A Comprehensive Taxonomy of User Intents for Search Queries

Jasmine Diaconu
Supervisors: Gaole He, Ujwal Gadiraju
EEMCS, Delft University of Technology, The Netherlands
20-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Search engines operate as an oracle between user queries and information access: the user types the input and receives back the information requested. To accomplish the task, search engines need to interpret human language and, most importantly, comprehend the underlying user intents of a query. With this process, they can retrieve the most appropriate sources of information.

The purpose of our research is to introduce a new, hierarchical taxonomy that better depicts the underlying intents of users asking questions online (on search engines and Q&A platforms). Throughout our study, we first review the prior work and findings on the topic. We assemble a new dataset with queries aggregated from MS Marco, AskReddit and Quora. We examine its questions and label them to construct a new fine-grained ontology. Our examination continues with the integration of Deep Learning (DL) models and Active Learning (AL) to evaluate the quality of our work. The results show that the taxonomy can effectively assess users' goals. Our taxonomy, the dataset composed and the codebase are publicly available[1] to support future research.

1 Introduction

People browse the Internet to find new recipes, book flights, get help with homework, or acquire a new skill to apply for their dream job. The increasing need for information and the expansion of online resources lead to a new problem: topics need to be separated into categories, and the most relevant results should be supplied to the users. Efficient identification and classification of user goals aid search engines to provide accurate results, while also reducing the research time. However, one of the obstacles to achieving the goal is that user intents are dynamic and change over time based on technological advancement or switches of interests [1], [2].

The research question *How to categorize queries into user intents?* has been examined extensively over the years, and various studies endeavoured to organize searches based on the underlying goals of the users. The areas of Information Retrieval (IR) and Natural Language Processing (NLP) have been intensively explored, and numerous taxonomies have been elicited. Early studies, such as the one proposed by Broder [3], depict a trichotomy of online search classes: *transactional*, *navigational* and *informational*. Rose and Levinson [4] refined Broder's approach by further subdividing the informational and transactional categories. Subsequently, Li and Roth [5] suggested the first Machine Learning (ML) approach for semantic classification. Other ideas employ Deep Learning (DL) algorithms [6], nevertheless, there is no universal taxonomy that totally answers the problem. Users' needs are continuously evolving and taxonomies need to evolve with them. When there is no appropriate label for a user intent, a new one has to be created and the taxonomy needs to be revised.

With this paper, we intend to propose a sophisticated taxonomy that incorporates the features of the preceding efforts, while addressing the present needs of online users. We are also interested in answering the aforementioned research question and its sub-questions:

1. How are queries correlated and how can we group them into categories?
2. How can we derive a more representative taxonomy from the user inputs?
3. How can we evaluate the correctness of our taxonomy?

To answer Question 1, we examine three different datasets: MSMarco, Quora and AskReddit. We analyze 5,000 queries (50% from MSMarco, 25% from Quora and 25% from AskReddit) and detect a correlation between queries containing similar keywords, such as *meaning*, *definition* or wh-words [2].

Once the correlation between queries is clear, we start grouping queries based on the closure in user intents and strive to create a hierarchical structure for the discerned classes. This step serves as input for Question 2.

Here we review existing taxonomies to understand how closely they resemble the identified user intents. We recognize their limitations and attempt to compensate for those in our work.

As a response to Question 3, we adopt technologies such as Active Learning [7] and Deep Learning models, such as MLP (Multilayer perceptron) [8], LSTM (Long Short Term Memory) [9] and BERT (Bidirectional Encoder Representations from Transformers) [10] to assess the quality of our work. The aforementioned tools are known to be widespread in the IR and NLP fields, yet little attention has been paid to their application in user intent categorization.

The rest of the paper is structured as follows. In Section 2, a summary of the previous work and a comparison with the devised taxonomy is provided. In Section 3, the design of a novel taxonomy is presented together with the phases constituting its ideation and final evaluation. Section 4 follows with the Deep Learning models elaborated, the conducted experiments and their outcomes. In Section 5 we highlight further discussion on the findings and the limitations of the study. Section 6 is dedicated to the ethical implications and reproducibility of the results. Finally, the paper is concluded in Section 7 with ideas for future improvements.

2 Literature Review

The identification of user intents has been a topic under study since the early 2000s. Different systems for categorizing search queries have been proposed, the first being Broder's

[1] https://github.com/jasminediaconu/CSE3000-ResearchProject

[2] Wh-words are the class of words used to introduce a question. Common examples of wh-words are: why, who, which, what, where, when, and how.

taxonomy [3]. Broder classified queries based on two methods: user surveys and manual query log classification. However, an imbalance among the three groups has been discovered, with *transactional* queries being the most difficult to detect. In our work, we use Broder's classification as a baseline and extend it with an additional label, *Human* queries.

Rose and Levinson [4] renamed the Transactional category as *Resource* to enclose all queries related to resource retrieval: downloading files, searching for music or watching videos. Furthermore, the Informational category has been subdivided to achieve a more detailed classification. Rose and Levinson questioned the feasibility of deriving user intents without directly surveying the users. This led to an alternative perspective on intents extraction, by looking at user behaviour such as the query itself, search engine results, user clicks or related actions by the same user. As more sophisticated methods aroused, we employ a similar approach in collecting samples. We also strive to label them according to commonsense knowledge and query interpretation from an outsider's perspective. While our taxonomy differs in the labels we identify, we similarly consider a subdivision of the Informational queries. We first classify them based on the type of answer the user seeks for. A question such as: *When was Shakespeare born?* is expecting a date, namely Shakespeare's birthdate. However, a question such as: *How many credits are needed for a bachelor's degree?* asks for a specific quantity and *What did Germany do after the Treaty of Versailles?* is looking for a more elaborate answer. Based on these assumptions, we classify Informational queries into *Textual*, *Numeric* and *Other* to indicate the type of answer the user seeks for. While the first two categories are straightforward, the *Other* category comprises queries with alpha-numeric answers and/or special characters.

