Improving search results with machine learning
- Classifying multi-source data with supervised machine learning to improve search results

Author: Meri Stakovska
Supervisor: Johan Hagelbäck
External supervisor: Hans Runehov
Semester: Spring 2018
Subject: Computer Science
Abstract

Sony’s Support Application team wanted an experiment to be conducted by which they could determine if it was suitable to use Machine Learning to improve the quantity and quality of search results of the in-application search tool. By improving the quantity and quality of the results the team wanted to improve the customer’s journey.

A supervised machine learning model was created to classify articles into four categories; Wi-Fi & Connectivity, Apps & Settings, System & Performance, and Battery Power & Charging. The same model was used to create a service that categorized the search terms into one of the four categories. The classified articles and the classified search terms were used to complement the existing search tool. The baseline for the experiment was the result of the search tool without classification.

The results of the experiment show that the quantity of articles did indeed increase but due mainly to the broadness of the categories the search results held low quality.

Keywords: Searcher Frustration, Information Retrieval, Search Results, Topic Classification, Machine Learning, Supervised Classification, Naive Bayes
Preface

Thanks to the Support Application team at Sony in Lund for the welcoming and supportive energy they contributed with during the course of my thesis project. Special thanks to my supervisor at Sony, Hans Runehov, for his encouragement, positive spirit and for being an inexhaustible source of ideas during our brainstorming sessions.
Contents

1 Introduction 4
  1.1 Background 5
    1.1.1 Artificial Intelligence 5
    1.1.2 Machine Learning 5
    1.1.3 Supervised Classification 6
    1.1.4 Sony Mobile Communications Support Application 6
  1.2 Related work 6
  1.3 Problem formulation 7
  1.4 Motivation 7
  1.5 Objectives 8
  1.6 Scope/Limitation 8
  1.7 Target group 9
  1.8 Outline 9

2 Method 10
  2.1 Categories 10
  2.2 Measurement control 10
  2.3 Tools 12
  2.4 Reliability and Validity 12

3 Implementation 13
  3.1 Data 13
  3.2 Model 14
  3.3 Prediction 14

4 Results 16

5 Analysis 19

6 Discussion 21

7 Conclusion 22
  7.1 Future work 22

References 23

A Qualitative Comparison 25
  A.1 Baseline example 25
  A.2 Mixed example 28
1 Introduction

Have you ever found yourself struggling to find the right keyword because a search engine does not seem to understand just what you are looking for? You are definitely not alone. Numerous studies have been conducted to examine the concept Searcher Frustration (SF), i.e. the frustration that occurs when dealing with obstruction of progress in Information Retrieval (IR). [1]

In these studies, the obstruction of progress could be manifested as no search results for a keyword search or results that were not relevant to the user. [1] Subsequently the ability to recover from SF varied heavily and the emotional state of the user quite often resulted in search abandonment. The younger population especially was much less prone to continue querying after being affected [1].

Sony’s Support Application (SA) team aims to make their customers journey’s easier by relieving them of SF when using the in-application search tool. With this project we aim to contribute to the improvement by hopefully minimizing the occurrence of no, little or irrelevant search results. This is to be done by classifying the data that is to be presented as search result. Classification is to be conducted because topicality is one of the most distinguishable concepts for determining relevance of content [2].

1.1 Background

This degree projects aims to improve the search result in the SA by using an existing category structure to classify new data. In the subsections that follow concepts that will pervade the project will be explained briefly.

1.1.1 Artificial Intelligence

Artificial Intelligence (AI) is a manmade entity that is capable of thought process, reasoning, and behavior. These actions can be formed off of a human as well as a rational model. [3]

Some contemporary examples of AI are:

- self-driving cars that have learned to obey traffic rules, avoid pedestrians and navigate in traffic at large [3]
- computer opponents in games that have learned rules and strategies turning them into serious competition for human players [3]
- software that machine translates text content to video content [4]
- software that transforms raw data into stories and articles [5]

1.1.2 Machine Learning

Machine Learning (ML) is a subset of the vast field of AI and aims at programming a machine or a system in such a manner that it learns to self-improve [6]. The process of self-improvement can be seen as a learning problem that can be divided into tasks (T), performance measure (P) and training experience (E) [6]. This can be exemplified with the following:

**Treadmill speed learning problem**

\[ T = \text{automatically set speed on treadmill based on human pulse} \]
\[ P = \text{meters progressed before error occurs} \]
\[ E = \text{data from pulse sensors and max values from database} \]
**Emotion detector learning problem**

\[ T = \text{recognizing human emotions in photographies} \]
\[ P = \text{percentage of emotions correctly classified} \]
\[ E = \text{a database of photographies with classifications} \]

Figure 1.1: Learning problems

After dividing the learning problem into these components (figure 1.1) one should choose a target function and a representation for the target function i.e. determine exactly what knowledge is to be used and how it will be used to improve performance [6].

The following step is to choose an algorithm that will enable the learning from the training examples and estimating training values. Thereafter the adjustments are made until the outcome of the training is satisfactory [6].

1.1.3 **Supervised Classification**

Using prior information to categorize data means using a supervised learning technique. A supervised learning technique implies that the training data that will be supplied will hold predefined categories. These categories will be used to create a model that will classify data from other sources in accordance with the categories of the training data [7].

There are numerous classification algorithms such as Artificial Neural Networks, Bayesian Network, Decision Trees, k-Nearest Neighbors, Logistic Regression, Random Forests, Support Vector Machine [7].

