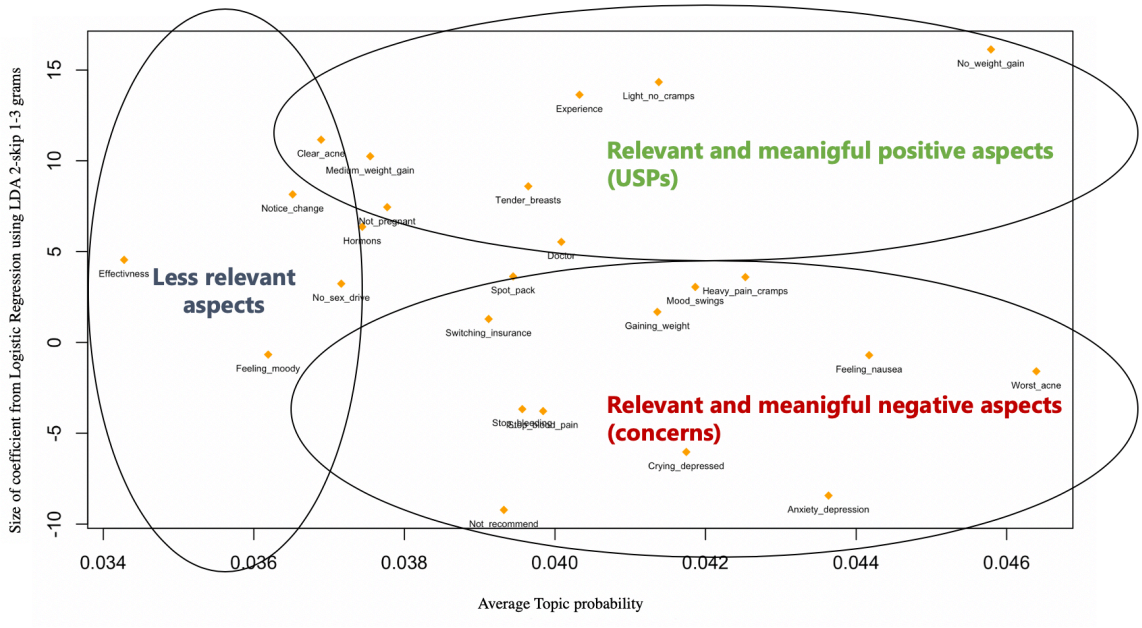


Figure 11. Perception Map for oral contraceptives for a high-level marketing interpretation[6].

The approach introduced in this chapter demonstrates how the use of skip-grams and negations in LDA improves aspect extraction from reviews, increasing the accuracy of the LDA model by three percentage points, from 78% to 81%, and improving the interpretability of the aspects. Further, the approach encourages the use of the Logistic Regression as a classifier due to its comparable performance to more complex classifiers, while offering more interpretability. Lastly, the approach offers a novel way to summarize and present the findings from aspect-level sentiment analysis to decision-makers in marketing, the so-called Perception Map. The map plots the relevance of the aspects (average topic probability from LDA) and the sentiment of the aspects (coefficient size of the Logistic Regression). The Perception Map allows the quick identification of USPs and concerns from customers. In addition, it helps companies identify new aspects or find out whether the sentiment/relevance is other than expected. Overall, the approach offers an improved extraction and classification process to find meaningful and relevant aspects from online reviews and brings them across in a simplified way to support decision-making.

[6] Circles are manually added to support interpretation. Size of circles depends on the context

6 Discussion and Conclusion

The patient-authored information available on the Internet on contraceptive methods is extremely valuable, both for women, to make a more informed decision, and for companies to understand what their customers talk about and how they feel towards a contraceptive method or brand. Therefore, there is a need to find new solutions on how to extract insights, not just based on five or ten reviews, but on the aggregated information available, which represents the summary of experiences rather than individual opinions. This results in the challenge to bridge between three areas; the latest developments in the text mining field, marketing, and the women's health domain, with the overall purpose to leverage the sheer endless amount of information on the Internet to improve the overall access to information on contraceptives.