Lee et al. [11] led further research on the categorization process to switch from a manual process to an automated one. This has been done by filtering out the queries with predictable goals and extracting the user goals based on *past-click behaviour* and *anchor-click distribution*. While the obtained results appeared accurate, the queries used were exclusively technical, which led to a biased dataset. For the aim of our research, we also use unsupervised (thus automated) models. However, we randomly select samples from datasets tailored specifically for IR and NLP tasks. This results in a diverse set of samples and, therefore, a less biased body of information.

From the methodology perspective, Li and Roth [5] posed their effort in devising the first Machine Learning approach. Despite the prior studies, they elicited six coarse classes, namely *Abbreviation*, *Description*, *Entity*, *Human*, *Location* and *Numeric*. Similarly, we are also exerting ML models, which proves their capabilities and impact in the field.

While a number of studies has examined user inquiries on search engines, others have put the effort into targeting QA (Question Answering) queries. This is the case for Bu et al. [12] who developed a function-based classifier. Real questions were revealed to be more suitable for categorization and closer to natural language compared to search queries, which often involve the use of keywords. The proposed taxonomy grouped questions based on six types: *Fact*, *List*, *Reason*, *Solution*, *Definition* and *Navigation*. To decide the question's type, MLN (Markov Logic Network) was applied. This technique tries to match questions through probabilities, to the most suitable type. While we are not using MLN for our case of study, we are embedding other techniques such as MLP, LSTM and BERT. Our dataset composition also differs: we incorporate both WS and cQA queries, rather than favouring one type.

A more probabilistic approach was introduced with Probase [13], which attempts to classify queries based on their *plausibility* and *typicality*. The innovation of this framework was the use of probabilities and knowledge types to conduct the classification. These crucial aspects enabled Probase to be resilient to ambiguity and inconsistency in queries, unlike the prior taxonomies. While our case of study does not involve a probabilistic classification, we strive to reduce ambiguity by providing a clear description for each label. We also aim to lessen the classification disagreement between annotators, by considering the most acceptable label as the winner in a tie.

Additional research in the QA field conveyed an interesting discovery: questions do not always have the absolute best response [14]. Moreover, there are questions solely intended for conversation engagement. Therefore, the *Social* label has been introduced. We consider this discovery valuable for our taxonomy, and we rename the Social label to Human. We believe there is a broader group of questions that require answers from individuals. Some of those questions are trying to acquire diverse viewpoints on a topic: *Who is the best actor in the world?*, others are looking for guidance or recommendations relating to a personal situation: *How do I get better at storytelling?*. Users might want to learn about a specific occurrence encountered in others' lives: *What is it like to work in Japan?* and lastly, we have queries purely intended for human interaction: *What is your hobby?*. We label these four patterns as *Opinion*, *Advice*, *Experience* and *Engagement*, respectively.

To complete our taxonomy, we ramify the Informational trichotomy according to the examples proposed by Gupta et al. [6] and Cambazoglu et al. [15]. The categories we recover are *Quantification*, *Entity*, *Definition*, *Description*, *List* and *Language*, *Location*, *Process*, *Reason*, *Temporal*, *Weather*, *Money*, *Date*, *Duration*, *Percent*, *Range*, *Time*, respectively.

3 Method

In this section, we present the hierarchical classification we construct and explain the enclosed categories. We, then, continue with the setup and tools utilized. Therefore, we delineate the procedure undertaken for collecting, labeling and pre-processing the data. Lastly, we introduce the algorithms applied for categorizing the queries.

3.1 Taxonomy composition

The taxonomy is illustrated in Figure 1 and is made of four layers. The first one is split into:

- **Informational**: the intent is to find objective information about a topic. These questions are usually addressed to search engines;

- **Human**: the intent is to find subjective information about a topic. These questions are usually addressed to other users;
- **Transactional**: "the intent is to perform some web-mediated activity" [3];
- **Navigational**: "the immediate intent is to reach a particular site" [3];

We decide to distinguish between Informational and Human questions as we believe not all of them have the intent to receive a correct answer [14]. Informational queries are related to objective information, therefore, they are based on facts, observations, measurements or unbiased analysis. Human queries are related to subjective information: personal interpretations, opinions, feelings and thus, personal biases [16]. As an example, the question: *When did the Flintstones cartoon come out?* is fact-based and all individuals can agree on a specific date to identify the Flintstones release. Whereas, the question: *Where should I start when learning a new language?* shows that the user is interested in learning a new language and wants suggestions on how to accomplish it. Language-learning techniques are subjective: they can depend on prior knowledge of the language, similarity with the native language or age. Regarding the Transactional and Navigational categories, we decide to incorporate them in our taxonomy, however, we actually do not have a sufficient amount of samples to include them in the evaluation process. Therefore, we decide to repropose Broder's definition of those labels and center our analysis on a corpus of Informational and Human queries.

