Financial Banking Dataset for Supervised Machine Learning Classification

Irina RAICU
The Bucharest University of Economic Studies
irina.raicu@ie.ase.ro

Social media has opened new avenues and opportunities for financial banking institutions to improve the quality of their products and services and to understand and to adapt to their customers' needs. By directly analyzing the feedback of its customers, financial banking institutions can provide personalized products and services tailored to their customer needs. This paper presents a research framework for creation of a financial banking dataset in order to be used for Sentiment Classification using various Machine Learning methods and techniques. The dataset contains 2234 financial banking comments from Romanian financial banking social media collected via web scraping technique.
Keywords: *Dataset, Financial banking, Web scraping, Opinion mining, Machine learning*

1 Introduction

With the explosive growth of social media (i.e. reviews, forum discussions, blogs, microblogs and social networks), customers are encouraged to express their thoughts online and also exchange their opinions regarding financial banking products and services. Thus, there is a large amount of financial banking data containing opinions generated from a variety of social media sources. Customers or potential customers are significantly influenced in their decision making regarding a product or a service by reading the online reviews to find out others' opinion about that product or service. [1]

The Romanian financial banking marketplace is constantly evolving to a digital marketplace. Financial banking customers are able to share their thoughts online and also perceive and receive others' opinions on different portals. Therefore, there is a need to explore the opinions insights from Romanian financial banking comments and to find out what people discuss on Romanian financial banking forums.

The information available on the Web consists predominantly of unstructured text. A significant challenge is collecting the needed information from different web pages with very heterogeneous formats in a structured way. Gathering data is the most important step in solving any machine learning problem. In the past few years, various researchers and practitioners have investigated different methods for data collection from online sources. One of the most popular methods to retrieve Web content at scale is **web scraping** i.e. the automated and targeted extraction of data. Various frameworks and Application Programming Interfaces to develop customized scrapers, as well as configurable ready-to-use scraping tools exist. Comprehensive overviews of frameworks and tools for different extraction tasks are presented by Glez-Peña et al. in [2] and Haddaway in [3]. Using Scrapy framework, an extraction of 2234 financial banking comments in Romanian language between June 2009 and April 2018 has been performed. The financial banking dataset contains data from the Conso portal, which is the most popular forum in financial banking social media from Romania.

This research paper presents the framework used for creation of financial banking dataset from Romanian marketplace in order to be used in Machine Learning context for Sentiment Classification and social media analytics.

To the best of our knowledge, there is no available dataset for sentiment classification of Romanian financial banking customer reviews.

The financial banking dataset is available in Romanian and English on Kaggle, one of the largest and diverse data communities for Machine Learning researchers and practitioners. The Romanian dataset is available at

DOI: 10.12948/issn14531305/23.1.2019.04

https://www.kaggle.com/iryna13raicu/financialbankingcommentsro and at https://www.kaggle.com/iryna13raicu/financialbankingcommentsen for English version.

2. Research Framework for Financial Banking Dataset Creation

The main research goal is to identify the relevant financial banking data from Romanian online sources in order to obtain consistent data of financial banking customers in a structured format to be analyzed and classified in a Machine Learning context. Thus, a Financial-Banking Dataset Creation framework is proposed. The framework is illustrated in Figure 1.

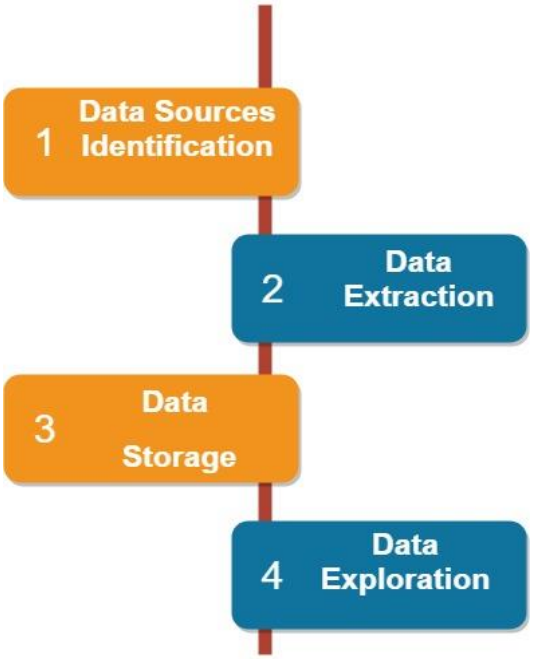

Fig. 1. Financial-Banking Dataset Creation Framework

The main stages of the framework are:

