Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science
Data Mining Research Group

Framework for multi-view predictive visual analytics for a Customer Due Diligence report

Master's Thesis

Juan Manuel González Huesca

Graduation committee:
prof. dr. Mykola Pechenizkiy (supervisor, TU/e)
ir. Simon B. van der Zon (supervisor, TU/e)
Werner van Ipenburg MSc (supervisor, Rabobank)
dr. Massimiliano de Leoni (TU/e)

Final version

Eindhoven, August 2018

Abstract

In this thesis, we propose a multi-view semi-supervised framework for a risk classification task in a Customer Due Diligence (CDD) report. The CDD report helps the financial institutions to screen new applicants and existing customers against restricted entities, sanctions lists and high-risk parties to label high risk customers; it is a vital element to protect and shield the financial system against illegal activities.

While financial crime is growing in scale and complexity, we need to keep the pace and develop more efficient solutions that leverage all the available growing customer data. This leaded us to study and choose the co-training algorithm as it has been demonstrated in the past that it outperforms single-view learning schemas. It also allows us to leverage a big amount of unlabeled data while having the opportunity to choose the most suitable classifier for each view (text and transactions).

The proposed framework aims to augment the capabilities of CDD experts to perform their work more efficiently and in a more effective way. We want to do this by providing not only accurate classifications but also an interpretable model, that explains why and how the prediction was done. Further, we exploit the properties of a visual analytics design to provide meaningful insight and to have humans (domain experts) in the CDD process.

We demonstrated that our co-training proposal outperforms a single classifier schema and other setups. It also provides a good trade-off between accuracy and interpretability. Based on empirical results, we also found the combination of initial training instances and number of iterations that produces the best performance. In addition to this, we showed empirical evidence that our visual analytics tool improves the knowledge available to CDD experts by showing relevant information in a simple and intuitive manner.

The framework will increase the insight and the tools available to CDD experts, and will allow them to optimize the resources by focusing on high-risk customers that are highlighted by our co-training proposal. Furthermore, the model will support one of the main financial institution's goals: be a rock solid bank.

Acknowledgements

This was quite an experience. This thesis was a great opportunity to do machine learning research to help solving an important challenge in the banking industry.

I owe my deepest gratitude to my supervisor prof. dr. Mykola Pechenizkiy, he provided not only outstanding technical guidance but also the motivation and encouragement needed to keep giving my best every day. dr. Pechenizkiy always found the time in his agenda to provide valuable insight and feedback, even on weekends. I would also like to thank ir. Simon B. van der Zon who constantly challenged me and guided me through the whole thesis. We had very interesting conversations and we shared many experiences that strengthen our relationship. I am deeply grateful for the opportunity I received from Werner van Ipenburg and Jan Veldsink to join their team, and also for all the relevant and prompt feedback provided along the project and during our weekly meetings. I also highly appreciated the time and effort put by the CDD expert in the whole thesis, from the scope until the final experimentation. Furthermore, I would like to thanks all my friends and colleagues for the help and support during my whole master studies, both in France last year and in the Netherlands this year. Finally, I want to express my greatest gratitude to my family who has supported me through all the years of my study, they have motivated me and encouraged to pursuit my goals.

Contents

List of Figures

Glossary

CDD Customer Due Diligence. iii, iv, 1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 18, 19, 20, 23, 24, 26, 28, 30, 31, 32, 33, 34, 35, 36, 37, 38

DNB De Nederlandsche Bank. 4

EROCS Entity RecOgnition in Context of Structured data. 11, 14

HTML Hypertext Markup Language. 24

IR Information Retrieval. 13, 14, 15, 26

KYC Know Your Customer. 4, 14

MAP Maximum A Posteriori. 16

MNB Multinomial Naive Bayes. 12, 16, 27, 38

POS Point of Sale. 18

ROC Receiver Operating Characteristic. vii, 2, 8, 27, 28, 29, 30, 31, 32, 35, 38

SQL Structured Query Language. 24

TF-IDF Term Frequency–Inverse Document Frequency. 11, 15, 18, 26

URL Uniform Resource Locator. 14

List of Algorithms

Chapter 1

Introduction

The Customer Due Diligence (CDD) report is the cornerstone for effective anti-money laundering and counter-terrorism financing programs, as well as for other illegal activities. It provides a complete overview and allow financial institutions to know their customers. This same importance persuaded us to find new solutions to improve the current CDD process by using all the available customer data (multi-view data).

This multi-view learning paradigm has been explored in many fields over the past years [41]. In most of them, the results have outperformed significantly the single-view learning scenarios. In this thesis we investigate how we can provide a reliable semi-supervised classification schema which will boost the knowledge and insight available for domain experts working in the banking industry. We want to provide a framework using different customer data sources for a risk classification task in a CDD report.

The most popular multi-view learning algorithms can be categorized as co-training, multiple kernel learning, and subspace learning [21]. But, focused on the quest to leverage a high amount of unlabeled customer data from different sources and to improve the CDD report, while including the domain experts in the knowledge discovery process, the one that got all our attention was co-training. In a co-training schema, the labels predicted by two classifiers (one per data view) help each other to retrain themselves and augment the labeled dataset.

We demonstrated that not only multi-view learning provides more significant insight than single-view learning, but also that it can be applied under this specific banking domain scenario, where we have text and transactions as the different customer views.

Another important aspect to notice, as explained in [25], is the trade-off between the classification accuracy and the model interpretability. While in some cases you just want to know what was predicted, in this particular scenario we are interested in why the prediction was made, the model also has to give an explanation on how the prediction was done. This will boost the knowledge and insight for the future cases of the CDD experts.

Visual analytics constitute a central component of the framework. When there is a task or process with a high degree of complexity and uncertainty, such as the CDD report, there is a need to expand the tools available to domain experts to carry out their activities in a more productive and effective way. The design territory of these visual tools is massive and includes many considerations like how to create and how to interact with these visual representations. We considered all these factors to proposed a visual analytics tool which will present the framework insight to the CDD experts. This tool will highlight the knowledge needed for an informed decision-making CDD process.

This thesis was written in collaboration with a major Dutch bank.

1.1 Motivation

By understanding the importance of the Customer Due Diligence report, not only for this financial institution but to shield and protect all the financial system against illegal activities, we are eager to provide a robust classification framework that will help the Customer Due Diligence performing their job more efficiently.

The large amount of data generated over the last decade allow us to research and create new solutions to old problems, in this case, to leverage all the available customer data (from different sources) to improve the CDD process by providing reliable customer risks prediction but specially, keeping the model interpretable so CDD can learn from the model prediction explanations. This is not just a novel integration of data types in a co-training schema (text and transactions) but also, a new approach in the banking industry to tackle this challenge and put the CDD experts in the knowledge discovery process.

1.2 Thesis objective and methodology

The main thesis objective is to demonstrate through an experimental analysis that a multi-view semi-supervised learning framework can outperform a single-view schema in the specific context of a Customer Due Diligence report, in this case, while dealing with text (unstructured data) and transactions (structured data). We also want to show how we can effectively include the domain experts in the knowledge discovery process, by providing relevant insight (top features for a class). Part of the previous objective is related to the creation of an effective transactions feature extraction process, which can characterize the customer transactions and provide discriminatory attributes to build a classification model.

We want to find the correlation between the text and transactions features, because in the assumption that the transactions data is ubiquitous for every customer, the correlation coefficient could give the CDD experts some hints of where to start the inquiries for a new case, if there is no text available at the time.

Visualization is an important part of the knowledge discovery process. For that reason, we also want to include best practices in visual analytics to support the insight obtained from the framework.

We will divide the experiments in two parts: the first is a technical analysis where we will measure correlation (for text and transactions features), accuracy and Receiver Operating Characteristic (ROC) of the models; and the second is focused on the application side, where we meet with domain experts and conduct interviews to compare our framework results with their analysis. As there is a class imbalance in the dataset, the ROC curve will show us the trade-off between sensitivity and specificity, by measuring the relation between the false positive and true positive rates.

1.3 Results

The results are promising. On one side we demonstrated that multi-view learning can be used to improve the overall performance of a single-view classifier; and on the other side, we confirmed that we can build a robust semi-supervised learning framework where domain experts can leverage valuable insight extracted from an interpretable model.

We also discovered that the transactions classification plays an important role in the process, and that some text and transactions features are correlated. One crucial remark to make is

the existence of a trade-off between accuracy and interpretability, which should be taken into consideration when defining the overall CDD process goal.

The results also showed the importance of a visual analytics tool, as the domain experts can gain more knowledge and insight when the information is presented using best practices for visualization design and validation. The outcome truly showed how we tackle this opportunity area in the banking industry with the development of this novel multi-view framework.

1.4 Thesis structure

This thesis is organized as follows. In Chapter 2, we present the circumstances in the banking industry that motivated the current research and framework. In Chapter 3, the problem is defined in detail and the challenges are introduced, both on the academic and on the application side. In Chapter 4, we introduce our framework and discuss in detail its two main parts: the text classification (unstructured data) and the transactions classification (structured data). In this Chapter, we also discuss the co-training algorithm and how we want to enhance it. Then, in Chapter 5, we move to introduce and explain all the experiments and results that help us to respond our research questions. Additionally, in this Chapter we also present a case study. Finally, Chapter 6 concludes with the highlights of the framework, the limitations and the description of the future work.

Chapter 2

Background

In recent years, financial institutions have strengthened the Know Your Customer (KYC) [30] processes. These processes help to identify the behaviour of customers, to assess risk and furthermore, to protect the financial system from money laundry and prevent the financing of terrorism. One of the key requirements in these processes is the CDD report [14], which is applied on every new customer as well as on selected existing ones. This report provides an overall outlook about the risk that a customer depicts, based on different characteristics such as: geographical locations, economic activities, relation with political organizations, and any other significant aspect that may be a threat to the bank interests. After this analysis, a risk classification is set (low, medium and high) based on the information collected and analyzed by CDD experts.

2.1 Key stakeholders

Within Fraud & Compliance, there is a highly qualified team of CDD experts. This team performs customer analysis while onboarding; but also on already existing customers based on signals detected through automatic systems or any other suspicious signals detected by the experts. There are three risks level (low, medium and high) and several risks being assessed by the bank.

The bank as a whole is very interested in the results. Following the compliance strategy, it is looking to be a healthy and sustainable financial institution providing secure services to customers. Also, this benefits the whole financial system as this process protects and shields it against illegal activities. Last but not least, the customers profit by having their money and investments in a rock solid bank.

The bank has selected and defined a set of significant customer risks, following the guidance provided by the *De Nederlandsche Bank (DNB)*, the central bank of the Netherlands. According to DNB [7], when working on relevant risks identification, among other information, we should consider "The institution's customers, services, products and delivery channels are the starting point for the identification of integrity risks. Country and geographical risks are also important. These relate to countries and regions where the institution is active itself or where its customers are established or conduct activities. When preparing an integrity risk assessment, the institution looks at the characteristics of different types of customer, such as sectors or professions, residency or assets and source of income. It also looks at how the contact with customers is generally established and how services are offered (the 'delivery channels', e.g. in person or otherwise, via intermediaries, by telephone or online)."

The Fraud & Compliance unit will leverage our framework as it will provide speed, prediction explanations (insight) and it can potentially save manpower, so resources can be allocated on other key areas. Our framework will also allow to increase the efficiency of the resources allocation, in

this way the CDD experts can focus on the right customers, the ones that represent a bigger risk to the bank.

2.2 Positioning of the project work

The insight provided by the framework would be very useful to allocate team's resources in the cases where it's more relevant. As indicated by a CDD expert, the framework has the potential to discover high risk clients among a portfolio of 10 million, especially potential high risk entities that may go unnoticed.

The potentiality of the project is very relevant. It will indirectly help to ensure that the financial institution is aligned with arising stronger regulatory compliance needs [12], as the volume and complexity of them have increased. In this way, we ensure the whole financial system benefits.

2.3 Framework overview

The proposed framework leverage the power of multi-view learning, as it's assumed that this learning paradigm gives a broader understanding of a specific task that leads to a better performance [34]. The goal is to explore and mine effectively the information from multiple views of the same instance to improve the learning efficiency. As it can be observed in Figure 2.1, the setup contains two classifier, one per view (text and transactions). This allow us to choose the most suitable classifier per data type.

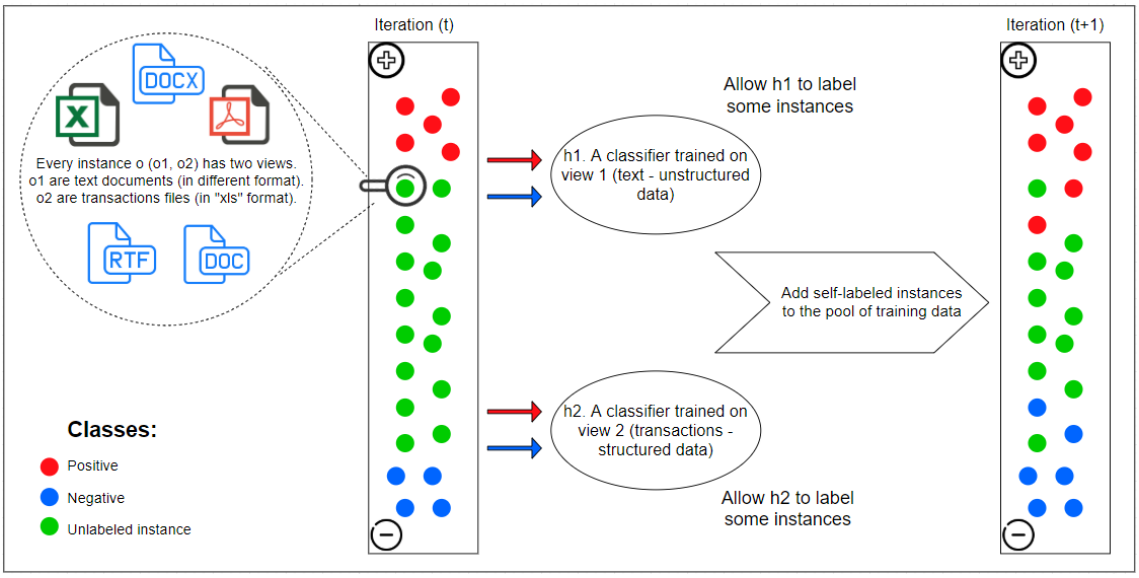

Figure 2.1: Simplified operation of a Co-training learning schema

As we have a high amount of unlabeled data, this learning paradigm allow us to construct a robust classification schema by augmenting the labeled data. This is done by predicting iteratively and labeling new instances, which are then used again to retrain the models. As we have more iteration rounds we have more labeled data which allow us to build a more robust and reliable classification. The prediction of each classifier helps each other to gain more knowledge and have more available labeled data. It's worth to mention that every instance o has two views: o_1 is a text document which describes a customer (the files can be in "doc", "docx", "pdf" or "rtf" format), and o_2 is a a file or a set of file with transactions (the files are in "xls" format).