As for the second layer, we subdivide the Informational class into: *Textual*, *Numeric* and *Other*. Typically, users seek for different types of information. Understanding their needs and returning the correct type has the effects of reducing their research time. As an example, a user asking *What is the speed of light in kilometers/sec?* wants to quickly retrieve a number. When a user asks *How is solar energy used to generate electricity?* the search engine should propose an explanation. Whereas, the question *How to calculate surface coverage?* is asking for a formula, which is an alpha-numeric information.

The Human queries are split into *Opinion*, *Advice*, *Engagement* and *Experience*. As these questions are addressed to other users, it is important to target the appropriate audience. The question *What is the best new movie to watch?* is based on a personal opinion and anyone could answer it. A question like *How can I fit into the German lifestyle?* seeks for advice from users who are acquainted with German habits. A user might search for experiences, regarding a situation, so they can make a decision: *Have you stayed in a studio apartment with your family before? How was it like?*. Finally, there are questions with the pure scope of engaging into a conversation *Do you believe in having New Year's Resolutions?*.

The third layer of the taxonomy is occupied by Textual, Numeric and Other subcategories:

- **Textual**: Definition, Description, Type, List, Entity, Process, Explanation, Language, Comparison, Example, Selection (Table 4);
- **Numeric**: Percentage, Quantity, Duration, Money, Frequency, Age, Phone, Conversion (Table 3);

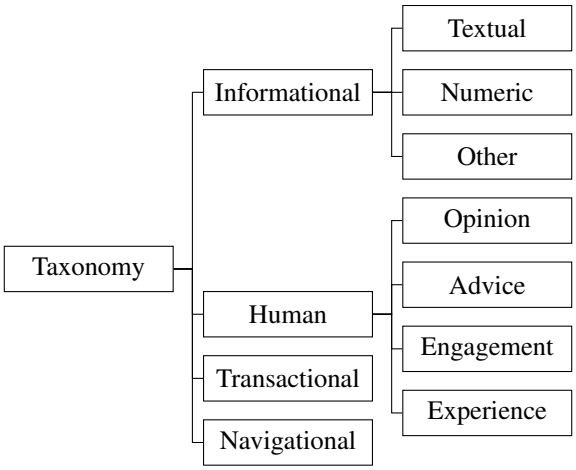

Figure 1: User intents taxonomy displaying the first and second layer constructed. The full taxonomy is available here.

- **Other**: Date, Range, Time, Boolean, Code and Formula (Table 5);

We do not mean to digress on their meaning, as we believe the naming we appoint is unequivocal. However, a full description and examples for each label are available as references in the appendix.

The fourth layer contains only subcategories of the *Entity* type (in the Textual category). We consider this distinction appropriate as there is a large corpus of such queries and they have characteristic patterns. The subclasses we consider are:

- **Location**: the intent is to retrieve a specific location. This can be a city, state, country, region or other geographical position;
- **Temporal**: the intent is to retrieve a defined sequence of time. This can be an epoch, an era, or a shorter period of time;
- **Person**: the intent is to identify a specific human, typically by retrieving their name;
- **Animal**: the intent is to identify a specific animal or species;
- **Weather**: the intent is to retrieve a particular climate;
- **Other**: everything that is considered as an entity, but cannot be placed in the aforementioned categories;

To zoom into some examples, Location questions can be structured as: *What [continent/country/state/region] is [name of the place] in?*. Temporal queries can follow the pattern: *What century is/was [name]?* Person queries are typically phrased as: *Who is/was [name]?*, while Animal queries can be: *What animal is [name]?*. An example of Weather questions can be: *How is the weather in [name of the place]?*. Finally, all queries that are not placed in the previous categories, fall into Other.

3.2 Setup

Datasets: we gathered samples from MS Marco [17], Quora and AskReddit. The first two datasets are retrieved from BEIR [18], a benchmark tool to assess NLP-based retrieval models. The AskReddit dataset is collected from HuggingFace [3]. While the first dataset adds to the WS (Web Search) samples, the other two make up for the cQA (Community Question Answering) queries. All samples come from real questions typed by users. Further details on the datasets can be found in Table 1. We select samples from these sources for two main reasons:

- The datasets capture different types of user intents. MS Marco mainly contributes to the Informational queries, given by traditional search methods. Quora and AskReddit contain more Human queries, thus resonate with the engagement with social networks [19], [20];
- The datasets are designed and used for IR and NLP tasks in previous studies [15], [18], [21];

Annotators: two annotators label the queries. Both annotators are aware of the categories of the taxonomy and the definition of each label.

Platform: labeling is performed with spreadsheet software. The annotators label the queries individually and, in case of disagreement, a discussion is undertaken to agree on the most suitable label for the ambiguous queries.

Software and other tools: We use Jupyter notebooks [22] to analyze the data distribution, preprocess the queries, train models and evaluate the taxonomy. The programming language in use is Python [23] along with a few libraries, such as:

- **Scikit-learn [24]:** used for label encoding;
- **TensorFlow [25]:** for text preprocessing, tokenization, transformers and model training;

Additionally, we use Paperspace[4] for the notebooks execution.

Dataset	Total # queries	Samples collected	Year	Type of query
MS Marco	509,962	2,500 (50%)	2021	WS
Quora	15,000	1,250 (25%)	2021	cQA
AskReddit	1,000,000	1,250 (25%)	2021	cQA

Table 1: The datasets used for sample collection and their data distribution.

