Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science
Data Mining Research Group

Detection of emerging compliance issues by matching structured and unstructured data

2IMM00 - Final report

Juan Manuel Gonzalez Huesca

Supervisor:
prof. dr. Mykola Pechenizkiy

version 1.0

Eindhoven, January 2018

Contents

Chapter 1

Introduction

Data is known as the new 21th century most valuable commodity[1]. Information extracted from data plays a very important role in the success of companies as it supports in making decisions, understanding and being compliant with regulations and improving other business functions[2].

Banks are no exception, and as stronger regulatory compliance needs emerge, it is important to find an effective and automated process for knowledge discovery. This will be achieved by leveraging the high volume of available data in both its forms: structured and unstructured (which sometimes represents up to 80% of the data), which are generated by systems and bank employees. For example, matching the insight extracted from employees emails (unstructured data) with their trading or other commercial activity (structure data). This will allow us to monitor traders compliance and, further, find emerging compliance issues as some communication patterns are discovered by using text mining technologies and statistical methods.

A framework proposal will be presented, where information from every electronic communication (e.g. emails) will be extracted, modelled and correlated to trade events (time series) and other commercial activities from different data sources, trying to find entities to link these data, and possible current or new compliance issues. It is worth to mention that a demo of the framework operation will be introduced and explained along the report, but further development is needed to achieve the desired functionality.

1.1 Motivation

The banking industry has struggle with regulatory compliance in recent years[3]. It has become a very challenging task for banks because the volume and complexity have drastically increased. So there is a big need of a reliable, accurate and automate process to detect and, more importantly, predict compliance issues in a banking environment.

Although there has been several attempts to tackle this challenge, most of them just focus on the analysis of numeric data. The linkage between structured and unstructured data, as it is a relatively new research field, has not being implemented to improve regulatory compliance. This is where this work is relevant, as it will develop further a new field in data science while helping the banking industry to improve its compliance monitoring.

[1] https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valu

[2] https://www.inc.com/jason-albanese/top-digital-experts-share-how-current-data-trends-can-drive-business-success.html

[3] https://www.infosys.com/industries/financial-services/Documents/regulatory-compliance-management.pdf

1.2 Problem statement

The proposed framework will link structured with unstructured data and find correlations to detect and predict compliance issues in a banking environment. The emails will represent the unstructured data and the trades the structure one.

To illustrate the process, lets consider a trader who constantly uses electronic communication to exchange information across peers and other individuals and organizations. This same trader is also leading the trading activities for company A. Through text mining techniques applied on the traders emails we discover that he is disclosing sensitive data about company A with other people, thus incurring in a compliance issue. We can also detect patterns that lead to this behavior, allowing us to predict future case, and in this way we can avoid them.

1.3 Approach

The unstructured data (emails) will be analyzed using text mining technologies and statistical methods, and as a start, the linkage between this data and the structured data (trades numeric data) will be the named entities (using NLTK python library), but further approaches will be implemented once the official dataset from the bank is obtained. Not only the named entities will be extracted from the emails, but also events will be detected, and all this information will create a new table with columns: contentID, timestamp, entities, event, actors (sender,receiver), email-Content. In the case of the trades, they will be modeled as: eventID, event, entities, timestamp, deviation, trader, and the task will be to match the corresponding entities with the actor (trader) within a time frame (using the timestamp) to detect patterns that lead to compliance issues.

For this report we will present a demo, where, based on the ENRON email public dataset, entities are extracted and their frequency is mapped across 24 months. A heatmap is then created with the most relevant entities found in 50 thousand emails over the year 2000 and 2001.

1.4 Results

The heatmap is a very informative way to present emerging patterns because it clearly shows, on a time frame, where more occurrences of an entity or event happen. The demo proved this point, the created heatmap highlights the time frame where a particular entity appears and also the frequency. This information can be compared with the trades deviations (difference between forecast and real) and find correlations that would indicate compliance issues.

1.5 Content structure

Organizations that can benefit from the main characteristics of data (volume, velocity, variety, veracity and value) can gain a huge competitive advantage while automating business processes and making sure it stays compliant. The remainder of this report will showcase and explain how text mining techniques can be used to extract knowledge from data and how this, in principle, unstructured and unorganized data can be linked with current records on databases to find interesting patterns. Section 2 introduces the current state of the art in the field by conducting a literature review. Section 3 describes in more details the question to be answered during the report. Section 4 gives an overview of the methodology and the approach followed for the framework demo. On section 5 the results are presented and discussed. We present the conclusions of the current work on section 6. And finally, insight for future development is introduced on section 7.

