Using Natural Language Processing to automatically classify search queries based on user intent

Bouwmeesters, G.C.A.

Award date:
2020

Using Natural Language Processing to automatically classify search queries based on user intent

By: G.C.A. (Guido) Bouwmeesters
Student number: 0904415
November 2020

Written in partial fulfilment of the requirements of the degree of Master of Science in Innovation Management.

Department of Industrial Engineering and Innovation Science, Eindhoven University of Technology.

Supervisors:

Dr. N. (Néomie) Raassens	TU/e, ITEM
Dr. S. (Shantanu) Mullick	TU/e, ITEM
A. (Annes) Alajmovic, MSc.	Greenhouse Group
R. (Ruben) Mak, MSc.	Greenhouse Group

Preface

Dear reader,

This thesis marks the end of my master Innovation Management at Eindhoven University of Technology and subsequently the end of my student life. In these last years and months, I had the opportunity to develop many skills, of which this thesis is the end product. While challenging at times, it has been a lot of fun to work on this project.

It would not have been possible for me to complete this thesis without the help of others. That's why I would like to express my gratitude to a few people.

First of all, I would like to thank Néomie Raassens, who acted as my mentor and first supervisor. You have given me many new insights after our meetings, and made sure I was keeping my work academic and to the point. Although challenging, I believe this has made substantial contributions to the quality of this research paper. Secondly, I would like to thank Shantanu Mullick as my second supervisor for providing me with more feedback to work with, which has helped to improve me in writing a better thesis.

Then, I would like to thank my colleagues from the AI department at Greenhouse for their support and assistance whenever I need it. In particular, I would like to thank Annes Alajmovic and Ruben Mak. Ruben, thanks for introducing me to the possibilities Greenhouse offered and your remarks on my research. Annes, thank you for introducing me to the world of AI and online marketing and always being able to answer my questions whenever I had any. I enjoyed the discussions we had which has opened my eyes for the practical use of AI, in the online marketing domain and elsewhere.

Last but not least, I have to thank my friends and family for supporting me during these times. In particular my parents for their support during all those years, and my friends for making my time as a student in Eindhoven unforgettable.

Thank you,

Guido Bouwmeesters

Executive summary

This report reveals the results of a research project to classify search queries on search intent. The research has been conducted at Greenhouse. Greenhouse is a leading online marketing company in the Netherlands, which helps its clients in their online marketing strategy. This research will offer insight to Greenhouse on what search intent is, how the models can be trained to classify on search intent and how this can help with the development of a new business proposition.

Introduction
People are more connected to the Internet than ever before. Search engines like Google are used on a daily basis, and the total amount of search queries handled per day rises. The business model of search engines like Google is built around the behavior of its users. When looking for something, Google wants to know what searchers are looking for and will try to show the best results, both paid and non-paid results. To do this, the search engine needs to understand what it is searchers are looking for, they need to understand why the search query is used. Therefore, the search engine does not only need an understanding of the individual words used, but also of the words combined in a context. This means a search engine should understand the intention of why someone is searching using a specific query.

In this research, it has been tried to create a model which can help to predict the search intent of a search query. Researchers have been trying to improve the accuracy of the search intent classification tasks for years. The recent developments of NLP models have led to human-quality results on benchmark tests. With the advent of transformers, it has become possible to pre-train a model on a large dataset and fine-tune it to perform specific tasks. Because of this, it should be possible to increase accuracy once again.
This resulted in the following research question: *How can Natural Language Processing (NLP) techniques be used to automatically classify search queries based on user intentions?*

Theoretical background
In the search for better information retrieval systems, researchers have increasingly been focusing on what a searcher is actually looking for when using search engines (Shneiderman, Byrd & Croft, 1997; Zhang et al., 2020). For humans to understand better how to potentially categorize search intent, researchers have developed taxonomies (Broder, 2002; Rose and Levinson, 2004; Jansen et al., 2008; Lewandowski et al., 2012), which are broadly used in both theory and practice (e.g. Yoast, 2018; Dean, 2019b; Agius, 2019; Searchmetrics, 2020). Broder (2002) was the first to introduce taxonomies, as he argued that search is not always informational but can be navigational (e.g. ‘I want to visit a certain website’) or transactional (e.g. ’I want to buy something’) as well. Rose and Levinson (2004), Jansen et al. (2008), and Lewandowski et al. (2012) built on this.

This has resulted in the search intent classification framework as presented in Figure 0.1 and is used in this research.
Since the development of transformers, pre-trained language models can be made and such a model, BERT, is used in this research. These models achieve state of the art results in NLP benchmark tasks, and will increase accuracy on intent classification tasks again.

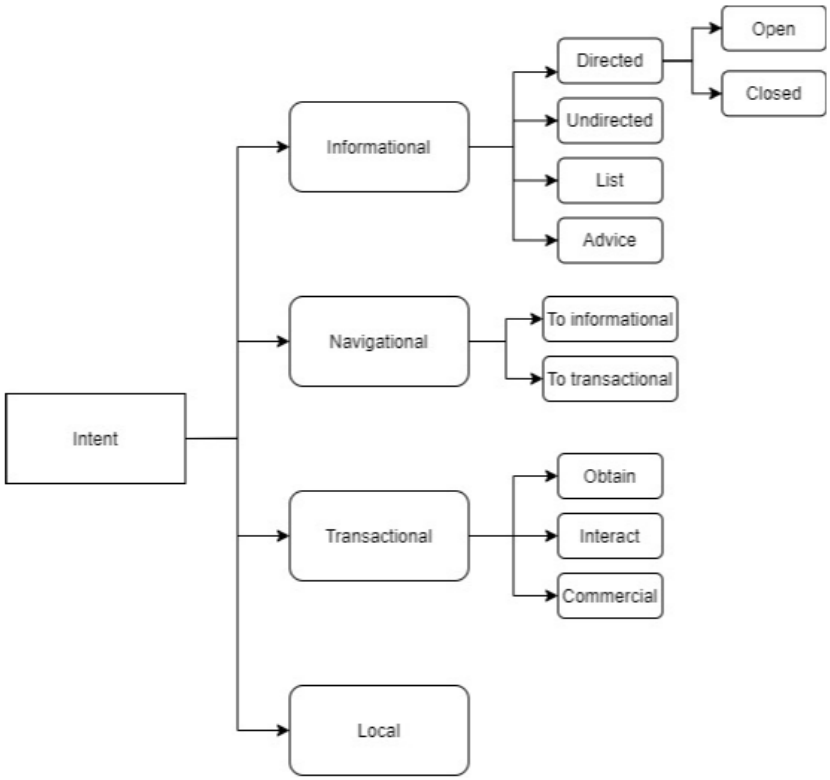


Figure 0.1: Search intent classification framework

Methodology

Since the introduction of BERT, many different transformer models have been developed. BERT has a family of models which work in the same way. For the Dutch language, these are BERTje and RobBERT. As they both claim to be better than the other, a small test has been conducted which shows BERTje is the better pre-trained model to use for this particular search intent classification task.

To collect training data for the model, third-party tools such as SEMrush, GetStat and Google Search Console have been used. These keywords have been labeled on the search intent classification framework as shown in Figure 0.1. A check was performed whether the intent was labeled correctly. Navigational intent was checked with Google Search Console data on high click-through rates, transactional intent was checked using Google Ads conversion data to see if this intent has the highest conversions. Furthermore, since the dataset was labeled by multiple persons, an intercoder reliability check was performed.

Results

The final result of the model is shown a confusion matrix in Figure 0.2. The accuracy on the test set of the model is 94.7%, over 9 different categories. When considering only the top 4 higher-level search intents, the accuracy is 96.3%.
Important note is that the classes ‘undirected’ and ‘obtain’ have not been used in the final model. This was due to the low number of keywords available for both classifications.

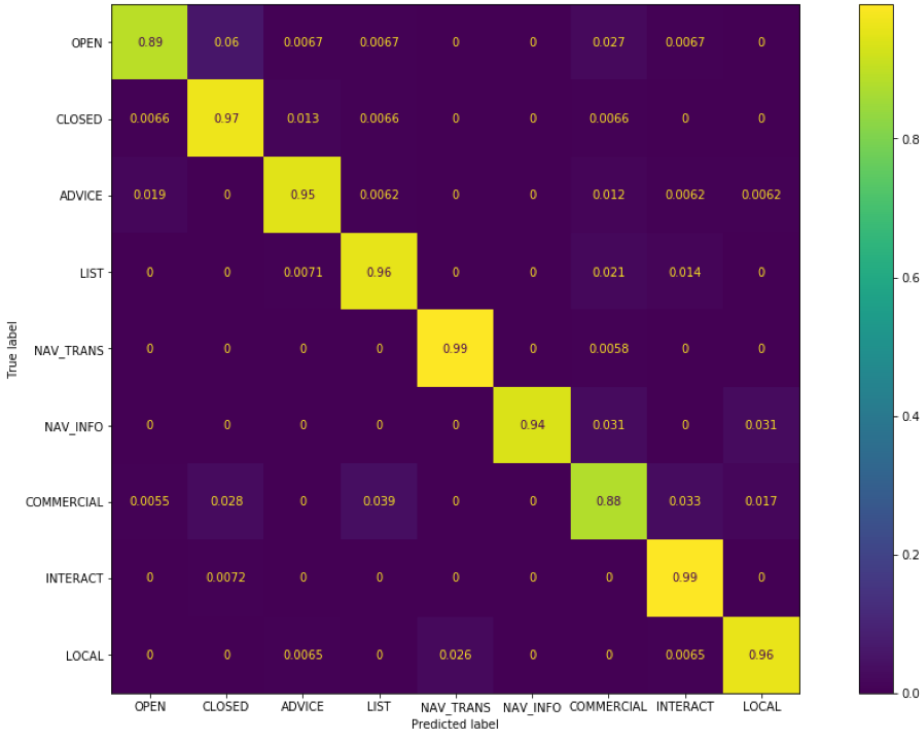


Figure 0.2: Model result

Conclusion

Based on this research, it can be concluded that using a pre-trained transformer-based language model it is possible to outperform the models used by other researchers on accuracy for search intent classification.

In addition, it has been demonstrated that using synonyms in the training set is a good way for data augmentation for search intent, which resulted in an increased test accuracy.

From a business perspective, this model will allow for high accuracy predictions of search intent in the tested domain (telecom). This will help Greenhouse to predict search intent on a larger scale. In turn, this knowledge will help to optimize pages for SEO or ads for SEA to have a better fit with user intent. This allows for a more effective spending of marketing budgets and the classification can be done more efficiently because of automatic classification which will save time (compared to manual classification).

Furthermore, it should be easy to deploy this in different domains (by fine-tuning on a different training dataset), so all clients in all domains can be supported better in their online search marketing endeavors.

Table of contents

List of figures

List of tables

List of abbreviations

BERT	-	Bidirectional Encoder Representations from Transformers
CNN	-	Convolutional Neural Network
CTR	-	Click Through Rate
CPC	-	Cost Per Click
GSC	-	Google Search Console
IR	-	Information Retrieval
LSTM	-	Long Short-Term Memory
NLP	-	Natural Language Processing
RNN	-	Recurrent Neural Network
SEA	-	Search Engine Advertising
SEM	-	Search Engine Marketing
SEO	-	Search Engine Optimization
SERP	-	Search Engine Result Page
SVM	-	Support Vector Machine
TF-IDF	-	Term Frequency-Inverse Document Frequency

1. Introduction

As of 2019, there are over 4.5 billion internet users which account for almost 60 percent of the world's population (Miniwatts Marketing Group, 2019). These numbers keep increasing, as in the past year the population grew with 1.1% while the internet users grew with 7.0%. When looking at developed countries, the numbers are even more impressive, with internet penetration rates of 95% in Northern Europe, 92% in Western Europe and 88% in Northern America (Clement, 2020a).

People are also no longer limited to an Internet (Wi-Fi) connection at home. Nowadays, people are always connected due to the adoption of mobile internet and smartphones.

The Internet is used for many different goals. These include educational, recreational, job-related and personal use (Slone, 2003). Based on the most visited websites in the world, searching is also a large part of how time is spent on the Internet.

When looking at the most visited website in the world, Google.com tops the rankings. In the top 50 there are eight other search engine websites (of which three are local Google sites) (SimilarWeb, 2020), showing many people use a search engine in their online journey. Of the search engines, Google has a dominant market position with over 86% of all searches (Desjardins, 2018) going through its platform while others report a market share of over 92% (StatCounter, 2020). Because Google has such a large market share, the focus in this report regarding search engines is on Google.

On a typical day, Google handles at least 63,000 search queries per second (Sullivan, 2016). That means at least 5.5 billion queries a day, in 2016. Of those queries, about 15% is unique and has not been seen before on a daily basis (Nayak, 2019). Furthermore, its search volume keeps on growing, with 10 to 15% each year (Internet Live Stats, 2020).

1.1 Search results

The business model of Google (and other search engines) has been built around search behavior of users. About 60% of the 2019 revenue came from Google Search. Included with other ad related income from the Google Ad Network (e.g. YouTube ads and Google Network Member ads), this amount increases to over 83% (Alphabet, 2020). The revenue has been growing almost exclusively in the double digits each year (Clement, 2020b). This has made Google's mother company Alphabet one of the most profitable (Fortune, 2020) and most valuable by market capitalization (PwC, 2019) companies in the world. Google on its own has become one of the most valuable brands in the world (Forbes, 2019).

Thus, the main income for Google, but also other search engines, is generated by the advertisements in its search engine. Advertisers can bid on a certain keyword, where the ads are shown when a user inputs the query which contains this keyword. For example, when an advertiser wants to be shown on the word 'Eindhoven', a sponsored link will be shown each time a user performs a search including that word (see Appendix A for an example). When the user clicks on the sponsored link, the advertiser pays a fee to Google. This is why this type of advertising is also called pay-per-click (PPC) (Google, 2020b).

This digital search ads market is large. Total digital search ads spent in 2019 was $127.9 billion and is expected to grow to $162.7 billion by 2023 (Statista, 2020). In comparison, ad spending on the TV accounted for $140 billion in 2018 (Friedman, 2018). Supposedly, TV ad spending had peaked in 2018 (in the US) and is on a decline for the following years, whereas digital advertisement is on a rise and accounts already for more than half of all ad spending (MAGNA, 2018).

Next to the paid search results, there are organic search results. These organic results are the 'normal' results when searching for a query. As can be imagined, businesses would like to rank as high as possible on these keywords, especially because the top three collects over 75% of all the clicks (Dean, 2019a). There are businesses that sell a service focused on gaining a higher ranking and more ranked keywords. This service is called Search Engine Optimization (SEO). SEO is different from digital advertising, as one cannot buy high positions (that is what the search ads are for). In terms of revenue, SEO accounts for $72 billion in 2018 in the US alone (McCue, 2018).

In conclusion, there is much money spent on getting traffic and results from search engines.

Academics try to understand the algorithm of Google Search as well, for both paid and non-paid results for years (Sedigh & Roudaki, 2003; Furnell & Evans, 2007; Killoran, 2013; Aswani et al., 2018).

The search results shown by search engines are nothing more than information retrieval (IR) systems (Burges et al., 2005), as IR is providing a ranking or list of documents given a specific query (Wang et al., 2017). The central belief of IR has been that the user is motivated by an information need (Broder, 2002). However, this might not always be the case.

An increasing interest is about understanding the intention behind a search query, on a broader domain than just information need. Broder (2002) was the first to introduce taxonomies for web search. He argued that the need behind a search query is not always informational, but can be navigational or transactional as well. For instance, it seems obvious that a keyword like 'buy iPhone 12' has some kind of a transactional intent as the user using this query is looking to buy and iPhone 12 (because they explicitly let the search engine know with their query). Rose and Levinson (2004), Jansen, Booth and Spink (2008), and Lewandowski, Drechsler and von Mach (2012) extended the search intent model with new (sub)categories (which will be discussed later).

Schultz (2019) combined search intent and PPC advertising, showing different intentions have an effect on conversion metrics. However, Schultz (2019) uses manual labelling for a small (321) number of keywords. Automatic classification should make it easier to test larger datasets. The research of Schultz (2019) does show the importance of knowing the intention of search queries, as different intents correspond with different objectives (e.g. conversions or awareness/reach). Businesses can assign their budgets accordingly. This should result in a more effective converse with the customer target groups. In turn, more potential customers which are in line with the business objective will find their way to the website. This will in turn increase the amount of conversions and/or decrease the average cost per conversion (Ludwig et al., 2013).

