CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Nuclear Sciences and Physical Engineering

Active Learning for Text Classification

Aktivní učení pro klasifikaci textů

Masters's Degree Project

Author:	**Marko Sahan**
Supervisor:	**doc. Ing. Václav Šmídl, Ph.D.**
Academic year:	2019/2020

Acknowledgment:

I would like to thank **doc. Ing. Václav Šmídl, Ph.D.** for his expert guidance and express my gratitude to .... for (his/her language assistance).

Author's declaration:

I declare that this Masters's Degree Project is entirely my own work and I have listed all the used sources in the bibliography.

Prague, July 8, 2019 Jméno Autora

Název práce:

Aktivní učení pro klasifikaci textů

Autor: Marko Sahan

Obor: Aplikované matematicko-stochastické metody

Druh práce: Diplomová práce

Vedoucí práce: **doc. Ing. Václav Šmídl, Ph.D.**, Ústav teorie informace a automatizace

Abstrakt:

Klíčová slova:

Title:

Active learning for text classification

Author: Marko Sahan

Abstract:

Key words:

Contents

Notation

Symbol	Definition
$\mathbf{x} \in \mathcal{X}$	vector of instance
$\mathbf{y} \in \mathcal{Y}$	one hot encoded label of a specific instance
$\mathbf{X} = \{\mathbf{x}_1, \dots \mathbf{x}_M\}$	set of available instances
$\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_M\}$	set of labels that can be provided by an annotator
$\tilde{\mathbf{X}}$	set of training instances
$\tilde{\mathbf{Y}}$	set of training labels
$[x_1, x_2, ..., x_S]^T$	transposed vector notation
$(\mathbf{x}, \mathbf{y})$	tuple notation
$\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_R\}$	set notation
$a \in \mathcal{A}$	action from a set of all possible actions
$\theta \in \Theta$	decision theory uncertainty parameter
L	loss function
π^*	probability density function of variable θ
$p(\mathbf{y}\|\mathbf{x})$	probability density function of label **y** given instance **x**
$\hat{\mathbf{y}}$	estimate of **y** given **x**
$\mathbf{w}$	vector of SVM or decision tree weights
b	SVM bias (scalar)
$\mathbf{W}$	matrix of Naive Bayes, random forest or neural network single layer parameters
$\mathbf{b}$	vector of neural network single layer biases
Ω	set of neural network parameters (all weights and biases)
δ	Dirac delta function

Introduction

Active Learning strategy lets the machine learning models iteratively and strategically query the labels of some instances for reducing human labeling efforts. This project shows how it is possible to connect active learning and text data. People has been already solving same problem for anomaly detection [6], image processing [8], etc..

If we take a look on modern approach of automating the labeling process of huge amount of unlabeled data is not optimal. People are randomly choosing unlabeled text data. These data are annotated by the subject matter experts and used for training and testing the models. If the model performance is weak after the training, more text documents are selected and annotated. This approach is costly because nobody knows how much text documents is needed to have good model scores. Our active learning strategy proposes selection of unlabeled text data that the model is not certain about. Unlabeled text data are given to a subject matter expert to provide the labels. This problem has already been solving for long period of time. Some active learning approaches for text classification dates back to 2001 [24] where are shown different querying strategies and results of the active learning superiority over randoms sampling strategies. There is no clue that active learning strategy brings a lot advantages. First of all, we are able to start with lower amount of training data and iteratively extend the dataset. The dataset is extended using the data, which the model is not certain about. In this work we are extending our dataset with only one sample per active learning iteration. However, it was also show that strategies which sample batches with more than one sample also can perform good results [2]. Thus, basing on the active learning approach the model will get much more information from non-randomly chosen text samples.

The project describes how we can formulate different algorithms with respect to decision theory and then connect all the methods to active learning theory. All the methods used in this work can be represented in ensembles way. Basing on [23] it was show that ensembles deep learning algorithms show the best performance both for Text and Image Processing data. Plenty of related active learning works for text classifications follow [8] with using acquisition functions and dropout algorithm for uncertainty representations e.g. named entity recognition [22] and text classification [13], [5]. We are concerned that ensembles outperform dropout uncertainty representation for text data as it was shown in [11] for image classification.

This project also provides link to python implementation of active learning algorithms and comparison of different results gathered with respect to different data. We believe that active learning approach is able to significantly reduce amount of time and expenses needed for automating text labeling process.

Chapter 1

Introduction to Decision Theory

The process of decision making is defined as a selection of the optimal action from a set of possibilities that can be applied at some operating conditions. The criteria of optimality is formalized by a loss function. The process of selection of the optimal action is thus formalized as the optimization problem. However, the operating conditions are often not known exactly since we have incomplete information about them. Therefore, we will use the theory of decision making under uncertainty [3], where the uncertainty is represented by probability density functions. We will now briefly review the theoretical background.

1.1 Decision Theory

The theory of decision making has three basic elements: i) the set of possible actions $\mathcal{A}$ from which we should select an optimal action $a^* \in \mathcal{A}$, ii) the vector $\theta \in \Theta$ defining operating conditions under which we make the decision, where Θ is parameters space, iii) the loss function

$$L = L(\theta, a), \tag{1.1}$$

that defines our preference of the action, in the sense that action a which has the lowest value from the action set of the loss function is preferred. For complete knowledge of the operating conditions θ the task is turned into simple optimization of (1.1). However, with incomplete information, we have to consider a range of possible states θ. The theory of decision making under uncertainty [3] defines the expected loss function that takes into account the uncertainty.

Definition 1. [3]If $\pi^*(\theta)$ is believed probability distribution of θ at the time of decision making, the *Bayesian expected loss* of an action a is

$$\rho(\pi^*, a) = \mathbb{E}_{\pi^*}[L(\theta, a)] = \int_{\Theta} L(\theta, a)\pi^* d\theta. \tag{1.2}$$

Based on definition 1.2, the optimal action is defined as the one that minimizes the expected loss:

$$a^* = \operatorname*{argmin}_{a\in\mathcal{A}} \mathbb{E}_{\pi^*}[L(\theta, a)]. \tag{1.3}$$

The key task of application of decision theory is the choice of the action space, parameter space, loss function and method of evaluating the probability measure. In the following Sections, we discuss examples of application of the theory to the problem of supervised learning and active learning, respectively.

1.2 Decision Theory for Supervised Learning

Supervised learning is defined as learning from the data with known target value. Specifically, for the classification problem, the target value is the class where each data point belongs.

We would like to commence our formal definition with the data. Let $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$ and $\mathbf{y} \in \mathcal{Y} = \{[0,1]^T, [1,0]^T\}$, where $\mathbf{x}$ is feature vector of size n and $\mathbf{y}$ is its label assigned to the data instance $\mathbf{x}$ from space $\mathcal{X}$. Each value from space $\mathcal{Y}$ can be represented as a one hot representation which is a vector consisting from ones and zeros. In the case of binary classification $\mathbf{y} \in \{[0,1]^T, [1,0]^T\}$ where, the first class is represented as $\mathbf{y} = [1,0]^T$ and the second class is represented as $\mathbf{y} = [0,1]^T$. As a good example of previous definition, $\mathbf{x}$ can be a text document (represented in a mathematical form in order to meet a definition above) and $\mathbf{y}$ can be its category such as sports or comedy. As seen from this example, the label and text are forming a tuple. In this work we are considering our data as tuples of variables $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$.

Basing on the data definitions from previous part, we can assume that $\mathcal{X} \times \mathcal{Y}$ is an infinite set and $(\mathbf{x}, \mathbf{y})$ is a sample from this set. We assume that all available data tuples are sampled independently from a joint probability density function $p(\mathbf{x}, \mathbf{y})$. If $p(\mathbf{x}, \mathbf{y})$ was known, the optimal classifier $p(\mathbf{y}|\mathbf{x})$ can be obtained by the chain rule of probability

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}). \tag{1.4}$$

However, since we do not know analytical form of the joint probability distribution we aim at selecting the best possible approximation within a chosen class. Specifically, we choose a parametric form $p(\mathbf{y}|\mathbf{x}, a)$ where a is the parameter to be optimized.

The uncertainty of the decision task is representation of the joint probability distribution. We will consider the uncertainty θ to be represented by empirical distribution:

$$\pi^* = p(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i, \mathbf{y} - \mathbf{y}_i) \tag{1.5}$$

where $\mathbf{x}_i, \mathbf{y}_i$ are elements of the training set ($\tilde{\mathbf{X}} \subset \mathcal{X}, \tilde{\mathbf{Y}} \subset \mathcal{Y}$). The training set is usually a subset of all available data on which the optimization is performed, the rest of the data is used for validation [25].

1.2.1 Decision Theory and Support Vector Machine Algorithm

In this subsection we will continue construction of the decision theory on the example of Support Vector Machine (SVM) method. For simplicity lets consider linearly separable dataset. From the theoretical perspective SVM constructs hyperplane in high dimensional space that separates two classes. In this case our decision (action) is a hyperplane that will separate two classes. Equation of the hyperplane can be written as $f(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T\mathbf{x} + b$ where $\mathbf{w} \in \mathbb{R}^n$ is a set of hyperplane parameters and $b \in \mathbb{R}$ is a bias. As a result, action space is represented as $(\mathbb{R}^n, \mathbb{R}) = \mathcal{A}$ and as a consequence tuple $(\mathbf{w}, b) \in \mathcal{A}$. From this knowledge we consider the uncertainty θ described with (1.5) that meets the condition of the limitation on $\Theta = (\tilde{\mathbf{X}} \subset \mathcal{X}, \tilde{\mathbf{Y}} \subset \mathcal{Y})$. Considering loss function (1.1) that can be written as

$$L = L(\mathbf{x}, \mathbf{y}, \mathbf{w}, b). \tag{1.6}$$

Following task is to understand how good is our action (hyperplane estimation) with respect to the dataset. We can choose different types of loss functions such as cross entropy, hinge loss, etc.. The most basic approach for SVM method is the hinge loss function [20] which is defined as

$$L(\mathbf{x}, \mathbf{y}, \mathbf{w}, b) = \max(0, 1 - y\hat{y}(\mathbf{x}, \mathbf{w}, b)) \tag{1.7}$$

where $\hat{y}(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T\mathbf{x} + b$ and $y = \mathbf{y}_1$.

In terms of SVM method we want to find such hyperplane that will label input values as a first class if it is "above" the hyperplane and as a second class if it is "below" the hyperplane. At this point very important assumption will be introduced. In order to find an optimal hyperplane we assume that the data $\tilde{\mathbf{X}}$ and its labels $\tilde{\mathbf{Y}}$ fully describe spaces $\mathcal{X}$ and $\mathcal{Y}$. Thus, the uncertainty of the decision task can be defined as (1.5).

Using (1.2) we can evaluate expected loss function for SVM as follows

$$\begin{aligned}
\mathbb{E}_{\pi^*}L &= \int_{\mathcal{X}\times\mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \mathbf{w}, b)p(\mathbf{x}, \mathbf{y})d(\mathbf{x}, \mathbf{y}), \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \max(0, 1 - y_1\hat{y}(\mathbf{x}, \mathbf{w}, b))\frac{1}{N}\sum_{i=1}^{N}\delta(\mathbf{x} - \mathbf{x}_i, \mathbf{y} - \mathbf{y}_i)d(\mathbf{x}, \mathbf{y}), \\
&= \frac{1}{N}\sum_{i=1}^{N}\max(0, 1 - y_{1,i}\hat{y}(\mathbf{x}_i, \mathbf{w}, b))
\end{aligned}$$

where $\hat{y}(\mathbf{x}_i, \mathbf{w}, b) = \mathbf{w}^T\mathbf{x}_i + b$ and $y_{1,i}$ is first component of $i-$ th vector $\mathbf{y}_i$ Expect loss function for SVM can be written as

$$\rho(\mathbf{x}_i, \mathbf{w}, b) = \frac{1}{N}\sum_{i=1}^{N}\max(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)). \tag{1.8}$$

1.2.2 Decision Theory and Algorithm Based on Neural Network Function

1.2.2.1 Neural Network

Given the instance $\mathbf{x} \in \mathcal{X}$ with further application of Feed Forward Neural Network we are able to predict an output value $\mathbf{y} \in \mathcal{Y}$. In this part of the work our prior interest is around Feed Forward Neural Network algorithm that assigns input value to a specific class.

First layer of NN is defined as

$$\mathbf{a}_1 = \mathbf{W}_1^T\mathbf{x} + \mathbf{b}_1 \tag{1.9}$$

where $\mathbf{W}_1$ is matrix of weights and $\mathbf{b}_1$ is a vector of bias values. First layer is called the input layer.

Further layers of NN are formed as

$$\mathbf{a}_k = \mathbf{W}_k^T f(\mathbf{a}_{k-1}) + \mathbf{b}_k,\ k = \{2, ..., K-1\}. \tag{1.10}$$

As seen from equation (1.10), neurons from each layer (except of input layer) take linear combination of the neurons from the previous layer. Function f is an activation function. Activation function is defined as a non-decreasing, continuous function. The most commonly used activation functions are sigmoid, relu, elu, and hyperbolic tangence functions [4].

Output values are computed with

$$\hat{\mathbf{y}} = f_{sm}\left(\mathbf{W}_K^T f(\mathbf{a}_{K-1}) + \mathbf{b}_K\right). \tag{1.11}$$

where f_{sm} is the softmax function that is typically used for classification problems. The softmax function is defined as

$$f_{sm,i} = \frac{\exp(\mathbf{z}_i)}{\sum_{i=1}^{2}\exp(\mathbf{z}_i)},$$

where

$$\mathbf{z} = \mathbf{W}_K^T f(\mathbf{a}_{K-1}) + \mathbf{b}_K$$

is an output vector before activation function is applied. Output vector has same size a label $\mathbf{y}$.

1.2.2.2 Decision Theory

Decision Theory construction for the algorithm, based on a neural network function, is mostly the same as in 1.2.1. However in this case our decision is to find estimate $\hat{\mathbf{y}} = \hat{\mathbf{y}}(\mathbf{x}, \Omega)$ of the probability density function $p(\mathbf{y}|\mathbf{x})$ where $\mathbf{x}$ is the input data, $\Omega = \{\mathbf{W}_1, ...\mathbf{W}_K, \mathbf{b}_1..., \mathbf{b}_K\}$ is a set of all neural network function parameters and biases. Action space $\mathcal{A}$ will be parameters' and biases' space of $\hat{\mathbf{y}}$. Same as in 1.2.1 we can define $(\mathbf{x}, \mathbf{y})$ are parameters of the loss function and $\mathcal{X} \times \mathcal{Y}$ is a parameters' space. Another example of loss functions that we will use is cross entropy loss function, which is defined as

$$L(\mathbf{x}, \mathbf{y}, \Omega) = -y_1 \ln(\hat{y}_1(\mathbf{x}, \Omega)) - y_2 \ln((\hat{y}_2(\mathbf{x}, \Omega)), \tag{1.12}$$

where $\mathbf{y} = [y_1, y_2]^T$ and $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2]^T$. With the usage of the given dataset where $\forall i \in \{1, .., N\}$, $(\mathbf{x}_i, \mathbf{y}_i) \in (\tilde{\mathbf{X}} \subset \mathcal{X}, \tilde{\mathbf{Y}} \subset \mathcal{Y})$ are independent identically distributed we can approximate $p(\mathbf{x}, \mathbf{y})$ as (1.5). Applying definition (1), expected loss for the algorithm based on a neural network function is evaluated as

$$\begin{aligned}
\mathbb{E}_{\pi^*} L &= \int_{\mathcal{X}\times\mathcal{Y}} L(\mathbf{x}, y, \Omega) p(\mathbf{x}, \mathbf{y}) d(\mathbf{x}, \mathbf{y}), \\
&= -\int_{\mathcal{X}\times\mathcal{Y}} \Big(y_1 \ln(\hat{y}_1(\mathbf{x}, \Omega)) + y_2 \ln((\hat{y}_2(\mathbf{x}, \Omega))\Big) \frac{1}{N} \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i, \mathbf{y} - \mathbf{y}_i) d(\mathbf{x}, \mathbf{y}), \\
&= -\frac{1}{N} \sum_{i=1}^{N} \Big(y_1 \ln(\hat{y}_1(\mathbf{x}, \Omega)) + y_2 \ln((\hat{y}_2(\mathbf{x}, \Omega))\Big),
\end{aligned} \tag{1.13}$$

where $\mathbf{y} = [y_1, y_2]^T$ and $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2]^T$.

