FOCA: A System for Classification, Digitalization and Information Retrieval of Trial Balance Documents

Gokce Aydugan Baydar and Seçil Arslan
R&D and Special Projects Department of Yapı Kredi Teknoloji, Istanbul, Turkey

Keywords: Pattern Recognition, Document Digitalization, Information Retrieval, Classification, ElasticSearch.

Abstract: Credit risk evaluation and sales target optimization are core businesses for financial institutions. Financial documents like t-balances, balance sheets, income statements are the most important inputs for both of these core businesses. T-balance is a semi-structured financial document which is constructed periodically by accountants and contains detailed accounting transactions. FOCA is an end to end system which first classifies financial documents in order to recognize t-balances, then digitalizes them into a tree-structured form and finally extracts valuable information such as bank names, human-company distinction, deposit type and liability term from free format text fields of t-balances. The information extracted is also enriched by matching human and company names who are in a relationship with existing customers of the bank from the customer database. Pattern recognition, natural language processing, and information retrieval techniques are utilized for these capabilities. FOCA supports both decision/operational processes of corporate/commercial/SME sales and financial analysis departments in order to empower new customer engagement, cross-sell and up-sell to the existing customers and ease financial analysis operations by digitalizing t-balances.

1 INTRODUCTION

Corporate, commercial and Small/Medium Enterprise (SME) banking requires customers to periodically provide financial documents, which are balance sheets, income statements and trial balances (t-balances), in order to evaluate their credibility and also convenience for the bank's products. Balance sheets and income statements are structured tables and demonstrate periodic snapshots of the company's financial situation. T-balance, on the other hand, records all assets, liabilities and shareholders′ equity in details. In other words, a company's t-balance shows its all detailed transactions.

T-balance is used by two different business units by YapıKredi Bank in Turkey with different purposes. In the Credit Risk Evaluation Process Department, it is used by expert financial analysts to check out the financial situation of the customers. At the end of the evaluation, they decide the credibility of customer loans in a considerable way. Customer Relation Management and Sales Department use t-balance for both extracting cross-sell and up-sell opportunities for current customers and detecting new customers who are in a relationship with existing customers of the bank.

In Turkey, a t-balance is a semi-structured free text format document which has around 1000 rows on average. This document contains debit and credit amounts of accounting items such as current and non-current assets, long and short term liabilities, stocks and so on, (Williams et al., 2005), (Warren, 2014). Therefore, reading and processing t-balances manually is highly time-consuming for both financial analysts and relational managers (RMs). Furthermore, in the bank's previous document management system, all financial documents were stored in the same place as a bulk of customer documents without any descriptive annotations. Thus; an employee who wants to access customer′s t-balances which were provided before the system change has to look through each customer document until finding the desired one. A typical credit risk evaluation process takes about 6 days because financial analysts examine all t-balances of the past three years. On the other hand, value chains, credit line distribution among competing banks, deposit distribution, cheque/note values all serve as sale targets for RMs. Processing all these documents and extracting all valuable information by matching external resources, such as a customer database, are time-consuming procedures and even infeasible with human effort in limited time.

Aydugan Baydar, G. and Arslan, S.
FOCA: A System for Classification, Digitalization and Information Retrieval of Trial Balance Documents.
DOI: 10.5220/0007843201740181
In *Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA 2019)*, pages 174-181
ISBN: 978-989-758-377-3

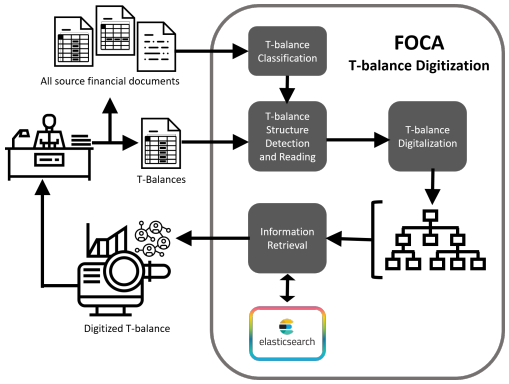

Figure 1: Framework of FOCA.

In this paper, we present a new system called FOCA for t-balances′ classification and digitalization, Figure 1. We classify t-balances and digitalize them with valuable information by using a combination of Pattern Recognition, Natural Language Processing and Information Retrieval techniques. In digital t-balance, we provide all accounting items with mapped bank names, currencies, demand/time deposits, liability terms, and human/company distinction. We match the mentioned human and company names with the bank's customer database and extend the knowledge about the related customer. We also present annotations such as factoring, leasing, and goodwill, which are crucial for financial analysis.