There has been done work on automatic intent classification using machine learning models with a focus on increasing the accuracy of classification (Lee et al. 2005; Kathuria, Jansen, Hafernik & Spink, 2010; Mendoza & Zamora, 2009; Tsukuda, Sakai, Dou & Tanaka, 2013; Figueroa, 2015; Qiu, Chen, Jia & Zhang, 2018). Progress in machine learning and natural language processing (NLP) develops really fast, already achieving human-quality results on benchmarks, and increasing benchmark scores every year (Zhou, Duan, Liu & Shum, 2020).

Just like previous research on automatic intent classification has focused on optimizing for accuracy, this research will aim for higher accuracy levels using updated state-of-the-art developments in NLP. This in turn will help to improve research on search intent for paid and non-paid search results, by being able to use larger datasets with higher degrees of accuracy.

Furthermore, when the model is capable of labeling queries automatically with high accuracy, it might be possible to unravel keywords from larger datasets which can be interesting for an advertiser to use in their search campaigns. The previous example of 'buy iPhone 12' is pretty obvious for every advertiser trying to sell an iPhone 12. But perhaps there might be keywords which show a transactional intent, but are not as obvious and not in sight for the advertiser yet.

1.2 Company introduction

The company at which this master thesis is performed is Greenhouse, located in Eindhoven. Greenhouse is a leading online marketing and advertising company in the Netherlands. It helps clients perform online in terms of search (paid and non-paid) marketing, social media marketing, media buying and creatives. They have a focus on creative, innovative, and data driven approaches.

Many large businesses outsource their digital media advertising to specialized bureaus such as Greenhouse, as this is their core activity and know how it works and are able to get a positive return on investment.

Currently, a new business proposal from Greenhouse lies around 'intent marketing'. This means a better understanding of the intentions of the user, in this particular case the query. They believe search as an advertising channel is underestimated (or at least misjudged), as searchers explicitly tell what they want and are looking for. This is a lot different than other channels such as display (banners), email or social media marketing. By gathering queries and working on the intent behind a query, insights can be delivered for clients and inhouse practices.

Clients are eager to have this information, as this could help them with optimization of their web pages, resulting in a better alignment between what the page offers and what the user is looking for. This should improve attracting the right users visiting those pages as well as helping convert those users, as companies have a better understanding of what their users are looking for.

Furthermore, the understanding of query intent can help optimize the internal ways of working in Greenhouse. For instance, to optimize the paid ads even further. When understanding what the user might be looking for, ad copies and keyword selections can be adjusted accordingly, and can be automatically assigned to ad campaigns.

Lastly, third-party cookies will be blocked by major browsers which alters the landscape of digital advertising, as it currently relies on third-party cookie data for personalization of ads (Deloitte, 2020). By anticipating on blocking on other data, Greenhouse believes (search) intent will play a large role in making audience target groups.

Of course, this labeling on search intent is all possible already, and is done to some extent manually. However, this is not scalable and thus this should be standardized and automated to be able to scale with ease which allows Greenhouse to use this for all its clients.

1.3 Research question

The main goal of this research is to find a way to be able to automatically label search queries on user intentions.

Hence, the following main research question is proposed:

"How can Natural Language Processing (NLP) techniques be used to automatically classify search queries based on user intentions?"

The objective of this research is to use search query data and label them based on classifications of intentions. As a result, the model used can be evaluated and compared with previous academic work. Furthermore, Greenhouse should gain valuable insights in order to further innovate and develop its value proposition for its clients in search engine marketing.

In order to answer the main research question, several specific questions need to be answered. This will help in understanding the concepts in this research, the proposed value for Greenhouse, the added value from an academic perspective, and the NLP techniques which have been developed over the years. The specific questions are proposed as follows:

- What is search intent?
- How to classify search queries into categories of search intent?
- How does understanding search intent help with increased paid search results?
- How does understanding search intent help with increased organic search results?
- What Natural Language Processing techniques are currently helpful to interpret search intent?

1.4 Academic and practical relevance

This research is relevant for both academic and practical parties. In the following paragraphs, the relevance is explained.

1.4.1 Academic relevance

Over the past decades, there has been done much research relating to query understanding (Zhang et al., 2020). At first, researchers tried to put much effort with human analysis into identifying search intent (Shneiderman, Byrd & Croft, 1997). Later on, more automated analysis was done. Search intent was attempted to be discovered by clustering, but it was difficult for humans to clearly understand the clusters and thus be interpretable (Beeferman & Berger, 2000; Wen, Nie & Zhang, 2002). Classification might be better suited, as with classification, the queries are classified into predefined categories.

Broder (2002) was the first who introduced taxonomies for web search. These are defined as informational, navigational, and transactional. In his research, it was reported that over 70% of queries had an informational type (with overlap). This might not be useful, as one can achieve an accuracy of 0.7 by putting all queries in this category. It should be noted that only 3,190 keywords were used. Rose and Levinson (2004) used a larger dataset and noted 62% keywords fall in the informational category.

Rose and Levinson (2004) and Jansen et al. (2008) extended the main three categories with subcategories. Lewandowski, Drechsler and von Mach (2012) extended this framework with two new categories. Jansen et al. (2008) reported an informational level classification on more than 80% of occurrences (on a 1.5 million keywords dataset), showing the importance of sublevels. By using sublevels, those 80% of keywords can be broken into the sublevels, which will help to diversify more on intent. This in turn can improve web search. Furthermore, high accuracy scores can be achieved if almost all keywords are in one category. Jansen et al. (2008) got an automatic classification accuracy of 74%, which is worse than just assigning every keyword to the informational classification.

Current research labeled intent based on what users said they did (e.g. Broder, 2002), manual classification (e.g. Rose and Levinson, 2004), automatic labeling (e.g. Jansen et al., 2008) or user clickthrough data (e.g. Lee, Liu & Cho, 2005). No user behavior data (e.g. buying a product) has been used. This research will contribute on that level, as search intent has been mostly guessed before when labeling, based on the query, or asked what the searchers behavior was even though there is a difference in what people say they want to do and actually do (intent-behavior gap (Blake, 1999)). No additional checks on whether it indeed corresponds with the user's intent have been performed. By using behavioral data, it is possible to check what the true behavior and intent of a query was and if the labeling thus corresponds with real intent.

Furthermore, the current research has a focus on all the searches performed in a language, without being restricted by a specific market (e.g. automotive, households, telecom and health searches). However, it is very likely that for different markets different results will be found. This can explain the difference between Broder (2002), Rose and Levinson (2004), and Jansen et al. (2008) on percentages in informational queries.

Also, many of the previous work done on intent classification is trying to improve the accuracy of classification models (Lee et al. 2005; Kathuria, Jansen, Hafernik & Spink, 2010; Mendoza & Zamora, 2009; Tsukuda, Sakai, Dou & Tanaka, 2013; Figueroa, 2015). As machine learning algorithms advance quickly, accuracy can be improved even further. For

example, by using state of the art machine learning techniques used by Google themselves, the search results can be optimized (see the Methodology section for detailed information). Furthermore, the previous work has been done several years ago, in machine learning terms it is almost ancient as developments are following each other very quickly. Almost every year, NLP algorithms reach new high scores on benchmark scores (Zhou et al., 2020) and the papers published on NLP increase year after year (Mohammad, 2020). In particular, research on search intent classification has not adopted the use of transformer-based language models, which allows for much larger (pre-trained) models (Vaswani et al., 2017).

Lastly, many researchers on this topic conclude with possible ways to improve search engines (Athukorala et al., 2016). This research, however, takes the search engines 'as is' without a need for change, but can help businesses by providing an understanding on how to increase their search rankings. There has been done research on SEO, but not in combination with search intent.

1.4.2 Practical relevance

There is also practical relevance for this research. First of all, in recent years search queries have changed, as people understand that search engines understand them better and from developments such as voice search (e.g. Google Home or Alexa, which results in longer queries) (Patel, 2020). The so called 'long tail keywords' have become very important, because (a) they are more specific (e.g. intention should be clearer), (b) it is generally easier to rank on those keywords and (c) about 40% of all search volume comes from long-term keywords (Soulo, 2019).

Furthermore, in recent years, intent became a hot topic in search engine marketing (SEM). Many respectable blogs came up with their views on this topic and how to implement this for a website strategy (e.g. Yoast, 2018; Dean, 2019b; Agius, 2019; Searchmetrics, 2020). These all require many manual tasks, which does not allow to get the full possibilities as there is an enormous number of keywords. They do, however, all agree that an understanding of intent will help with attaining higher rankings. Hamlet Batista from Search Engine Journal did automate classification using machine learning algorithms, but based on categories instead of intentions (Batista, 2019). Categories might be useful for a specific website. Batista (2019) used categories from BBC news articles. This, however, is not generalizable to other websites. Hence, intent might be a better classification as this works for all websites.

In addition, knowing intent is very valuable information as this might reveal a user's mindset which can be used to design texts, creatives, bids, and landing pages for all possible channels, even offline (Sullivan, 2020).

By automating these tasks, a much larger set of data can be classified. This should result in more and deeper insights for businesses who wish to take their online presence to the next level.

A current practical problem is that there is no model available which can automatically classify search queries on search intent. Models have been developed in the past (as discussed in 1.4.1), however they are either not available for everyone to use (closed

source) or lack behind on recent developments in NLP, suggesting the accuracy is not optimal. As discussed in 1.2, such a model can help Greenhouse with their business.

1.5 Outline

This report is split into several chapters. In the first chapter, the reader has been introduced to the topic and scope of SEM. Furthermore, the company at which this research was conducted has been introduced, along with the research questions. Next to that, the academic and practical relevance have been covered.

The second chapter is a theoretical background which will help understand the concepts used in this research. It also gives answers to the sub questions proposed in the research question section.

Chapter three discusses the methodology, model and data used to get results for the main research question.

The fourth chapter discusses the validation of the dataset as well as the main results resulting from the model.

In chapter five, the discussion on the research is laid out. This includes a conclusion from the results, theoretical and managerial implications, the limitations of this research and possible future research directions.

2. Theoretical background

This chapter will help to get a better understanding of the concepts used in this research by discussing the theoretical literature. Also, this will help in answering the research questions which have been proposed in the previous chapter.

The first section will explain search query intent in more detail, and will show classifications of search queries for search intent. Then, an explanation will be given as to why intent should help with getting better results in the search engines rankings, both paid and non-paid. The final section will elaborate on the machine learning and natural language processing technique strengths and weaknesses and how these techniques have developed over the years. This should result in a method which can be used to form the basics of the Methodology chapter.

2.1 Online information seeking behavior

Information seeking behavior is the active engagement with texts or people who provide access to texts to find the information needed (Belkin 1993). This has been studied for some time (Schutz & Luckmann, 1973; Wersig, 1979; Dervin, 1983; Belkin, Seeger & Wersig, 1983). For a very long time, academics believed that the information demand of users were defined as information classes in a narrow sense (Qiu, Chen, Jia & Zhang, 2018). Broder (2002) was one of the first who proposed that users do not only perform information retrieval for information when on the web, and this idea has gained traction from academics since. Jansen et al. (2008) argue as well that web search is a whole lot different. While web search has the same aspects as defined by Belkin (1993), there are important differences. This is because of the direct availability, the scale at which web search is performed and the variety of content, users and systems (Jansen et al., 2008). Therefore, information seeking behavior for web search has its own domain of study.

As online information retrieval in combination with queries only became relevant in the internet era, research has been done for only a few years on this subject, but it has received much attention. Research on types of web queries has been done in a couple of ways: query topics (Li, Zheng & Dai, 2005), query ambiguity (Song, Luo, Wen, Yu & Hon, 2007) and query intent/task type (Broder, 2002). For this research, the research stream of query intent is used. This is because query topics are not very useful for a business who already knows in what domain the business operates and query ambiguity is not going to help improve search performance from a business perspective as much.

2.2 Search query intent

One of the first fields in research on query intent was aimed at the strategies and tactics of users. O'Day and Jeffries (1993) came up with three strategies: monitoring, following a plan and exploring. Others have categorized search tasks as fact finding and exploration (Navarro-Prieto, Scaife, and Rogers, 1999), or as finding, exploring, monitoring and collecting (Morrison, Pirolli and Chard, 2001). Following research on strategies and tactics in web search, researchers started looking at search goals. Rozanski, Bollman and Lipman

(2001) had categorized the goals as a single mission, do it again, quickies, information please, loitering, just the facts, and surfing. Others classified the goals as finding, information gathering, browsing and transacting (Sellen, Murphy & Shaw, 2002).

Those researchers all used lab or panel studies rather than real search query logs. Broder (2002) was one of the first who used real search query logs in his research, and was able to introduce taxonomies for web search. In his paper, he states that classic information retrieval is based on only the informational need of users. However, the need for web search is not only informational, but can also be navigational ('I want to reach a particular website') or transactional ('show me where I can buy, download, perform a transaction'). These taxonomies have been widely used to identify intent for search queries.

Rose and Levinson (2004) and Jansen et al. (2008) built on the taxonomies from Broder (2002) by adding hierarchical subcategories for all three categories. The next section will focus on the classification of Broder (2002) and the additions of Rose and Levinson (2004), Jansen et al. (2008), and to a smaller extent Lewandowski et al. (2012) as these are the most common referred to additions.

2.2.1 Categories of queries

Three main categories were defined by Broder (2002). These are informational, navigational and transactional, where a query could belong to more than one category. The informational queries have an intent to find information. It is assumed that this information is available on the web already. Also, no further interaction is required to obtain the information and the information is already out there (e.g. it is not created in response to the query). In addition, the queries in this category are usually general, such as 'cars' or 'San Francisco'. However, they can also be more specific if for instance the intention is to find an answer to a specific question (Broder, 2002). Navigational queries have the intent to reach a particular site. Either because the user visited it in the past, or assumes that such a site exists. Therefore, navigational queries usually have only one right destination. Example queries would be 'Hewlett Packard' or 'American airlines home' (Broder, 2002). For transactional queries, the intent is to find a website with the goal to eventually obtain a product or service. An example would be the purchase of a product online (e.g. the query 'buy table clock') (Jansen et al., 2008).

As mentioned, research has been done to further extend these main three levels of intent. This will be explained in the next section.

2.2.2 Extension of classifications

The three top level classifications identified by Broder (2002) have been extended with sublevel classifications by Rose and Levinson (2004) and Jansen et al. (2008). The top levels and sublevels can be seen in Figure 2.1. Rose and Levinson (2004) extended the Informational queries with Directed-open, Directed-closed, Undirected, List, Advice, and Locate. The Transactional level was renamed Resource, and included Download, Entertainment, Interact, and Obtain. Jansen et al. (2008) renamed Resource back to Transactional, and added additional sublevels: Download-free, Download-not free, Obtain-online, Obtain-offline, Results page-links, and Results page-other. Also, they added

sublevels for the Navigational intent, Navigational-to-Informational and Navigational-to-Transactional (see Figure 2.1). Furthermore, the authors added characteristics for each category to help define which query belongs to which category (Jansen et al. 2008).

