



**FACULTY
OF MATHEMATICS
AND PHYSICS**
Charles University

MASTER THESIS

Mitchell Borchers

**Active learning in E-Commerce
Merchant Classification using Website
Information**

Department of Theoretical Computer Science and Mathematical Logic

Supervisor of the master thesis: Mgr. Marta Vomlelová, Ph.D.

Study programme: Artificial Intelligence

Study branch: IUIPA

Prague 2023

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In date
Author's signature

Dedication.

Title: Active learning in E-Commerce Merchant Classification using Website Information

Author: Mitchell Borchers

Department: Department of Theoretical Computer Science and Mathematical Logic

Supervisor: Mgr. Marta Vomlelová, Ph.D., Department of Theoretical Computer Science and Mathematical Logic

Abstract: Data and the collection and analysis of data has become an important part of everyday life. For example, navigation, e-commerce, and social media all make use of immense amounts of data to provide users with suggestions on the best routes to take, which new items they might be most interested in, and which content might fit best with their interests. A variety of algorithms and methods exist to process the data and use it to make predictions. One such algorithm is xPAL, which is a decision-theoretic approach to measure the usefulness of a labeling candidate in terms of its expected performance gain. With xPAL and other active machine learning methods an optimal strategy can be explored to classify new data points.

Keywords: probabilistic active learning xPAL machine learning multi-class classification active learning

Contents

Introduction	2
1 Active Learning	3
1.1 Introduction	3
1.2 Query Function Construction	4
1.2.1 Pool-Based	4
1.2.2 Stream-Based	4
1.2.3 Membership Queries	4
1.3 Sampling Strategies	4
1.3.1 Random Sampling	4
1.3.2 Diversity Sampling	4
1.3.3 Uncertainty Sampling	5
1.3.4 xPAL	5
1.3.5 PAL	5
1.3.6 ALCE	5
1.3.7 QBC	6
1.3.8 EER	6
1.4 Summary	6
2 Understanding xPAL	7
2.1 Pool Based Query	7
2.2 xPAL Explained	7
3 Analysis	8
4 Results	12
Conclusion	14
Bibliography	15
List of Figures	16
List of Tables	17
List of Abbreviations	18
A Attachments	19
A.1 First Attachment	19

Introduction

One of the main challenges of creating a successful machine learning model is obtaining labeled data. With easy access to a variety of modern tools, devices, and sensors, we are able to rapidly collect unlabeled data. But, in supervised learning, prediction models are trained using labeled data. The problem is that acquiring labels for the collected data can be expensive, time-consuming, or even impossible in some cases.

However, methods have been developed to help reduce the time required to label this data. Active learning is a semi-supervised machine learning method where the model is trained with a smaller set of labeled data but also aims to exploit trends within the unlabeled data. Active learning has been heavily researched in the past but typically with binary data. Multi-class active learning research is still in its infancy.

Active learning is different from other machine learning methods because it uses the unlabeled data and some evaluation criteria to determine which candidate could be the most beneficial to the model if it was given a label. In summary, the model requests the label from some oracle that provides the label then it takes this new labeled data point and rebuilds the model. We describe it as semi supervised active learning because of the oracle (typically a human) involved in the process that provides the label for the requested candidate data.

In our case we have some data (website urls) for some company or business that are given to us from our partner. From this data our partner currently utilizes human labor to browse the website and then label the url with a category (23 labels) and a sub-category (234+ tags) that branch from the main category but still have some relation. This is a repetitive and expensive task that could be automated using active learning.

To reduce the burden of human labeling we propose creating a workflow using Scrapy, Postgres, translation services, and semi supervised Active Learning (bayesian, expected error, etc.) that only require occasional interaction where a human can label a candidate that is most beneficial to the model such as using a decision-theoretic approach to measure the usefulness of a labeling candidate in terms of expected performance gain. The workflow takes the website as input, navigates to the webpage, collects and translates the text, and adds it to a database. We then run the model using the data from the database and return the model.

1. Active Learning

1.1 Introduction

Russel and Norvig succinctly define an agent and different types of learning in their book "Artificial Intelligence: A Modern Approach" (Russell and Norvig [2009]), their definition is paraphrased here. They define an agent as something that acts and a rational agent as one that acts so as to achieve the best outcome. If there is uncertainty, then the agent tries to achieve the best expected outcome. Any component of an agent can be improved by learning from data. The improvements and techniques used to make them depend on four major factors:

- Which component is to be improved.
- What prior knowledge the agent already has.
- What representation is used for the data and the component.
- What feedback is available to learn from.