It's important to notice that as we make more iterations, the number of labeled training instances increases and therefore, the number of unlabeled instances decreases, but there are two trade-offs, there is no linear relation between the number of iterations and the model performance, and also interpretability has an accuracy price to pay.

Chapter 3

Research questions

Over the last decade, regulatory compliance for the banking industry has become more strict [11]. For that reason, banks are investing in training courses for employees and new systems to improve the current compliance processes. This leads us for the need of a reliable system to classify customers based on the available data (text on different domains and transactions), but with the aim to have humans (CDD experts) in the loop.

The bank is interested in finding a new, effective and automated process for knowledge discovery, so it's employees can target correctly customers with a higher risk potential (ones that fit with suspicious patterns), and manpower can be employed efficiently. Also, as each individual within the Fraud & Compliance team has different knowledge and expertise, the outcome (most relevant features per risk) of the framework will provide a general overview of the most important elements to consider when assessing customer risks on a CDD report.

Nowadays, there are many systems to support the work done by CDD experts, but the increasing amount of available customer data brings new improvement opportunities. We want to introduce a first step in automating the whole process while leveraging text (reports describing the customer) and customer transactions. And, at the same time, we want to take advantage from all the customer data, from different domains, in different formats; which will also help to decrease the false positive rates.

The main question we will answer with our proposal is: How can we leverage all the available current customer data (structured and unstructured) to build a framework in order to predict customer risks in a CDD report? But more importantly, how can we integrate CDD experts in the loop so they can benefit from the provided insight?

3.1 Description

In a more formal and precise manner, we want to address the correlation among features, the quality of our features extraction proposal for transactions, the effectiveness of the multi-view classification with a co-training schema and how the visual analytics design can boost knowledge acquisition. We also want to validate the framework interpretability and make sure there is a good trade-off between the prediction explanation and the accuracy.

We target to answer the following questions:

1. Is the correlation between multi-view (text and transactions) features significant enough to make the assumption that one may imply the other?
    - The data for a customer sometimes can be incomplete, for a particular customer we may just have the transactions. This analysis can provide further steps in how to approach

these scenarios. We can focus in some specific text features (key words) correlated with transactions features, these can be used as an initial inquiry lead.

- We use the Pearson correlation coefficient to measure the association between top multi-view features in a numeric and visual way.

2. Is the transactions feature extraction process (number and quality of features) good enough to provide good classification scores and relevant insight to CDD experts?

   - We are proposing features based on feedback from CDD experts, literature review and statistics. This will allow us to construct a classification model. We want to confirm we can construct a sound model based on the extracted features, so we can predict risky customers.
   - The plan is to create a transactions classification model and compare the prediction and top features with the analysis of a CDD expert. We expect to find a high correlation between the CDD expert advice and our model prediction.

3. Does a multi-view co-training classification algorithm improve the scores and insight provided to CDD experts? Is it better and more relevant than showing separate classifiers' results?

   - This will provide awareness of whether is valid to invest time and effort in looking for more customer data (multi-view), in order to create a continual classification improvement cycle.
   - One senior CDD expert will be interviewed to validate the insight provided. We will also measure the classification scores (accuracy and ROC).

4. We want to analyze two trade-offs. The first related to the number of co-training iterations, the number of initial training instances and the classification performance as explained in [40]. The second one related to the compromise between interpretability (model explanations) and accuracy.

   - In the case of the first trade-off, it is demonstrated in [4] and [40] that, based on the nature of the classification task, the prediction performance fluctuates as we change the number of initial training instances and the number of iterations; this is because having more initial training instances helps the model to learn more, and intuitively, having more iterations allows the model to become stronger (but we should consider the performance threshold). This previous analysis can provide the best parameter setup in building a robust and reliable framework. On the other hand, in the case of the second trade-off, it is demonstrated in [25] what and why this happens. The analysis on last statement will help us to get to an agreement on how much accuracy we are willing to give up in order to gain model explanation.
   - We will repeat the experiments 10 times, to make sure we have consistent and steady results.

5. Following best practices in design, what is the best approach and layout for a useful visual analytics setup?

   - We want to include humans in the knowledge discovery process. We want CDD experts to leverage and benefit the most from the insight provided by the framework. Our design should consider best practices in visual analytics and decrease as much as possible the influence of cognitive biases [15] and visual noise.
   - We will create a visual analytics dashboard using best practices in visualization. Via an interview, we can measure the comfort of a senior CDD expert towards each demonstration.

In order to achieve relevant conclusions and successfully answer our research questions, our experiments are divided in two main sections as outlined in the introduction. The technical experiments are done by measuring the Pearson correlation coefficient (for question 1) and the accuracy and ROC value (for questions 2, 3 and 4) where we compare these five configurations:

1. Two separate classifiers (one for text and the other for transactions) with a simple score combination.
2. Classification using a simple concatenation of text and transactions features.
3. Classifiers self-training ("co-training" without multi-view).
4. Original co-training algorithm.
5. Our proposed co-training algorithm enhanced.
6. Our proposed co-training algorithm enhanced (only when both classifiers agree).

3.2 Challenges

Besides the predicted risks for new customers, CDD experts want to know what the key features are that lead to every classification. This is very important, because at this point, we do not just want to classify new customers, but we want to provide useful insight to the researchers so they can focus (based on the most relevant features per risk) on customers which have these or some of these features describing them. We also want to prove that we can leverage the multi-view existing customer data in this specific case, so we have a set of different challenges and opportunity areas. So, in order to better classify the challenges, we can divide them in two groups: the technical and the research.

- Technical challenges
  - There is a challenge to understand the data itself, besides the text data is in Dutch, we need to understand the CDD process and translate it into technical requirements and actions.
  - As we want to leverage all the available customer data, which is in different formats ("pdf", "docx", "rtf", "xls", etc.) and comes from different domains, we need to find an effective and efficient way to extract the data and transform it so we can apply machine learning techniques to find patterns and build a model.
  - In some cases, the customer file names don't follow a naming convention, so it's very difficult to identify the relevant files associate to a customer. This is more important when we have many files related to a specific customer, and we want to focus the main and most decisive one.
  - We want to design experiments with CDD experts, where they are not told in advance the real purpose of the experiment (to reduce biases as much as possible, but they are presented to the framework's insight (from different algorithms) for assessment. The challenge is to consider all (or as much as possible) cognitive biases in order to design useful experiments to prove our framework.
- Research challenges
  - As we want to include the researchers in the knowledge discovery process loop, there is a need to create interactive visual interfaces. This visual analytics tool will allow people to combine their expertise, background knowledge, human flexibility and other capacities to make well informed decisions on future complex customer analysis. So, the challenge is to use state-of-the-art visual analytics design techniques to show the framework insight to the CDD experts.

- There is no record that co-training has been used before for a CDD report, so there are many unexpected challenges to sort out, from features extraction to data modeling.
- The features from one customer data view were extracted by us. This is very important to consider as we had to understand what is the analysis the CDD experts do in order to propose the most relevant features that could capture every customer characteristic needed to make a correct classification.

Chapter 4

Framework

The current framework was created for a semi-supervised risk classification task, while leveraging multi-view customer data: text (unstructured) and customer transactions (structured). The aim is to construct a robust predictive model that have good classification scores and, at the same time, that is interpretable. The interpretability plays an important role as it will help, in a visual way, the CDD experts to gain more knowledge and improve the efficiency of the CDD process.

As it can be seen in Figure 4.1, the framework allow us to have a multi-view predictive model with two classifiers, one for text and the other for the transactions. We used the power of co-training [4] as it's proven that we can boost the performance of a learning algorithm and leverage inexpensive *unlabeled* data to enlarge a small set of reliable labeled data, while learning from a multi-view dataset. The two classifiers are trained separately, and the predictions of each algorithm are used to augment the training set of the other. This is very important when not all the data is labeled or when the available labels are not completely trustworthy.

The framework gives us flexibility as we only take labeled instances with a certain degree of accuracy. In this way, while having several iterations to train and classify our instances, our model is more robust and the performance is better. It is worth to mention that the different views of the same instance (customer) are linked by the customer name extracted from the reports and the customer repository folder name (finding the closest match when some misspelling [31]). Another method similar to Entity RecOgnition in Context of Structured data (EROCS) [6] was implemented and considered, where we heuristically looked for numbers in an account numbers format while mining the text files, but at the end the first approach was more effective. This allowed us to create a relation between TextFile-CustomerName. The other relationship needed (CustomerAccount-CustomerName) was extracted from the transactions, as well as the TransactionsFile-CustomerAccount. In this way, we can link text and transactions data with a particular customer.

Visual analytics is a significant element of the framework as it helps people to carry out tasks more effectively, it augments human capabilities. We have built a dashboard which shows at a glance the risk labels of a particular customer. It also shows the prediction explanations, for both, the model and the specific customer. This dashboard will be the foundation of a more complex and interactive learning system like the one presented in [18].

Formally speaking, we assume we haven an instance space $O = O_1 + O_2$, where O_1 and O_2 represents different "views" (one for the text and one for the transactions) of the same object (customer). Which means, every training instance o is (o_1, o_2). Also, it's important to mention that we assume every view would be enough to create a classification model. Then, we have the feature extraction section. For the text data, we used Term Frequency–Inverse Document Frequency (TF-IDF) [22] to create a vector space representation. In the case of the transactions,

we used several characteristics provided by CDD experts, literature review and statistics to extract features from the transactions.

Once the features are extracted, we have two predictive models, Multinomial Naive Bayes (MNB) for the text data and Random Forest for transactions. When the models are trained, we can classify new instances, and this is when we leverage the multi-view customer data: the new classified instance from classifier 1 is used to re-train classifier 2, and so on.

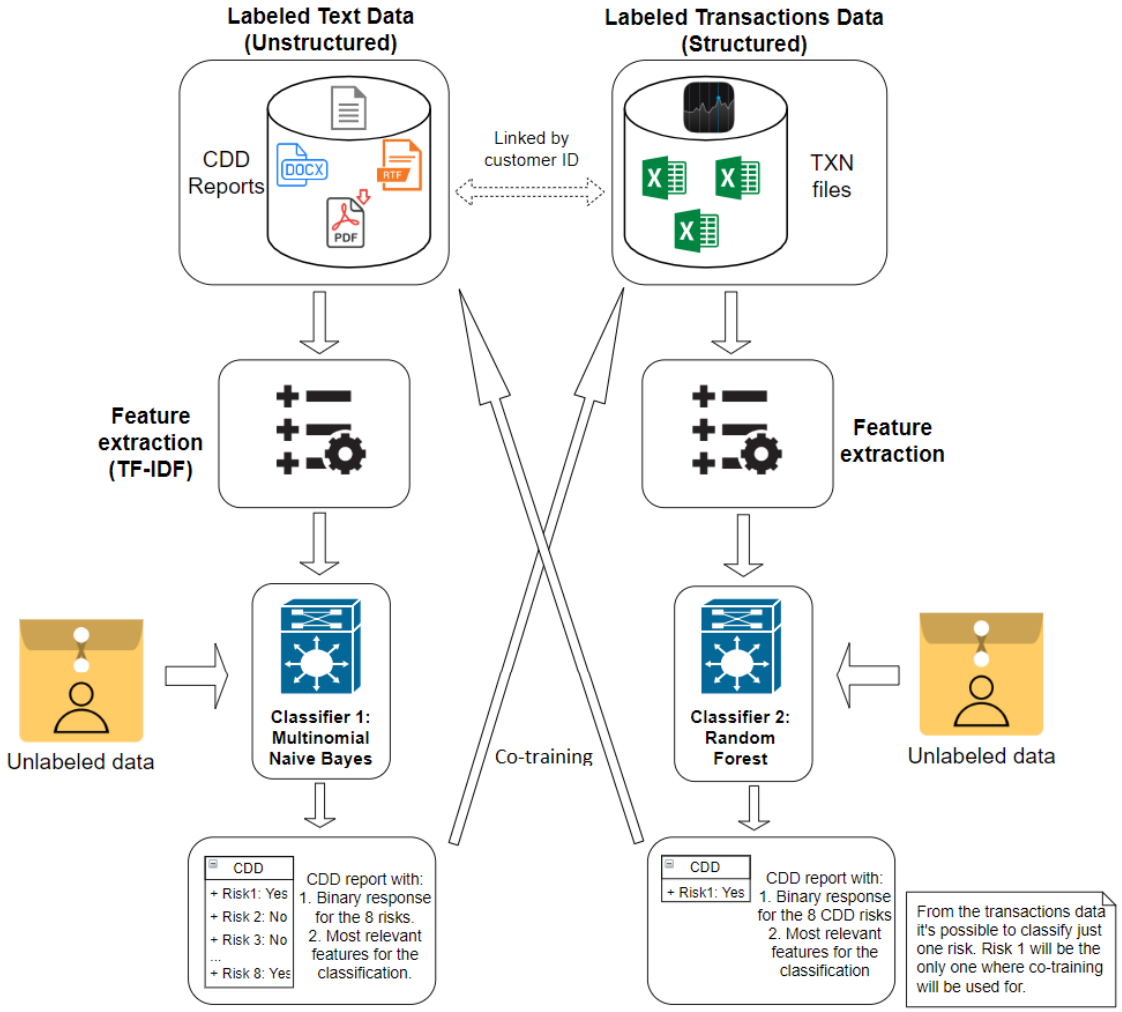

Figure 4.1: Framework for multi-view predictive analytics

There is a remark to make. Out of all the risks in the CDD report, the transactions data can only provide features for one of them, so the co-training algorithm will be implemented for the classification of just this risk. The other risks will be classified entirely based on text data.

4.1 Related work

In the past, many research efforts have been made in the banking industry to identify and understand better the behavior of customers. Most of these endeavors have been done to combat money laundering and to prevent financing of terrorism, but also to assess different risks and for

credit scoring. In [19] and [26], researchers focused on customer credit scoring but the analysis target only transactions patterns and the end result is a simple denial or approval for a customer credit request. Other approaches have been fully focused on anti-money laundering. For example in [8], the researchers presented a framework which detects outliers and unusual activities based on transactions, and also, they use the power of links analysis in their proposal.

There are also other approaches focusing more in profiling the customers. An unsupervised machine learning approach is presented in [1], where the aim is to cluster customers based on transactions patterns (volume and frequency). On the other hand, in [35] the researchers introduced a novel approach combining network analysis and supervised learning, but it just analyzes customer transactions.

All of these previous studies have provided valuable information. We have learned several customer characteristics and features to consider while creating our framework, and it has also allowed us to understand better what are the specific business needs and requirements in the banking sector. Unlike previous approaches, in our framework proposal we leverage the power of the multi-view classification so we can analyze and benefit from all available customer data (text and transactions) in order to have a better characterization of the customers. This will not only allow us to build a more accurate predictive model, but also to gain fruitful insight that can help and put CDD experts in the knowledge discovery loop.