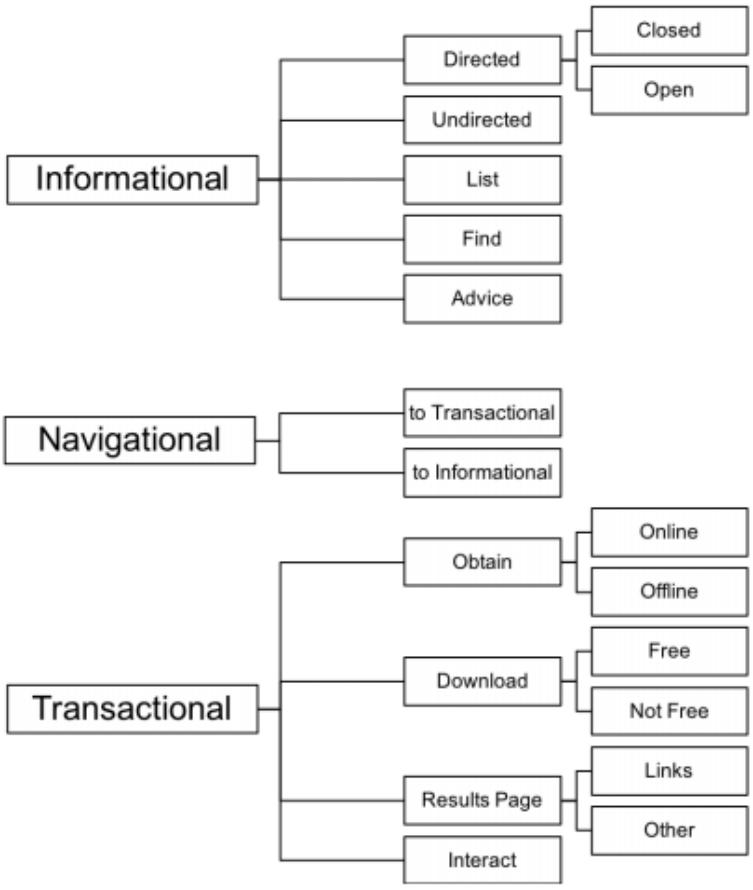


Figure 2.1: Sub level search query intent classification

In this research, some changes will be made to the model as shown in Figure 2.1. This is because the sublevels have changed through time, which will be explained below.

To begin with the Results Page sublevel. This sublevel is determined by Jansen et al. (2008) as “obtain a resource that one can print, save, or read from the search engine results page" (p. 1260). As this sublevel talks about obtaining, this could be added in the Obtain sublevel. Furthermore, search engines results page (SERP) have changed, e.g. by including featured snippets which help gaining information without interacting further with the website (Strzelecki & Rutecka, 2020).

As for Download, this sublevel has been described by Jansen et al. (2008) as “find a file to download" (p. 1260), which again has much in common with obtaining something, and therefore should be considered as the Obtain sublevel.

As for Obtaining-offline, this can be removed as well as the search query entered is online and the object is obtained online. A user could save it (e.g. a wallpaper) offline, but most of the action is done online. Hence this sublevel is dropped.

Much research on query intent also includes a commercial intent (e.g. Kang, 2005; Dai, Zhao, Nie & Wen, 2006; Ashkan & Clarke, 2009). This is different from current transactional levels, as commercial intent might be interested in commercial offerings (Lewandowski et al., 2012). For example, a query such as ‘best frying pan’ can be considered commercial, as a person is gathering information and might perform an interaction into buying one online. Commercial intent is therefore placed as a sublevel under Transactional.

Furthermore, the Informational-find sublevel is changed to a top-level classification Local, as Lewandowski et al. (2012) also have. This is because local search is booming and not necessarily informational, navigational or transactional (Hardwick, 2019). Therefore, a new top level is created for Local. The intent for local queries is to find information and or locate a business near his or her geographic position (Lewandowski et al., 2012). Examples would be ‘supermarket Eindhoven’ or ‘Apple store Amsterdam’

This also has similarities with the Google Quality Rater Guidelines, which defines four major (e.g. top level) intents searchers can have (Google, 2019):

- To know
- Finding a website
- To do
- Visit in person

These guidelines are for human evaluators of search results. They are extensively trained and help with the machine learning algorithms of the Google search engine by providing input on, for instance, expertise of a website, quality of the page, and intent.

This results in the following top levels and sublevels for classifying search queries, as shown in Figure 2.2.

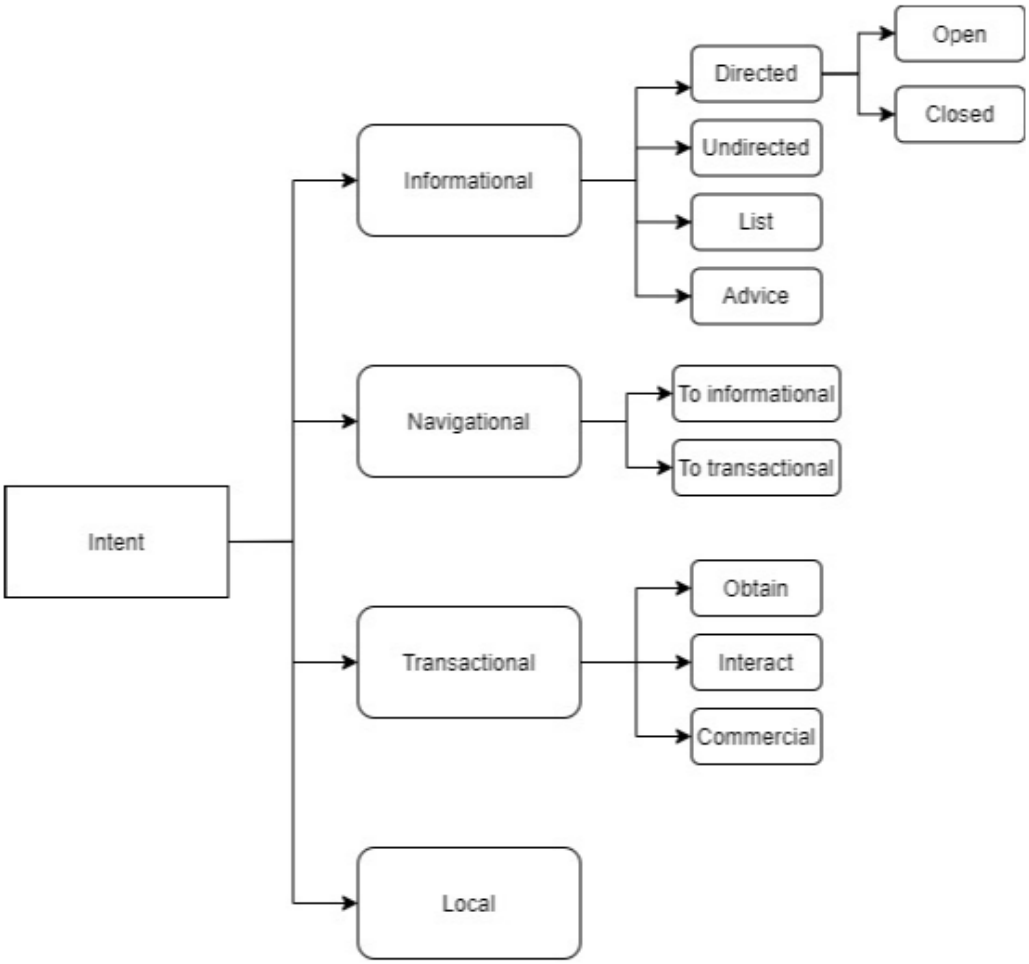


Figure 2.2: Search query intent classification framework

2.2.3 Sublevel explanation

In Table 2.1, an overview of the sublevels is given together with an explanation of the sublevel, and possible queries that fit in that particular sublevel.

Table 2.1

Intent Sublevel Explanation and Example Queries.

Sublevel	Description	Examples	References
Informational Directed Open	Find information on open-ended questions or with unconstrained depth.	'honeybee communication' or 'why are metals shiny?'	Rose & Levinson (2004) and Mohasseb et al. (2014).
Informational Directed Closed	Searching for an answer to a question with one unambiguous answer.	'Nine supreme court justices' or '2020 election dates'	Rose & Levinson (2004) and Jansen et al. (2008).
Informational Undirected	Find information about everything on a topic.	'Color blindness' or 'hypertension'	Rose & Levinson (2004) and Mohasseb et al.

			(2014).
Informational List	The goal is to find a list of plausible suggestions.	‘Things to do in Hollywood’ or ‘Amsterdam universities’	Rose & Levinson (2004) and Jansen et al. (2008).
Informational Advice	Find advice, ideas, suggestions or instructions on a topic.	‘Help quitting smoking’ or ‘What to serve with pork’	Rose & Levinson (2004) and Jansen et al. (2008).
Navigational to Informational	The URL searchers want to visit is for informational purposes.	‘Yahoo.com’ or ‘Facebook.com’	Jansen et al. (2008).
Navigational to Transactional	The URL searchers want to visit is for transactional purposes.	‘Match.com’ or ‘Amazon.com’	Jansen et al. (2008).
Transactional Obtain	Goal is to obtain a resource, which may be printed out, to use. Not to learn some information, but to use the resource.	‘House document no. 587’ or ‘Music lyrics’	Rose & Levinson (2004) and Jansen et al. (2008).
Transactional Interact	The goal is to interact with a resource on another website.	‘Buy table clock’ or ‘Buy cell phones’	Jansen et al. (2008) and Mohasseb et al. (2014).
Transactional Commercial	The goal is to find more information on a product or service with a potential of buying eventually.	‘Shoe sale’ or ‘best cameras for astrology’	Lewandowski et al. (2012).

2.3 Intent in search engines

In order to rank higher in Google, and therefore increasing total search query clicks and visitors, it is key to understand how Google works. In the next section, there will be an explanation on how search algorithms work and how an understanding of search query intent can influence the rankings in paid and organic search results.

Two main ways on how to get visitors from Google will be explained, with paid search (Search Engine Advertising, SEA) and organic search (Search Engine Optimization, SEO).

2.3.1 SEA: Paid results algorithm

The business model of Google is built on the advertisements in the search results. Search ads are still growing by revenue, as shown in the introduction, meaning advertisers see the benefits of related search ads.

Google tries to show the most relevant information for their users, and that includes ads. On the other hand, they want to generate as much revenue and/or profit as possible. Those ads are therefore based on some sort of quality score, which combines relevancy and potential revenue (Liu, Chen & Whinston, 2010).

As Hillard et al. (2010) note, most search engines have a three-stage approach for sponsored ads:

1. Finding the relevant ads for the given search query;
2. Estimating the click through rate (CTR) on an ad;
3. Selecting how many ads to show.

The finding of the relevant ads is basically done in the same way a search engine retrieves organic results, but together with the input of the advertiser (e.g. on what words they want to rank). In the second stage, most search engines rank the ad. This is based on the estimated CTR and the cost per click (CPC) to maximize revenue (Hillard et al., 2010). Google calls this Ad Rank, which is based on the bid price, ad Quality Score (expected CTR, ad relevance, and landing page relevance), and context of a person's search (e.g. location, device, time, organic results) among other factors (Google, 2020a). It has also been empirically shown that an improved ad Quality Score leads to lower prices per click (Nabout & Skiera, 2012).

As said, the ad Quality Score is based on the expected CTR, relevance of the ad and relevance of the landing page. As searchers look at the title, summary and URL before clicking on an ad (Jansen & Resnick, 2005), improving the relevance can increase CTR as well. Therefore, much gain is to be found in relevance improvements.

Improving the Quality Score is what most advertisers should be doing to improve their paid search results. An improved Quality Score will decrease cost per clicks (Ghose & Yang 2009), increase rankings (Olbrich & Schultz, 2014), which increases CTR (Jansen, Liu & Simon, 2013) and thus potentially decrease the cost per conversion while increasing total conversions, which in turn will result in a higher return on investment and a higher revenue.

When understanding user intent, the relevance of the ad can be determined and adjusted in such a way the relevance has a better fit. This should increase the Quality Score, which will result in increased results for paid search.

2.3.2 SEO: Organic results algorithm

The goal of Google was to ‘improve the quality of web search engines’ by focusing on the most relevant results (Brin & Page, 1998). Originally, Google was found on an algorithm called PageRank, named after one of the founders, Larry Page. PageRank counted the quantity and quality of links pointing to a website, with the assumption that more important websites had more links pointed to them. Relevant sites were found (based on presence of query on the page) and ranked according to their PageRank. In comparison to the then competitor Altavista, “Altavista returns random looking web pages that match the query” (Page et al., 1999, p. 9).

Of course, Google has made major steps since. In recent times, many factors (reportedly over 200 (Furnell & Evans, 2007)) will influence the position in the SERPs with incoming links still being one of the strongest factors (Zhang & Cabage, 2017).

Google explains that their organic results algorithm is based on a few factors. First, the meaning of the query needs to be understood. “To return relevant results for your query, we first need to establish what information you’re looking for−the intent behind your query. Understanding intent is fundamentally about understanding language, and is a critical aspect of Search” (Google, 2020c). Therefore, it is very important for Google to understand the intent of the query, which is based on the search query input.

The next step is for the algorithm to determine the content on the webpage, and whether this is relevant for the search query (e.g. the intent is similar). “The most basic signal (...) is when a webpage contains the same keywords as your search query” (Google, 2020c). Google also includes interaction data from other users using similar queries to see if the page is relevant (Google, 2020c). Not only does the relevance of the webpage count, the algorithms of Google also rank sites to only show reliable sources. This is done with signals that help to determine whether a page has expertise, authoritativeness, and trustworthiness on a topic (Google, 2020c).

People can use many different browsers to access websites, as well as many different devices (e.g. desktop, tablet, mobile, all with possible different resolutions). Google wants their users to find what they are looking for in a fast way, which also includes the way how a page is used. For instance, users want to be able to easily read the text on a mobile phone, without zooming. This usability factor also has influence on the organic rank (Google, 2020c).

Finally, context and settings of a search query also help Google to determine the intent of the query and the relevance of the results (Google, 2020c). For instance, while a user is likely to look for Premier League results when searching for ‘football’ in England, the user is likely to find American football results when the user looks for ‘football’ in the USA. In fact, about 40% of English words are polysemous, meaning a word having multiple meanings (Shoebottom, 2020).

Also, previous queries help. When searching for ‘Barcelona versus Arsenal’, and the next query is just ‘Barcelona’, it is likely that the intention is to learn more about the football club, not necessarily only the city.

Google says it had 3,234 updates to the search algorithm in 2018 alone (Google, 2020d). Most of them are minor updates which will not show a large effect. Some of the updates Google rolls out are major. These major updates (since 2011) include:

- Google Panda (2011): to remove sites with small amounts of content (‘thin content’) and give sites with quality content more presence in the SERPs.
- Google Penguin (2012): to remove sites who intended to manipulate their incoming links to secure higher positions in the SERPs.
- Google Hummingbird (2013): to get a better understanding of natural language queries and considering context.

- RankBrain (2015): to get a better understanding of the query by applying machine learning, which should result in more relevant results for users.
- BERT (2019): to even better understand context and small nuances in text to improve the results. Dubbed “one of the biggest leaps forward in the history of Search” (Nayak, 2019).

To summarize, those updates help Google to get a better understanding of what the searcher is looking for, and what quality and relevant results will help the users in their search goal. Therefore, the updates should help at getting a better understanding of search intent from users.

Basically, it all comes down to relevance, both for SEO and SEA. And to understand relevance, an understanding of user intent is needed (Google, 2020c).

By understanding what the searcher is looking for, pages can be optimized to show more and more relevant information. This should thus help with increased rankings in the search results.

2.4 Text classification and machine learning

Sebastiani (2002) explains text classification as the task of assigning a Boolean value (e.g. only two possible values) to each pair $<d_j, c_i> \in D \times C$, where D is a domain of documents and $C = \{c1, \ldots, c_{|C|}\}$ is a set of predefined classifications. When the value is T (True), d_j is filed under classification c_i, otherwise (with value F) this is not the case.

Therefore, a text can be classified in a predetermined classification class, or not.

There has been done quite some work on text classification in the past, described in the following paragraphs.

The Bag-of-words (Harris, 1954) has been a starting point for many different types of classification tasks (e.g. question (Zhang & Lee, 2003) or sentiment classification (Maas et al., 2011)). It works by making vectors from pieces of text. The vector has a value of 1 if the word the vector represents was in the text, a 0 if it was not in the text. A problem with Bag-of-words is that it does not give words more importance. Words like ‘a’, ‘the’, and ‘or’ have therefore equal values as ‘balloon’. Even though ‘balloon’ probably gives more meaning to the text than the stop words do. One can of course account for stop words, but then do words like ‘big’ and ‘balloon’ still have equal values.

This has been mitigated with methods such as term frequency inverse document frequency (TF-IDF) which appropriately weights terms (Salton & Buckley, 1988). The TF-IDF formula gives words which appear in a document but not in other documents of the corpus a larger weight. Stop words used in almost every document will therefore have a lower weight.

Methods like TF-IDF still have the problem that there is no semantic understanding. For instance, a conjugation of a word which has not been discovered in the corpus, is unknown for the system. Stemming is a way to handle this, by taking roots of the words (e.g. removing prefixes and suffixes). However, this still has issues, as universal, university, universities and

universe would all be stemmed to ‘univers’. This is an issue as universal and universe are semantically different from university and universities.