Our system has been integrated into the bank in June 2018 and is being used since then. For CRM usage of FOCA, we have classified and digitalized around 170,000 historical customer documents and extracted around 50 million lines of information from 100,000 t-balances. This output is consumed by RMs and 1700 new/potential companies, 540 of them are high volume companies, are extracted as a sales target. We have also served for financial analysis purposes; around 250,000 t-balances are digitalized by FOCA. We detect t-balances with approximately 93% recall rate and digitalize them with a 94% success rate as of February 2019. Furthermore, we match human and company names from the customer database in milliseconds with 74% success. Since FOCA classifies and digitalizes a t-balance in few seconds, it eases RMs′ and financial analysts′ job and shortens the processes.

The rest of the paper is as follows. T-balance concept and difficulties caused by both the structure and content of the t-balance are given in Section 2. Section 3 explains datasets, classification model, customer matching, t-balance digitalization algorithms, and information retrieval techniques. Experimental setup and results are given in Section 4 and Section 5 gives the conclusion of the paper.

2 CONCEPT AND DIFFICULTIES

2.1 Concept

The Italian accounting system is adopted in Turkey. T-balances consist of at least four main columns; a specific account code, account explanation, total debit, and total credit amounts.

Account codes are globally standard indicators in order to show certain accounting items. Most common account codes are given in Table 1.

Table 1: Common Accounting items and their globally standard account codes.

100	Cash
101/103	Cheques
102	Bank Accounts
108	Other Liquid Assets
120	Account Receivables
121/321	Bonds
301/401	Leasing
320	Account Payables
500	Shareholders

Account explanation is a free format short text which summarizes the related accounting item. For instance, 102 bank account of the company may contain bank name, deposit type, time vs. demand type and currency. T-balance is a valuable document because it contains detailed financial information about the t-balance owner. In the credit risk evaluation, cheatings over the other financial documents are caught thanks to t-balances. It also contains a company's all business relationships and this nature make t-balance a golden mine for new sales opportunities and potential customer extraction. All this valuable information can be extracted from account explanation cells.

Total debit and credit amounts state the volume of the record.

2.2 Structure based Difficulties

Although the Italian accounting system is adopted in Turkey, there is not a standard for bookkeeping. T-balances are semi-structured documents and this nature causes lots of problems.

Account code starts with a globally specific three digit number but breakdowns, in which account details are listed, depending on the accountant's styling. For instance, bank accounts are recorded with "102" account code but breakdowns can be given as "102-01-01" or "102.01" for the same record.

T-balance is a semi-structured table and there is not any constraint about the maximum number of

rows or columns. According to our knowledge, t-balances have mostly four to eight columns. Besides the main four columns, transaction currency, monthly debit and credit columns and debit, and credit balance columns are commonly given. Since the number of columns is not standardized; detecting the positions of the main t-balance columns becomes a difficulty. Some t-balances have header as a row which contains a definition of columns and headers ease to find the positions of the main t-balance columns. The ones which do not have any header bring the problem of analyzing the data in each column in order to detect whether it is a target column or not. The number of rows depends on both the size of the company and the accountants' style. For instance, t-balances have the same number of bank accounts may contain a different number of records depending on the detailing levels. The beginning of a t-balance is also not standard; some directly start with bookkeeping and some contain other information such as t-balance period, owner's title and so on. The rows that locate above the t-balance origins should be detected to find the beginning of the bookkeeping.

T-balance is a free-format document, which means that a t-balance can be constructed as any type of documents like PDF or Excel. This causes the problem of tackling different document type problems.

2.3 Information Retrieval Difficulties

Account explanations contain valuable information but there is not any linguistic standard for this field. The data here is shaped by accountants; thus, the explanation column may contain lots of abbreviations and is highly prone to mistakes such as a misspelling. Furthermore, although there are not any character constraints, the text given in this field is highly short,2 to 10 words on average. This weakness is the basic problem of information retrieval part of the t-balance digitalization.

