



**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

I want to thank my supervisor, Mgr. Marta Vomlelová, Ph.D., for her guidance and support throughout this project. I would also like to thank my parents and my sisters for their support and encouragement throughout my studies.

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

<b>Introduction</b>	<b>3</b>
<b>1 Active Learning</b>	<b>6</b>
1.1 Introduction . . . . .	6
1.2 Query Function Construction . . . . .	7
1.2.1 Pool-Based . . . . .	7
1.2.2 Stream-Based . . . . .	7
1.2.3 Membership Queries . . . . .	7
1.3 Sampling Strategies . . . . .	7
1.3.1 Random Sampling . . . . .	8
1.3.2 Diversity Sampling . . . . .	8
1.3.3 Uncertainty Sampling . . . . .	8
1.3.4 PAL . . . . .	8
1.3.5 xPAL . . . . .	8
1.3.6 ALCE . . . . .	9
1.3.7 QBC . . . . .	9
1.3.8 EER . . . . .	9
1.4 Classifiers . . . . .	9
1.5 Summary . . . . .	10
<b>2 Understanding xPAL</b>	<b>11</b>
2.1 Kernel . . . . .	11
2.2 Risk . . . . .	12
2.3 Conjugate Prior . . . . .	12
2.4 Risk Difference Using the Conjugate Prior . . . . .	13
2.5 Expected Probabilistic Gain . . . . .	14
<b>3 Data Review</b>	<b>15</b>
3.1 Overview . . . . .	15
3.2 Collection . . . . .	16
3.3 Processing . . . . .	16
<b>4 Testing</b>	<b>19</b>
4.1 Original Data . . . . .	19
4.1.1 Active Learning with PWC and RBF Kernel . . . . .	19
4.1.2 Active Learning with PWC and Cosine Kernel . . . . .	20
4.1.3 Classifier Evaluation . . . . .	21
4.2 All Data . . . . .	26
4.2.1 Active Learning with Additional Data . . . . .	26
4.2.2 LinearSVC with Additional Data . . . . .	27
4.2.3 Active Learning Revisited . . . . .	28
4.2.4 Active Learning with Less Categories . . . . .	28
<b>Conclusion</b>	<b>29</b>
<b>Bibliography</b>	<b>30</b>

<b>List of Figures</b>	<b>31</b>
<b>List of Tables</b>	<b>32</b>
<b>List of Abbreviations</b>	<b>33</b>
<b>A Attachments</b>	<b>34</b>

# 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 impossibly difficult in some cases.

However, methods have been developed to help reduce the number of labeled data required to train the classifier. Active learning is a semi-supervised machine learning framework where the model is trained with a smaller set of labeled data but which also aims to exploit trends within the unlabeled data. Active learning is a framework in which the learner has the freedom to select which data points are added to its training set (Roy and McCallum [2001]).

Active learning is different from other frameworks 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 classifier. 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 will provide a set of labeled data to the active learning framework (or sampling strategy). The sampling strategy will assume all the data is unlabeled and then choose a candidate from the unlabeled data pool. Then the label is revealed and the classifier is updated using the new data point. The newly labeled data point is then added to the labeled data pool and the process repeats.

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 supplemented using active learning.

To reduce the amount of data required to train the classifier we consider a combination of tools and frameworks, namely: Scrapy, Postgres, a translation service, and an active learning sampling strategy paired with a classifier. We also explore the use of different classifiers to determine if there is some optimal classifier.

A website is required as input, then we use the Scrapy framework to navigate to the webpage, and collect then store the scraped data into the database. Next we access the data, translate the text, and add the translated data back into the database. During this process we also remove some html tags and other unnecessary data.

Once the the data is close to just pure text we use TF-IDF to transform the it into a vectorized representation. This data is used by the active learning sampling strategy and the model is trained iteratively. We also explore different classifiers to determine if there is some optimal classifier other than the one implemented

within the sampling strategy framework.

In the first section we introduce active learning and the different components of active learning. In the second section we look more into the details of xPAL and how it works. In the third section we discuss the data and the steps we took to collect and process the data. In the fourth section we conduct a variety of experiments to explore the performance of the sampling strategies and alternative classifiers.

Our goal is to understand the data (plus the collection process), and the performance of the selection strategies and classifiers so that we can help our partner make an informed decision on how to proceed with the labeling process.

## Notable Definitions

In this section we define some terms and ideas that will be helpful in understanding the upcoming sections.

**Definition 1** (Beta Prior). *A beta prior is a conjugate prior for the binomial distribution. It is a continuous probability distribution defined on the interval  $[0, 1]$  and is parameterized by two positive shape parameters,  $\alpha$  and  $\beta$ . The beta distribution is defined as:*

$$\text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

where  $\Gamma$  is the gamma function and  $x$  is a random variable. The gamma function is defined as:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

The gamma function is used as a normalizing constant to ensure that the probability density function integrates to 1 over the simplex, which is the space of all probability vectors that sum to 1.

**Definition 2** (Conjugate Prior). *A conjugate prior is a prior distribution that is in the same family of distributions as the likelihood function. In other words, the posterior distribution will have a similar functional form to the prior distribution.*

**Definition 3** (Decision-Theoretic). *Decision-theoretic active learning is a framework that uses the expected performance gain of a candidate to determine which candidate to label. The expected performance gain is the expected performance of the classifier after labeling the candidate minus the expected performance of the classifier before labeling the candidate. The expected performance of the classifier is the expected value of the performance measure given the posterior distribution of the classifier.*

**Definition 4** (Dirichlet Distribution). *The Dirichlet distribution is a multivariate generalization of the beta distribution. It is a continuous probability distribution defined on the  $K$ -simplex,  $\Delta_K = \{x \in \mathbb{R}^K : x_i \geq 0, \sum_{i=1}^K x_i = 1\}$ . The Dirichlet distribution is parameterized by a vector of positive shape parameters,  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$ . The Dirichlet distribution is defined as:*

$$\text{Dir}(\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i-1}$$



where  $\Gamma$  is the gamma function as defined in Definition 1 and where  $x$  is a random vector.

**Definition 5** (Ground Truth). *Ground truth is the true label of a data point.*

**Definition 6** (Posterior Probabilities). *Posterior probability is a type of conditional probability that results from updating the prior probability with information summarized by the likelihood via an application of Bayes' rule. The posterior probability is the probability of an event occurring given that another event has occurred.*

**Definition 7** (Omniscient Oracles). *Omniscient oracle is a hypothetical entity that has complete knowledge of the true labels of all data points in a given dataset. An omniscient oracle knows the ground truth labels of all data points.*

# 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". However, we will also discuss the importance of the second and third points because of our use of Bayesian learning and how the form and quality of the data affects the experiments. There are three main types of feedback that determine the three main types of learning, which are: unsupervised, reinforcement, and supervised.

In unsupervised learning an agents goal is to discover patterns in the data even though no feedback or labels are provided. In reinforcement learning, the agent learns from a series of rewards or punishments that are dealt out based on its decisions. In supervised learning, an agent learns from input-output pairs, which can be discrete or continuous, to find a function that maps the pairs as best as possible.

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  $f$ .

In reality, the lines separating the types of learning aren't so clear. Semi-supervised learning is also an important and widely used approach. 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 select 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 an evaluation methodology working with a classifier.

Many different sampling strategies 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 other 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 (Huang and Lin [2016]). 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.

### 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 (Baram et al. [2004]). Note that the stream-based model can be viewed as an online version of the pool-based model.

### 1.2.3 Membership Queries

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

## 1.3 Sampling Strategies

Sampling strategies, also referred to as selection strategies, are the core of the active learning process. The goal of sampling is to select the most useful data points from the unlabeled pool to label. The most useful data points are those that are expected to improve the classifier the most.

### 1.3.1 Random Sampling

Random sampling is self explanatory as it randomly selects an unlabeled data point from the pool and requests to have it labeled then used in the model. Random sampling is good to use as a baseline to compare other sampling strategies with.

### 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 (Munro [2021]). 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.

### 1.3.3 Uncertainty Sampling

Uncertainty sampling is the set of strategies for identifying unlabeled items that are near a decision boundary in the current machine learning model (Munro [2021]). Uncertainty sampling is simple given a classifier that estimates  $P(C|w)$  (Lewis and Gale [1994]). On each iteration, the current version of classifier can be applied to each data point, and the data with estimated  $P(C|w)$  values closest to 0.5 are selected, since 0.5 corresponds to the classifier being most uncertain of the class label.

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.