That problem has been solved with word vectors. Word vectors like Word2vec encode similarities between words in word vectors in high dimensional space (Maas et al., 2011). This has made it possible to give semantical meaning to vectors, which can be added and subtracted. For instance, ‘King’ - ‘Man’ + ‘Woman’ = ‘Queen’. Still, word vectors have some issues regarding language interpretation. That is because a word can have different meanings in different contexts in natural language, but there is one vector per word. For instance, the word ‘bank’ has different meanings in ‘the bank of the river’ and ‘the bank of America’, but the same value in the vector.

Recurrent neural networks (RNN) deal with that issue. That is because RNNs are able to dynamically change the input variables (Sak, Senior & Beaufays, 2014), and therefore can give different meanings to the same word. This makes these kinds of networks better suited for natural language processing.

2.4.1 Text classification and NLP

Natural Language Processing (NLP) is an area of research which looks at how computers can understand natural language in text or speech (Chowdhury, 2003). NLP is generally difficult for computers as human language is vastly different from computer languages. Due to increased computational power, NLP is being used more and more.

Deep learning and deep neural networks have been used extensively to classify texts since the 2000s (Gers & Schmidhuber, 2001). For deep learning algorithms such as NLP to work properly, large datasets and a lot of computational power is required (Zhang, Wang & Liu, 2018). This will be further handled in the Methodology chapter.

NLP has been used in different types of classification for texts, for instance in question answering (e.g. Zhang & Lee, 2003; Hardy & Cheah, 2013), fraud or bot detection (e.g. Ngai et al., 2011; Chu, Gianvecchio, Wang & Jajodia, 2012), sentiment analysis (e.g. Altrabsheh, Cocea & Fallahkhair, 2014; Yang et al., 2017), short text classification (Paalman, Mullick, Zervanou & Zhang, 2019), fake news detection (Vishwakarma, Varshney & Yadav, 2019) and web search (e.g. Hernández, Gupta, Rosso, & Rocha, 2012; Højgaard, Sejr & Cheong, 2016; Huang, Wang, Zhang & Liu, 2020).

This last domain of research (NLP and web search) has a perfect fit with classifying search queries.

Researchers have been using machine learning algorithms extensively for (query) intent classification in the last couple of years, using all kinds of different machine learning methods for natural language processing.

Baeza-yates, Calderón-Benavides and González-Caro (2006) used Support Vector Machine (SVM) to classify user goals (as informational, not-informational and ambiguous) and categories. Kathuria et al. (2010) used k-means clustering to cluster on navigational, informational, and transactional queries. Kim (2014) and Hashemi, Asiaee and Kraft (2015) used convolution neural networks on top of Word2vec and Figueroa et al. (2015) did so with

ensembling multiple classifiers (SVM and Naive Bayes). More recently, researchers started using Long short-term memory (LSTM) (Ravuri and Stolcke, 2015) convolutional neural networks (CNN) (Zhang et al., 2019), and bidirectional approaches (Qiu et al., 2018).

The results reported by researchers differ, and have been summarized in Table 2.2.

Table 2.2
Reported accuracy, precision, recall and F1 scores of automated search intent classifications.

Authors	**Accuracy**	**Precision**	**Recall**	**F1**
Baeza-yates et al. (2006)	-	0.55	0.50	-
Jansen et al. (2008)	74%	-	-	-
Hernández et al. (2012)	-	0.68	0.76	0.68
Hashemi et al. (2015)	90.3%	0.56	0.43	0.47
Mohassed et al. (2017)	-	0.94	0.94	-
Qiu et al. (2018)	84.6%	-	0.92	0.88

The reason the scores are so different can be attributed towards the different datasets used and the developments in machine learning and NLP. Baeza-yates et al. (2006) did a good job on classifying informational searchers with results (precision and recall) of over 50%. For the other classifications not so much. Jansen et al. (2008) reported accuracy levels of 74%, but informational classifications totaled 80% of all queries, so this does not say much as it also does not report precision and recall levels. Hernández et al. (2012) got results around 80% for precision and recall, but scores very low on navigational queries. Hashemi et al. (2015) report precision and recall scores around 50%, but has this on an extended 14 high-level classification. Mohasseb, Bader-El-Den, Liu & Cocea (2017) score a very remarkable score of 94% and above for precision and recall for three main query types (informational, navigational, transactional), however their classification of data might be off (e.g. navigational queries are only classified if there is a domain suffix or prefix). Qiu et al. (2018) report F1 values of 88%, showing bidirectional approaches can lead to high scores.

2.4.2 Bidirectional NLP method

A new promising and already highly cited method for NLP is called BERT, which stands for Bidirectional Encoder Representations from Transformers (Devlin, Chang, Lee & Toutanova, 2018). This is a paper published by Google AI Language researchers which has reached state-of-the-art scores in eleven standard NLP tasks (Devlin et al., 2018).

The big difference of why a model like BERT performs so well is because of the bidirectional transformers. Normally, inputs are being read from left-to-right or right-to-left (directional). The transformer encoder reads the sequence of words at once. This helps the model to learn context based on its surroundings (Devlin et al., 2018).

BERT has been trained with two unsupervised tasks, i.e., masked language model (MLM) and next sentence prediction (NSP). The training data consist of the corpus of BooksCorpus and English Wikipedia (excluding lists, tables, and headers), totaling over 3.3 billion words.

MLM, also referred to as Cloze (Taylor, 1953), masks 15% of the words at random. The model aims to predict the masked words. A downside of this is that the model would not be good at producing tokens for non-masked words, as the model will be optimized for finding masked words. This has been mitigated by replacing the word with [MASK] 80% of time, giving a random token 10% of time and not changing a thing 10% of time.

NSP is a very simple method to train the model in understanding the relationship between sentences. It generates two sentences from the corpus, A and B. In half of the cases B is the actual next sentence following A, in other cases it is not. Even though very simple, the model is able to reach up to 98% accuracy (Devlin et al., 2018).

Furthermore, it has been shown by Pires, Schlinger & Garrette (2019) that BERT is surprisingly robust with multilingual tasks, such as Dutch, German, and Spanish compared to English results.

This pre-trained model is available as an open source model on GitHub (GitHub, 2018), making it a useful potential start for this research.

3. Methodology

This chapter will describe the methodology which in turn will help to answer the main research question.

First, a model will be discussed which can be used to help understand language and to help with labeling queries based on their intent categories. The method of data collection will be discussed, as well as the way to train the model with manual labelling. Furthermore, two sections are dedicated to explain the validation of the (labeled) data and to validate the model.

3.1 Model selection

The last section in the previous chapter gave an overview of the methods and models used in NLP and text classification. Currently, there are open-source bidirectional NLP models such as BERT. This is a method built particularly for the English language, but has multilingual capabilities (mBERT). Recent studies, however, show that models trained on data from one single language perform better in that particular language (de Vries et al., 2019). As the company Greenhouse mainly has Dutch clients, it would seem to be best to have a model particular for the Dutch language.

De Vries et al. (2019) developed a model like this, named BERTje. This model has the same 340 million parameters as the original BERT, but is trained on Dutch corpora rather than English ones. Accordingly, its results significantly outperformed the multilingual BERT.

Very recently, Delobelle, Winters & Berendt (2020) came up with RobBERT, which again increases scores on various NLP tasks in comparison to other models. Again, the training set used is from the Dutch language, but from a larger size (12 GB vs 39 GB of text).

In the paper of Delobelle et al. (2020), RobBERT is compared to BERTje and mBERT (multilingual BERT) on two different tasks, sentimental analysis to show the performance on classification tasks and a Die/Dat pronoun analysis. The results are shown in Table 3.1.

Because of the very recent date and the increased scores on various NLP tasks, these are, to my knowledge, the best NLP models of its kind for the Dutch language. However, one needs to be chosen to be used for the final model.

RoBERT is also (just like BERT and BERTje) available as an open source model, and thus can be used for the classification task discussed in this research.

Table 3.1: Model performance on different tasks

Model	**Sentiment analysis accuracy**	**Die/Dat accuracy**
van der Burgh and Verberne (2019)	93.8	-
Allein et al. (2020)	-	75.03
mBERT	-	98.285
BERTje	93.0	98.268
RobBERT	**94.4**	**98.406**

As can be seen from Table 3.1, RobBERT outperforms the other models on the given tasks, even though just by a small margin in comparison to BERTje. On the Die/Dat accuracy, there is almost no difference between the three different BERT models.

However, BERTje has also made comparisons on certain named entity recognition (classification) benchmarks, comparing its performance to other (including the newer RobBERT) models. This is shown in Table 3.2.

Table 3.2: Model performance on different Named Entity Recognition tasks

Model	**CoNLL-2002**	**SoNaR-1**	**spaCy UD LassySmall**
BERTje	**90.24**	**84.93**	86.10
mBERT	88.61	84.19	**86.77**
RobBERT	84.72	81.98	79.84

Both models (BERTje and RobBERT) claim to be superior to one another. Depending on the benchmark at hand, they are. To determine which one to use, a small test will be conducted to see which model performs better on the search intent classification task. The test will not tweak any architectural hyperparameters, as this is unimportant compared to the scale of BERT (Kaplan et al., 2020).

At Greenhouse, the SEO team has been working on intent segmentation for a while, as this is what the clients of Greenhouse want to have. They have manually labeled a dataset for a particular kitchen supplier in the Netherlands, which is used to test the different models (RobBERT and BERTje).

The labeling has been done a bit different to what is discussed in this thesis. Instead of the identified eleven sublevels, the kitchen dataset has been labeled on four higher level classifications: Transactional, Commercial, Informational, and Local. Because these

classifications do have a lot of similarities with the higher-level categories discussed in section 2.2.2, it has been decided to use this dataset as a test to get an understanding of which model to potentially use. This dataset contains 38,664 manually labeled keywords, of which about 73% is classified as 'Transactional'. For the model, a balanced dataset will be used, so the total amount of keywords used is 11,564 (2,891 per higher level label).

Figure 3.1 shows a confusion matrix on how the model performed on the classification task using RobBERT. This resulted in an accuracy of 0.67.

Figure 3.2 shows a confusion matrix from the same dataset and classification task, but using BERTje as its model. This resulted in an accuracy of 0.91.

Because the keywords are equally divided over the labels, it can be concluded that, based on accuracy, BERTje is the better model for this classification task.

Figure 3.1
Test set using RobBERT

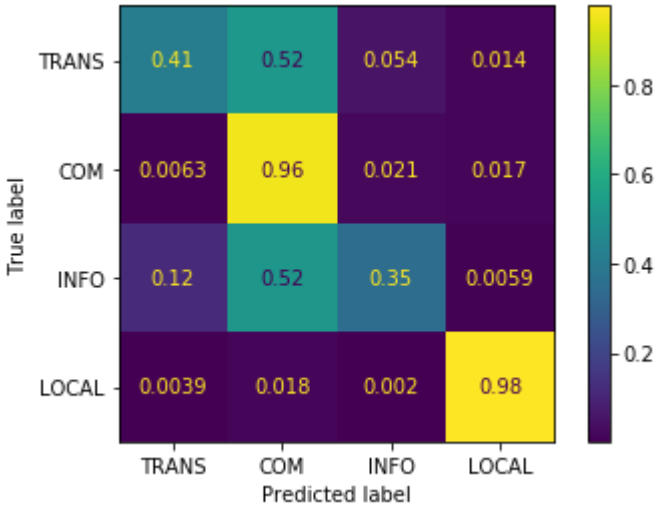
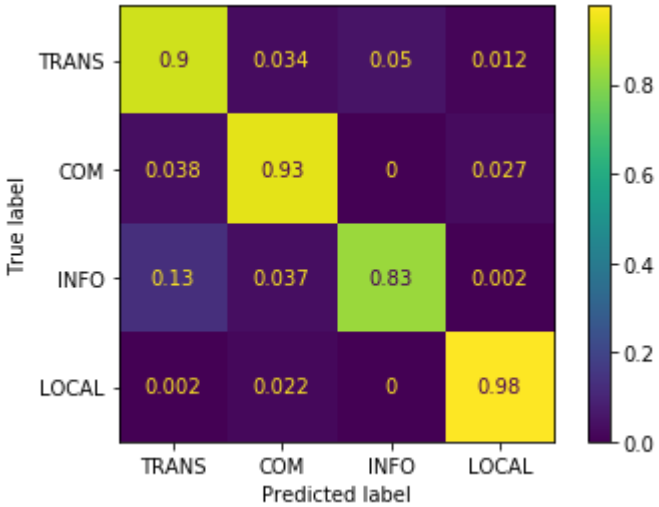


Figure 3.2
Test set using BERTje



3.2 Domain of research

In this report, keywords from a large telecom provider in the Dutch market will be used. There are a few reasons this has become the domain of the keywords.

For once, this particular business contains clients from Greenhouse for both SEO and SEA campaigns. This means there is data which can be used in this thesis to build the dataset which can be used, and to validate the labeled keywords with Google Search Console and Google Ads data on intent (see section 3.5 for more information).

Also, this telecom provider is already interested in search intent insight and reporting. Therefore, if this will work in their domain (i.e. telecom), they are thought to be one of the first paying clients for this.

Furthermore, basically everyone has a smartphone, knows how to search for a new phone or information on a phone related problem and is aware of the brands and possibilities. This should make the (search) market easier to understand, and thus easier to determine search intent.

Because search queries in general are usually short and ambiguous (Hu et al., 2009), it is assumed that the model will perform equally well on a dataset regarding telecom search queries, as it did on a dataset containing kitchen queries.

3.3 Data collection

To conduct the research, data is needed from search queries. The search queries will be gathered from three different third-party data sources: SEMrush, GetStat, and Google Search Console.

Both SEMrush and GetStat are large data software companies in the online marketing domain. They have large databases filled with (for instance) keyword data, which will be used. This has been collected using a combination of scraping the Google search results, Google Ads API, and clickstream data of real users.

With Google Search Console, a website owner can see on which keywords the website has been viewed and clicked on in Google. This also gives an overview of real queries used by searchers for a particular website. However, this is limited to 1,000 rows.

The combination of these three data sources will be checked for duplicates, so each keyword only appears once in the dataset. As the data which is used is from third party businesses who collect their data from real users, the keywords are also used by real persons. This resulted in 19,724 keywords.

3.4 Data labeling

In order to train the model, keywords need to be labeled on their intent. This is a manual task, and requires human thinking. This labeling of the data will be similar to what has been

done in the paper by Ashkan & Clarke (2013) by using Amazon Mechanical Turk, but with using native Dutch speakers. Because of budget reasons due to covid-19, this task could not be outsourced. Instead, colleagues at Greenhouse helped.

In order to have an understanding on how to label the keywords, examples will have to be made. There are existing examples in the literature, but they are in English. As the model will be trained on Dutch keywords, it would be better to have Dutch examples as well. These examples will be made in a Delphi method (Linstone & Turoff, 1975) in cooperation with four SEA experts of Greenhouse. Although Ludwig (1997) reports that most Delphi studies have between 15 and 20 panelists, this was not feasible due to employee constraints. However, as they all are experts on search, it is believed they provide a representative pooling of judgements.

The Delphi method starts with a round of open-ended questions and is useful to gather content specific information (Custer, Scarcella, & Stewart, 1999). The SEA experts will be informed on the theoretical background (as shown in Figure 2.2) with English examples for each category. They will be asked to individually come up with Dutch keywords for each category. In the next round, the Delphi panelists are asked to review the keywords provided by the other panelists (Hsu & Sandford, 2007). This will result in a consensus being formed between the panelists (Jacobs, 1996). The third round gives the panelist the possibility to specify why they made the decisions they made, and thus to further clarify their judgements and the relative importance of items (Hsu & Sandford, 2007). In the final round, the panelist had the opportunity to revise their judgements, and to reach a consensus.

Because the available SEA specialists worked on automotive campaigns, the example keywords for each intent group are related to automotive. However, it is assumed these examples will work just as fine for any other market in the Netherlands. Just like previous research (e.g. Rose & Levinson, 2004; Jansen et al., 2008) had example keywords which would work for other markets as well.