4.2 Functional requirements

The following are characteristics that capture in a more detailed way the scope, objectives and requirements of the proposed framework. Here, we will emphasize on what the system will do and accomplish.

- The framework must include an IR system with a very high precision to retrieve just relevant text documents and transactions files related to a target customer.
- The transactions feature extraction process must have a high degree of soundness to provide relevant and meaningful characteristics to be used by an ensemble method.
- The framework must provide accurate classification scores for new instances (customers) with two views: text and transactions. It should do it by leveraging a high amount of unlabeled data.
- An interpretable predictive model needs to be created as an outcome. The top important features per classifier (text and transactions) will be displayed, as well as the specific impact of these on every particular customer.
- A visual analytics dashboard should be used to present the framework results: predictions and model explanations.
- The model must provide a Pearson correlation matrix of top text and transactions features. The domain experts should analyze this information (Pearson coefficient) and make a decision whether transactions feature may imply text features.

4.3 Text classification analysis (unstructured data)

As stated in [24], "In general, most information available in the real world exists as written or recorded words. This is probably the reason why text mining is capturing significant attention in the business and research communities concerned with practical applications". While this may seems obvious, it is important to understand the role that text has played in history, from the Code

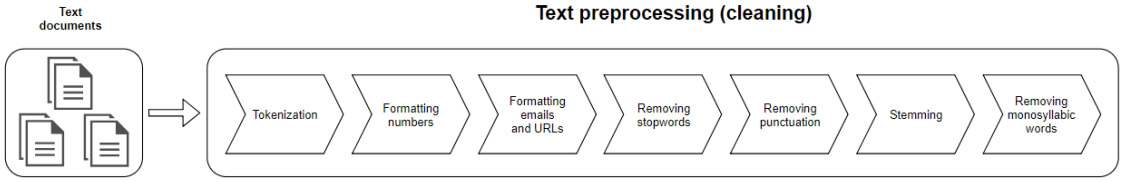

Figure 4.2: Text preprocessing pipeline

of Hammurabi (one of the oldest writings in the world), to the 500 millions tweets per day [13], and realize the huge potential it has for a knowledge discovery process.

This leads us to find efficient and effective techniques and algorithms, that will allow us to extract meaningful and useful information from the exponentially growing amount of text that is being generated on a daily basis.

The banking industry is no exception. Financial institutions want to leverage this tremendous volume of mostly unstructured text, that is why we will take advantage of the current available customer data to gain some insight and improve the KYC processes.

The text classification pipeline will be divided on three main steps: text preprocessing, the creation of a vector space representation and the classification algorithm.

4.3.1 Text preprocessing

A very important step in building a text classification model is the preprocessing. Because of its nature, raw text is very difficult to process, therefore we have defined seven steps to clean up text documents, as can be seen in Figure 4.2.

Below we explain every preprocessing step:

- Tokenization. This step implies the task of chopping a defined document into pieces, called tokens. As stated by The Stanford Natural Language Processing Group: "These tokens are often loosely referred to as terms or words, but it is sometimes important to make a type/token distinction. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. A type is the class of all tokens containing the same character sequence. A term is a (perhaps normalized) type that is included in the Information Retrieval (IR) system's dictionary."

- Formatting numbers. As account numbers are important to link a text document with a customer, we implemented a heuristic approach similar to EROCS to format all the consecutive numbers and try to find matches.

- Formatting emails and Uniform Resource Locator (URL). While performing many experiments, we follow an approach similar to [28] to replace any email or URL to specific terms ("emailaddr" and "httpaddr" respectively). This was done after making experiments and proving this information is not relevant in our classification.

- Removing stopwords. As stop words does not add any significant value to discriminate and describe a document, we filtered them out.

- Removing punctuation. Similar to stop words, punctuation does not provide any significant value to the classification task, so they were removed to eliminate noise.

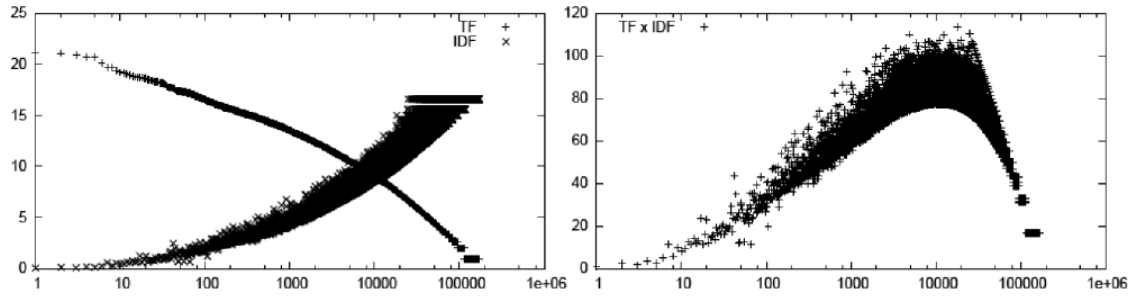

Figure 4.3: TF-IDF properties [3]

- Stemming. In this step, we try to reduce inflectional forms to get words to a common base form. Heuristically, stemming chop off the end of words to have a single term when a word is repeated multiple times over the documents but with multiple endings.
- Removing monosyllabic words. As most of the words with a length of 3 letters or less do not provide any useful data, they were removed from the documents.

4.3.2 Vector space representation

One of the first approaches to perform machine learning (borrowed from IR systems) on text documents was to represent the presence or absence (Boolean logic) of every word w in every document d, part of the vocabulary. So we would have a vector of size (i, j) where i is the number of documents and j is the number of words in the vocabulary (all the unique words in the set of documents). This Boolean approach was useful but has many limitations: it is insufficient to capture the richness of our language and it is very difficult to discriminate among document vectors as many have the same words, among others.

Then, another more sophisticated approach was introduced. Now, we were counting the number of occurrences of each word w in every document d. The shape of the matrix was the same as the Boolean counterpart, but now the number of occurrences is taken into account. The problem with this approach arises when the document lengths are different, so the probability calculations could be inconsistent and have many discrepancies. So now, the next improvement was the term frequencies (TF), which is the result of dividing the number of occurrences of each word in a document by the total number of words in the document.

There were many words with very little meaningful information about the content of the document, which were hindering the frequencies of rare and more interesting words that define the content. So, in order to discriminate documents in a more meaningful way, TF-IDF [29] was introduced (TF for term-frequency and IDF for inverse document-frequency), which is computed as:

$$tf - idf(t,d) = tf(t,d) \times idf(t) \tag{4.1}$$

$$idf(t) = log\frac{1+n_d}{1+df(d,t)} + 1 \tag{4.2}$$

Where n_d is the total number of documents, and *df(d,t)* is the number of documents that contain term *t*. TF-IDF can be used to create a vector space representation of text. It has the following properties: it presents a power-law behavior, and the terms of intermediate *idf*

values display maximum *tf-idf* weights, resulting in the most interesting for ranking, as seen in Figure 4.3. This latter properties are used, and confirmed what Luhn (1958) suggested: both extremely common and extremely uncommon words were not very useful for indexing.

4.3.3 Text classification algorithm

In order to prove our framework, we used a canonical text classification model [2], Multinomial Naive Bayes (MNB) [22]. This model, based on the Bayes' theorem, was chosen because it usually performs better than multi-variate Bernoulli providing an average of 27% reduction in error [23]. The algorithm works as described in Algorithm 1 (to train) and Algorithm 2 (to apply).

This supervised learning algorithm has the "naive" assumption of independence between every pair of features. Despite that when working with text on real-world application this latter assumption is rarely true, empirical experiments have shown a competitive performance with state-of-the-art classifiers [43], when under certain circumstances, that dependence among attributes may cancel out each other.

While looking for interpretability and faithfulness to create a human-in-the-loop system, this classifier provides the proper means to achieve them.

Given a class variable y and a dependent feature vector x_1 through x_n, Bayes' theorem expresses the coming connection [29]:

$$P(y|x_1, ..., x_n) = \frac{P(y)P(x_1, ..., x_n|y)}{P(x_1, ..., x_n)} \tag{4.3}$$

And by using the naive independence assumption we know that:

$$P(x_i|y, x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = P(x_i|y) \tag{4.4}$$

for all i, this connection is simplified to:

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, ..., x_n)} \tag{4.5}$$

Since $P(x_1,...,x_n)$ is constant given the input, the following classification rule can be used:

$$P(y|x_1, ..., x_n)\alpha P(y)\prod_{i=1}^{n} P(x_i|y) \tag{4.6}$$

$$\hat{y} = argmaxP(y)\prod_{i=1}^{n} P(x_i|y) \tag{4.7}$$

so Maximum A Posteriori (MAP) estimation can be used to estimate $P(y)$ and P(x_i|y)

The advantage of this classifier lies in the nature of the algorithm, Naive Bayes is an already interpretable model. We compute the empirical log probability of features given a class ($P(x_i,y)$), which can be used to obtained the most relevant classification features (words) that lead to a decision.

Data: C, D
Result: V, $prior$, $condprob$
begin
  $V \longleftarrow ExtractVocabulary(D)$
  $N \longleftarrow CountDocs(D)$
  for $c \in C$ **do**
    $N_c \longleftarrow CountDocsInClass(D, c)$
    $prior[c] \longleftarrow N_c/N$
    $text_c \longleftarrow ConcatenateTextOfAllDocsInClass(D, c)$
    for $t \in V$ **do**
      $T_{ct} \longleftarrow CountTokensOfTerm(text_c, t)$
    end
    for $t \in V$ **do**
      $condprob[t][c] \longleftarrow (T_{ct} + 1)/(\sum T_{ct} + 1)$
    end
  end
end

Algorithm 1: Multinomial Naive Bayes: Train Multinomial [29]

Data: C, V, $prior$, $condprob$, d
Result: $argmax_{c \in C} score[c]$
begin
  $W \longleftarrow ExtractTokensFromDoc(V, d)$
  for $c \in C$ **do**
    $score[c] \longleftarrow logprior[c]$
    for $t \in W$ **do**
      $score[c] += logcondprob[t][c]$
    end
  end
end

Algorithm 2: Multinomial Naive Bayes: Apply Multinomial [29]

4.4 Transactions classification analysis (structured data)

In a financial institution, transactions data is ubiquitous. Every customer has at least one account where they are entitled to generate transactions in a double-entry bookkeeping fashion, where every entry is either identified as debits ("D") or credits ("C"). In a broad sense the source of spending money in the account is credit, and what the money obtained with the credit is described as debit. It is important to acknowledge the difference, as most of the subsequent analysis considers this entry tag.

Due to the nature of the transactions, we proposed a time series analysis as the data is a set of values from different variables taken at successive points in time.

It's imperative to mention that there is no default standard way to extract features from transactions. Every time series analysis is unique and, in this case, we needed to understand the essence of the data. This will help us to find the most effective and efficient way to extract relevant characteristics, that could be used as features to train a classification model.

There has been many approaches in the past that have tried to extract features from time series. In [9], the author proposed a technique to find and extract patterns from time series for classification problems. Another very interesting approaches are in [20] and [16]. The first one proposes a clustering based anomaly detection where every transaction is ranked. The second proposes a Bayesian approach to detect suspicious activities. Even though the approach is different, these two latter papers define a set of important rules and transactions characteristics to examine when doing this analysis, which were considered in our framework proposal. Another helpful approach is discussed in [37], where a classification system for time series is proposed using features extracted applying TF-IDF.

Among all the different approaches, a new time series primitive discussed in [42] got our attention. The authors proposed *time series shapelets*, which are time series subsequences that in some sense maximally represents a class, but due to time and technical constraints this new primitive will be implemented in a future development.

4.4.1 Time series dataset

At the bank, we have access to customer's transactions that involved any of the accounts managed by the financial institutions. The entries are tagged as "Debit" or "Credit" depending on the situation, and are recorded in chronological order. Among the characteristics registered per transaction we can find the time stamp, transaction type, amount of money, currency, balance, Point of Sale (POS), and many others.

Out of all the transactions attributes, after some advice from a CDD expert, we chose the most relevant ones for our analysis. This allowed us to focus our effort and create one python DataFrame per customer with only key columns, which was then "cleaned" by filtering out irrelevant (based on domain expert input) transactions.

For this project, the transactions files are provided in ".xls" format, but it is possible to find transactions in other formats.

4.4.2 Feature extraction

The main component of the transactions classification analysis is the feature extraction process. We needed to understand the business requirements to be able to propose meaningful, informative and non-redundant measurements and derived values from the time series data. It was a long process with several iterations and inquiries with domain experts, but at the end we managed to extract relevant attributes that characterize the data.

Due to the nature of the data and the feature extraction process, the approach was proposed following a ensemble method, where we combine the predictions of several decision trees in order to improve generalization and robustness over a single estimator.

- CDD best practices. Along the whole project, we have met several times with CDD experts which have provided valuable input on what are the key transactions. We have analyzed their reasoning and have proposed an automated approach to extract the proper researched patterns.
- Literature research. We also considered a set of rules and transactions features described in [20] and [16]. Most of the relevant literature research scrutinize the transaction amount and the frequency as two very important aspects to consider while doing this analysis. There are many rules constructed with withdrawals, deposits, dormant accounts, balance, etc., in combination with the amount of each one of them and the frequency.
- Statistics. We have proposed a set of meaningful statistics and other characteristics describing the time series data, considering the features advised by the CDD domain experts.

4.4.3 Transactions classification algorithm

The chosen algorithm was Random Forest [5] because of its performance. It has been proven than we can build several classifiers independently and the average prediction is better than any of the single classifier, due to a variance reduction.

In order to achieve good results with ensemble classifiers two conditions needs to be presented [10]. Each individual classifier is accurate and diverse. An accurate classifier is the one that has an accuracy rate better than random guessing of new instances. Two classifiers are diverse if the errors made by the classifiers are uncorrelated. These two assumptions are taken into consideration by the random forest algorithm.

We will start explaining how a single decision tree works. The basic idea behind is to split the set of instances in subsets looking to minimize the variation within each subset. The goal is to reduce the entropy (degree of uncertainty) in leaves of trees to improve the predictability, and so maximize the information gain by the choice of the right tree leaves.

In a more formal sense the mathematical formulation is the following [29]. Given training vectors $x_i \in \mathrm{R^n}$ i=1,...,l and a label vector $y \in \mathrm{R^l}$, a decision tree is partitioning the space recursively so the samples with the same labels are bring together.

Let the data at node m be described by Q. For every candidate split $\theta = (j,t_m)$ that is composed of a feature j and a threshold $\mathrm{t_m}$, partition the data into $\mathrm{Q_{left}}(\theta)$ and $\mathrm{Q_{right}}(\theta)$ subsets:

$$Q_{left}(\theta) = (x,y)|x_j <= t_m \tag{4.8}$$

$$Q_{right}(\theta) = Q/Q_{left}(\theta) \tag{4.9}$$

The impurity at m is measured using an impurity function $H()$, the choice of which depends on the task we are working out:

$$G(Q,\theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta)) \tag{4.10}$$

Here, the parameters selections that minimizes the impurity is done:

$$\theta^* = argmin_\theta G(Q,\theta) \tag{4.11}$$

Recursively, we compute the subsets $Q_{\text{left}}(\theta*)$ and $Q_{\text{right}}(\theta*)$ until the maximum admissible depth is reached, $N_m < min_{samples}$ or $N_m = 1$.