Chapter 2

Literature analysis

A literature review, while researching on main scientific databases such as arXiv[1] and CiteSeerX[2], was conducted to find state of the art methods and techniques to link unstructured with structured data, and then find interesting patterns for knowledge discovery and detection of emerging compliance issues in a banking environment. The proposed link between the two types of data is the named entities categorized as organizations (company names). Nevertheless, in this report we will just present an overall demo of the framework operation as part of the seminar in data mining.

2.1 Structured and unstructured data

Structured data refers to data stored with a high degree of organization, such as the one recorded on relational databases in columns and rows. In this study, I will refer to structured data to the trades and commercial activities recorded on the bank databases and it is assumed to be in the form of a time series encoding metrics or performance events over a specific period of time. On the other hand, unstructured data does not have any internal structure, such as documents, news, emails, social media posts, digital images, videos, etc., and because of its nature it requires a different set of techniques to analyze it and extract value from it. For this study purposes, the traders emails and any other electronic communication data will be considered unstructured data.

2.2 Framework proposal

There has been some efforts in the past to link structured with unstructured data with different approaches. EROCS (Entity RecOgnition in Context of Structured data) [1] is a system to link a giving text document (unstructured data) with relevant structured data such a trades, in an external database. In EROCS, structured data is predefined as a set of entities and the system identifies the entities that best match the given document, which is seen as a sequence of sentences. The approach is similar to a dictionary-based entity-recognition but EROCS can identify the entity even if it is not mentioned in the document, and also, the terms matching the entity can be spread in all the document. The system uses part-of-speech tagging to identify key noun-phrases in the sentences, and then the entities around them.

Another interesting tool that associates text data with information from relational databases is LIPTUS [2]. Unlike EROCS, LIPTUS can potentially associates employees interactions (emails) with trades or customers profiles, by analyzing the identifiers, which are heuristically extracted

[1] https://arxiv.org/

[2] http://citeseerx.ist.psu.edu/index

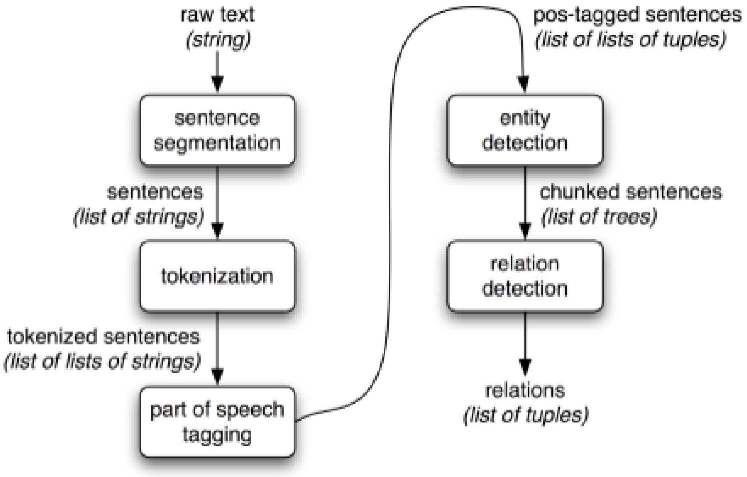


Figure 2.1: Simple Pipeline Architecture for an Information Extraction System. This system takes the raw text of a document as its input, and generates a list of (entity, relation, entity) tuples as its output. For example, given a document that indicates that the company Georgia-Pacific is located in Atlanta, it might generate the tuple ([ORG: 'Georgia-Pacific'] 'in' [LOC: 'Atlanta']).

from the text. This system can also be used to extract events from data, which is an important activity in the framework proposal, as the pivot of the correlation of data.

As a first approach for the demo, we will focus on Named Entity Recognition (NER) but in the future this will be complemented with other tools like LIPTUS to expand and make a more robust solution. Figure 2.1[3] shows the architecture of a simple system for information extraction, where one of the main goals is to extract the named entities. Overall speaking this same approach will be used during the demo to extract the entities from the emails which will be tagged with an ID and a timestamp to find a correlation with the trades or other commercial activity.