1.2.2.3 Parameters Estimation

In further sections we are going to introduce more methods based on Neural Networks which will slightly differ between each other. Thus, we would like to cover more theory around parameters estimation. The very simple but efficient method is gradient descent. This method is using equation

$$\hat{\Omega}_{n+1} = \hat{\Omega}_n - \eta_n \nabla L(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \hat{\Omega}_n, Z_n) \tag{1.14}$$

where $\hat{\mathbf{\Omega}}_n$ is the $n-$th iteration value of gradient descent of NN weights and its biases that converges to $\hat{\Omega}$. Value of $\nabla L(\mathbf{X}, \mathbf{Y}, \hat{\Omega}_n, Z_n)$ is a gradient of the loss function and η_n is $n-$th iteration of a value that in terms of NN is defined as learning rate with a decay. Term Z_n represents indices from X and Y that are used in $n-$th iteration of loss function. Learning rate decay is not obligatory feature. It can be constant as well. However, it is very strong feature that increases performance. In this work we are using ADAM optimization [10] which includes learning rate decay.

The usual gradient descent is extremely efficient method for complex $\hat{\mathbf{y}}(\mathbf{x}, \Omega)$ but may struggle with local minima. In this case, the algorithm may stop iterating even in a very shallow local minimum. Due to the complex functions $\hat{\mathbf{y}}(\mathbf{x}, \Omega)$, the loss function of its approximation will be non convex with large amount of local minima and maxima. Solution of this problem is that algorithm must jump over or walk around the local minimum. Stochastic gradient descent (SGD) [4] lets us to provide this operations. In comparison to standard gradient descent the key difference is that SGD allows us to use only one piece of data (minibatch) from the training dataset in order to calculate the step [4]. The minibatch is represented with value Z_n that says which training indices to use. The data samples (minibatch) is picked randomly at each step.

1.2.3 Decision Theory and Naive Bayes Algorithm

Naive Bayes algorithm is a bit different to the algorithm based on Neural Networks and SVM. In the case of Naive Bayes we want to estimate $p(\mathbf{W}|\mathbf{x},\mathbf{y})$ where $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n\} \subset \mathcal{W}$ is an action ($a = \mathbf{W},\ a \in \mathcal{A}$). The reason why we look for an estimate of the $p(\mathbf{W}|\mathbf{x},\mathbf{y})$ but not $p(\mathbf{y}|\mathbf{x},\mathbf{w})$ is due to the fact that in case of estimating $p(\mathbf{y}|\mathbf{x},\mathbf{W})$ we would have to work with normalization constant would be dependent on the set of parameters $\mathbf{W}$. That fact would make our computations very complicated. Before continuing with a loss function construction we would like to go trough Naive Bayes (NB) method.

1.2.3.1 Naive Bayes

Consider a binary classification problem. With the usage of the Bayes rule we can rewrite $p(\mathbf{W}|\mathbf{x},\mathbf{y})$ as follows

$$p(\mathbf{W}|\mathbf{x},\mathbf{y}) = \frac{p(\mathbf{y})p(\mathbf{x}|\mathbf{y},\mathbf{W})p(\mathbf{W})}{\int_{\mathcal{W}} p(\mathbf{x},\mathbf{y}|\mathbf{W})p(\mathbf{W})d\mathbf{W}} \tag{1.15}$$

where $\mathcal{W}$ is the space of possible values of $\mathbf{W}$.

Naive Bayes method introduces very strong assumption in equation (1.15). This assumption says that features of vector $\mathbf{x} = [x_1, x_2, ..., x_n]^T$ are conditionally independent. As a result estimation of $p(\mathbf{W}|\mathbf{x},\mathbf{y})$ can be estimated as

$$\tilde{p}(\mathbf{W}|\mathbf{x},\mathbf{y}) = \frac{1}{Z}p(\mathbf{y})p(\mathbf{W})\prod_{i=1}^{n}(p(x_i|y_1,\mathbf{w}_i)^{y_1}p(x_i|y_2,\mathbf{w}_i)^{y_2}), \tag{1.16}$$

where $\mathbf{y} = [y_1, y_2]$, $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n\}$, Z is a normalizing constant.

1.2.3.2 Decision Theory

We want to maximize probability $\tilde{p}(\mathbf{W}|\mathbf{x},\mathbf{y})$. As a result, using (1.16), loss function L will be represented as

$$\begin{aligned} L(\mathbf{y},\mathbf{x},\mathbf{w}) &= -\log(\tilde{p}(\mathbf{W}|\mathbf{x},\mathbf{y}), & (1.17)\\ &= \log(Z) - \log(p(\mathbf{y})) - \log(p(\mathbf{W})) - \sum_{i=1}^{n}\log(p(x_i|y_1,\mathbf{w}_i)^{y_1}p(x_i|y_2,\mathbf{w}_i)^{y_2}). & (1.18)\end{aligned}$$

Same as in 1.2.1 and 1.2.2 we will assume that we can approximate $p(\mathbf{x},\mathbf{y})$ as 1.5. From this moment everything is ready for deriving expected loss function. Expected loss function for Naive Bayes method is derived as

$$\begin{aligned} &\mathbb{E}_{\pi^*}L = \int_{\mathcal{X}\times\mathcal{Y}} L(\mathbf{x},\mathbf{y},\mathbf{w})p(\mathbf{x},\mathbf{y})d(\mathbf{x},\mathbf{y}),\\ &\int_{\mathcal{X}\times\mathcal{Y}}\Big(\xi(\mathbf{W},\mathbf{y}) - \sum_{i=1}^{n}\log(p(x_i|y_1,\mathbf{w}_i)^{y_1}p(x_i|y_2,\mathbf{w}_i)^{y_2}\Big)\frac{1}{N}\sum_{j=1}^{N}\delta(\mathbf{x}-\mathbf{x}_j,\mathbf{y}-\mathbf{y}_j)d(\mathbf{x},\mathbf{y}),\\ &\frac{1}{N}\sum_{j=1}^{N}\Big(\xi_j(\mathbf{W},\mathbf{y}_j) - \sum_{i=1}^{n}\log(p(x_{i,j}|y_{1,j},\mathbf{w}_i)^{y_{1,j}}p(x_{i,j}|y_{2,j},\mathbf{w}_i)^{y_{2,j}})\Big) \end{aligned} \tag{1.19}$$

where $\xi(\mathbf{W}, \mathrm{y}) = \log(Z) - \log(p(\mathbf{y})) - \log(p(\mathbf{W}))$ and $\xi_j(\mathbf{W},\mathbf{y}_j) = \log(Z) - \log(p(\mathbf{y}_j) - \log(p(\mathbf{W}))$.

1.2.4 Decision Theory and Random Forest Algorithm

In order to work with random forests we must precisely define decision trees and only then construct theory for random forests.

1.2.4.1 Decision Tree

In this section we expect our decision three to give us an estimate $\hat{\mathbf{y}}(\mathbf{x}, \mathbf{w}) \in \{[0, 1]^T, [1, 0]^T\}$ where $\mathbf{w}$ is a vector that describes tree (depth, branches, etc.), $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. It is very important to mention that for different trees $\mathbf{w}$ can have different dimensionality. Thus, for consistency we will assume that for all $\mathbf{w} \in \mathcal{W}$ exists upper bound, where $\mathcal{W}$ is redefined as a space of tree parameters. As a result we will make all $\mathbf{w}$ to have the same length. If $\mathbf{w}$ has spare elements, they will be filled with zeros. Parameter space will be same as in 1.2.1-1.2.3, whereas action $a \in \mathcal{A}$ will be represented as $\mathcal{A} = \mathcal{W}$. In order to understand when our tree is optimal we can use zero-one loss function. Zero-one loss function is defined as

$$L(\mathbf{x}, \mathbf{y}, \mathbf{w}) = \begin{cases} 1, & \mathbf{y} \neq \hat{\mathbf{y}}(\mathbf{x}, \mathbf{w}) \\ 0, & \mathbf{y} = \hat{\mathbf{y}}(\mathbf{x}, \mathbf{w}) \end{cases}. \tag{1.20}$$

With the usage of the given data where $\forall i \in \{1, .., N\}$, $(\mathbf{x}_i, \mathbf{y}_i) \in \tilde{\mathbf{X}} \times \tilde{\mathbf{Y}}$ are independent identically distributed we can approximate $p(\mathbf{x}, \mathbf{y})$ as (1.5). As a result, expected loss function for a decision tree can be derived as

$$\begin{aligned}
\mathbb{E}_{\pi^*} L &= \int_{\mathcal{X} \times \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \mathbf{w}) p(\mathbf{x}, \mathbf{y}) d(\mathbf{x}, \mathbf{y}), \\
&= \int_{\mathcal{X} \times \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \mathbf{w}) \frac{1}{N} \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i, \mathbf{y} - \mathbf{y}_i) d(\mathbf{x}, \mathbf{y}), \\
&= \frac{1}{N} \sum_{i=1}^{N} L(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w})
\end{aligned}$$

where $\sum_{i=1}^{N} L(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w})$ is (1.20).

If we want to minimize expected loss we have to follow next steps. While constructing a decision tree we choose such feature $x_i \in [x_1, ..., x_n]^T = \mathbf{x}$ that will bring the highest information about the system. This feature will form first layer, then we add another feature with the highest information gain and construct second layer. Using this method, we construct nodes and add more and more layers (branches).

In the following part, we are going to work with a set of decision trees. For this purposes we will define our decision tree as $\hat{\mathbf{y}} = [T_1(\mathbf{x}, \mathbf{w}_l), T_2(\mathbf{x}, \mathbf{w}_l)]^T$ where index l represents set of parameters for $l-$th three and indices $1, 2$ represent first and second value of the one-hot vector.

1.2.4.2 Random Forest

Considering $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ as random variables with joint probability density function $p(\mathbf{x}, \mathbf{y})$. We will also assume that $\forall i \in \{1, .., N\}$, $(\mathbf{x}_i, \mathbf{y}_i) \in \tilde{\mathbf{X}} \times \tilde{\mathbf{Y}}$ are independent identically distributed.

With the usage of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ we will make $\{1, ..., V\}$, $V \in \mathbb{N}$ sets where $\forall v \in V$, $\tilde{\mathbf{X}}_v \subset \tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}_v \subset \tilde{\mathbf{Y}}$. The data $\tilde{\mathbf{X}}_v$ and $\tilde{\mathbf{Y}}_v$ are created with random uniform sampling from $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$. We also want each subset to contain strictly 80% of the data from $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$. As a result parameter space for random forests will form tuples of sets $(\tilde{\mathbf{X}}_v, \tilde{\mathbf{Y}}_v)$. Using this theory we will construct L decision trees $\hat{y} = T(\mathbf{x}, \mathbf{w}_l)$, where

$\mathbf{x} \in \tilde{\mathbf{X}}_l$. As a result for $v-$ th decision tree

$$\mathbb{E}_{\pi^*} L = \frac{1}{N_l} \sum_{i=1}^{N_l} L(\mathbf{x}_{i,v}, \mathbf{y}_{i,v}, \mathbf{w}_l) \delta(\mathbf{x} - \mathbf{x}_{i,v}, \mathbf{y} - \mathbf{y}_{i,v}) \tag{1.21}$$

where $(\mathbf{x}_{i,v}, \mathbf{y}_{i,v}) \in (\tilde{\mathbf{X}}_v, \tilde{\mathbf{Y}}_v)$ and N_v is a number of the data in $\tilde{\mathbf{X}}_v$ and $\tilde{\mathbf{Y}}_v$. If we assume $\mathbf{w}_v$ as a random variable then L decision trees form samples from probability density function $p(\mathrm{y}|\mathbf{x}, \mathbf{w})$. In other words

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}_v) = T_1(\mathbf{x}, \mathbf{w}_v)^{y_1} T_2(\mathbf{x}, \mathbf{w}_v)^{y_2} \tag{1.22}$$

where label $\mathbf{y}$ is written as a one-hot representation $\mathbf{y} = (y_1, y_2)^T$. Thus, we can say that classification probability $p(\mathbf{y}|\mathbf{x})$ can be written as

$$p(\mathbf{y}|\mathbf{x}) = \int_{\mathbf{w} \in \mathcal{A}} p(\mathbf{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}. \tag{1.23}$$

where $\mathcal{A}$ is an action space. With the usage of samples $\mathbf{w}_v$ we can approximate $p(\mathbf{y}|\mathbf{x})$ as

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{V} \sum_{l=1}^{V} T_1(\mathbf{x}, \mathbf{w}_v)^{y_1} T_2(\mathbf{x}, \mathbf{w}_v)^{y_2} \tag{1.24}$$

where each decision tree $[T_1(\mathbf{x}, \mathbf{w}_v), T_2(\mathbf{x}, \mathbf{w}_v)]^T$ is constructed with the use of (1.21).

Before continuing with further sections we define output of the Random Forest as $\hat{\mathbf{y}} = \hat{\mathbf{y}}(\mathbf{x}, \mathbf{W})$, where $\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_V\}$ is set of parameters of specific Random Forest algorithm. We define vector $\hat{\mathbf{y}}$ as

$$\hat{\mathbf{y}} = \frac{1}{V} \sum_{v=1}^{V} (T_1(\mathbf{x}, \mathbf{w}_v), T_2(\mathbf{x}, \mathbf{w}_v)). \tag{1.25}$$

1.3 Decision Theory for Active Learning

As mentioned in previous sections $\tilde{\mathbf{X}} \times \tilde{\mathbf{Y}}$ is the training set. When we finish a model training, we may think that we need more training data. Thus, we can choose the data from $\mathbf{X} \subset \mathcal{X} \backslash \tilde{\mathbf{X}}$. However it is important to understand that we have no labels for the set $\mathbf{X}$. We can ask for a help from an annotator that can give us those labels. We assume that getting labels needs some time and is very expensive.

Active learning problem is defined as a sequence of Supervised learning problems. Specifically, we assume that labels $\mathbf{y} \in \tilde{\mathbf{Y}}$ are available only for $\mathbf{x} \in \tilde{\mathbf{X}}$. We have the possibility to select an unlabeled element from $\mathbf{X}$ and ask for its label. Since it is expensive, we aim to have such questions that will maximize scores as fast as possible. Formally, we denote $J_0 = \{j_{01}, j_{02}, \ldots j_{0N}\} = \{1, \ldots, N\}$ the initial set of N available labels. using only the labeled data, the supervised learning task is defined on sets $\mathbf{X}_0 = \{\mathbf{x}_i\}_{i \in J_0}$ and $\mathbf{Y}_0 = \{\mathbf{y}_i\}_{i \in J_0}$. This set is sequentially extended with new labels gained from $\mathbf{X}$. We consider a sequence of U questions $u = \{1, \ldots, U\}$, in each question, we select an index j_u and ask to obtain the label $\mathbf{y}_{j_u}$ for data record $\mathbf{x}_{j_u}$. The index set and the data sets are extended as follows

$$J_u = \{J_{u-1}, j_u\}, \qquad \mathbf{X}_u = \{\mathbf{X}_{u-1}, \mathbf{x}_{j_u}\}, \qquad \mathbf{Y}_u = \{\mathbf{Y}_{u-1}, \mathbf{y}_{j_u}\}.$$

The task of active learning is to optimize the selection of indices j_u to reach as good classification metrics with as low number of questions as possible. As a result we have to define the expected loss for each question u that will be dependent on the action and parameter spaces. In this case we can define our action space $\mathcal{A}_u$ as a space of the data indices with respect to the parameters space Θ_u for each question

u. Parameters space is defined as a set of possible parameters from a specific model. It is very important to understand that we will need not only one set of parameters but parameters distribution. We need parameters distribution because we will integrate over the parameters space. As an example, if we talk about SVM method, then parameters space for active learning problem will be defined as a set of weights that form a hyperplane. If we talk about the algorithm that is based on a Neural Network function, then parameters space of the active learning problem will form weights from neurons. We wanted to highlight that parameters space will be different for each problem but the idea for each algorithm is the same. .