This thesis contributes to solving the above challenge by applying a novel approach to answer the following research question, "*What are the relevant positive and negative aspects expressed in online reviews for oral contraceptives?*". Specifically, the findings show that the relevant and negative aspects are about having bad acne, feeling nausea and having anxieties or depressions. Relevant and positive aspects are about not gaining weight, having lighter cramps and the overall experience of taking oral contraceptives. The results also shed light on aspects which are rarely discussed, such as the loss of sex drive or having tender breasts, while indicating that some topics, such as effectiveness, are of lesser interest to women than expected.

The approach presented in this thesis contributes to the field of aspect-level sentiment analysis in the health-domain. First, an extended topic modelling method is presented to extract more meaningful aspects. The extension, using skip-grams and negations in the corpus, increases the interpretability, distinctiveness and meaning of an aspect. In addition, it improves the predictive accuracy to distinguish between positive and negative aspects by three percentage points, from 78% to 81%. For the classification task, five different classifiers were tested. The Logistic Regression is recommended due to its comparable performance and better interpretability compared to other, more complex methods. It achieved an accuracy of 81%, which is among the highest results found in literature for aspect-level sentiment analysis. Lastly, the results are mapped on a Perception Map, which shows how relevant an aspect is, i.e. how often do reviewers talk about that aspect, and the sentiment of an aspect, i.e. how positive or negative do the reviewers perceive that aspect.

From a marketing point of view, the approach provides a simple yet insightful way to extract and illustrate the findings. Although this thesis focuses on oral contraceptives, it may also be transferable to gain insights from other contraceptive methods. By comparing different Perception Maps, it may be indicated as to how customers perceive the USPs and concerns of different contraceptive methods. Additional applications may be to compare two Perceptual Maps, reflecting the results from different product groups, brands or timeframes, which may reveal how the sentiment and relevance of certain aspects changed over time. The insights gathered can be used in many marketing-related applications, such as to update brochures/website towards topics that are of great concern or benefit to patients, to monitor topics that affect negative reviews most substantially, to focus on aspects not yet addressed appropriately or to improve communication, for example by updating SEO-related keywords based on the findings.

It is important to be aware of the limitations of this thesis. In textual online review data, some opinions and thoughts might not be expressed but are still relevant. In addition, the context in which reviews are written needs to be taken into account, for example, anonymity, which influences how and what customers write. In this thesis, the reviews used were anonymously written, which can lead to the question of how trustworthy the reviews are. It is assumed that they are trustworthy, but this needs to be further validated. Furthermore, there is a selection bias. People expressing opinions in reviews usually represent a polarized point of view; thus, when evaluating them, this bias needs to be taken into account. In addition, this thesis used only extreme opinions for the analysis, while neutral reviews were excluded. Therefore, the results present an extreme point of view rather than a reflection of reality.

In general, future research should focus on bridging between text mining and its application in the field of marketing for the health-domain. Building on the findings in this thesis, future work should include other health-related domains in order to increase the validity of the approach. Another interesting research avenue could be to compare the similarity of text created by the company itself (e.g. marketing booklets, website) to the patient-authored texts and investigate the impact on the sentiment. Further, the results from the presented approach could be compared to other data, for example, clinical studies or findings from pharmacovigilance.

7 Bibliography

Airoldi, E. M., Blei, D., Erosheva, E. A., & Fienberg, S. E. (Eds.). (2014). Handbook of mixed membership models and their applications. In *CRC press*.

Arndt, J. (1967). Role of product-related conversations in the diffusion of a new product. In *Journal of marketing Research, 4(3), 291-295*.

Barbado, R., Araque, O., & Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. In *Information Processing & Management,* 56(4), 1234-1244.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. In *Journal of Marketing, 84(1), 1-25*.

Blei, D., & Lafferty, J. (2006). Correlated topic models. Advances in neural information processing systems, 18, 147.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. In *Journal of machine Learning research*, 3(Jan), 993-1022.