Figure 2.2 illustrates the overall proposed framework to match structured with unstructured data and look for emerging compliance issues while trading. The framework was inspired in the approach taken in [4] but it simplifies some blocks and uses different methods like the combination of NER and LIPTUS for the data matching, while keeping a robust solution. It is worth to mention that further analysis and testing is required with a real bank dataset to confirm the approach or make the required changes.

The functional details of each component of the framework is explained below:

A. Data sources.

Data is gathered from different sources. On one side, the text data from emails (if required from other text sources as well like blogs, news or IM tools); and on the other, the data from trades. These will be the core data to be linked in further steps.

B. Data processing.

A standard text cleaning approach is used to remove special characters as most of the time they represent noise in the emails. It is important also do spelling check as it is highly likely to have spelling mistakes in text entered by employees on emails. As mentioned on figure 2.1, POS tagging is very important to discover the named entities so this step is implemented in every email..

[3] http://www.nltk.org/book/ch07.html

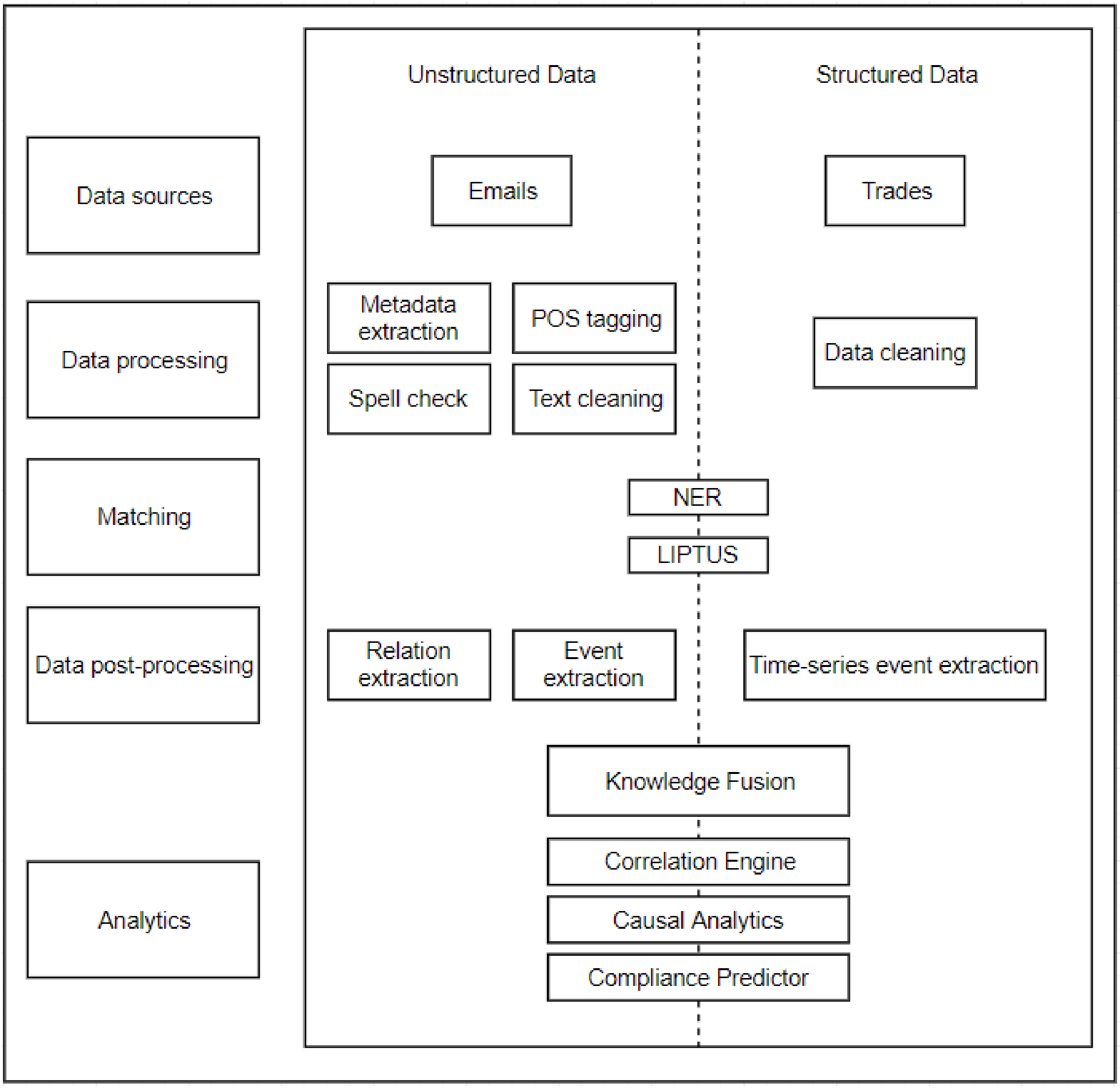


Figure 2.2: Framework for matching structured with unstructured data and detection of emerging compliance issues (based on [4]).