3.3 Procedure

The procedure we adopt can be subdivided into two phases:

1. **Phase one:** dedicated to the research and examination of previous work on user intents taxonomies;
2. **Phase two:** dedicated to the process of collecting samples, labeling them, constructing the taxonomy and evaluating it;

[3]https://huggingface.co/datasets/SocialGrep/one-million-reddit-questions

[4]https://paperspace.com/

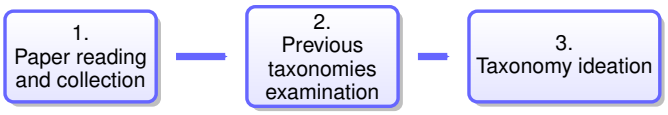

Figure 2: The steps of **Phase one**: previous work examination and ideas for a new taxonomy.

Phase one

We design the taxonomy by considering three main steps, as illustrated in Figure 2:

1. **Paper reading and collection:** to learn more about the research topic, we are initially given two anchor papers [26], [27]. We derive the relevant keywords from them and start searching for related papers on taxonomies, search queries, user intents and classification models. We collect the papers and organize them by topic, to best select the ones inspiring our work.
2. **Previous taxonomies examination:** we examine the taxonomies gathered and compare them. We then try to extract their shared categories to integrate them into our work. We also identify the classification models used and the process brought to their ideation to use as reference.
3. **Taxonomy ideation:** once the previous work is clear, we start drafting ideas for a new taxonomy. We agree on different aspects:
   - **Taxonomy structure**: it can be faceted or hierarchical. Our taxonomy is hierarchical;
   - **Number of layers**: four layers in total;
   - **Amount of samples to collect**: we collect a sample of 5,000 queries;
   - **Type of queries**: previous studies focus on either WS or cQA queries. We collect a combination of both;
   - **Datasets**: we collect samples from MS Marco, Quora and AskReddit;

Phase two

Phase two is illustrated in Figure 3 and follows nine steps:

1. **Sample collection:** the sample queries are randomly picked from the datasets. We, then, filter out inappropriate, irrelevant or ambiguous queries. For example, we decide to completely discard queries containing multiple questions, as they cannot fit in a single category of our taxonomy. We also agree on merging the utterances into a single dataset to reduce labeling bias.
2. **Taxonomy elicitation:** once the samples are ready, the first annotator analyzes them and drafts the initial taxonomy. In the first draft four main categories are present: *Informational*, *Human*, *Transactional* and *Navigational*. However, due to the scarcity of Transactional and Navigational queries, we examine only Informational and Human queries. After adding the first label, we identify new patterns among queries, and ramify the taxonomy furtherly, to efficiently capture the user intents.

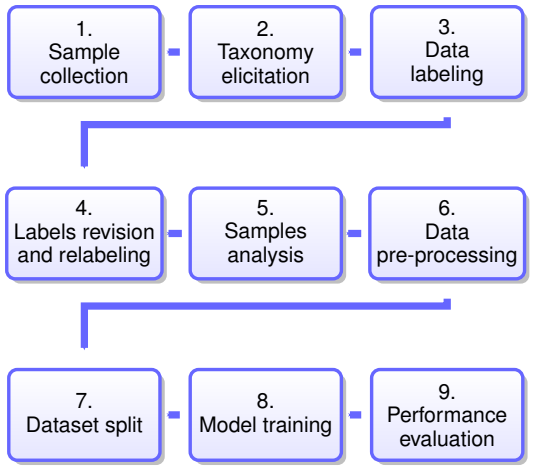

Figure 3: The steps belonging to **Phase two**: data processing, taxonomy creation and evaluation.

3. **Data labeling:** with the taxonomy in mind, the annotators label the samples independently.
4. **Labels revision and relabeling:** after the labeling step, the annotators discuss the samples with divergent labels. The discussion brings two viable solutions: an agreement on the label is reached or the query is discarded from the selection and replaced with a new one.
5. **Samples analysis:** we analyze the label distribution of the queries to ensure that no label is underrepresented. Two cases follow:
   - Undersampled labels for which we can collect additional samples;
   - Undersampled labels for which we can not find additional samples;

   In the first case, we increase the number of samples for the label, whereas in the second one we completely discard it from the classification.
6. **Data pre-processing:** during this phase we clean up the dataset before the DL models are applied. The preprocessing step involves:
   - Lower-case all words inside the query;
   - Spelling correction;
   - Punctuation and special characters removal;
   - Labels encoding;
7. **Dataset split:** we propose three different strategies to split our dataset before training the models. We call the first scenario *full dataset* as we use all the collected samples to compose the training, validation and test sets. The second and third scenarios involve the application of Active Learning (AL) and we refer to them as *AL with uncertainty sampling* and *AL with random sampling*.

   AL is a widespread strategy when unlabeled data is abundant, yet manual labelling is expensive. There are various types of AL, the one presented in our paper is the *Pool-Based Sampling* illustrated in Figure 4. The tactic works under the premise that there is a small pool of labeled data $\mathcal{L}$ and a large pool of unlabeled one $\mathcal{U}$, such that $|\mathcal{L}| \ll |\mathcal{U}|$ [7]. Moreover, the labeled pool comprises a consistent number of examples across the labels, which are fed to the AL model. Thus, the model is able to select the most informative instances from the unlabeled pool, and query an oracle [5] to obtain their actual labels. Therefore, the new samples are added to the labeled pool and the model gets trained on the new set. The procedure continues iteratively until a stopping criterion is met.

   In our study, we apply AL only to the first layer of the taxonomy, to emphasize its potential. Whereas, we apply the full dataset split to all layers.