The decision task for this particular problem can be written as

$$j_u^* = \underset{j \in J \setminus J_u}{\operatorname{argmin}} (\mathbb{E}_{\pi_u^*} L^*) \tag{1.26}$$

where $\mathbb{E}_{\pi_u^*} L^*$ is the expected loss that is dependent on an action given question u, and J is the space of all indices. The expected loss for the active learning problem is defined as

$$\mathbb{E}_{\pi_u^*} L^* = \int_{\Theta_u} L^*(a, \theta) \pi_u^* d\theta \tag{1.27}$$

where $a \in \mathcal{A}_u$, $\theta \in \Theta_u$ ($\theta = (\mathbf{w}, \mathbf{b})$ for SVM, $\theta = \Omega$ for NN, $\theta = \mathbf{W}$ for RF and NB) and L^* is a loss function for the active learning problem. Character "$*$" is used only for distinguishing active learning loss from the loss function which is used for different models. We will specify action space because it will be the same for all models that are used in the active learning section. Action space $\mathcal{A}$ is a set of possible indices $j_u \in J_u$ where u is a specific question. Thus, (1.27) can be written as

$$\mathbb{E}_{\pi_u^*} L^* = \int_{\Theta_u} L^*(j_u, \theta) \pi_u^* d\theta. \tag{1.28}$$

Using this approach we will be able sequentially select indices from $\mathbf{X}$ and ask for a label from $\mathbf{Y}$ that will help us to get higher scores faster than in the case of random choice of indices.

1.3.1 Bayesian Approach of Classifiers' Parameters Sampling

Considering that $\mathbf{y} \in \mathcal{Y}$. Let $\hat{\mathbf{y}} = \hat{\mathbf{y}}(\mathbf{x}_{j_u}, \theta_u)$ is an estimate of $\mathbf{y}$. However, in this case output estimate $\hat{\mathbf{y}}$ is represented as a vector of probabilities that $\mathbf{x}_{j_u}$ is assigned to different classes. As an example for a well trained binary classifier, for specific $\mathbf{x}$ that is assigned to $\mathbf{y} = [1, 0]^T$, classifiers estimate of $\mathbf{x}$ can be $\hat{\mathbf{y}} = [0.95, 0.05]^T$. It is very interesting that before we can solve the optimization problem with choosing the index j_u we have to solve the optimization problem of finding $\hat{\mathbf{y}}$. This leads us to supervised learning models that we have discussed in previous sections.

In this section we would like to construct theory around π_u^* from equation (1.28). Mentioned distribution is a distribution of the models' parameters given the training data that can be written as

$$\pi_u^* = p_u(\theta_u | \mathbf{X}_u, \mathbf{Y}_u). \tag{1.29}$$

We do not have explicit form of the pdf. However, we assume that we have Q_u samples $\theta_{u,q} \in \{1_u, ..., Q_u\}$ from $p_u(\theta_u | \mathbf{X}_u, \mathbf{Y}_u)$. As a result $p_u(\theta_u | \mathbf{X}_u, \mathbf{Y}_u)$ can be approximated as

$$\pi_u^* = \frac{1}{Q_u} \sum_{q=1}^{Q_u} \delta(\theta_u - \theta_u^{(q)}), \tag{1.30}$$

where $\delta(\theta_u - \theta_u^{(q)})$ is Dirac delta function centered in $\theta_u^{(q)}$.

1.3.1.1 Parameters Sampling Based on Training Data Subsets

This method is quite general and can be applied to all types of classifiers in this work (Random Forest, SVM, Neural Network). The idea is very simple. We consider that some data samples in training dataset $\tilde{\mathbf{X}} \times \tilde{\mathbf{Y}}$ are noise corrupted. Thus, it is obvious that we do not want our models to learn from noise corrupted data. As a result, we would like to randomly sample Q_u subsets from $\tilde{\mathbf{X}}$ with their labels from $\tilde{\mathbf{Y}}$. Lets rewrite it in more mathematical form.

Assuming N_u is amount of samples in $\mathbf{X}_u$. Let $Z_u = \{z_1, ..., z_{N_u^{sub}}\} \subset J_u$, where $N_u^{sub} < N_u$. Let $p(\theta_u | \mathbf{X}_u, \mathbf{Y}_u, Z_u)$ be a probability of model parameters θ_u given $\mathbf{X}_u, \mathbf{Y}_u$ and Z_u. The conditioning in the pdf is defined as restriction of sets $\mathbf{X}_u$, $\mathbf{Y}_u$ on indices from Z_u. As a result we can approximate (1.29) as

$$\begin{aligned} \pi_u^* &= p(\theta_u | \mathbf{X}_u, \mathbf{Y}_u) && (1.31) \\ &= \int p(\theta_u | \mathbf{X}_u, \mathbf{Y}_u, Z_u) p(Z_u) dZ_u && (1.32) \\ &= \frac{1}{Q_u} \sum_{q=1}^{Q_u} p(\theta_u | \mathbf{X}_u, \mathbf{Y}_u, Z_u^{(q)}) && (1.33) \\ &= \frac{1}{Q_u} \sum_{q=1}^{Q_u} \delta(\theta_u - \theta_u^{(q)}). && (1.34) \end{aligned}$$

Sampling from $p(Z_u)$ is very simple. The only thing that must be predefined is N_u^{sub}. After training the model using $\mathbf{X}_u$ and Y_u under restriction Z_u ,vector of model parameters will represent a single sample from π_u^*.

1.3.1.2 SGLD

Unlike previous section method, SGLD sampling is designed only for neural network based classifiers. SGLD modifies Neural Network learning algorithm by adding noise in Stochastic Gradient descent. In comparison to Normal Feed Forward Neural Network in [27] is introduced Stochastic Gradient Langevin Dynamics Neural Network (SGLD). This is type of Bayesian Neural Network algorithm. In comparison to normal Feed Forward NN, [27] proposes to add Gaussian noise to the system while doing gradient descent. In fact, when the algorithm is almost trained, additional noise lets us to sample i.i.d parameters values in a neighborhood of a minimum. With the usage of Bayes rule. we can rewrit (1.29) as

$$\begin{aligned} \pi_u^* &= p(\theta_u | \mathbf{X}_u, \mathbf{Y}_u) \\ &\propto p(\mathbf{X}_u, \mathbf{Y}_u | \theta_u) p(\theta_u). && (1.35) \end{aligned}$$

Next, we can approximate (1.35) as

$$\pi_u^* = \frac{1}{Q_u} \sum_{q=1}^{Q_u} p(\mathbf{X}_u, \mathbf{Y}_u | \theta_u^{(q)}) \qquad (1.36)$$

that results in (1.34). However, for this case θ must meet some constraints. If we want to sample parameters from its distribution θ must be independent identically distributed. Basing on [27] the updated gradient for SGLD can be written as

$$\begin{aligned} \hat{\Omega}_{n+1} &= \frac{N}{N_{minibatch}} \frac{\eta_n}{2} \Big(\hat{\Omega}_n - \eta_n \nabla L(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \hat{\Omega}_n, Z_n) \Big) + \epsilon_n, && (1.37) \\ \epsilon_n &\sim \mathcal{N}(0, \eta_n) && (1.38) \end{aligned}$$

where N is amount of training data, $N_{minibatch}$ is amount of the data in minibatch and η_n is $n-$th iteration of a learning rate.

1.3.1.3 Dropout

Dropout is another technique similar to SGLD how to sample from a neural network parameters distribution. This type of sampling can be also considered as a Bayesian Neural Network [8]. The idea is randomly turn off some neurons while training. When the algorithm is almost trained we are able to start sampling the i.i.d. parameters in the same way as it was done for SGLD case. Thus, an estimate of

$$\pi_u^* = p(\theta_u | \mathbf{X}_u, \mathbf{Y}_u)$$

can be written with the usage of 1.34.

1.3.1.4 DENFI

In case of DENFI algorithm we can approximate (1.29) with the usage of equations (1.35) and (1.34). The equations are same but sampling from $p(\theta_u)$ is different. In this work we are using modification of the initial DENFI algorithm. The idea of this algorithm is to train ensemble of Q_u Feed Forward Neural Networks with the usage of stochastic gradient descent. In theory, each neural network will find different local minimum due to different initial weights of neurons and stochastic gradient descent. The beauty of this algorithm is in further training iterations that will be described and shown in further sections. For now, we are only interested in parameters sampling from $p(\theta_u)$, that has been already covered and described.

1.3.2 Active Learning Loss Function

The acquisition function is a function that helps us to decide which data sample is the best for model learning given models uncertainty. Wide overview of different acquisition functions is shown in [8] e.g. entropy, information or mean standard deviation maximization. However in our work we decided to use only the entropy loss.

1.3.2.1 Entropy Based Active Learning Loss

First approach of defining Active Learning loss function is negative entropy. Basing of the formal definition of the entropy [21] we can write it as

$$-H(\hat{\mathbf{y}} | \mathbf{x}_{j_u}, \theta_u) = \sum_{r=1}^{R} \hat{y}_r(\mathbf{x}_{j_u}, \theta_u) \log (\hat{y}_r(\mathbf{x}_{j_u}, \theta_u)), \tag{1.39}$$

where $\hat{y}_r$ is $r-$th element of the output estimate $\hat{\mathbf{y}}$ and θ is a vector of parameters for specific model. As done in Passive Learning sections we want to find expected loss based on entropy function.

With the usage of previous knowledge we can derive expected entropy loss as

$$\begin{aligned}
\mathbb{E}_{\pi_u^*} L^* &= \int_{\Theta_u} -H(\hat{\mathbf{y}}|\mathbf{x}_{j_u}, \theta_u) p_u(\theta_u|\mathbf{X}_u, \mathbf{Y}_u) d\theta_u \\
&= \int_{\Theta_u} -H(\hat{\mathbf{y}}|\mathbf{x}_{j_u}, \theta_u) \frac{1}{Q_u} \sum_{q=1}^{Q_u} \delta(\theta_u - \theta_{u,q}) d\theta_u \\
&= \frac{1}{Q_u} \sum_{q=1}^{Q_u} -H(\hat{\mathbf{y}}|\mathbf{x}_{j_u}, \theta_{u,q})) \\
&= \frac{1}{Q_u} \sum_{q=1}^{Q_u} \sum_{r=1}^{R} \hat{y}_r(\mathbf{x}_{j_u}, \theta_{u,q}) \log(\hat{y}_r(\mathbf{x}_{j_u}, \theta_{u,q})). \qquad (1.40)
\end{aligned}$$

As a result, minimization of given expected loss will lead us to a sample with the highest entropy. Thus, we are seeking for a an index of maximal entropy data sample. In other words $j_u^* = \text{argmin}_{j \in J \setminus J_u}(\mathbb{E}_{\pi_u^*} L^*)$.

1.3.3 Active Learning

We would like to generalize active learning part for all described algorithms in order to estimates $p(\mathbf{y}|\mathbf{x}, \mathbf{X}_u, \mathbf{Y}_u)$ basing on samples from π_u^* in (1.29). In Supervised Learning section we have derived estimate $\hat{\mathbf{y}}$ of $\mathbf{y}$ for SVMs, Random Forests and Feed Forward Neural Networks. Active Learning algorithm requires distribution over the parameters of the algorithms. We will solve this problem the way that we will get samples from π_u^* and then approximate probability distribution as (1.29).

In order to estimate $p(\mathbf{y}|\mathbf{x}, \mathbf{X}_u, \mathbf{Y}_u)$ we define Generalized Ensemble Algorithm. SGLD and DENFI can be also represented as generalized ensembles models because parameters sampling (neuron weights sampling) represents different configurations of neural networks. Thus, each sample from π_u^* can be assumed as i.i.d. ensemble. As a result, we will use Q_u ensembles in each step of Active Learning algorithm. Therefore, basing on the previous theory we can approximate $p(\mathbf{y}|\mathbf{x}, \mathbf{X}_u, \mathbf{Y}_u)$ as

$$p(\mathbf{y}|\mathbf{x}, \mathbf{X}_u, \mathbf{Y}_u) = \frac{1}{Q_u} \sum_{q=1}^{Q_u} \hat{\mathbf{y}}_{q,u}. \qquad (1.41)$$

1.4 Conclusion

In this section we have covered decision theory for both passive and active learning with respect to different algorithms and ensemble approach. Passive Learning section showed how it is possible to view SVM, Random Forest and Neural Networks in terms of decision theory problem setup, whereas Active Learning section showed how to represent uncertainty of the model and parameters sampling for ensembles representation.

Chapter 2

Natural Language Processing Theory

2.1 Text Representation

According to [12], Natural Language Processing (NLP) is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

In this work we are focused on two techniques such as TF-IDF [19] and Fast Text Word Embeddings [15]. These methods are used for representation of text in a mathematical form (vectors, matrices). Even though TF-IDF is quite old method for text representation, it is still widely used. However, primary method, that is used in the thesis is the Fast Text Word Embeddings. In this project we are working with text documents (articles and tweets) and their labels. In the beginning of chapter 1 we defined value $\mathbf{x} \in \mathcal{X}$ as text features vector. By features vector we mean any kind of text encoding (TF-IDF, Fast Text Word Embedding, etc..).

2.1.1 TF-IDF

Term Frequency - Inverse Document Frequency (TF-IDF) is extremely powerful tool. This text encoding tool is quite simple and powerful. Method's advantage is its popularity. Plenty of packages in different programming languages have implementations of this algorithm. As mentioned in the name of this method, it is composed from two parts Term Frequency and Inverse Document Frequency. Term Frequency is defined as

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}},$$

where $f_{t,d}$ is number of times of word t in a document d. Inverse Document Frequency is defined as

$$IDF(t,d) = \log \frac{|D|}{|\{d \in D : t \in d\}|},$$

where numerator stand for total number of documents in the corpus and denominator is number of documents where the term t appears. We are assuming using only words that from corpus D. Thus, the denominator is always greater than zero.

Finally,

$$TF - IDF(t,d) = TF(t,d) \cdot IDF(t,d).$$

2.1.1.1 TF-IDF and Information Theory

In this part is shown the connection of TF-IDF to Information theory [1]. Lets first take a look on documents' entropy given word t,

$$\begin{aligned}
H(D|T=t) &= -\sum_{d} p(d|t)\log p(d|t) \\
&= \log \frac{1}{|\{d \in D : t \in D\}|} \\
&= -\log \frac{|\{d \in D : t \in D\}|}{D|} + \log |D| \\
&= -IDF(t,d) + \log |D|, \qquad (2.1)
\end{aligned}$$

where D is a documents' random variable and T is words' random variable. Equation (2.1) is correct under condition that we have no duplicate documents in the text corpus. Next step is to derive an equation of mutual information of documents and words as follows

$$\begin{aligned}
M(D,T) &= H(D) - H(D|T) \\
&= -\sum_{d} p(d)\log p(d) - \sum H(D|T=t)\cdot p(t)_t \\
&= \sum_{t} p(t)\cdot\Big(\log\frac{1}{|D|} + IDF(t,d) - \log|D|\Big) \\
&= \sum_{t} p(t)\cdot IDF(t,d) \\
&= \sum_{t,d} p(t|d)\cdot p(d)\cdot IDF(t,d) \\
&= \frac{1}{|D|}\sum_{t,d} TF(t,d)\cdot IDF(t,d). \qquad (2.2)
\end{aligned}$$

As seen from (2.2) TF-IDF has really good explanatory definition based on Information Theory. As a result, it is one more advantage of this method usage. However, here is one big disadvantage that can be very crucial. The higher amount of words is, the bigger and sparser vectors, that represent each document, will be.