### 1.3.4 PAL

Probabilistic Active Learning (PAL) follows a smoothness assumption and models for a candidate instance both the true posterior in its neighborhood and its label as random variables (Krempel et al. [2014]). By computing for each candidate its expected gain in classification performance over both variables, PAL selects the candidate for labeling 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.

### 1.3.5 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 (Kottke et al. [2021]). 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.

### 1.3.6 ALCE

Active Learning with Cost Embedding (ALCE) is a non-probabilistic uncertainty sampling algorithm for cost-sensitive multiclass active learning (Huang and Lin [2016]). 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.

### 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]). The idea is to train a committee of classifiers on the available labeled data and then use the committee to select the most informative unlabeled data for labeling (Freund et al. [1997]). The committee consists of several classifiers, each trained on a slightly different subset of the available labeled data.

The QBC algorithm measures the disagreement of the committee’s predictions on each unlabeled data point. The intuition is that if the committee members disagree then it is likely to be a difficult data point for the current classifier and thus informative for labeling.

The algorithm selects a fixed number of the most informative examples and asks the user or oracle to label them. The labeled examples are then added to the labeled dataset, and the committee is retrained on the expanded labeled dataset. This process is repeated until the algorithm achieves a desired level of accuracy or the available labeling budget is exhausted.

### 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 Classifiers

The main classifier used with the active learning sampling strategies is the Parzen Window Classifier (PWC). It is a non-parametric method used for classification

and density estimation in machine learning. It works by estimating the probability density function of a given class using a kernel density estimator, and then using Bayes' theorem to classify new instances based on their estimated probability densities.

We will also explore using other classifiers from Scikit-Learn and TensorFlow and compare their performance on the data without using active learning to see if there is any improvement beyond the PWC classifier.

## 1.5 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 sampling strategies in the previous section, and we will use them to test which strategy performs best with our data. However, we will mainly focus on using the xPAL sampling strategy and a pool based query function. The xPAL sampling strategy is a decision-theoretic approach to measure the usefulness of a labeling candidate in terms of its expected performance gain (Kottke et al. [2021]). We can estimate the data distribution but we are uncertain about the true class posterior probabilities. The class posterior probabilities are modeled as a random variable based on the current observations. Therefore a Bayesian approach is used by incorporating a conjugate prior to the observations. In general, the idea is to estimate the expected performance gain for the classifier, using the unlabeled data, and then select the best data point and request or reveal its label. Descriptions of the variables used throughout the paper are listed in Table 2.1.

	<b>Descriptions</b>
$C$	Number of classes
$x$	Input $x \in \mathbb{R}^D$ (D-dimensional vector space)
$y$	Class label $y \in \mathcal{Y}$
$\mathcal{Y}$	Set of all labels $\mathcal{Y} = \{1, \dots, C\}$
$f^{\mathcal{L}}$	Classifier that maps input $x$ to label $y$
$L$	Loss
$R$	Risk $R(f^{\mathcal{L}}) \in \mathbb{R}_0^x$
$R_{\mathcal{E}}$	Empirical risk
$\mathcal{L}$	Set of labeled data $\{(x_1, y_1), \dots, (x_n, y_n)\}$
$\mathcal{U}$	Set of unlabeled data $\{x_1, \dots, x_n\}$
$\mathcal{E}$	Set of available labeled and unlabeled data $\{x : (x, y) \in \mathcal{L}\} \cup \mathcal{U}$

Table 2.1: Variable names and descriptions.

### 2.1 Kernel

A kernel based classifier is used in xPAL which determines the similarity of two data points. The kernel function  $\mathbf{K}(x, x')$  is a function that maps two data points to a real number, which is then used to estimate the probability density of the data. The kernel frequency estimate  $\mathbf{k}_x^{\mathcal{L}}$  of an instance  $x$  is calculated using the labeled instances  $\mathcal{L}$ . The  $y$ -th element of that C-dimensional vector describes the similarity-weighted number of labels of class  $y$ .

$$\mathbf{k}_{x,y}^{\mathcal{L}} = \sum_{(x',y') \in \mathcal{L}} \mathbb{1}_{y=y'} \mathbf{K}(x, x') \quad (2.1)$$

The Parzen Window Classifier uses the labeled data for training and predicts the most frequent class and was selected by Kottke et al. to use because of its

speed and ability to implement different kernels depending on the data (Kottke et al. [2021]). It was used for all the selection strategies in their experiments.

$$f^{\mathcal{L}}(x) = \arg \max_{y \in \mathcal{Y}} (\mathbf{k}_{x,y}^{\mathcal{L}}) \quad (2.2)$$

We will use the PWC classifier for our experiments because that is what is implemented with the active learning strategies, but we will also evaluate other classifiers and compare their performance less active learning.

## 2.2 Risk

For xPAL, Kottke et al. use the classifications error as the performance measure and minimize the zero-one loss. The risk describes the expected value of the loss relative to the joint distribution given some classifier. The zero-one loss returns 0 if the prediction from the classifier is equal to the true class else it returns 1. The risk is a theoretical concept that cannot be computed directly since it requires knowledge of the entire population distribution. Instead, we typically try to approximate the risk using the empirical risk.

$$R(f^{\mathcal{L}}) = \mathbb{E}_{p(x,y)} [\mathbf{L}(y, f^{\mathcal{L}}(x))] \quad (2.3)$$

$$= \mathbb{E}_{p(x)} \left[ \mathbb{E}_{p(y|x)} [\mathbf{L}(y, f^{\mathcal{L}}(x))] \right] \quad (2.4)$$

$$\mathbf{L}(y, f^{\mathcal{L}}(x)) = \mathbb{1}_{f^{\mathcal{L}}(x) \neq y} \quad (2.5)$$

Because it is not known how the data is generated Kottke et al. use a Monte-Carlo integration with all available data  $\mathcal{E}$  to represent the generator. The empirical risk  $R_{\mathcal{E}}$  is the average of the loss over all the data points in the dataset. It refers to the average value of a given loss function over a finite set of observed data points.

$$R_{\mathcal{E}}(f^{\mathcal{L}}) = \frac{1}{|\mathcal{E}|} \sum_{x \in \mathcal{E}} \mathbb{E}_{p(y|x)} [\mathbf{L}(y, f^{\mathcal{L}}(x))] \quad (2.6)$$

$$= \frac{1}{|\mathcal{E}|} \sum_{x \in \mathcal{E}} \sum_{y \in \mathcal{Y}} p(y|x) \mathbf{L}(y, f^{\mathcal{L}}(x)) \quad (2.7)$$

The empirical risk is a computable quantity that can be used as an estimate of the risk. However, it is only an approximation and is subject to sampling error.

## 2.3 Conjugate Prior

The conditional class probability  $p(y|x)$  depends on the ground truth which is unknown. As a result the conditional class probability is exactly the  $y$ -th element of the unknown ground truth vector  $\mathbf{p}$ . The nearby labels from  $\mathcal{L}$  can be used to estimate the ground truth  $\mathbf{p}$  because the oracle provides the labels according to



$\mathbf{p}$ . If we assume a smooth distribution then the estimate is relatively close to the ground truth if we have enough labeled instances.

$$p(y|x) = p(y|t(x)) = p(y|\mathbf{p}) = \text{Cat}(y|\mathbf{p}) = p_y \quad (2.8)$$

A Bayesian approach is used for estimation by calculating the posterior predictive distribution (calculating the expected value over all possible ground truth values). The probability of  $y$  given some  $x$  is approximately equal to the kernel frequency estimate of  $x$ .

$$p(y|x) \approx p(y|\mathbf{k}_x^\mathcal{L}) = \mathbb{E}_{p(\mathbf{p}|\mathbf{k}_x^\mathcal{L})} [p_y] = \int p(\mathbf{p}|\mathbf{k}_x^\mathcal{L}) p_y d\mathbf{p} \quad (2.9)$$

Bayes theorem is then used to determine the posterior probability of the ground truth at instance  $x$  in Equation 2.10. The likelihood  $p(\mathbf{k}_x^\mathcal{L}|\mathbf{p})$  is a multinomial distribution because each label has been drawn from  $\text{Cat}(y|\mathbf{p})$ . A prior is introduced and selected as a Dirichlet distribution with  $\alpha \in \mathbb{R}^C$  as this is the conjugate prior of the multinomial distribution. An indifferent prior is chosen and each element of alpha is set to the same value. The Dirichlet distribution is an analytical solution for the posterior when the conjugate prior of the multinomial likelihood are used.

