Predicting User Intent based on Search Queries

I.R. Muller

Layout: typeset by the author using LaTeX.

Predicting User Intent based on Search Queries

I.R. Muller
12177970

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Dr. S. (Siamak) Farshidi
N. (Na) Li
Dr. Z. (Zhiming) Zhao

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

Semester 1, 2022

findable ~> قابل یافتن

interaperable ~> قابل همکاری

accessible ~> در دسترس

reusable ~> قابل استفاده مجدد

Abstract

Understanding and predicting user intent is a challenging task that has a great field of interest as it can be used to improve users' experience by making personal recommendations. This thesis aims to research the approaches of recognizing and predicting the intent of a user when browsing a scientific search engine. Researchers often need data or tools from heterogeneous sources making it difficult to efficiently retrieve the right information from a search engine. The ENVRI community is a community aiming to create a platform in which scientific sources are brought together for more efficient information retrieval and to make data and services findable, accessible, interoperable, and reusable (FAIR). For this thesis, it has been researched how to make predictions based on the users' intent to improve the users' experience while searching for information in such a scientific database. Therefore, two models, the BERT intent recognition model and the Ludwig model, were selected for predicting user intent based on submitted queries. Both models are built to solve one of the main challenges in Natural Language Processing (NLP): the shortage of training data. A common problem in classifying queries is that it differs from traditional text classification mainly caused by the following issues: queries are usually very short, queries can be ambiguous, a query often belongs to multiple categories and a query is less topic-focused than a document. Based on these outcomes, the two selected models are trained to achieve high accuracy for predicting intent on the SNIPS and ATIS datasets. Lastly, both models were tested on a real-life dataset. However, they did not perform as well as on the pre-labelled datasets achieving an accuracy of less than 30%. It is argued that this is mainly caused by the problem of classifying web queries and therefore an approach is discussed to cope with this issue.

ناهمگون

retake

Contents

Chapter 1

Introduction

1.1 User Intent

Understanding and predicting user intent is a challenging task that has a great field of interest as it can be used to improve users' experience by making personal recommendations. This thesis aims to research the approaches of recognizing the intent of a user when browsing a scientific search engine. Researchers often need data or tools from heterogeneous sources making it difficult to efficiently retrieve the right information from a search engine. Virtual Research Environments are cohesive platforms where these scientific sources are brought together for more efficient information retrieval [18]. The ENVRI community is a community that aims to create such a platform in which they aim to make data and services findable, accessible, interoperable, and reusable (FAIR). Scientists are stimulated to contribute with data, models, instruments, algorithms, and discoveries for this understanding. By connecting all these data and services, a large database arises. By recognizing the intent of a user, predictions can be made to improve the users' experience while searching for information in this database. This thesis will study how the users' intent can be recognized based on submitted queries and which models can be used to make relevant predictions based on this intent.

منسجم

Users' experience could be improved by predicting the behaviour and making recommendations about the needs of the user. When navigating the internet a user performs several actions that could provide hints on his/her future activities. Understanding the user intent by analyzing search behaviour can be approached in many ways, such as clicked URLs or submitted queries [3]. The way a user reaches the website, which URLs a user clicks on and the visiting time of the URLs are all hints to what the user is intending to retrieve from the search engine. Agichtein et. al [2] proposed a real-world study modelling the behaviour of web search to predict search result preferences. The model analyzes user activities, such as clicks, scrolls and dwell time to predict user intention during web page navigation. The most straightforward way to recognize intent is the query a user submits. However, query classification based on user intent differs from traditional text classification because of the following issues: web queries are usually very short, many queries are ambiguous, a query often belongs to multiple categories and queries are less topic-focused than documents. [27] [15] Classifying queries is a complicated task that requires a substantial amount of data about the users' behaviour and intent. A common issue in Natural Language Processing (NLP) is the shortage of training data.

قابل توجه

Recognizing intent from a search query requires a large amount of training data, since it is a supervised problem, meaning the data must be labelled in order to train a machine learning model. However, this kind of data is often unavailable. Finding solutions to overcome this issue is one of the main aims of this study. Furthermore, query classification faces many issues as web queries usually are very short, ambiguous and can belong to multiple categories. [27] There have been studies on how to overcome data scarcity for this issue which often results in extending the data by labelling more queries. However, these experiments are time-consuming and not suitable for real-time applications. Thus, this thesis studies how to efficiently classify these queries and what models are suitable for making a prediction based on these classified queries.

lack

1.2 Objectives

This thesis aims to analyze the approaches to recognize the user's intent based on search queries while browsing. When intent is recognized, it will be studied how to predict the intent of a user by evaluating two user intent prediction models. The BERT intent recognition model that is based on the original Transformer model [8] and the Ludwig model which is a framework introduced by Uber for training and testing deep learning [23]. Both models will be tested on two datasets, the SNIPS dataset [9] and the ATIS dataset[14]. Then, this study aims to apply a real-life dataset containing queries extracted from the ENVRI search engine on the models. By applying a real-life dataset the limitations of the models can be shown. Modelling user intent is a supervised technique which is a common problem in recognizing user intent as it needs to be understood what the intent of a user is and what kind of behaviour is applicable to which intent. One of the main challenges will be to find a method to turn unsupervised data into supervised data. The performance of intent recognition has been hindered by data scarcity since it is significant to collect sufficient examples for new intents. Furthermore, two models that make predictions based on this intent will be evaluated on their features for future work. The questions that are aimed to be answered are formulated as follows:

قابل توجه

1. *How can users' intent be recognized based on a search query when browsing search results?*
2. *How can users' intent be predicted when browsing search results?*

   2.1. *What are the existing intent prediction models?*

   2.2. *How to select a good prediction model?*

This thesis aims to contribute a step forward in understanding the users' intent in order to improve users' experience when browsing a search engine. Therefore, two intent prediction models will be tested on their performance on the SNIPS and ATIS datasets. The input of the models is a dataset containing search queries labelled with intent and the desired output is a prediction of the intent based on unseen data. The output of the models will be evaluated and the advantages and disadvantages will be discussed. Also, it will be researched how to recognize the intent in a real-life application by generating potential queries from the ENVRI search engine and thus how to make predictions based on unseen data. Hence, this study motivates future investigations that focus on intent prediction systems.

1.3 Contributions

Firstly, this thesis aims to make a contribution in recognizing and predicting the users' intent in order to improve users' experience when browsing the ENVRI search engine. In order to do this, two models are tested for predicting user intent. Both models are tested on real-life data from the ENVRI search engine. This resulted in the issue of classifying search queries. It is found that recognizing intent is a supervised task, for which a labelled dataset is needed. Therefore, approaches are proposed to overcome this problem. Lastly, a dataset of generated queries from the ENVRI search engine is created, which can be used for future work. However, a method for efficiently labelling this dataset should be applied. This thesis shows the importance of having qualitative data.

کیفی

Chapter 2

Background

This chapter studies earlier research to understand the challenges of recognizing and predicting user intent and to define common terms in this field of research. Also a brief description of the technical background will be discussed in section 2.3 for understanding the implemented models. Finally, in section 2.4 a gap analysis will be made.

2.1 Understanding user intent

This thesis will focus on recognizing intent as well as predicting intent. While predicting intent is done by a machine learning model, capturing intent, or intent recognition, is the task of classifying a written or spoken input based on what the user wants to achieve. Thus, in order to make an intent prediction the intent must first be recognized. Recognizing intent is a challenging task since a goal can be achieved with different words and users do not stay in the same semantic territory. In robotics, intent recognition can be performed using several methods such as video, gestures, eye movements, which affect the information in speech or speech. The recognized intentions are raw input and must be transformed into data structures that are suitable for inference such as action frames. [26] In NLP there are many more difficulties such as ambiguity and out-of-domain utterances. Intent recognition is a supervised technique that requires examples of text alongside their intents and trains it with a machine learning (ML) model. Once the data is labelled the model can be trained.

گفته ها

Intent recognition can be *single-intended* and*multi-intended.* Multi-intent classification aims at identifying multiple user goals in a sentence. Xu et. al [29] try to achieve this by adding class features and hidden variables to identify segments belonging to each intent. They show techniques where it is aimed to exploit shared information across different intent combinations. In their proposal, it is shown that the class-based and the hidden variable-based approaches both lead to a higher classification accuracy than the baseline approach in which the combinations are treated as atomic labels, as well as the binarization-based technique commonly used for multi-label learning. Therefore multi-intent classification could improve the predictions of the user intent.

بهره برداری

Jansen and Booth [17], define user intent as the expression of an affective, cognitive, or situational goal in an interaction with a Web Search Engine. Intent can be recognized in many ways that must be analysed and defined. The most straightforward way for capturing the intent is by analyzing the queries that were used to find certain web pages. There are promising methods that are able to recognize the interests of the user and determine the user's intent. The intent and search behaviour will be different when a user is shopping online than when a user is searching for data in a scientific search engine since the goals differ, nevertheless, it can be argued that similarities can be found. Broder [6] introduces three types of general intents: informational, transactional and navigational intent, called Broder's taxonomy.

معاملاتی

اطلاعاتی

طبقه بندی گسترده تر

relating to the action of moving around a website

1. Informational intent is when a user retrieves information by reading or viewing web pages.
2. Transactional intent is when a user retrieves a resource available on a web page. (E.g. downloading a resource, watching an image or locating a service or product)
3. Navigational intent is when a user aims to reach a certain web page, thus the goal is known.