- Data Sources Identification
- Data Extraction
- Data Storage
- Data Exploration

In the following subsections, the stages for dataset creation are detailed.

2.1. Data Sources Identification

The first step of research framework is the identification of relevant financial banking data sources which contain various discussion topics where users' experiences about various institutions, their offered conditions and their customer services with focus on Romanian marketplace. The main data sources which contain information regarding products and services from Romanian financial-banking marketplace are illustrated in Table1.

The most relevant financial reviews were found on the Conso (www.conso.ro) portal. Conso is an extremely popular and reliable website in search of financial banking knowledge with reasonable number of users. The portal provides a general overview regarding financial and banking products and services from Romanian marketplace such as:

- Different types of loans (personal loan, mortgage, car loan, leasing, real estate loan) - Savings (deposit, saving accounts, SME deposit)
- Payments (debit card, internet banking, free cash withdrawals)
- Investment (investment funds, pensions, insurance)
- Other kind of information provided by Conso is related to utilities (electricity), exchange rate, and evolution of exchange rate, guides for various types of loan, cards, and savings.

DOI: 10.12948/issn14531305/23.1.2019.04

Table 1. Main Data Sources

Data Source	Web Source
Finzoom	http://www.finzoom.ro
Conso	http://www.conso.ro
Vreau credit	http://www.vreaucredit.ro/
Vreau depozite	http://www.vreaudepozite.ro/
Vreau card	http://www.vreaucard.ro/
Efin	http://www.efin.ro
eCredite	https://ecredite.net/tag/scoring-credite/
SolutieCredit	http://solutiecredit.ro/
BugetulFamiliei	http://www.bugetulfamiliei.ro
InCont	http://www.incont.ro
Info Bancare	http://www.infobancare.ro
Ghiseul Bancar	http://www.ghiseulbancar.ro
Financiare	http://www.financiare.ro
Bank news	http://www.banknews.ro
Money center	http://www.moneycenter.ro
Bancherul	http://www.bancherul.ro
CursDeGuvernare.ro	http://cursdeguvernare.ro/
Piata financiara	http://www.piatafinanciara.ro/
Ziarul financiar	http://www.zf.ro/
Lichidatori	http://www.lichidatori.com

Conso portal is structured as a standard website with a similar section to a web forum dedicated to users' reviews (Vocea Clientului in Romanian) where the users can express their own opinion regarding a bank or a financial institution and their financial banking products or services. An overview of financial banking forum from Conso portal is illustrated in Figure 2.