$$p(\mathbf{p}|\mathbf{k}_x^\mathcal{L}) = \frac{p(\mathbf{k}_x^\mathcal{L}|\mathbf{p})p(\mathbf{p})}{p(\mathbf{k}_x^\mathcal{L})} \quad (2.10)$$

$$= \frac{\text{Mult}(\mathbf{k}_x^\mathcal{L}|\mathbf{p}) \cdot \text{Dir}(\mathbf{p}|\alpha)}{\int \text{Mult}(\mathbf{k}_x^\mathcal{L}|\mathbf{p}) \cdot \text{Dir}(\mathbf{p}|\alpha) d\mathbf{p}} \quad (2.11)$$

$$= \text{Dir}(\mathbf{p}|\mathbf{k}_x^\mathcal{L} + \alpha) \quad (2.12)$$

The conditional class probability is determined next from Equation 2.9. It is calculated with the expected value of the Dirichlet distribution.

$$p(y|\mathbf{k}_x^\mathcal{L}) = \mathbb{E}_{\text{Dir}(\mathbf{p}|\mathbf{k}_x^\mathcal{L} + \alpha)} [p_y] \quad (2.13)$$

$$= \int \text{Dir}(\mathbf{p}|\mathbf{k}_x^\mathcal{L} + \alpha) p_y d\mathbf{p} \quad (2.14)$$

$$= \frac{(\mathbf{k}_x^\mathcal{L} + \alpha)_y}{\|\mathbf{k}_x^\mathcal{L} + \alpha\|_1} \quad (2.15)$$

The last term is the  $y$ -th element of the normalized vector. The 1-norm is used to normalize the vector.

## 2.4 Risk Difference Using the Conjugate Prior

Next, we insert equation 2.15 into the empirical risk equation 2.7. We are approximating  $p(y|x)$  with  $p(y|\mathbf{k}_x^\mathcal{L})$  which is the empirical risk based on the labeled data  $\mathcal{L}$ .

$$\hat{R}_{\mathcal{E}}(f^{\mathcal{L}}, \mathcal{L}) = \frac{1}{|\mathcal{E}|} \sum_{x \in \mathcal{E}} \sum_{y \in \mathcal{Y}} \frac{(\mathbf{k}_x^{\mathcal{L}} + \alpha)_y}{\|\mathbf{k}_x^{\mathcal{L}} + \alpha\|_1} \cdot L(y, f^{\mathcal{L}}(x)) \quad (2.16)$$

Now let's assume we add a new labeled candidate  $(x_c, y_c)$  to the labeled data set  $\mathcal{L}$ . We will now denote the set with the newly labeled data point  $\mathcal{L}^+ = \mathcal{L} \cup \{(x_c, y_c)\}$ . Next we need to determine how much this new data point improved our classifier. We then make an estimate of the gain in terms of risk difference using the probability to estimate the ground truth.

$$\Delta \hat{R}_{\mathcal{E}}(f^{\mathcal{L}^+}, f^{\mathcal{L}}, \mathcal{L}^+) = \hat{R}_{\mathcal{E}}(f^{\mathcal{L}^+}, \mathcal{L}^+) - \hat{R}_{\mathcal{E}}(f^{\mathcal{L}}, \mathcal{L}^+) \quad (2.17)$$

$$= \frac{1}{|\mathcal{E}|} \sum_{x \in \mathcal{E}} \sum_{y \in \mathcal{Y}} \frac{(\mathbf{k}_x^{\mathcal{L}^+} + \alpha)_y}{\|\mathbf{k}_x^{\mathcal{L}^+} + \alpha\|_1} \cdot \left( L(y, f^{\mathcal{L}^+}(x)) - L(y, f^{\mathcal{L}}(x)) \right) \quad (2.18)$$

The observations used to estimate the risk are the same for both the old and new classifiers. We do this because we assume that adding labeled data will make the classifier better, so this allows us to more accurately compare the current classifier and the new one.

## 2.5 Expected Probabilistic Gain

If we are able to reduce the error with the new  $\mathcal{L}^+$  model then equation 2.18 will be negative. As a result, we negate this term and maximize the expected probabilistic gain. To simplify we set  $\alpha = \beta$ .

$$\text{xgain}(x_c, \mathcal{L}, \mathcal{E}) = \mathbb{E}_{p(y_c | \mathbf{k}_{x_c}^{\mathcal{L}})} \left[ -\Delta \hat{R}_{\mathcal{E}}(f^{\mathcal{L}^+}, f^{\mathcal{L}}, \mathcal{L}^+) \right] \quad (2.19)$$

$$= - \sum_{y \in \mathcal{Y}} \frac{(\mathbf{k}_x^{\mathcal{L}} + \beta)_y}{\|\mathbf{k}_x^{\mathcal{L}} + \beta\|_1} \cdot \frac{1}{|\mathcal{E}|} \sum_{x \in \mathcal{E}} \sum_{y \in \mathcal{Y}} \frac{(\mathbf{k}_x^{\mathcal{L}^+} + \alpha)_y}{\|\mathbf{k}_x^{\mathcal{L}^+} + \alpha\|_1} \cdot \left( L(y, f^{\mathcal{L}^+}(x)) - L(y, f^{\mathcal{L}}(x)) \right) \quad (2.20)$$

Finally, for the xPAL selection strategy, we simply choose this candidate  $x_c^* \in \mathcal{U}$  where the gain is maximized:

$$x_c^* = \arg \max_{x_c \in \mathcal{U}} (\text{xgain}(x_c, \mathcal{L}, \mathcal{E})). \quad (2.21)$$

## 3. Data Review

In this chapter we take a deeper look into the data and the process of collecting, translating, and encoding the data. Our partner has provided a small sample of 1000 labeled data points. This data was manually labeled by an human annotator.

### 3.1 Overview

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 from the original data set.

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

The annotator simply needs to view then mentally process the text and images from the website and make some reasonable decisions on how the site should be classified. However, the annotator does not record what the content on the site said or what drove them to make their decision. As a result we are missing a key portion of data for the classification process.

Tags are also when labeling the data to provide further granularity. The merchant tags are ordered by specificity, 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 Tag	Level 2 Tag	Level 3 Tag	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.

The tags are important because they allow us to separate the data even further and group or sort the data differently. However, in our work we didn't opt to include the tags in the classification or selection strategy process at this time.

## 3.2 Collection

Similarly to the annotator our goal is to automate the navigation, collection/storing process, and classification of the website. This process speeds up the browsing process by navigating to websites that have already been labeled and scraping the text data and storing it into a database.

The initial 1000 data points we received were labels with a pointer (a url) to where the text data is located (the website). 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 only to scrape the top level (main or home) page text because of the results published in another study where it was observed that adding more pages to the data set does not necessarily mean obtaining better results (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. Two websites were particularly problematic. Facebook and Instagram both are used by business as main information webpages. However, neither site allows for simple text scraping and a more advanced approach was needed to extract data from business with information on these platforms. Out of the remaining 821 data points 274 of them were in English. Out of the remaining 274 English data points the data was distributed into the categories as shown in Figure 3.1.

## 3.3 Processing

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 data set. 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 (Pedregosa et al. [2011]). This holds true for our data set as well.

At this stage, it was clear that our data set 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, see Figure 3.1. 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 and test sets.

At this point we had a significant amount of data that wasn’t being used (the untranslated data). We decided to find a way to translate the existing data. We tried various libraries available on GitHub but 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 viable option and we were able to use this API to translate the remaining data. We limited the number of characters to 1000 per data point from the scraped text to avoid maxing out the API.

An example of the first 100 characters of raw scraped text data from a website is shown in Table 3.3. The scraped text data is a single string of text that is a concatenation of all the text data pulled from the website url.