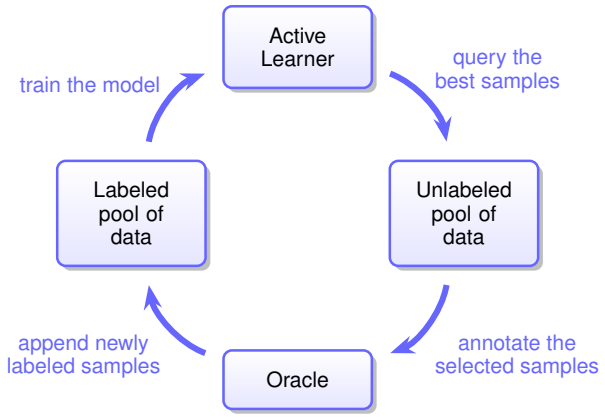

Figure 4: The Pool-based Sampling employed.

8. **Model training:** during this phase we train three different DL models over the taxonomy: MLP, LSTM and BERT. They are all known to perform well on text classification problems [28], [29], [30]. Therefore, they are powerful for backtracking user intents. The aim of training multiple models on the taxonomy is to reduce model bias and understand the benefits and drawbacks of favouring one model over the other.
9. **Performance evaluation:** once the models are trained, we compare their performance. We have nine options to compare, based on the cartesian product between each model M (MLP, LSTM or BERT) and each strategy S (AL with uncertainty sampling, AL with random sampling or full dataset): $M \times S$. The metrics we report are Loss, Accuracy, False Negatives and False Positives. A detailed description on the performance evaluation follows in Section 4.

4 Experiments and Results

In this section we review the data distribution over the layers, and describe the dataset splits for the models training. We continue with the technical setup of each model. Finally, we elicit the performance results and their meaning.

[5] In Active Learning, an oracle is a human annotator that provides the labels of the samples to the Machine Learning model.

4.1 Experimental Setup

The experimental setup and results refer to the first layer of the taxonomy. Each sub-layer contains a limited amount of samples ($< 5,000$ queries) and applying AL for each of them would be costly. This is why we train only the full dataset on the other layers. The experiments are executed on Paperspace: we use a setup of 8 CPUs, 30GB of RAM and a Quadro M4000 graphic card with 8 GB of GPU.

Dataset distribution

For each label of the taxonomy, we ensure to have at least 25 examples. When the threshold is not met, we randomly select new queries (from the initial datasets) and try to increase the number. When the operation fails, we discard the label. The procedure is applied to each layer of the taxonomy. In total, we identify 2,900 Informational queries (58%) and 2,100 Human queries (42%). While the first layer is fairly balanced, this is not the case for the sub-layers. For the Human layer, we have a distribution of 26% Advice, 29% Engagement, 12% Experience and 33% Opinion. The Informational queries are distributed as 24% Numeric, 59% Textual and 17% Other.

The third layer is also unbalanced: the Textual queries Example, Selection, Purpose, Language, Process, Composition and Description account for 2-6% each, whereas we have 9% as List, 10% as Explanation, 18% as Definition and 35% as Entity. For Numeric queries we have a total of 31% spread over Frequency, Percentage, Phone, Conversion and Age. Another 15% is composed by Duration, 24% by Quantity and, finally, 30% by Money. In the Other category, Boolean samples account for 43% of the total, Date for 15%, Code for 12%, Range for 11%, Formula for 10% and Time for 9%. Finally, the fourth layer has a distribution of 34% as Location, 32% as Other, 18% as Person, 9% as Animal and 7% as Temporal.

As a consequence of the data imbalance, the results of the experiments (Section 4.2) and their discussion (Section 5.1) refer to the first layer only. The results we obtain from the other layers can be found in the code repository and are left for the reader's curiosity.

Dataset splits

As described in Section 3.3, we adopt three different dataset splits and compare their impact on the model performance. In the first case, we use the full dataset strategy. We apply a 70%-10%-20% split to the dataset into training set, validation set and test set, respectively. The second case involves AL with uncertainty sampling. Here, we first separate positive and negative samples and then we evenly distribute them over three sets: train set (100 samples), validation set (500 samples), test set (1,000 samples). The leftover samples are kept in two separate sets, namely the *pool of positive samples* and the *pool of negative samples*. Separating positive and negative samples (which correspond to Informational and Human queries) allows the model to learn the label it comprehends the least. We compute the quantity of negative and positive samples to be selected based on the ratio[6] of False Negatives and False Positives:

$$Negative\ samples : \frac{False\ Negatives}{All\ False\ Predictions}$$
$$Positive\ samples : \frac{False\ Positives}{All\ False\ Predictions}$$

AL with random sampling has the same ratio of samples (across the three sets) as uncertainty sampling. However, instead of separating positive and negative samples, we place them in a single *pool of unlabeled samples*. Therefore, at each iteration we randomly select 100 queries from the pool and add them to the train set, without considering the positive and negative ratio.

MLP model configurations

To configure the MLP model we set both batch size and maximum sequence of tokens to 128. We define a sequential model for the MLP to ease the neural network construction. We then apply *max pooling* to prevent overfitting and reduce the number of parameters the model needs to learn. Finally, we apply two dense layers: the first one of 12 units with the *relu* activation function and the second of 1 unit and *sigmoid* activation function. To evaluate the model, we use the *binary cross-entropy* loss function and *binary accuracy* as the accuracy metric. The optimizer used is *rmsprop* and the amount of epochs is set to 5.