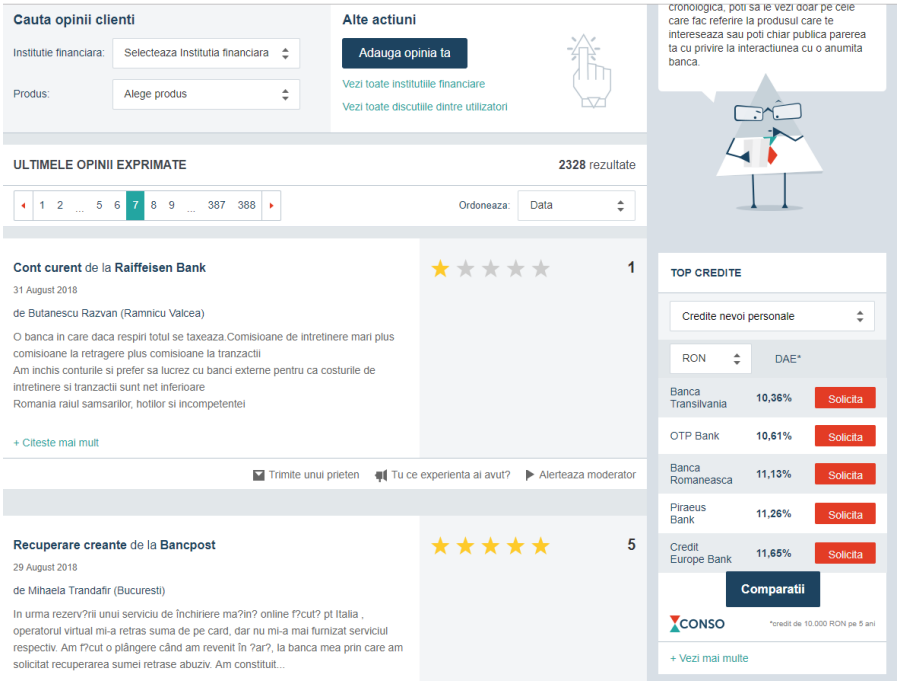

Fig. 2. Overview of Financial Banking Forum on Conso portal

DOI: 10.12948/issn14531305/23.1.2019.04

The customers interested in different financial banking products or services are saving time when visiting this website and compare the offers. They can compare offers regarding real estate loan, mortgage, and personal loan, refinancing credits, private pensions, deposits, cards, personal investment, auto loans, SME loan and electricity. Also, the users can express their own opinion regarding a bank or a financial institution and their financial banking products or services.

2.2. Data Extraction

Recently, web scraping methods have emerged as a promising way for online data collection. Scraping technique is considered as one solution to collect data in an automatic way from various Internet resources. Web scarping known also as data extraction or web crawling is the process for finding and extracting data from the web page in a structured format such as files or database. Another definition of web scraping states transforming process of the semi-structured documents from the web pages into a markup language such as HTML or XHTML, and analyzing the document in order to obtain the needed information. [4]

A web crawler also known in other terms such as indexers, web spiders or web robots is an automated program, application or script that automatically scans through web pages to collect information. The web crawler searches and extracts data from web pages, navigating from URL to URL, according to some predefined algorithms. The main role of the web crawler is to simplify and to automate the entire crawling process and makes the data crawling easy and accessible to everyone. According to the [5], crawlers are classified into two main categories: classical (traditional) crawlers and focused crawlers. The classical (traditional) web crawlers navigate the webpages and gather both relevant and irrelevant information which is a huge waste of crawling time and the storage of the downloaded information [6] Focused crawlers do not crawl the whole web as opposed to the traditional crawlers, as they only crawl to the deepest the specific part of the web that is related to the given topic.

The web scraping process consists of two main steps: fetching, downloading the web pages and parsing the obtained data to the desired information. Web scrapers are software programs with similar functionalities of web crawlers also with some differences illustrated in Table 2.

Table 2. Main differences between a web crawler and a web scraper

	Crawler	**Scraper**
Data Source	Downloading pages from the Internet	Extracting data from various sources including Internet
Scalability	Large scale	Any scale
Deduplication	Essential part	Not an essential part
Intelligent Agents	Crawler	Crawler and parser.

A significant challenge is to collect in an automatically way the financial banking information from Conso portal using a web scraper. Various frameworks and Application Programming Interfaces to develop customized scrapers, as well as configurable ready-to-use scraping tools exist. The platforms for scraper implementation differ from one another in terms of scalability, flexibility and their performance in different scenarios. Robustness and Politeness are properties that every web scraper must provide. [7] Other properties that should be considered when implementing a web scraper are: performance and efficiency, distributed, scalability, quality, and extensibility.