1.1.4 **Sony Mobile Communications Support Application**

Sony’s smartphones, tablets, and projectors feature a preloaded self-support application called Support. SA offers users insight about their specific devices, e.g. IMEI number, software updates. It contains information about features e.g. articles on device-specific applications such as Camera, Email and it also provides users with means by which they can come in contact with customer support by e.g. chat, email or phone call.

SA features a search tool which enables the user to retrieve information by entering a search term instead of browsing the application. The information is retrieved from a backend service via API endpoints. The result is collected from across several sources, independent of each other in form as well as content. The data retrieved by this search tool will be the focus of this thesis.

1.2 **Related work**

This degree project overlaps two research areas: topic classification (TC) and IR. Related studies that overlap these two areas have conducted on Search Engines for Books, News articles, and Microblogs [8] [9] [10] [11].

A study conducted by T. T. S. Nguyen compares a book search using keyword search alone and a book search that makes use of word-vector comparison. Word-vector comparison implies that all words in a document are vectorized and the distance between them is computed. Thereafter the word-vector representation is used to predict the context of the word. By doing so the study concluded that the book search with word-vector comparison gave more relevant results than when using keywords alone [8].
M. Scharkow conducted an experimental study on classification of German online news articles where the articles were thematically classified using a supervised machine with a Naive Bayes (NB) classifier. The outcome of the study shows that supervised automatic coding is about 15% less reliable than manually coding themes. In most cases, it is robust and reliable. The machine learning approach had the drawback of encountering difficulties with contextual knowledge. In general, Scharkow writes that most preprocessing steps did not improve supervised classification, exception for the exception of text extraction for web content [9].

In a study in news articles by Z. Li, W. Shang and M. Yan a news text classification model based on Supervised Latent Dirichlet Allocation (SLDA) is proposed. SLDA is based on a probability model that identifies topics in a large data set. The classification is effectual but there were some shortcomings, such as the choices made for the topic model [10].

A study made on topic classification of Microblogs (namely Sina Weibo and Twitter) examined how to retrieve more relevant results when searching by classifying data. The researcher encountered some issues with the quick introduction of new topics on microblogs that a traditional model struggles with and introduced data from external search engines to complement the training model in order to solve this. Their semi-supervised Bayesian network (SSBN) performed slightly better than other algorithms for classification [11].

1.3 Problem formulation
The SA backend service collects data from several sources e.g. knowledge management systems, forums and so forth and more sources are to be added in the near future. The data is currently not categorized in a uniform manner and due to the growing amount of sources, a need has arisen to classify it. Classifying data would create structure and improve the search result because the team would be able to present related search result if no or little result is found by matching keywords.

One of the existing data sources has a favorable category structure that would be well suited as a foundation from which the classification can be derived. Since this template source exists we would like to investigate how we could make use of machine learning to classify the rest of the data by its category structure.

We will investigate which algorithm and framework would be best suited to complement the current service and thereby create an experimental service for evaluation purposes, Proof-of-Concept (POC). By doing this we are testing the hypothesis:

“If we classify multi-source data through machine learning, then our search result will improve by the minimizing of occurrences of no, slim or irrelevant search result. “

It is expected that by classifying the data the user will be presented with related search results in the cases where no or few results were previously displayed.

1.4 Motivation
By classifying the searchable data it is likely that the number of searches ending in no or
little result would be lowered. This as SA would be able to show related results as a
compliment. The improvement is in comparison to only making use of keywords in title
or description [8].

The SA team would also like to make use of supervised machine learning to
classify data as it is likely to provide a satisfactory level of accuracy while
simultaneously saving resources as opposed to categorizing all current and future
sources manually [10] [12].

1.5 Objectives
The objectives for this degree projects are as follows:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>Decide on a machine learning algorithm</td>
</tr>
<tr>
<td>O2</td>
<td>Decide on framework for machine learning</td>
</tr>
<tr>
<td>O3</td>
<td>Prepare data to be used for training machine learning model</td>
</tr>
<tr>
<td>O4</td>
<td>Train the machine learning model</td>
</tr>
<tr>
<td>O5</td>
<td>Optimize objective functions and algorithm of machine learning model</td>
</tr>
<tr>
<td>O6</td>
<td>Create service(s) that make use of the model</td>
</tr>
<tr>
<td>O7</td>
<td>Classify data</td>
</tr>
<tr>
<td>O8</td>
<td>Evaluate results</td>
</tr>
</tbody>
</table>

Table 1.1: Objectives

Objectives O4-5 (table 1.1) will be subject to an iterative process but are described as
separate objectives to clarify the process.

An investigating will be conducted to determine which algorithm and framework
will be best suited to complement the current service. By creating an experimental
service for evaluation purposes we expect to be able to present a related search result in
the case where no or little result was previously displayed.

1.6 Scope/Limitation
The aim of this degree project is to supply Sony’s current platform with a POC for a
complimentary service that is to classify data in accordance with set categories. By
classifying the data by category the aim is to improve the search results by minimizing
no or few results for the end user.

The scope of this degree project is to in the most suitable way incorporate
machine learning in the process of improving search results. It is not in the scope of this
degree project to index the data post classification nor is it in the scope of this project to
implement the final service.

The research amongst suitable frameworks will be limited to frameworks that
are most suitable to integrate with the ecosystem.

1.7 Target group
In addition to being of interest for Sony, the degree project could also be of scientific value. This as it can serve as a case study for organizations or individuals who are interested in how supervised machine learning can be used to improve search services by data classification.

1.8 Outline
The following chapters include Method, Implementation, Results, Discussion, and Conclusion.

In the Method chapter we will account for how the study will be conducted. We will also acknowledge the constraints to validity and reliability that this project is subject to.