Pirvu et. al [28] train their model based on these 3 intents. These intent types are useful for classifying general interaction features. When analysing the behaviour it can be classified what the general goal of a user is for an accurate prediction. Although these intents are enough to make a prediction, they provide a global classification and this study will aim to divide the intent into more classes for a more specific prediction.

The behaviour of a user can be modelled in many ways. Modelling behaviour in this case means how to represent the search behaviour of the user. This thesis has chosen to model the intent as a search query since it is an often-used approach. Nevertheless, the behaviour of a user shows its needs and can lead to certain patterns which can be used to train a machine-learning algorithm for making new predictions. [5] Thus besides queries, the user behaviour can be useful for predicting the intent. Kawazu et. al [19] propose a method to classify web user behaviour based on hidden intention, interest, or motivation states. Hidden Markov Models (HMM) are commonly used for modelling web user behaviour. It can deal with clickstream sequential data and latent states of users behind the observed data. They found an approach to classify latent states of users as the following sessions: enthusiasm for the main contents, low motivation, and light users. This could be applied for services with many complicated pages, especially a scientific search engine.

پنهان

A model for classifying user interaction actions by interpreting mouse cursor actions, such as scrolling, movement, text selection while reading on the web is proposed by Deufemia et. al [10]. A *relevance value* indicating how a user found the document useful for his/her search purposes is aimed to retrieve. The Yet Another Ranker (YAR) is implemented, which re-ranks the web pages retrieved by a search engine based on the relevance values computed from the interaction actions of previous visitors. The presented model infers user interests about web page contents from his/her mouse cursor actions, such as scrolling, movement, text selection, and the time spent on the page, while reading web documents. Their results demonstrate that the proposed model is able to predict user feedback with an acceptable level of accuracy.

2.2 Predicting user intent

When the intent is recognized new predictions can be made with a machine learning model. For this, several methods are proposed in which various factors must be considered. Cristian Pîrvu et. al [28] discuss the problem of predicting the user intent, only based on the queries that were used to access a certain webpage. They propose a method where both recurrent and convolutional networks are used as models while representing the words in the query with multiple embedding methods. For labelling the data they used a similar approach, also making a distinction between single and multi-intent classifiers, labelling the data automatically and rule-based. The models which are used are recurrent and convolutional networks, fine-tuned with their own requirements. The RNN is tested three times with different amounts of layers. A word embedding is the input for the neural networks which results in a probabilistic output for each of the three intents. When testing the model, for single-intent prediction the one with the highest probability is chosen, or for multi-intent prediction, the probabilities themselves are chosen. They conclude that the learning models outperform the rule-based system on both accuracy and percentage of items labelled. It is also shown that the task of multi-intent prediction from queries can be modelled with both recurrent and convolutional neural networks, trained as a regular classification problem.

Agichtein [2] proposed a real-world study modelling the behaviour of web search users to predict web search result preferences. The model analyzes user activities, such as clicks, scrolls and dwell times to predict user intention during web page navigation. They use a random sample of 3000 queries from their search engine. For every query 30 results were labelled by human judges on a scale ranging from "Perfect" down to "Bad". The interactions, such as clicking on a result link or going back to search results, with the search engine were also recognized and aggregated in the data. These interactions were converted to features as *Clickthrough features*, *Browsing features* and *Query-text features*.

Caruccio [7] proposes to mine user interaction activities to predict the intent of the user during a navigation session. A model for User Intention Understanding (UIU) is proposed, focusing on both interactions with *Search Engine Result Pages* (SERPs) results and on the visited web pages. The intent of the user is recognized by giving participants several goals in different sessions while their interactions were logged by the YAR plug-in for Google Chrome/Chromium [10]. These interaction features, such as query keywords and contextual information are all feeding a classification algorithm to understand the user intent. They built a dataset by recruiting 31 participants with a user profile. The user profile represents their gender, age and experience in using the Web. The participants were requested to perform several search sessions, guided and free, but with a certain goal. These sessions were recorded and analyzed in order to classify the intent of a user. This way a labelled dataset was created with features such as *query, search, interaction and context* and its corresponding intent.

However, to make these predictions based on Natural Language Processing (NLP) it is found two common issues are faced, the lack of training data and recognizing intent behind a short search query. For all discussed approaches data needs to be gathered and labelled in order to train the model. Several studies have been done in dealing with these issues.

Research that aims to investigate how much training data is actually needed for high performance in an intent classification task has been done by Huggins et. al. [16] They have trained and evaluated BiLSTM and BERT models on various subsets of the ATIS and SNIPS datasets. The BERT model results in higher accuracy than the BiLSTM model on both datasets with 25 training examples per intent. This model is built so that no large amounts of labelled data are required for high performance. Nevertheless, they also train the models on a real-world dataset which results in lower accuracy and higher error. Making a good prediction relies on many factors and several methods can be applied to improve a prediction. A common problem of recognizing behaviour behind web queries is that queries are usually short. This makes it hard to recognize user intent. Phan et. al [27] present a general framework for building a classifier that can deal with such short and sparse text, as can be seen in figure 2.1. A large external data collection (the universal dataset) is collected per classification task and a classification model is built on a small set of labelled training data and a rich set of hidden topics discovered from the same data collection. Hidden topics are discovered with the Latent Dichirlet Allocation (LDA) topic model. Their proposed framework aims to reduce data sparseness, expand the coverage of the classifier. This can be seen as flexible semi-supervised learning and it is easy to implement.

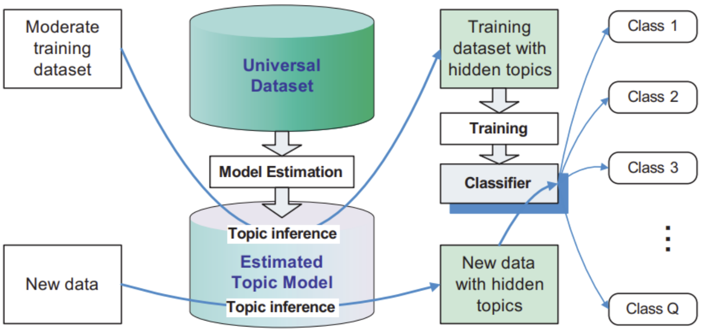


Figure 2.1: The general framework proposed by Phan et. al [27]

Pirvu et. al [28] address the problem in intent recognition of categorizing queries. They argue two approaches can be taken taken, a rule-based approach or a statistical approach. The rule-based approach applies various rules at word and query level mainly based on human intuition. This is a reliable approach, but it is unfeasible for a large dataset. The statistical approach is more feasible for making predictions on new data. First, they annotate a small amount of their dataset to build a ground-truth labeled dataset and then use various automatic labeling methods. Labeling the ground truth dataset is done by several annotaters and both a single- and multi- intent approach is used to increase reliability. For labeling the data automatically various rules are applied such as checking various keywords or looking for similar words. They base their labeling on Broder's taxonomy, thus queries are classified as informational, transactional or navigational. An advantage of classifying on these classes is that general keywords exist, such as *buy* or *rent* almost certainly belong to *transactional*.

Another method that could be applied in improving the prediction of intent is multimodal search. Multimodal search is a type of search that uses different methods to get relevant results. This can be any kind of search, search by keyword, search by concept, search by example, etc. Etzold et. al [12] propose a unified framework for multimodal indexing, sharing, search and retrieval. Specific types of multimedia and multimodal content, such as text, 2D images, hand-drawn sketches, videos, 3D objects, audio files, and also real-world information will be handled by the framework. The search results of these multimodal queries can include any available relevant content.

Lee et. al [21] propose a feature-based model for the automatic identification of search goals, focusing on navigational and informational queries. Their experimental evaluation showed that using a combination of the proposed features allows the model to correctly identify the goals for 90 % of the queries that were studied. These two features are *past user-click behaviour*, to infer users intent from their past interactions with results, and *anchor-link distribution*, which uses possible targets of links sharing the same text with the query.

2.3 Technical background

This section provides a short description of the technical background for understanding the implementation of the intent recognition models. First, Natural Language Processing (NLP) will be described since this is an important part of preparing the model. Secondly, Intent Recognition (IR) will be explained which is a form of NLP. The intent recognition models are based on Recurrent Neural Networks (RNNs) and multi-layer bidirectional transformer (BLTSM) encoders and these will be discussed briefly. Finally, Latent Dirichlet Allocation (LDA) topic modelling will be discussed since this technique will be applied for generating queries.

2.3.1 Natural Language Processing

Natural Language Processing (NLP) is a research area, specifically in the branch of Artificial Intelligence (AI), which explores the way computers can be used to understand and manipulate natural language text or speech. It uses rule-based modelling of human language with statistics, machine learning and deep learning models. It is aimed to understand the voice of text as humans do, including the intent and sentiment. It can be used in real-world applications like spam recognition, Google Translate or virtual agents like Siri or Alexa.

There are common sub-problems in NLP that will also be encountered in preparing the data for this study. Nadkarni et. al distinguish low-level and high-level problems in NLP. [25]

- *Sentence boundary recognition*: abbreviations and titles like 'Dr.' or listed items must be distinguished.
  نقطه‌گذاری
- *Tokenization*: punctuation must be recognized within a sentence. A lexer, which turns a meaningless string into a flat list of words, plays a crucial role in this task.
- *Part-of-speech (POS) tagging*: an important task is POS-tagging. It categorizes words in a corpus in correspondence with a particular part of speech.
- *Morphological decomposition*: complicated words need to be decompositioned to be comprehended. Lemmatization is a useful sub-task that converses a word to a root by removing suffixes.
- *Shallow parsing*: in shallow parsing phrases are identified from constituent part-of-speech tagged tokens.
  تشکیل دهنده

- *Problem-specific segmentation*: text will be segmented into meaningful groups.