Scrapy is an open-source web crawling framework based on Python programming language for scraping massive amounts of data from various sources in a robust and efficient manner. [8] Scrapy is an integrated system that includes an engine for controlling the data flow

DOI: 10.12948/issn14531305/23.1.2019.04

between all the components, a scheduler for receiving requests, a downloader for fetching web pages and custom classes (called spiders) written by users to parse responses and to extract data. [9] In the literature, there are many approaches which use web scraping based on Scrapy framework for data collection. Landers et all [10] proposed an approach called theory-driven web scraping in order to collect data regarding substantive theory for psychologies. An interesting domain where web scraping gained attention is criminal justice [11]. Web scraping is widely used also in eCommerce applications. [12] [13]. Data scraping was successfully applied in real-estates domain. [14]

Extracting information in a structured format from Conso portal involves retrieving automatically the links that lead to posts and obtaining the actual data objects of those posts. For this purpose, a web scraper has been implemented using Scrapy framework. For the implementation of the web scraper, Anaconda distribution has been used. The web scraper implementation is available at https://github.com/irina-raicu/ATLAS/scraping

The web scraper is based on a focused crawler for collection of all web pages from Conso portal where financial banking posts are posted and on a parser for extraction of needed text from the entire website.

The task of capturing and structuring data extracted from the Web is divided into two parts: crawling and data scraping. Crawling effortless multiple URLs from Conso portal by avoiding non-informative data and duplicate pages, scraping a huge amount of data automatically using CSS selectors to select relevant data related to different banking services and products such as: loans, deposits and cards, storage of data into a specific structured format such as JSON, an exploitable structure to facilitate data processing and analysis, are provided by Scrapy. Others relevant requirements for implementation of a web spider for our application are related to CPU and memory usage, and speed. Memory and CPU requirements of Scrapy follow the amount of data that is needed for a multithread application, also speed requirement of Scrapy in automatically navigation of dynamic URLs and data extraction is satisfied.

2.3. Data Storage

As described to the previous stage, an extraction of 2234 financial banking posts in Romanian language from Conso portal between June 2009 and April 2018 has been performed using Scrapy framework.

For each review, the following information has been collected into a JSON format.

(1). Entire text of the review
(2). Date of the review
(3). Financial banking institution
(4). Financial banking product
(5). Characteristics that have been evaluated by the users
(6). Rate of each characteristic
(7). Star-rating of the reviews

The following example is a review in Romanian language from Conso portal in JSON format:


```
{
    "text": "Neserioasa banca! Prima banca care comisioneaza incasarea salariului, desi isi fac publicitate ca nu au nici un cost.",
    "autor_opinie": "de Florin Stefan (Braila)",
    "data_opinie": "20 Aprilie 2018",
    "banca": "Banca Transilvania",
    "produs_bancar": "incasarea salariului",
    "review_total": "1",
    "caracteristica1": "Transparenta costurilor",
    "nota_caracteristica1": "1",
    "caracteristica2": "Timpul de asteptare",
    "nota_caracteristica2": "1",
    "caracteristica3": "Functionarii institutiei financiare",
    "nota_caracteristica3": "1 ",
    "caracteristica4": "Procedura de lucru",
    "nota_caracteristica4": "1",
    }
```


DOI: 10.12948/issn14531305/23.1.2019.04

As aforementioned before, the aim of the extracted data is to be used in Sentiment Classification using supervised machine learning methods and techniques. Supervised learning mechanism identifies specific relationships or structure in the data received as input in order to effectively predict correct output data. Therefore, the importance to supervised learning of having access to labeled data is paramount. [15]

On Conso portal, each post has associated a star value provided by the financial banking user. Comments posted on Conso portal uses a star system based on a scale one to five for a review. Also, it is important to notice that not all the posts represent opinions regarding a financial banking service or products. Some users post updates regarding legal ordinances or other relevant legislation modifications, petitions, suggestions. In addition, financial banking representatives have a dynamic interaction with users through Conso and they offer responses to users regarding financial banking product or service or other relevant information.

The collected data serves the input for supervised classification and thus, there is a need that each post from Conso portal to be labelled (concept known also as "annotation"). Most common annotation for sentiment classification is classification in polarity classes positive or negative or positive, negative or neutral. Thus, in the dataset, each post is annotated with opinion labels (positive, negative or neutral in order to capture the polarity of subjective texts. To deal with objective texts, each post is annotated with factuality labels (opinions, facts and experiences).