The Implementation chapter will contain detailed information about how our study is conducted. It will include the search results prior to the experiment.

The Results chapter will show a comparison between the results prior to the study and the results post study.

In the Discussion chapter, we will compare the results and the implementation process to related studies. We will also account for possible challenges during implementation and other interesting finding.

The Conclusion chapter will feature the answer to our research question.
2 Method

The aim of this project is to see if it is possible to improve the search result by incorporating machine learning for classification of data. Firstly, we will create an experimental service to classify data based on predefined categories. Secondly, we will verify, by using the support application search tool, if the desired effect is reached and to what degree. The method we will use is that of human-oriented controlled experiment [13].

In the context of this degree project, it means that we will inspect the outcome of our experiment by comparing the search result prior to the experiment with the results post-experiment. The dependent variable is the search result. The search result prior to the experiment will serve as the baseline (control) of the independent variable and the post-experiment result is used for evaluation. Partly, we will compare the number of search results but also make a qualitative study where we evaluate the content and order of the search results both with and without the additional categories.

2.1 Categories

The articles will be categorised into four categories; Wi-Fi & Connectivity, Apps & Settings, System & Performance, and Battery Power & Charging. The categories are part of a business decision.

2.2 Measurement control

When classifying the data by using a supervised machine learning approach we have the key beforehand i.e. we know how an entry should be classified by category through machine learning. By comparing the outcome against our predefined categories we can establish the rate of accuracy.

When analyzing the search result we will use the twenty most common search terms in English (table 2.1). We will compare the ten first hits for each these terms before and after our experiment to analyze the impact of the experiment. The measurement will be conducted on five Sony devices running on different platforms and/or operating systems (tables 2.2-4). The reason that we have chosen to display results for different devices separately is that the results are filtered depending on which device is used and not all articles are available for all devices.

<table>
<thead>
<tr>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
</tr>
<tr>
<td>hotspot</td>
</tr>
<tr>
<td>update</td>
</tr>
<tr>
<td>keyboard</td>
</tr>
<tr>
<td>screen</td>
</tr>
<tr>
<td>battery</td>
</tr>
<tr>
<td>camera</td>
</tr>
<tr>
<td>Device</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Xperia X Performance</td>
</tr>
<tr>
<td>Xperia XZ Premium</td>
</tr>
<tr>
<td>Xperia XZ2</td>
</tr>
</tbody>
</table>

Table 2.2: Smartphones for evaluation

<table>
<thead>
<tr>
<th>Device</th>
<th>Android version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xperia Z4 Tablet</td>
<td>7.0 (Nougat)</td>
</tr>
</tbody>
</table>

Table 2.3: Tablet for evaluation

<table>
<thead>
<tr>
<th>Device</th>
<th>Android version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xperia Touch</td>
<td>8.0 (Oreo)</td>
</tr>
</tbody>
</table>

Table 2.4: Projector for evaluation
2.3 Tools

Firstly, we will extract relevant data from our content database using MySQL Workbench [14].

Thereafter we will conduct the core machine learning experiment using SciKit-learn as it is a lightweight machine learning framework for Python [15] [16]. We will also use NLTK for preprocessing the data. [17] The last part of the experiment will be to launch a microservice launched on Amazon Lambda with API Gateway and Amazon S3 [18] [19] [20]. The microservice will make use of the machine learning model we have created to determine the category of a search term. We will use the Amazon services as they are compatible with the rest of the platforms and services used for the search tool.

2.4 Reliability and Validity

By conducting a controlled experiment we will follow a study strategy that should be easy to replicate and that holds high measurement and execution control [13].

The results of this study are only valid for the search tool of the Sony support application and the data that it uses.
3 Implementation

The implementation was divided into three iterative main stages and result was separated into four standalone applications where one is deployed on Amazon Lambda and is integrated with Sony’s search platform.

3.1 Data

Firstly, we extracted data from different databases. As we had data from four sources with varying structure we closely evaluated the content and exported the data as a JSON array with article object. The structure of the JSON array is shown in figure 3.1.

```json
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Article data set",
  "type": "array",
  "items": {
    "title": "Article",
    "type": "object",
    "properties": {
      "id": {
        "description": "The unique identifier for an article",
        "type": "string"
      },
      "title": {
        "description": "The title of an article",
        "type": "string"
      },
      "content": {
        "description": "The selected content of an article",
        "type": "string"
      },
      "tags": {
        "description": "The tags added to an article",
        "type": "string"
      },
      "categories": {
        "description": "The categories given to an article",
        "type": "string"
      }
    },
    "required": ["id", "title", "content", "tags", "categories"]
  }
}
```

Figure 3.1: JSON scheme for articles.
Secondly, we created an application for preprocessing the data in the JSON array by sanitizing, removing stop words and stemming words. The sanitation consisted of removing HTML tags, specific recurring words such as “xperia” and special characters. The NLTK library was used for stop words and stemming, namely the English stop word library and the SnowballStemmer [17].

The application outputs two JSONs per source - one with categorized articles and one with uncategorized articles. We ended up with 1172 categorized articles.

It should be noted that during this step we could identify approximately 32 categories by analyzing tags and similar in the articles. However we did not pursue this since the four categories were a business decision.

3.2 Model

We have created an application for training the model. The application input is all the categorized articles that are joined into a dataframe and subsequently a records array using the Pandas Library [21].

The data is split into 80% training data and 20% test data for evaluation. Thereafter we have created a pipeline that combines feature extraction of all properties separately and converting their values into a matrix of token counts.

Thereafter we used sklearn’s OneVsRestClassifier strategy with sklearn’s multinomial Naive Bayes classifier. The strategy selected is a multilabel strategy that fits one classifier per class and is therefore highly suitable for the project.