LSTM model configurations

The configurations of the LSTM model is identical to the MLP one, aforementioned. In addition to that, it comprises a bidirectional LSTM layer of 32 units. We consider similar setups to ease the performance comparison between the two models.

BERT model configurations

There are different BERT pre-trained models available, in our case we employ $\text{BERT}_{\text{BASE}}$ uncased[7]. We opt for the uncased version as it lower-cases the text and removes accent marks, as cased information is not needed for our task. Additionally, working with a pre-trained model ensures a solid performance and it spares the time of manually normalizing, cleaning up and tokenizing the queries' corpus. To complete the pre-processing phase, we encode each word with a maximal sequence length of 128 tokens. We determine that this setting, together with a batch size of 32, yields the best results. Finally, we exclude the BERT layer from the training due to the elevated time and memory cost that the operation entails. Once the model is ready, we apply the *rmsprop* optimizer. We take inspiration from the analysis presented in [31] to determine the hyper parameter tuning configurations.

4.2 Results and performance

We present the results obtained in Table 2. To measure the quality of the models, we apply different metrics such as Loss, Accuracy, False Negatives and False Positives.

[6] https://keras.io/examples/nlp/active_learning_review_classification

[7] https://huggingface.co/bert-base-uncased

Loss: refers to the output of the loss function. A loss function quantifies the error between the predicted label and the true label of the sample. The loss function we apply is *binary cross-entropy* which is tailored for classification problems with two labels. When the loss is close to 1, the error between predicted label and actual label is high. The error is low when the loss is closer to 0.

Accuracy: is the ratio[8] between correct predictions and total predictions the model performs. As for the loss, we employ *binary accuracy*: the closer the value is to 1, the higher is the percentage of correctly predicted labels.

False Negatives: refers to the number of samples that are incorrectly classified as "0". In our case, this corresponds to the amount of Informational queries that are incorrectly classified as Human.

False Positives: refers to the number of samples that are incorrectly classified as "1". In our case, this corresponds to the amount of Human queries that are incorrectly classified as Informational.

The results displayed in Table 2 show that BERT is the model with the best performance overall. In all three scenarios, its loss is closer to 0 and accuracy is closer to 1 compared to its competitors. Likewise, the total sum of misclassified utterances is smaller when using BERT compared to the other models. However, BERT's high performance comes with a cost: while MLP and LSTM take on average 1 min of execution, BERT takes about 72 min.

5 Discussion

In this section we delve into the meaning of the results obtained and the limitations of the experiments undertaken.

5.1 Results implication

When analyzing the results there are two points of consideration: the model in use and the sampling strategy. From the model perspective, BERT is outperforming MLP and LSTM. BERT is using *masked language model* (MLM) to extract the bidirectional context when performing NLP processing. It also uses *next sentence prediction* (NSP) to concatenate phrases and predicting their sequence. These features increase the time complexity of the model, however, they ensure a better comprehension of human language. Despite of using bidirectional LSTM, the results still show a performance gap between LSTM and BERT. This is because bidirectional LSTMs are separately capturing left to right and right to left context, whereas BERT is integrating both simultaneously. We also notice that the performance of MLP and LSTM does not differ drastically when employing AL with random sampling or full dataset. This is expected, as the two models have similar configurations. Surprisingly, a consistent outcome is not observed when applying uncertainty sampling. A possible explanation for this divergence is that MLP follows unidirectional context, which can cause overfitting when calculating uncertainty. Contrastingly, the bidirectional context of LSTM can make the model resilient to such performance decrease.

Across the strategies, the best scores are achieved by the full dataset. This is expected since the training dataset is larger. We observe a minor accuracy variance among the strategies. The major difference between employing AL or not is visible in the loss value and the sum of incorrectly labeled samples. However, the performance gap of AL in this scenario is justified by an 86% reduction in the number of samples to label.

5.2 Limitations of the research

While the results demonstrate the potential of the tools in use, further analysis needs to be conducted to ensure their reliability. Due to the time constraints, we collect a sample of 5,000 queries, leaving out transactional and navigational queries. A larger corpus of data could assess the effectiveness of the taxonomy and the performance of the algorithms. Additionally, some labeling bias can be introduced by the limited amount of annotators (only two) at our disposal. The current models are not optimized for detecting transactional and navigational queries. Without this consideration, we cannot ensure that they would predict these unseen labels. Moreover, the AL strategy is not working as expected when combined with BERT: while the final test accuracy and loss are ideal, the ones on the train and validation set drop at each iteration (see BERT implementation[9]). Moreover, we focus our analysis on the first layer of the taxonomy. A thorough examination, involving the application of different models and sampling strategies, should be conducted for the other layers as well.

6 Responsible Research

The research is conducted in compliance with the Netherlands Code of Conduct for Research Integrity [32] to guarantee the validity of the results and the trustworthiness of our study. The ethical concerns are proposed in Section 6.1, whereas the reproducibility of our results is discussed in Section 6.2.

6.1 Scientific Integrity

In light of the scientific integrity, the datasets collected and their use conform to the rights and permissions attributed by their original authors. The samples collected from MS Marco are real, anonymized user queries. They are collected from Bing's search logs and do not contain private information that could backtrack the users. The queries retrieved from Quora and AskReddit are publically available on the respective platforms. Moreover, the sample collection is randomized to avoid bias in selecting given samples. To avoid plagiarism, we cite all literature inspiring our work by following the IEEE style. We also provide references to libraries, tools and third-party code that constitute the foundation of our experiments.