The annotation for opinion labels is performed automatically using SentiWordNet lexicon. SentiWordNet is a lexical resource that associates to each sense of a term scores according to the notions of positivity, negativity and objectivity. [16]

The application used for labelling the customer comments is described in Figure 3.

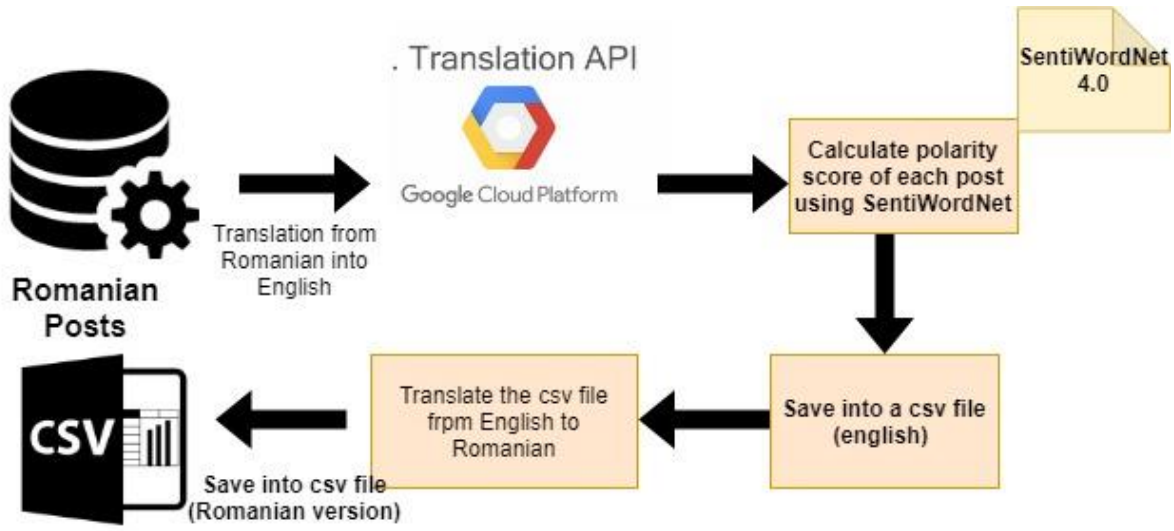

Fig. 3. Automatic Opinion Labels Annotation of Romanian posts

Factuality is successfully applied in sentiment classification in order to capture the polarity of objective texts. [17] Even if fact, opinion and experience concepts are very similar; there is a distinction among these categories. In certain application domains facts may also have polar orientations, since they may have negative/positive implications for users (financial banking customers, in this case). For instance, adoption of a new legal ordinance to increase customer interest has a negative impact on financial banking customer from his point of view, because the costs will increase. Facts are considered **objective** information whilst opinions and experiences are **subjective**.

A **fact** represents information used as evidence. A fact can be proved.

DOI: 10.12948/issn14531305/23.1.2019.04

Examples of facts from the extracted data are "You have to read the contract carefully" (translation from Romanian original post "Trebuie sa citesti atent contractul ») and "Regarding the message of Mr. Narciz Bejinariu, we will communicate the following: Loan interest rate is retained. Depending on the supporting documents submitted by the client on request for restructuring, a total grace period (no principal or interest) or partial interest (up to 12 months) can be granted. The Bank charges the unique commission for services rendered at the request of the respective clients 0 5% of the current credit balance at the request date (minimum 100 maximum of 1,000 lei)" (translation from Romanian orginal post "Referitor la mesajul domnului Narciz Bejinariu va comunicam urmatoarele: La restructurarea creditului se pastreaza dobanda in vigoare. In functie de documentele justificative prezentate de client la solicitarea restructurarii se poate acorda o perioada de gratie totala (nu se plateste nici principalul si nici dobanda) sau partiala (se plateste dobanda) de pana la 12 luni. Banca percepe comisionul unic pentru servicii prestate la cererea clientilor respectiv 0 5% din soldul curent al creditului la data solicitarii (minimum 100 maximum 1.000 lei) »)

An **opinion** is a judgment, viewpoint, or statement about something not necessarily based on fact or knowledge.