Here we will mostly be focused on the final point, "What feedback is available to learn from" but also slightly on the second point because we will incorporate Bayesian learning. There are three main types of feedback that determine the three main types of learning which are unsupervised, reinforcement, and supervised learning.

In unsupervised learning an agent learns patterns even though no feedback is provided. In reinforcement learning, the agent learns from a series of rewards or punishments. In supervised learning, an agent learns from input-output pairs, which can be discrete or continuous, to find a function that maps the pairs.

The goal of supervised learning is given a training set of N example input-output pairs:

$$(x_1, y_1), (x_2, y_2), \dots (x_N, y_N),$$

where each y_j was generated by some unknown function $y = f(x)$, find a function h that approximates the true function (Russell and Norvig [2009]).

In reality, the lines separating the types of learning aren't so clear. Semi-supervised learning is also an important and widely used method. In semi-supervised learning we are given a few labeled examples that were labeled by some oracle (labeler, data annotator, etc.) and we must then make the most of a large collection of unlabeled examples. But what can we do with the unlabeled data?

Supervised learning models almost always get more accurate with more labeled data. Active learning is the process of deciding which data to sample for annotation (Munro [2021]). In other words, the central component of an active learning algorithm is the selection strategy, or deciding which of the unlabeled data could be the most useful to the model if it was labeled. Active learning uses a selection strategy that augments the existing classifier, it is not itself a classifier but rather a tool paired with a classifier.

Many strategies for choosing the next points to label exist. First we will discuss query functions then we will briefly define three basic sampling approaches: uncertainty, diversity, and random sampling to get an idea of sampling. We will then discuss some more advanced sampling approaches that are used in our experiments. When sampling the unlabeled data an ordered list is returned and the top candidate is the candidate that is expected to be most valuable for the model, but we are not strictly limited to taking just one candidate.

1.2 Query Function Construction

There are various techniques used to construct the querying functions we have discussed. We will focus on pool-based active learning, but a number of interesting and relevant ideas appear within other active-learning frameworks that are worth mentioning.

1.2.1 Pool-Based

The learner calculates the potential gain of all the unlabeled points in the pool, then requests the label for the point that maximizes the expected information gain for the classifier. For pool-based multiclass active learning, a labeled pool and an unlabeled pool are presented to the algorithm. In each iteration, the algorithm selects one instance from the unlabeled pool to query its label (Huang and Lin [2016]).

1.2.2 Stream-Based

The learner is provided with a stream of unlabeled points. On each trial, a new unlabeled point is drawn and introduced to the learner who must decide whether or not to request its label. Note that the stream-based model can be viewed as an online version of the pool-based model (Baram et al. [2004]).

1.2.3 Membership Queries

On each trial the learner constructs a point in input space and requests its label. This model can be viewed as a pool-based game where the pool consists of all possible points in the domain (Baram et al. [2004]).

1.3 Sampling Strategies

1.3.1 Random Sampling

Random sampling is rather self explanatory as we randomly select an unlabeled data point from the pool and have an oracle provide a label.

1.3.2 Diversity Sampling

Diversity sampling is the set of strategies for identifying unlabeled items that are underrepresented or unknown to the machine learning model in its current state.

The items may have features that are unique or obscure in the training data, or they might represent data that are currently under-represented in the model.

Either way this can result in poor or uneven performance when the model is applied or the data is changing over time. The goal of diversity sampling is to target new, unusual, or underrepresented items for annotation to give the algorithm a more complete picture of the problem space (Munro [2021]).

1.3.3 Uncertainty Sampling

Uncertainty sampling is the set of strategies for identifying unlabeled items that are near a decision boundary in your current machine learning model. If you have a binary classification task, these items will have close to a 50% probability of belonging to either label; therefore, the model is called uncertain or confused.

These items are most likely to be wrongly classified, so they are the most likely to result in a label that differs from the predicted label, moving the decision boundary after they have been added to the training data and the model has been retrained (Munro [2021]).

1.3.4 xPAL

Extended probabilistic gain for active learning (xPAL) is a decision-theoretic selection strategy that directly optimizes the gain and misclassification error, and uses a Bayesian approach by introducing a conjugate prior distribution to determine the class posterior to deal with uncertainties. Although the data distribution can be estimated, there is still uncertainty about the true class posterior probabilities.