Higher-level tasks include *spelling/grammatical error identification, Named entity recognition, word sense disambiguation, negation and uncertainty identification, relationship extraction, temporal inferences or information extraction.* This study used simple NLP low-level tasks.

2.3.2 Artificial Neural Networks

Artificial Neural Networks (ANNs), or neural networks, are computational algorithms intended to simulate the behaviour of biological systems composed of neurons. An ANN is capable of machine learning as well as pattern recognition or classification problems. They have been developed as generalizations of mathematical models of biological nervous systems. Neural networks use artificial neurons or nodes which are input and output devices for the ANN as can be seen in figure 2(a). It receives the input from data and other nodes. [1] Each neuron has a weight, which implies the importance of the input, and a bias that ensures that data can be fit even if the input is 0. The output of the neuron is represented as:

$$O = f(net) = f(\sum_{j=1}^{n} w_j x_j) \tag{2.1}$$

where w_j is the weight vector and f(net) is the activation function. Based on a computation of the weight and the bias an activation function is applied. Often used activations functions are sigmoid, ReLU, Leaky ReLU and tanh. The output of the activation function is either sent to the next layer of the network or it is the final output. An ANN has a layer structure where each layer performs a specific function where the number of layers increases as the complexity of the model increases, therefore it is known as the multi-layer perceptron.

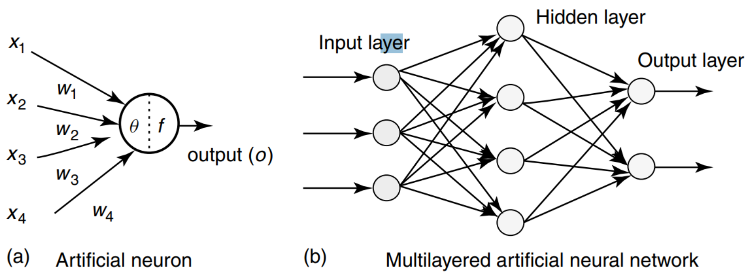


Figure 2.2: The architecture of an artificial neuron and a multilayered network [1]

An ANN consists of three layers: the input layer, the hidden layer and the output layer, as shown in figure 2.2.

To train neural networks, data is important, poor training data leads to an unreliable and unpredictable network. To create a reliable network often noise or randomness is added to the training data. Training the network requires a number of epochs to decrease the output error below a particular threshold. A fixed number of epochs is necessary since a model can be overtrained and will be unable to correctly classify samples outside the training set.

دوره‌ها

2.3.3 Recurrent Neural Networks

A Recurrent Neural Network (RNN) model has been shown to significantly outperform many competitive language modelling techniques in terms of accuracy and is often used to make predictions. [22] An RNN is a class of artificial neural networks where connections between nodes form a directed or undirected graph along a temporal sequence. An RNN refers to the class of networks with an infinite impulse response, whereas a convolutional neural network (CNN) refers to the class of finite impulse response. It means that RNNs use information from previous outputs for the current in- and output, in other words, they do not assume that inputs and outputs are independent of each other. [30] In a bidirectional neural network, the output layers are connected forwards and backwards. The output layer, therefore, receives more data, which increases the performance of the neural network. RNNs have a short memory and cannot hold all information over a longer period of time. For this problem, long short-term memories (LSTMs) are applied which can carry sequential data for a longer period of time. [20] Bidirectional long-short term memory (Bidirectional LSTM) is the process of making a neural network to have the sequence information in both directions backwards (future to past) or forward (past to future), as can be seen in figure 2.3. A BLSTM network can be used in text classification, speech recognition and forecasting models since the input flows in both directions to preserve the future and the past information.

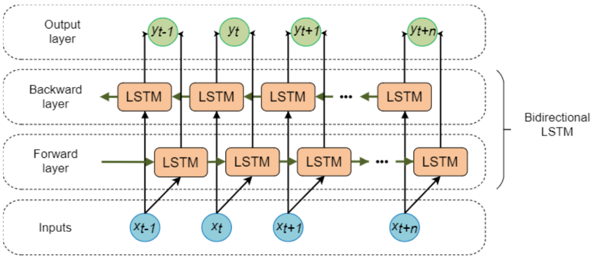


Figure 2.3: BLSTM architecture [20]

2.3.4 Latent Dirichlet Allocation (LDA) Topic Modelling

A subproblem in NLP is topic modelling, which will be used for generating queries from a large collection of data. A method called Latent Dirichlet Allocation (LDA) will be used to discover hidden topics in large data collections. LDA a is a three-level hierarchical Bayesian model where each item of a collection is modelled as a finite mixture over an underlying set of topics [4]. Every topic is considered as a set of terms sharing a theme. The general idea is to discover a set of terms based on the co-occurrence of individual terms.

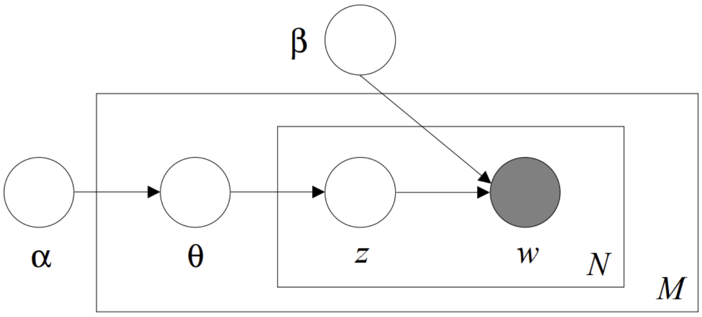


Figure 2.4: Graphical model representation of LDA [4]

In figure 3.2 the boxes are "plates" representing replicates. The outer plate M represents the number of documents, while the inner plate N represents the number of words in a given document (document i has N_i words). Further, α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the parameter of the Dirichlet prior on the per-topic word distribution, θ_i is the topic distribution for document i, φ_k is the word distribution for topic k, z_{ij} is the topic for j-th word in document i, w_{ij} is the specific word.

The following generative process for each document w in a corpus D is assumed by LDA:

1. Choose N $\sim$ Poisson(ϵ).
2. Choose $\theta \sim$ Dir(α).
3. For each of the N words w_n:
   - (a) Choose a topic $z_n \sim$ Multinomial(θ).
   - (b) Choose a word w_n from p(w_n $|z_n, \beta$), a multinomial probability conditioned on the topic z_n

Bayesian Statistics

2.4 Gap analysis

تجارت الکترونیک

Firstly, most of the user intent past research is based on ~~e-commerce~~, while this thesis aims to recognize the intent in an environmental search engine. The goal of both research areas is the same; predicting what the user wants, nevertheless the intent features can be different. When someone is aiming to buy something from the internet their queries are stated differently than when someone is searching for scientific papers for example. Moreover, their click behaviour may be different than when a user is writing an algorithm and is searching for data. Since an environmental database consists of many data and data types it is difficult to efficiently search in the search engine. Many studies are based on Broder's taxonomy which is the taxonomy for search terms in web searches that divides the need behind a search query into three classes: informational, navigational and transactional which is a general method of classifying, while this research aims to classify the intent with specific classes. For these three classes, general keywords can be thought of. (E.g. *buy* or *rent* for transactional) Thus, this study will research the challenges of recognizing an intent with specified classes.

Furthermore, a user intent prediction model will be introduced to which not much research has been done. The model will be compared to the BERT model, which has been researched broadly for predicting user intent. Thus, this thesis aims to evaluate a new model for predicting intent.

Thus, this thesis aims to bridge the gap between recognizing user intent in e-commerce and recognizing user intent in another situation, specifically browsing a scientific search engine. Additionally, a model which has not been researched broadly will be introduced and evaluated.

Chapter 3

Methodology

This chapter describes the method used to research how to recognize and predict the intent of a user when browsing a search engine. First, the state of the art of existing user intent models is researched and several user intent approaches are found. It is found that not many open-source models are available, therefore two pre-trained models are selected, the BERT intent recognition model and the Ludwig model. Both models are built to solve one of the biggest challenges in NLP, namely the shortage of training data. Both models are trained for dealing with data consisting of queries labelled with a certain intent. Therefore this study will aim to recognize the intent by training the model on submitted queries. Section 3.1 will describe the experimental setup. Then section 3.2 provides information about the data that will be used to make predictions. Finally, in section 3.3 the selected models will be evaluated.

3.1 Experimental setup

Predicting the intent of a user is a supervised learning process for which a labelled dataset is needed to understand the intent behind the behaviour of a user, which is an intensive process. First, in order to evaluate the two models' best performance, they are trained on both the SNIPS and ATIS datasets, which is shown in figure 3.1. The SNIPS dataset is a well-balanced dataset, while ATIS is an unbalanced dataset, therefore it is chosen to test the models on both datasets. Furthermore, these datasets are chosen since they are open-source and often used in user intent prediction.

Then, based on these results it is determined how the models can be tested in a real-world application. To test the models in a real-world application, queries will be generated from all web pages of the ENVRI search engine, aiming to create a similar database as SNIPS and ATIS. By performing Latent Dirichlet Allocation (LDA) topic modelling on the collection of all documents, potential queries from the search results are retrieved. Topic modelling is an unsupervised machine learning technique that scans a set of documents, recognizes words and phrase patterns within these documents and clusters word groups that characterize a set of documents. The topics that are returned by the LDA algorithm are unsupervised single words. These are all classified manually into the 7 ENVRI classes. This will be further explained in section 3.2.3. Finally, both models will be tested on unseen data by making new predictions on all three datasets. Based on these results a comparison is made.