Examples of opinionated posts are "An unbearable bank! It is the first bank which takes fee for wages, but it is declaring as having no cost" (translation from Romanian orginal post "Neserioasa banca! Prima banca care comisioneaza incasarea salariului desi isi fac publicitate ca nu au nici un cost ») and BCR beneath criticism! Terrible!!! So NOOO, NO, NO !!!! (translation from Romanian orginal post "BCR sub orice critica! Groaznic!!! Deci NUUUU, NU, NU!!!!»)

An **experience** is something someone has lived through and that have an effect on her. It is expected to be true (as in the case of a fact), but may be affected by personal feelings (closer to opinion concept).

Examples of sentences describing experiences are "I regained my trust in BCR. We had few troubles with First House loan but I believe this bank appreciates their clients and takes into account their opinions. Mrs. Counselor was very nice and welcoming us with smiles every time even in difficult moments. I particularly appreciate the professional and polite way in which I received explanations and answers to my questions. I recommend Orizont Brasov Branch. At the beginning I gave 1 star but now I give them 5 stars! Thank you!" (translation from Romanian orginal post "Mi-am recapatat increderea in BCR. In Creditul Prima Casa am avut cateva impotmoliri din care am iesit cu bine si m-am lamurit ca aceasta banca tine cont de parerile clientilor. In sucursala doamna consilier a fost foarte draguta intampinandu- ma cu zambet chiar si in momente mai tensionate. Apreciez in mod deosebit modul profesionist si politicos in care am primit explicatii si raspunsuri la intrebarile mele. Recomand Agentia Orizont Brasov.La inceput am dat 1 steluta acum dau 5 stelute! Multumesc!") and "Hello. I have the same unpleaseant experience with BCR. Internal rate of 8 9% plus margin. I join with those who want to do something against BCR. We propose to mediate this case of BCR and through media channels to send a joint news to newspapers. We should set up a common meeting for all parties together with a lawyer and file complaints at the bank and possibly newspapers."(translation from Romanian orginal post "Buna ziua. Si eu am aceeasi experienta trista cu BCR-ul. Dobanda interna de 8 9% plus marja. Ma alatur si eu celor care vor sa faca ceva impotriva BCR-ului. Propun sa mediatizam acest caz al BCR-ului si prin canalele mass-media eventual sa trimitem o sesizare comuna posturilor de televiziune ziarelor. Ar trebui sa stabilim o intalnire comuna toti cei patiti impreuna cu un avocat si sa depunem sesizari la anpc banca si eventual televiziuni ziare. »)

As aforementioned, experiences are widely shared among the financial banking community users (even more than opinions). Therefore, a new category to the traditional categorization of facts vs. opinions is added. It is possible that, when describing an experience,

DOI: 10.12948/issn14531305/23.1.2019.04

the user also expresses an opinion, so that experiences are not always mutually exclusive from opinions.
If an experience includes an opinion, the annotator is asked to label the sentence as "experience". On the other hand, "facts", "opinions" and "experiences" may be positive, negative or neutral, depending on whether they express or arose positive, negative or neutral sentiments and feelings, respectively.
The annotation process for extracted data is performed manually by a specialized person in financial banking domain. There is a possibility that, when describing an experience, the customer also expresses an opinion, so that experiences are not always mutually exclusive from opinions. If an experience includes an opinion, the sentence is labelled as "experience".
Table 3 shows examples of positive, negative and neutral facts, opinions and experiences.