When the categories have been assigned keywords unanimously, there are specific examples in the Dutch language for the labelers to consult. They then will have example keywords for each category, and will be more precise in their labeling.

This resulted in the examples shown in Table 3.3.

Table 3.3
Dutch examples intent keywords

Sublevel	**Examples**
Informational - Directed - Open	Wat is private lease, wat kost een elektrische auto,
Informational - Directed - Closed	Peugeot 208 prijs, technische specificaties [merk][model]
Informational - Undirected	Private lease, SUV, tanken
Informational - List	Beste private lease auto’s, lijst autodealers

Informational - Advice	Private lease reviews, wat is de beste gezinsauto, hoe verwissel ik een band
Navigational to Informational	Opel.nl, Autoweek.nl, ANWB
Navigational to Transactional	Auto kopen opel.nl, autoscout24.nl
Transactional - Obtain	Peugeot 208 brochure, verbruikcijfers SEAT Ibiza
Transactional - Interact	Tweedehands auto kopen, nieuwe auto kopen
Transactional - Commercial	Private lease aanbieding, goedkoopste private lease
Local	Autodealer Eindhoven, Dealer Noord Brabant

Due to stricter budget reasons, the actual labelling of the data which will be used for the model will be done by myself and a colleague of Greenhouse. This will be done while being supervised and helped where possible and necessary by the SEO team of Greenhouse.

3.5 Data validation

In order to check whether the labelers have the same understanding of the intent categories and search queries, cross checking will be performed. With intercoder reliability "the extent to which the different judges tend to assign exactly the same rating to each object" is measured (Tinsley & Weiss, 2000, p. 98). This intercoder reliability is performed between the labelers. To what extent an acceptable level of agreement is reached, is hard to say. Krippendorff (2004) mentions that it depends on the costs of drawing invalid conclusions from the data. For instance, when human lives are dependent on the results, the criteria should be far higher. Krippendorff (2004) does suggest a level of acceptance. Reliabilities below 0.667 should not be accepted.

Furthermore, to verify that a keyword does belong to a true intent, verification will be done using search and purchase behavior from users. For users who have a navigational intent (e.g. knowing what website they want to visit), it is expected they have a very high (>70%) click through rate in the search results. Normally, a number one position in the organic search results can expect to get 17-33% of the clicks (Kim, 2016). Thus, by deliberately choosing a much higher number, it is hoped that this shows true user intent for visiting that particular website. This can be verified using click through rate data from Google Search Console of a specific website.

Moreover, as Greenhouse also performs the search advertising for its clients, it is possible to track which keywords lead to purchases, and thus which keywords belong to the transactional intent. This can be verified with the conversion metrics from Google Ads. It is assumed that the keywords with a transactional intent have a higher conversion rate than keywords with another intent.

This leaves local and informational. For local, all keywords with a city name should be considered. All other keywords are deemed to be informational.

3.6 Model validation

As a final part, the model itself needs to be validated. This will be done with the labeled dataset.

The dataset will be split in two parts:

- Training set
- Test set

The training set (80% of the set) is used to train the model on. After training the model, the model is tested using the test set (20% of the set), to see if the outputs of the model (e.g. the given intent) correspond with the intent as classified by the labelers. Based on the results, certain adjustments can be made (e.g. adding new data) to train the model again, and to see what this does to the test set. This way, it is possible to verify that an increase in accuracy for the training set will also yield in an increase for data the model is not trained on.

4. Results

This chapter will describe the main results of the model and help with answering the main research question. In the first part, descriptive statistics of the data will be shown.

Furthermore, the data is validated in two ways. One side of the validation is about the intercoder reliability, which is to what extent the different labelers agree on the given labels to a keyword. The other side of the validation has to do with the true intent of the users, and is checked using click through and conversion data.

4.1 Descriptive statistics

In Table 4.1 an overview of the number of keywords per intent group is given. This shows that some intent groups are more common than others (e.g. INF_ADVICE and INF_OPEN vs. INF_UNDIRECTED). Table 4.2 gives an overview of the top-level intent groups. It is clear to see here that 'INFORMATIONAL' keywords are far more often present in the dataset, just as Broder (2002), Rose & Levinson (2004), and Jansen et al. (2008) reported.

As this results in an unbalanced dataset, the dataset has been balanced when used for training and validating the model. Keywords will be included at random until the max sample size per intent group is met. This should result in a more effective accuracy (Han, Kamber & Pei, 2012). As INF_UNDIRECTED and TRANS_OBTAIN are both very small in size, it has been decided to cut them out of the final dataset on which the model is trained on. This is because the model would otherwise be trained on very few samples (61 per intent group vs more than 400 per intent group when leaving both out).

Figure 4.1 shows the range of searches per month for keywords in the particular intent group. It is easy to see that the highest volume keywords are the one in the 'Undirected' intent. These include keywords such as 'iphone', '5g', 'wifi', and 'glasvezel'. The intent 'Obtain' also has some very high-volume keywords such as 'specificaties iphone 8' and 'internet snelheid testen'.
However, when adding all the keywords together, the most search volume by far is in the 'Commercial' intent group, as visualized in Figure 4.2 (and log scaled in Figure 4.3).

In Figure 4.4, the average CPC is given for each intent group. Many of the CPC's are close to each other. It is worth noting that 'Navigational' keywords have the lowest CPC's, which can be a result of low competition and clear intent (and thus a high quality score). Furthermore, 'Commercial' intent has the highest CPC per keyword.

Table 4.1: overview of labeled keywords per intent group

Intent	**Count**
INF_ADVICE	4387
INF_CLOSED	834
TRANS_COMMERCIAL	3031
TRANS_INTERACT	629
INF_LIST	602
LOCAL	3448
NAV_INFO	427
NAV_TRANS	2306
TRANS_OBTAIN	107
INF_OPEN	3892
INF_UNDIRECTED	61
TOTAL	**19724**

Table 4.2: overview of labeled keyword per top intent

Intent	**Count**
INFORMATIONAL	9776
NAVIGATIONAL	2733
TRANSACTIONAL	3767
LOCAL	3448
TOTAL	19724

Figure 4.1: range of search volume per intent group

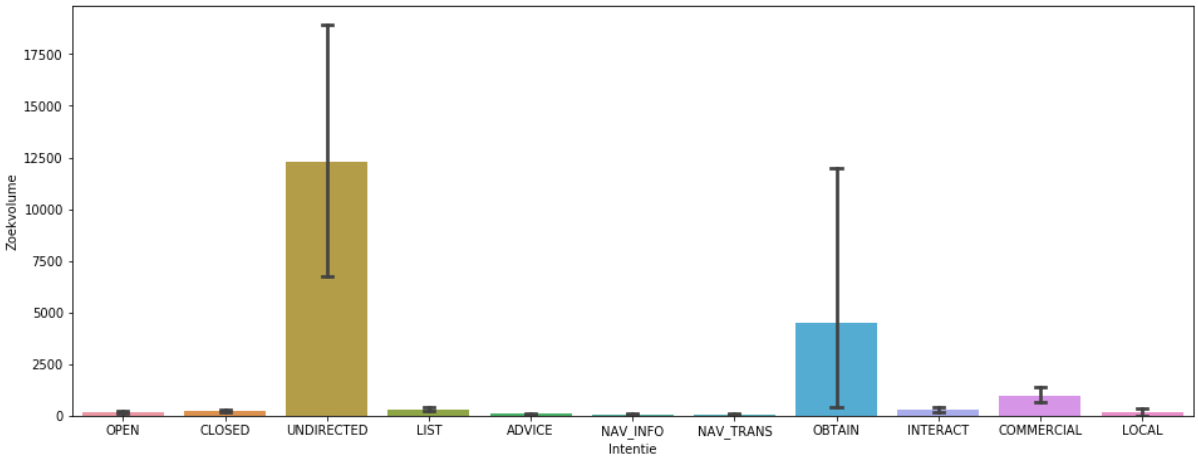


Figure 4.2: sum of search volume per intent group

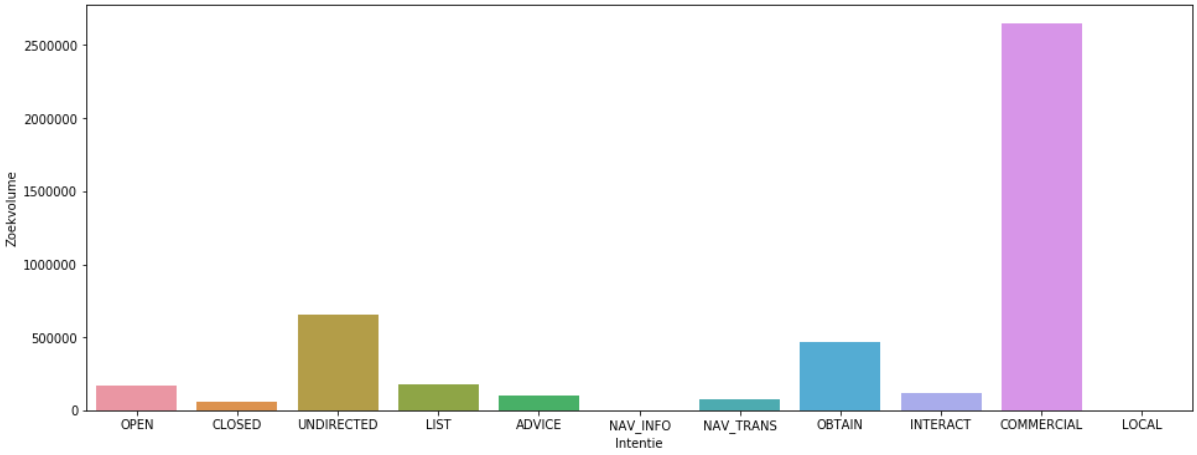


Figure 4.3: sum of search volume per intent group LOG SCALE

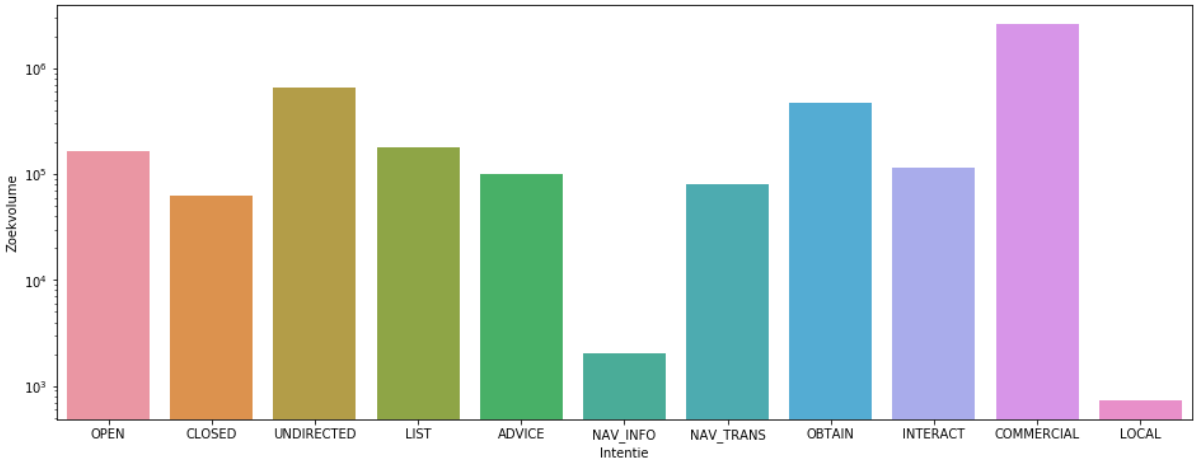
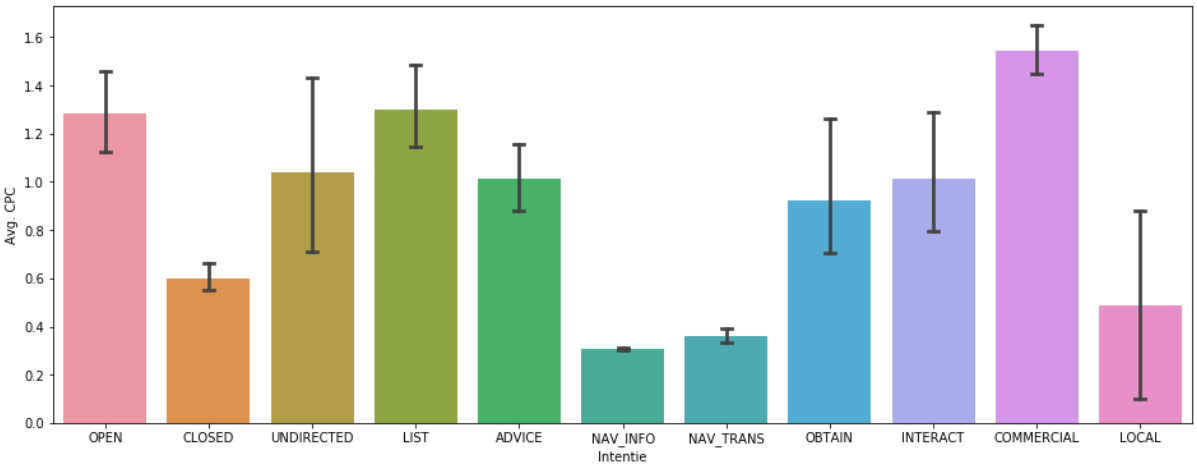


Figure 4.4: overview of cost per click per intent group



4.2 Data label validation

In the next sections, the validation of the dataset is given. The dataset is validated on two aspects:

- Whether both labelers agree on what labels to give to what keywords (i.e. intercoder reliability)
- Whether the intent label given to a keyword truly corresponds with the intent of the searchers.

Both aspects will be discussed in the next sections.

4.2.1 Intercoder reliability

The label validation has been done on a random set of 300 keywords which have been labeled by both labelers. To measure the intercoder reliability, Cohen's Kappa can be used. Cohen's Kappa resulting from this is 0.92, which suggests an almost perfect level of agreement between the labelers (McHugh, 2012). The keywords on which not an agreement was found have been discussed between the two labelers. Main reason for not mentioning the same intent category was the way the keyword was interpreted. As keywords can be ambiguous with possible multiple intents, the labelers have labeled differently in some cases but could understand why the other did what they did. An example keyword which was labeled differently is 'hoe verander ik mijn wifi wachtwoord', which has been labeled as advice and as closed (as there is often only one way to change the password).

Therefore, the labelers have both the same sense of which intent label belongs to which keyword, and thus the dataset is ready to be used for the model.

4.2.2 Intent validation

The next section is about the validation of intent. This is done using data from Google Ads for conversions and Google Search Console for CTR.

4.2.2.1 Validation of Transactional keywords

When looking at the validation of the Transactional keywords, something interesting happens. In Figure 4.5, boxplots are shown for the conversions per keyword per intent. It is clear to see that many of the conversions also happen within ‘Undirected’ keywords, and not just ‘Commercial’ keywords.

This can be because of the limited number of keywords in the ‘Undirected’ category (of which all are of high volume) and the market specifics. A keyword such as ‘iPhone’ or ‘mobiel’ have been labeled as ‘Undirected’, because it does not give a true intent. However, many advertisers do pay to be visible on this particular keyword, resulting in conversions. Perhaps in other markets (outside commercial markets in which the clients of Greenhouse are present), ‘Undirected’ keywords are much less interesting to advertise on (e.g. ‘color blindness’ or ‘hypertension’ from section 2.2.3).

If the conversions are summed for each intent group, the ‘Commercial’ intent group has the most conversions, as expected. This expectation comes from the fact that the commercial intent is explained in 2.2.3 as “The goal (...) to find more information on a product or service with a potential of buying”. The conversions per intent group is shown in Figure 4.6. Based on this, there is some confidence that the transactional keywords are labeled correctly, as they show the most conversions.

It is worth noting that the dataset with labeled keywords and the dataset from Google Ads (with conversion metrics) do not fully overlap. This is simply because the keywords used in the Google Ads campaign and the dataset extracted from third party sources are different. This also means there is room for improvement on the Google Ads campaign, as they might be missing out on some opportunities. Therefore, there are keywords present in the dataset from Google Ads which are not present in the dataset which is used to train the model on, and vice versa.