Besides later techniques, meta-data is extracted to find other useful information like timestamp, sender, recipient, and in the case of the trades we will just focus on relevant data to get every trade in the format: eventID, entity, timestamp, deviation, trader. It is worth to mention that as this is mainly a cleaning email interaction another steps are also needed: removal of the stock replies, removal of the history text, and removal of the advertisements and disclaimers.

C. Matching.

The techniques described on figure 2.1 and on LIPTUS are used in this stage to find the links between the two types of data. Another technique for unsupervised automatic hashtag annotation [3] will be tested in this section, as it has proven good results for extracting information from text data, and transforming it into structured relations that can be also linked with information on relational databases.

D. Data post-processing.

This section provides the methods to give structure to the data, for further usage when finding

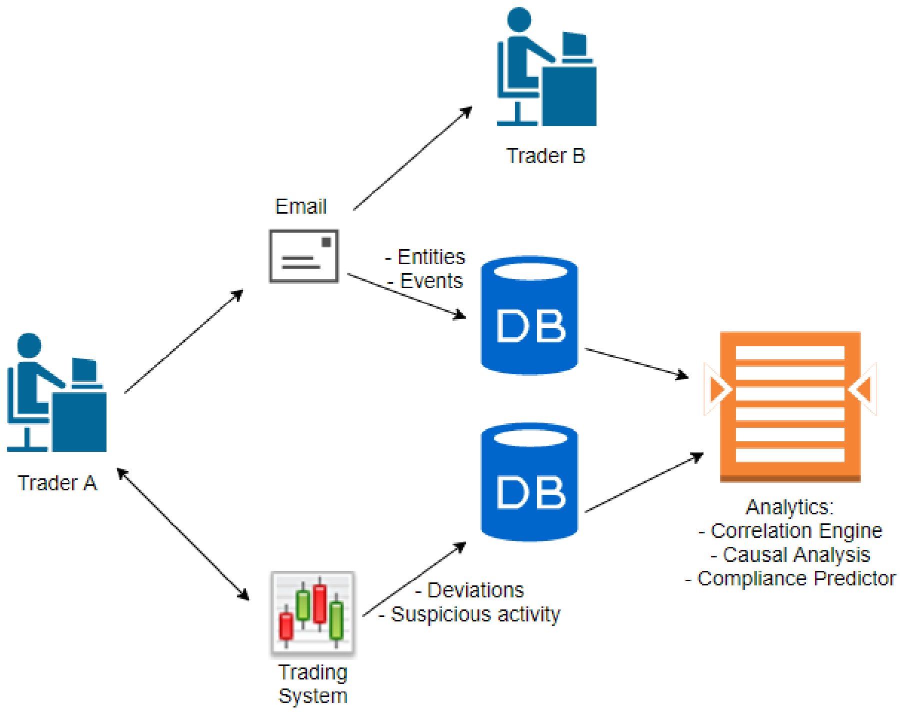


Figure 2.3: Running example of the proposed framework.

correlations between events and emerging compliance issues. This is where the relations and events are extracted from the text data, and also the time series events from the trades are extracted. As something additional, it is also possible to do a sentiment analysis to have more elements for a causal analysis.

F. Analytics.

Once all the data is available in a proper format, it is possible to create a correlation engine to find causal relations and predict compliance issues.