These class posterior probabilities can be modeled as a random variable based on the current observations in the dataset. For this model, a Bayesian approach is used by incorporating a conjugate prior to the observations. This produces more robust usefulness estimates for the candidates Kottke et al. [2021].

1.3.5 PAL

Probabilistic Active Learning (PAL) follows a smoothness assumption and models for a candidate instance both the true posterior in its neighbourhood and its label as random variables. By computing for each candidate its expected gain in classification performance over both variables, PAL selects the candidate for labelling that is optimal in expectation. PAL shows comparable or better classification performance than error reduction and uncertainty sampling, has the same asymptotic linear time complexity as uncertainty sampling, and is faster than error reduction (Kreml et al. [2014]).

1.3.6 ALCE

Active Learning with Cost Embedding (ALCE) is a non-probabilistic uncertainty sampling algorithm for cost-sensitive multiclass active learning. They first designed a cost-sensitive multiclass classification algorithm called cost embedding (CE), which embeds the cost information in the distance measure in a special hidden space by non-metric multidimensional scaling. They then use a mirroring

trick to let CE embed the possibly asymmetric cost information in the symmetric distance measure (Huang and Lin [2016]).

1.3.7 QBC

Query by committee uses an ensemble of classifiers that are trained on bootstrapped replicates of the labeled set (Seung et al. [1992]).

1.3.8 EER

Monte Carlo estimation of error reduction (EER) estimates future error rate by log-loss, using the entropy of the posterior class distribution on a sample of the unlabeled examples, or by 0-1 loss, using the posterior probabilities of the most probable class for the sampled unlabeled examples Roy and McCallum [2001].

1.4 Summary

Now it should be more clear how the sampling strategy is the major component of active learning. The query function construction is also important but it is just a means of routing the data to be sampled. In the next chapter we will look into the specifics of xPAL.

2. Understanding xPAL

We have introduced many different active learning models in the previous section, and we will test these models on our data. However, we will focus on xPAL using a pool based query function.

2.1 Pool Based Query

2.2 xPAL Explained

3. Analysis

In this section we take a deeper look into the data, the process of augmenting the data, the experiment, and the results. Our partner has provided a small sample of 1000 labeled data points. This data was manually labeled by an annotator. The data consists of a merchant name, merchant website (url), merchant category, and merchant tag as shown in Table 3.1.

merchant name	merchant url	merchant category	merchant tags
State Hospital	http://hospital.com/	Health	'{"Clinic"}'

Table 3.1: This is an example of a single data point.

The current process consists of giving the merchant url to an annotator and the annotator then views the website and either can instantly provide a label and tags for the website or in some cases may need to browse further into the website (by viewing sibling pages such as the 'About Us' sections or product pages) to get an idea of how the website should be classified.

The merchant tags are ordered by relevance, with the first tag in the list being the most general and the final being the most specific. An example of the tag hierarchy is show in Table. 3.2 where we can see that this sample consists of data from various categories all contained within the 'Eco' side tag grouping.

Category	Level 1	Level 2	Level 3	Side Tag
Travel	Local Transport	Micro-mobility	Bike Sharing	Eco
		Public Transport		Eco
Fashion	Clothing - Other	Second Hand		Eco
Car	Charging Station			Eco
	Car Sharing			Eco

Table 3.2: This is an example of how the tags use different levels.

Similarly to the annotator our goal is to automate the navigation, collection/storing process, and classification of the website. This pipeline speeds up the browsing process and can allow the annotator to spend much less time annotating and require the annotator to only anotate data expected to drastically improve the classifier.

The initial 1000 data points we received were essentially just labels with a pointer (a url) to where the data is. The labels needed the text from the websites that the annotator viewed to begin the classifying process. To gather the text data from the websites we used the Scrapy framework to extract text data from a single top level page of a website. We chose to only scrape the top level page text because of the results published in another study where it was observed that adding more pages to the dataset does not necessarily mean obtaining better resutls (Sahid et al. [2019]).

Out of these initial data points 179 contained links that could not be accessed or links that provided no text data that could be scraped. Out of the remaining 821 data points 274 of them were in English.

It is important for us to have the data in English as it allows us to exploit stop words when using the Scikit-Learn TF-IDF Vectorizer to construct our dataset. Stop words are words like “and”, “the”, “him”, which are presumed to be uninformative in representing the content of a text, and which may be removed to avoid them being construed as signal for prediction (skl).