Table 3. Examples of sentences from the extracted data according to their factuality (Romanian version)

Fact	
Positive	Gusti Gusti faptul ca tu nu casti ochii in momentul semnarii unui contract nu inseamna ca cineva ti-a tras teapa.Comisionul de1% si primele de la stat de 25% pt.depunerile anuale NU SE PRIMESC daca se reziliaza contractul inainte de perioada minima de 5 ani STIPULATA in contract. Primele de la stat pot intarzia sa fie virate in contul tau BPL dar chiar daca se vireaza anual la timp tie ti se vor retine daca reziliezi contractul inainte de implinirea celor 5 ani...in rest daca te tine sa poti duce cei 5 ani de depuneri sumele acumulate (garantate de fondul de garantare a depozitelor) iti vor aduce un castig care nu ti-l aduc la ora nici o forma de depozit la nici-o banca din Romania. In concluzie, cititi bine contractul, depozitul este avantajos.
Negative	Banca nu respecta prevederile OUG 50/2010: marja fixa din contract de 1 2% a fost modificata unilateral de banca la valoarea de 8 96%; 2. Contractul de refinantare propus cu o marja de 4 5% contine multe clauze ilegale/ incorecte. La propunerea mea de corectare a contractului nici nu au vrut sa auda de asa ceva.
Neutral	OUG 50/2010 spune ca comisionul de administrare este unul ce poate fi perceput de banci. Mai mult OUG 50/2010 nu reglementeaza nivelul maxim sau minim al comisioanelor percepute. Adica tu ai semnat binemersi contractul.
Experience	
Positive	CEC Bank ag.Amzei. Prima Casa O experienta placuta tradusa printr-un mod de lucru foarte efficient al comisarului de credit. Nu am intampinat dificultati sau intarzieri neprogramate tot procesul s-a desfasurat in timp util. CEC Bank ag.Amzei. Prima Casa O experienta placuta tradusa printr-un mod de lucru foarte efficient al comisarului de credit. Nu am intampinat dificultati sau intarzieri neprogramate tot procesul s-a desfasurat in timp util.
Negative	Am aplicat pentru un credit refinantare cu ipoteca pe data de 06.11.2009. Nci pana astazi 10.12.2009 nu am reusit sa semnez contractul de credit desi dosarul a fost aprobat de conducerea bancii. Insistente pe langa functionari etc. nu au ajutat cu

DOI: 10.12948/issn14531305/23.1.2019.04

	nimic. Slab pregatiti si in dispret total fata de clienti.Din posturile de aici constat ca de fapt ceea ce mie mi se pare o exceptie este de fapt o regula a bancii.
Neutral	Buna ziua doresc sa inchid un contract cu Cetelem ca urmare a achitarii integrale a ratelor. De 2 zile stau in asteptare la telefon pentru a intra in legatura cu serviciul clienti interior 331 fara nici un raspuns.
Opinion	
Positive	Recomand cu incredere, sunt ok!
Negative	Neserioasa banca! Prima banca care comisioneaza incasarea salariului desi isi fac publicitate ca nu au nici un cost.
Neutral	Va multumesc pentru impartasirea experientei si pentru ca m-ati avertizat. E incredibil ce se intampla. Suntem foarte multi in aceasta situatie. Cu siguranta ma voi constitui parte civila intr-un eventual proces.

After performing the labelling processes, the data is saved into a csv format as a financial banking dataset. The financial banking dataset is available in Romanian and English on Kaggle, one of the largest and diverse data communities for Machine Learning researchers and practitioners. The Romanian dataset is available at https://www.kaggle.com/iryna13raicu/financialbankingcommentsro and at https://www.kaggle.com/iryna13raicu/financialbankingcommentsen for English version.

2.4. Data Exploration

The aim of this stage is to ensure that dataset is reliable for training a machine learning model and obtaining useful predictions. Thus, several metrics such as duplicated text, completeness of text values, text format, and ambiguity text detection have been applied to the financial banking dataset.

The chart in Figure 4 shows the distribution of "fact", "opinion" and "experience" classes for each comment. As expected, customers are posting comments about their experiences (1324 of reviews) comparing to facts (421 reviews) and opinions (488 reviews).