In the case of a classification outcome taking values on 0, 1,...,K-1, for node m, representing a region R_m with N_m observations, let

$$p_{mk} = 1/N_m \sum_{x_i \in R_m} I(y_i = k) \tag{4.12}$$

be the fraction of class k observations in node m.

The impurity measure used is Gini, computed as:

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk}) \tag{4.13}$$

where X_m is the training data in the node m.

In contrast, as explained in [36], each tree in the Random Forest (ensemble) is built from a sample drawn with replacement fro the training set. Also, when the split is done during the construction of the tree, the chosen split is no longer the best split among all features. Rather, the split that is chosen is the best split among a random subset of the features. This randomness increases a little the bias (compared the bias of a single non-random tree) of the forest but, due to averaging, the variance decreases, commonly more than compensating for the increase in bias, thus resulting in an overall better model.

As mentioned before in the text classification analysis, in our framework proposal we want to include the CDD experts in the knowledge discovery loop, so a simple prediction is not enough. We also want to discover and display the features that have a bigger impact in the classification, so the CDD team can leverage of this insight and increase their knowledge for future customer investigations. This will be achieved by an attribute of the classifier that shows the importance that every feature had in determining the which split will most effectively help distinguish the classes.

4.5 Co-training

In most of today's data analytics challenges, there is a high amount of data collected or obtained from various sources that shows heterogeneous properties. And while this later could be seen as a threat, it's proven [41] that by exploring several properties of different views of the same data, multi-view learning is more effective, promising and has better generalization ability than single-view learning. This is the main reason of this framework proposal, leverage all the available data to construct a more robust classifier.

4.5.1 Classification of data integration methodologies

There are many methodologies of data integration as shown in [33]. The chosen procedure depends on the stage of integration that suits every case as described in Figure 4.4:

- Early integration. This method reside in the data concatenation of different views in a single feature space, but it does not change neither the format nor the nature of the data. There are some disadvantages of this approach, the dimensionality increases and so the performance of the similarity function decreases. It's important to mention that in order to use this approach, the data should share the same common feature space, such as text from the same domain.

- Intermediate integration. This method basically transforms all the data sources in a common feature space before combining them.

- Late integration. There are many approaches nowadays which uses this method. It resides in a separate analysis of every view, and then a combination of the results. Among the advantages we can see that it is possible to choose the best algorithm for every view based on the nature of the data.

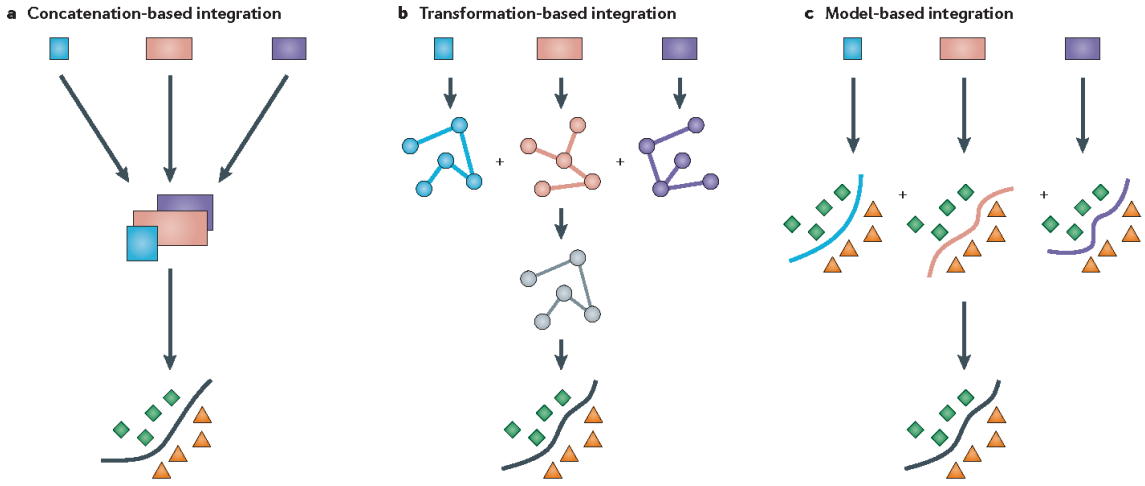

Figure 4.4: Classification of data integration methodologies [33]

4.5.2 Representative groups of multi-view learning algorithms

On the other hand, besides the mere data integration methodologies, there are three main groups of representative multi-view learning algorithms [41]: 1) co-training, 2) multiple kernel learning, and 3) subspace learning.

- Co-training. This algorithm was one of the earliest schemas for multi-view learning [4]. It is a semi-supervised learning that trains alternately two learners to maximize the mutual agreement on two different views, and lets the learners labels some unlabeled examples for each other. Its success relies on three assumptions: a) sufficiency, each view is enough for classification on its own, b) compatibility, the target function of both learners predict the same labels for co-occurring features with a high probability, and c), conditional independence, views are conditionally independent for a given label.

- Multiple kernel learning. These learning algorithms leverage kernels that naturally correspond to different views, and then merge the kernels (linearly or non-linearly) to improve the learning performance. This approach was originally designed to control the search space capacity of possible kernels to improve generalization, but it has been applied in multi-view problems lately.

- Subspace learning. The objective of these algorithms is to get a latent subspace shared by the multiple views, with the assumption that every view is generated from this latent subspace. The dimensionality of this latent subspace is lower than of any of the individual views, consequently reducing the "curse of dimensionality".

4.5.3 Co-training algorithm

After a deep analysis of the alternatives for data integration and multi-view learning algorithms we have chosen co-training to integrate the two customers' views at the financial institution. This algorithm will allow us to leverage the high amount of unlabeled data by building this semi-supervised model, and also to choose the best classifier for each view. This last statement is very important, as different data types need different types of models.

In more detail, the co-training algorithms works as described in Algorithm 3.

Data: L, U
Result: Trained classifiers: h_1, h_2
begin
  Given:
    - a set L of labeled training examples
    - a set U of unlabeled examples

  Create a pool U' of examples by choosing u examples at random from U
1 **while** $k \neq K$ **do**
    Use L to train a classifier h_1 that considers only the x_1 portion of x
    Use L to train a classifier h_2 that considers only the x_2 portion of x
    Allow h_1 to label p positive and n negative examples from U'
    Allow h_2 to label p positive and n negative examples from U'
    Add these self-labeled examples to L
    Randomly choose $2p + 2n$ examples from U to replenish U'
    k += 1
  end
end

Algorithm 3: The Co-training algorithm [4]

The co-training algorithm proposes the usage of an unlabeled set *U'* because the authors demonstrated that by using a smaller pool set the results were better, presumably because it forces the classifiers h_1 and h_1 to select examples that are more representative of the latent distribution D that generates U. Another important remark to make is that the values for p and n are chosen based on the data labels distribution.

4.5.4 Co-training enhancement

Taking the original co-training algorithm as a base, we want to propose two modifications. The first adjustment is to change the order of the operations, instead of training both classifiers and labeling the data at the same time, we propose to do it alternately, in this way, literally, the prediction from classifier h_1 helps to train the classifier h_2 within the same iteration, and then the prediction from h_2 helps to train h_1, and the cycle repeats. The second is related to the handling of sets L and U. We want to treat them separately (one per view) so labeled instances from h_1 can only go to the training instances for h_2, contrary to what happens in the original algorithm.

A detailed description of these two main modifications is presented in Algorithm 4. One necessary observation to make is that the chosen instances for L_1 and L_2 ensure that the data labels distribution is kept.

It's worth to mention that L_1 and L_2, as well as U_1 and U_2 are linked; they create the instance

Data: L_1, L_2, U_1, U_2
Result: Trained classifiers: h_1, h_2
begin
  Given:
  - a set L_1 of labeled training examples for h_1
  - a set L_2 of labeled training examples for h_2
  - a set U_1 of unlabeled examples for h_1
  - a set U_2 of unlabeled examples for h_2

1 **while** $k \neq K$ **do**
    Use L_1 to train a classifier h_1
    Allow h_1 to label 1 positive and 1 negative examples from U_1
    Add these self-labeled examples to L_2
    Remove self-labeled examples from U_1 and U_2
    Use L_2 to train a classifier h_2
    Allow h_2 to label 1 positive and 1 negative examples from U_2
    Add these self-labeled examples to L_1
    Remove self-labeled examples from U_1 and U_2
    k += 1
  end
end

Algorithm 4: The Co-training algorithm. Enhancement proposal

space $O = O_1 + O_2$, where O_1 and O_2 represents different "views" (L_1/U_1 for the text and L_2/U_2 for the transactions) of the same object (customer). Which means, every training instance o is (o_1, o_2).

In a formal way, let dataset Ω consist of L labeled records of the form $(a_1, a_2, ..., a_n, b_1, b_2, ..., b_m, l)$ where n is the number of text features, m is the number of transactions features and l is the label. Now, we train two models m_1 and m_2; m_1 on records of the form $(a_1, a_2, ..., a_n, l)$ and m_2 on records of the form $(b_1, b_2, ..., b_m, l)$. We have also a set U of unlabeled samples of the form $(c_1, c_2, ..., c_n, d_1, d_2, ..., d_m, l)$. Let our co-training enhancement proposal be defined as the process of selecting the most certain positive and negative classified unlabeled sample from U. We feed back these two new samples (one positive and one negative) labeled by m_1 into the record of the form $(b_1, b_2, ..., b_m, l)$. Now, m_2 classify two new unlabeled samples from U, and feed them back to $(a_1, a_2, ..., a_n, l)$, then m_1 is retrained and the whole process starts again. This process is repeated for K iterations.

With these new configuration, we expect to build a robust co-training schema with very few labeled data. This last assumption is crucial for the project because in the real-life application (CDD process) the amount unlabeled data greatly exceeds the labeled one.

As an additional setup, based on 4, we will test the effect of adding just the classified samples when both classifiers agree on the label.

4.6 Visual analytics

One of the main project goal is to have a human in the loop, in this case, to have domain experts in the CDD process. And, because of the nature of this process (it is not always certain and sometimes there are many possible questions to ask), the best path forward, as proposed in [27],

is to analyze where you can exploit the properties of a visual analytics design. In this way, we augment the CDD experts capabilities, rather than replace the human in the loop.

This premise above was decisive in the design of our visual analytics dashboard. We started the conceptualization with the creation of four wireframes (which can be seen in Figure 4.5 and in Appendix A) that were presented to domain experts and were modified as required, to guarantee that the wireframes are informative and consolidate the main information at a glance. These visualization tools will be a stepping stone to gain a deeper understanding of analysis requirements before formal research or models are developed. In this sense, the tool would be used very early in the CDD process in a highly exploratory way.

Recognizing that this is a first approach in the development of a much complex and robust visual analytics solution, as the one developed in [18], we created a first dashboard (Figure 4.6) where we can see an overview of the risk labels for a customer, but more importantly, we can examine the model explanations. We can see the top important features for the model and for the specific customer. These information provides valuable insight to CDD experts as they notice quickly the relation between the labels and the features in one view.

Prof.dr.ir. Jack J. van Wijk said: "The main challenge is to understand needs, strengths, and limitations of people, in such a way that we can develop powerful methods to support them in their exploration of large and complex datasets. Data science is about people, not about data". Following the latter, we will continue working together with the domain experts to improve the visualization analytics tools as required so they can perform a better job.

The design is focused on the tasks, as its well known that no visual representation supports all tasks. Validating the effectiveness of a design is both necessary and difficult, but considering our particular task and the intended user, we will focus in a simple and intuitive layout which can showcase at a glance the customer risk labels and, most importantly, the main model explanations with relevant thresholds where needed. The user will be able to search for specific customers, and the tool will retrieve the risks labels and the top features (for transactions and for text, from the model and describing the specific customer).

On a more technical side, for the proof of concept, the data will be stored in a Structured Query Language (SQL) Database and will be retrieved from a Hypertext Markup Language (HTML) website using SQL queries. This architecture was chosen for the simplicity and because it is fast to implement, ideal for the proof of concept.