Bank names pass through the accounting items are mostly misspelled in t-balances. For instance "Yapıkredi Bankası" commonly written as "Yapı ve Kredi Bankası", "Y.K.B", "Yapıkredi" and "Yk bank". According to our research, we have detected 33 different phrasings of this bank name and this number does not include the misspelled variations such as "Ypı Kredi". In addition, there are banks with similar names in Turkey. For instance "Finansbank" and "Türkiye Finans Katılım Bankası" are two different banks. "Türkiye Finans Katılım Bankası" is generally abbreviated as "T. Finans Bank" and this abbreviation is highly similar to "Finansbank" for information retrieval algorithms. Another bank name related problem is the short bank names such as "Ing Bank" and "HSBC".

Sectors and types of company titles are the most abbreviated parts in account explanations. It is a common behavior to abbreviate these but there is more than one abbreviation and even they are prone to misspelling. For instance, the sector "sanayi" is commonly abbreviated as "san", "sn" and even "sana". Misspelling on short texts causes information algorithms to fail to find the similarities.

Table 2: Example of common proper name problem from bank′s customer database.

ABC ağaç paletçilik makine ve tekstil sanayi aş
ABC international hijyen ltd şti
ABC dayanıklı tüketim mamülleri ticaret ltd şti
ABC dağıtım gıda pazarlama ve ticaret aş
ABC spor eğitim ve sosyal yardim vakfı işletmesi

The most important ability FOCA provides is customer matching through the mentioned human and company names in t-balance but there are many problems here. One of them is that there are lots of different companies with the same proper name. Table 2 shows some of 125 different customers of the bank[1] which starts with the same proper name and differentiate with sector and type parts. Companies have lots of sectors in their official title and bank records customers with the official title. However, accountants′ generally use only a few sectors in t-balances. These problems cause decreasing of similarity algorithms′ success and bring a need to manipulate the data before using these algorithms. Another problem is that there are around 25 million customers in the bank and they are stored in a relational database. Searching hundreds of fuzzy queries in a big relational database is not feasible with a time constraint.

3 METHODOLOGY

3.1 Dataset Preparation and Feature Engineering

3.1.1 Corpora for T-Balance Classification

We collect around 5000 PDF and Excel financial documents of commercial/corporate/SME customers and human annotators label them with binary labels "T-BALANCE" and "NOT T-BALANCE". PDF documents in this set contain one financial document but

[1]Proper name "ABC" given in Table 2 is not the real name, it has changed due to the privacy-preserving concerns.

Excel documents may contain lots of financial documents on different sheets. Therefore, annotators also have labeled each sheet of Excel documents. At the end of the labeling stage, we obtain around 5300 financial documents where almost half of them (47%) are t-balances.

Feature Engineering. In order to apply machine learning techniques, we extract features that differentiate t-balances from other financial documents by focusing on the differences between t-balances and other financial documents.

According to our realization, the word "mizan" (t-balance in Turkish) may be present in the sheet name of Excel documents. (i)We use the existence of this word as a boolean feature. As we indicated before, balance sheets and income statements are structured documents which means the shape is standardized. Moreover, positions and the total number of numeric columns are also fixed for these financial documents. T-balances, on the other hand, are semi-structured and both shape and positions differ from t-balance to t-balance. Therefore, (ii, iii) the shape of the documents, (iv)the position of the first numeric column and (v) the total number of numeric columns are used as features. Focusing on only the structure of the documents is inadequate because t-balances and other financial documents are table formatted documents and have too many structural similarities. Furthermore, companies may also provide some documents which contain specific part of t-balance such as cheques and these documents are even more confusing than other financial documents for machine learning algorithms. According to our knowledge, t-balances must contain at least the majority of the following account codes in order to show assets, liabilities and shareholder's equity; 100-cash, 101/103-checks, 102-bank accounts, 108-other liquid assets 120-account receivables, 121-321 bonds, 320-account payables, and 500-shareholders. (vi)How much of these account codes exist is used as a feature.

At the end of data collection and feature extraction, we obtain a $[5300, 7]$ sized corpora where the seventh columns indicate the label. We divide our data into two parts where two-thirds of it is for training and one third is for testing.

3.1.2 Corpora for Dividing Company Titles

In order to divide company titles into proper name, sector and type parts, we collect 2500 random Turkish human and company names. Human annotators have labeled the data for named entity recognition models. An example is shown in Figure 2.

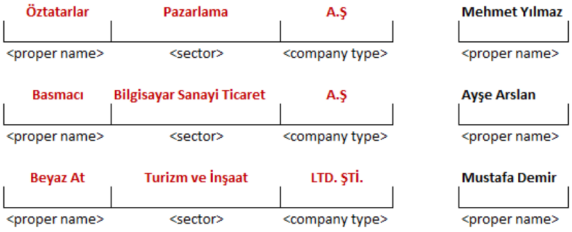

Figure 2: Human and company name data examples. Red examples are companies and black examples are human names.