Out of the remaining 275 english data points the data was distributed into the categories as shown in Figure 3.1.

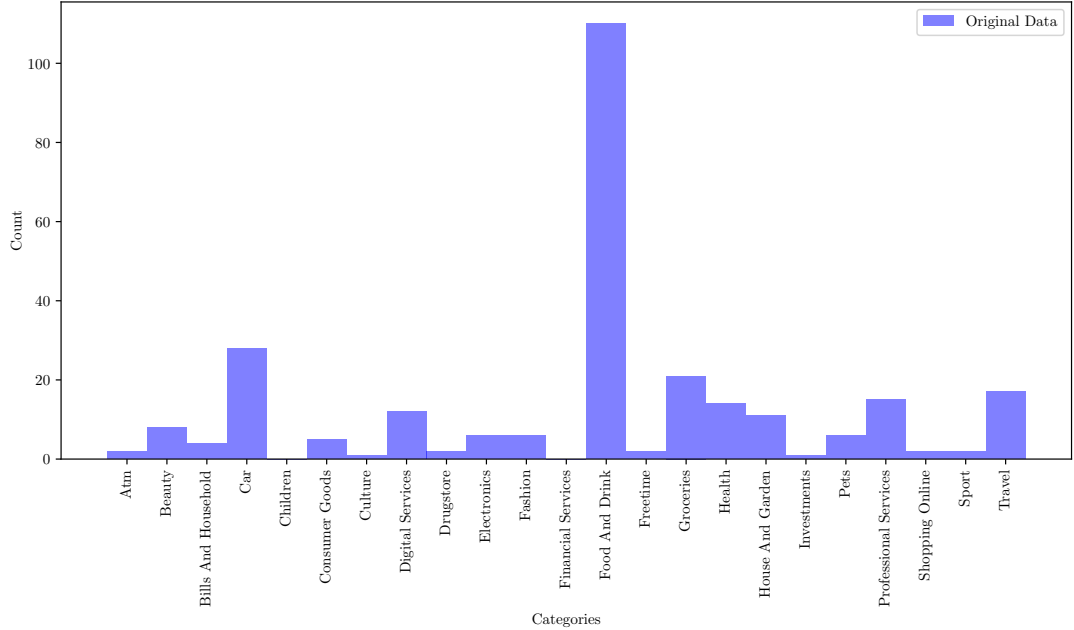


Figure 3.1: The histograms for the original usable english data.

At this stage, it was clear that our dataset wasn't representing all categories equally. The 'Food and Drink' category has many more data points than the 'Culture' and 'Investments' categories which each had a single data point. The 'Children' and 'Financial Services' categories weren't represented at all. Obviously this was problematic because we would like to have, minimum, three data points in each category to build train, test, and validation datasets, albeit limited as this would be.

At this point we had a significant amount of data that wasn't being used. We decided to find a way to translate the existing data. We tried various libraries available on GitHub but we weren't getting good results. We found that Azure had a service available and a free option of up to 2 million characters translated per month. This was a perfect option and we were able to use this api to translate the remaining data.

In addition to the translated data we also manually collected and labeled 141 additional data points to use mainly for validation. All the original data and additional data are shown in Figure 3.2

In addition to translating the data we used the TF-IDF vectorizer from Scikit-Learn. With just the original data we were able to find the highly correlated words for each category, shown below in Table 3.3. Some categories such as 'Culture', 'Digital Services', 'Shopping Online' that have few data points have words such as 'kihnu', 'td', 'patria', respectively, have no relative meaning to the category in