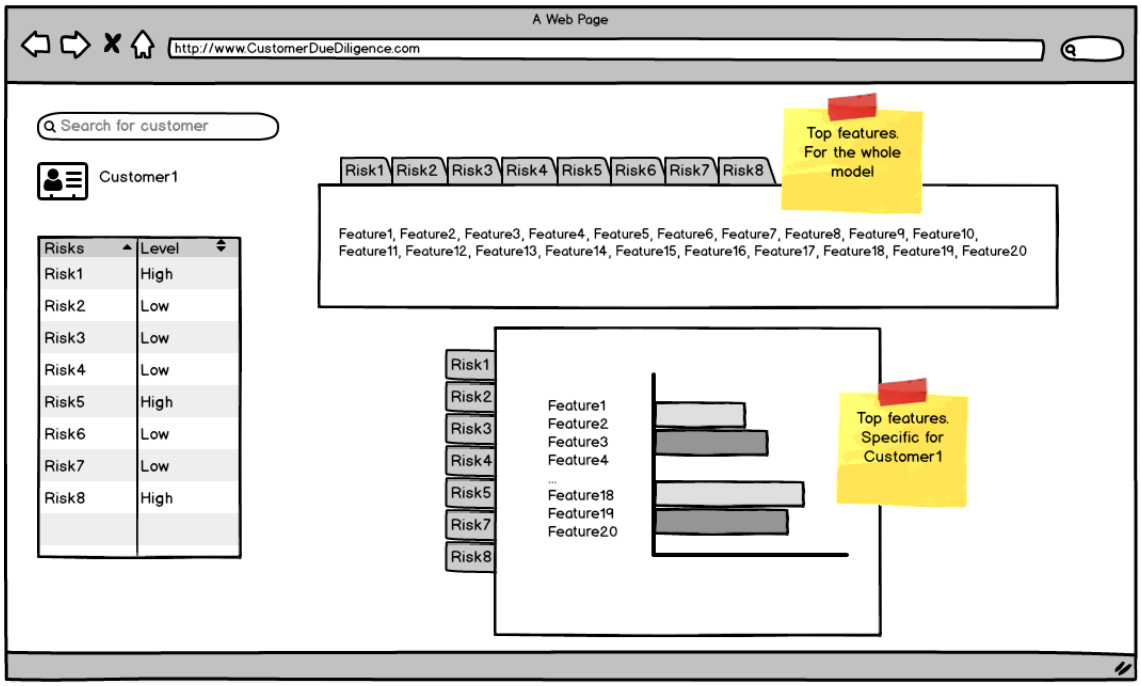

Figure 4.5: Wireframe 1. CDD risks overview for a customer

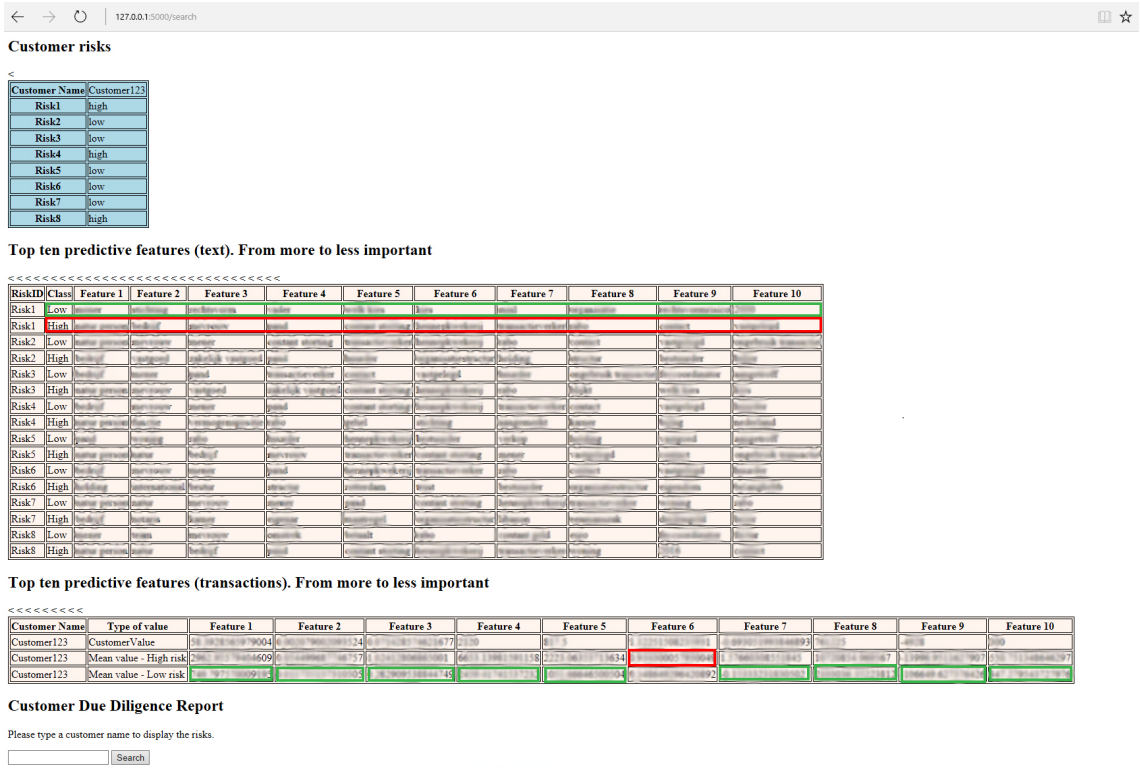

Figure 4.6: Main visual analytics dashboard (from wireframe 1)

Chapter 5

Experimental evaluation

In this section we try to answer all the research questions described in Chapter 3. On one side, we demonstrate that a co-training schema is very useful to increase the insight provided to CDD experts while keeping a competent classification score. On the other side, we present different classification setups where we can observe the two mentioned trade-offs: between number of iterations, initial size of training set and the performance score; and between interpretability (model explanations) and accuracy.

We conducted two types of experiments: the technical and the operational experiments(which were focused on the application of the framework in the banking industry).

Along the set of experiments, we tested the hypothesis on [40] which states that "the co-training process could not improve the performance further after a number of rounds", but we could not get conclusive results in this matter, highly likely due to the low amount of available data.

5.1 Data

Our dataset consist of 693 instances, each having the two views: text and transactions features. It's worth to mention the process to get the data in this format was very long and complex. We started with raw text documents and transactions in excel files, but at the end we were able to extract text features using TF-IDF, and 113 features from the transactions. Another relevant aspect to comment is that the text is in Dutch language, so sometimes we validated the relevance of words with domain experts.

The success in the dataset creation is due to an IR system with a very high precision. We wanted to make sure that all the retrieved documents were relevant and related to the target customers. As mentioned earlier, we needed to understand the business rules and gain domain expertise to determine which were the relevant documents.

Out of an universe of $\approx$ 35 thousands files, we selected 693 text documents and 2565 transactions files to construct the dataset. The amount of transactions files is higher because there are customers with several accounts or having other circumstances which make them have various transactions files.

5.2 Experiment design

In a co-training schema, there is the assumption that we only have a few labeled instances. In the original co-training paper [4], the authors used 12 initial labeled instances, and from there the training data is augmented by the new predicted instances. In our case, as we consider our data

more complex than in the experiments of the original co-training paper, we selected three different initial training size: with 12, 24 and 50 instances.

Another factor that we varied was the number of iterations. The experiment was done with 10, 20, 30, 40, 50, 60, 70, 80 and 100 iterations. These were chosen to understand what is the best trade-off between the number of initial training instances, the iterations and the classification performance. In the original co-training experiment, the authors tested with 30 iterations.

As the labels distribution in the dataset is $\frac{1}{3}$ for low risk and $\frac{2}{3}$ for high risk, in all the experiments, the initial training size has this same distribution. Furthermore, even though the instances are shuffled every time, the seed used to shuffle the index is the same for all the experiments in each repetition. Having this last configuration allows us to compare experiments in every repetition and also to compare the overall performance by averaging the scores.

Accuracy is an important metric in the experiments, but due to the class labels distribution, we are also interested in the ROC score because it provides the connection between the sensitivity (true-positive rate) and the specificity (false-positive rate).

The framework was programmed in Python. The library to transform text into a vector space representation was *TfidfVectorizer* from scikit learn, in which we used "uni-grams", "bi-grams" and "three-grams". We also just considers the words presented in at least 10% of the documents and as much as 90% of them. This configuration gave the best scores and the best representation of top features (words), while removing the template words.

In the case of the classification models, we used Multinomial Naive Bayes (Python library scikit learn) for text, with an alpha of 0.001 (smoothing parameter). In the case of the transactions, we used Random Forest (Python library scikit learn) with 500 trees and with a max tree depth of 500. In order to tune the hyper-parameters of the classifiers we used a grid search.

The scores for accuracy and ROC will be presented in the next section. The combined score from both classifiers was calculated with a sum-rule. Interesting theoretical results have been inferred for simple combination schemas, such as sum-rule. For instance, in [17], the authors show that sum-rule is less sensitive to noise than other rules. Although this rule is simple, it has a high recognition rate, and it has been demonstrated that more complex rules are not by default superior to simple ones, such as sum-rule [39].

The configurations to be tested, as mentioned in section 3.1, are described below:

1. Two separate classifiers with a simple score combination. We used the classifiers MNB for text and Random Forest for transactions as a baseline for the experiment. These classifiers do not use iterations and do not interact between each other (named as "Single classifier" in Figures 5.1 and 5.2 and subsequent experiment results).

2. Classification using a simple concatenation of text and transactions features. In this configuration we used the classifier Random Forest due to its properties (flexible, simple to implement, avoid overfitting, among others).

3. Classifiers self-training ("co-training" without multi-view). Each classifier is being retrained using its predicted samples, similar to a co-training schema but the classifiers do not interact, they are retrained in isolation (named as "Self-training" in Figures 5.1 and 5.2 and subsequent experiment results).

4. Original co-training algorithm. We used the operation described in Algorithm 3 (named as "Co-training original" in Figures 5.1 and 5.2 and subsequent experiment results).

5. Our proposed co-training algorithm enhanced. We used the operation described in Algorithm 4 (named as "Co-training proposal 1" in Figures 5.1 and 5.2 and subsequent experiment results).

6. Our proposed co-training algorithm enhanced (only when both classifiers agree). We used the operation described in Algorithm 4, but we just added the new labeled instances only when both classifiers predict the same label (named as "Co-training proposal 2" in Figures 5.1 and 5.2 and subsequent experiment results).

5.3 Results

The results were satisfactory, we can show empirical evidence that our co-training proposal outperform the other experiments (but the mere features concatenation), and also that the insight provided, as a combination with text and features transactions, it is very relevant and useful to the work of the CDD experts.

We surprisingly found that the mere concatenation of text and transactions features had a very good performance, always with the best accuracy and ROC scores. The drawback of this setup is that we are not able to extract meaningful insights,: the model looses its interpretability and none of the transactions features (which are very crucial to the CDD experts) are highlighted as relevant for the model. Here it's possible to see the trade-off between interpretability (model explanation) and accuracy. The classification of concatenated features (text and transactions) has a high accuracy but low interpretability, but the outcome of our co-training proposal has a high interpretability but lower accuracy score.

The visual analytics tool was very useful to the CDD expert. We conducted an interview with the domain expert where we showcased the tool. This helped us to gain important input and feedback related to the tool. The empirical evidence confirm that our statement is correct, a CDD expert can carry out their tasks more effectively when you present relevant and meaningful insight (for this specific task) at a glance.

Another important insight was the discovery of correlation between text and transactions features. We found a moderate correlation (using the Pearson correlation coefficient) between some text and transactions features, which can guide the initial inquiries of CDD experts in a new case that lack text data.

5.3.1 Technical analysis

As it can be seen in Figures 5.1 and 5.2, with 24 initial training instances, our co-training proposal outperformed the single-view learning schema, the self-training approach, as well as the original co-training algorithm. The changes done were significant enough to provide a better results while increasing as well the interpretability.

In the other configurations, with 12 and 50 initial training instances, in most of the cases, our proposal also outperformed the other experimental configurations (described in section 5.2), but we chose the configuration with 24 as it has the most stable and consistent results (the other results can be found in Appendix B).

We can clearly see how co-training boosted the performance of the transactions classifier, because this classifier by itself was not providing any important predictions. With only a few training instances it had a very poor performance. Experiments showed that this model with 24 instances (as well as with 12 and 50), has a very high variance, which is solved with more training

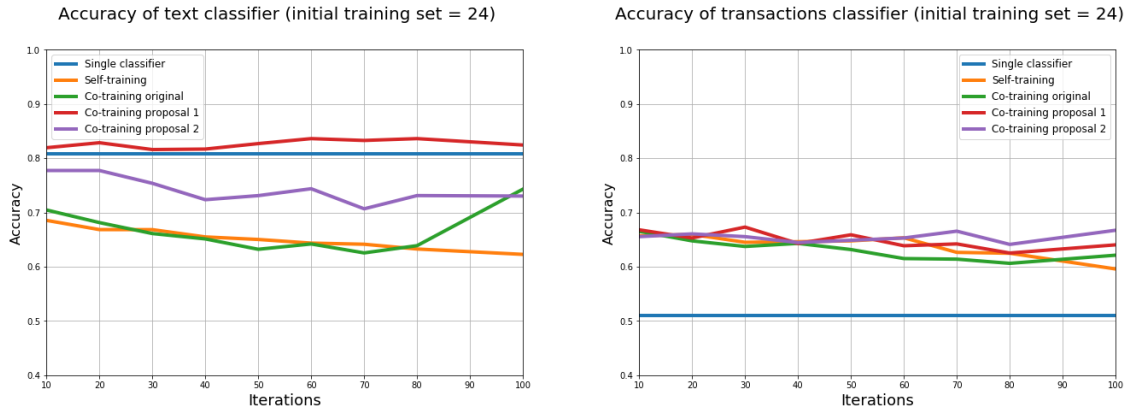

Figure 5.1: Accuracy for the classification with 24 initial training instances

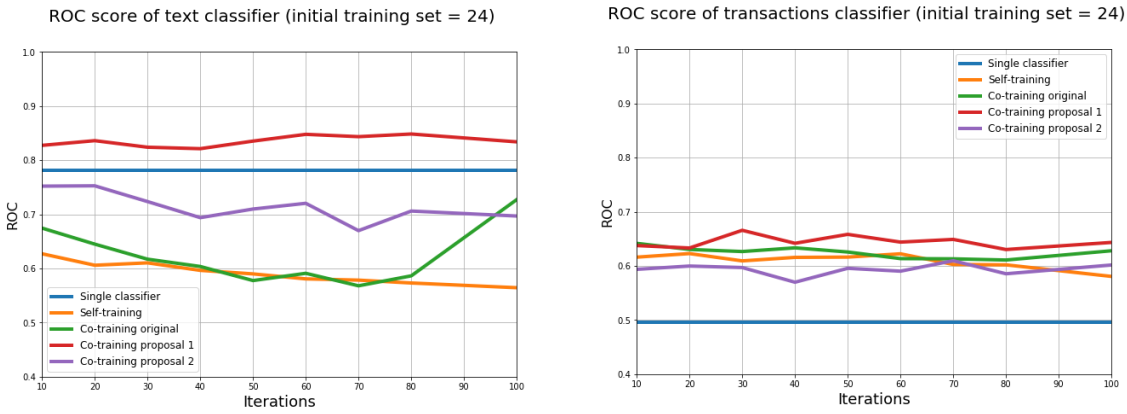

Figure 5.2: Receiver Operating Characteristic (ROC) score for the classification with 24 initial training instances

examples. In some cases we can also notice that the performance of the text classifier does not improve drastically, this is due to the poor initial performance of the transactions classifier.

While comparing the insight obtained from single separated classifiers and the co-training schemas, it was very obvious the co-training schemas provided much more relevant model explanations. For instance, in the case of the single transactions classifier, the top important features were different and not meaningful every time the experiment was executed, something that did not happen when it was done under a co-training schema, specially using our co-training proposal (Algorithm 4).

One interesting phenomenon was discovered during the experimental analysis. Sometimes the performance (accuracy and ROC scores) of the self-training configuration was better than the original co-training algorithm, as it can be seen in Figures 5.1 and 5.2. This is unexpected as a multi-view schema is supposed to give a larger understanding of the task, so a better performance is expected [34]. One possible explanation is that (in the presence of just a few initial training instances) the performance of the transactions classifier has a negative initial impact in the overall classification scores. We need to study deeper this behaviour to reach a definitive conclusion.

Another very important point to notice is that we did not have a drop in performance during the experiment. This performance drop is theoretically explained as follows: "As the Co-Training

Process proceeds, more and more unlabeled data are labeled for the learners each other, which makes the difference between the two learners become smaller and smaller. Thus, after a number of learning rounds, the Co-Training Process could not improve the performance further" [40]. Our hypothesis is that we did not have enough data and the number of iterations did not reach that threshold where the classification starts to deteriorate, but further research needs to be done.

The highest scores were reached with a simple concatenation of features (text and transactions). We reached an accuracy of 91.2% (with ROC score of 91.9%), 95.7% (with ROC score of 96.5%) and 97.3% (with ROC score of 97.5%) respectively for 12, 24 and 50 initial training instances. This is very important to consider because if we would only be interested in what was predicted, then this configuration would be the best, but as we are interested mainly in why the prediction was done (model explanations), then this configuration is not the best as none of the 50 (or even more) top model features is a transactions feature. And as we discovered during the project, transactions features play a very important role in the analysis of the CDD experts.