3.1.3 Corpora for Testing Customer Matching

We collect six different datasets in two different collections in order to test customer matching.

The first collection consists of five different sets, A, B, C, D and E, which include existing bank customers. Datasets A, B, and D were created from RMs while C and E datasets were generated from the data sets.

The second collection consists of 2000 labeled queries; CIF is used for bank customers, positive samples, and -1 is used for others, negative samples.

3.1.4 Dictionaries for Information Mapping

We use dictionary-based algorithms in order to extract valuable information. For this purpose, we have four different dictionaries for bank names, company indicators, annotations and sectors'-company types' abbreviations. These dictionaries contain synonyms, abbreviations and common misspelling versions of each word.

3.2 Classification of Financial Documents

In order to detect t-balances among the financial documents, we need to classify them. For this purpose, we focus on document structure and layout analysis, document/text and binary classification algorithms. Document structure and layout analysis algorithms mostly work on picture-based datasets. Text classification algorithms classify text by analyzing words and frequencies. Our data neither consist of pictures nor contain long texts. Hence, we deeply focus on feature engineering to represent our data with numbers and then practice binary classification methods.

We perform following state of the art binary classification algorithms; Multi-Layer Perceptron, Support Vector Machines, Random Forrest Classifiers, KNN and Decision Trees, (Shawe-Taylor et al., 2004), (Natarajan, 2014).

3.3 Dividing Company Titles

Searching company titles as a whole decreases the success because of the reasons we state in Section 2.3. Thus, we divide each company title into three parts as proper name, sectors, and type.

Company titles in Turkey have a specific pattern; each title starts with a proper name and followed by either another proper name, a sector or a type. After a sector, either another sector or a type can come and company type is the end of a title, sector or proper name comes after this is irrelevant to the title.

In order to divide company titles, we utilize named entity recognition techniques. The data that we want to divide into parts is short text and the rules between parts are strict and one way. Hence, we name entities with a Hidden Markov Model (HMM), (Morwal et al., 2012). Calculated initial probabilities of our HMM model, $P(\pi)$ are given in Table 3. This table supports the rule that a company title always starts with proper names. The deviation is caused because of the synonym words; for instance “demir” is a common sector and proper name.

Table 3: HMM initial Probabilities for dividing company names.

Part	$P(\pi)$
Proper Name	0.8832
Sector	0.1151
Company Type	0.0017

3.4 Database Preparation with ElasticSearch

We create a NoSQL, search engine based database with ElasticSearch due to the reasons indicated Section 2.3.

For ElasticSearch, we collect target customers, commercial, corporate and SME, with customer identification number (CIF), national identification number (TCKN), tax number (VKN) and name/title (UNVAN). On the collected data, we first expand the abbreviations in order to standardize sectors and company types. Then, we divide company title into three parts with the HMM model. Once the data is ready, we index it to ElasticSearch where each property is represented in a field.

3.5 Digitalizing T-Balance

3.5.1 Reading and Extracting T-Balances

There are many frameworks for reading Excel documents but none of them covers all versions of Excel. Thus, FOCA uses two different frameworks; Apache POI for documents created with Excel version after then 2007 and JExcelApi for older ones. The output of both frameworks protects the table structure of Excel.

PDF documents have no indicators for the border of the table columns and commercial off-the-shelf PDF readers are not adequate for protecting table structure while reading. Thus, we implement an algorithm that takes a raw text from PDFBox and finds the column borders with the rules learned before. This algorithm searches for boarders by checking the distinct change on the sequence of characters. Since there is no character limit for account code and explanation columns, borders are located starting from the end.

Regardless of the further detail columns, FOCA is interested in four main t-balance columns; account codes, account explanation, total debit, and total credit amount columns. In order to detect the positions of these four columns on the documents that are classified as t-balance, we implement an algorithm which searches the positions by using both type and context of the columns. If there is a header, we find it with string approximation algorithms. In the case of not founding the header, we search for ”text - text - numeric - numeric” columns sequence. On the found sequence, we examine the content of the first text column; the number of found account codes of common account codes shows whether the sequence indicates the desired columns or not. If the majority is found, then the position sequence is used as the main t-balance columns.