After adjustments we computed an accuracy score of ≈ 0.983 when comparing against the test data.

Finally the pipeline is serialized, saved and ready for usage for predicting the categories of articles and search terms.

3.3 Prediction

We have created an application that predicts the data of the remaining 10 315 uncategorized articles. The application input is all the uncategorized articles that are joined into a dataframe and subsequently a records array using the Pandas Library. The pipeline is a model and used to predict the categories of the articles. The output is a JSON structure in figure 3.2.

```json
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Article source",
  "type": "object",
  "properties": {
    "source": {
      "description": "The article source",
      "type": "array",
      "items": {
        "title": "An article",
        "type": "object",
        "properties": {
          "id": {
            "description": "The unique identifier for an article",
            "type": "string"
          },
          "categories": {
```
"description": "Predicted categories of an article",
"type": "string"
},

"required": [
 "id",
 "categories"
 ]
}
}
}
}

Figure 3.2: JSON scheme for categorized articles

We have also created a microservice deployed on Amazon Lambda that takes a query as input and returns a predicted category of type string. The microservice is integrated with the current search platform. The search platform takes the predicted category into account when serving search results as an additional parameter.
4 Results

Firstly, we used the lambda microservice to predict the category for the twenty most common search terms. The results are shown in table 4.1.

<table>
<thead>
<tr>
<th>Query</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>Apps &amp; Settings</td>
</tr>
<tr>
<td>hotspot</td>
<td>Apps &amp; Settings</td>
</tr>
<tr>
<td>update</td>
<td>System &amp; Performance</td>
</tr>
<tr>
<td>keyboard</td>
<td>Apps &amp; Settings</td>
</tr>
<tr>
<td>screen</td>
<td>Apps &amp; Settings</td>
</tr>
<tr>
<td>battery</td>
<td>Battery, Power &amp; Charging</td>
</tr>
<tr>
<td>camera</td>
<td>Apps &amp; Settings</td>
</tr>
<tr>
<td>google</td>
<td>Apps &amp; Settings</td>
</tr>
<tr>
<td>storage</td>
<td>System &amp; Performance</td>
</tr>
<tr>
<td>backup</td>
<td>Battery, Power &amp; Charging</td>
</tr>
<tr>
<td>language</td>
<td>Apps &amp; Settings</td>
</tr>
<tr>
<td>safe mode</td>
<td>System &amp; Performance</td>
</tr>
<tr>
<td>finger</td>
<td>Apps &amp; Settings</td>
</tr>
<tr>
<td>security</td>
<td>Wifi &amp; Connectivity</td>
</tr>
<tr>
<td>notification</td>
<td>Wifi &amp; Connectivity</td>
</tr>
<tr>
<td>Bluetooth</td>
<td>Wifi &amp; Connectivity</td>
</tr>
<tr>
<td>software</td>
<td>System &amp; Performance</td>
</tr>
<tr>
<td>WiFi calling</td>
<td>Wifi &amp; Connectivity</td>
</tr>
<tr>
<td>vibrate</td>
<td>System &amp; Performance</td>
</tr>
<tr>
<td>download</td>
<td>Apps &amp; Settings</td>
</tr>
</tbody>
</table>