We can see that the symbols were removed and the majority of the words

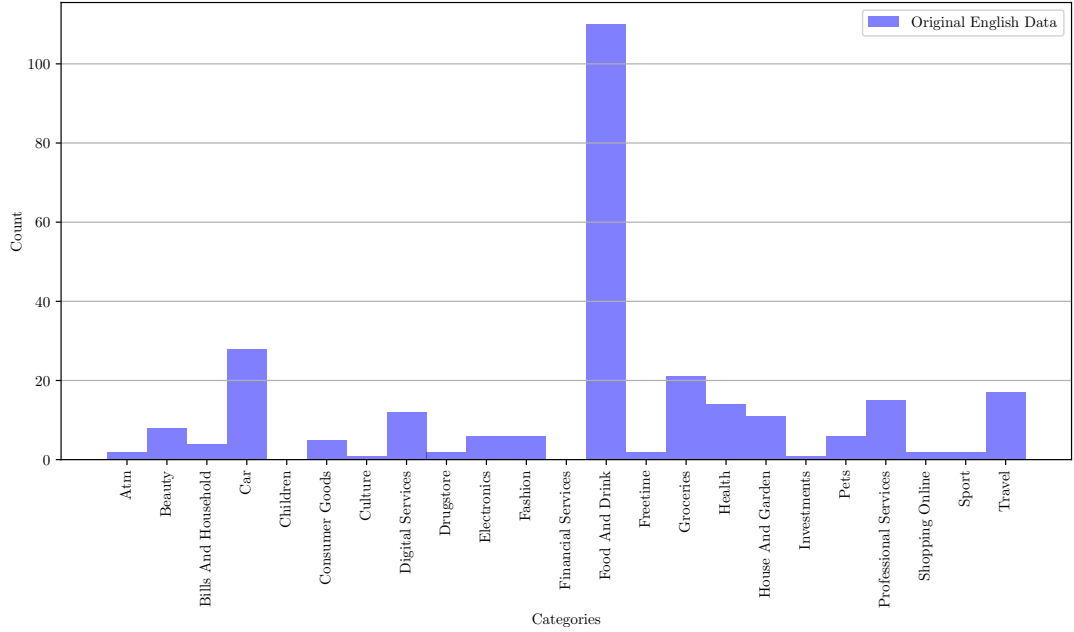


Figure 3.1: The histograms for the original usable english data.

Table 3.3: Raw text collected by scraper and the translated text.

Raw	DentalVision - Profesionální soukromá zubní klinika v centru Hradce Králové ÚvodSlužby a ceníkOrdina
Translated	DentalVision Professional private dental clinic in the center of Hradec Králové IntroductionServices

were translated. There are still some issues with words being concatenated, but we attempt to break these words apart before passing the text to the TF-IDF vectorizer.

In addition to the original data we also manually collected and labeled 141 additional data points. All the original data and additional data counts are shown in Figure 3.2 and discretely in Appendix Table A.1.

After translating the data we used the TF-IDF vectorizer from Scikit-Learn. We were able to find highly correlated words for each category using Chi Squared analysis with only the original data, shown below in Table 3.4. Some categories such as 'Culture', 'Digital Services', 'Shopping Online' that have few data points have words such as 'kihnu', 'td', 'patria', respectively, which have no relative meaning to the category in English.

From what we discussed in the previous section, we can see that the 'Culture' category has only one data point and the 'Digital Services' category has only two data points. This is problematic because if the single data point we have is doesn't represent the category well then we will continue to have difficulty classifying until we have more robust data.

We also calculated the variable importance using the RandomForestRegressor from Scikit-Learn and provided a list of the top 20 most important words from the TF-IDF vectorizer. This helps orient ourselves within the data and check if there may be any anomalies. This list can be found in Appendix Table A.2.

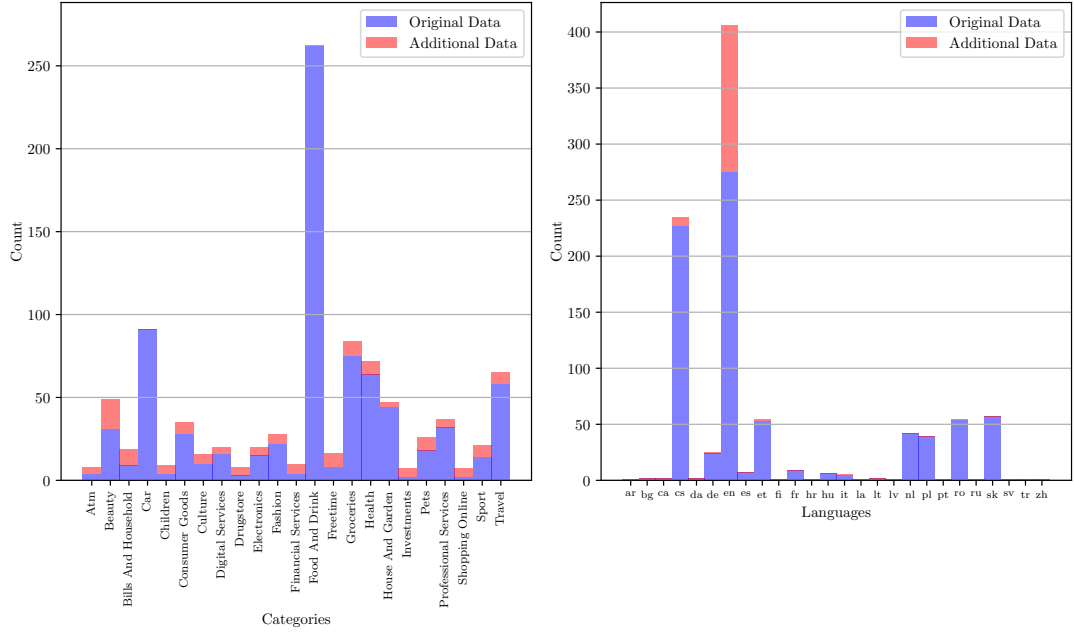


Figure 3.2: The histograms for the original and additional data for all languages.

Table 3.4: Keywords from TF-IDF with Chi Squared using the original data.

	Keyword 1	Keyword 2	Keyword 3
Atm	banking	individuals	caixa
Beauty	hairdressing	hairdresser	hair
Bills And Household	liberty	internet	fullness
Car	cars	auto	car
Children	toy	sold	toysrus
Consumer Goods	kiosk	flowers	flower
Culture	theater	museum	kihnu
Digital Services	synnex	bitly	servers
Drugstore	esodrogeria	detergent	drimble
Electronics	laptops	onoff	computers
Fashion	rings	jewellery	women
Financial Services	pre	nissan	insurance
Food And Drink	cafe	bar	restaurant
Freetime	representation	casino	likely
Groceries	liquor	maso	bakery
Health	drug	dental	pharmacy
House And Garden	skylights	paints	hardware
Investments	mistakes	investment	patria
Pets	mat	pet	veterinary
Professional Services	toilet	faculty	parcel
Shopping Online	owner	default	joom
Sport	adidas	singltrek	functional
Travel	rooms	accommodation	hotel

## 4. Testing

### 4.1 Original Data

Here we will explore the results of different classifiers and active learning strategies on the original data. The original data consisted of data shown previously in Figure 3.1 and discretely in Table A.1 in Appendix A.

#### 4.1.1 Active Learning with PWC and RBF Kernel

In Figure 4.1 we have the train and test errors for four different active learning sampling strategies. XPAL appears to perform slightly better than PAL but not significantly better. With our data we found that PAL runs slower than xPAL, which was unexpected based on the mean computation time results published by Kottke et al. [2021]. We expect this drop in calculation time is a result of the high dimensional data.

We weren't satisfied with the testing error which leveled out to about 70% for each sampling strategy. PAL and xPAL were able to rapidly reduce the testing error early on in the training process while random selection and QBC weren't able to determine the data with the highest information gain.

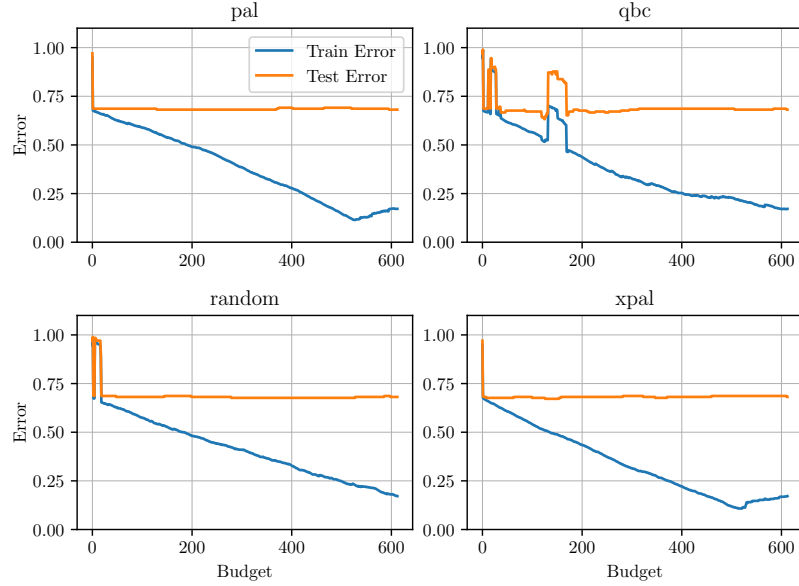


Figure 4.1: Train and test error using different query strategies and RBF kernel for the PWC classifier.