Once the positions are found, t-balance is read into a tree structure in which parents represent the main account code as total, like ”102” for bank account total, and leaves represent the breakdown details, such as ”102.01.01 ING Bank USD”. This tree structure provides a unique representation for t-balances regardless of its length and detail level.

3.6 Information Retrieval

We map bank names, distinguish human and company names, annotate financial transactions, match bank customers and enrich the knowledge through the digitalized t-balance in FOCA. In order to perform information retrieval algorithms, we first apply a prepossessing on raw account explanation text in order to standardize the representations. In this step, we convert all characters to lower case, map Turkish characters to the Latin alphabet and remove non-alphabetic characters. Then we utilize information retrieval techniques in order to extract valuable information.

3.6.1 Mapping and String Matching

FOCA performs a cascaded algorithm that is based on edit distance and String-Searching(Melichar et al., 2005) algorithms. The algorithm firstly looks a match for account explanation as a whole. If the similarity between account explanation and any of dictionary items is not enough, θ, then the last word is cropped and the same process is repeated fractionally. The Turkish language tends to give the most valuable information first, hence the algorithm uses a reverse n-gram technique. This process is repeated from the size-gram to 1-gram until a match is found.

FOCA utilizes this algorithm in order to make financially crucial annotations, represent bank names and cities/regions in a unique way, expand abbreviations and distinguish human and company names using dictionaries.

3.6.2 Customer Matching with ElasticSearch

ElasticSearch provides a fuzzy search option. It calculates the similarity between each item in the index and given query with BM25 scoring and sorts the results in descending order. We search mentioned human and company names in the ElasticSearch index. If there is a match, the knowledge is extended with CIF, TCKN, and VKN. For the ones who are not a customer of the bank are assigned with a unique negative number in order to highlight potential customers.

Similar to ElasticSearch indexing, the same preprocessing steps are also applied to the queries. We search the query with different importance levels. The proper name is searched without any toleration because this part is the most distinguishing part and ElasticSearch already tolerates misspelling. Sectors are searched with some toleration due to the reasons indicated in Section 2.3; if 60% of the words in query has found in the index then it is accepted as a match. Company types may not be given in each account explanation, so this field is searched with a should importance. The highest scored one among the matches is accepted as the result.

We map found customers all over the t-balance. This utility extends the knowledge about the t-balance owner. For instance, some t-balance account codes refer to business partners and shareholders, namely the group of the company. Detected business partners are marked in t-balance to highlight the transactions in the group members. Furthermore, the companies and humans are assigned with a negative number are considered as potential customers.

4 EXPERIMENTAL RESULTS

4.1 T-Balance Classification

We test machine learning models on the dataset explained in Section 3.1.1. Among the experimented algorithms, Multi-layer Perceptron(MLP) performs the best scores on average, Figure 3. In terms of recall score, Desicion Tree (DT) model provides the highest score bu its precision score is not good enough. We need a classification model in order to correctly distinguish t-balances from similar documents. This means that both recall and precision measures are crucial for the system. As MLP outperforms all other alternatives on average, we prefer the MLP model.

FOCA has classified approximately 100,000 t-balances among 170,000 financial documents and extracted 50 million rows of information for 8 months. The user feedbacks also show that the classification model works.

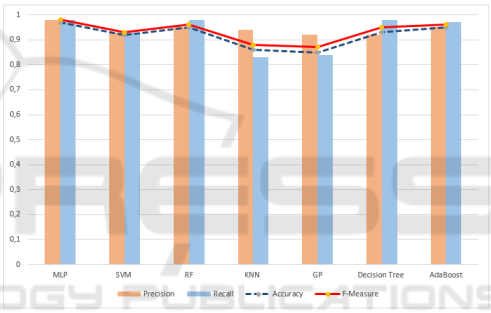

Figure 3: T-balance classification experiment results.

4.2 Customer Name Matching

We evaluate customer matching with two experiments using collections explained in Section 3.1.3. In the first experiment, we use the first collection which consists of only existing bank customers. Since datasets A, B and D contain t-balance records, the queries are prone to human mistakes and may contain unusual abbreviations. C and E datasets are created by RMs so the queries are more clear than other datasets. The results show that our ElasticSearch based matching mechanism is not only successful but also fast. Searching for a query takes approximately 1 millisecond and the result is correct for t-balance entries with 74% probability.

Secondly, we experiment with our system on the second collection that contains both customers of the bank and others who have no records in the bank's database. Table 5 demonstrates the confusion matrix of this experiment. Here, p′ - n′ stands for actual labels while p - n stands for the predicted ones. As a

Table 4: Customer matching evaluation experiments and their results in terms of accuracy and elapsed time.