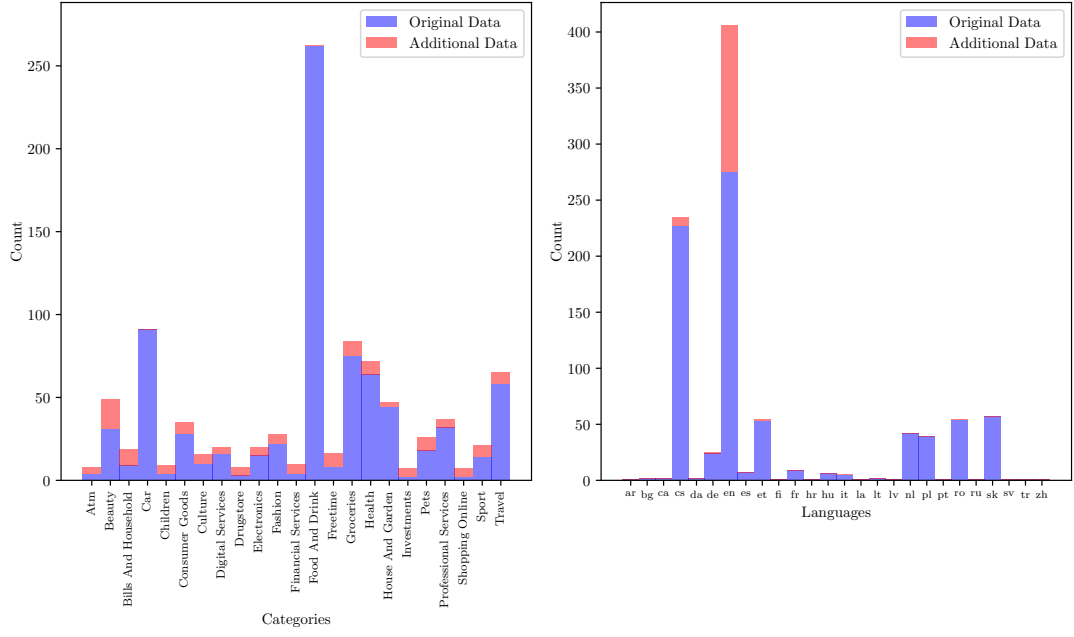


Figure 3.2: The histograms for the original and additional data for all languages.

English. This is likely because the collected data is sparse and the quality of the scraped text data is low.

Table 3.3: Keywords from TF-IDF with Chi Squared using the original data.

	Keyword 1	Keyword 2	Keyword 3
Atm	banking	zu	bank
Beauty	hairdressing	hairdresser	hair
Bills And Household	recycling	liberty	internet
Car	stations	auto	car
Children	toy	sold	toysrus
Consumer Goods	wallpaper	flowers	flower
Culture	museum	theater	kihnu
Digital Services	td	bitly	servers
Drugstore	alert	enable	detergent
Electronics	electronic	onoff	computers
Fashion	jewellery	men	women
Financial Services	pre	nissan	insurance
Food And Drink	bar	cafe	restaurant
Freetime	searched	casino	smartflex
Groceries	functioning	liquor	bakery
Health	optics	dental	pharmacy
House And Garden	wallpapers	paints	hardware
Investments	strive	investment	patria
Pets	breeding	pet	veterinary
Professional Services	faculty	parcel	school
Shopping Online	web	owner	joom
Sport	functional	singltrek	adidas
Travel	rooms	accommodation	hotel

4. Results

Here we will explore the results of the probabilistic and other active learning classifiers. Check out Figure 4.1.