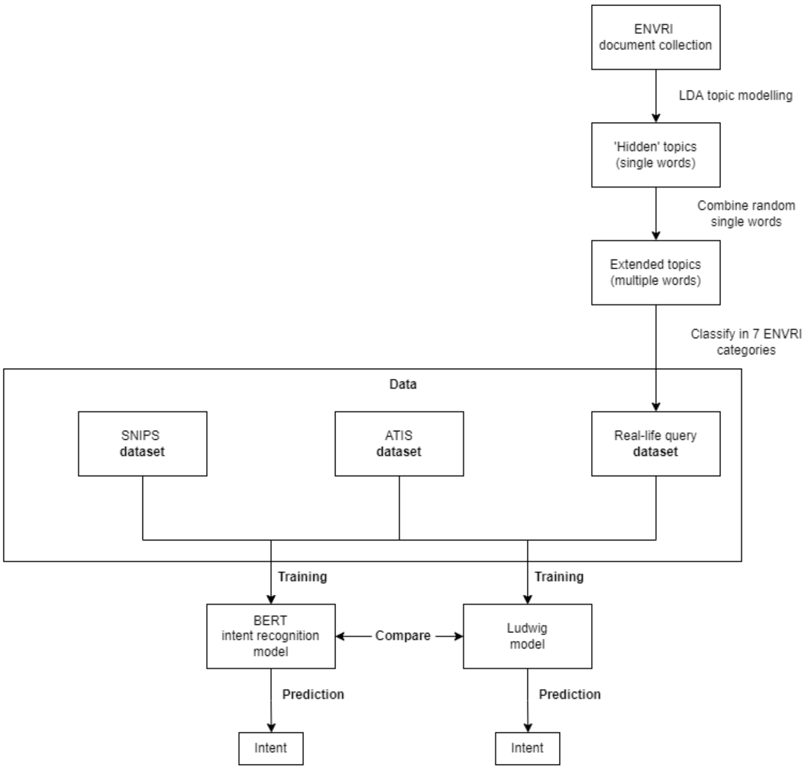

ENVRI document collection
LDA topic modelling
'Hidden' topics (single words)
Combine random single words
Extended topics (multiple words)
Classify in 7 ENVRI categories
Data
SNIPS dataset
ATIS dataset
Real-life query dataset
Training
Training
BERT intent recognition model
Compare
Ludwig model
Prediction
Prediction
Intent
Intent

Figure 3.1: Method diagram

3.2 Data

Both models are trained on the SNIPS Natural Language Understanding benchmark dataset and the ATIS dataset. Then, another dataset, which consists of potential queries generated from the ENVRI search engine labelled with a corresponding intent will be tested on the models.

3.2.1 The SNIPS Natural Language Understanding benchmark

The SNIPS Natural Language Understanding benchmark is a dataset containing 16,000 queries categorized into seven unique intents of various complexity:

- SearchCreativeWork (e.g. Find me the I, Robot television show)

- GetWeather (e.g. Is it windy in Boston, MA right now?)
- BookRestaurant (e.g. I want to book a highly rated restaurant for me and my boyfriend tomorrow night)
- PlayMusic (e.g. Play the last track from Beyoncé off Spotify)
- AddToPlaylist (e.g. Add Diamonds to my road trip playlist)
- RateBook (e.g. Give 6 stars to Of Mice and Men)
- SearchScreeningEvent (e.g. Check the showtimes for Wonder Woman in Paris)

This pre-labelled dataset is often used for making model comparisons therefore, it is chosen to use this data for the selected models. The training set consists of 13.784 examples, the test and validation set consist of 700 examples. The data turns out to be balanced so no further preprocessing needs to be done.

3.2.2 ATIS dataset

The ATIS (Airline Travel Information Systems) dataset consists of audio recordings and corresponding manual transcripts about humans asking for flight information on automated airline travel inquiry systems. It contains 8 unique intent categories.

- atis_flight
- atis_flight_time
- atis_airfare
- atis_aircraft
- atis_ground_service
- atis_airline
- atis_abbreviation
- atis_quantity

It is originally split into 4478 training examples, 500 validation examples and 893 test examples. This dataset is chosen since it is an unbalanced dataset and therefore the performance of the models can be compared.

3.2.3 Real-life queries

It is aimed to create a similar database as the SNIPS and ATIS dataset. *Latent Dirichlet Allocation* (LDA) topic modelling is used for discovering the abstract "topics" that occur in a collection of documents. [4] This will be applied over the collection of all documents in the ENVRI search engine in order to generate topics that will serve as potential queries. This method is chosen since the potential queries will serve as general queries for the whole collection of all documents. In machine learning and natural language processing, topic models are generative models which provide a probabilistic framework. A document is seen as a mixture of hidden topics that can be found by applying LDA. There are several topic modelling methods, but LDA is the most complete approach as it follows a full generation process for document collection. [13] Topic modelling is an unsupervised method which means determining the number of topics is done by trial and error. In this case, it is decided to generate 20 topics per document in the first round and 400 topics in the second round in order to generate queries from the database as can be seen in figure 3.2.

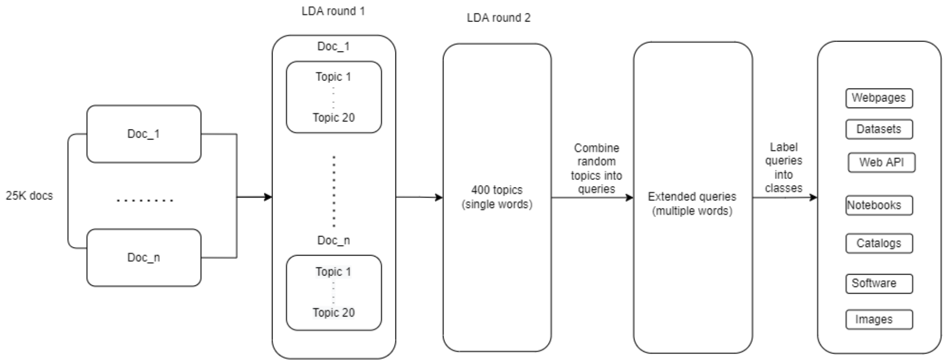
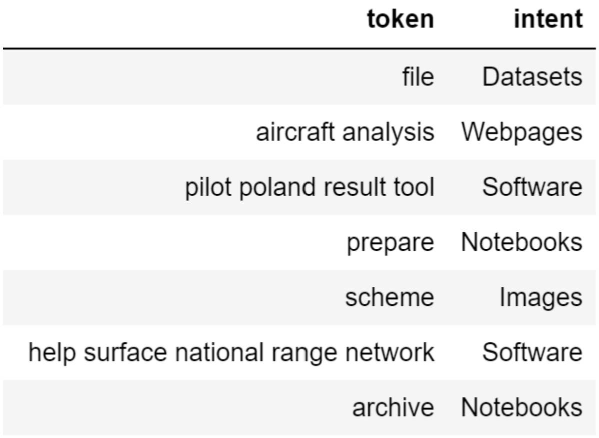


Figure 3.2: Process of generating queries with LDA topic modelling

In order to perform the LDA modelling on the data, some preprocessing steps are executed. Tokenization splits the text into sentences and the sentences into words, words are lowercased and punctuation is removed, words that have fewer than 3 characters are removed, stopwords and unnecessary words are removed and words are stemmed meaning they are reduced to their root form.

To generate general queries, two LDA rounds are performed. The first round is performed for every document of the database. 20 topics per document are retrieved and therefore 25k lists of 20 topics in total are returned. In the second round, LDA is performed over the 25k lists of topics and it is chosen to return 400 topics. These topics are the top 400 'hidden' topics that occur in the 25k documents. The LDA topic model returns topics containing hidden themes which are still unsupervised and unlabeled. The intents are in this case the 7 classes in which the ENVRI search engine is categorized: *webpages,* *datasets*, *Web APIs*, *notebooks,* *catalogs*, *software* or *image*. From the LDA model, 400 unique words are returned which is not enough to train the model on, therefore by combining these topics more queries are generated. 600 queries of 3,4 and 5 words are added to the database by combining the 400 single unique words. Two methods are tested for labelling the data. First, all queries are manually labelled with an intent class but can belong to multiple classes (multi-intent labelling) so that the classifier can predict for multiple labels. In addition, a labelling method was tested in which single words were classified manually. Then, queries consisting of multiple words were classified by aggregating the classes of the single contributing words. This led to queries that were not able to discriminate between classes. Hence, these queries had to be removed from the dataset. However, the resulting dataset was too small to accurately train the models. For this reason, the former approach was chosen. The final dataset that is returned consisted of 1000 queries labelled with intent as shown in figure 3.3.



token	intent
file	Datasets
aircraft analysis	Webpages
pilot poland result tool	Software
prepare	Notebooks
scheme	Images
help surface national range network	Software
archive	Notebooks

Figure 3.3: Example of queries

3.3 User intent recognition models

There are many methods to predict user intent but all tasks require different sources of information, such as geo-location, user profile or web surfing history. Moreover, a labelled dataset is required. Therefore it is important to analyse what kind of data is required for the specified model. This study will evaluate two models: the BERT intent recognition and the Ludwig Model, both pre-trained on unlabeled data. They are used for different natural language processing tasks e.g. predicting user intent based on by a user-submitted search query. Both models will be trained on a pre-labelled dataset and the dataset that is described in section 3.2.3.