2.1.2 Fast Text and CBOW Word Embeddings

Term "embedding" means a set of language modeling and feature learning techniques in natural language processing where words or phrases from the vocabulary are mapped to vectors of real numbers. Nowadays exist plenty of word embedding methods based on neural networks and co-occurrence matrices. Word embeddings are used as pretrained models. Words' encoding is used in order to encode the text and then text encoding is used for different purposes such as classification, clustering, etc..

The principle of word embeddings based on neural networks is explained in this section. We decided to describe Continuous Bag of Words Model (CBOW), because Fast Text word embeddings model is a modification of this method and CBOW covers all main theoretical aspects.

2.1.2.1 CBOW Word Embeddings

[14]

We would like to treat text {"The", "cat", 'over", "the', "puddle"} as a context and basing on the context we would like to predict or generate the center word "jumped". This type of model is Continuous Bag of Words (CBOW) Model. Let the known parameters in our model be the sentence represented with one-hot encoded word vectors. The input one hot encoded context vector is defined as $\mathbf{x}^{(c)}$, where c is position of a word in the context. The predicted context word is defined as $\mathbf{h}$. We create two matrices, $\mathcal{V} \in \mathbb{R}^{n \times |V|}$ and $\mathcal{U} \in \mathbb{R}^{|V| \times n}$. Where n is an arbitrary size which defines the size of our embedding space. $\mathcal{V}$ is the input word matrix such that the $i-$th column of $\mathcal{V}$ is the $n-$dimensional embedded vector for word w_i when it is an input to this model. We denote this $n \times 1$ vector as $\mathbf{v}_i$. Similarly, $\mathcal{U}$ is the output word matrix. The $k-$th row of $\mathcal{U}$ is an $n-$dimensional embedded vector for word w_k when it is an output of the model. We denote this row of $\mathcal{U}$ as $\mathbf{u}_k$.

For this method sequence of actions can be written as follows:

- We generate our one hot word vectors $(\mathbf{x}^{(c-m)}, ..., \mathbf{x}^{(c-1)}, \mathbf{x}^{(c+1)}, ..., \mathbf{x}^{(c+m)})$ for the input context of size $2m$.

- We get our embedded word vectors for the context $(\mathbf{v}_{c-m} = \mathcal{V}\mathbf{x}^{(c-m)}, \mathbf{v}_{c-m+1} = \mathcal{V}\mathbf{x}_{(c-m+1)}, ..., \mathbf{v}_{c+m} = \mathcal{V}\mathbf{x}_{(c+m)})$

- Average these vectors to get $\tilde{\mathbf{v}} = \frac{\mathbf{v}_{c-m}+\mathbf{v}_{c-m+1}+...+\mathbf{v}_{c+m}}{2m}$

- Generate a score vector $\mathbf{z} = \mathcal{U}\tilde{\mathbf{v}}$

- Turn the scores into probabilities
$$\hat{\mathbf{h}} = \text{softmax}(\mathbf{z}) \tag{2.3}$$

- We desire our probabilities generated, $\hat{\mathbf{h}}$, to match the true probabilities, $\mathbf{y}$, which also happens to be the one hot vector of the actual word.

Described method can be interpreted as a feed forward neural network with one hidden layer that do not use activation function. As a loss function for this algorithm can be chosen cross-entropy loss function

$$L = \sum_{i=1}^{|V|} \mathbf{h_i} \log(\hat{\mathbf{h}}_\mathbf{i}) \tag{2.4}$$

where $\hat{\mathbf{y}}$ is softmax (2.3) function.

2.1.2.2 Fast Text Word Embeddings

As mentioned previously, Fast Text method is a CBOW modification. Main modification is that Fast Text is taking into account not only words but also suffixes of words. The words are splitted into suffixes and as a result they can handle understanding of the context better.

In this theses we used pretrained Fast Text models [15] consisting of 1 million word vectors trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (16B tokens).

Chapter 3

Data and Evaluation Metrics

3.1 Evaluation metrics

When the models are implemented and trained we have to compare them. This part is very important because we want to define such metrics that will not be biased and which will have high discriminability. In this project experiments are separated into two parts. First part is supervised classification with big amount of data. This is done for understating what is the maximal upper bound of specific classifiers. These upper bounds are used as maxima to which our active learning algorithms should be converging.

3.1.1 Receiver Operating Curve metric

In section 1 we mentioned that we are limiting our problem only on binary classification. Plenty of metrics such as recall, accuracy, precision, etc. exists for binary classification. However we decided to find a metic that is able to unify all metrics discussed before and do not underperform each of them. For these purposes we chose Receiver Operating Curve metric. ROC visualizes the tradeoff between true positive rate (TPR)

$$\text{TPR} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

and false positive rate (FPR)

$$\text{FPR} = \frac{\text{false positive}}{\text{false positive} + \text{true negative}},$$

where terminology true/false positive/negative, true or false refers to the assigned classification being correct or incorrect, while positive or negative refers to assignment to the positive or the negative category [7]. This means that for every threshold, we are able to calculate TPR and FPR and plot it on one figure.

We are working with balanced datasets. Thus, there is no problem in using ROC metrics. ROC metric is also very good when we care equally about positive and negative class. Another advantage is that if we notice small changes in ROC it will not result in big changes in other binary classification metrics.

For unique comparison, the metric should be compacted into a single number. This is typically done by integrating the area under the ROC, which is known as the area under the curve (AUC). The probabilistic interpretation of ROC score is that if a positive case and a negative case are chosen randomly, the probability that the positive case outranks the negative case according to the classifier is given by the AUC [7].

3.1.2 Supervised Learning Results Validation

As mentioned above, supervised learning results are used as maximal upper bound of specific classifiers. In order to make results statistically valid we used k-fold cross validation. For each batch from k-fold cross validation we calculated both ROC and AUC. As an output result of a classifier performance we calculated mean value through all ROC and AUC results. All results are calculated with respect to balanced data classes.

3.1.3 Active Learning Results Validation

Active learning model evolution is based on supervised learning algorithms that are sequentially retrained. Thus, we are not able to display ROC for active learning algorithms because the amount of results is too big. We decided to aggregate results and display evolution of AUC metric for each step of the active learning sequence. AUC sequences can be well compared between different classifiers. Another aspect of data validation is making the results statistically significant. We are not able to use k-fold cross validation for active learning algorithms. Therefore we run the active learning algorithm $H \in \mathbb{N}$ times with different random selection of the initial training set. Due to random initializations, we are able to determine uncertainty bounds that are calculated as standard deviations from the mean value.

3.2 Data

This chapter is dedicated to dataset description. We used two datasets for algorithms training and testings. We consider these datasets big and diverse enough for getting unbiased results. We took into account size of texts (articles and tweets), diversity and at the same time similarity of topics.

3.2.1 HuffPost 200k Articles Dataset

HuffPost 200k Articles Dataset is publicly available at Kaggle competition webpage and can be found as News Category Dataset [16] here `https://www.kaggle.com/rmisra/news-category-dataset`. Described dataset contains around 200k news headlines from the year 2012 to 2018 obtained from HuffPost. Following dataset includes url address and label to each article. HuffPost dataset has 200k articles assigned to 41 categories. We used only 10 categories. We are interested in binary classification, as a result we have to make pairs from chosen categories. These categories and pairs are listed in table 3.1.

Tuple Id	Category Pairs	
1	Crime	Good News
2	Sports	Comedy
3	Politics	Business
4	Science	Tech
5	Education	College

Table 3.1: HuffPost Dataset Categories which were chosen for algorithms' training and testing

Due to the fact that there is no raw article included in the dataset, we used url links in order to find the articles. For each category we scraped 500 original articles from `www.huffpost.com`. Listed categories are chosen with respect to diversity and classification complexity. We sorted the categories in table 3.1 with respect to ascending classification complexity order. By classification complexity we mean two sets intersections in feature spaces. If the classification complexity is high, then majority of feature space dimensions have intersections between two datasets. Thus, it is harder to find such set of features that can be used for high classification performance.

3.2.2 1600k Tweets Dataset

Another dataset that is used in this work is 1600k Tweets dataset which is publicly available and is also taken from the Kaggle competition webpage. The dataset can be found as sentiment140 dataset [9] here `https://www.kaggle.com/kazanova/sentiment140`. This dataset contains 1,600,000 tweets extracted using the twitter api. The tweets have been annotated as negative (0), positive (4) and they are used for sentiment detection. Same as with HuffPost dataset we used 500 records for each category for training and testing purposes. The reason of choosing this dataset is that we wanted to show how our algorithms can handle data that consist of little texts.

Chapter 4

Project Implementation and Architecture

Even though this work is quite theoretical with experiments that proof theoretical concepts, we consider implementational part interesting as well. In this chapter we show the architecture of the project and explain how different dependencies cooperate with each other. The codebase of this project can be easily found at `https://github.com/sahanmar/Peony`. We expect this project going to continue grow and will be used not only in terms of master theses.

This project was written in Python 3.7 programming language with the usage of Conda environment. The project combines a lot of different tools and programs such as Docker, Docker-Compose, MongoDb, Jupyter, etc..

In this theses we used two main components that represent the database and computational part. We tried to unify all methods as much as possible and make the utilization process very easy.

4.1 Database

In order to make everything consistent and let the models work with the same input and output format we decided to create a database that will store all the data in the JSON format. This unification let us connect the database to machine learning and visualizing components. In this project we decided to work with NoSQL database. Our choice was MongoDb. The reason why we have chosen MongoDb is because of its simplicity and possibility of maintaining through Docker. Since Docker and MongoDb is a perfect combination, the database can be deployed with two lines of code through Docker-Compose as explained in documentation on GitHub. Of course it is easier to use MongoDb without Docker but our motivation was measured on simplicity of creating and working with the database. All experiments were run on Google Cloud Platform Virtual Machine instance. Thus, we could start working with the models right away without any complications with installation.

4.1.1 MongoDb Data Format

MongoDb represents the data in BSON format behind the scenes but we are send and get there JSON format data. Despite the fact that we are having different text datasets that we store in the database, we decided to create a unified JSON scheme that will let us preserve the structure of the data stored in MongoDb. JSON schema of how the data are stored and what a user will get as an output from a database is shown in figure 4.1. Deeper explanation of JSON schema can be found at `https://json-schema.org/understanding-json-schema/`.

```
{
  "title": "Database",
  "type": "object",
  "properties": {
    "datasetName": {
      "type": "string",
      "description": "Name of the dataset"
    },
    "datasetId": {
      "type": "int",
      "description": "Unique hash id that will be created automatically"
    },
    "record": {
      "type": "object",
      "description": "All information about an instance",
      "properties": {
        "id": {
          "type": "string",
          "description": "Unique hash id that will be created automatically
              "
        },
        "snippet": {
              "type": "string",
              "description": "Snippet of a text. Can be empty"
        },
        "text": {
          "type": "object",
          "description": "Text instance that is used for a model",
          "properties" : {
            "title": {
              "type": "string",
              "description": "Title of a text. Can be empty"
            },
            "body": {
              "type": "string",
              "description": "Body of a text"
            },
          },
        },
        "label": {
          "type": "string",
          "description": "Label for an instance. Can be empty if this is
              not validation data"
        },
        "metadata": {
          "type": "object",
          "description": "Any additional metadata. Can be empty field"
        },
      },
    },
  },
}
```

Figure 4.1: MongoDb JSON schema visualization

4.2 Computations

All computations were done with the usage of virtual instance on Google Cloud Platform. We used configuration with 2 CPU, 7.5Gb memory that was running on Debian GNU/Linux 10. All versions of python, python packages, docker, mongo, etc. can be found in .yml files in GitHub folder with the project.

4.3 Machine Learning Component

Machine Learning (ML) Component is fully implemented in Python with the usage of open source libraries. In order to understand how to use the models, it is possible to find the code and its usage in Jupyter notebook that are stored in showcase folder. Showcase folder has four Jupyter notebooks that show both how to get the data for the models from the database and how to start using the models.

4.3.1 Data Transformers

Before models training and testing user has to fit the data transformer that transforms text into tensors form. Tensors are used as input values for models. As mentioned in chapter 2, we are working only with TF-IDF and Fast Text text encodings. Both TF-IDF and Fast Text models are created from the documents that are given to the transformer.

4.3.1.1 TF-IDF Transformer

TF-IDF transformer represents one article as a vector. From the articles are extracted all words that exist in the vocabulary and then for a document is calculated TF-IDF encoding. As a result, if we make TF-IDF encoding of a set of articles, we will get a matrix where each row represents specific document and each column represents word from a dictionary.

4.3.1.2 Fast Text Transformer

Fast Text model is a pretrained model that consists of one million words mapped to vectors. These words are stored in MongoDb. When a user starts to use Fast Text model, ML component creates a words' vocabulary from the texts taken for model training/testing. This vocabulary is created in a form of a hash map (word -> vector) where word embeddings are downloaded from MongoDb. It is important to remember that Fast Text encoding represents each word as a vector with predefined number of components. We are using word embeddings that represent each word with 300 float values. We introduce article encoding as a mean value through all words from a text that is given for encoding. Thus, if we make Fast Text encoding of a set of articles, we will get a matrix where each row represents specific document. Huge advantage of this method in comparison to TF-IDF, is that we are working only with 300 float values (300 columns if we provide encoding of set of articles) than with huge dictionary. Therefore we get lower features dimensionality and better context understanding.

4.3.2 Machine Learning

In this work we created a Generalized Model that unifies all models. Generalized Model allows to work with each Machine Learning algorithm in the same way. Generalized model is able to take a data transformer as an input argument. This feature makes it easier to work with models. In first chapter of this theses we introduced our models the way that we want to sample from their parameters' distributions.

That means that we are aiming to work with ensembles. In figure 4.2 is shown a generalized diagram of machine learning structure.

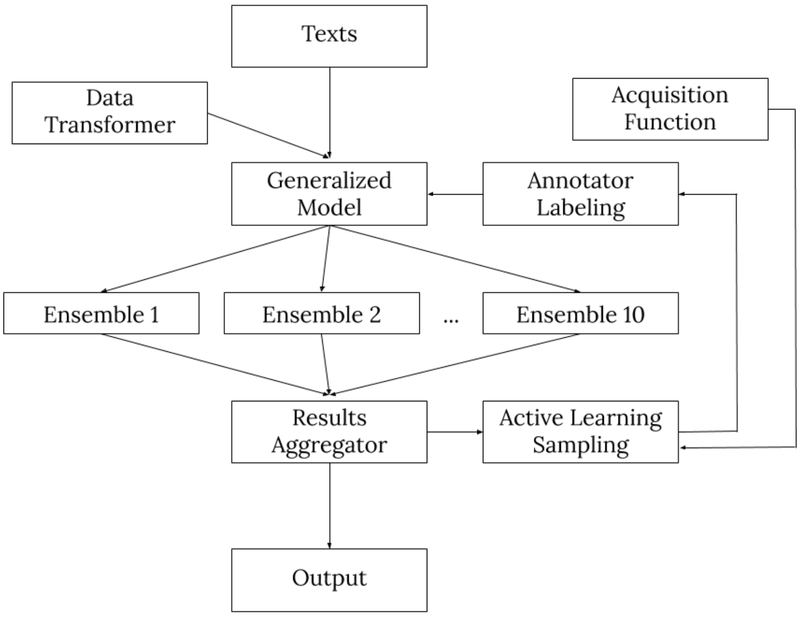

Figure 4.2: Machine Learning Workflow

We used scikit-learn [18] for basic algorithms such as Random Forests and SVMs. However, core of this project is constructed around Neural Networks. We used PyTorch [17] as a Neural Networks framework. In this work we have implemented and tested five classification algorithms. Three of them, such as SVM, Random Forest and Feed Forward Neural Network ensembles are trained on a randomly chosen subsets from the training data. For each ensemble are randomly chosen 80% from training data that are used for training. Two algorithms such as SGLD and DENFI are using full training dataset for their ensembles. The variability in SGLD and DENFI ensembles is reached though adding gaussian noise while ensembles training. We hardcoded amount of ensembles for all models to 10.