The radial bias function (RBF) kernel is a popular kernel function. It is defined as:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (4.1)$$

where  $\sigma$  is a parameter that controls the smoothness of the kernel and  $x_i$  and  $x_j$  are the two points in the feature space to compare. As seen in the Figure 4.1

when the PWC classifier uses the RBF kernel it doesn't perform well with this data.

Starting with the data in Figure 4.1 we made an error while constructing the  $x$  dataset. While preparing the data, the TF-IDF vectorizer was fitted using all the data instead of just the training data. This resulted in the TF-IDF matrix having more information about the test data than it should have so our testing error results are likely lower than they should be. This error persists throughout the rest of the experiments in this chapter up until the experiments in Section 4.2.

### 4.1.2 Active Learning with PWC and Cosine Kernel

In Figure 4.2 we have the train and test errors for the same four different active learning sampling strategies tested on the same data. The only change was that we used Cosine kernel instead of the RBF kernel. The Cosine kernel is another important kernel function that is used in many machine learning algorithms. It is defined as:

$$K(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|} \quad (4.2)$$

where  $x_i$  and  $x_j$  are the two points in the feature space to compare. We found that using the Cosine kernel reduced the test error across the board by  $\sim 15\%$ .

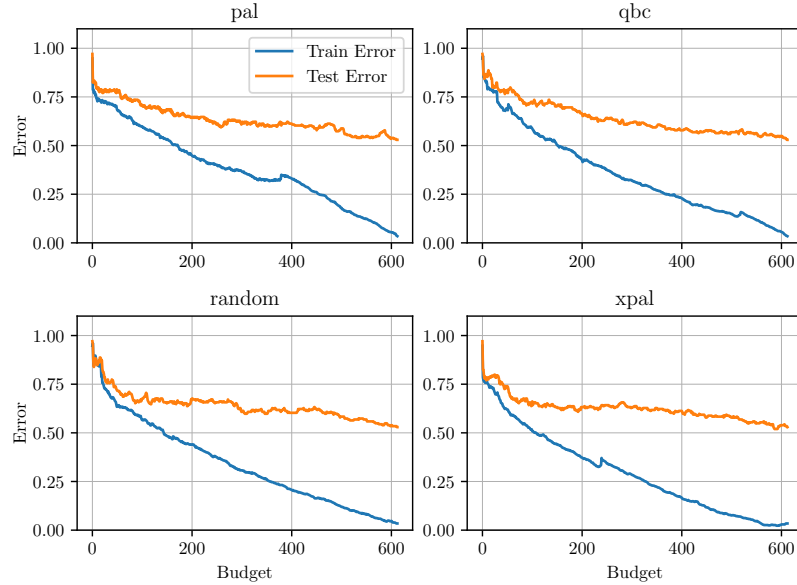


Figure 4.2: Train and test error using different query strategies and Cosine kernel for the PWC classifier.

In Figure 4.2 PAL and xPAL were able to reduce the training error, by about 20% and 25% respectively, early in the training process compared to random selection and QBC. We also tested the other sampling strategies with the Cosine kernel and found that the results were similar. The other sampling strategies and their test data results are shown in Table 4.3 along with the test data from Figure 4.2.



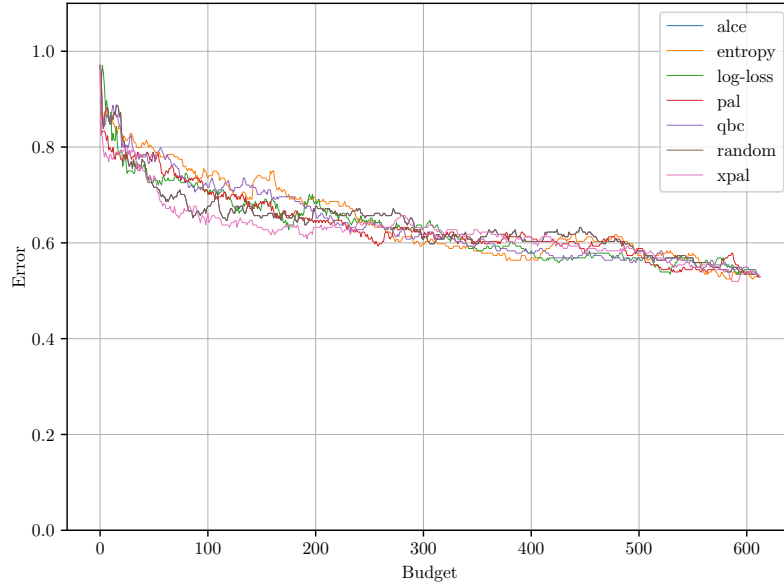


Figure 4.3: Comparing test error with one data split using different query strategies and Cosine kernel for the PWC classifier.

We can see that the sampling strategies test performance converges over time (as we are using the same data and classifier) but xPAL appears to have an absolute minimum near the 600 budget mark in comparison to all sampling strategies. XPAL also appears to be performing well early on in the training process, in the 100-200 budget range. However, this test is only showing the results of one data split. We can get a better idea if we run more tests with different train-test splits and see how the results average out.

Figure 4.3 shows that xPAL seems to be performing the best with our data but we wanted to see if we ran more tests with different train-test splits how the results would average out and which sampling strategy would perform the best on average. We ran 10 different data splits with each of the 7 sampling strategies and then took the average to get a smoother curve compared to the single run results shown in Figure 4.3. The results for this experiment are shown in Figure 4.4.

It is clear in Figure 4.4 that xPAL is performing the best early on (budget from 0-100) in the sample selection process. XPAL selects the data that minimizes the test error and builds the strongest classifier quickly while it takes the other sampling strategies more data points to get to the same level of performance. Around the 100 budget mark we can see that the other selection strategies catch up to xPAL performance wise.

### 4.1.3 Classifier Evaluation

We also decided to test out some classifiers from the Scikit-Learn library to compare performances. Again we used the original data with the same TF-IDF vectorizer as used with the previous active learning models to stay consistent. It should be noted that cross validation was used here for evaluation but it was not used in the previous sections.

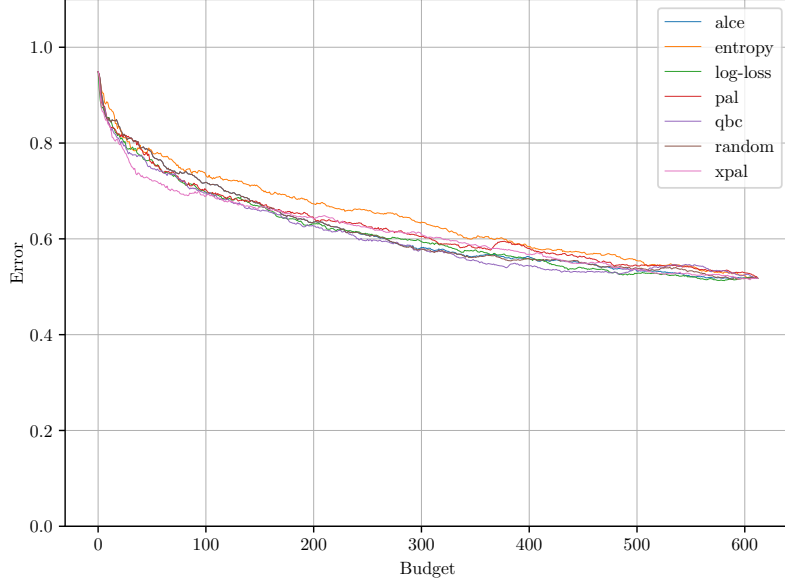


Figure 4.4: Comparing test error using different query strategies and Cosine kernel for the PWC classifier with results averaged over ten different data splits.

The goal of this exploratory phase was to try and decide which classifier to conduct more thorough testing with. As a result, we didn't use GridSearchCV for each classifier at this stage and we mostly used the default parameters and their cross validation scores with all of the original data (i.e. additional data not included). In some cases where using weights was an option for the classifier we included the precomputed Cosine decay weights. A table of the parameters used for each classifier is shown in Appendix Table A.4.

Initially we made a minor error while evaluating the classifiers that used the precomputed weights in the following experiments. While computing the accuracy scores for the classifiers, we failed to incorporate the class weights that were used in training the classifier. This resulted in the accuracy scores being incorrectly calculated. This error persists throughout the rest of the experiments in this chapter up until the experiments in Section 4.2.