Table 4.1: Categorized queries

Secondly, we used the search platform to determine the baseline that is how many hits there were for each search term. If there were zero hits we made a new search were the category replaced the search term. We also created a scenario where we always included the category as part of the search query, but with a lower weight. The latter is referred to as Mixed. The hits are displayed in tables 4.2-3.
<table>
<thead>
<tr>
<th></th>
<th>Xperia X Performance</th>
<th></th>
<th></th>
<th>Xperia XZ Premium</th>
<th></th>
<th></th>
<th>Xperia XZ2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>Zero hits</td>
<td>Mixed</td>
<td>Baseline</td>
<td>Zero hits</td>
<td>Mixed</td>
<td>Baseline</td>
<td>Zero hits</td>
<td>Mixed</td>
</tr>
<tr>
<td>location</td>
<td>318</td>
<td>-</td>
<td>1498</td>
<td>318</td>
<td>-</td>
<td>1498</td>
<td>318</td>
<td>-</td>
<td>1499</td>
</tr>
<tr>
<td>hotspot</td>
<td>84</td>
<td>-</td>
<td>6095</td>
<td>84</td>
<td>-</td>
<td>6096</td>
<td>85</td>
<td>-</td>
<td>6097</td>
</tr>
<tr>
<td>update</td>
<td>4980</td>
<td>-</td>
<td>5530</td>
<td>4980</td>
<td>-</td>
<td>5530</td>
<td>4981</td>
<td>-</td>
<td>5532</td>
</tr>
<tr>
<td>keyboard</td>
<td>349</td>
<td>-</td>
<td>6204</td>
<td>349</td>
<td>-</td>
<td>6205</td>
<td>349</td>
<td>-</td>
<td>6204</td>
</tr>
<tr>
<td>screen</td>
<td>3148</td>
<td>-</td>
<td>3960</td>
<td>3149</td>
<td>-</td>
<td>3961</td>
<td>3150</td>
<td>-</td>
<td>3962</td>
</tr>
<tr>
<td>battery</td>
<td>1432</td>
<td>-</td>
<td>2528</td>
<td>1432</td>
<td>-</td>
<td>2528</td>
<td>1431</td>
<td>-</td>
<td>2526</td>
</tr>
<tr>
<td>camera</td>
<td>904</td>
<td>-</td>
<td>6415</td>
<td>904</td>
<td>-</td>
<td>6415</td>
<td>905</td>
<td>-</td>
<td>6415</td>
</tr>
<tr>
<td>google</td>
<td>1454</td>
<td>-</td>
<td>6442</td>
<td>1454</td>
<td>-</td>
<td>6442</td>
<td>1454</td>
<td>-</td>
<td>6442</td>
</tr>
<tr>
<td>storage</td>
<td>787</td>
<td>-</td>
<td>1780</td>
<td>788</td>
<td>-</td>
<td>1781</td>
<td>788</td>
<td>-</td>
<td>1782</td>
</tr>
<tr>
<td>backup</td>
<td>989</td>
<td>-</td>
<td>2041</td>
<td>989</td>
<td>-</td>
<td>2041</td>
<td>989</td>
<td>-</td>
<td>2041</td>
</tr>
<tr>
<td>language</td>
<td>189</td>
<td>-</td>
<td>6079</td>
<td>189</td>
<td>-</td>
<td>6080</td>
<td>189</td>
<td>-</td>
<td>6080</td>
</tr>
<tr>
<td>safe mode</td>
<td>2521</td>
<td>-</td>
<td>7239</td>
<td>2521</td>
<td>-</td>
<td>7239</td>
<td>2522</td>
<td>-</td>
<td>7240</td>
</tr>
<tr>
<td>finger</td>
<td>233</td>
<td>-</td>
<td>6103</td>
<td>234</td>
<td>-</td>
<td>6104</td>
<td>234</td>
<td>-</td>
<td>6103</td>
</tr>
<tr>
<td>security</td>
<td>391</td>
<td>-</td>
<td>1485</td>
<td>391</td>
<td>-</td>
<td>1485</td>
<td>391</td>
<td>-</td>
<td>1489</td>
</tr>
<tr>
<td>notification</td>
<td>989</td>
<td>-</td>
<td>6330</td>
<td>989</td>
<td>-</td>
<td>6330</td>
<td>989</td>
<td>-</td>
<td>6330</td>
</tr>
<tr>
<td>Bluetooth</td>
<td>528</td>
<td>-</td>
<td>1471</td>
<td>528</td>
<td>-</td>
<td>1471</td>
<td>528</td>
<td>-</td>
<td>1470</td>
</tr>
<tr>
<td>software</td>
<td>2740</td>
<td>-</td>
<td>3543</td>
<td>2740</td>
<td>-</td>
<td>3543</td>
<td>2740</td>
<td>-</td>
<td>3543</td>
</tr>
<tr>
<td>WiFi calling</td>
<td>998</td>
<td>-</td>
<td>6435</td>
<td>997</td>
<td>-</td>
<td>6434</td>
<td>998</td>
<td>-</td>
<td>6435</td>
</tr>
<tr>
<td>vibrate</td>
<td>407</td>
<td>-</td>
<td>1564</td>
<td>407</td>
<td>-</td>
<td>1564</td>
<td>407</td>
<td>-</td>
<td>1568</td>
</tr>
<tr>
<td>download</td>
<td>2235</td>
<td>-</td>
<td>3157</td>
<td>2235</td>
<td>-</td>
<td>3157</td>
<td>2236</td>
<td>-</td>
<td>3158</td>
</tr>
</tbody>
</table>

Table 4.2: Results for smartphones.
<table>
<thead>
<tr>
<th></th>
<th>Xperia Z4 Tablet</th>
<th></th>
<th></th>
<th>Xperia Touch (0)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>Zero hits</td>
<td>Mixed</td>
<td>Baseline</td>
<td>Zero hits</td>
<td>Mixed</td>
</tr>
<tr>
<td>location</td>
<td>318</td>
<td>-</td>
<td>1498</td>
<td>318</td>
<td>-</td>
<td>2990</td>
</tr>
<tr>
<td>hotspot</td>
<td>84</td>
<td>-</td>
<td>6096</td>
<td>84</td>
<td>-</td>
<td>6096</td>
</tr>
<tr>
<td>update</td>
<td>4980</td>
<td>-</td>
<td>5530</td>
<td>4982</td>
<td>-</td>
<td>5533</td>
</tr>
<tr>
<td>keyboard</td>
<td>349</td>
<td>-</td>
<td>6205</td>
<td>350</td>
<td>-</td>
<td>6205</td>
</tr>
<tr>
<td>screen</td>
<td>3149</td>
<td>-</td>
<td>3961</td>
<td>3152</td>
<td>-</td>
<td>3964</td>
</tr>
<tr>
<td>battery</td>
<td>1432</td>
<td>-</td>
<td>2528</td>
<td>1430</td>
<td>-</td>
<td>2524</td>
</tr>
<tr>
<td>camera</td>
<td>905</td>
<td>-</td>
<td>6415</td>
<td>903</td>
<td>-</td>
<td>6412</td>
</tr>
<tr>
<td>google</td>
<td>1454</td>
<td>-</td>
<td>6442</td>
<td>1455</td>
<td>-</td>
<td>6443</td>
</tr>
<tr>
<td>storage</td>
<td>788</td>
<td>-</td>
<td>1781</td>
<td>787</td>
<td>-</td>
<td>1782</td>
</tr>
<tr>
<td>backup</td>
<td>989</td>
<td>-</td>
<td>2041</td>
<td>990</td>
<td>-</td>
<td>2041</td>
</tr>
<tr>
<td>language</td>
<td>189</td>
<td>-</td>
<td>6080</td>
<td>192</td>
<td>-</td>
<td>6082</td>
</tr>
<tr>
<td>safe mode</td>
<td>2521</td>
<td>-</td>
<td>7239</td>
<td>2519</td>
<td>-</td>
<td>7238</td>
</tr>
<tr>
<td>finger</td>
<td>234</td>
<td>-</td>
<td>6104</td>
<td>234</td>
<td>-</td>
<td>6104</td>
</tr>
<tr>
<td>security</td>
<td>391</td>
<td>-</td>
<td>1485</td>
<td>391</td>
<td>-</td>
<td>1488</td>
</tr>
<tr>
<td>notification</td>
<td>989</td>
<td>-</td>
<td>6330</td>
<td>986</td>
<td>-</td>
<td>6327</td>
</tr>
<tr>
<td>Bluetooth</td>
<td>528</td>
<td>-</td>
<td>1471</td>
<td>528</td>
<td>-</td>
<td>1470</td>
</tr>
<tr>
<td>software</td>
<td>2740</td>
<td>-</td>
<td>3543</td>
<td>2742</td>
<td>-</td>
<td>3545</td>
</tr>
<tr>
<td>WiFi calling</td>
<td>997</td>
<td>-</td>
<td>6434</td>
<td>1000</td>
<td>-</td>
<td>6436</td>
</tr>
<tr>
<td>vibrate</td>
<td>407</td>
<td>-</td>
<td>1564</td>
<td>407</td>
<td>-</td>
<td>1567</td>
</tr>
<tr>
<td>download</td>
<td>2235</td>
<td>-</td>
<td>3157</td>
<td>2237</td>
<td>-</td>
<td>3159</td>
</tr>
</tbody>
</table>