4.3.2.1 SVM and Random Forest Ensemble Setup

Both SVM and Random Forest models were taken and used out of the box. We created 10 SVM and 10 Random Forest ensembles with default scikit-learn setting. No modifications were provided.

4.3.2.2 Feed Forward Neural Network

Despite the fact that we used very simple Neural Networks architecture, it showed very good results. We implemented Feed Forward Neural Network with the usage of PyTorch python package with one hidden layer that consists of 100 neurons with sigmoid activation function. We chose softmax activation function for output layer.

Algoritmus 4.1 DENFI modification algorithm pseudocode

```
def active_learning_iteration_training(all_ensembles, training_data):
  for ensemble in all_ensembles:
    if first_active_learning_iteration is False:
      ensemble.weights = ensemble.weights_prev_iteration
      training_epochs = 500
    else:
      ensemble.weights = gaussian_initialization(mean=0, var=0.1)
      training_epochs = 2000
    ensemble.train(training_data)
        ensemble.weights = ensemble_weights
                        + gaussian_noise(mean=0, var=0.1)
```

4.3.2.3 SGLD

For SGLD we used the same configuration as for Feed Forward Neural Network. One significant difference is that we used whole training dataset and were adding gaussian noise to a gradient descent [27] as shown in (1.37). Precise configuration of SGLD can be found in github project here `https://github.com/sahanmar/Peony/blob/master/Peony_project/Peony_box/src/peony_adjusted_models/sgld_nn.py`.

4.3.2.4 DENFI

In this work we have simplified original DENFI algorithm. The pseudocode of this algorithm is show in algorithm 4.1. The main idea is that the algorithm finds different local minimums due to the random weights initialization in first active learning iteration. When the training is finished, gaussian noise is added to the output weights in order to increase the variability. In further active learning iterations, weights from previous iterations with extended training dataset is used. After the training we also add gaussian noise to the weights. When the learning iterations are done, we can use the algorithm for predicting.

Chapter 5

Passive Learning Classification

Before we can start active learning part we have to understand if implemented algorithms are capable to solve the classification task. Thus, we decided to test the models on Sports and Comedy categories from HuffPost Dataset and on Tweets from Tweets Dataset. The classification was done both for TF-IDF and Fast Text Encodings. We separated experiments with respect to the dataset types.

5.1 Passive Learning HuffPost Dataset

In this section we are illustrating ROC and AUC metrics with respect to $10-$fold cross validation and 500 Sports, 500 Comedy articles. Vocabulary, that is created from 1000 articles corpus consists of 20 thousand unique words. We would like to mention that all algorithms are trained and tested with respect to 10 ensembles. Moreover, the ratio of randomly chosen training data for ensembles (SMV, Random Forest, Feed Forward NN ensembles) is set to 80%.

5.1.1 SVM Ensembles

Results for SVM Ensembles shown in figures 5.1 and 5.2. These are results both for TF-IDF and Fast Text encodings. As seen on these plots ROC and AUC values of $10-$fold cross validation are extremely high. This means that our model works very good. Another interesting point is that standard deviation with respect to all runs is very low. It means that SVM ensembles could linearly separate Sports and Comedy sets with acceptable classification error.

It is also seen that ROC and AUC metrics are almost same for TF-IDF and Fast Text encodings. However, there is one significant difference. The difference is in computational time because Fast Text encoded text document consists of 300 float components and TF-IDF encoded text document consists of 20 thousand float components. Even though we are using algorithms for sparse matrix computation for TF-IDF method, computations for Fast Text encoding are five times faster. The higher amount of unique words is, the higher elapsed time per article for TF-IDF based encoding algorithm will be. Another good aspect is that because of algorithm's simplicity, SVM is trained much faster than Neural Network based models.

It is possible to conclude that the model shows good results for this classification problem and can be used for active learning experiments.

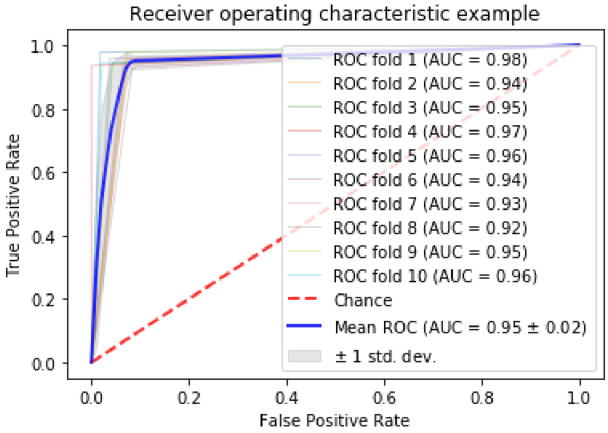

Figure 5.1: TF-IDF ROC and AUC for 10 SVM Ensembles trained and tested on Sports and Comedy data where each ensemble is trained on 80% of randomly chosen training data

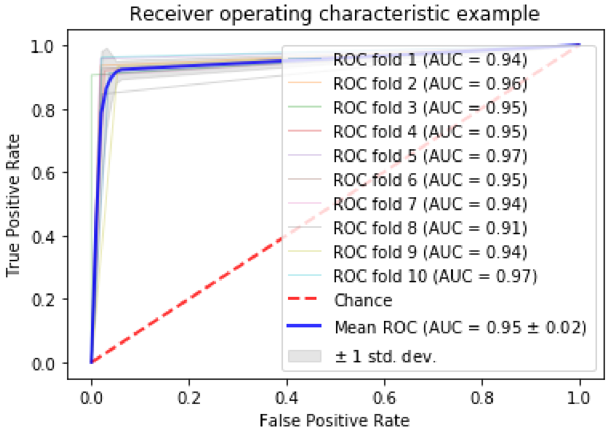

Figure 5.2: Fast Text ROC and AUC for 10 SVM Ensembles trained and tested on Sports and Comedy data where each ensemble is trained on 80% of randomly chosen training data

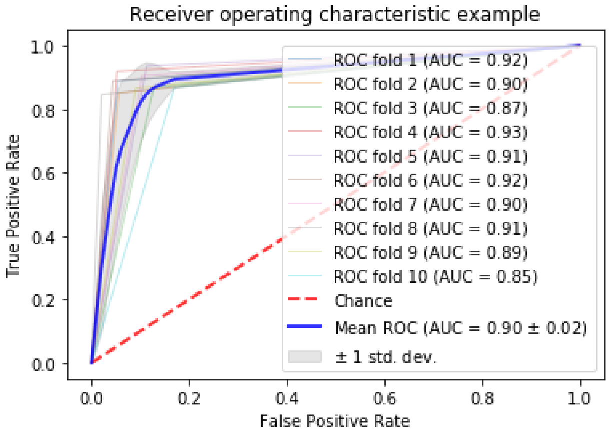

Figure 5.3: TF-IDF ROC and AUC for 10 Random Forest Ensembles trained and tested on Sports and Comedy data where each ensemble is trained on 80% of randomly chosen training data

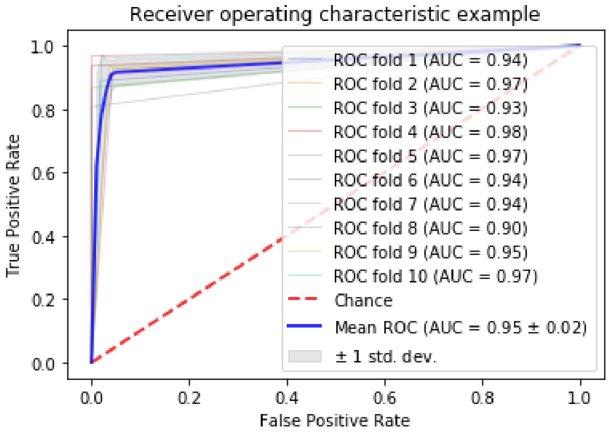

Figure 5.4: Fast Text ROC and AUC for 10 Random Forest Ensembles Ensembles trained and tested on Sports and Comedy data where each ensemble is trained on 80% of randomly chosen training data

5.1.2 Random Forest Ensembles

Results for Random Forest Ensembles shown in figures 5.3 and 5.4. These are results both for TF-IDF and Fast Text encodings. Same as in SVM section, ROC and AUC values of 10−fold cross validation for Random Forest are high as well. This means that our model works very good. Standard deviation with respect to all runs is also very low.

If we compare results for TF-IDF and Fast Text encodings, it is seen that Fast Text based model outperforms results with respect to TF-IDF encoding model. Mean AUC value for Fast Text article encoding model is higher by 5%. It is interesting that if we apply sparse matrix algorithms for TF-IDF in Random Forest ensembles model, then computational speed for Fast Text and TF-IDF text will approximately same. This observation was made on the basis of processing 1000 articles from Comedy and Sports sets. As mentioned in SVM section we can also say that Random Forest algorithm is trained much faster than Neural Network based models.

It is possible to conclude that the model shows good results for this classification problem and can be used for active learning experiments.

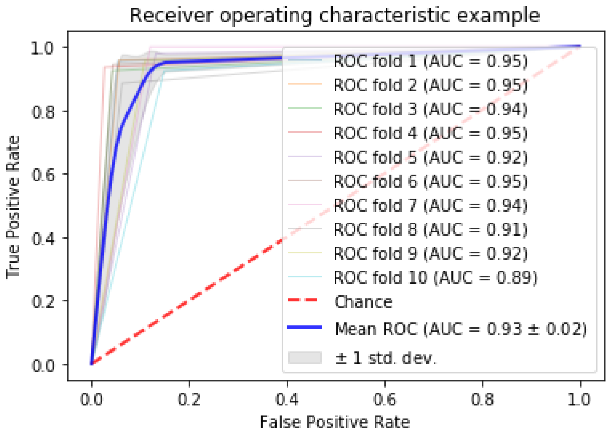

Figure 5.5: TF-IDF ROC and AUC for 10 Neural Networks Ensembles trained and tested on Sports and Comedy data where each ensemble is trained on 80% of randomly chosen training data

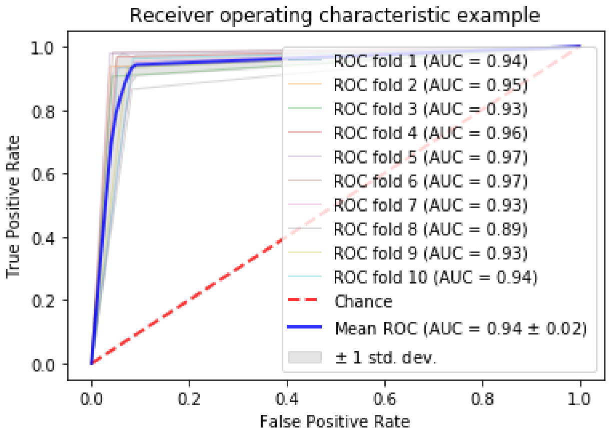

Figure 5.6: Fast Text ROC and AUC for 10 Neural Networks Ensembles trained and tested on Sports and Comedy data where each ensemble is trained on 80% of randomly chosen training data

5.1.3 Feed Forward Neural Network Ensembles

In this work we have implemented thee algorithms based on neural networks such as Neural Networks Ensemble, SGLD and DENFI algorithms. The difference between these methods is in parameters sampling. Neural Networks Ensemble takes a randomly selected subset from training data for each ensemble. Thus, we decided do not show SGLD and DENFI results in this section because they use whole set of training data. As a result if Neural Networks Ensemble shows good results, we assume that SGLD and DENFI will also perform good passive learning results.

Results for Neural Networks Ensembles shown in figures 5.5 and 5.6. These are results both for TF-IDF and Fast Text encodings. As seen in previous section, ROC and AUC values of 10−fold cross validation for Neural Networks are also high.

Comparing the results between TF-IDF and Fast Text encoding based models we can observe that Fast Text based model gives slightly better results. We have already discussed fastness of algorithm training with respect to different embedding models in Random Forest section. In case of Neural Networks, this difference is even more significant. Model that is based on Fast Text word embeddings takes

20 times less time than TF-IDF encoding based model.

5.2 Conclusion

All of the tested models gave really good results in classification of Sports and Comedy categories. It means that we are able to continue working with these algorithms in active learning section. We would like to highlight that Fast Text encoding based algorithms showed a bit better results than TF-IDF. Fast Text based algorithms also showed significant improvement in evaluation speed of algorithm training.

Chapter 6

Active Learning Classification

In Passive Learning section we showed that all implemented algorithms are good for solving passive text classification task. Active learning section results is the main part of this theses. Therefore, we tested all five algorithms on the data mentioned in the Data section. The results are shown and described in further subsections. However, before starting with text classification results, we would like to show how our models represent uncertainty. As written in theoretical introduction to active learning, we use models' uncertainty for active learning loop.

6.1 Active Learning Models' Uncertainty

Due to the fact that text encoding features space dimensionality is extremely high, we introduce a 2-dimensional toy problem for uncertainty visualization. In figure 6.1 is shown a dataset that is used for toy problem classification. We generated this dataset with adding gaussian noise in order to make the task similar to real world problems.

We generated 1000 data points where 500 are assigned to Class 0 and another 500 points are assigned to Class 1. We used 50% randomly chosen data samples as a training dataset. Next step, was creating a two dimensional grid that will be used for model predictions. We assign data sample to Class 1 if prediction value is higher than 0.5. If prediction value is lower than 0.5 than the value is assigned to Class 0. We consider that model is uncertain about specific data sample if its prediction values is close

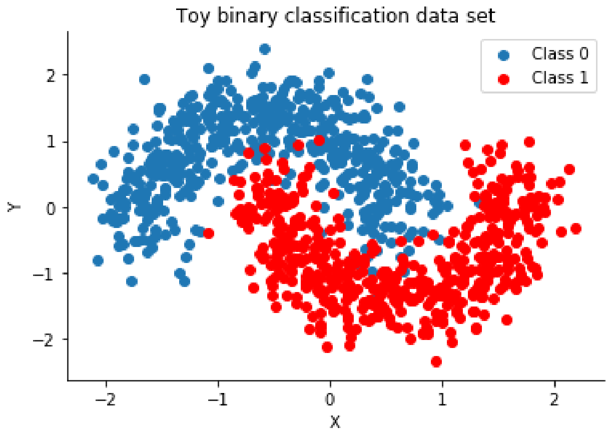

Figure 6.1: Toy problem dataset visualization

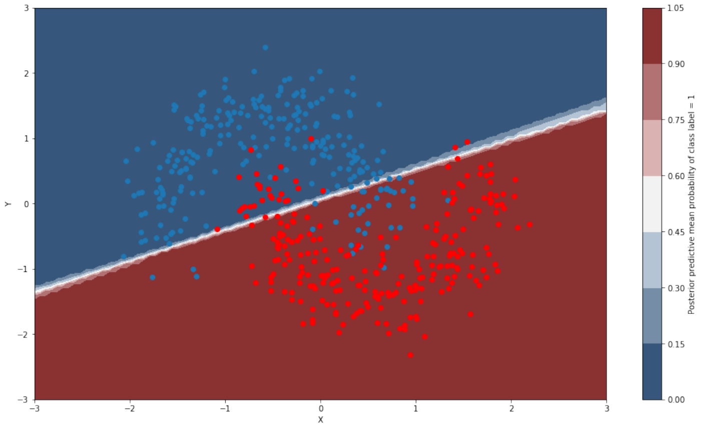

Figure 6.2: SVM Ensembles posterior predictive mean probability of Class 1

to 0.5. As a result, sampling values from maximal uncertainty region will bring maximum information about the dataset. Models set up for Toy Problem is the same as it is defined in the Machine Learning Component section 4.3 .