3.3.1 BERT intent recognition model

Bidirectional Encoder Representations from Transformers (BERT) is a pre-training NLP technique and is designed to pre-train deep bidirectional representations from unlabelled text. It can be used for many tasks, such as question answering and language inference. A large advantage of the BERT model is that it can be used without having a large dataset. Being bidirectional means considering context to its left and right which leads to a remarkable performance on various language tasks. BERT consists of two stages, the pre-training stage and fine-tuning stage. In the first stage, the model is trained on a vast unlabeled amount of text. In the second stage, the model is initialized with weights of the pre-trained model and fine-tuned with task-specific labelled data.

3.3.1.1 Model architecture

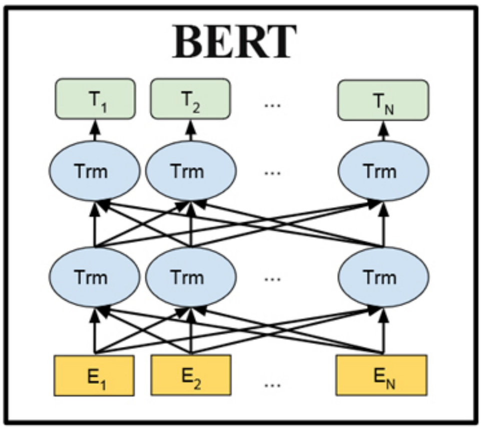


Figure 3.4: Multi-layer bidirectional Transformer encoder BERT [11]

BERT embeddings are trained with two training tasks. The classification task determines which category the input sentence should fall into and the Next Sentence Prediction Task determines if the second sentence naturally follows the first sentence. During fine-tuning, all parameters are fine-tuned. For the classification task, a special symbol [CLS] is added in front of every input example to preprocess the input text data and [SEP] is a special separator token (e.g. separating questions/answers) which is applied in the Next Sentence Prediction Task.

به خوبی تنظیم شده است.

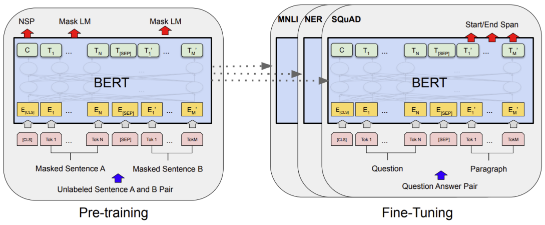


Figure 3.5: Pre-training and fine-tuning stages [11]

3.3.1.2 Specifications

The model is now trained on the two pre-processed labelled text datasets with seven intents. The data is divided into a train, test and validity set. The BERT model is fine-tuned by using the inputs (text and intent). Dropout is added, which is a technique where randomly selected neurons are ignored during training, with two fully-connected layers. The last layer has a softmax activation function. To compile the model, the Adam optimiser with sparse categorical cross-entropy has been used.

3.3.2 Ludwig model

Ludwig is a toolbox that trains and tests deep learning models for making predictions. The model is based on datatype abstraction, so that the same data preprocessing and postprocessing will be performed on different datasets that share datatypes and the same encoding and decoding models developed can be re-used across several tasks. The motivation behind this abstraction are recurring patterns in deep learning projects, given certain types of inputs pre-processing code is often the same in specific tasks, as is the code implementing models and training loops. Models can be hard to compare and reused because of small differences. [24]

3.3.2.1 Model architecture

The model adopts an Encoders-Combiner-Decoders architecture (ECD) as can be seen in figure 3.6. Encoders encode different features of the input data, the combiner combines information received from the encoders and decoders decode the input into one or more output features. Ludwig adopts this architecture since it maps most deep learning models architectures and allows modular composition.

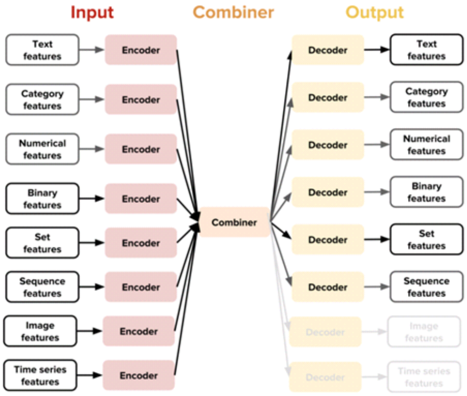
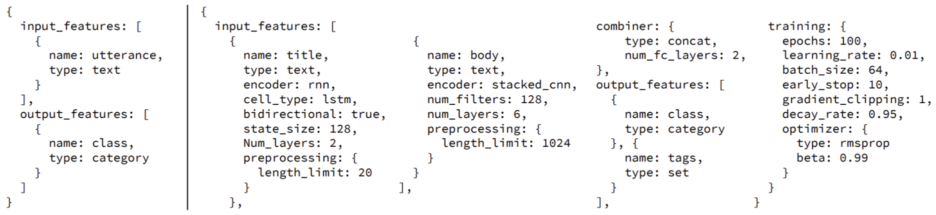


Figure 3.6: Encoder-Combiner-Decoder Architecture [24]

Therefore, to run the model simply a dataset file needs to be provided and input and output features have to be defined in a model definition YAML file as shown in figure 3.7. The model definition contains a list of input features and output features, where the names of the columns in the dataset that are input to the model alongside their datatypes, and names of columns in the dataset that will be output, the target variables which the model will learn to predict the need to be specified.



```
{
  input_features: [
    {
      name: utterance,
      type: text
    }
  ],
  output_features: [
    {
      name: class,
      type: category
    }
  ]
}
```

```
{
  input_features: [
    {
      name: title,
      type: text,
      encoder: rnn,
      cell_type: lstm,
      bidirectional: true,
      state_size: 128,
      Num_layers: 2,
      preprocessing: {
        length_limit: 20
      }
    },
```

```
    {
      name: body,
      type: text,
      encoder: stacked_cnn,
      num_filters: 128,
      num_layers: 6,
      preprocessing: {
        length_limit: 1024
      }
    }
  ],
```

```
combiner: {
    type: concat,
    num_fc_layers: 2,
},
output_features: [
  {
    name: class,
    type: category
  }, {
    name: tags,
    type: set
  }
],
```

```
  training: {
    epochs: 100,
    learning_rate: 0.01,
    batch_size: 64,
    early_stop: 10,
    gradient_clipping: 1,
    decay_rate: 0.95,
    optimizer: {
      type: rmsprop
      beta: 0.99
    }
  }
}
```

Figure 3.7: Left: minimal model definition. Right: complex model definition. [24]

3.3.2.2 Specifications

→ مشخصات فنی

The dataset that is used contains a column with tokens, these are the queries in the used dataset and the labelled intent of every query. The chosen encoder is a Recurrent Neural Network (RNN) with Long Short-Term Memory (LTSM) Networks and 2 layers. The encoder is a process that converts information from one format to another to standardize or reduce for speed, analysis or prediction. The RNN encoder first maps the input integer sequence into a sequence of embeddings, then passes the embedding through a stack of recurrent layers, followed by a reduced operation that by default returns the last output. The model runs until it notices that the best results are in a past epoch and accuracy has gone down since then.

Chapter 4

Experiment results

The obtained results of both models are described in this chapter. Firstly, per model the results of the pre-labelled dataset will be shown and examples of potential queries with the predicted intent will be given. In section 4.1 the results of the BERT model will be shown with an evaluation metric and a confusion matrix. Secondly, the Ludwig model will be shown in section 4.2. The results of the real-life dataset trained on both models will be discussed in section 4.3.

4.1 BERT intent recognition model

As can be seen in table 4.1 BERT achieves a high accuracy of 96% on the SNIPS dataset. Precision, recall and F1-score are all close to 100%. Except SearchCreativeWork is occasionally misclassified as PlayMusic or SearchScreeningEvent, therefore recall and F1-score is lower for that class.

	precision	**recall**	**f1-score**	**support**
PlayMusic	0.89	1.00	0.94	86
AddToPlaylist	1.00	1.00	1.00	124
RateBook	1.00	1.00	1.00	80
SearchscreeningEvent	0.89	0.94	0.92	107
BookRestaurant	0.99	1.00	0.99	92
GetWeather	0.99	0.99	0.99	104
SearchCreativeWork	0.94	0.79	0.86	107
accuracy			**0.96**	700
macro avg	0.96	0.96	0.96	700
weighted avg	0.96	0.96	0.96	700

Table 4.1: BERT evaluation metric SNIPS

The confusion matrix in figure 4.1 gives insight into the predicted results. The number of correct and incorrect predictions are summarized with count values and broken down by each class. It can be seen that AddToPlaylist is predicted correctly most often and RateBook is predicted correctly the least. The amount of misclassifications is low since most columns are 0, which means they are not classified as that particular class. It is aimed to achieve 0 misclassifications.