We found a moderate correlation between some text and transactions features as it can be seen in Figure 5.3. We used the top 50 important features (using a co-training schema), the first 40 were text features and the last 10 were transactions features. We can clearly see a very strong correlation among text features and among transactions features, but we are interested in the correlation of text features with transactions features. By checking with a domain expert, we can confirm that these correlated features are relevant and interesting (text with transactions).

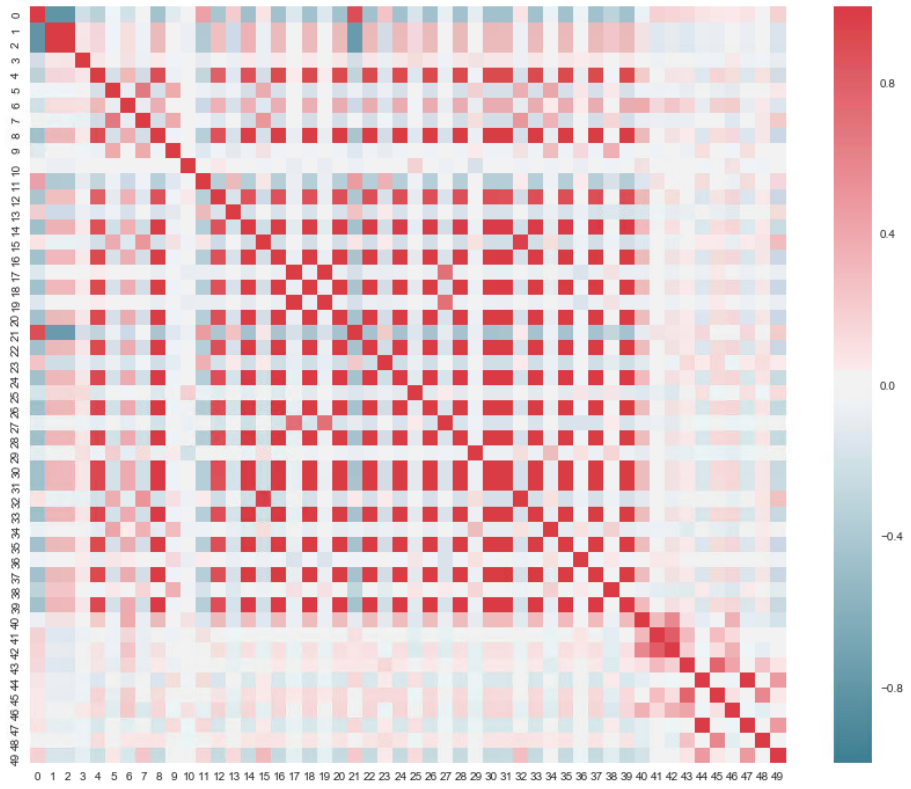

Figure 5.3: Text and transactions features correlation (Pearson)

As an outcome of previous analysis, when a CDD expert has a new customer research case, and this target customer is described by those top transactions features for a high risk customer, then CDD expert could start the inquiries by looking at those correlated text features. This is

very useful when there is just transactions data and the domain experts are looking for some hints to start the analysis. This will be triggered and validated when the Pearson correlation coefficient is above 0.3 (moderate), but further experiments and analysis with more data needs to be done to confirm these first results.

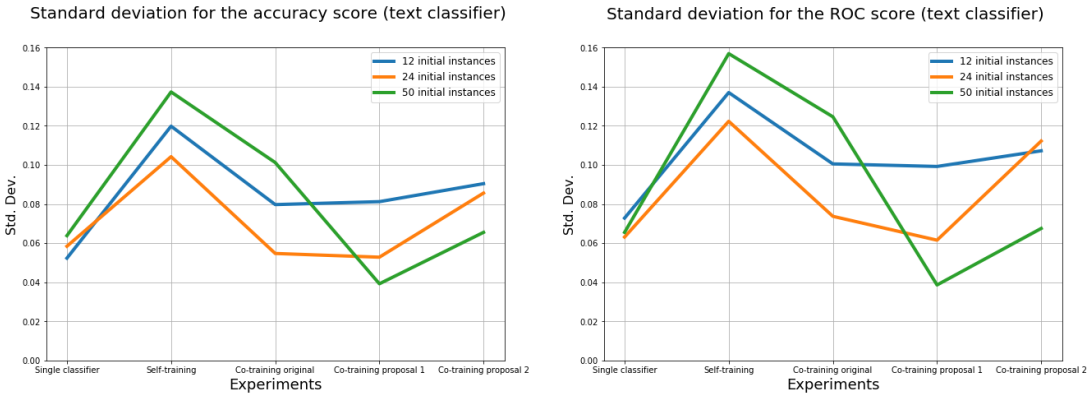

Figure 5.4: Standard deviation (σ) for accuracy and ROC scores of text classifier

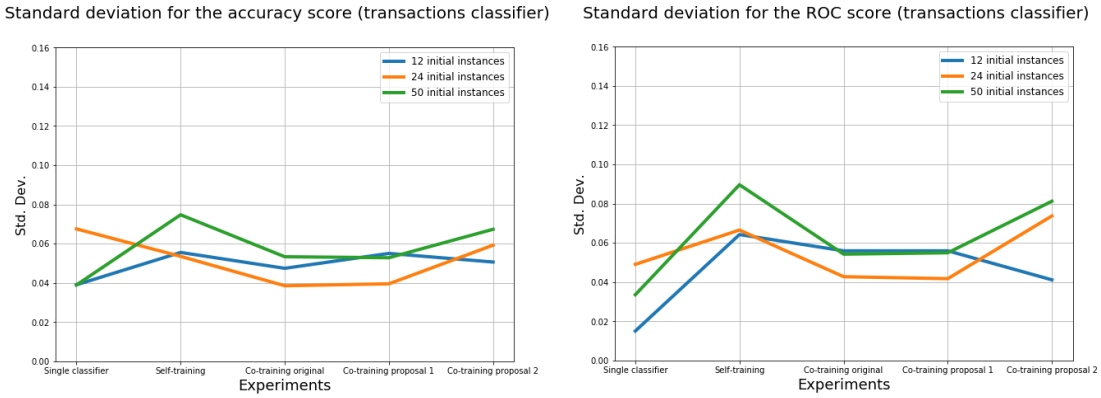

Figure 5.5: Standard deviation (σ) for accuracy and ROC scores of transactions classifier

For the overall CDD report, we decided to consider just three risks: one predicted using the proposed framework, but the other two were predicted using just text data as the information extracted from the transactions was not relevant for these two risks. In the case of the risks classified with just text, the distribution was very imbalanced with 85% of the instances labeled as low risk and 15% as high risk; for this risk the accuracy was 89% and the ROC score was 79%. In the case of the other risk, the distribution was 93% labeled as low risk and only 7% as high; the accuracy was 74% and the ROC score 58%. All of the other risks were not considered because the risk labels distribution was extremely imbalanced, in some cases some labels had less than 1% of the instances.

While looking for coherence in the results, we measured the standard deviation (σ) to quantify the degree of dispersion during the repetition of the experiments. We found interesting patterns: overall speaking, our co-training proposal 1 was the second most consistent experiment, the variation among repetitions was the second lowest just after the single classifier configuration. This

	Text classifier		Transactions classifier		Combined classifier	
Configuration	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**
Single classifier	0.8088	0.7816	0.5104	0.4966	0.8093	0.7823
Self-training	0.6684	0.6104	0.6440	0.6134	0.6606	0.6017
Co-training original	0.6482	0.6007	0.6254	0.6147	0.6456	0.5986
Co-training proposal 1	0.8143	0.8172	0.6563	0.6439	0.8101	0.8129
Co-training proposal 2	0.7412	0.7195	0.6664	0.6084	0.7445	0.7257

Figure 5.6: Accuracy and ROC scores for the experiments with the following setup: 24 initial training instances and 30 iterations

can be seen in Figures 5.4 and 5.5.

Also, the experiments showed that the most consistent results were found with 24 initial instances. Additionally, the best improvement in performance (accuracy and ROC scores) was also with 24 initial instances, so we decided to explore further this setup in combination with 30 iterations. The number of iterations was chosen due to its good overall behaviour plus it is the number of iterations chosen as baseline in the original co-training paper.

In Figure 5.6 we can observe that our co-training proposals gave the best results in all the experiments with initial training size of 24, and with 30 iterations (the average of 10 repetition is recorded). This was just to showcase the usefulness of our proposal with this particular setup, but as seen in Figures 5.1 and 5.2, this happened in most of the experiments, our co-training proposals outperformed the other configurations.

5.3.2 Operational analysis

In this section, we conducted two experiments and one interview with a CDD expert. The results were very good, a senior CDD expert validated the predictions and the model explanations. This domain expert was enthusiastic about the results, as the insight provided was very useful for their actual research analysis. The expert was very interested for the framework implementation in production, and a few uses cases were discussed.

Analysis 1: Transactions feature extraction process

This experiment is related to the research question number 2: "Is the transactions feature extraction process (number and quality of features) good enough to provide good classification scores and relevant insight to CDD experts?".

To conduct this experiment, we interviewed one CDD expert. We randomly selected 10 transactions instances (customer cases) to evaluate. The prediction (and the top most important features) of our transactions classification algorithm was compared with the domain expert analysis for each one of these 10 instances, and the results are summarized below:

The accuracy of the algorithm was 70%; it predicted 6 instances as "high" risk and 4 as "low" risk. On the feature importance side, we highlighted the top 10 model important features and, the mean value of this features for "low" risk and "high" risk cases. We then showed the value of this features for each of the 10 predicted instances so we can compare with the model values.

The top important features were relevant to the CDD expert, but there were other characteristics identified on the CDD expert analysis which are not captured by the current feature extraction process, so these will be added as a future work. These new features are related to the customer type, their main activity sector, and other characteristics that cannot be disclosed due to confidentiality reasons.

On a qualitative analysis, the CDD expert was excited about the results and is looking forward for the feedback incorporation and a future framework implementation in their day to day activities.

Analysis 2: Multi-view co-training classification

This experiment is related to the research question 3: "Does a multi-view co-training classification algorithm improve the scores and insight provided to CDD experts?".

During this experiment, we analyzed in detail one full customer case (with its corresponding set of text and transactions files). We wanted to compare the CDD expert analysis with the prediction and insight (top important features) provided by our framework. To analyze this case, we predicted three risks, one obtained with co-training and the other two just with text data. It's important also to mention that the trigger for this CDD report was an external source.

Text classification analysis. The framework accuracy for the text part was 66%, out of three risks two were correctly classified. On the other hand, the CDD expert mentioned that most of the text top features (words) are not useful, but there are a few that are very important and informative. Another finding was that many top features (words) are repeated among risks. This is because many these risks share the same labels, in many instances. This needs to be analyzed in the future work.

Risk	Ground truth	Model prediction
Risk1 (only text data)	**Low**	**Low**
Risk2 (only text data)	**Low**	**High**
Risk3 (co-training)	**High**	**High**

Figure 5.7: Text classification results

Transactions classification analysis. The framework performed very well in this section. The prediction was correct, and four out of the top five transactions features were very informative to the CDD expert. On table 2, we can see that for this customer, 3 features were above the average value for "high" risk classification, one was near this threshold and just one was not relevant for this classification.

Top important features (TXN)	Relevance	Comments
Feature1	★★★	Above the "high" threshold
Feature2		Not informative
Feature3	★★★★★	Significantly above the "high" threshold
Feature4	★	Near the "high" threshold
Feature5	★★★★★	Significantly above the "high" threshold

Figure 5.8: Top important transactions features

As a very important remark, the CDD expert validated and confirmed that the top transactions features were crucial and very relevant for this analysis, although there was another decisive feature identified during the analysis which is not capture at the moment by the framework, but which will be incorporated in the future work.

Conclusion of analysis 2. As this CDD report was trigger by an external source, it's highly likely we could have identified this issue before with our framework, using just the customer

transactions. It's also worth to mention that the correlation between transactions and text features can lead to have an initial inquiry starting point, as we may focus, at the beginning, in those features (words). It will potentially increase the efficiency and effectiveness of the CDD process.

As an outcome, it would be important to know what is the minimum number of transactions or time window of transactions necessary to capture the transactions features.

Overall speaking, the framework demonstrates its validity and its importance for one of the compliance business goals: protecting and shielding the financial system.

Interview: Visual analytics

The purpose of this questionnaire is to gather feedback from a senior CDD expert to validate the current visual analytics tool (Figures 4.6 and A.4). We also want to gain insight and recommendations to be implemented in the next development cycle.

1. How easy is our visual analytics tool to use and understand?
    - Easy
    - Reasons: The vocabulary used is self-explanatory, there is no difficult "transcription" needed to understand the tool. The information and the design is explicit and simple.
2. What components of our visualization proposal are most important to you? And why?
    - The top ten predictive features:
        - Text. It's very time efficient to be able to observe and analyze the top features per risk at the same time, we do not need to check the whole CDD report to find important characteristics.
        - Transactions. It saves a lot of time; besides the fact that we have the main features, we also have the mean values per risk label and for specific customers. This information is very useful as valid arguments for the whole CDD investigation, it gives more credibility.
3. What components of our visualization proposal are less important to you? And why?
    - The table with the customer risk's labels because it already exists.
4. What is the most important component you think we should add? And why?
    - It would be very interesting to add more text from different sources.
    - Also, if we can add a view where we can find all the customers that share some specific top high-risk features. It would be similar to the table where we see all the customer that share the same risks, but in this case, for the top features.
    - We could change the colors of the table, to make an easier to visualize the low and the high features (i.e. green for low, red for high).
    - We can add more sources of data, we can integrate internal and external sources (i.e. news).
5. How often would you use the tool?
    - Everyday.
6. What are the most frequent tasks you will do when using our tool?
    - While doing a research for a CDD report, there are many tasks done manually at this point. This visual analytics tool can save time and can help us focusing on high-risk customers. The prediction and the top important features will allow us to target those customers that share these traits.

7. How and how much do you feel the current classification framework and visual analytics tool can help you with your work?

    - It will definitely be very useful, it will save us a lot of time and effort so we can focus on those tasks that add more value to the business. Further, it will not only provide relevant predictions but also insight, we will learn the reason behind the risks and increase our domain knowledge for future cases.

5.4 Case study

The purpose of this section is to present the applicability and usage of this framework in a financial institution.

The experimental results certainly demonstrate that the framework can be used also in the following ways:

- The framework as a whole can be implemented to focus on just the top risky customers, by giving to the framework as input a bigger amount of customer data, a more representative dataset of the financial institution customers. This, added to a visual analytics tool can speed up and boost the CDD team efficiency.
- The feature extraction process for the transactions can be used now to analyze automatically the current transactions. The outcome could be reflected also in new rules for the anti-money laundering team or in other system alerts.
- The features correlation can be used to find initial inquiry leads when a new CDD research starts.