6.1.1 SVM Ensembles

Uncertainty for the SVM Ensembles model is visualized in figure 6.2. Uncertainty bounds are linear and quite narrow. Linearity of uncertainty bounds is explained with the fact that we are using SVMs with linear kernel. Narrowness can be explained with richness of the training dataset and linear limitations of SVM decision bound.

6.1.2 Random Forest Ensembles

In figure 6.3 is visualized uncertainty for Random Forest Ensembles model. In the case of Random Forest Ensembles we can see that uncertainty bounds are not linear and lay near the region where two classes are splitted. We see that uncertainty region is becoming wider near the places where datapoint of two different classes lay near each other. This is a behavior that we wanted to observe.

It is also seen that the curve is not smooth enough. This is caused by Random Forest decision boundary .

6.1.3 Neural Network Ensembles

In figure 6.4 is visualized uncertainty for Neural Network Ensembles model. In comparison to non-neural networks based methods that were shown above, uncertainty bounds are quite smooth. We can also observe that uncertainty bounds become wider when they go to further from the data points. This behavior can be explained with the fact that in these places algorithm did not get any training data samples. We could also see this behavior in SVM Ensembles but in this case, Neural Network models represent the uncertainties much better.

6.1.4 SGLD

In figure 6.5 is visualized uncertainty for SGLD model. We see that SGLD algorithm has similar behavior to Neural Network Ensembles. The difference is that all uncertainty bound curves has similar

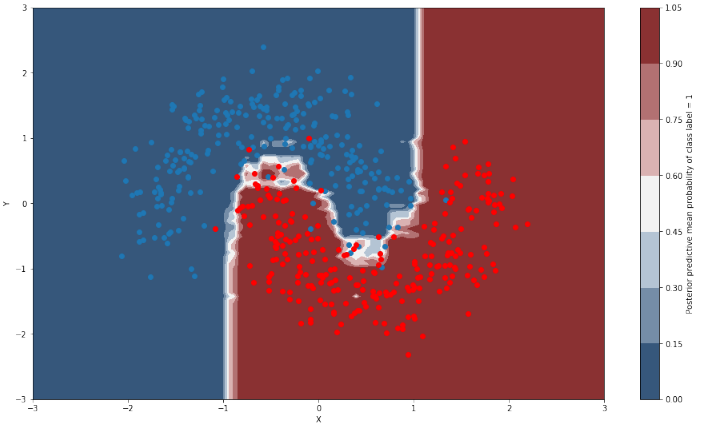

Figure 6.3: Random Forest Ensembles posterior predictive mean probability of Class 1

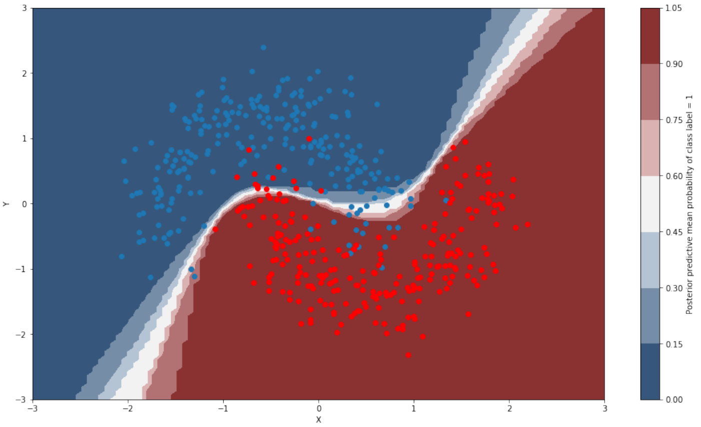

Figure 6.4: Neural Network Ensembles posterior predictive mean probability of Class 1

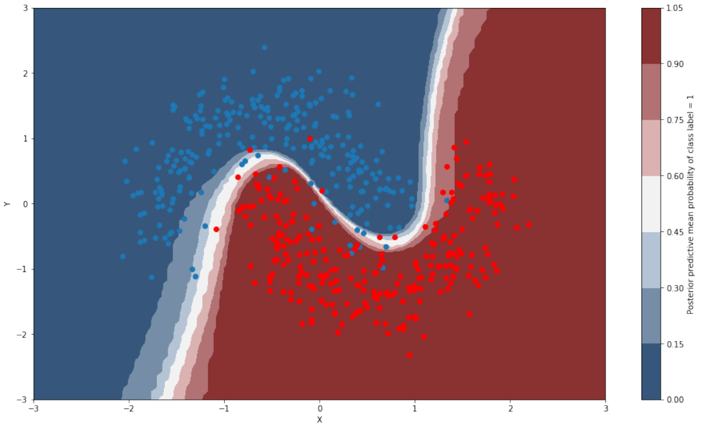

Figure 6.5: SGLD posterior predictive mean probability of Class 1

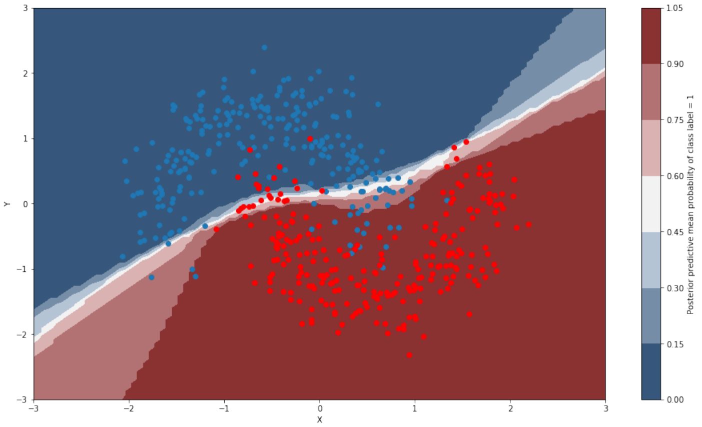

Figure 6.6: DENFI posterior predictive mean probability of Class 1

curvature. This is happening due to the fact that SGLD finds loss function minimum and than samples parameters' values in a neighborhood of the minimum. As a result we expect decision bound for each parameters sample be similar to each other.

6.1.5 DENFI

In figure 6.6 is visualized uncertainty for the DENFI model. We see that uncertainty bounds are very similar to Neural Network Ensembles algorithms but still a bit different. As told in pseudocode 4.1, algorithm founds different local minimums and then we add some gaussian noise to parameters values. In next learning iterations the algorithm continues training using the weights from the previous step.

The last 100 loss function values with respect to each DENFI ensemble are shown in figure 6.7. It is seen that each ensemble found its own local minimum. Additional gaussian noise adds more variability to the algorithm that makes samples more diverse.

6.1.6 Conclusion

To conclude, we can say that the best variability representation was seen in Neural Network models. Of-course, we could test SVM algorithm with the usage of different kernels but our prior interest was

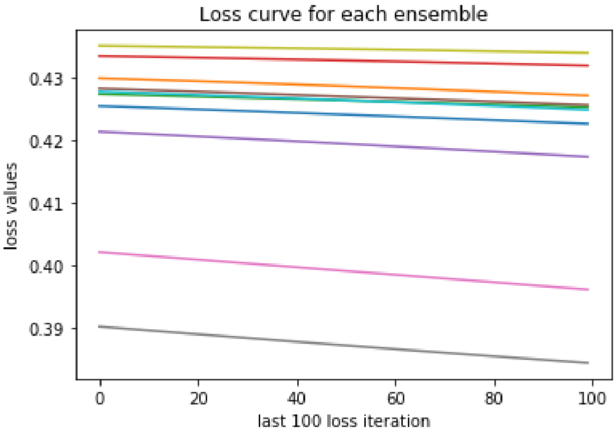

Figure 6.7: The last 100 loss function values with respect to each DENFI ensemble

around Neural Networks. We have showed that all models are able to represent variability and can be used for further tests.

6.2 Active Learning Simulation Set Up

6.2.1 Simulation Loop

The active learning simulation is done for comparing two strategies such as random and ”smart” selection of text documents. We randomly choose initial training set that has 10 samples. Described 10 random samples are chosen from 1000 text documents (500 text documents per category). We define 1000 documents dataset as validation set.

Next we start two runs. One run is based on random selection of text documents and second run is based on acquisition function selection of the documents. Both runs start with the same training dataset and then they chose additional training documents based on their strategies. We consider continuous new data selection from validation set and imitating annotators labeling process. All in all we imitate text samples selection 200 times. We select 10 random samples in the beginning, train our model, make a prediction on the complement to the training dataset (990 text documents from validation set). As a further step, we select 1 new sample from the set on which the prediction was done. New text sample for labeling is chosen with respect to acquisition function or random choice. Before we extend our training dataset with a new labeled sample we calculate the AUC metrics on the complement to the dataset (990 text documents). This process is repeated 200 times. Thus, by the end of the simulation, our training set will have 210 text documents and testing set (complement to a training set) will have 790 data samples. In order to make our results statistically valid, we repeat the described simulation loop 10 times.

6.2.2 Epsilon Greedy Strategy

In this work we decided that we do not want to sample with respect to an acquisition function from the beginning but follow with epsilon greedy strategy [26]. We decided that both active learning and random strategy are going to start with random sampling. The algorithms have a coefficient $\epsilon \in [0, 1)$ that represents probability of data sampling with respect to acquisition function. In this work we define

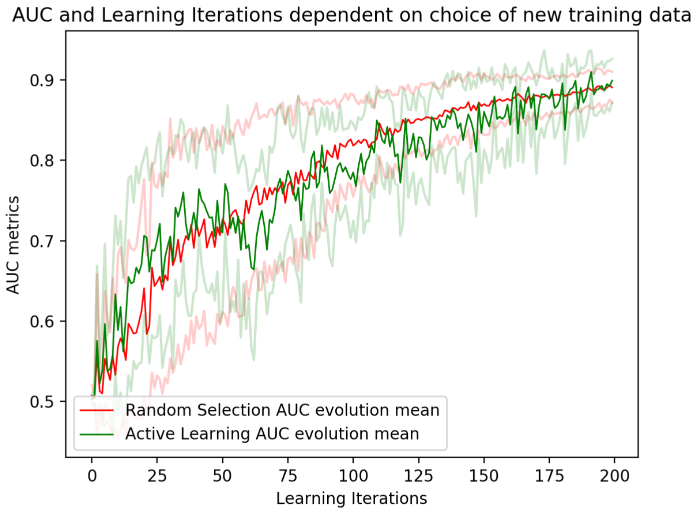

Figure 6.8: Comparison of random and active learning text data selection mean and one standard deviation for SVM ensembles with respect to 10 samples as initial training configuration, 200 data selection iterations and TF-IDF encoding

ϵ as

$$\epsilon = \begin{cases} \frac{\exp(u-40)}{\exp(u-40)+1}, & u \in \{1, ..., U\} \\ 0, & u = 0, \end{cases} \tag{6.1}$$

where $u = \{0, 1, ..., U\}$ is set of question that results in number of text documents which will be labeled by an annotator. As mentioned in previous section, the number iterations (question U) is set to 200. From equation 6.1 is seen that when we reach $40-$th document, the probability of random sampling is 50%. This strategy showed really good results which will be shown in further sections.

6.3 Active Learning on Texts with TF-IDF Encoding Based Models

We decided to show the results for TF-IDF encoding based model because described text encoding provides high discriminability and relatively simple at the same time. However, as mentioned in passive learning section due to the high amount of features, models that are using this type of encoding are harder to train. We were not able to train Neural Network Ensembles due to the high dimensional feature space. Thus, we decided to show the results only with respect to SVM and Random Forest with respect to Sports and Comedy categories. Therefore, there is no need to split further section to different datasets because all the experiments based on TF-IDF encoding are assumed to be done only for Sports and Comedy categories.

6.3.1 SVM Ensembles

In figure 6.8 is shown active learning simulation results for SVM Ensembles and TF-IDF encoding. Active learning strategy is based on entropy acquisition function.

As seen from figure 6.8, the results for entropy are not better than for random selection. This can be explained due to low variability of SVM ensembles. For this case we can conclude that there are no

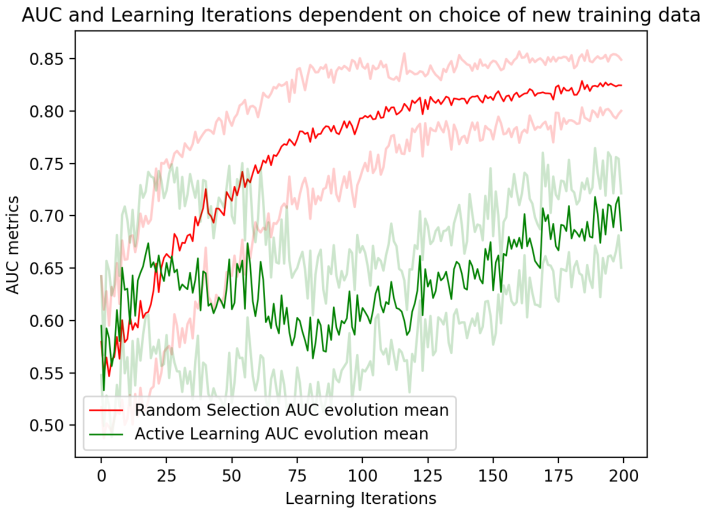

Figure 6.9: Comparison of random and active learning text data selection mean and one standard deviation for Random Forest ensembles with respect to 10 samples as initial training configuration, 200 data selection iterations and TF-IDF encoding

significant difference between entropy and random data selection strategy. First and second strategy has almost same uncertainty bounds and converge to the same AUC results.

6.3.2 Random Forest Ensembles

In figure 6.9 is shown active learning simulation results for Random Forest Ensembles and TF-IDF encoding. Active learning strategy is based on entropy acquisition function.

In comparison to SVM Ensembles, active learning strategy for Random Forest Ensembles show really poor results. Random data selection overcomes entropy selection and it can be said that active learning strategy is not working at all for this algorithm. The reason of such behavior can be explained due to very little amount of training data in the beginning. This fact may make algorithm to start selecting the data which have high entropy but are close to each other. As a result, it will not lead to high performance results.

6.3.3 Conclusion

Even though we were able to see extremely good results for TF-IDF encoding in Passive Learning section, we saw that due to the features space high dimensionality we are not able to train Neural Networks models and test active learning there. Moreover, SVM and Random Forest Ensembles' results were unsatisfying and in comparison to random sampling, active learning strategy did not show good results.

6.4 Active Learning on Texts with Fast Text Encoding Based Models

In comparison to section 6.3, in this section we show results to all models because Fast Text encoding maps texts to 300 dimensional feature space. This is much lower than in the case of TF-IDF encoding. Thus, we were able to provide computations with respect to all models.