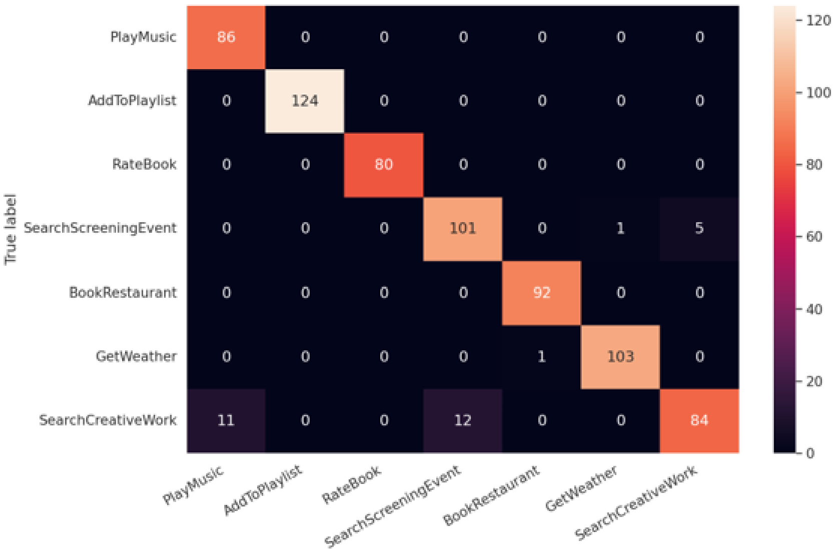


Figure 4.1: BERT confusion matrix SNIPS

As can be seen in table 4.2 BERT achieves an accuracy of 79% on the ATIS dataset, which is lower than the achieved accuracy on the SNIPS dataset. The ATIS dataset is highly unbalanced which can be seen in the confusion matrix in figure 4.2. The atis_flight is predicted correctly the most but is also misclassified the most. Most columns are 0, which means they are never classified, but also not misclassified. This is because most of the queries are labelled as atis_flight making the accuracy high, since the chance of a query actually being atis_flight is high.

	precision	**recall**	**f1-score**	**support**
intents	0.00	0.00	0.00	1
atis_flight	0.79	1.00	0.88	632
atis_flight_time	0.00	0.00	0.00	1
atis_airfare	0.00	0.00	0.00	48
atis_aircraft	0.00	0.00	0.00	9
atis_ground_service	0.00	0.00	0.00	36
atis_airline	0.00	0.00	0.00	38
atis_abbreviation	0.00	0.00	0.00	33
atis_quantity	0.00	0.00	0.00	3
accuracy			**0.79**	801
macro avg	0.09	0.11	0.10	801
weighted avg	0.62	0.79	0.70	801

Table 4.2: BERT evaluation metric ATIS

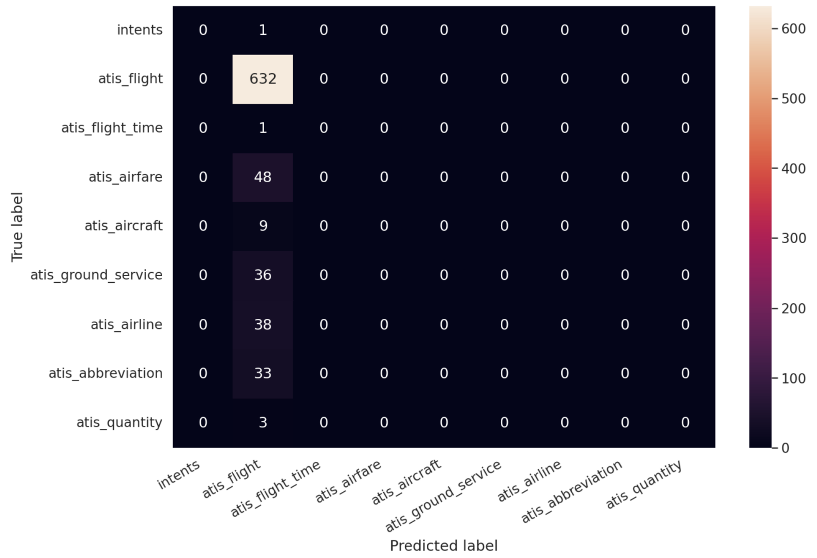
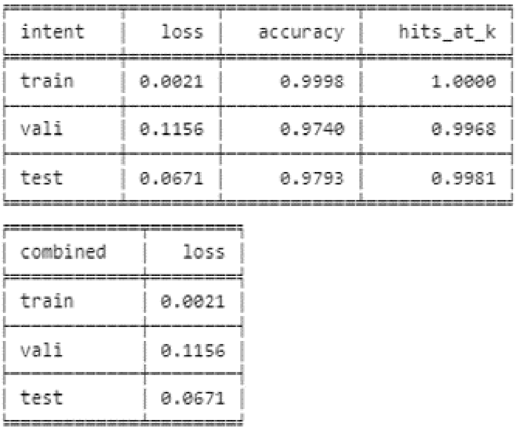


Figure 4.2: BERT confusion matrix ATIS

4.2 Ludwig model

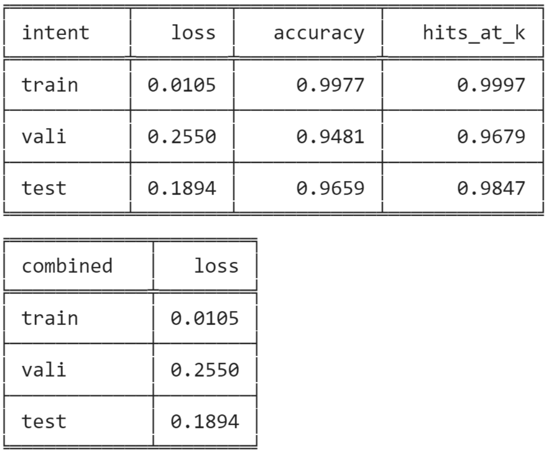
As can be seen in figure 4.6 Ludwig achieves a high accuracy of 97.93% on the SNIPS dataset. The hits_at_k is the number of times that a correct prediction was made in ratio to the number of total prediction names, which also has a high score. Another important measurement is the loss, which is the penalty for a bad prediction, therefore a low loss means a low penalty. The loss on the test set is 6.71 % meaning Ludwig performs well on this dataset.



```
╒══════════╤════════╤════════════╤═════════════╕
│ intent   │   loss │   accuracy │   hits_at_k │
╞══════════╪════════╪════════════╪═════════════╡
│ train    │ 0.0021 │     0.9998 │      1.0000 │
├──────────┼────────┼────────────┼─────────────┤
│ vali     │ 0.1156 │     0.9740 │      0.9968 │
├──────────┼────────┼────────────┼─────────────┤
│ test     │ 0.0671 │     0.9793 │      0.9981 │
╘══════════╧════════╧════════════╧═════════════╛
╒════════════╤════════╕
│ combined   │   loss │
╞════════════╪════════╡
│ train      │ 0.0021 │
├────────────┼────────┤
│ vali       │ 0.1156 │
├────────────┼────────┤
│ test       │ 0.0671 │
╘════════════╧════════╛
```

Figure 4.3: Ludwig evaluation metric SNIPS

When training Ludwig on the ATIS dataset, an accuracy of 96,59% is achieved. This is comparable with the accuracy when trained on the SNIPS dataset. Furthermore, the loss is higher with a loss of 18.94 %. However, hits_at_k is 98.47 % on the test set which means a correct prediction was made a 98.47 % in the total of prediction names .



```
intent        loss    accuracy    hits_at_k
train       0.0105      0.9977       0.9997
vali        0.2550      0.9481       0.9679
test        0.1894      0.9659       0.9847

combined      loss
train       0.0105
vali        0.2550
test        0.1894
```

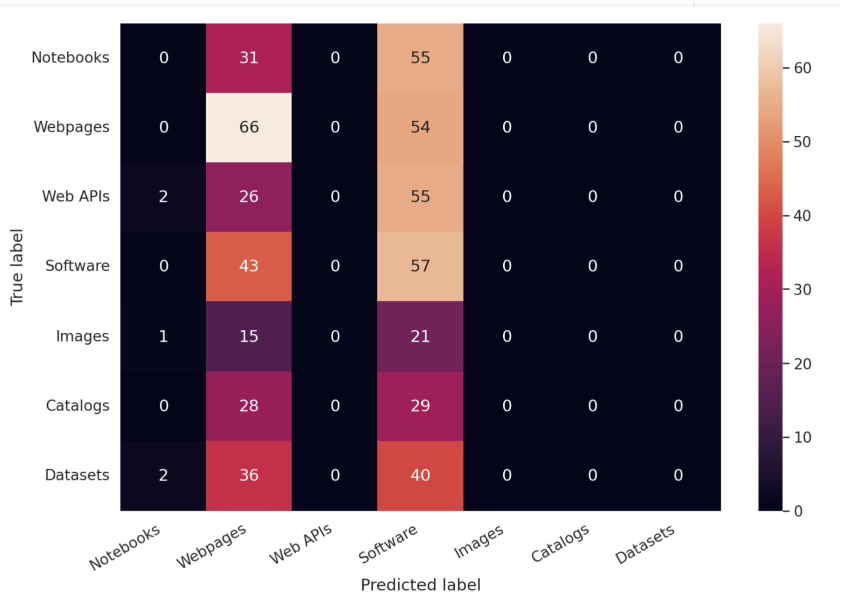
Figure 4.4: Ludwig evaluation metric ATIS

4.3 BERT and Ludwig trained on real-life data

As can be seen in table 4.3 and figure 4.5 BERT does not achieve high accuracy on the real-life dataset. The confusion matrix shows that the classes Software and Webpages are in some cases classified correctly, but are also the classes that are misclassified most often.

	precision	**recall**	**f1-score**	**support**
Notebooks	0.00	0.00	0.00	86
Webpages	0.27	0.55	0.36	120
Web APIs	0.00	0.00	0.00	83
Software	0.18	0.57	0.28	100
Images	0.00	0.00	0.00	37
Catalogs	0.00	0.00	0.00	57
Datasets	0.00	0.00	0.00	78
accuracy			**0.22**	561
macro avg	0.06	0.16	0.09	561
weighted avg	0.09	0.22	0.13	561