The results prove that the CDD team is moving in the right direction towards the implementation of new tools to leverage all the current available customer data. This framework is the first step to create a machine learning ecosystem in which CDD experts play an important role. The multi-view framework can be seen as a tool to support and help the domain experts while doing the CDD report.

On the other hand, there are a few points to improve. Even though most of the files follow an specific naming convention, it's important to ensure and strengthen a more strict naming convention, so the analysis of the files can be done easier. But in general the future looks bright for the whole CDD team.

5.5 Summary

The experimental analysis allows us to show empirical evidence that a multi-view learning scenario (text and transactions data) with a co-training schema outperforms a single-view learning scenario for a risk classification task in a CDD report. We demonstrated that our co-training proposal provides better performance (accuracy and ROC scores) while also improving the interpretability (model explanations). We also developed a proof of concept for a visual analytics tool that augment domain experts capabilities and include them in the loop, in the CDD process.

To summarize, based on the research questions, our findings are described below.

1. Is the correlation between multi-view (text and transactions) features significant enough to make the assumption that one may imply the other?

- We found a moderate (> 0.3 Pearson's correlation coefficient) correlation among some text and transactions features, this can be observed in Figure 5.3.
- While inspecting these features deeper with a CDD expert, we found that these correlated features are interesting and this provided useful insight to the domain expert. This knowledge could be used to have some leads when starting inquiries in a new customer case (mainly when only transactions data is available at the moment). But further experiments needs to be done in order to confirm the findings and propose a new business process in the banking industry.

2. Is the transactions feature extraction process (number and quality of features) good enough to provide good classification scores and relevant insight to CDD experts?

   - In the presence of just a few initial training instances (12, 24 and 50 in the experiments), it was demonstrated that under a co-training learning schema we can have relevant performance scores (as seen in Figures 5.1 and 5.2), but also meaningful model explanations (demonstrated in Subsection 5.3.2) that produce new knowledge to CDD experts.

3. Does a multi-view co-training classification algorithm improve the scores and insight provided to CDD experts? Is it better and more relevant than showing separate classifiers' results?

   - In Subsection 5.3.2, we demonstrated that a multi-view co-training schema provides better performance and insight to CDD experts. While the simple concatenation of text and transactions features gives much better performance, a co-training schema have a trade-off: the classification scores are lower but the interpretability is much higher than the simple concatenation.

4. We want to analyze two trade-offs. The first related to the number of co-training iterations, the number of initial training instances and the classification performance as explained in [40]. The second one related to the compromise between interpretability (model explanations) and accuracy.

   - Throughout the section 5.3.1, we found empirical evidence that a setup with 24 initial training instances with 30 iterations has better performance, and has also more consistent results (the standard deviation is the lowest for the 10 repetitions, shown in Figures 5.4 and 5.5).
   - It was also demonstrated, that co-training has a good trade-off between interpretability and performance. While the classification scores are better than the single-view classifiers (baseline), we have good model explanations. Even though the performance of the simple concatenation model is much better, the interpretability is much higher when using a co-training schema.

5. Following best practices in design, what is the best approach and layout for a useful visual analytics setup?

   - One of the main thesis objective is to have humans (domain experts) in the CDD loop. This is achieved by augmenting the capabilities of the experts with the development of visual analytics tools.
   - Our proof of concept presented in Section 4.6 demonstrated good results. In Subsection 5.3.2, we validated the utility of the tool by interviewing a senior CDD expert. The expert confirmed that, not only the provided information was useful, but also the format was easy and intuitive to follow.
   - This is a good first step, but further improvement is needed. The target is to have an interactive learning approach as the one presented in [18].

Chapter 6

Conclusions

This thesis provided a novel framework for multi-view predictive visual analytics in three ways: we improved the original co-training algorithm, we applied a co-training learning schema with a new input data paradigm (text and transactions), and we provided a new approach to tackle a challenging and relevant problem, the CDD report.

The results were very important and significant, but further research needs to be conducted in order to improve the current framework. But, we can foreseen many opportunity areas in the future.

6.1 Contribution

The contribution can be divided in two parts: the academic and the application side.

On the academic part we contributed in the following:

- We created a transactions features extraction process (in Section 4.4.2), which gave meaningful and relevant insight to the CDD domain experts.
- We presented a performance improvement of the original co-training algorithm introduced in [4]. The improvement was based in the operation, and by avoiding any kind of self-learning (none of the labeled samples for a classifier is feeding the training set of the same classifier). Our proposal outperforms other co-training configuration, and, in most of the cases, also the baseline (single classifiers). In other cases this baseline was not reached due to the transactions classifier's poor performance in the first iterations, which in the long run affects the text performance. So we need to improve the transactions classification model to leverage the whole framework in a more effective and efficient way.
- We presented a semi-supervised multi-view framework which is able to have a good interpretability and accuracy trade-off while learning just with a few initial labeled instances.

On the application side, our contribution was:

- The framework boosted the knowledge available for CDD experts, and put them in a new knowledge discovery process, where they are a key element. The framework will help the CDD team to focus and allocate the resources on the most risky customers based on the prediction and model explanation from our framework.
- The visual analytics tools will boost and make more efficient their time and effort while working on different CDD cases.

- The features correlation can give some initial inquiries leads when an investigation is starting. We could find a very strong correlation with some text features based on discovered transactions patterns.
- A new set of rules and alerts can be implemented using the outcome from the transactions feature extraction process, this can improve the effectiveness of the compliance and other teams which goal is to protect the financial system against illegal activities.

6.2 Limitations

At this point, there are some improvements that need to be done in order to increase the performance of the framework. The main ones are in the domain of the data gathering and features extraction. We need to gather more data and make sure we are able to make a distinction between the files strictly related to our target customer, and the files related to entities which are linked to the target customers. We can potentially find new and interesting insight in these relationships.

The domain and nature of the current text data is not the ideal at the moment, we would like to have some text documents that were generated before the CDD report even starts. This new data would be more relevant and it could potentially save a lot of effort and speed up the risky-customer identification.

Also, at this point, the chosen text classifier is a good start, but it has some limitations. So in the future we want to implement the improvements to the MNB classifier presented in [32]. We also want to test different state-of-the-art text classifiers and choose the one that has a better trade-off between interpretability and classification scores (accuracy and ROC scores).

Due to the small size of the dataset, the accuracy cannot be extrapolated when the dataset size is in the millions. Further research and experiments would need to be done to confirm and validate the current framework behavior.

6.3 Future work

One of the main point to focus on is in the improvement of the text classification model. We will analyze different state-of-the-art text classifiers and choose the best one for this particular scenario. We will also improve the transactions feature extraction process based on the feedback provided by the CDD expert, and other characteristics discovered during the experimental analysis. Furthermore, as mentioned before, we will add a new time series primitive called *time series Shapelets* (discussed in [42]) in this process, as very good results have been achieved in the past.

We can also go deeper into the transactions analysis, and analyze the text available on it. We could potentially make categories and "topics" out of the text analyzed in different fields.

Another very promising path is to include interactive learning, and integrate even more the CDD experts into the knowledge discovery process. We want to implement a similar solution as in [18]. This will allow us to improve in parallel, the analysis done by the domain experts and the performance of the framework. Also, the feedback provided by the CDD experts will be incorporated, for example, we will create a dashboard that will show the customers that share the same top features.

Further, as recommended by the CDD expert during the interview, we can integrate other internal and external sources of data (i.e. news), which can boost and create, potentially, different customer views. Another additional improvement would be to keep a document identifier so we can also trace back not only the top features, but also the documents that contain them (i.e. emails, reports, questionnaires, notes, etc.).

In the direction of the co-training schema improvement, we will also analyzed the integration of co-training with a canonical correlation analysis (CCA) as presented in [38]. The author claims that with this new addition, the co-training algorithm is more robust.

In a second and later phase, we would like to do network analysis and construct a graph with the transactions information. The nodes will be represented by customers and the links by transfers (incoming and outgoing). Each node will have a risk score based on the number and types of risks the customer has, and a graph clustering approach will be researched to see how close is a customer of interested from known risks entities (or any other risky customer) in the network. This can give us insight to see if the customers are naturally group by risks or there is no relation. A threshold of links can be determined to measure the distance of a customer to a risky entity.

Bibliography

[1] Alexandre, C., and Balsa, J. Client profiling for an anti-money laundering system. *CoRR abs/1510.00878* (2015).

[2] Allahyari, M., Pouriyeh, S. A., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques. *CoRR abs/1707.02919* (2017).

[3] Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval: The Concepts and Technology Behind Search*, 2nd ed. Addison-Wesley Publishing Company, USA, 2008.

[4] Blum, A., and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (New York, NY, USA, 1998), COLT' 98, ACM, pp. 92–100.

[5] Breiman, L. Random forests. *Machine Learning 45*, 1 (Oct 2001), 5–32.

[6] Chakaravarthy, V. T., Gupta, H., Roy, P., and Mohania, M. Efficiently linking text documents with relevant structured information. In *Proceedings of the 32Nd International Conference on Very Large Data Bases* (2006), VLDB '06, VLDB Endowment, pp. 667–678.

[7] De Nederlandsche Bank. DNB guidance on the anti-money laundering and counter-terrorist financing act and the sanctions act, Preventing the misuse of the financial system for money laundering and terrorist financing purposes and controlling integrity risks, 2015. `http://www.toezicht.dnb.nl/en/binaries/51-212353.pdf`, Last accessed on 2018-05-28.

[8] Gao, Z., and Ye, M. A framework for data mining-based anti-money laundering research. *Journal of Money Laundering Control 10*, 2 (2007), 170–179.

[9] Geurts, P. Pattern extraction for time series classification. In *Principles of Data Mining and Knowledge Discovery* (Berlin, Heidelberg, 2001), L. De Raedt and A. Siebes, Eds., Springer Berlin Heidelberg, pp. 115–127.

[10] Hansen, L. K., and Salamon, P. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell. 12*, 10 (Oct. 1990), 993–1001.

[11] Hubbard, R. G. Financial regulatory reform: a progress report. *Review*, May (2013), 181–198.

[12] Infosys Limited. Regulatory compliance management in banks: Challenges and complexities, 2017. `https://www.infosys.com/industries/financial-services/Documents/regulatory-compliance-management.pdf`, Last accessed on 2018-05-28.

[13] Internet Live Stats. Twitter usage statistics. `http://www.internetlivestats.com/twitter-statistics/`, Last accessed on 2018-05-13.

[14] Isa, Y. M., Sanusi, Z. M., Haniff, M. N., and Barnes, P. A. Money laundering risk: From the bankers' and regulators perspectives. *Procedia Economics and Finance 28* (2015), 7 – 13. 7th INTERNATIONAL CONFERENCE ON FINANCIAL CRIMINOLOGY 2015, 7th ICFC 2015, 13-14 April 2015,Wadham College, Oxford University, United Kingdom.

[15] Kahneman, D. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York, 2011.

[16] Khan, N. S., Larik, A. S., Rajput, Q., and Haider, S. A bayesian approach for suspicious financial activity reporting. *International Journal of Computers and Applications 35*, 4 (2013), 181–187.

[17] Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 3 (Mar 1998), 226–239.

[18] Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2015), IUI '15, ACM, pp. 126–137.

[19] Lanzarini, L. C., Monte, A. V., Fernandez Bariviera, A., and Santana, P. J. Simplifying credit scoring rules using lvq+pso. Papers, arXiv.org, 2017.

[20] Larik, A. S., and Haider, S. Clustering based anomalous transaction reporting. *Procedia Computer Science 3* (2011), 606 – 610. World Conference on Information Technology.

[21] Li, S., Li, Y., and Fu, Y. Multi-view time series classification: A discriminative bilinear projection approach. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (New York, NY, USA, 2016), CIKM '16, ACM, pp. 989–998.

[22] Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA, 2008.

[23] McCallum, A., and Nigam, K. A comparison of event models for naive Bayes text classification. In *Learning for Text Categorization: Papers from the 1998 AAAI Workshop* (1998), pp. 41–48.

[24] Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., and Delen, D. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Elsevier Science, 2012.

[25] Molnar, C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* Bookdown, USA, 2018.

[26] Mortezapour, R., and Afzali, M. Assessment of customer credit through combined clustering of artificial neural networks, genetics algorithm and bayesian probabilities. *CoRR abs/1312.7740* (2013).

[27] Munzner, T., and Maguire, E. *Visualization analysis and design.* AK Peters visualization series. CRC Press, Boca Raton, FL, 2015.

[28] Patidar, V., Singh, D., and Singh, A. A novel technique of email classification for spam detection. *International Journal of Applied Information Systems 5*, 10 (August 2013), 15–19. Published by Foundation of Computer Science, New York, USA.

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

[30] PricewaterhouseCoopers. Know your customer: Quick reference guide, 2014. https://www.pwc.com/gx/en/financial-services/publications/assets/pwc-anti-money-laundering-know-your-customer-quick-reference-guide.pdf, Last accessed on 2018-05-28.

[31] Ratcliff, J. W., and Metzener, D. Pattern matching: The gestalt approach. *Dr. Dobb's Journal* (1988), 46.

[32] Rennie, J. D. M., Shih, L., Teevan, J., and Karger, D. R. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning* (2003), ICML'03, AAAI Press, pp. 616–623.

[33] Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics 16* (2015), 85–97.

[34] Rüping, S., and Scheffer, T. Learning with multiple views. In *Proceedings of ICML Workshop on Learning with Multiple Views* (2005).

[35] Savage, D., Wang, Q., Chou, P. L., Zhang, X., and Yu, X. Detection of money laundering groups using supervised learning in networks. *CoRR abs/1608.00708* (2016).

[36] scikit-learn, Machine Learning in Python. Ensemble methods. http://scikit-learn.org/stable/modules/ensemble.html, Last accessed on 2018-06-11.

[37] Sugimura, H., and Matsumoto, K. Classification system for time series data based on feature pattern extraction. In *2011 IEEE International Conference on Systems, Man, and Cybernetics* (Oct 2011), pp. 1340–1345.

[38] Sun, S., and Jin, F. Robust co-training. *IJPRAI 25* (2011), 1113–1126.

[39] Tulyakov, S., Jaeger, S., Govindaraju, V., and Doermann, D. S. Review of classifier combination methods. In *Machine Learning in Document Analysis and Recognition* (2008).

[40] Wang, W., and Zhou, Z.-H. Analyzing co-training style algorithms. In *Proceedings of the 18th European Conference on Machine Learning* (Berlin, Heidelberg, 2007), ECML '07, Springer-Verlag, pp. 454–465.

[41] Xu, C., Tao, D., and Xu, C. A survey on multi-view learning. *CoRR abs/1304.5634* (2013).

[42] Ye, L., and Keogh, E. Time series shapelets: A new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2009), KDD '09, ACM, pp. 947–956.

[43] Zhang, H. The optimality of naive bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)* (2004), V. Barr and Z. Markov, Eds., AAAI Press.