Dataset Name	Dataset Size	True Match	Precision	Elapsed Time (ms)
A	581	413	0.71	768
B	6926	4932	0.71	8.106
C	387	367	0.95	695
D	105	75	71.4	247
E	364	360	0.99	472

result, FOCA matches customers with 0.96 precision, 0.76 recall and 0.85 F1-score in this experiment. In other words, if there is a match for someone, this one is 96% actually a customer of the bank.

FOCA enriches knowledge about t-balance owner using the customer number obtained from this matching. Therefore, matching someone who is actually not a customer is dangerous for further analysis. FOCA searches entries with strict rules in order to decrease false-positive ratio and this is the reason behind the recall score.

Table 5: Confusion matrix for customer matching on a mixed dataset.

	n	p	
n′	582	41	623
p′	335	1083	1418
	917	1124	

4.3 T-Balance Digitalization and Information Retrieval

In the new document management system, RMs have the responsibility of obtaining and uploading customer t-balances. We assume that the document an RM uploads is a t-balance. In the past 8 months, approximately 250,000 t-balances are digitalized by FOCA and the extracted information is used for financial analysis. Figure 4 shows the number of documents FOCA digitalized per month with red bars, how many of the uploaded t-balances are recognized with blue bar and the precision values with green line. FOCA detects t-balances with 92% precision.

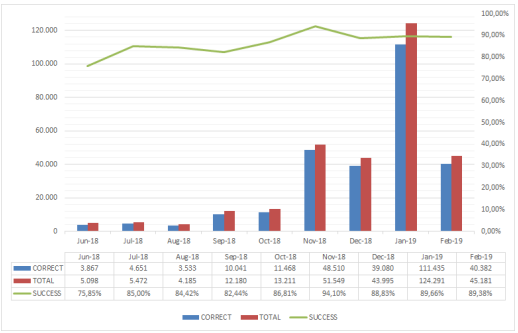

Figure 4: Number of T-balances which is digitalized for financial analysis usage and the success rate among months.

We ask users to check the digitalized t-balance. If there is a mistake in digital t-balance, such as wrong annotation, users can edit the output and we keep the edit ratio in order to measure how correctly we digitalize t-balances. The edit ratio is approximately 1.6% which means that the information FOCA retrieves is correct with 98.4% probability.

At sales opportunity extraction phase, we processed all historical t-balances and retrieved demand and time deposit items are investigated by corporate and commercial sales business units.After this investigation, RMs are engaged for 1,700 new companies, of which 540 are high volume. This analysis is infeasible with human effort in such short time.

5 CONCLUSION

We present FOCA which is a new end to end system. FOCA takes raw t-balances as input and returns digitalized t-balance as output. Digitalized t-balance represents it in a tree structure, where parents are account totals and leaves are breakdown details, and contains valuable information. Mapped bank names, currencies, bank account deposit/demand types, liability terms, annotations and human/company distinction are extracted from t-balance. Furthermore, mentioned human and company names are matched from customer database and the knowledge about the relevant customer is extended in digitalized t-balance. We test our system on different datasets and our results show that the system works fast and with high accuracy.

Our system is consumed by two units of the bank, Credit Risk Evaluation and Customer Relation Management and Sales departments. Before FOCA, examining and interpreting t-balance procedure was highly time consuming and it was also highly prone to human mistakes. FOCA completes classification, digitalization and information retrieval steps in few seconds which is infeasible with human effort. As future work, we plan to use the output of this project in order to automatize financial analysis and sales opportunity extraction with human effort.

ACKNOWLEDGEMENTS

The authors would like to thank Bilge Köroğlu and Mert Basmacı for their contributions to the system. This work is supported by TUBITAK 3170677.

REFERENCES

Melichar, B., Holub, J., and Polcar, J. (2005). Text searching algorithms. *Available on: http://stringology. org/athens*.

Morwal, S., Jahan, N., and Chopra, D. (2012). Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing (IJNLC)*, 1(4):15–23.

Natarajan, B. K. (2014). *Machine learning: a theoretical approach*. Elsevier.

Shawe-Taylor, J., Cristianini, N., et al. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Warren, C. (2014). *Survey of accounting*. Nelson Education.

Williams, J. R., Haka, S. F., Bettner, M. S., and Carcello, J. V. (2005). *Financial and managerial accounting*. China Machine Press.