Table 4.3: Results for tablet and projector.

We made a qualitative comparison of the ten first hits of all the terms on all devices and found that the order differed somewhat using the baseline approach and the mixed approach. Appendix A.1 and A.2 shows an example of the difference between the baseline and mixed approach for the term language on a Xperia XZ2.
5 Analysis

Table 4.1 shows that queries security and notification were predicted to belong to the Wifi & Connectivity category. At first glance, this might seem odd however most of the forum articles that were used as training material include content that relates to terms that are similar to wifi and connectivity. The remaining terms are matched in a more apparently suitable way.

Tables 4.2-3 show that none of the search terms were eligible for the zero hits approach. When using the Mixed approach it is clear that the number of hits has increased dramatically for all of the search terms.

The analysis of the quantitative study shows the relevance of the additional hits in the Mixed approach is not particularly accurate. In the example, shown in Appendix A.1 and A.2 it is evident that articles from the sources INDEVICE and KB have been added in A.2 where there are none in A.1. However, none of the articles are related to the search term “language”. They are rather related to the predicted category Apps and settings (see table 4.1 and tables 5.1-2).

<table>
<thead>
<tr>
<th>INDEVICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;To keep your apps up-to-date&quot;</td>
</tr>
<tr>
<td>&quot;To add an exception to power-saving features&quot;</td>
</tr>
<tr>
<td>&quot;What are permissions?&quot;</td>
</tr>
</tbody>
</table>

Table 5.1: Article titles added to INDEVICE with the Mixed approach (Appendix A.2).

<table>
<thead>
<tr>
<th>KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;How do I move apps from my device's internal storage to a memory card? (Xperia V)&quot;</td>
</tr>
</tbody>
</table>

Table 5.2: Article titles added to KB with the Mixed approach (Appendix A.2).

The quantitative study also shows that the ordering of articles is effected. By analyzing the baseline in Appendix A.1 and the mixed approach in Appendix A.2 we can see that the articles from the source FORUM have changed in order (see table 5.3). Both options show relevant articles however it can be argued that at least the first hit is more general in the mixed approach.
<table>
<thead>
<tr>
<th>A.1</th>
<th>A.2</th>
</tr>
</thead>
</table>
| "Arabic language"                        | "Change language while writing messages."
| "Change language while writing messages."| "How do I change the language back to English?"
| "How do I change the language back to English?" | "keyboard language" |
| "cant read language!!"                   | "Arabic language" |
| "keyboard language"                      | "cant read language!!"                   |
| "Chinese language by mistake"            | "Changing Language"                      |
| "Changing Language"                      | "Chinese language by mistake"            |
| "PC Companion changed language after upgrade" | "PC Companion changed language after upgrade" |
| "Weird language"                         | "language problem, xperia x2"            |
| "Sinhala language support for Xperia"    | "Weird language"                         |

Table 5.3: Article titles from the source FORUM have been reordered when using the Mixed approach (A.2) as opposed to Baseline (A.1).
6 Discussion

The thesis of this project degree was that “if we classify multi-source data through machine learning, then our search result will improve by the minimizing of occurrences of no, slim or irrelevant search result.” By analyzing the results we can definitely state classifying the data and the search terms increased the sheer number of hits in our results. The relevance of the search result cannot be said to have improved. The quality of the results shows that the four predetermined categories; Wi-Fi & Connectivity, Apps & Settings, System & Performance, and Battery Power & Charging are too few to fully represent the range of articles featured when used to improve the results.

Using the four categories works well for solely classifying articles in categories and displaying them in accordance with these categories for the user to browse through. For the more complex task of using the category to assist in finding more relevant results to a search term, four categories do not seem to be sufficient.

The occurrence of seemingly unfitting classification as seen in table 4.1 with the queries security and notification that is predicted to belong to the Wifi & Connectivity category is another issue. If we would have had more categories it would most likely have been easier to divide these into more appropriate categories than these very broad categories.

While preprocessing the data we could make out 32 categories by analyzing tags and similar content. These could perhaps have served better for the purpose of improving the search result as the categories would not be as broad. However, since the four categories were a business decision we did not pursue this approach.