As a running example, figure 2.3 describes the overall process. As it is establish in the bank policies, all kind of electronic communication is being recorded and monitored, so when trader A sends an email to trader B, the entities and events are extracted (along with other data as explained before) from the email. On the other hand, traders commercial activity is being recorder for anomalies, deviations or suspicious activity which is then recorded on a database, which fetch the analytics block (correlation engine, causal analysis, compliance predictor) to find correlations between the email and the trades, leading to compliance issues detection and prediction.

2.3 Approaches

There are many approaches in how to represent and model text data, and also how to extract information from it. These techniques and approaches will be tested and the chosen one will be based on the results for this particular bank dataset, if necessary further research will be done.

Among the choices that needs to be made are:

- Named entity recognition has proved to be an effective way to extract insight automatically to link and correlate text and numeric data [6]. Although entities are effective, other approaches to link data will be tested further such as events, and if necessary further research will be done with the final dataset.

- There are also severals ways to analyze the text, some of them are determined by the system to be used. EROCS for instance is based on a sentence approach for the documents. On the other hand there are also another approaches like the unsupervised methods for automatic hashtag annotation mentioned on the framework. Another very important approach [5], in an study following a similar overall objective, introduces an identification of the document's topic by using a vector space model.

These approaches has proven to work under different scenarios so further testing and evaluation needs to be performed under this particular banking environment.

2.4 Lines of research

There are many questions and possibilities that arise when the unstructured information is in a more organized format, but the main one we want to address here is to find emerging compliance issues based on the patterns discovered with the timeseries events (trades) and the global events extracted from the text data (emails). Here, it will be possible to answer questions like: how likely is for this employee to do trading for a particular customer?, or what is the probability this employee fall into compliance issues?, etc.

Also, this new structure of information for unstructured data can be used to find new predictive models and causal relations that can be implemented in the trader compliance monitoring department and in risk assessment, as it will be possible to know if employees are following the regulations and if not, why and how can we predict it based on their electronic communication and trading and commercial activities.

Last but not least, sentiment analysis can be used to find patterns in the employees' daily and commercial activities with their state of attitude.

2.5 Questions

Every bank has its own set of tools, processes and data formats; so it will be very useful to get a dataset of the electronic communication and some trade samples in order to understand the nature and wonder what other questions can be answered from it.

Also, labeled data can provide more insight while working on building a predictive model, and also to discover emerging compliance issues. Overall, we need to understand the business rules, the processes, the frequency of these events, the current tools being in placed and the employee constraints regarding electronic communication and their trades/commercial activities. All this data will be useful to move forward.

Chapter 3

Question proposed

The proposed framework will link unstructured data (emails) with structured data (trades) to find correlations and detect and predict compliance issues. As mentioned earlier, this is a relatively new area, which has gained relevance due to the high volume of heterogeneous data and the need to create more robust and reliable solutions. And this is why we are interested in the implementation of an information fusion system to ensure regulatory compliance in a banking environment.

As explained on figure 2.3, internal electronic communication is being monitored to detect compliances issues caused by the employees (traders) when they exchange information about a company or organization they are also doing trading for. For example, if trader A is trading is dealing with financial instruments from company Z, then he is not supposed to disclose any information regarding these particular trades with anybody on any electronic communication mean, this is what is to be ensure.

Chapter 4

Methodology

Jupyter notebook (python): https://github.com/juanmangh/Seminar-Data-Mining

We created a basic demonstration in python of the main sections in the framework proposal, up to the Data post-processing block (not including the Analytics) due to time constraints. Its also worth to mention that the unstructured data section was done using a the ENRON email dataset with over 500 thousand emails, but just 50 thousand were used due to hardware processing limitation. In the case of the structured data, no suitable database was found which could be used to link with the results obtained during the text mining process, so we decided to make the assumption that this data exists, as it will be explained below.