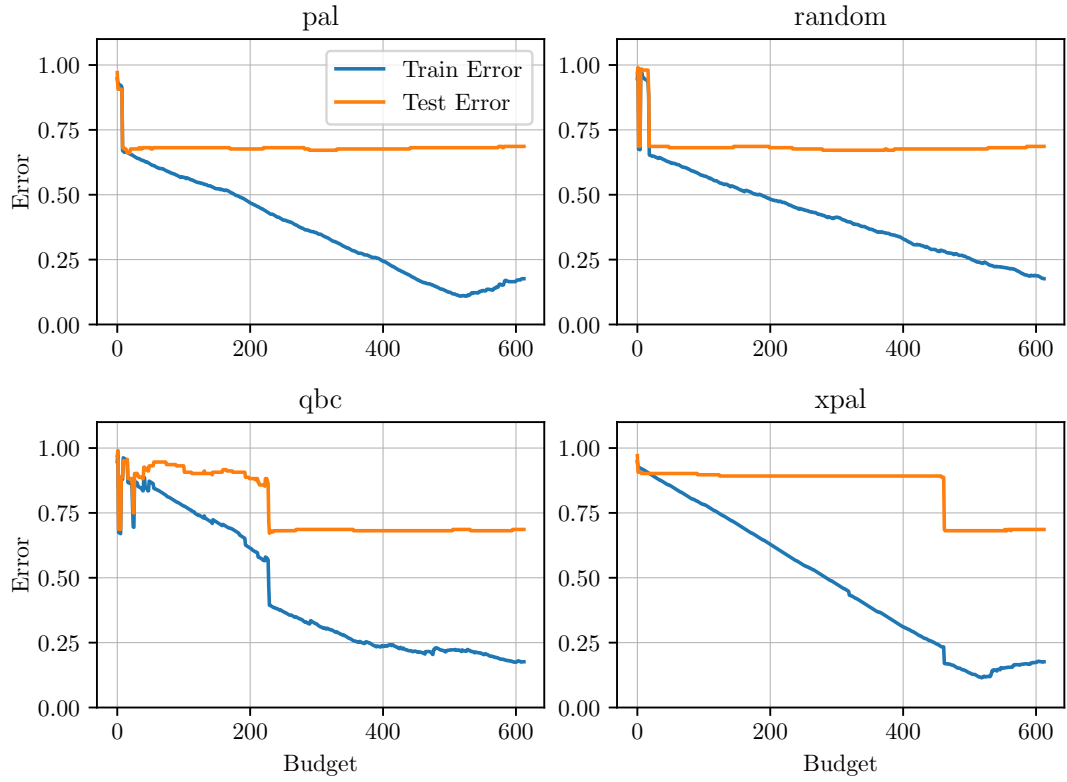


Figure 4.1: Train and test (test set size 25%) results using with the original data set.

Tester 4.1

Table 4.1: Classification Report GBDT

	precision	recall	f1-score	support
Atm	0.00	0.00	0.00	1.00
Beauty	0.75	0.33	0.46	9.00
Bills And Household	0.00	0.00	0.00	3.00
Car	0.67	0.52	0.58	27.00
Children	0.00	0.00	0.00	1.00
Consumer Goods	1.00	0.50	0.67	8.00
Culture	0.50	0.33	0.40	3.00
Digital Services	1.00	0.40	0.57	5.00
Drugstore	0.00	0.00	0.00	1.00
Electronics	0.50	0.20	0.29	5.00
Fashion	0.50	0.14	0.22	7.00
Financial Services	0.00	0.00	0.00	1.00
Food And Drink	0.48	0.90	0.62	79.00
Freetime	0.00	0.00	0.00	2.00
Groceries	0.17	0.09	0.11	23.00
Health	0.86	0.63	0.73	19.00
House And Garden	0.50	0.31	0.38	13.00
Investments	0.00	0.00	0.00	1.00
Pets	0.00	0.00	0.00	5.00
Professional Services	0.75	0.30	0.43	10.00
Shopping Online	0.00	0.00	0.00	1.00
Sport	0.00	0.00	0.00	4.00
Travel	0.67	0.59	0.62	17.00
accuracy	0.52	0.52	0.52	0.52
macro avg	0.36	0.23	0.26	245.00
weighted avg	0.52	0.52	0.48	245.00

Conclusion

Bibliography

- 6.2. Feature extraction — scikit-learn.org. https://scikit-learn.org/stable/modules/feature_extraction.html. [Accessed 03-Mar-2023].
- Yoram Baram, Ran El Yaniv, and Kobi Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5(Mar):255–291, 2004.
- Kuan-Hao Huang and Hsuan-Tien Lin. A novel uncertainty sampling algorithm for cost-sensitive multiclass active learning. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 925–930. IEEE, 2016.
- Daniel Kottke, Marek Herde, Christoph Sandrock, Denis Huseljic, Georg Kreml, and Bernhard Sick. Toward optimal probabilistic active learning using a bayesian approach. *Machine Learning*, 110(6):1199–1231, 2021.
- Georg Kreml, Daniel Kottke, and Myra Spiliopoulou. Probabilistic active learning: A short proposition. In Torsten Schaub, Gerhard Friedrich, and Barry O’Sullivan, editors, *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI2014)*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 1049–1050, Prague, Czech Republic, August 2014. IOS Press. Short Paper.
- Robert Munro. *Human-in-the-loop machine learning*. Manning Publications, New York, NY, July 2021.
- Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2:441–448, 2001.
- Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2009. ISBN 9780136042594.
- Galuh Tunggadewi Sahid, Rahmad Mahendra, and Indra Budi. E-commerce merchant classification using website information. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*, pages 1–10, 2019.
- H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

List of Figures

3.1	The histograms for the original usable english data.	9
3.2	The histograms for the original and additional data for all languages.	10
4.1	Train and test (test set size 25%) results using with the original data set.	12

List of Tables

3.1	This is an example of a single data point.	8
3.2	This is an example of how the tags use different levels.	8
3.3	Keywords from TF-IDF with Chi Squared using the original data.	11
4.1	Classification Report GBDT	13

List of Abbreviations

A. Attachments

A.1 First Attachment