[8] Accuracy formula: https://developers.google.com/machine-learning/crash-course/classification/accuracy

[9] BERT with uncertainty sampling: here, and BERT with random sampling: here.

Model	Strategy	Loss	Accuracy	False Negatives	False Positives
MLP		0.51	0.73	269	0
LSTM	AL with uncertainty sampling	0.33	0.90	3	96
BERT		0.12	0.97	33	36
MLP		0.33	0.89	46	65
LSTM	AL with random sampling	0.32	0.88	82	43
BERT		0.18	0.96	31	42
MLP		0.34	0.89	38	74
LSTM	Full dataset	0.25	0.92	51	29
BERT		0.04	0.99	9	10

Table 2: Comparison between MLP, LSTM and BERT models over three different sampling strategies.

6.2 Research Reproducibility

To support the reproducibility of the results, the taxonomy, the complete code base and the dataset utilized are available on a public GitHub repository. We thoroughly explain the research method in Section 3 and the experimental configurations and results in Section 4. Moreover, the algorithms we propose are integrating random seeds to guarantee the consistency of the results over multiple iterations. Thus, all the experiments we conduct are reproducible.

7 Conclusions and Future Work

In this paper, we collect WS and cQA queries, group them according to the underlying user intents and derive a novel taxonomy. We also compare our efforts to related work and underline the differences and similarities with our study. We, then, propose three deep learning models: MLP, LSTM and BERT to evaluate the effectiveness of the taxonomy. Follow-up work on the research should aim to fill in the gaps mentioned in Section 5.2. More specifically, the dataset could be enlarged and include a representative amount of transactional and navigational queries. Moreover, attempts to balance the data distribution would improve the accuracy of the models adopted. Additional considerations would be broadening the number of models utilized in the evaluation process and the metrics involved. In this work, we examine the performance of the AL model with uncertainty sampling and random sampling. The formula for the uncertainty sampling strategy could be revised, as suggested by [33]. Additional work may involve the adoption of unexplored strategies such as In-Context Learning [34] and Few-Shots Learning [35], [36]. A similar procedure might be followed for the models applied: in our research, we utilize MLP, LSTM and $\text{BERT}_{\text{BASE}}$ uncased. In the future, other deep learning models and different hyper-parameter tuning strategies might yield to a more precise classification.

References

[1] J. Jansen and D. Booth, "Classifying web queries by topic and user intent," 04 2010, pp. 4285–4290.

[2] A. B. Siddique, F. T. Jamour, L. Xu, and V. Hristidis, "Generalized zero-shot intent detection via commonsense knowledge," *CoRR*, vol. abs/2102.02925, 2021. [Online]. Available: https://arxiv.org/abs/2102.02925

[3] A. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, p. 3–10, September 2002. [Online]. Available: https://doi.org/10.1145/792550.792552

[4] D. E. Rose and D. Levinson, "Understanding user goals in web search," in *Proceedings of the 13th International Conference on World Wide Web*, ser. WWW '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 13–19. [Online]. Available: https://doi.org/10.1145/988672.988675

[5] X. Li and D. Roth, "Learning question classifiers: the role of semantic information," *Natural Language Engineering*, vol. 12, no. 3, p. 229–249, 2006.

[6] D. Gupta, R. Pujari, A. Ekbal, P. Bhattacharyya, A. Maitra, T. Jain, and S. Sengupta, "Can taxonomy help? improving semantic question matching using question taxonomy," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 499–513. [Online]. Available: https://aclanthology.org/C18-1042

[7] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[8] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958. [Online]. Available: http://dx.doi.org/10.1037/h0042519

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[11] U. Lee, Z. Liu, and J. Cho, *Automatic Identification of User Goals in Web Search*. New York, NY, USA: Association for Computing Machinery, 2005, p.

391–400. [Online]. Available: https://doi.org/10.1145/1060745.1060804

[12] F. Bu, X. Zhu, Y. Hao, and X. Zhu, "Function-based question classification for general QA," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, Oct. 2010, pp. 1119–1128. [Online]. Available: https://aclanthology.org/D10-1109

[13] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 481–492. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/2213836.2213891

[14] Y. Liu, S. Li, Y. Cao, C.-Y. Lin, D. Han, and Y. Yu, "Understanding and summarizing answers in community-based question answering services," in *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, ser. COLING '08. USA: Association for Computational Linguistics, 2008, p. 497–504.

[15] B. B. Cambazoglu, L. Tavakoli, F. Scholer, M. Sanderson, and B. Croft, "An intent taxonomy for questions asked in web search," in *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, ser. CHIIR '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 85–94. [Online]. Available: https://doi.org/10.1145/3406522.3446027

[16] "Subjective" vs. "Objective": What's the difference? Accessed: Jun. 17, 2022. [Online]. Available: https://www.dictionary.com/e/subjective-vs-objective/

[17] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "Ms marco: A human generated machine reading comprehension dataset," November 2016. [Online]. Available: https://www.microsoft.com/en-us/research/publication/ms-marco-human-generated-machine-reading-comprehension-dataset/

[18] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: https://openreview.net/forum?id=wCu6T5xFjeJ

[19] M. R. Morris, J. Teevan, and K. Panovich, "What do people ask their social networks, and why? a survey study of status message qamp;a behavior," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1739–1748. [Online]. Available: https://doi.org/10.1145/1753326.1753587

[20] Z. Liu and B. J. Jansen, "A taxonomy for classifying questions asked in social question and answering," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1947–1952. [Online]. Available: https://doi.org/10.1145/2702613.2732928

[21] H. Wu, W. Wu, M. Zhou, E. Chen, L. Duan, and H.-Y. Shum, "Improving search relevance for short queries in community question answering," in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, ser. WSDM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 43–52. [Online]. Available: https://doi.org/10.1145/2556195.2556239

[22] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing, "Jupyter notebooks – a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt, Eds. IOS Press, 2016, pp. 87 – 90.