The code was divided into 6 sections:

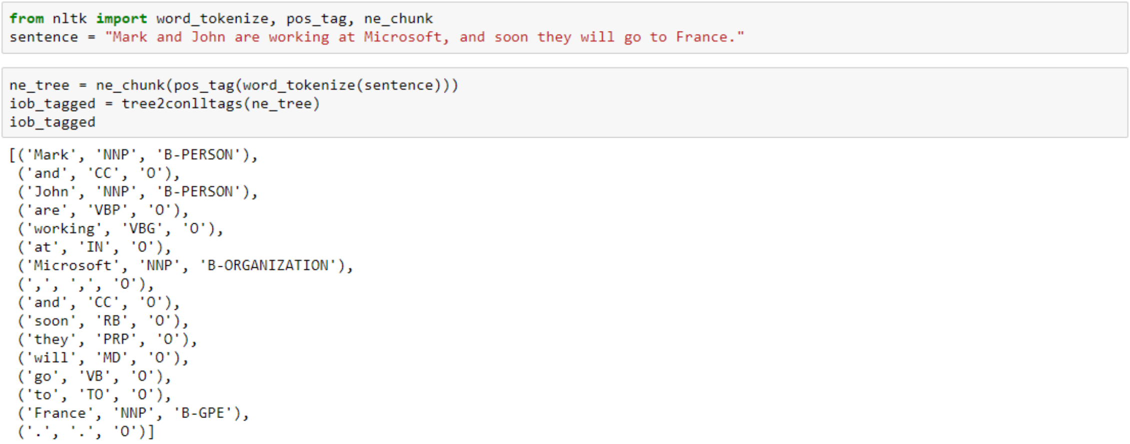
1. Import of python libraries. Here, the main libraries are imported to be later used during the heatmap creation. The main libraries used were: pandas, nltk, numpy, seaborn and re.

2. Function to extract email body from dataframe. ENRON email dataset has only two columns, the email ID and the email content information itself with its metadata. In this case, we focus on the second column, to extract just the email body, and for later use, the timestamp, sender and recipient.

3. Entity extraction with NLTK library. A function was created to extract one organization entity per email were available. This step was done with the NLTK library tuned to detect just organizations, as it is possible also to extract entities for persons and locations, but in this case we are interested just in organizations. At the end of the process, the function give us a dataframe with 50 thousand rows (one per email analyzed) with the format: ID, Date, Email-Body, Entity, Sender, Recipient; resulting in 5818 unique entities. Figure 4.1 shows a running example of the functionality of the entity extraction and POS tagging of NLTK.

4. CSV creation with columns ”ID”, ”Date” and ”Entity”. For the heatmap creation we are interested just in the columns ID, Date and Entity, so for the time being we created a .csv file to be imported into Microsoft SQL Server SMS for further computation.

5. External processing. This section is important because the entity column is cleaned from special characters and was also truncated for an easier processing in Microsoft SQL Server SMS. It is important also because the SQL server allow us to query the dataframe and group the occurrences of each main entity per month. At the end, we have a table with the columns Entity, Date (in months), and Frequency (number of occurrences per month), which was created in Excel using the output from the SQL query. Its worth to mention that the selection of entities was done manually, as per their relevance as an example.



```
from nltk import word_tokenize, pos_tag, ne_chunk
sentence = "Mark and John are working at Microsoft, and soon they will go to France."
```

```
ne_tree = ne_chunk(pos_tag(word_tokenize(sentence)))
iob_tagged = tree2conlltags(ne_tree)
iob_tagged
```

```
[('Mark', 'NNP', 'B-PERSON'),
 ('and', 'CC', 'O'),
 ('John', 'NNP', 'B-PERSON'),
 ('are', 'VBP', 'O'),
 ('working', 'VBG', 'O'),
 ('at', 'IN', 'O'),
 ('Microsoft', 'NNP', 'B-ORGANIZATION'),
 (',', ',', 'O'),
 ('and', 'CC', 'O'),
 ('soon', 'RB', 'O'),
 ('they', 'PRP', 'O'),
 ('will', 'MD', 'O'),
 ('go', 'VB', 'O'),
 ('to', 'TO', 'O'),
 ('France', 'NNP', 'B-GPE'),
 ('.', '.', 'O')]
```

Figure 4.1: Example of named entity recognition using NLTK.

6. Heatmap creation. The final table is then imported into the jupyter notebook again, and then a pivot function is used to create a table in heatmap format. This new table is then used as as input for the seaborn heatmap library and heatmap is created.

During the whole demo, two different methods for named entity recognition were used: NLTK and Stanford NER, but at the end we decided to implement the one from NLTK because of the performance.

For further development, we plan to capture more than one entity per email. Also, we want to add the capability to create a heatmap per user and display the results even per day.