Bobicev, V., Sokolova, M., Jafer, Y., & Schramm, D. (2012). Learning sentiments from tweets with personal health information. In *Canadian conference on artificial intelligence* (pp. 37-48). Springer, Berlin, Heidelberg.

Carrillo-de-Albornoz, J., Vidal, J. R., & Plaza, L. (2018). Feature engineering for sentiment analysis in e-health forums. In *PloS one*, 13(11).

Chorowski, J., Wang, J., & Zurada, J. M. (2014). Review and performance comparison of SVM- and ELM-based classifiers. In *Neurocomputing*, 128, 507-516.

Claringbold, L., Sanci, L., & Temple-Smith, M. (2019). Factors influencing young women's contraceptive choices. *Age (years), 18(20), 21-22*.

Collins, W. (2015). Collins English Dictionary–Complete & Unabridged 2012 Digital Edition. *Diunduh dari http://dictionary. reference. com/browse/togetherness*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. In *Machine learning*, 20(3), 273-297.

Denecke, K., & Nejdl, W. (2009). How valuable is medical social media data? Content analysis of the medical web. In *Information Sciences,* 179(12), 1870-1880.

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. In *Procedia Computer Science, 87, 44-49*.

Duan, W., Gu, B., & Whinston, A. (2008). The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry. In *Journal of Retailing, 233-242*.

Duric, A., & Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. In *Decision support systems*, 53(4), 704-711.

Fernández, J., Gutiérrez, Y., Gómez, J. M., & Martínez-Barco, P. (2014, August). Gplsi: Supervised sentiment analysis in twitter using skipgrams. In *Proceedings of the 8th International Workshop on Semantic Evaluation* (SemEval 2014) (pp. 294-299).

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. In M*achine learning*, 29(2-3), 131-163.

Fürnkranz, J. (1998). A study using n-gram features for text categorization. Austrian Research Institute for Artifical Intelligence, 3(1998), 1-10.

Goeuriot, L., Na, J. C., Min Kyaing, W. Y., Khoo, C., Chang, Y. K., Theng, Y. L., & Kim, J. J. (2012). Sentiment lexicons for health-related opinion mining. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 219-226).

Guthrie, D., Allison, B., Liu, W., Guthrie, L., & Wilks, Y. (2006). A closer look at skip-gram modelling. In *LREC* (Vol. 6, pp. 1222-1225).

Hai, Z., Chang, K., & Kim, J. J. (2011, February). Implicit feature identification via co-occurrence association rule mining. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 393-404). Springer, Berlin, Heidelberg.

Hai, Z., Cong, G., Chang, K., Liu, W., & Cheng, P. (2014, July). Coarse-to-fine review selection via supervised joint aspect and sentiment model. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 617-626).

HaCohen-Kerner, Y., Ido, Z., & Ya'akobov, R. (2017). Stance classification of tweets using skip char ngrams. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 266-278). Springer, Cham.

Hoffmann, J. P. (2004). Generalized linear models: An applied approach. Pearson College Division.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177).*

Jindal, N., & Liu, B. (2007). Review spam detection. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1189-1190).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer, Berlin, Heidelberg.

Johnson, S., Pion, C., & Jennings, V. (2013). Current methods and attitudes of women towards contraception in Europe and America. In *Reproductive health, 10(1), 7.*

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. In *SMU Data Science Review, 1(3), 9*.

Kornfield, R., Donohue, J., Berndt, E. R., & Alexander, G. C. (2013). Promotion of prescription drugs to consumers and providers, 2001–2010. *PloS one, 8(3)*.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling (*Vol. 26). New York: Springer.

Kumabam, R. S., Meitei, C. I., Singh, S. S., & Singh, T. P. (2017). EXPANDING THE HORIZON OF MARKETING: CONTEMPLATING THE SYNERGY OF BOTH TRADITIONAL WORD OF MOUTH AND E-WORD OF MOUTH. *International Journal of Management (IJM), 8(3)*.