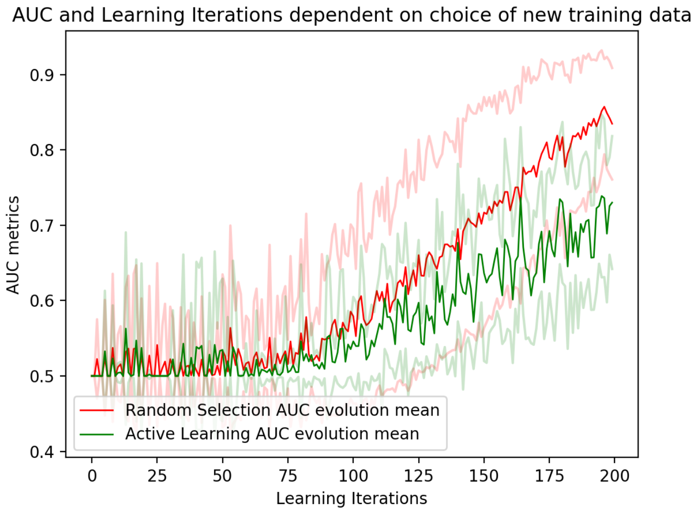

Figure 6.10: Comparison of random and active learning text data selection mean and one standard deviation for SVM ensembles with respect to 10 samples as initial training configuration, 200 data selection iterations and Fast Text encoding

6.4.1 SVM Ensembles

6.4.1.1 Sports and Comedy Categories

In figure 6.10 is shown active learning simulation results for SVM Ensembles and Fast Text encoding. Active learning strategy is based on entropy acquisition function.

The behavior which we observe in figure 6.10 is similar to figure 6.8. We can see that active learning strategy is not working for the case of Fast Text encoding as well.

6.4.1.2 Conclusion

We can conclude that even despite good performance in Passive Learning section the active learning algorithm is not working as we expected. Thus, we will not continue testing this algorithm on further datasets.

6.4.2 Random Forest Ensembles

6.4.2.1 Sports and Comedy Categories

In figure 6.11 is shown active learning simulation results for Random Forest Ensembles and Fast Text encoding. Active learning strategy is based on entropy acquisition function.

As illustrated in figure 6.10 we can see the strategies results are similar to each other but and in active learning case, entropy based sampling is overcoming random sampling strategy. We can also see that in active learning case, uncertainty bounds are much more narrow.

6.4.2.2 Conclusion

Even though we could observe that active learning strategy shows better results than random sampling strategy, we do not continue with further experiments because the results are not significant enough.

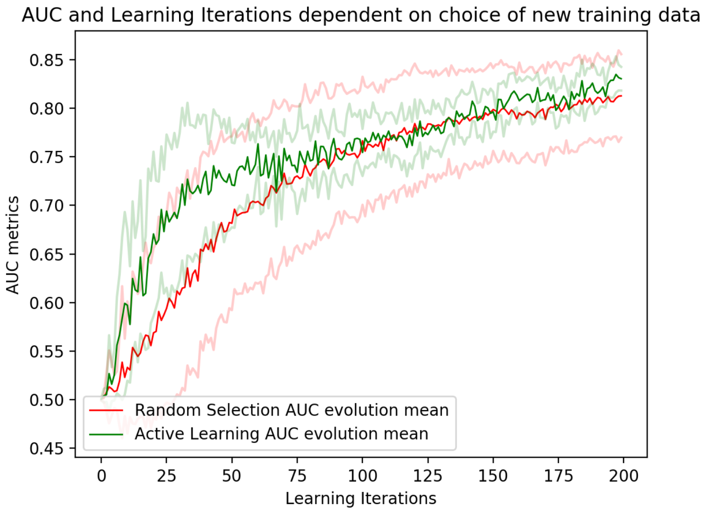

Figure 6.11: Comparison of random and active learning text data selection mean and one standard deviation for Random Forest ensembles with respect to 10 samples as initial training configuration, 200 data selection iterations and Fast Text encoding

6.4.3 Neural Network Ensembles

6.4.3.1 Sports and Comedy Categories

In figure 6.12 is shown active learning simulation results for Neural Networks Ensembles and Fast Text encoding. Active learning strategy is based on entropy acquisition function.

If we compare output metrics from figure 6.12 and figures 6.10, 6.10, we can see significant difference in the algorithms performance. In comparison to previous methods, algorithm based ob neural networks shows same results by $50-$th iteration. That means that the probability of using non-random acquisition function equals to 50%. We are able to observe significant overcome of entropy based sampling than random sampling strategy. We can also see, that random sampling strategy is converging very slow to the results of the active learning strategy. Moreover, active learning uncertainty bounds are much more narrow than the uncertainty bounds of the random sampling algorithm.

6.4.3.2 Conclusion

One significant disadvantage of this method is that it is very slow and hard to train. Each time, we have to train 10 neural networks with cold start. That means training neural networks with random weights initialization and no prior information from the previous training. Despite the fact that the result are really good, we decided not to continue with the algorithm testing because it is very slow and computationally costly. In next sections we introduce results based on neural networks but different way of uncertainty representation.

6.4.4 SGLD

Stochastic Gradient Langevin Dynamics algorithm [27] is one of two algorithms that are assumed as the best algorithms (in terms of this project) with good ratio of training time consumption and performance. As mentioned previously, gaussian noise is added to a gradient descent while training. Thus

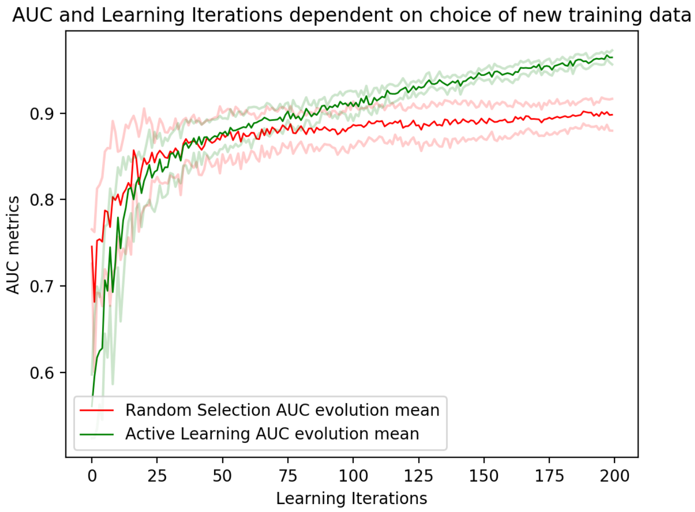

Figure 6.12: Comparison of random and active learning text data selection mean and one standard deviation for Feed Forward Neural Network ensembles with respect to 10 samples as initial training configuration, 200 data selection iterations and Fast Text encoding

approach helps us to sample different parameters vectors around the loss minimum by simply continuing training. In comparison to Neural Networks Ensembles, we do not have to train neural network ensembles separately, but train only one neural network with some additional training epochs.

6.4.4.1 Sports and Comedy Categories

In figure 6.13 is shown active learning simulation results for SGLD, Fast Text encoding and Sport vs Comedy categories. Active learning strategy is based on entropy acquisition function.

We can see that in figure 6.13 both mean and uncertainty bound curves are not smooth enough in comparison to Neural Network Ensembles. This behavior can be explained with low number of samples from SGLD. However, we are able to observe same analogy as with Neural Network Ensembles. The active learning strategy overcomes random sampling and has more narrow uncertainty bounds. Moreover, SGLD was trained almost three times faster than Neural Network Ensembles algorithm.

6.4.4.2 Crime and Good News Categories

In figure 6.14 is shown active learning simulation results for SGLD, Fast Text encoding and Crime vs Good News categories. Active learning strategy is based on entropy acquisition function.

We expect Crime and Good News categories have small intersection. As seen from 6.14 it is really true because first iteration of the simulation starts at around $85\% - 90\%$ AUC. This means that basing on only 10 training data samples algorithm could reach high performance results. The evolution of two strategies is quite same as in figure 6.13. We would like to pay attention to the place where the upper uncertainty bound reaches the value that is greater than one. It not possible for AUC to be greater than one because AUC $\in [0, 1]$. However, as mentioned previously, we construct uncertainty bounds as a standard deviation from mean value. In this case it is possible that upper uncertainty bound will be greater than one. Thus, while interpreting the results, reader has to remember that AUC metric cannot be greater than one.

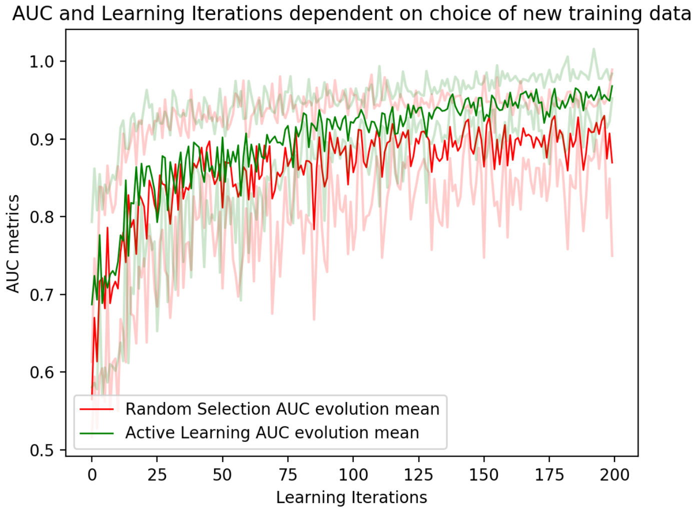

Figure 6.13: Comparison of random and active learning text data selection mean and one standard deviation for SGLD algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and Sports vs Comedy categories

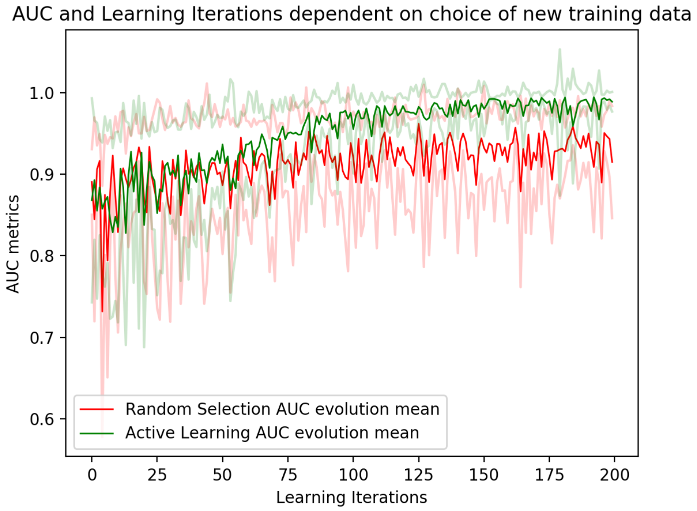

Figure 6.14: Comparison of random and active learning text data selection mean and one standard deviation for SGLD algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and Crime vs Good News categories

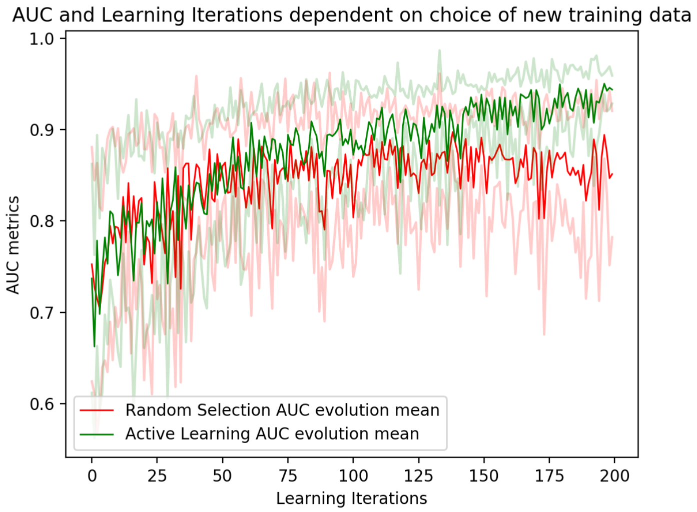

Figure 6.15: Comparison of random and active learning text data selection mean and one standard deviation for SGLD algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and Politics vs Business categories

6.4.4.3 Politics and Business Categories

In figure 6.15 is shown active learning simulation results for SGLD, Fast Text encoding and Politics vs Business categories. Active learning strategy is based on entropy acquisition function.

These two categories have bigger intersection what means that it is harder classification task. The results show again that active learning strategy easily overcomes random sampling strategy. Even though, the topics are harder to distinguish, active learning strategy finds the patterns in the data and show much better results with more narrow uncertainty bounds.

6.4.4.4 Tech and Science Categories

In figure 6.16 is shown active learning simulation results for SGLD, Fast Text encoding and Tech vs Science categories. Active learning strategy is based on entropy acquisition function.

We observe again that active learning strategy is better even though the categories are quite similar. As seen from the plot uncertainty upper bound reaches the values which is greater than one. This is exactly the case which was already covered in 6.4.4.2.

6.4.4.5 College and Education Categories

In figure 6.17 is shown active learning simulation results for SGLD, Fast Text encoding and College vs Education categories. Active learning strategy is based on entropy acquisition function.

The collected results based on College and Education category do not differ to the results from previous parts of SGLD algorithm. We can still see that active learning strategy outperforms random sampling even despite the fact that these categories are very similar.

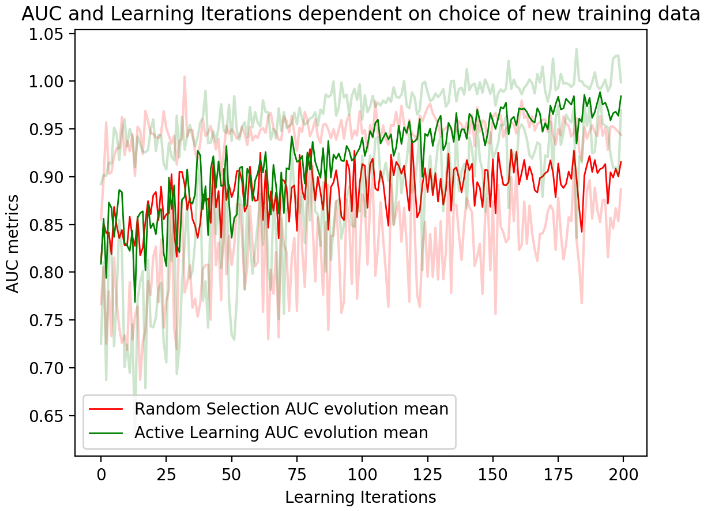

Figure 6.16: Comparison of random and active learning text data selection mean and one standard deviation for SGLD algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and Tech vs Science categories

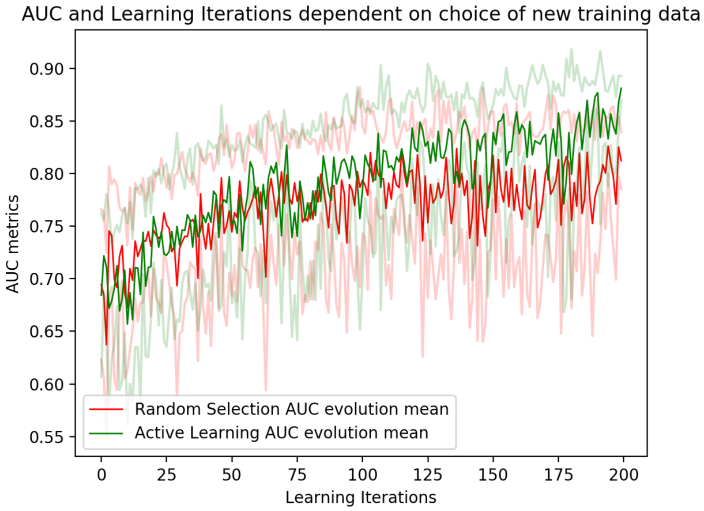

Figure 6.17: Comparison of random and active learning text data selection mean and one standard deviation for SGLD algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and College vs Education categories

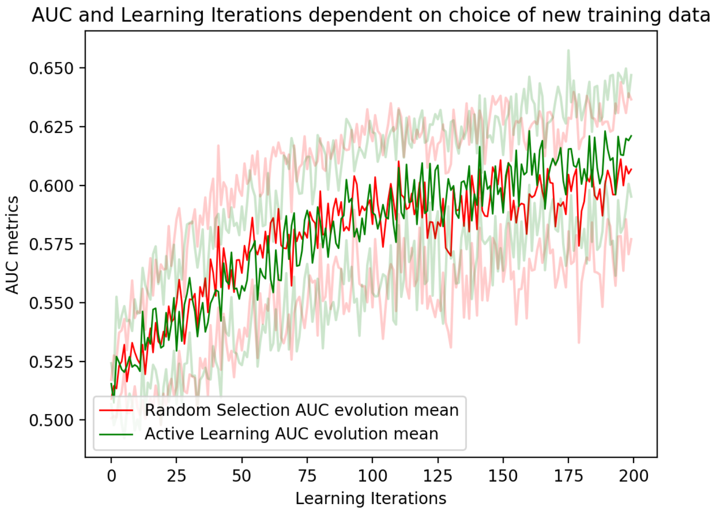

Figure 6.18: Comparison of random and active learning text data selection mean and one standard deviation for SGLD algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and Positive vs Negative tweets categories

6.4.4.6 Positive and Negative Tweets Categories

In figure 6.18 is shown active learning simulation results for SGLD, Fast Text encoding and Positive vs Negative tweets categories. Active learning strategy is based on entropy acquisition function.