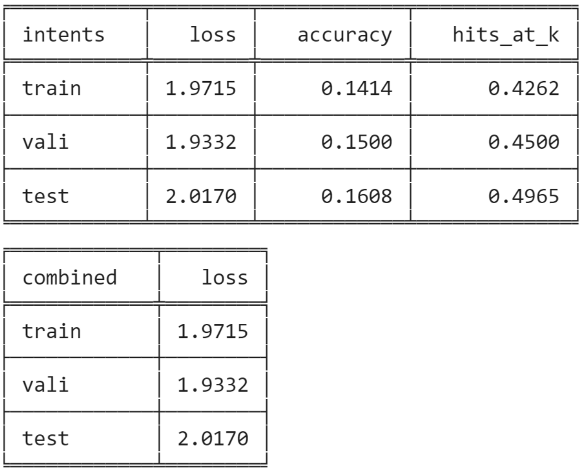
Table 4.3: BERT evaluation metric real-life queries



	Notebooks	Webpages	Web APIs	Software	Images	Catalogs	Datasets
Notebooks	0	31	0	55	0	0	0
Webpages	0	66	0	54	0	0	0
Web APIs	2	26	0	55	0	0	0
Software	0	43	0	57	0	0	0
Images	1	15	0	21	0	0	0
Catalogs	0	28	0	29	0	0	0
Datasets	2	36	0	40	0	0	0

Figure 4.5: BERT confusion matrix real-life data

The Ludwig model also does not achieve high accuracy. The hits_at_k is almost 50 %, and the loss is high. Both models aim to deal with unbalanced data. Some explanations for these results can be thought of which will be discussed in section 5.1.



```
╒═══════════╤════════╤════════════╤════════════╕
│ intents   │   loss │   accuracy │   hits_at_k │
╞═══════════╪════════╪════════════╪════════════╡
│ train     │ 1.9715 │     0.1414 │     0.4262 │
├───────────┼────────┼────────────┼────────────┤
│ vali      │ 1.9332 │     0.1500 │     0.4500 │
├───────────┼────────┼────────────┼────────────┤
│ test      │ 2.0170 │     0.1608 │     0.4965 │
╘═══════════╧════════╧════════════╧════════════╛
╒════════════╤════════╕
│ combined   │   loss │
╞════════════╪════════╡
│ train      │ 1.9715 │
├────────────┼────────┤
│ vali       │ 1.9332 │
├────────────┼────────┤
│ test       │ 2.0170 │
╘════════════╧════════╛
```

Figure 4.6: Ludwig evaluation metric real-life data

Chapter 5

Discussion

This section discusses the limits and possible improvements of this thesis' results. The Ludwig model and BERT intent recognition model are evaluated in section 5.1 and 5.2, as is the method of modelling the user intent as search queries in section 5.3. In section 5.4 will discuss future work.

5.1 Model comparison

Both models are trained on the same datasets, BERT achieves an accuracy of 96 % and Ludwig an accuracy of 97.93 % on the SNIPS test set. When trained on the ATIS dataset, BERT achieves an accuracy of 79% and Ludwig an accuracy of 98.47 %. Although, for an unbalanced dataset like ATIS, accuracy is not always a reliable measurement. This is because if atis_flight is 90% of the dataset and atis_airline is 10% of the dataset, the model will still achieve an accuracy of 90% if it fails to predict atis_airline. So, accuracy does not hold for unbalanced data. Therefore, the loss and hits_at_k are considered for evaluating the Ludwig model.

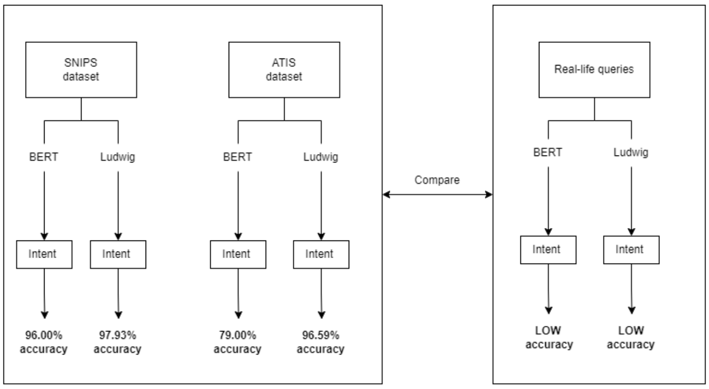


Figure 5.1: Models' performance

An example of a prediction for both models on the SNIPS dataset with a given input query is as follows:

Query: i want to bring four people to a place that s close to downtown that serves churrascaria cuisine
Intent: BookRestaurant

Query: put lindsey cardinale into my hillary clinton s women s history month playlist
Intent: AddToPlaylist

Both models achieve a similarly high accuracy and predict an equal intent on the SNIPS dataset. A few advantages and disadvantages are found. Due to the pre-training high speed training on unseen data is provided for both models, which saves time and memory. Furthermore, pre-training results in the possibility of training the model on a relatively small dataset. BERT has been tested on small datasets achieving high accuracy, despite dealing with an unbalanced dataset. BERT is conceptually simple and empirically powerful and can be fine-tuned with one additional output layer. It can be used for a wide range of tasks, such as question answering and language inference, without many architecture modifications.

استنتاج زبان

خودکفا

Ludwig model as well is a complete self-contained pipeline and can be trained without a huge amount of data. It is easy to use and has many possible options for adjustments. تنظیمات Ludwig requires almost no coding, which also could be seen as a disadvantage. The data is automatically split and the amount of epochs is established while training, which is helpful but rather makes it a black box since it is hard to evaluate the model when a specific issue occurs. By training the models on the ATIS dataset, it can be concluded Ludwig performs better on unbalanced data than BERT.

	BERT model	**Ludwig model**
Training on small dataset	✓	✓
Easy fine-tuning	✓	×
Fast training	✓	✓
Clear evaluation	✓	×
Dealing with unbalanced data	×	✓

Table 5.1: Comparison BERT and Ludwig.

Thus, both models are suitable for predicting the intent. The main advantage of BERT is that it is easier to make adjustments. While Ludwig is also easy to use, limited information is available on the model besides the documentation. This can be explained since it is a relatively new model, thus not much research has been done. Nevertheless, it is found Ludwig performs better with unbalanced data than BERT. Both models are pre-trained and therefore able to achieve high accuracy with a small dataset and achieve high-speed training. These are all features solving the general issue of data scarcity, however, a disadvantage is that being pre-trained makes it impossible to combine their features. Therefore, a model must be built. Furthermore, training the models on real-life data raised issues that will be discussed in the next section.

When preparing the real-life data for the real-world application an issue occurred that could be argued as a limitation of both models. The labelling method described in section 3.2.3 in which single words were classified manually faced some issues. Queries consisting of multiple words were classified by aggregating the classes of the single contributing words. This led to queries not being able to discriminate between classes. Hence, these queries had to be removed from the dataset. Which resulted in a small dataset leading the model to inaccurately train the models. The main issue of labelling data this way means that a longer query will often belong to more classes than a short query. Thus, labelling data this way will achieve a high accuracy since longer queries belong to multiple classes and therefore have a higher chance of being classified correctly. This is a disadvantage since in a real-life application submitting a longer query will give an inaccurate prediction. In reality, combining single words into a sentence changes the meaning of the separate words. This brings back the earlier discussed issue of labelling queries, which addresses the issue of scarcity of training data. Thus, this shows the importance of having qualitative data. Both models have especially been developed for dealing with the issue of few training data, therefore the challenge of recognizing intent behind a search query is emphasized by this study.

تَبعیضِ
قائل شدن

5.2 Intent recognition methods

This thesis aimed to research how intent can be recognized. The first research question is formulated as how a users' intent can be recognized based on a search query when browsing search results. To answer this question, an experiment was set up to recognize intent on a real-life dataset. The classes were defined as the classes in which the ENVRI search engine is categorized. Many previous approaches for recognizing intent are based on Broder's taxonomy, thus it is analyzed how to label this real-life data. It is found query classification often faces issues as web queries are usually very short, ambiguous and often belong to multiple categories. A large dataset is required to train a machine learning model since a search engine has a substantial amount of information available and therefore many possible queries. Low accuracy was achieved when training the model on real-life data. This could be explained by queries being too short or queries being ambiguously labelled since the data was labelled manually. Thus, it is important to have a dataset that is labelled correctly in order to recognize the intent.

While this thesis focused on predicting intent based on queries, other experiments can be performed to capture the users' behaviour and make a more significant prediction. Studies have been done to predict user intent, however, those studies used real-world data labelled by people. In this case, the available data was unlabeled and unsupervised. It was chosen to use Latent Dirichlet Allocation (LDA) topic modelling to generate the queries. This is one of the possibilities for using topic modelling, while there are more methods like Latent Semantic Analysis (LSA) or Non-Negative Matrix Factorization (NMF), the most often used approach is to manually label the data by multiple persons. Here, the real-life dataset, based upon the queries for the ENVRI search engine, was manually labelled. As a result, the dataset was well balanced, however, overfitted the data achieved and achieved a low accuracy. When the model overfits it tries to learn the data based on a few samples and is more likely to see patterns that do not exist, which resulted in high variance and high error on a test set.