[23] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[26] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the user intent of web search engine queries," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 1149–1150. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/1242572.1242739

[27] C. Kofler, M. Larson, and A. Hanjalic, "User intent in multimedia search: A survey of the state of the art and future challenges," *ACM Comput.*

Surv., vol. 49, no. 2, aug 2016. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/2954930

[28] M. Zhang, "Applications of deep learning in news text classification," *Scientific Programming*, vol. 2021, 2021.

[29] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, and J. W. Kim, "Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism," *Applied Sciences*, vol. 10, no. 17, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/17/5841

[30] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," *CoRR*, vol. abs/2005.13012, 2020. [Online]. Available: https://arxiv.org/abs/2005.13012

[31] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *China National Conference on Chinese Computational Linguistics*. Springer, 2019, pp. 194–206.

[32] N. T.-f. V. H. KNAW, NFU and VSNU, "Netherlands code of conduct for research integrity (2018)," 2018.

[33] V.-L. Nguyen, M. H. Shaker, and E. Hüllermeier, "How to measure uncertainty in uncertainty sampling for active learning," *Mach. Learn.*, vol. 111, no. 1, p. 89–122, jan 2022. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1007/s10994-021-06003-9

[34] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma, "An explanation of in-context learning as implicit bayesian inference," 2021. [Online]. Available: https://arxiv.org/abs/2111.02080

[35] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[36] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

A Appendix

Label	Description	Examples
Percentage	The user is seeking for a ratio in proportion to a whole.	*- what is the fatality rate of ebola* *- canada fresh water percent*
Quantity	The user is seeking for the amount of something.	*- average number of tornadoes per year* *- how many carbs in organic gummy bears*
Duration	The user is seeking for the time duration of a particular action.	*- how long does it take to paint a fire hydrant* *- length of time for short-term memory*
Money	The user is asking questions related salaries or prices.	*- current egg price in philippines* *- average house cost 2016*
Frequency	The user is seeking the number of occurrences of a certain action or event.	*- exchange brake fluid how often* *- average adult heart rate*
Age	The user is asking age-related questions.	*- what age is passport required* *- how old is raymond romano*
Phone	The user is seeking for a specific phone number.	*- boost mobile care number* *- support number for verizon*
Conversion	The user is seeking for the conversion of a number from one unit to another.	*- how many cups in a shot* *- what's one dollar equal in pesos*

Table 3: Numeric category sub-classification

Label	Description	Examples
Definition	The user is seeking for the meaning of a word or a generic idea about a topic.	*- what is autism?* *- phytonutrients definition* *- what is the dragon dance*
Description	The user is seeking for a detailed account about a certain topic that they already have a generic idea about.	*- what does tobacco smell like* *- what is the managers role in choosing equipment?*
Type	The user is seeking for the category or type a word belongs to.	*- what bonds are polymers* *- what type of eruption was the mayon volcano*
List	The user is seeking for a list of elements.	*- who were the opponents of the new deal* *- what are the most commonly used batteries*
Entity	The user is seeking for a specific name of something they can describe or they have some background knowledge about.	*- who directed doctor strange love?* *- what is the language of kenya?*
Process	The user is seeking for a series of actions or steps taken in order to achieve a particular end.	*- how to grill raw shrimp on the grill* *- onenote how to get a link to a page*
Explanation	The user is seeking for a statement or account to clarify a certain topic or a reason/justification given for an action or belief.	*- why were the thirteen colonies founded* *- why are proteins important nutrients*
Language	The user is seeking for a translation in a different language, the semantic meaning of a word or the correct grammatical structure of a sentence.	*- how to say happy birthday in french language* *- another name for glass noodles*
Comparison	The user is seeking for an estimate of the similarities and dissimilarities between two or more things.	*- what is difference between debt and liability* *- what do random assignment and matching have in common?*
Example	The user is seeking for an instance of a specific category or rule.	*- example of jargon sentence* *- what is a correct guiding principle of crm*
Selection	The user is seeking for the correct option given a few alternatives or a description.	*- _____ is an indirect technique used by interest groups to influence government policy.* *- are eggs or grapes better to fight colds*

Table 4: Textual category sub-classification

Label	Description	Examples
Date	The user is seeking for a date related to a specific event.	- *when is cesar chavez holiday* - *when did porter wagner die*
Range	The user is seeking for a range of items or numbers.	- *what years was korean war* - *hilux price range*
Time	The user is seeking for a time related to a specific event.	- *what is the time for canada* - *conn what time zone*
Boolean	The user is asking a questions with a yes or no answer.	- *did princess margaret marry* - *is google chrome java enabled*
Code	The user is seeking for the code of a specific entity.	- *airport code mont tremblant* - *icd code early pregnancy*
Formula	The user is seeking for a scientific or mathematical rule.	- *how do you calculate gpa*

Table 5: Other category sub-classification