We faced many challenges during the whole development like noisy data in the emails body, and also some format challenges with the output of the NLTK library but they were are handled properly.

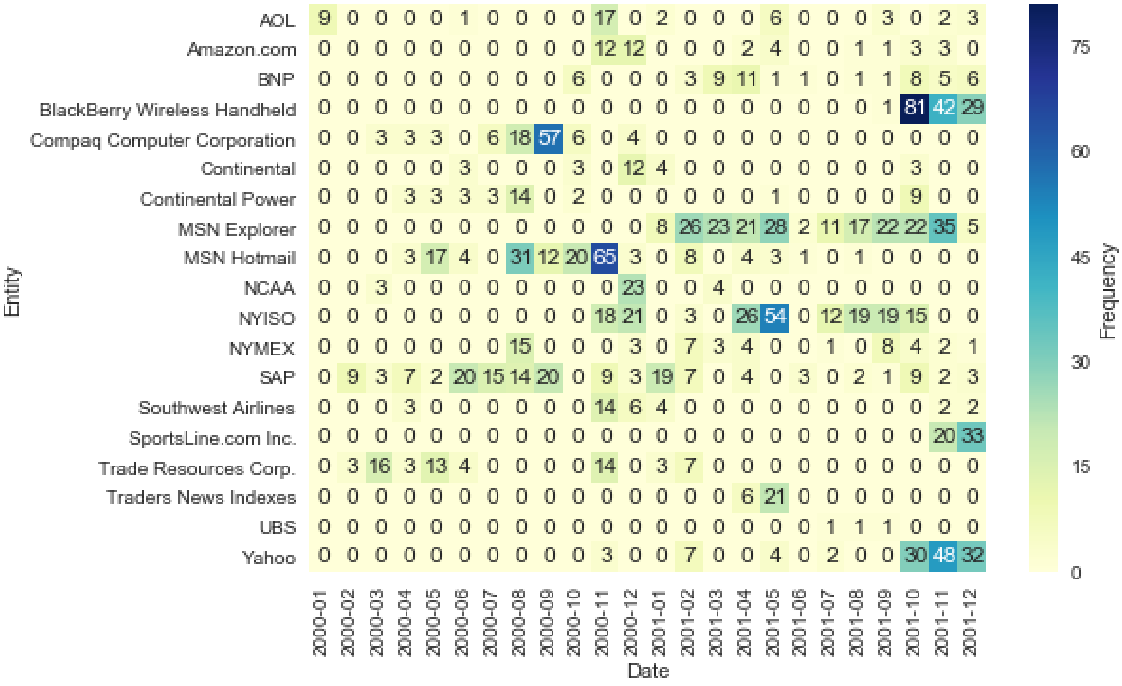
Chapter 5

Results

The output of the whole demo is a heatmap where we can see the frequency of the appearance of some entities per month, from January 200 to December 2001 (24 months). Its very interesting to see that once we have this information, if it is properly assigned to the sender, we can correlate it with the respective trade information for the same entity and confirm whether the trader fell into a compliance issues.

For example, out of the ENRON email dataset (figure 5.1) we can observe that during October 2001 the entity Blackberry Wireless Handheld had a high number of occurrences, which indicates this was a big issue for the company at the time.

With the proper trading dataset, we could correlate this peak in the Blackberry entity with the commercial activities a trader is working on. But for the time being, the demo proved its possible to extract useful insight from email using text mining techniques.