A combination of the approaches used in the related studies that have been conducted on Search Engines for Books, News articles, and Microblogs [8] [9] [10] [11] have been used with success in our project. We have combined word vectorization and a Naive Bayes classifier and reached an accuracy rate of over 90%. The results can, however, be questioned as we only had 1172 classified articles and most of these came overwhelmingly from the same source (Forum). The 10 315 unclassified articles also mostly came from the source Forum. On the other hand, one could argue that forum posts mostly resemble the way in which a customer would formulate themselves and therefore are most suitable for the task of acting as a foundation.
7 Conclusion

By conducting this study we can draw the conclusion that it is suitable to use machine learning to classify articles into categories. It will also most likely improve the end user’s journey. The number of hits will increase, there is a slim chance of no results being shown. However, using only a limited number of categories does in no way guarantee relevant results.

The project would most likely have benefited from a far larger number of articles/posts and more categories in order to improve the search results.

These findings could benefit others attempting to categorize their search engine content by showing a concrete example of the limits of introducing too few categories namely that they become too broad and generally not very useful for optimizing search results.

We think that the project would have reached different results if we had access to a larger body of content and more categories were used.

7.1 Future work

As we see the small body of content and the few and broad categories as the main obstacle for the low relevance of the search results we would propose the following paths for future studies:

It would be interesting to see what result could be reached with an unsupervised approach. By using this approach we would not force articles into preset categories and the results would perhaps be more exact. We also suspect that we would reach a larger number of categories.

It would also be interesting to see what would happen if the approximately 32 categories we identified by tags and similar were used for classification instead, with a supervised approach. We think that it would definitely lead to a lower accuracy rate whilst categorizing but it would perhaps be of greater help for improving search results.
References


A  Qualitative Comparison

A.1 Baseline example

Search term “language” on Xperia XZ2

```json
{
    "responseCode": 0,
    "message": "",
    "result": [
        {
            "infoSource": "INDEVICE",
            "found": 0,
            "start": 0,
            "documents": []
        },
        {
            "infoSource": "KB",
            "found": 0,
            "start": 0,
            "documents": []
        },
        {
            "infoSource": "FORUM",
            "found": 160,
            "start": 0,
            "documents": [
                {
                    "uniqueId": "e135c400-709c-481b-b4fb-cf6704ee685d",
                    "documentId": null,
                    "version": null,
                    "title": "Arabic language",
                    "body": "The x series now uses SwiftKey as the default keyboard instead of xperia keyboard, which includes downloadable languages that includes arabic.",
                    "modifiedDate": "2016-08-06T01:43:00.000Z",
                    "priority": 0,
                    "url": "https://talk.sonymobile.com/t5/Xperia-X-Performance/Arabic-language/m-p/1149471"
                },
                {
                    "uniqueId": "c94c04af-68f4-4309-a433-ea060def5335",
                    "documentId": null,
                    "version": null,
                    "title": "Change language while writing messages.",
                    "body": "Hi and thank you for your reply. As i had finished writing a response i gave it another whirl. I managed to solve it. It seems like when i added Swedish and Norwegian, the keyboard layout set the default to English (UK) layout, instead of Norwegian and Swedish. As a result the virtual keyboard didn't detect other languages and i couldn't switch. I tapped the languages and changed the layout to the correct ones",
```
and voilà, everything was ok. What i would recommend (if it's a general thing an ...",
"ModifiedDate": "2017-05-24T22:30:00.000Z",
"priority": 0,
"url": "https://talk.sonymobile.com/t5/Xperia-XA1/Change-language-while-writing-mess-ages/m-p/1225967"
},
{
  "UniqueId": "ebe63139-cd6a-46e7-8d58-bfa954beb0b5",
  "documentId": null,
  "version": null,
  "title": "How do I change the language back to English?",
  "body": "Press MENU (rite hard button)>Settings(hammer & spanner)>Language & keyboard (11th option)>Select locale (1st option)>you should be able to read and select the rite language from there onwards. Have fun."
  "ModifiedDate": "2011-01-08T13:05:00.000Z",
  "priority": 0,
},
{
  "UniqueId": "2bdbe0e5-81f4-4510-9904-cd7f5e9f0c9e",
  "documentId": null,
  "version": null,
  "title": "cant read language!!",
  "body": "Goto Settings (hammer & spanner icon)>Language & keyboard(11th option>Select Locale (1st option), you should be able settle the rest from here. Good luck."
    "ModifiedDate": "2011-01-18T12:31:00.000Z",
  "priority": 0,
},
{
  "UniqueId": "2a4fc186-5ffb-4237-9d14-60c3212ef38c",
  "documentId": null,
  "version": null,
  "title": "keyboard language",
  "body": "two ways to do this 1) open the messaging app -> open a conversation -> once you see the keyboard -> tap on the input box -> long press on the input box -> select keyboard or 2) settings -> writting & language -> select the keyboard Don't forget to mark the Correct Answers & Helpful Answers"
    "ModifiedDate": "2012-05-28T18:55:00.000Z",
  "priority": 0,
The following message tells you what to do if you have changed the language of the whole phone: http://talk.sonyericsson.com/message/9489#948 If you just mean for text input, go to where you would normally enter text and touch and hold the text input area until options appear on screen. One of those is Input method. Tap that and that select the language you want. Standard keyboard or Android keyboard are both English. Other options, on mine at least, are Japanese, Chinese and Korean. Hope t ...

Hi Mackis, To change the language to English, from the menu screen, tap on Settings (the icon with the tools) then slide the screen to the eleventh option and tap on it, once there, tap the first option that appears and by sliding the screen up and down, you should see and tap on English (United States).