The results for the different classifiers are shown in Figure 4.5. In the box-plot, the whiskers extend from the box to the furthest data points that are within 1.5 times the inter-quartile range (IQR) of the box. Any data points that are beyond the whiskers are considered outliers and are plotted as individual points or symbols (diamonds) as seen in the figure.

The base LinearSVC classifier model performed best compared to other classifiers and it is a fast running algorithm even with data that has many features. We decided to look further into LinearSVC because it performed so well. We also wanted to conduct more testing with KNeighborsClassifier and Neural Networks.

We created three models for LinearSVC, the first was a boilerplate LinearSVC with no argument modifications, the second model used the class weights parameter set to 'balanced'. The 'balanced' mode uses the values of  $y$  to automatically adjust weights inversely proportional to class frequencies in the input data as  $n\_samples / (n\_classes * np.bincount(y))$ .

For the third test we created a dictionary of weights for each class using the

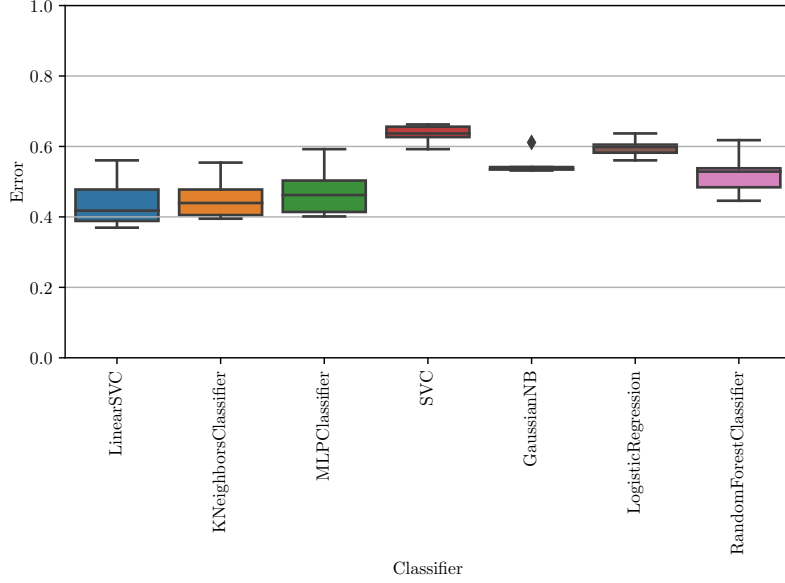


Figure 4.5: Performance of classifiers without GridSearchCV optimization.

Cosine decay function. The weights for each category ranged from 0.1 to 1.0 where the most frequent classes had smaller weights. The Cosine decay function is defined as:

$$w_i = \frac{1}{2} \left( 1 + \cos \left( \frac{\pi t}{T} \right) \right) \quad (4.3)$$

where  $w_i$  is the weight for the  $i^{th}$  class,  $t$  is the current iteration, and  $T$  is the total number of iterations. The Cosine decay function is a common function used for weights in machine learning algorithms. The calculated cosine weights are shown in the appendix in Table A.5.

The results for the LinearSVC classifier experiments are shown in Table 4.1 where the LinearSVC with the Cosine decay weights didn't perform better than the other LinearSVC models.

Table 4.1: Error for three differing LinearSVC models.

Model	Error
Boilerplate One-vs-rest	0.426
Crammer Singer	0.447
Cosine Decay Weights One-vs-rest	0.486

Using K-Neighbors Classifier (KNN) and a Neural Network from TensorFlow we conducted additional experiments. For the KNN we found that using 8 neighbors and the cosine distance metric provided the lowest error.

For the Neural Network we used a dense hidden-layer with 1000 neurons with Sigmoid activation and 23 output neurons with Softmax activation. For the NN optimizer we used Adamax with Cosine decay with an initial learning rate of 0.1, alpha value of 0.1 and 915 decay steps ( $X\_data\_size // batch\_size * num\_epochs$ ). The results are shown in Table 4.2. We can see that the K-Nearest Neighbors

classifier and the Neural Network classifier performed slightly worse compared to LinearSVC.

The LinearSVC outperformed the other classifiers and we decided to experiment with it further. We attempted to boost performance of the LinearSVC classifier using multiple cross validation grid searches with the bagging. Bagging (bootstrap aggregating) is a type of ensemble learning, where multiple models are trained on different subsets of the training data and their predictions are combined to make the final prediction. In Scikit-Learn we used the BaggingClassifier to implement bagging. An example of our setup and parameters are shown in the code snippet.

```
base_classifier = LinearSVC()
bagging_classifier = BaggingClassifier(
    base_estimator = base_classifier,
    n_estimators = 20,
    random_state = args.seed)
params = {
    'base_estimator__random_state': [args.seed],
    'base_estimator__max_iter': [10000],
    'base_estimator__intercept_scaling':
        np.linspace(0.1, 1, 20),
    'base_estimator__loss': ['hinge', 'squared_hinge'],
    'base_estimator__penalty': ['l1', 'l2'],
    'base_estimator__class_weight': [None, weights],
    'base_estimator__C': np.linspace(0.1, 1000, 50),
    'base_estimator__multi_class': ['ovr',
        'crammer_singer']
}
```

Performance was not improved from what we had already seen. Using bagging may not be the best approach at this stage because there are some categories that have very few samples so bagging may be unable to create a good model. We will revisit bagging in future chapters when we have more data at our disposal. We also didn't use the balanced class weights parameter because we had already seen that it was the worst performing class weight parameter in previous tests. The BaggingClassifier and GridSearchCV combination didn't improve the performance of the LinearSVC classifier beyond what we had already achieved.

Table 4.2: Testing errors for best performing classifiers.

Model	Error
LinearSVC	0.426
Tensor Flow Neural Network	0.452
K Neighbors Classifier	0.462

The precision-recall curve for the best performing classifier (LinearSVC) is shown in Figure 4.6 and the confusion matrix is shown in Figure 4.7 for the best performing LinearSVC classifier. The precision-recall curve gives us an idea at how well our classifier can correctly categorize the data. It also gives us a visualization

of how unbalanced our categories are. We can see this imbalance clearly in Figure 4.6 where we have straight lines and large steps for some categories. This is a result of having a small number of data in a class. However, we can also see that for some classes the precision is relatively high even though we have few data points. Here we are namely concerned with the 'Culture' and 'Beauty' categories which have 10 and 31 data points respectively. It may be that the keywords in the 'Culture' and 'Beauty' categories are drastically different from the other categories so the performance is better.

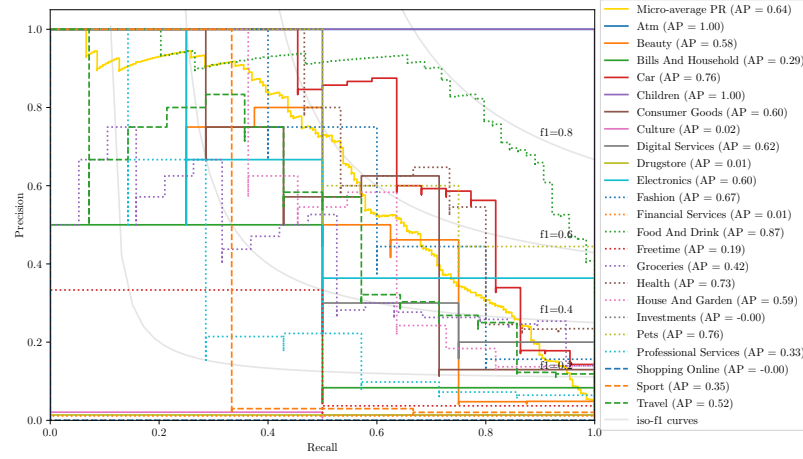


Figure 4.6: Precision-recall curve for the best performing LinearSVC classifier.

The confusion matrix shown in Figure 4.7 may be a better metric for visualizing this data. In addition, the classification report for the LinearSVC classifier is shown in Appendix Table A.3 with F1, accuracy, precision, recall, and support scores.