Entity	2000-01	2000-02	2000-03	2000-04	2000-05	2000-06	2000-07	2000-08	2000-09	2000-10	2000-11	2000-12	2001-01	2001-02	2001-03	2001-04	2001-05	2001-06	2001-07	2001-08	2001-09	2001-10	2001-11	2001-12
AOL	9	0	0	0	0	1	0	0	0	0	17	0	2	0	0	0	6	0	0	0	3	0	2	3
Amazon.com	0	0	0	0	0	0	0	0	0	0	12	12	0	0	0	2	4	0	0	1	1	3	3	0
BNP	0	0	0	0	0	0	0	0	0	6	0	0	0	3	9	11	1	1	0	1	1	8	5	6
BlackBerry Wireless Handheld	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	81	42	29
Compaq Computer Corporation	0	0	3	3	3	0	6	18	57	6	0	4	0	0	0	0	0	0	0	0	0	0	0	0
Continental	0	0	0	0	0	3	0	0	0	3	0	12	4	0	0	0	0	0	0	0	0	3	0	0
Continental Power	0	0	0	3	3	3	3	14	0	2	0	0	0	0	0	0	1	0	0	0	0	9	0	0
MSN Explorer	0	0	0	0	0	0	0	0	0	0	0	0	8	26	23	21	28	2	11	17	22	22	35	5
MSN Hotmail	0	0	0	3	17	4	0	31	12	20	65	3	0	8	0	4	3	1	0	1	0	0	0	0
NCAA	0	0	3	0	0	0	0	0	0	0	0	23	0	0	4	0	0	0	0	0	0	0	0	0
NYISO	0	0	0	0	0	0	0	0	0	0	18	21	0	3	0	26	54	0	12	19	19	15	0	0
NYMEX	0	0	0	0	0	0	0	15	0	0	0	3	0	7	3	4	0	0	1	0	8	4	2	1
SAP	0	9	3	7	2	20	15	14	20	0	9	3	19	7	0	4	0	3	0	2	1	9	2	3
Southwest Airlines	0	0	0	3	0	0	0	0	0	0	14	6	4	0	0	0	0	0	0	0	0	0	2	2
SportsLine.com Inc.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	33
Trade Resources Corp.	0	3	16	3	13	4	0	0	0	0	14	0	3	7	0	0	0	0	0	0	0	0	0	0
Traders News Indexes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	21	0	0	0	0	0	0	0
UBS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
Yahoo	0	0	0	0	0	0	0	0	0	0	3	0	0	7	0	0	4	0	2	0	0	30	48	32

Figure 5.1: Heatmap for entities occurrences from January 2000 to December 2001.

Chapter 6

Conclusion

The demonstration gives us a solid foundation for further development as the concept proved the capability to extract entities to be used as the linkage between the unstructured data and the structured data. This results will allow us to keep building on top to reach the goal, propose a framework and methodology to match unstructured and structured data to find correlations, and detect and predict compliance issues in the banking industry.

The proposed framework can be implemented also to other data sources, where extraction, classification, evaluation and correlation of events are needed.

Its important to mention the relevance of modelling the data as time series, because this allow us to find the correlation in predefined time windows.

Chapter 7

Future work

The demo proved the overall concept, but further work needs to be implemented on top and a more detailed analysis of the target dataset needs to be done to determine the final methods.

We need to extract events from both, the text and numeric data, because this information can add value in the correlation and give more contexts while matching the data. Also, another section to improve is in the entity extraction, as we need to add more than one entity if available. As said before, these approaches will be tested on the target data set and if necessary, other means to link the data will be researched and tested.

Also, once we have the target dataset, the approaches applied on [4] for the structured data will be implemented and tested. We need at the end to have a fully end-to-end running example with the target dataset to confirm the performance and the potential of this framework.

Chapter 8

Bibliography

[1] Venkatesan T. Chakaravarthy, Himanshu Gupta, Prasan Roy, Mukesh Mohania, ”Efficiently linking text documents with relevant structured information,” Proceedings of the 32nd international conference on Very large data bases, Seoul, Korea, September 12-15, 2006.

[2] M. Bhide, A. Gupta, R. Gupta, P. Roy, M. Mohania, Z. Ichhaporia, ”LIPTUS: associating structured and unstructured information in a banking environment,” Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Beijing, China, June 11-14, 2007.

[3] Lin Li, William M. Campbell, Cagri Dagli, Joseph P. Campbell, ”Making Sense of Unstructured Text Data”, 2017.

[4] Lipika Dey, Ishan Verma, Arpit Khurdiya, Sameera Bharadwaja H., ”A Framework to Integrate Unstructured and Structured Data for Enterprise Analytics”, Proceedings of the 16th International Conference on Information Fusion, 2013.

[5] Deovrat Kakde, Arin Chaudhuri, ”Leveraging Unstructured Data to Detect Emerging Reliability Issues”, 2016.

[6] Lipika Dey, Ishan Verma. Text-driven Multi-structured Data Analytics for Enterprise Intelligence, Proceedings of the 2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013.