Yes! Now the Swedish language works. You'll have to uninstall the old version, restart your PC and then install the new version.

PC Companion changed language after upgrade
"title": "Weird language",
"body": "Sounds like some encoding problem rather than language problem. SMS should be in english? What language / region is your sola set to? You should contact your phone provider, as they're the ones sending that SMS. What app is that doing the pop up?",
"modifiedDate": "2012-09-29T15:23:00.000Z",
"priority": 0,
"url": "https://talk.sonymobile.com/t5/Xperia-sola/Weird-language/m-p/242866"
}

"title": "Sinhala language support for Xperia",
"body": "It has been asked before Rickard: I just got information that support for Sinhala is planned to be added in a future maintenance release. I don’t have any release date of this update though.
"modifiedDate": "2015-07-21T07:15:00.000Z",
"priority": 0,
}

A.2 Mixed example
Search term “language” on Xperia XZ2

{"responseCode": 0,  
"message": "",
"result": [
{"infoSource": "INDEVICE",  
"found": 3, 
"start": 0, 
"documents": [
{ 
"uniqueId": "GUID-981A711A-BB48-4A03-A15F-DD245E09A903",
"title": "Sinhala language support for Xperia",
"body": "It has been asked before Rickard: I just got information that support for Sinhala is planned to be added in a future maintenance release. I don’t have any release date of this update though.
"modifiedDate": "2015-07-21T07:15:00.000Z",
"priority": 0,
} 
] 
} 
}
In Play Store, drag the left edge of the screen to the right and select Apps & Games > My apps & games. Under the Installed tab, you can see all your installed applications on your device. If there are available updates, they will show under Updates. Tap Update on the specific app you want to update or Update all to update all apps at the same time.

Go to Settings > Battery and tap the menu icon. Tap Power-saving exceptions. Select the Apps or the System tab. Select the app or feature you want to exclude from STAMINA mode and from other Android power-saving features. Apps that are excluded from STAMINA mode are allowed to sync in the background, but other restrictions still apply.

Apps are not allowed to use device resources, such as the camera or the phone, without asking. In earlier AndroidTM versions, you had to accept all or none of these permission requests. Now, you can decide which requests to accept for each app. Some of them are critical for the app to work, others are there to provide better functionality. Permissions can be modified at any time in Settings > Apps & notifications > App permissions. By tapping, for example, Camera, you can select which applic
On most older Android™ devices from Sony (Android™ 4.3 and earlier), it is not possible to move apps from the device's internal storage to a memory card. If you need more space on your device's internal storage, you can transfer media files such as photos, videos and music to the removable memory card or to a computer.

Hi and thank you for your reply. As I had finished writing a response i gave it another whirl. I managed to solve it. It seems like when I added Swedish and Norwegian, the keyboard layout set the default to English (UK) layout, instead of Norwegian and Swedish. As a result the virtual keyboard didn't detect other languages and I couldn't switch. I tapped the languages and changed the layout to the correct ones and voilå, everything was ok. What I would recommend (if it's a general thing an...

https://talk.sonymobile.com/t5/Xperia-XA1/Change-language-while-writing-messages/m-p/1225967
Press MENU (rite hard button)>Settings(hammer & spanner)>Language & keyboard (11th option)>Select locale (1st option)>you should be able to read and select the rite language from there onwards. Have fun.

Two ways to do this 1) open the messaging app -> open a conversation -> once you see the keyboard -> tap on the input box -> long press on the input box -> select keyboard or 2) settings -> writing & language -> select the keyboard Don’t forget to mark the Correct Answers & Helpful Answers.

The x series now uses SwiftKey as the default keyboard instead of xperia keyboard, which includes downloadable languages that includes arabic.

Goto Settings (hammer & spanner icon)>Language & keyboard(11th option>Language & keyboard(11th option)>Select Locale (1st option), you should be able settle the rest from here. Good luck.

Goto Settings (hammer & spanner icon)>Language & keyboard(11th option>Language & keyboard(11th option)>Select Locale (1st option), you should be able settle the rest from here. Good luck.
Hi Mackis, To change the language to English, from the menu screen, tap on Settings (the icon with the tools) then slide the screen to the eleventh option and tap on it, once there, tap the first option that appears and by sliding the screen up and down, you should see and tap on English (United States).
ge-after-upgrade/m-p/1051810"},
{
 "uniqueId": "fa1c2aa3-1dd2-4da1-8d33-ac15793a4627",
 "documentId": null,
 "version": null,
 "title": "language problem, xperia x2",
 "body": "solved...... long press on the homescreen and then choose (plus sign) add shortcut, it would probably appear in english,,,(hope so) then select settings>spb shell> ok.... explore a litle with just 3 choices. and u'll find the language good luck",
 "modifiedDate": "2012-08-29T03:27:00.000Z",
 "priority": 0,
},
{
 "uniqueId": "bb000de4-902a-4ca7-8ee6-a811c9a01eb5",
 "documentId": null,
 "version": null,
 "title": "Weird language",
 "body": "Sounds like some encoding problem rather than language problem. SMS should be in english ? What language / region is your sola set to ? You should contact you phone provider, as they're the ones sending that SMS. What app is that doing the pop up ?",
 "modifiedDate": "2012-09-29T15:23:00.000Z",
 "priority": 0,
 "url": "https://talk.sonymobile.com/t5/Xperia-sola/Weird-language/m-p/242866"
}
]}},
{
 "infoSource": "TRS",
 "found": 0,
 "start": 0,
 "documents": []
}]}