Figure 4.5: boxplots of average conversions per intent group

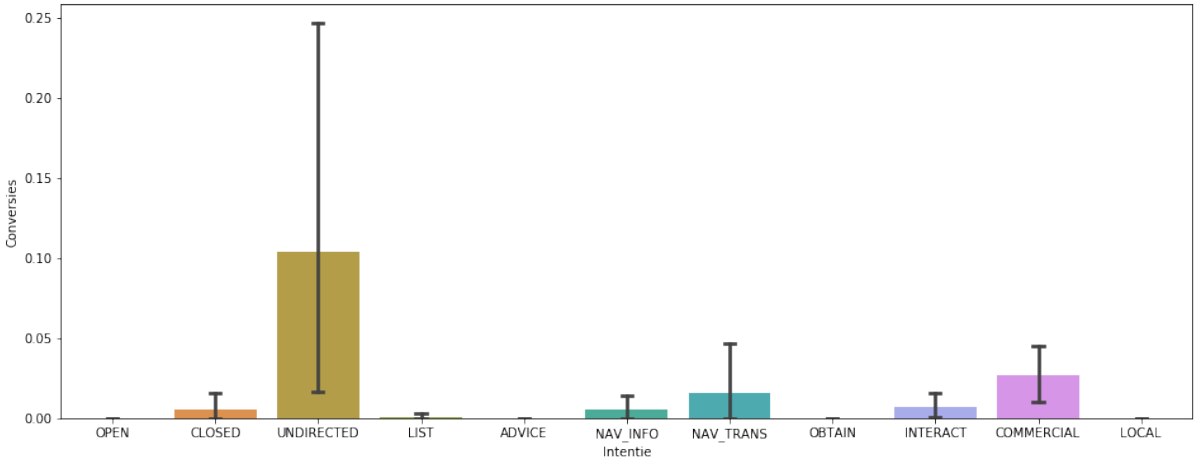
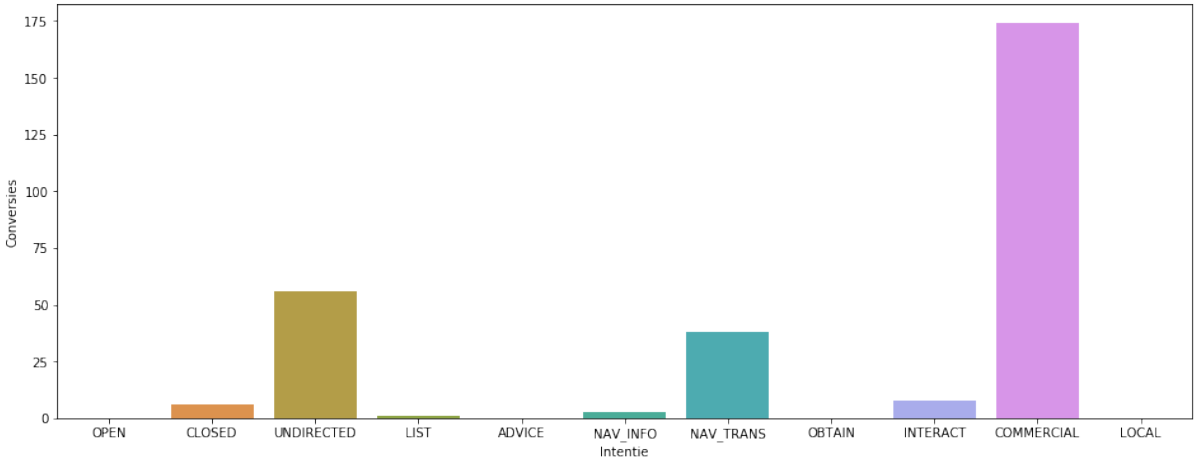


Figure 4.6: sum of conversions per intent group



4.2.2.2 Validation of Navigational keywords

In order to validate the keywords which have been labeled as Navigational ('Navigational to Informational' and 'Navigational to Transactional'), a check will be performed using data from Google Search Console (GSC).

Using GSC, a website owner can track the keywords for which the site ranks in Google, what the daily average position was, the number of clicks, and the number of views, among other things.

In the dataset, there are a total of 2,733 keywords labeled as Navigational. However, because access to GSC is limited to site owners and those with whom it is shared, not all Navigational keywords can be validated. For instance, Wikipedia related keywords are also included in the dataset (e.g. 'Apple iPhone Wikipedia') and labeled as Navigational to Informational. Other Navigational related keywords include for instance competitors in the industry.

Furthermore, GSC is limited to a certain number of keywords which can be shown (1,000). This is based on the amount of clicks each keyword receives. As a result, many long-tail keywords will not be given because they have too little views and clicks.

For the keywords labeled as Navigational which are present in the GSC dataset, the average CTR is 75%. The average CTR of the entire GSC dataset is 11%. This average is above the 70% cut-off mark (as mentioned in section 3.5). There are a few keywords which are below this cut-off mark. These include:

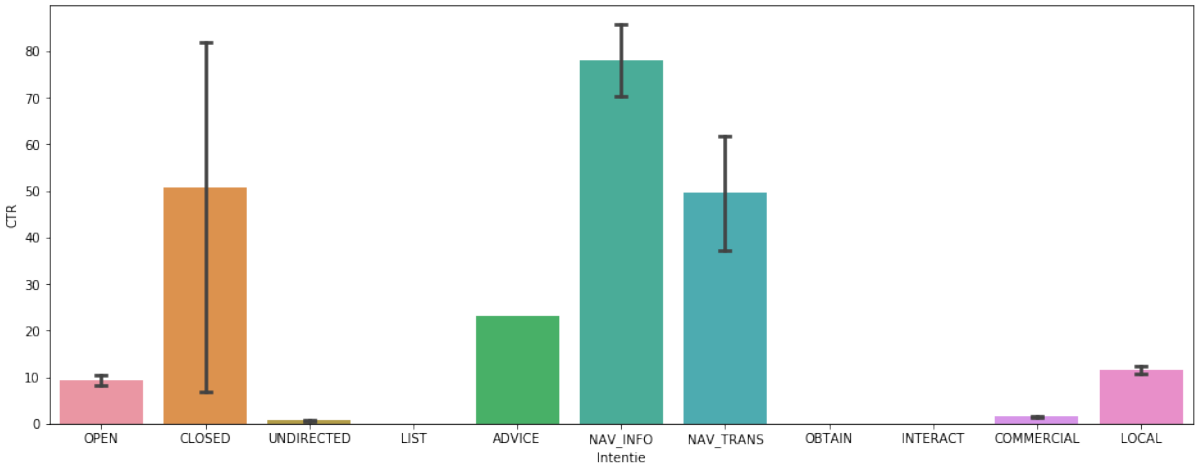
- [COMPANY] sim only (22%)
- [COMPANY] klantenservice (69%)
- [COMPANY] contact (53%)
- [COMPANY] live chat (61%)

This is due to several reasons. For instance, for keywords such as '[COMPANY] sim only' there are many more advertisers present in the SERP, including the company itself. These clicks are not accounted for in the GSC data.

Furthermore, Google gives an overview of possibilities to contact [COMPANY] in the SERP, by showing a phone number or location of the business. Lastly, there is also a business website, for which the GSC data is collected on a separate domain.

To visualize the fact that the intent group 'Navigational' has the highest CTR, Figure 4.7 gives an overview of the CTR per intent group. It is worth noting that the intent group 'Closed' also has a high CTR. This is due to keywords related to '[COMPANY] dekking'. These have been labeled as 'Closed', as there is only one answer (or map overview) for the search intent. However, due to the fact that it does include the company name, it can be seen as a 'Navigational' query as well, which can explain the high CTR. As explained in the theoretical discussion, it is possible for keywords to have multiple intents.

Figure 4.7: boxplots of CTR per intent group



In conclusion, it seems that the intent labels have been classified correctly by the manual labelers. The labelers have a Cohen's Kappa suggesting an almost perfect agreement between the labelers. Furthermore, the user intent seems to be in line with the labeling, as the click through rate is the highest for the navigational labeled keywords and the conversions are the highest for the transactional labeled keywords.

4.3 Model result

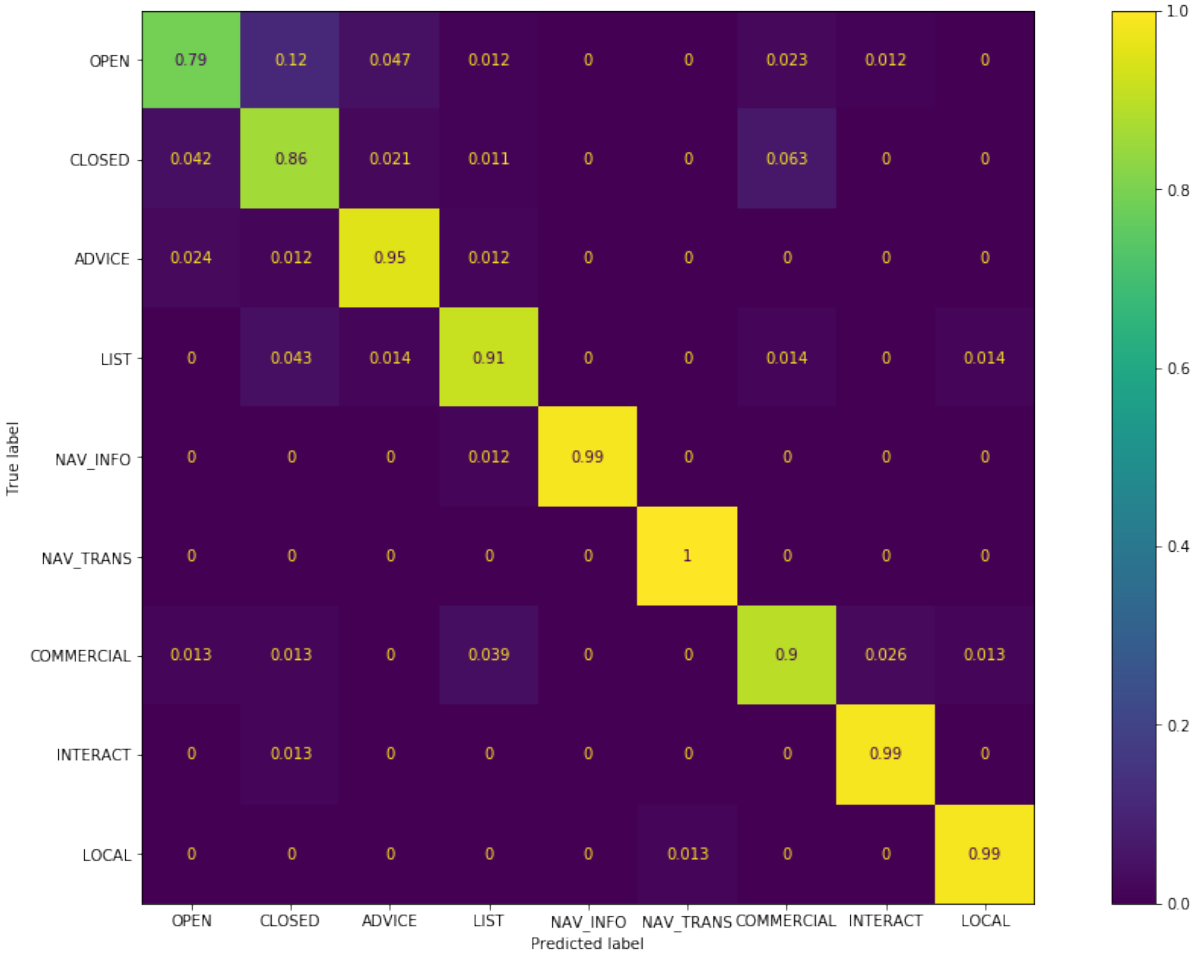
The model has been trained with BERTje using the dataset described earlier. It has been limited to 3,609 keywords, equally divided over the different intent labels.

Results of this model, after 8 epochs, are shown in Figure 4.8 in a confusion matrix. These are results on data the model has not been trained on, but the test set.

This model has an accuracy of 0.984 on the training set and 0.928 on the test set.

As can be seen in the confusion matrix from Figure 4.8, the model has most trouble with deciding between ‘Open’ and ‘Closed’ keywords. These categories have an accuracy of respectively 0.79 and 0.86.

Figure 4.8: Results model on test set



To try and increase the accuracy of the model even further, data augmentation has been added. Data augmentation has been described as ‘augmenting the observed data to make it easier to analyze’ (Tanner & Wong, 1987). In a machine learning context, it is often used to increase the amount of data that can be worked with and to uneven class balances (Mikołajczyk & Grochowski, 2018).

For instance, in image classification, this works by flipping the image, mirror the image or, change shades in an image. This way, the diversity of the training set will be increased with realistic examples which should help and improve the model performance.

Data augmentation with NLP is a bit different than with an image, as it is not feasible to rotate words. Although it is possible to change the position of words within a sentence, this could affect the semantics of the sentence.

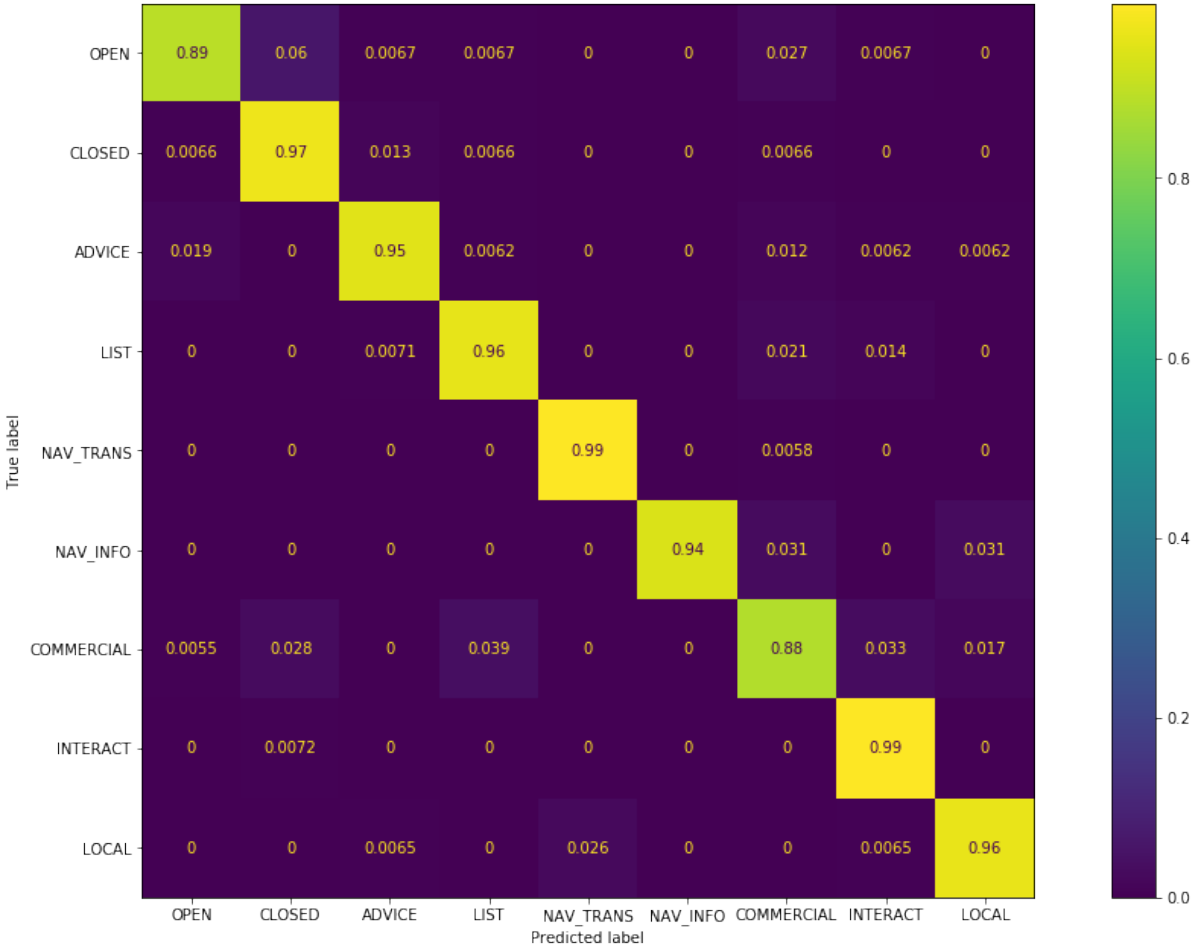
Instead, in an NLP context, synonyms can be used for data augmentation (Kobayashi, 2018). For this, WordNet has been used. This is a lexical database of semantic relationships between words from over 200 languages. Using this database, the total sample size of the

dataset has been increased to 6,750 keywords, equally divided over the intent groups. This is an increase of 87% in data. The newly added data has only been used to train the model on, the testing of the model is done with only original data from the dataset, without anything augmented data, resulting in a test set on actual user queries.