Appendix A

Visual analytics. Wireframes and dashboards

In this chapter we show the original wireframes and the visual analytics dashboards. The dashboards were created to display the insight provided by the multi-view framework. The wireframes were the source of inspiration for the real dashboards and will lead the future developments.

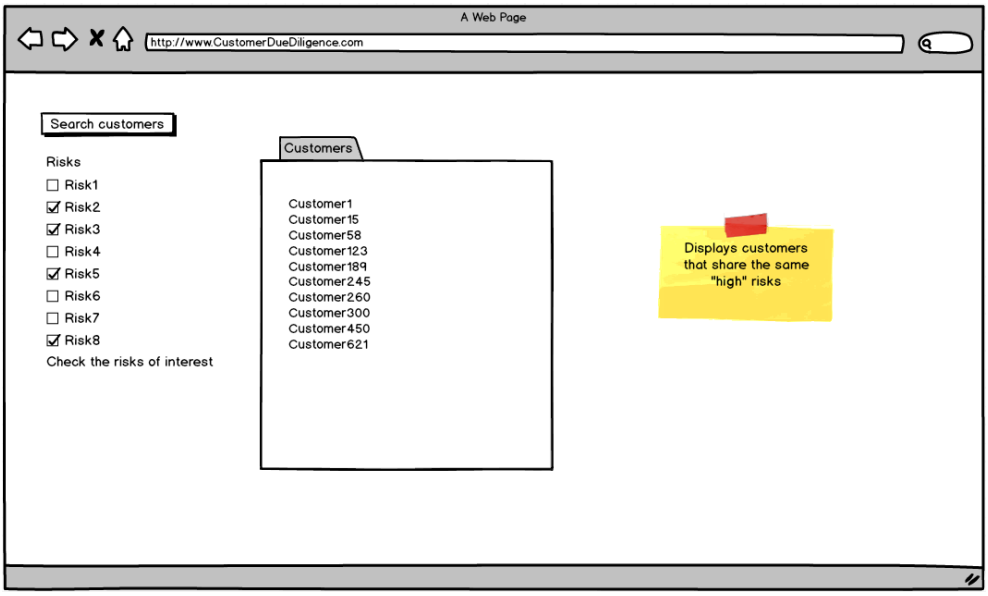

Figure A.1: Wireframe 2. Customers similarity outlook

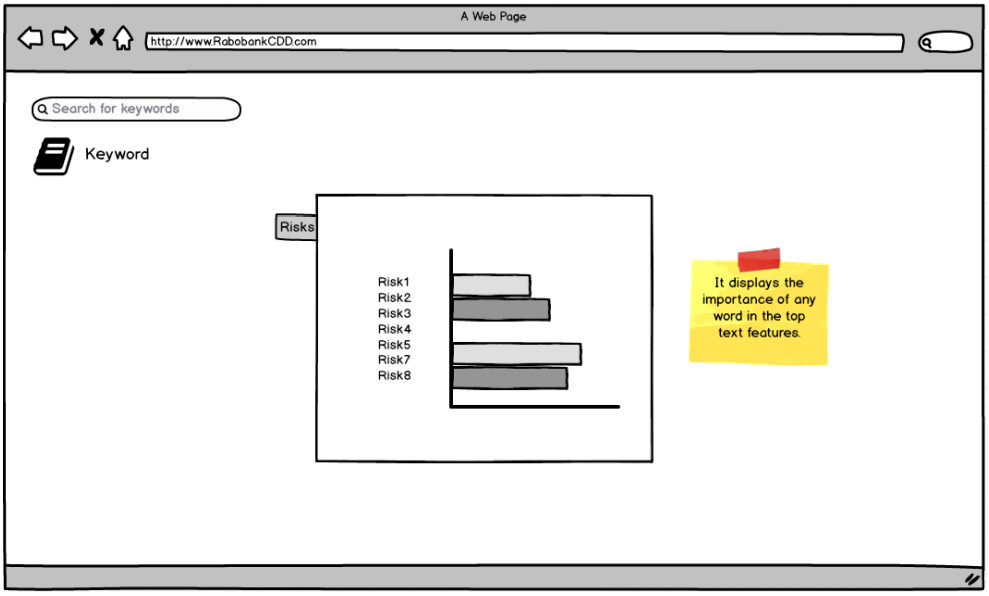

Figure A.2: Wireframe 3. Key features (words) importance per risk

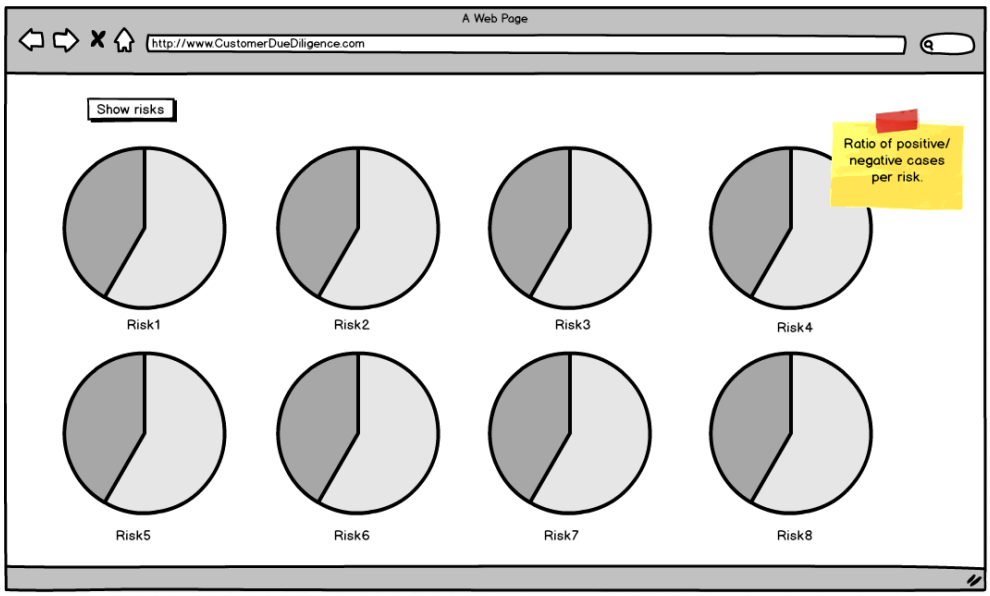

Figure A.3: Wireframe 4. Risks' labels distribution

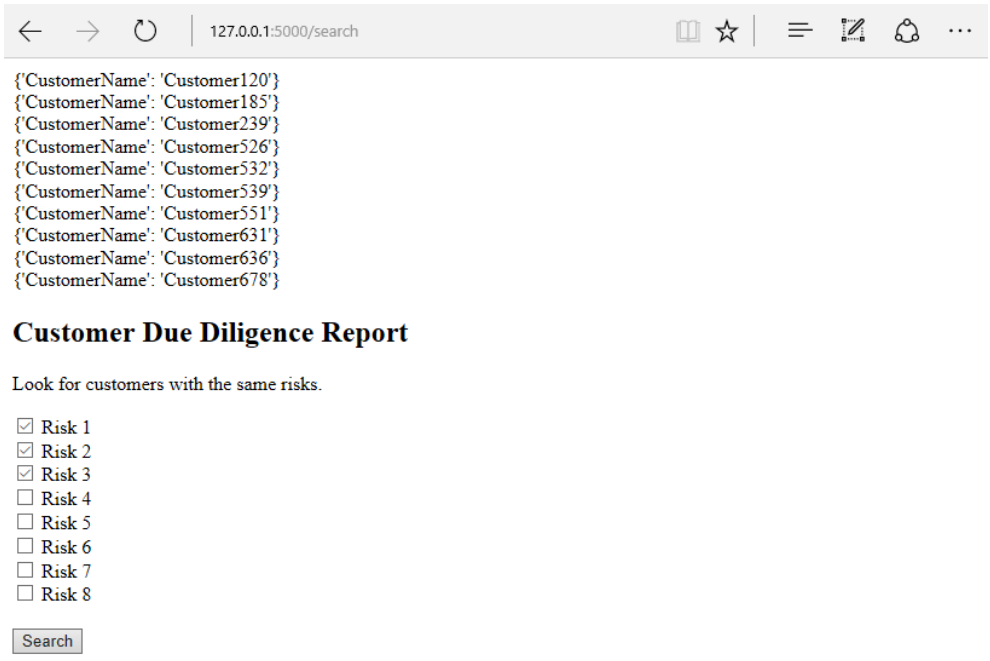

Figure A.4: Visual analytics for customer likeness (from wireframe 2)

Appendix B

Experimental results. Accuracy and Receiver Operating Characteristic (ROC) scores

The accuracy and Receiver Operating Characteristic (ROC) scores of the experiments with 12 and 50 initial training instances are presented here, but the main analysis is done in Chapter 5.

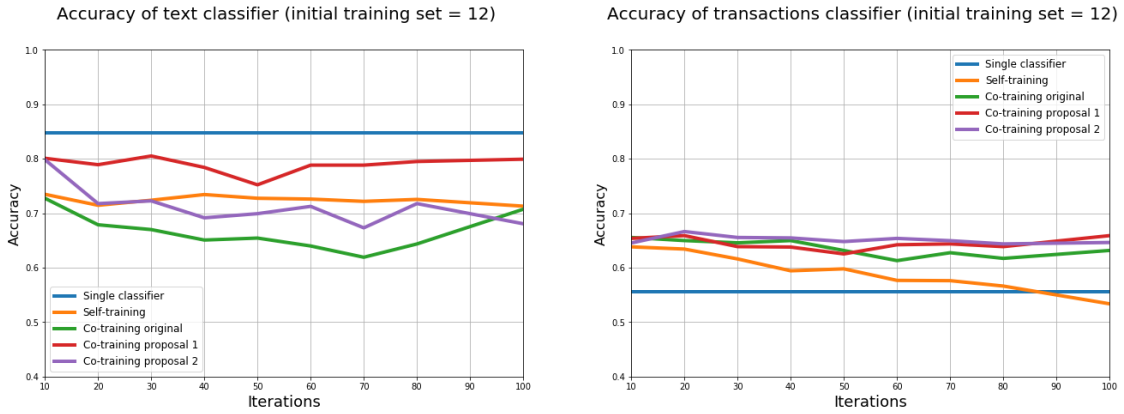

Figure B.1: Accuracy for the classification with 12 initial training instances

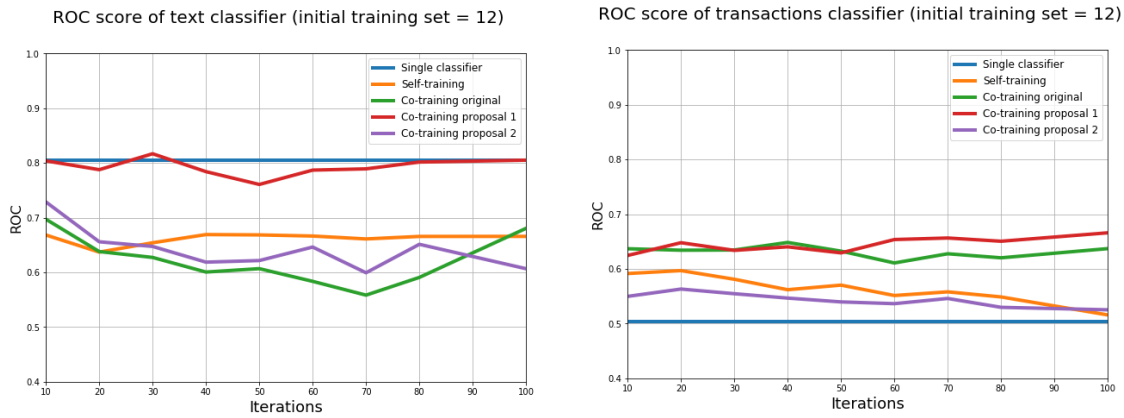

Figure B.2: Receiver Operating Characteristic (ROC) score for the classification with 12 initial training instances

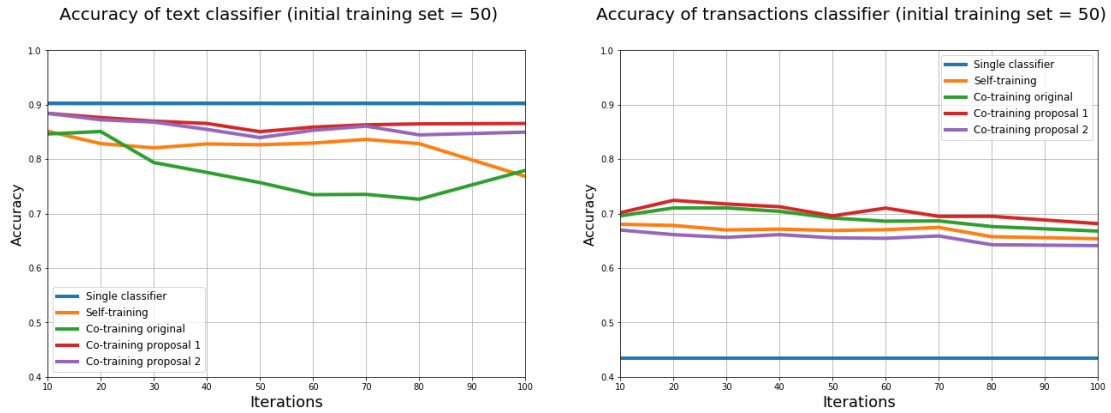

Figure B.3: Accuracy for the classification with 50 initial training instances

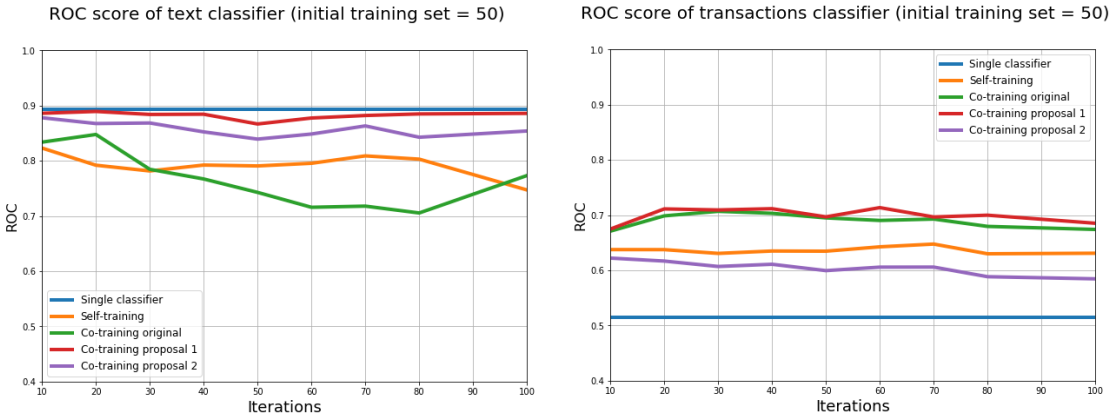

Figure B.4: Receiver Operating Characteristic (ROC) score for the classification with 50 initial training instances