In previous sections we tested SGLD algorithm on the data from HuffPost dataset. In figure 6.18 are illustrated results with respect to tweets dataset. Due to the fact that tweets are really short, it is quite hard to find the patterns that can be used for high performance categories separation. Thus, we can see that our classification results (AUC) is not that high in comparison to previous results. Moreover, we observed that for this dataset there is not difference between active learning and random sampling strategy.

6.4.4.7 Conclusion

To sum up, we can say that SGLD algorithm showed really impressive results and proved that the active learning strategy outperforms random sampling. However, there are several disadvantages. First disadvantage is that the resulted curves were not smooth. We think that this problem can be eliminated by sampling more parameters vectors while training the model. Another problem is that the algorithm do not show better active learning results when it is tested on the short and quite general text data. We believe that this problem can be solved with different encoding approach.

6.4.5 DENFI

DENFI algorithm is second algorithm which is not partitioning training dataset. In addition to this we use a modification of DENFI algorithm. Huge advantage of this algorithm is that it uses prior information from previous training round. In the first round of training, DENFI algorithms follows almost same strategy as Neural Network Ensembles method. It trains 10 ensembles with respect to all training data. The variability in ensembles is reached with different weights initialization. After the training is finished, we add some gaussian noise in order to increase the variability. When we add some training samples,

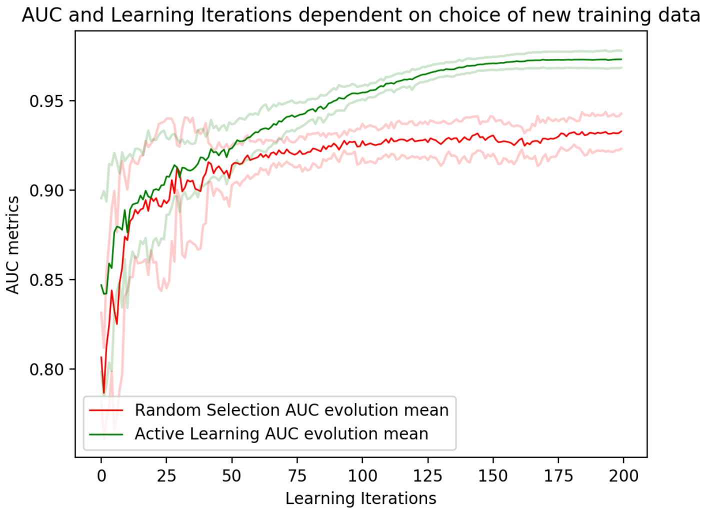

Figure 6.19: Comparison of random and active learning text data selection mean and one standard deviation for DENFI algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and Sports vs Comedy categories

we continue training, using the weights from the previous iteration with addition of gaussian noise in the end. Moreover we are using lower number of epochs. This kind of ”hot start” approach gives perfect variability that is shown in quality of results in further section.

6.4.5.1 Sports and Comedy Categories

In figure 6.19 is shown active learning simulation results for DENFI, Fast Text encoding and Sport vs Comedy categories. Active learning strategy is based on entropy acquisition function.

The active learning result, displayed in figure 6.19 outperform all the algorithms that were tested before. We observe that the uncertainty bounds are extremely narrow. In addition to this we see that active learning strategy has much higher AUC metrics and the curves are quite smooth. Moreover, due to hot start training, the time needed to fit the algorithm is much lower in comparison to Neural Network Ensembles.

6.4.5.2 Crime and Good News Categories

In figure 6.20 is shown active learning simulation results for DENFI, Fast Text encoding and Crime vs Good News categories. Active learning strategy is based on entropy acquisition function.

We have already talked in SGLD section that Crime and Good news categories can be well separated from each other. In this case we see same behavior for active learning strategy as in previous section. The uncertainty bounds are narrow and the AUC scores are very high in comparison to the random sampling strategy.

6.4.5.3 Politics and Business Categories

In figure 6.21 is shown active learning simulation results for DENFI, Fast Text encoding and Politics vs Business categories. Active learning strategy is based on entropy acquisition function.

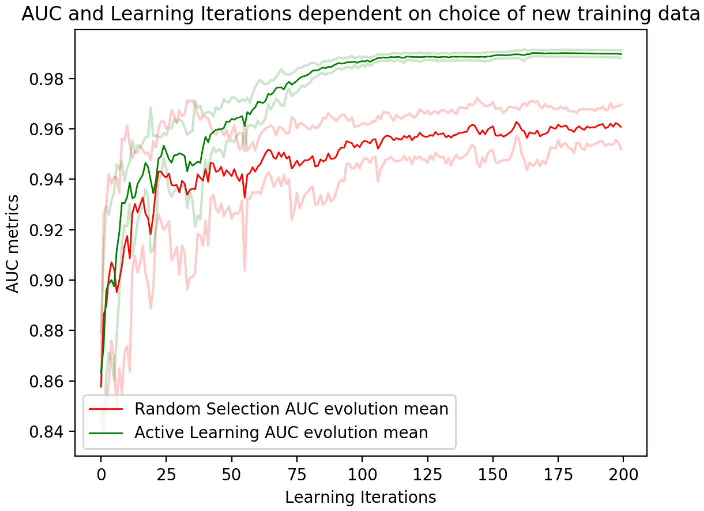

Figure 6.20: Comparison of random and active learning text data selection mean and one standard deviation for DENFI algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and Crime vs Good News categories

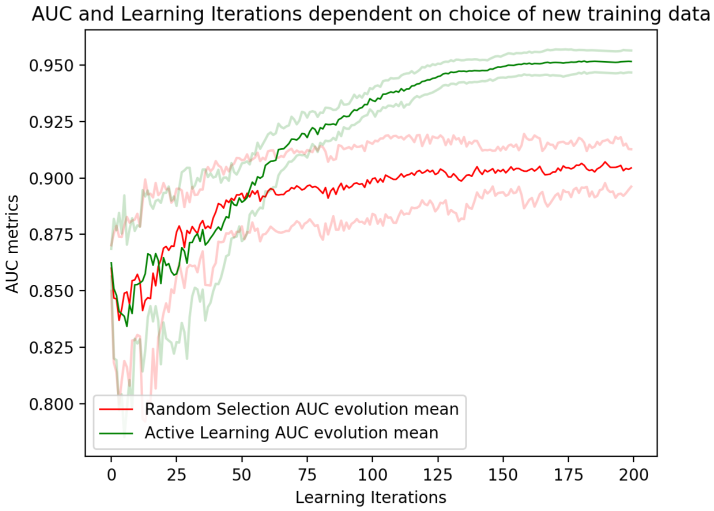

Figure 6.21: Comparison of random and active learning text data selection mean and one standard deviation for DENFI algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and Politics vs Business categories

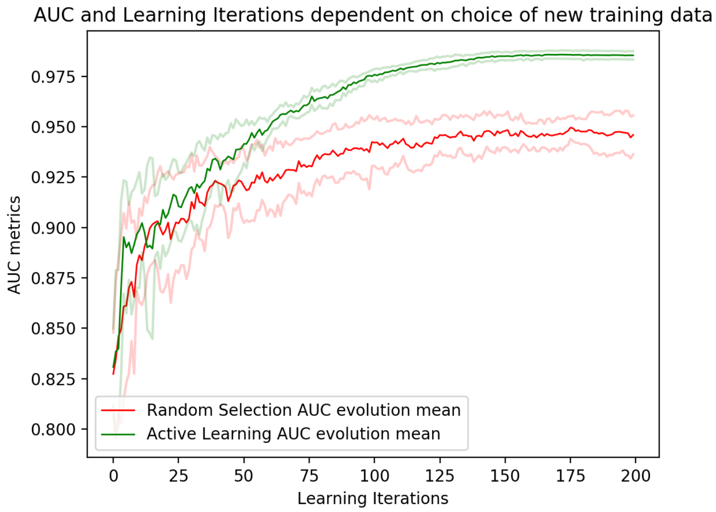

Figure 6.22: Comparison of random and active learning text data selection mean and one standard deviation for DENFI algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and Tech vs Science categories

The behavior, observed in figure 6.21, is same as in previous DENFI results. Moreover, we see good active learning performance, even though it is not easy to separate these categories. If we compare results in figure 6.21 to SGLD results in figure 6.15 we can see DENFI is better not only because its less corrupted with noise but also in higher AUC scores.

6.4.5.4 Tech and Science Categories

In figure 6.22 is shown active learning simulation results for DENFI, Fast Text encoding and Tech vs Science categories. Active learning strategy is based on entropy acquisition function.

The active learning strategy results for DENFI algorithm are much better in comparison to SGLD method that are shown in figure 6.16. We see that for active learning strategy case DENFI easily found the patterns which helped it to learn faster and show better scores with more narrow uncertainty bounds.

6.4.5.5 College and Education Categories

In figure 6.23 is shown active learning simulation results for DENFI, Fast Text encoding and College vs Education categories. Active learning strategy is based on entropy acquisition function.

Last but not least results for College and Education also proof that DENFI active learning strategy is not too much dependent on how big intersection is between the categories. The uncertainty bounds for active learning are still narrow and the AUC score is much higher than for random sampling case.

6.4.5.6 Positive and Negative Tweets Categories

In figure 6.24 is shown active learning simulation results for DENFI, Fast Text encoding and Positive vs Negative tweets categories. Active learning strategy is based on entropy acquisition function.

In comparison to SGLD, DENFI algorithm show better active learning performance for tweets dataset. DENFI active learning strategy is also working on tweets whereas SGLD active learning strategy was

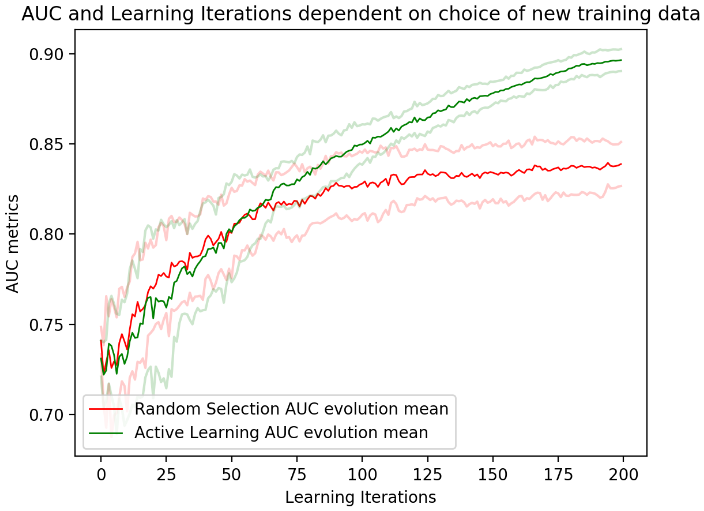

Figure 6.23: Comparison of random and active learning text data selection mean and one standard deviation for DENFI algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and College vs Education categories

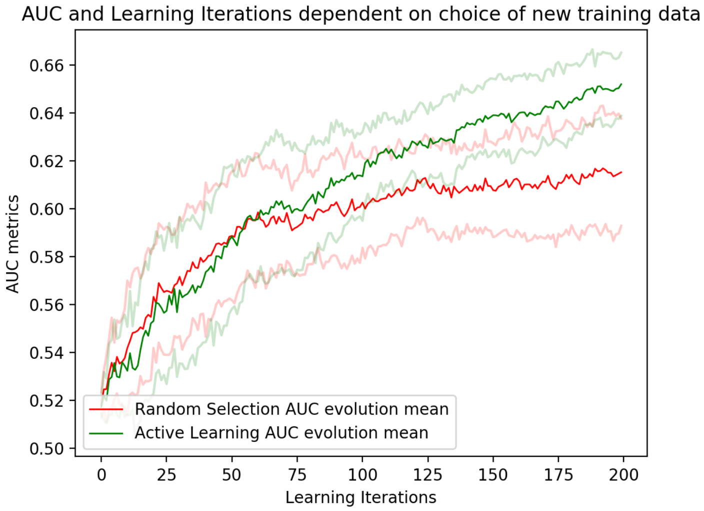

Figure 6.24: Comparison of random and active learning text data selection mean and one standard deviation for DENFI algorithm with respect to 10 samples as initial training configuration, 200 data selection iterations, Fast Text encoding and Positive vs Negative tweets categories

not able to overcome the random sampling. We can also see that DENFI AUC scores for active learning are higher with more narrow uncertainty bounds than for SGLD.

6.4.5.7 Conclusion

To conclude we can say that DENFI showed brilliant results for all problems which it was tested on. In all cases active learning strategy outperformed random sapling. The algorithm was tested both on the data that are easy to separate and hard to separate. In addition to this, the AUC scores for all active learning results were higher than for different algorithm. That makes DENFI algorithm the best one with the highest performance and the lowest uncertainty, which was tested in term of this project.

Conclusion

This project shows how active learning strategy of querying unlabeled text documents for further labeling and training can beat random selection strategy. We have provided not only high level theoretical description of the problem but also testing results that cover different scenarios and text document categories. Github link for Python implementation is also available in this project.

Based on the achieved result that were gathered from testing on 12 different categories, we were able to see that a modification of the DENFI algorithm shows great performance and overcomes other algorithms in all aspect. DENFI outperforms all other models both in higher AUC scores and more narrow uncertainty bounds. Another huge advantage that DENFI was the fastest neural network model, that was implemented and tested in this theses.

We see plenty of further opportunities how it is possible to improve the algorithm, starting from better text representation and ending using other DENFI modifications. To conclude, we would like to say, that the ensembles showed their power of solving active learning problem and in our opinion it is good field for continuing the research.

Text of the conclusion...

Bibliography

[1] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[2] Bang An, Wenjun Wu, and Huimin Han. Deep active learning for text classification. In *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, pages 1–6, 2018.

[3] James O Berger. *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer Series in Statistics. Springer, New York, 1985.

[4] Christopher M Bishop. Machine learning and pattern recognition. *Information Science and Statistics. Springer, Heidelberg*, 2006.

[5] Sophie Burkhardt, Julia Siekiera, and Stefan Kramer. Semisupervised bayesian active learning for text classification. In *Bayesian Deep Learning Workshop at NeurIPS*, 2018.

[6] Shubhomoy Das, Md Rakibul Islam, Nitthilan Kannappan Jayakodi, and Janardhan Rao Doppa. Active anomaly detection via ensembles. *arXiv preprint arXiv:1809.06477*, 2018.

[7] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[8] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017.

[9] Huang L. Go A., Bhayani R. Twitter sentiment classification using distant supervision, 2009.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017.

[12] Elizabeth D Liddy. Natural language processing. 2001.

[13] David Lowell, Zachary C Lipton, and Byron C Wallace. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30, 2019.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[15] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[16] Rishabh Misra. News category dataset, 06 2018.

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[18] 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.

[19] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[20] Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[21] Claude E Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[22] Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*, 2017.

[23] Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980, 2019.

[24] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[25] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[26] Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

[27] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.