This resulted in an accuracy of 0.964 for the training and 0.947 for the test set after 8 epochs. When only focusing on the four higher level categories of intent, the model scores a test accuracy of 0.963. The confusion matrix of the test set is shown in Figure 4.9.

Significant progression has been made in the 'Open' and 'Closed' intent groups, of which the accuracy increased respectively 0.1 and 0.11. Using the hypothesis testing from Johnson, Miller & Freund (2000), there is also a significant difference between the two test sets (before and after augmentation) ($Z = 2.04$, $p < .05$) with the result after data augmentation being the better one. Furthermore, the difference between train and test set accuracy is smaller, implying overfitting has been reduced (Yu et al., 2017).

Figure 4.9: Results model on test set after data augmentation



The model has run on 8 epochs. As can be seen in Figure 4.10 and Figure 4.11 the accuracy and loss of the model reach a plateau after these epochs. There is not optimal value of the number of epochs, but to avoid overfitting, the loss and accuracy are monitored (Yu et al., 2017). As overfitting is a central problem in machine learning using labeled training data, it needs to be monitored to be avoided (Dietterich, 1995).

Figure 4.10: Accuracy as a function of epoch

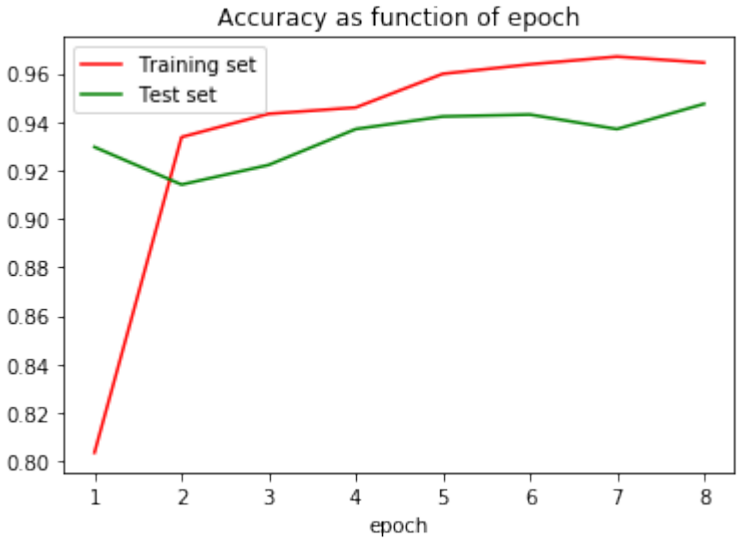


Figure 4.11: Loss as a function of epoch

5. Discussion

In this chapter, the main conclusions which can be drawn from the research will be discussed. This will include an answer to the sub questions and main research question as stated in section 1.3. Furthermore, both the theoretical and managerial implications are indicated. The chapter concludes by discussing the limitations of the research and potential future research directions.

5.1 Conclusion

The main goal of this research was to find a way to automatically classify search queries into pre-existing taxonomies based on search intent, and to do so with higher accuracy than previous research. These taxonomies have been developed by Broder (2002), Rose and Levinson (2004), and Jansen et al. (2008).

First, search intent has been explained as the information seeking behavior of searchers, specifically in search engines. Moreover, it has been shown that an understanding and implementation of search intent can help with obtaining better results from a business perspective (e.g. higher CTR, higher conversion) from search engine users. Finally, the developments in natural language processing have been discussed, concluding with the fact that recent developments with regard to transformer-based language models outperform other models.

With BERT, a language model developed by Google, it is possible to make use of such a transformer-based language model. This model is fine-tuned to the specific task of classification on search intent.

Based on the results of this study, it can be concluded that the model used in this study is able to achieve higher accuracy than the models of previous work on search intent classification (as discussed in section 2.4.1). This is likely due to the use of a transformer-based model which allows for much larger models because of increased parallel computing (Vaswani et al., 2017), which led to large pre-trained models such as BERT (Devlin et al., 2018).

Furthermore, it has been shown that data augmentation using synonyms is a good way to increase the training size of the dataset, which resulted in an increase in the accuracy of the model in this kind of NLP problem. This means that data augmentation is not only helpful for image classification tasks (Mikołajczyk & Grochowski, 2018), but for search intent classifications tasks as well. To my knowledge, this is the first application of data augmentation in research on intent classification.

5.2 Theoretical implications

From a theoretical perspective, many studies in the search intent classification domain have been trying to improve the accuracy scores of their own model (e.g. Mendoza & Zamora, 2009; Tsukuda et al., 2013; Figueroa, 2015). One goal of this research was to achieve an

improved accuracy, which has been accomplished. In NLP, almost every year new high scores on the benchmarks are set (Zhou et al., 2020). By increasing the score for a specific task (search intent classification), this research shows that NLP algorithms not only become better in benchmark tasks, but also on a task such as search intent classification.

Furthermore, in, for instance, the models of Jansen et al. (2008), Tsukuda et al. (2013), Figueroa (2015), Qiu et al. (2018), and Zhang et al. (2019), the researchers have built and trained their own models. It could be argued that fine-tuning large-scale pre-trained models (such as BERT) work better in this search intent classification task than the task specific models from the past. This fine-tuning (and subsequent transfer learning) of pre-trained models is becoming more prevalent in NLP (Sanh, Debut, Chaumond & Wolf, 2019), and, concluding from this research, has good application in search intent classification.

Moreover, this research is the first one to check whether the intent labeling of the keywords corresponds with actual behavior shown by searchers. This is important, because there is a difference between what people say and what people do, known as the value-action or intention-behavior gap (Blake, 1999). In previous research, Broder (2002) asked the searchers in what category their intention was, but did not measure their actual behavior. Other researchers built onwards on the ideas put forward by Broder (2002), but did not try to see if the intention-behavior gap was present (e.g. Rose and Levinson, 2004 & Jansen et al., 2008). This research did not find discrepancy between intent and behavior for the labeled keywords. This might be due to the fact that web search is already implicitly driven by high-level goal orientation of users (Strohmaier et al., 2007).

In addition, research with regards to SEO has been done to understand which factors influence the ranking in a search engine (e.g. Furnell & Evans, 2007; Berman & Katona, 2013; Zhang & Cabage, 2017). While it has already been concluded by Schultz (2019) that search intent has an impact on SEA effort, no research has been done one the combination between search intent and SEO. This research (although not a direct result from the results section) shows that having a focus on search intent seems to help with achieving SEO goals better. However, more research on this topic is required.

Lastly, this research has shown again that informational intent queries make up the most part of the search queries, also in telecom. The results from this research show about 50% of keywords are informational (over four classes). Earlier research from Broder (2002), Rose and Levinson (2004), and Jansen et al. (2008) show informational queries in 62-80% of instances (over three classes).

5.3 Managerial implications

The findings in this research show that it is possible to predict search intent using the final fine-tuned BERT model described in section 4.3 with high accuracy. Therefore, the practical problem of not such a search intent model existing has been solved for the telecom market. Using this model, it is possible to show how the telecom market is divided on search intent (as shown in Table 4.1 and 4.2, but perhaps more keywords can be added). Combining this with the current keywords a client of Greenhouse in this telecom market already ranks for, gaps can be discovered in which search intent the client is not present but might want to be

(e.g. on transactional keywords such as ‘buy iPhone 12’ or informational keywords such as ‘how to reset my iPhone’).

Furthermore, combined with search volume over time, the impact of seasonality can be shown to the clients. Perhaps there are interesting fluctuations. For instance, when the new iPhone is launched, more search volume is on the transactional keywords (as people want to buy the new iPhone). Or perhaps more informational keywords are used during the holidays, as people are looking for information regarding mobile phones on vacation. This could help clients in their marketing efforts, as they will be able to respond to those changes in search intent with dedicated content.

Moreover, the paid search campaigns in the telecom market can be examined to see if the search intent of the keywords which are used are in line with the campaign objective. If for instance many informational keywords are used in the targeted ad keywords, while the desired objective is to generate more conversions, the ad campaign needs to be arranged differently. Next to that, new keywords for a campaign might be discovered, by letting the model predict search intent on a dataset of unseen telecom-related keywords. If some of those keywords correspond with the campaign objective (e.g. transactional and more conversions, or informational and increased brand awareness), newly discovered keywords can be added (in bulk) to the campaign.

This research has only been conducted in the telecom market. Whether this model also works great for other markets, must be tested first. This should be fairly easy, as the model only needs new input for training. Then other markets outside telecom (and thus other clients of Greenhouse) can be examined as well.

As this model is capable of classifying with high accuracy, a next step for Greenhouse could be to explore possibilities to exploit this new tool and build it in a new proposition. Clients are looking for more insights in search data, and this model just might help them. For instance, based on this study, Greenhouse is able to show an overview of how the keywords of the market the client is present in (e.g. telecom) are divided in search intent groups. This can be combined with keywords the client does not rank for (in organic search and paid search), to help create a plan to better be present in an intent category which is important for the client. For instance, if the client has little presence with transactional keywords but wants more conversions and the market for telecom has much room for transactional keywords, a plan can be made to rank better for transactional keywords (e.g. paid search, improving organic search, creating new content or partner with publishers). A model like this is also very scalable, as the model can be used at the same time for multiple clients.

Furthermore, from this research a substantial number of keywords have been discovered, which have not been used in the Google Ads campaign. Including keywords which are labeled as having a transactional intent, e.g. people are looking to buy something. These keywords can be added to the existing Google Ads campaigns to drive extra conversions for the client.

5.4 Limitations & future research

Just like every research suffers from some limitations, this one does too. First of all, the total amount of keywords used to train the model (6,750) is rather limited. Especially when compared to the 1.5 million keywords dataset from Jansen et al. (2008). It is interesting to see that this model does achieve improved accuracy (compared to other research) even on a smaller dataset. This can be one of the advantages of a model like BERT, as it works well on classification tasks with small datasets (Sun, Qiu, Xu & Huang, 2019). It might be interesting for future research to see if a larger dataset has the ability to improve the accuracy of the model.

Secondly, there might be a problem due to the fact that search queries can have multiple intents, and they are different from person to person. This was part of the disagreement between the labelers. When two similar keywords are labeled differently (e.g. informational and transactional), the model might have difficulty learning from those examples.

The first and second limitation mentioned can possibly be researched by looking into a more unsupervised way of training for fine-tuning large-scale models such as BERT. This can be done using unlabeled data, just like BERT has been pre-trained on, which will allow for much larger datasets (Siddhant, Goyal & Metallinou, 2019). A way of unsupervised learning is clustering instead of classification. Current research on search intent uses classification, as this is easier for humans to interpret (Beeferman & Berger, 2000). However, search engines are the ones who perform the actual information retrieval, and should have no issue to interpret clustered data.

Thirdly, as the model has been trained in one specific market (telecom), it is more difficult to generalize towards other markets, where the accuracy of the model might be less. Future research can, perhaps with the help of a larger unsupervised dataset, reveal what the accuracy of a model like BERT is in different markets and different markets combined in one single model (e.g. one model trained on keywords from telecom, car and other markets combined).

Finally, validation was done for transactional and navigational keywords, but not for informational or local keywords. Future research can look into this by trying to figure out clues for informational or local keywords, for instance with search engine result features such as a map, featured snippet, knowledge card/panel, or related questions.

In addition to the future research related to the limitations mentioned above, an interesting direction for future research could be to further pre-train a model instead of only fine-tuning a pre-trained model. For example, Sun et al. (2019) found that within-task and in-domain further pre-training would significantly boost the performance of the classification model.

Lastly, there seems to be no research done on the connection between search intent and SEO. As mentioned in section 2.3.2, there does seem to be enough reason that there is a connection between SEO and search intent. Schultz (2019) already researched the connection between SEA and search intent, but none has been done with regards to search intent and SEO. Future researchers are encouraged to take up this issue.

References

Agius, A. (2019, June 12). A quick and easy guide to understanding search intent for SEO. Retrieved from https://searchengineland.com/a-quick-and-easy-guide-to-understanding-search-intent-for-seo-317841

Allein, L., Leeuwenberg, A., & Moens, M. F. (2020). Binary and multitask classification model for Dutch anaphora resolution: Die/Dat prediction. *arXiv preprint arXiv:2001.02943*.

Alphabet. (2020, February 3). Alphabet Announces Fourth Quarter and Fiscal Year 2019 Results. Retrieved from https://abc.xyz/investor/static/pdf/2019Q4_alphabet_earnings_release.pdf

Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2014, November). Sentiment analysis: towards a tool for analysing real-time students feedback. In 2014 IEEE 26th international conference on tools with artificial intelligence (pp. 419-423). IEEE.

Ashkan, A., & Clarke, C. L. (2009, July). Term-based commercial intent analysis. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (pp. 800-801).

Ashkan, A., & Clarke, C. L. (2013). Impact of query intent and search context on clickthrough behavior in sponsored search. *Knowledge and information systems, 34*(2), 425-452.

Aswani, R., Kar, A. K., Ilavarasan, P. V., & Dwivedi, Y. K. (2018). Search engine marketing is not all gold: Insights from Twitter and SEOClerks. *International Journal of Information Management*, *38*(1), 107-116.

Athukorala, K., Głowacka, D., Jacucci, G., Oulasvirta, A., & Vreeken, J. (2016). Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. Journal of the Association for Information Science and Technology, 67(11), 2635-2651.

Baeza-Yates, R., Calderón-Benavides, L., & González-Caro, C. (2006, October). The intention behind web queries. In International symposium on string processing and information retrieval (pp. 98-109). Springer, Berlin, Heidelberg.

Batista, H. (2019, June 24). Automated Intent Classification Using Deep Learning. Retrieved from https://www.searchenginejournal.com/automated-intent-classification-using-deep-learning/311309/

Beeferman, D., & Berger, A. (2000, August). Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 407-416).

Belkin, N. J. (1993). Interaction with texts: Information retrieval as information seeking behavior. Information retrieval, 93, 55-66.

Belkin, N.J., Seeger, T., & Wersig, G. (1983) Distributed expert problem treatment as a model for information system analysis and design. Journal of Information Science 5,5, 152-167.

Berman, R., & Katona, Z. (2013). The role of search engine optimization in search marketing. *Marketing Science*, *32*(4), 644-651.

Blake, J. (1999). Overcoming the 'value-action gap' in environmental policy: Tensions between national policy and local experience. *Local environment*, *4*(3), 257-278.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine.

Broder, A. (2002, September). A taxonomy of web search. In ACM Sigir forum (Vol. 36, No. 2, pp. 3-10). New York, NY, USA: ACM.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005, August). Learning to rank using gradient descent. In Proceedings of the 22nd international conference on Machine learning (pp. 89-96).

Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg?. IEEE Transactions on Dependable and Secure Computing, 9(6), 811-824.

Clement, J. (2020a, February 3). Global regional internet penetration rate 2020. Retrieved from https://www.statista.com/statistics/269329/penetration-rate-of-the-internet-by-region/

Clement, J. (2020b, February 5). Annual revenue of Google from 2002 to 2019. Retrieved from https://www.statista.com/statistics/266206/googles-annual-global-revenue/

Custer, R. L., Scarcella, J. A., & Stewart, B. R. (1999). The modified Delphi technique: A rotational modification. *Journal of Vocational and Technical Education, 15* (2), 1-10.

Dai, H., Zhao, L., Nie, Z., Wen, J. R., Wang, L., & Li, Y. (2006, May). Detecting online commercial intention (OCI). In Proceedings of the 15th international conference on World Wide Web (pp. 829-837).

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). BERTje: A Dutch BERT Model. *arXiv preprint arXiv:1912.09582*.