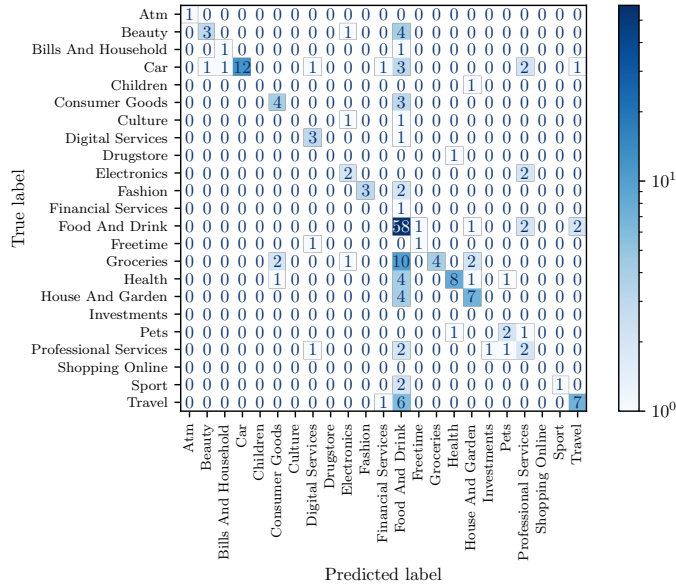


Figure 4.7: Confusion matrix for the best performing LinearSVC classifier using the original data.

## 4.2 All Data

In this section we will take what we have learned from the experiments in the previous sections and apply that knowledge to the original data set augmented with the additional data that we collected. This will provide us with a better understanding of how the data is interacting with the active learning sampling strategies and the PWC classifier. We will also removing some data from the set if the text length is below some threshold, evaluate the performance with a reduced number of categories, and explore the performance of xPAL with using all available data.

### 4.2.1 Active Learning with Additional Data

In this experiment we tested all the active learning methods with the corrected TF-IDF vectorizer transformation. In this section instead of incorrectly vectorizing the data then importing it to be used with Kottke et al. [2021] probabilistic active learning code, we exported the raw text data and then split the text data, fit and transformed the vectorizer to the train data, then transformed the test data and conducted our experiments with the modified code.

We then used the different active learning methods and PWC with the Cosine kernel to run our experiments. We ran 10 test runs for each method and then averaged the test error of all the runs resulting in a single curve for each sampling strategy. We only used the first 300 data points instead of using all available data. As a result, the test error doesn't converge in this plot because we are only using a subset of the data. The results are shown in Figure 4.8.

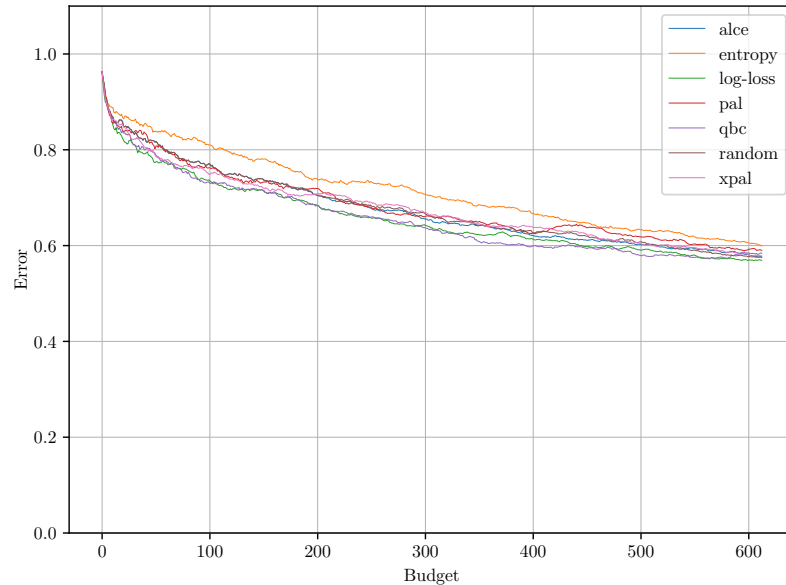


Figure 4.8: Active learning results with using all data.

Its clear that active learning sampling strategies found it more difficult with this additional data compared to our previous runs. We can clearly see that the xpal-original outperformed all other sampling strategies.

When making more data available to the classifier we assumed that the test error would decrease. However, it may be the case that the additional data is not as useful as we thought.

### 4.2.2 LinearSVC with Additional Data

We have some data that may not have enough text to classify correctly and others that have more than 1000 characters of text that may be noisy. We also know that sometimes while scraping the text data from a website we collected a non empty string that in reality provided no words that were related to the label.

However, using all available data, we now have the luxury of having more than the minimum of 2 data points in some categories and can afford to remove data from the set that may only have two or three strings in the text. We decided to explore if filtering the data based on text length could improve the performance of the classifier and use LinearSVC as it has performed well with this data previously and it is relatively fast to train.

We imported the data and either selected a minimum number of characters or we altered the maximum number of characters allowed and created a new dataset. For each run we selected the data based on this criteria and then built the TF-IDF array. For the minimum string size tests we incremented the string size by 1 character. While for the max string size tests we decremented the string size by 10. We again used a train test split of 25% which has been our standard for testing throughout our experiments. We found that around the 200 character mark we would filter out too much data and would not have a minimum representation (2 data points) for all the categories. The results for this experiment are shown in Figure 4.9.

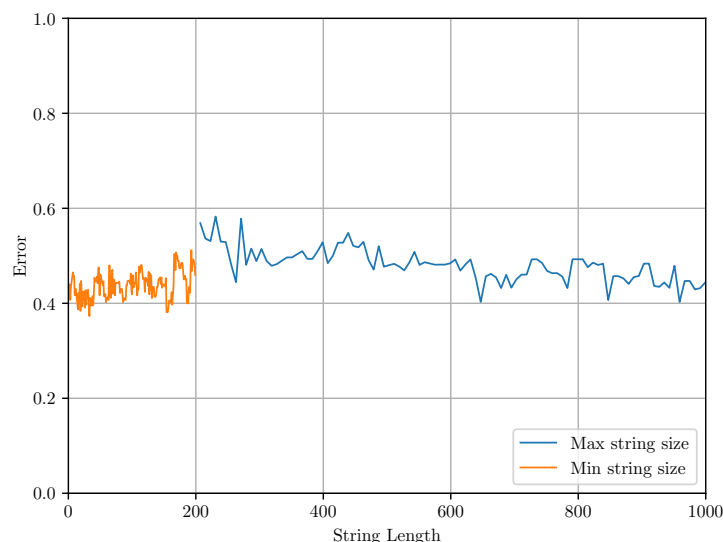


Figure 4.9: Test error results for best LinearSVC with varying text length parameters.

We also tested with a minimum text length and a maximum text length constraint implemented together but this combination did not yield any obvious

gains.

This naive approach to filtering the data did not yield any obvious improvements in the test error. At best it may have filtered out some bad data and provided a small improvement in the test error. However, when implementing xPAL we hope to have it filter out the bad data in a more calculated manner.

We can see that for the 414 data points results from the filtered data in the train split converge as we expect to a single point because the PWC has access to all of the labeled train data.

### 4.2.3 Active Learning Revisited

With the new filtered data set we also wanted to see how the PWC classifier with active learning would perform. We used the same text length constraints as mentioned in the previous section and evaluated the data with the PWC classifier with the Cosine kernel.

We chose to test with the QBC selection strategy in addition to xPAL because QBC appeared to perform well in previous tests and it is much faster to calculate in comparison to PAL. The results show that the PWC was able to make some small improvements in the test error overall but also that the new data allowed the xPAL selection strategy to decrease the test error even more rapidly than before when we used the original data, and it took less data.

TALK ABOUT DISTRIBUTION USING ORIFINAL AND ALL DATA INSERT PLOT HERE

### 4.2.4 Active Learning with Less Categories



# Conclusion

The entire process of data collection, storage, retrieval, processing, and testing allowed us to create a customized pipeline and explore active learning using website text data to classify online merchants and other websites. The quality of the collected data wasn't perfect but we were able to use the data to create a model that could classify websites into 1 of 23 categories with  $\sim 60\%$  accuracy using a linear support vector classifier. We also discovered that the xPAL sampling strategy discovered the best samples quickly to reduce the testing error the most.

Pairing an optimal classifier with xPAL allows us to create a pipeline to choose the most informative data points to add to the training set and reduce the amount of time it takes to train the model as well as the size of the classifier.

Many sites had multiple languages present on their websites and this could be seen in our scraped data. This posed a challenge for us because when using the API for translation it made it more difficult to detect the language while also adding non english words into the TF-IDF.

## Improvements

To improve the results of the active learning classifier we would suggest making a number of changes. One of the first changes would be to find a better way to profile a website and make sure the quality of the text data from the websites were better. For example, a stronger web scraper would have allowed us to avoid potential IP address restriction issues and scrape data from social media websites that refused to allow our scraper to collect any data and resulted in some unusable data. We could have also run our own tests regarding how the number of sibling pages could have fortified the data.