5.3 Model evaluation

The second research question aims to answer how users' intent can be predicted when browsing search results. Therefore, two models were selected to predict users' intent. It was chosen to test pre-trained models since not many open source models were available and both models deal with the issue of training data shortage, which answers subquestion 2.1 questioning the existing intent prediction models. Several intent recognition approaches are discussed in previous research but were not available to use. However, training such a model takes a lot of time and memory which already has been done with pre-trained models. Therefore, using pre-trained models saves time and memory. On these terms the models were selected, answering subquestion 2.2 on how to select a good prediction model. The two evaluated models achieved high accuracy and not much loss on the SNIPS and ATIS dataset, while on the real-life queries both models did not perform well. Both models are suitable for predicting the intent. BERT's main advantage is it is adjustable and several studies have chosen to use this model for user intent prediction. However, Ludwig is easy to use, but being a relatively new model, not many resources were available besides the documentation, which made it challenging to evaluate the model. Nevertheless, Ludwig performs better with unbalanced data than BERT. Both models are pre-trained and therefore able to achieve high accuracy with a small dataset and achieve high-speed training. These are all features solving the common problem of data scarcity, but since they are pre-trained it is not possible to merge the models. Furthermore, training the models on real-life data showed the importance of having qualitative data.

قابل تنظیم

کیفی

5.4 Future work

As discussed, the main issue this thesis faced was related to the quality of the data. Therefore, the framework proposed by Phan et. al which aims to deal with short and sparse text approaches, describing one of the limits of this thesis: the issue of classifying short queries, should be applied for future work. A large external data collection (the universal dataset) is collected per classification task and a classification model is built on a small set of labelled training data and a rich set of discovered hidden topics. Thus, this approach could be applied in order to deal with the issue of short queries.

For this thesis, the intent is modelled as search queries which is just a small fraction of the search behaviour. Recognizing user intentions in real-time can provide personalized recommendations in order to meet user needs and improve user experience. Recognizing these intents can be provided by capturing real-time data when users are interacting on websites. Agichtein et. al modelled user intent by capturing user behaviour such as eye-tracking, mouse movements or dwell time. Caruccio et. al built a labelled dataset that was created with features as *query, search, interaction and context* and its corresponding intent by performing real-life experiments. A deeper understanding of a user's intent could lead to a more detailed intent prediction in the future. Thus, for future work experiments can be done to capture data such as click behaviour, mouse movements or dwell time when a user is interacting with websites to recognize the intent.

However, when such data is available it must be labelled since user intent prediction is a supervised problem. To label data, a rule-based method should be applied for automatically labelling the data. Pirvu et. al have proposed a method for automatic labelling since manually labelling data is unfeasible with a large amount of data. The proposed method is based on various rules classifying the queries as informational, transactional or navigational. This study aimed to predict intent based on the 7 classes from the ENVRI search engine. Therefore, to accomplish such a rule-based method, a large dataset of keywords per intent class need to be gathered in order to automatically label the data according to their approach. If this can be achieved data can be classified fast and efficiently to achieve high accuracy.

If the data is labelled, predictions can be made. A user intent prediction method, without using pre-trained models, is proposed by Pirvu et. al where both recurrent and convolutional networks are used as models representing the words in the query with multiple embedding methods. An RNN is tested three times with different amounts of layers resulting in a probabilistic output for each of the three intents. It is concluded RNNs can be used for multi-intent prediction with queries. No pre-trained models are used since they can not be combined together and have little room for adjustments. Therefore, building an RNN would give more flexibility and thus can be researched for future work.

Chapter 6

Conclusion

This thesis aimed to research how to recognize user intent by a submitted search query and how to predict a new intent when browsing a scientific search engine as the ENVRI search engine. It is found predicting users' intent is a field that has an increasing amount of interest. In order to predict intent, intent must be recognized. For this thesis, it was chosen to model intent as search queries. As discussed, it is discovered a general issue in intent recognition is the scarcity of data, since intent recognition is a supervised problem meaning substantial training data is needed to predict intent on new data. Most intent models are based on search queries, therefore an experiment was set up to test the models on real-life queries extracted from the ENVRI search engine. Testing the models based on real-life data resulted in a conclusion that recognizing intent is challenging since queries are usually very short and ambiguous. A labelled dataset is required to train a machine learning model which is often not available. An approach is proposed which can be used to overcome this problem, such as extending the data by combining a large universal dataset and LDA topic modelling to discover hidden topics in a document. It is discussed how intent can be modelled without having a labelled dataset by using rule-based methods for automatically labelling the data. Thus, for future work, an approach for labelling a substantial amount of data should be applied since this was the main challenge that was faced in this thesis.

Furthermore, models for predicting user intent were selected based on availability: the BERT model and Ludwig model, both pre-trained models. Predicting intent faces the issue of data scarcity, however, both models are pre-trained aiming to solve this problem. Both models achieve high accuracy on a small dataset and, especially Ludwig, perform well on unbalanced data. These are beneficial features of both models. However, being pre-trained has disadvantages as it is harder to make specific adjustments and both models can not be merged. Both models are based on RNNs, since RNNs have been shown to significantly outperform many competitive language modelling techniques in terms of accuracy. Therefore, an RNN is often used for predicting user intent. The selected models performed well for predicting intent, however, for future work, it is argued an RNN model should be built for predicting intent.

Lastly, for this thesis, recognizing the intent was based on search queries, but for future work, it is argued other intent modelling approaches such as click behaviour, mouse movement or dwell time can be applied to make more accurate predictions.

References

[1] Ajith Abraham. Artificial neural networks. *Handbook of measuring system design*, 2005.

[2] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning user interaction models for predicting web search result preferences. pages 3–10, 01 2006.

[3] Steven M. Beitzel, Eric C. Jensen, Ophir Frieder, David D. Lewis, Abdur Chowdhury, and Aleksander Kolcz. Improving automatic query classification via semi-supervised learning. *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8 pp.–, 2005.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[5] Ludovico Boratto and Eloisa Vargiu. Data-driven user behavioral modeling: from real-world behavior to knowledge, algorithms, and systems. *Journal of Intelligent Information Systems*, 54(1):1–4, 2020.

[6] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM New York, NY, USA, 2002.

[7] Loredana Caruccio, Vincenzo Deufemia, and Giuseppe Polese. Understanding user intent on the web through interaction mining. *Journal of Visual Languages Computing*, 31:230–236, 2015. Special Issue on DMS2015.

[8] Qian Chen, Zhu Zhuo, and Wen Wang. BERT for joint intent classification and slot filling. *CoRR*, abs/1902.10909, 2019.

[9] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces, 2018.

[10] Vincenzo Deufemia, Massimiliano Giordano, Giuseppe Polese, and Genny Tortora. Inferring web page relevance from human-computer interaction logging. 2012.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Jonas Etzold, Arnaud Brousseau, Paul Grimm, and Thomas Steiner. Context-aware querying for multimodal search engines. In *International Conference on Multimedia Modeling*, pages 728–739. Springer, 2012.

[13] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[14] Charles T. Hemphill, John J. Godfrey, and George R. Doddington. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990.

[15] Irazú Hernández, Parth Gupta, Paolo Rosso, and Martha Rocha. A simple model for classifying web queries by user intent. In *Proc. 2nd Spanish Conf. Information Retrieval*, pages 235–240, 2012.

[16] Matthew Huggins, Sharifa Alghowinem, Sooyeon Jeong, Pedro Colon-Hernandez, Cynthia Breazeal, and Hae Won Park. Practical guidelines for intent recognition: Bert with minimal training data evaluated in real-world hri application. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21, page 341–350, New York, NY, USA, 2021. Association for Computing Machinery.

[17] Jim Jansen and Danielle Booth. Classifying web queries by topic and user intent. pages 4285–4290, 04 2010.

[18] Keith Jeffery, Antti Pursula, and Zhiming Zhao. *ICT Infrastructures for Environmental and Earth Sciences*, pages 17–29. Springer International Publishing, Cham, 2020.

[19] Hirotaka Kawazu, Fujio Toriumi, Masanori Takano, Kazuya Wada, and Ichiro Fukuda. Analytical method of web user behavior using hidden markov model. pages 2518–2524, 12 2016.

[20] Delanyo Kwame Bensah Kulevome, Hong Wang, and Xue gang Wang. A bidirectional lstm-based prognostication of electrolytic capacitor. *Progress in Electromagnetics Research C*, 109:139–152, 2021.

[21] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. page 391–400, 2005.

[22] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5528–5531, 2011.

[23] Piero Molino, Yaroslav Dudin, and Sai Sumanth Miryala. Ludwig: a type-based declarative deep learning toolbox, 2019.

[24] Piero Molino, Yaroslav Dudin, and Sai Sumanth Miryala. Ludwig: a type-based declarative deep learning toolbox, 2019.

[25] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[26] Michele Persiani and Thomas Hellström. Intent recognition from speech and plan recognition. In Yves Demazeau, Tom Holvoet, Juan M. Corchado, and Stefania Costantini, editors, *Advances in Practical Applications of Agents, Multi-Agent Systems, and Trustworthiness. The PAAMS Collection*, pages 212–223, Cham, 2020. Springer International Publishing.

[27] Xuan-Hieu Phan, Le Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text web with hidden topics from large-scale data collections. pages 91–100, 01 2008.

[28] Mihai Cristian Pîrvu, Alexandra Anghel, Ciprian Borodescu, and Alexandru Constantin. Predicting user intent from search queries using both cnns and rnns. *CoRR*, abs/1812.07324, 2018.

[29] Puyang Xu and Ruhi Sarikaya. Exploiting shared information for multi-intent natural language sentence classification. In *Interspeech*, pages 3785–3789, 2013.

[30] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *CoRR*, abs/1409.2329, 2014.