Kwartler, T. (2017). Text mining in practice with R. John Wiley & Sons.

Liu, B., Hu, M., & Cheng, J. (2005, May). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web* (pp. 342-351).

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010), 627-666.

Liu, B. (2012). Sentiment analysis and opinion mining. In *Synthesis lectures on human language technologies, 5(1), 1-167*.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). In *Springer, Boston, MA*.

Long, C., Zhang, J., & Zhu, X. (2010, August). A review selection approach for accurate feature rating estimation. In *Coling 2010: Posters* (pp. 766-774).

López, M., & Sicilia, M. (2011). The impact of e-WOM: determinants of influence. In *Advances in Advertising Research (Vol. 2)* (pp. 215-230). Gabler.

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011, December). Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th international conference on data mining workshops* (pp. 81-88). IEEE.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. In *Shams engineering journal, 5(4), 1093-1113*.

Melo, J., Peters, M., Teal, S., & Guiahi, M. (2015). Adolescent and young women's contraceptive decision-making processes: choosing "The Best Method for Her". In *Journal of pediatric and adolescent gynecology,* 28(4), 224-228.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv* preprint *arXiv:1301.3781*.

Moghaddam, S., & Ester, M. (2011). The FLDA model for aspect-based opinion mining: addressing the cold start problem. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 909-918).

Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text (pp. 26-34).* Association for Computational Linguistics.

Na, J. C., Kyaing, W. Y. M., Khoo, C. S., Foo, S., Chang, Y. K., & Theng, Y. L. (2012). Sentiment classification of drug reviews using a rule-based linguistic approach. In *International conference on asian digital libraries (pp. 189-198).* Springer, Berlin, Heidelberg.

Neslin, S., Grewal, D., Leghorn, R., Shankar, V., Teerling, M., Verhoef, P., & Thomas, J. (2006). Challenges and Opportunities in Multichannel Customer Management. In *Journal of Service Research.*

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems* (pp. 841-848).

Pang, B., & Lee, L. (2009). Opinion mining and sentiment analysis. In *Comput. Linguist,* 35(2), 311-312.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Prakash, P. K. S., & Rao, A. S. K. (2017). R deep learning cookbook. Packt Publishing Ltd.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009, August). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 248-256).

Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. In *Journal of Business Economics,* 89(3), 327-356.

Roscher, R. (2013). Sequential learning using Incremental Import Vector Machines for semantic segmentation.

Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M. A., & Valencia-Garcia, R. (2017). Sentiment analysis on tweets about diabetes: an aspect-level approach. In *Computational and mathematical methods in medicine*, 2017.

Schoenmüler, Verena, Oded Netzer, and Florian Stahl (2019), "The Extreme Distribution of Online Reviews: Prevalence, Drivers and Implications," In *Columbia Business School Research Paper*

Schouten, K., & Frasincar, F. (2015). Survey on aspect-level sentiment analysis. In *IEEE Transactions on Knowledge and Data Engineering, 28(3), 813-830.*

Sharma, A., & Dey, S. (2012, October). A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM research in applied computation symposium* (pp. 1-7).

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. In *Expert Systems with Applications*, 41(3), 853-860.

Stojanovski, D., Strezoski, G., Madjarov, G., Dimitrovski, I., & Chorbev, I. (2018). Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages. Multimedia Tools and Applications, 77(24), 32213-32242.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. In *Computational linguistics, 37(2), 267-307.*

Tam, Y. C., & Schultz, T. (2005). Dynamic language model adaptation using variational Bayes inference. In *Ninth European Conference on Speech Communication and Technology.*

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1555-1565).

Titov, I., & McDonald, R. (2008, April). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web* (pp. 111-120).

Tyrawski, J., & DeAndrea, D. C. (2015). Pharmaceutical companies and their drugs on social media: a content analysis of drug information on popular social media sites. In *Journal of medical Internet research, 17(6), e130.*

United Nations, Department of Economic and Social Affairs, Population Division (2019). Contraceptive Use by Method 2019: Data Booklet (ST/ESA/SER.A/435).