The scraper could be designed to scrape multiple pages of a website, scrape social media pages, and possibly scrape other websites that are linked to the original website. This could improve the quality of the data and allow for more accurate classification. Prioritizing the quality of the data is one of the key points to improving the performance of the classifier. It also seems that a small amount of good text data is better than a large amount of poor quality text data. Another way we could have made the scraped data better for the classifier is to take out all non english words (post translation).

To be more thorough, we could have run exhaustive tests for *all* the available classifiers within Scikit-Learn using GridSearchCV and other ensemble methods to see if any more performance gains were attainable. We could have also explored the performance of TensorFlow RNN's and LSTM's for text classification.

The current setup of the active learning xPAL process uses the PWC as it is fast (because it updates only parts of the classifier with each new data point) and relatively simple to implement. However, it is not the best method for classifying our data as we have seen from our experiments. We could improve the pipeline by implementing xPAL with other Scikit-Learn classifiers and running tests to choose the best active learning classifier for our data.

# Bibliography

- Yoram Baram, Ran El Yaniv, and Kobi Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5(Mar):255–291, 2004.
- Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133, 1997.
- 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.
- David D Lewis and William A Gale. A sequential algorithm for training text classifiers. *arXiv preprint cmp-lg/9407020*, 1994.
- Robert Munro. *Human-in-the-loop machine learning*. Manning Publications, New York, NY, July 2021.
- 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*, 12:2825–2830, 2011.
- 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 Oppel, 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. . . . .	17
3.2	The histograms for the original and additional data for all languages.	18
4.1	Train and test error using different query strategies and RBF kernel for the PWC classifier. . . . .	19
4.2	Train and test error using different query strategies and Cosine kernel for the PWC classifier. . . . .	20
4.3	Comparing test error with one data split using different query strategies and Cosine kernel for the PWC classifier. . . . .	21
4.4	Comparing test error using different query strategies and Cosine kernel for the PWC classifier with results averaged over ten different data splits. . . . .	22
4.5	Performance of classifiers without GridSearchCV optimization. . .	23
4.6	Precision-recall curve for the best performing LinearSVC classifier.	25
4.7	Confusion matrix for the best performing LinearSVC classifier using the original data. . . . .	25
4.8	Active learning results with using all data. . . . .	26
4.9	Test error results for best LinearSVC with varying text length parameters. . . . .	27

# List of Tables

2.1	Variable names and descriptions. . . . .	11
3.1	This is an example of a single data point from the original data set.	15
3.2	This is an example of how the tags use different levels. . . . .	15
3.3	Raw text collected by scraper and the translated text. . . . .	17
3.4	Keywords from TF-IDF with Chi Squared using the original data.	18
4.1	Error for three differing LinearSVC models. . . . .	23
4.2	Testing errors for best performing classifiers. . . . .	24
A.1	Category counts of usable given data sorted by originally given, originally given data that is English, originally given translated to English (+ English data), and supplementary data. . . . .	34
A.2	Variable importance, top 20 words from the vectorizer. . . . .	35
A.3	Best LinearSVC model classification report. . . . .	36
A.4	Set of parameters used to test all other Scikit-Learn classifier shown in Figure 4.5. . . . .	37
A.5	Cosine decay weights for each category. . . . .	38

# List of Abbreviations

# A. Attachments

Table A.1: Category counts of usable given data sorted by originally given, originally given data that is English, originally given translated to English (+ English data), and supplementary data.

Category	Original	Original English	Translated	Additonal
Atm	8	2	4	4
Beauty	49	8	31	18
Bills And Household	19	4	9	10
Car	91	28	91	0
Children	9	0	4	5
Consumer Goods	36	5	28	7
Culture	16	1	10	6
Digital Services	20	12	16	4
Drugstore	8	2	3	5
Electronics	20	6	15	5
Fashion	28	6	22	6
Financial Services	10	0	4	6
Food And Drink	265	110	262	0
Freetime	16	2	8	8
Groceries	85	21	75	9
Health	73	14	64	8
House And Garden	47	11	44	3
Investments	7	1	2	5
Pets	26	6	18	8
Professional Services	37	15	32	5
Shopping Online	7	2	2	5
Sport	21	2	14	7
Travel	65	17	58	7

Table A.2: Variable importance, top 20 words from the vectorizer.

	Importance
hotel	0.091429
car	0.053077
hair	0.029760
auto	0.028876
station	0.017934
hairstylist	0.016447
flower	0.015114
spa	0.013156
parking	0.013083
stations	0.012124
barber	0.011965
internet	0.011843
services	0.011364
rental	0.010331
salon	0.009946
accommodation	0.009681
service	0.009430
bank	0.009072
cookies	0.007105
gas	0.006985

Table A.3: Best LinearSVC model classification report.

	precision	recall	f1-score	support
Atm	0.00	0.00	0.00	1.00
Beauty	0.75	0.38	0.50	8.00
Bills And Household	0.00	0.00	0.00	2.00
Car	0.63	0.55	0.59	22.00
Children	0.00	0.00	0.00	1.00
Consumer Goods	1.00	0.29	0.44	7.00
Culture	0.00	0.00	0.00	2.00
Digital Services	1.00	0.50	0.67	4.00
Drugstore	0.00	0.00	0.00	1.00
Electronics	0.50	0.25	0.33	4.00
Fashion	1.00	0.40	0.57	5.00
Financial Services	0.00	0.00	0.00	1.00
Food And Drink	0.49	0.94	0.64	64.00
Freetime	0.00	0.00	0.00	2.00
Groceries	0.50	0.21	0.30	19.00
Health	0.73	0.53	0.62	15.00
House And Garden	0.50	0.45	0.48	11.00
Investments	0.00	0.00	0.00	0.00
Pets	1.00	0.50	0.67	4.00
Professional Services	0.50	0.14	0.22	7.00
Shopping Online	0.00	0.00	0.00	0.00
Sport	0.00	0.00	0.00	3.00
Travel	0.70	0.50	0.58	14.00
micro avg	0.55	0.55	0.55	197.00
macro avg	0.40	0.24	0.29	197.00
weighted avg	0.57	0.55	0.51	197.00



Table A.4: Set of parameters used to test all other Scikit-Learn classifier shown in Figure 4.5.

Classifier	Parameters
LinearSVC	{'C': 1.0, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'loss': 'squared_hinge', 'max_iter': 10000, 'multi_class': 'ovr', 'penalty': 'l2', 'random_state': 42, 'tol': 0.0001, 'verbose': 0}
KNeighborsClassifier	{'algorithm': 'auto', 'leaf_size': 30, 'metric': 'cosine', 'metric_params': None, 'n_jobs': None, 'n_neighbors': 8, 'p': 2, 'weights': 'uniform'}
MLPClassifier	{'activation': 'relu', 'alpha': 0.0001, 'batch_size': 'auto', 'beta_1': 0.9, 'beta_2': 0.999, 'early_stopping': False, 'epsilon': 1e-08, 'hidden_layer_sizes': 500, 'learning_rate': 'constant', 'learning_rate_init': 0.001, 'max_fun': 15000, 'max_iter': 100, 'momentum': 0.9, 'n_iter_no_change': 10, 'nesterovs_momentum': True, 'power_t': 0.5, 'random_state': 42, 'shuffle': True, 'solver': 'adam', 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': False, 'warm_start': False}
SVC	{'C': 1.0, 'break_ties': False, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 'scale', 'kernel': 'rbf', 'max_iter': -1, 'probability': False, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}
GaussianNB	{'priors': None, 'var_smoothing': 1e-09}
LogisticRegression	{'C': 1.0, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 1000, 'multi_class': 'auto', 'n_jobs': None, 'penalty': 'l2', 'random_state': 42, 'solver': 'lbfgs', 'tol': 0.0001, 'verbose': 0, 'warm_start': False}
RandomForestClassifier	{'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 200, 'n_jobs': None, 'oob_score': False, 'random_state': 42, 'verbose': 0, 'warm_start': False}

Table A.5: Cosine decay weights for each category.

	Weight
Shopping Online	1.000
Investments	0.996
Drugstore	0.983
Atm	0.963
Financial Services	0.934
Children	0.899
Freetime	0.857
Bills And Household	0.810
Culture	0.757
Sport	0.701
Electronics	0.642
Digital Services	0.581
Pets	0.519
Fashion	0.458
Consumer Goods	0.399
Beauty	0.343
Professional Services	0.290
House And Garden	0.243
Travel	0.201
Health	0.166
Groceries	0.137
Car	0.117
Food And Drink	0.104