Dean, B. (2019a, August 27). We Analyzed 5 Million Google Search Results. Here's What We Learned About Organic CTR. Retrieved from https://backlinko.com/google-ctr-stats

Dean, B. (2019b, September 5). Search Intent and SEO: A Complete Guide. Retrieved from https://backlinko.com/hub/seo/search-intent

Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model. *arXiv preprint arXiv:2001.06286.*

Deloitte. (2020). How the cookie crumbled: Marketing in a cookie-less world. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consultancy/deloitte-uk-cookie-less-marketing.pdf

Dervin, B. (1983) An overview of sense-making: concepts, methods and results to date. Paper presented at the International Communication Association Annual Meeting, Dallas TX, May 1983.

Desjardins, J. (2018, April 24). How Google retains more than 90% of market share. Retrieved from https://www.businessinsider.com/how-google-retains-more-than-90-of-market-share-2018-4

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, *27*(3), 326-327.

Figueroa, A. (2015). Exploring effective features for recognizing the user intent behind web queries. Computers in Industry, 68, 162-169.

Forbes. (2019). The World's Most Valuable Brands. Retrieved from https://www.forbes.com/powerful-brands/list/

Fortune. (2020, January 30). Fortune Global 500. Retrieved from https://fortune.com/global500/2019/search/?profitable=true&profits=desc

Friedman, W. (2018, December 18). Television News Daily: Global TV Ad Spend Up 1% In 2018. Retrieved from https://www.mediapost.com/publications/article/329489/global-tv-ad-spend-up-1-in-2018.html

Furnell, S., & Evans, M. P. (2007). Analysing Google rankings through search engine optimization data. *Internet research*.

Gers, F. A., & Schmidhuber, E. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. IEEE Transactions on Neural Networks, 12(6), 1333-1340.

Ghose, A., & Yang, S. (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. Management science, 55(10), 1605-1622.

GitHub. (2018, 31 oktober). GitHub - google-research/bert: TensorFlow code and pre-trained models for BERT. Retrieved from https://github.com/google-research/bert

Google. (2019, December 5). Search Quality Evaluator Guidelines. Retrieved from https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

Google. (2020a). Ad Rank - Google Ads Help. Retrieved from https://support.google.com/google-ads/answer/1752122?hl=en

Google. (2020b). Get More Customers with Pay Per Click (PPC) Search Ads - Google Ads. Retrieved from https://ads.google.com/intl/en/home/campaigns/search-ads/

Google. (2020c). Google's Search Algorithm and Ranking System - Google Search. Retrieved from https://www.google.com/search/howsearchworks/algorithms/

Google. (2020d). How Google's Algorithm is Focused on Its Users - Google Search. Retrieved from https://www.google.com/search/howsearchworks/mission/users/

Hardwick, J. (2019, December 5). Local SEO: A Simple (But Complete) Guide. Retrieved 2020, from https://ahrefs.com/blog/local-seo/

Hardy, H., & Cheah, Y. N. (2013). Question classification using extreme learning machine on semantic features. Journal of ICT Research and Applications, 7(1), 36-58.

Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

Hashemi, H. B., Asiaee, A., & Kraft, R. (2016, February). Query intent detection using convolutional neural networks. In *International Conference on Web Search and Data Mining, Workshop on Query Understanding*.

Hernández, I., Gupta, P., Rosso, P., & Rocha, M. (2012). A simple model for classifying web queries by user intent. In Proc. 2nd Spanish Conf. Information Retrieval (pp. 235-240).

Hillard, D., Schroedl, S., Manavoglu, E., Raghavan, H., & Leggetter, C. (2010, February). Improving ad relevance in sponsored search. In Proceedings of the third ACM international conference on Web search and data mining (pp. 361-370).

Højgaard, C., Sejr, J., & Cheong, Y. G. (2016). Query categorization from web search logs using machine learning algorithms. International Journal of Database Theory and Application, 9(9), 139-148.

Hsu, C. C., & Sandford, B. A. (2007). The Delphi technique: making sense of consensus. *Practical Assessment, Research, and Evaluation*, *12*(1), 10.

Hu, J., Wang, G., Lochovsky, F., Sun, J. T., & Chen, Z. (2009, April). Understanding user's query intent with wikipedia. In *Proceedings of the 18th international conference on World wide web* (pp. 471-480).

Huang, J., Wang, H., Zhang, W., & Liu, T. (2020). Multi-Task Learning for Entity Recommendation and Document Ranking in Web Search. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(5), 1-24.

Internet Live Stats. (2020). Google Search Statistics - Internet Live Stats. Retrieved from https://www.internetlivestats.com/google-search-statistics/#rate

Jacobs, J. M. (1996). *Essential assessment criteria for physical education teacher education programs: A Delphi study*. Unpublished doctoral dissertation, West Virginia University, Morgantown.

Jansen, B. J., & Resnick, M. (2005, June). Examining searcher perceptions of and interactions with sponsored results. In Workshop on Sponsored Search Auctions.

Jansen, B. J., Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. Information Processing & Management, 44(3), 1251-1266.

Jansen, B. J., Liu, Z., & Simon, Z. (2013). The effect of ad rank on the performance of keyword advertising campaigns. Journal of the american society for Information science and technology, 64(10), 2115-2132.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kang, I. H. (2005, October). Transactional query identification in web search. In Asia Information Retrieval Symposium (pp. 221-232). Springer, Berlin, Heidelberg.

Kathuria, A., Jansen, B. J., Hafernik, C., & Spink, A. (2010). Classifying the user intent of web queries using k-means clustering. Internet Research.

Killoran, J. B. (2013). How to use search engine optimization techniques to increase website visibility. *IEEE Transactions on professional communication*, *56*(1), 50-66.

Kim, L. (2016, April 18). Does Organic CTR Impact SEO Rankings? [New Data]. Retrieved October 21, 2020, from https://moz.com/blog/does-organic-ctr-impact-seo-rankings-new-data

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Lee, U., Liu, Z., & Cho, J. (2005, May). Automatic identification of user goals in web search. In Proceedings of the 14th international conference on World Wide Web (pp. 391-400).

Lewandowski, D., Drechsler, J., & Von Mach, S. (2012). Deriving query intents from web search engine queries. Journal of the American Society for Information Science and Technology, 63(9), 1773-1788.

Li, Y., Zheng, Z., & Dai, H. (2005). KDD CUP-2005 report: Facing a great challenge. ACM SIGKDD Explorations Newsletter, 7(2), 91-99.

Linstone, H. A., & Turoff, M. (Eds.). (1975). *The delphi method*(pp. 3-12). Reading, MA: Addison-Wesley.

Liu, D., Chen, J., & Whinston, A. B. (2010). Ex ante information and the design of keyword auctions. Information Systems Research, 21(1), 133-153.

Ludwig, B. (1997). Predicting the future: Have you considered using the Delphi methodology. *Journal of extension*, *35*(5), 1-4.

Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, *77*(1), 87-103.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1 (pp. 142-150). Association for Computational Linguistics.

MAGNA. (2018, September 20). Advertising Forecasts (Fall Update – Executive Summary). Retrieved from https://magnaglobal.com/magna-advertising-forecasts-fall-update-executive-summary/

McCue, T. J. (2018, July 30). SEO Industry Approaching $80 Billion But All You Want Is More Web Traffic. Retrieved from https://www.forbes.com/sites/tjmccue/2018/07/30/seo-industry-approaching-80-billion-but-all-you-want-is-more-web-traffic/#461749957337

McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Mendoza, M., & Zamora, J. (2009, August). Identifying the intent of a user query using support vector machines. In International symposium on string processing and information retrieval (pp. 131-142). Springer, Berlin, Heidelberg.

Mikołajczyk, A., & Grochowski, M. (2018, May). Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)* (pp. 117-122). IEEE.

Miniwatts Marketing Group. (2019). World Internet Users Statistics and 2019 World Population Stats. Retrieved from https://www.internetworldstats.com/stats.htm

Mohammad, S. M. (2020). Nlp scholar: An interactive visual explorer for natural language processing literature. *arXiv preprint arXiv:2006.01131*.

Mohasseb, A., Bader-El-Den, M., Liu, H., & Cocea, M. (2017, July). Domain specific syntax based approach for text classification in machine learning context. In 2017 international conference on machine learning and cybernetics (ICMLC) (Vol. 2, pp. 658-663). IEEE.

Mohasseb, A., El-Sayed, M., & Mahar, K. (2014, April). Automated identification of web queries using search type patterns. In *WEBIST (2)* (pp. 295-304).

Morrison, J. B., Pirolli, P., & Card, S. K. (2001, March). A taxonomic analysis of what World Wide Web activities significantly impact people's decisions and actions. In CHI'01 extended abstracts on Human factors in computing systems (pp. 163-164).

Nabout, N. A., & Skiera, B. (2012). Return on quality improvements in search engine marketing. Journal of Interactive Marketing, 26(3), 141-154.

Navarro-Prieto, R., Scaife, M., & Rogers, Y. (1999, July). Cognitive strategies in web searching. In Proceedings of the 5th Conference on Human Factors & the Web (pp. 43-56).

Nayak, P. (2019, October 25). Understanding searches better than ever before. Retrieved from https://www.blog.google/products/search/search-language-understanding-bert/

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision support systems, 50(3), 559-569.

O'Day, V. L., & Jeffries, R. (1993, May). Orienteering in an information landscape: how information seekers get from here to there. In Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems (pp. 438-445).

Olbrich, R., & Schultz, C. D. (2014). Multichannel advertising: Does print advertising affect search engine advertising?. European Journal of Marketing.

Paalman, J., Mullick, S., Zervanou, K., & Zhang, Y. (2019, September). Term based semantic clusters for very short text classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 878-887).

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Stanford InfoLab.

Patel, N. (2020, January 23). Type No More: How Voice Search is Going to Impact the SEO Landscape. Retrieved from https://neilpatel.com/blog/type-no-more-how-voice-search-is-going-to-impact-the-seo-landscape/

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT?. arXiv preprint arXiv:1906.01502.

PwC. (2019, July 1). Global Top 100 companies by market capitalisation. Retrieved from https://www.pwc.com/gx/en/audit-services/publications/assets/global-top-100-companies-2019.pdf

Qiu, L., Chen, Y., Jia, H., & Zhang, Z. (2018). Query intent recognition based on multi-class features. IEEE Access, 6, 52195-52204.

Ravuri, S., & Stolcke, A. (2015). Recurrent neural network and LSTM models for lexical utterance classification. In Sixteenth Annual Conference of the International Speech Communication Association.

Rose, D. E., & Levinson, D. (2004, May). Understanding user goals in web search. In Proceedings of the 13th international conference on World Wide Web (pp. 13-19).

Rozanski, H. D., Bollman, G., & Lipman, M. (2001). Seize the occasion! The seven-segment system for online marketing. Strategy and Business, 42-53.

Sak, H., Senior, A. W., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schultz, C. D. (2019). Informational, transactional, and navigational need of information: relevance of search intention in search engine advertising. Information Retrieval Journal, 1-19.

Schutz, A., & Luckmann, T. (1973). The structures of the life-world (Vol. 1). northwestern university press.

Searchmetrics. (2020). User/Search Intent – SEO Glossary. Retrieved from https://www.searchmetrics.com/glossary/user-intent/

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

Sedigh, A. K., & Roudaki, M. (2003). Identification of the dynamics of the google's ranking algorithm. In *13th IFAC Symposium On System Identification*.

Sellen, A. J., Murphy, R., & Shaw, K. L. (2002, April). How knowledge workers use the web. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 227-234).

Shneiderman, B., Byrd, D., & Croft, W. B. (1997). Clarifying search: A user-interface framework for text searches.

Shoebottom, P. (2020). Polysemy. Retrieved from http://esl.fis.edu/teachers/support/vocabPoly.htm

Siddhant, A., Goyal, A., & Metallinou, A. (2019, July). Unsupervised transfer learning for spoken language understanding in intelligent agents. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 4959-4966).

SimilarWeb. (2020). Website Ranking: Top Websites Rank In The World - SimilarWeb. Retrieved from https://www.similarweb.com/top-websites

Strohmaier, M., Lux, M., Granitzer, M., Scheir, P., Liaskos, S., & Yu, E. (2007, December). How do users express goals on the web?-an exploration of intentional structures in web search. In *International Conference on Web Information Systems Engineering* (pp. 67-78). Springer, Berlin, Heidelberg.

Slone, D. J. (2003). Internet search approaches: The influence of age, search goals, and experience. *Library & Information Science Research*, *25*(4), 403-418.

Song, R., Luo, Z., Wen, J. R., Yu, Y., & Hon, H. W. (2007, May). Identifying ambiguous queries in web search. In Proceedings of the 16th international conference on World Wide Web (pp. 1169-1170).

Soulo, T. (2019, October 16). Long-Tail Keywords: The ‘Secret’ to Getting TONS of Search Traffic. Retrieved from https://ahrefs.com/blog/long-tail-keywords/

StatCounter. (2020, January 1). Search Engine Market Share Worldwide. Retrieved from https://gs.statcounter.com/search-engine-market-share

Statista. (2020). Search Advertising - worldwide | Statista Market Forecast. Retrieved from https://www.statista.com/outlook/219/100/search-advertising/worldwide

Strzelecki, A., & Rutecka, P. (2020). Featured Snippets Results in Google Web Search: An Exploratory Study. In *Marketing and Smart Technologies* (pp. 9-18). Springer, Singapore.

Sullivan, D. (2016, May 24). Google now handles at least 2 trillion searches per year. Retrieved from https://searchengineland.com/google-now-handles-2-999-trillion-searches-per-year-250247

Sullivan, R. (2020, February 10). Why all marketing should be performance marketing. Retrieved from https://www.thinkwithgoogle.com/intl/en-154/insights-inspiration/research-data/full-funnel-marketing-performance/

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, October). How to fine-tune bert for text classification?. In *China National Conference on Chinese Computational Linguistics* (pp. 194-206). Springer, Cham.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528-540.

Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. Journalism quarterly, 30(4), 415-433.

Tinsley, H. E., & Weiss, D. J. (2000). Interrater reliability and agreement. In *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95-124). Academic Press.

Tsukuda, K., Sakai, T., Dou, Z., & Tanaka, K. (2013, December). Estimating intent types for search result diversification. In Asia Information Retrieval Symposium (pp. 25-37). Springer, Berlin, Heidelberg.

Wang, J., Yu, L., Zhang, W., Gong, Y., Xu, Y., Wang, B., ... & Zhang, D. (2017, August). Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 515-524).

Wen, J. R., Nie, J. Y., & Zhang, H. J. (2002). Query clustering using user logs. *ACM Transactions on Information Systems*, *20*(1), 59-81.

van der Burgh, B., & Verberne, S. (2019). The merits of Universal Language Model Fine-tuning for Small Datasets--a case with Dutch book reviews. *arXiv preprint arXiv:1910.00896*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

Wersig, G. (1979) The problematic situation as a basic concept of information science in the framework of the social sciences - a reply to N. Belkin. In: New Trends in Informatics and its Terminology (FID 568). Moscow, VINITI: 48-57.

Yang, K., Cai, Y., Huang, D., Li, J., Zhou, Z., & Lei, X. (2017, February). An effective hybrid model for opinion mining and sentiment analysis. In 2017 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 465-466). IEEE.

Yoast. (2018, January 15). What is search intent? Retrieved from https://yoast.com/search-intent/

Yu, N., Yu, Z., Gu, F., Li, T., Tian, X., & Pan, Y. (2017). Deep Learning in Genomic and Medical Image Data Analysis: Challenges and Approaches. *JIPS*, *13*(2), 204-214.

Zhang, S., & Cabage, N. (2017). Search engine optimization: Comparison of link building and social sharing. *Journal of Computer Information Systems*, *57*(2), 148-159.

Zhang, H., Song, X., Xiong, C., Rosset, C., Bennett, P. N., Craswell, N., & Tiwary, S. (2019, July). Generic intent representation in web search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 65-74).

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

Zhang, R., Guo, J., Fan, Y., Lan, Y., & Cheng, X. (2020). Query Understanding via Intent Description Generation. *arXiv preprint arXiv:2008.10889*.

Zhang, D., & Lee, W. S. (2003, July). Question classification using support vector machines. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 26-32).

Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in neural NLP: modeling, learning, and reasoning. *Engineering*, *6*(3), 275-290.

Appendices

Appendix A