UN DESA (2019), The Sustainable Development Goals Report 2019, UN, New York*, https://doi.org/10.18356/55eb9109-en*

Vapnik, V. (1998). The support vector method of function estimation. In *Nonlinear Modeling* (pp. 55-85). Springer, Boston, MA.

Varkaris, E. (2017). The influence of social media on the consumers' hotel decision journey. In *Journal of Hospitality and Tourism Technology, 100-118.*

Wang, T., Cai, Y., Leung, H. F., Lau, R. Y., Li, Q., & Min, H. (2014). Product aspect extraction supervised with online domain knowledge. In *Knowledge-Based Systems,* 71, 86-100.

Wu, M. H., Bartz, D., Avorn, J., & Seeger, J. D. (2016). Trends in direct-to-consumer advertising of prescription contraceptives. *Contraception, 93(5), 398-405.*

Yu, L. C., Wang, J., Lai, K. R., & Zhang, X. (2017). Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 534-539).

Zhu, X., & Davidson, I. (2007). Knowledge discovery and data mining: challenges and realities (p. 118). Hershey: Information Science Reference.

Zhu, J., Wang, H., Tsou, B. K., & Zhu, M. (2009). Multi-aspect opinion polling from textual reviews. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1799-1802).

8 Appendix

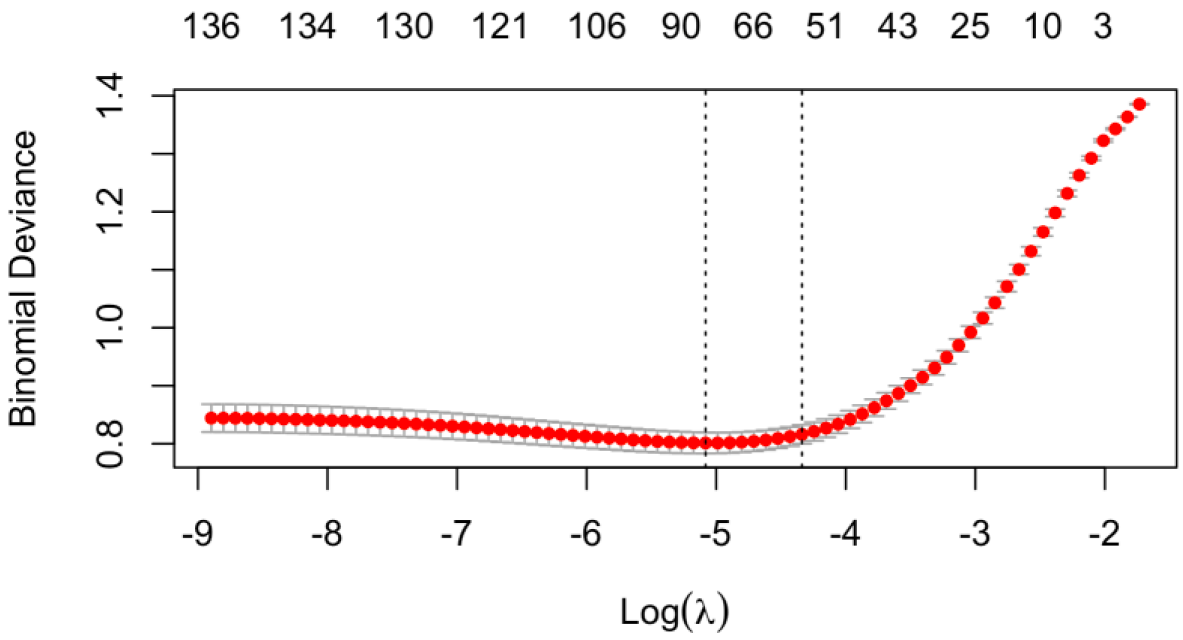


Figure 1. Cross-validated binominal deviance (error) according to log of lambda. Vertical line on the right finds lambda that provides the simplest model and lies within one standard error of the optimal value of lambda.