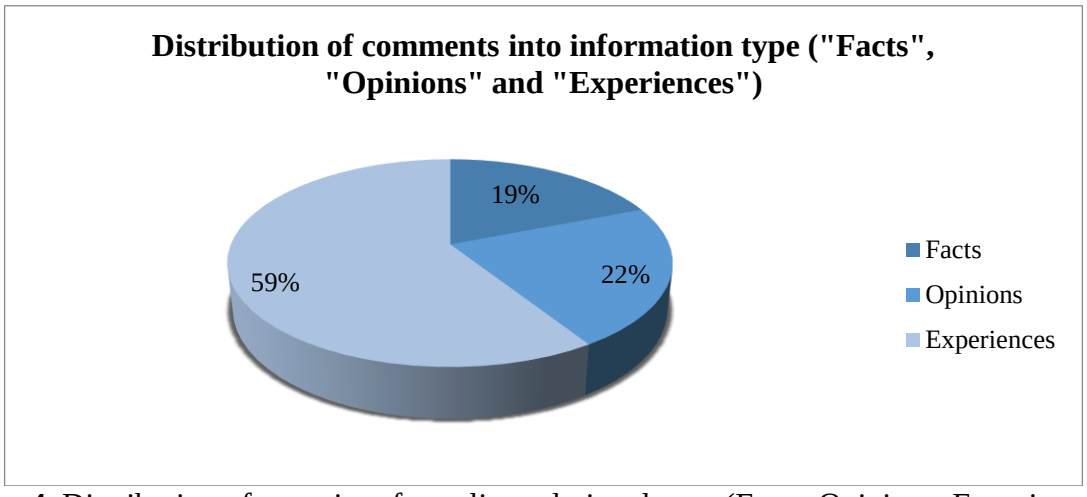

Fig. 4. Distribution of posts into factuality polarity classes (Facts, Opinions, Experiences)

The financial banking products and services were grouped into 30 of categories as shown in Figure 5.

DOI: 10.12948/issn14531305/23.1.2019.04

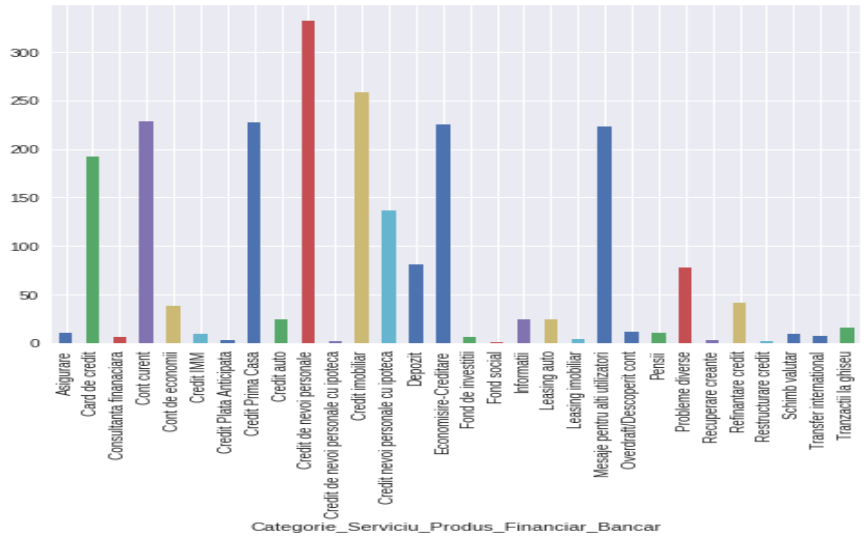

Fig. 5. Distribution of financial banking products and services

The most commented financial banking products are:

- Savings and loan (Translation from Romanian "Economisire-Creditare")
- First House loan (Translation from Romanian "Credit Prima Casa")
- Current account (Translation from Romanian "Cont curent")
- Real estate loan (Translation from Romanian "Credit imobiliar")
- Personal loan (Translation from Romanian "Credit de nevoi personale")

The figure 6 shows the distribution of financial banking institutions commented by the customers on Conso portal.

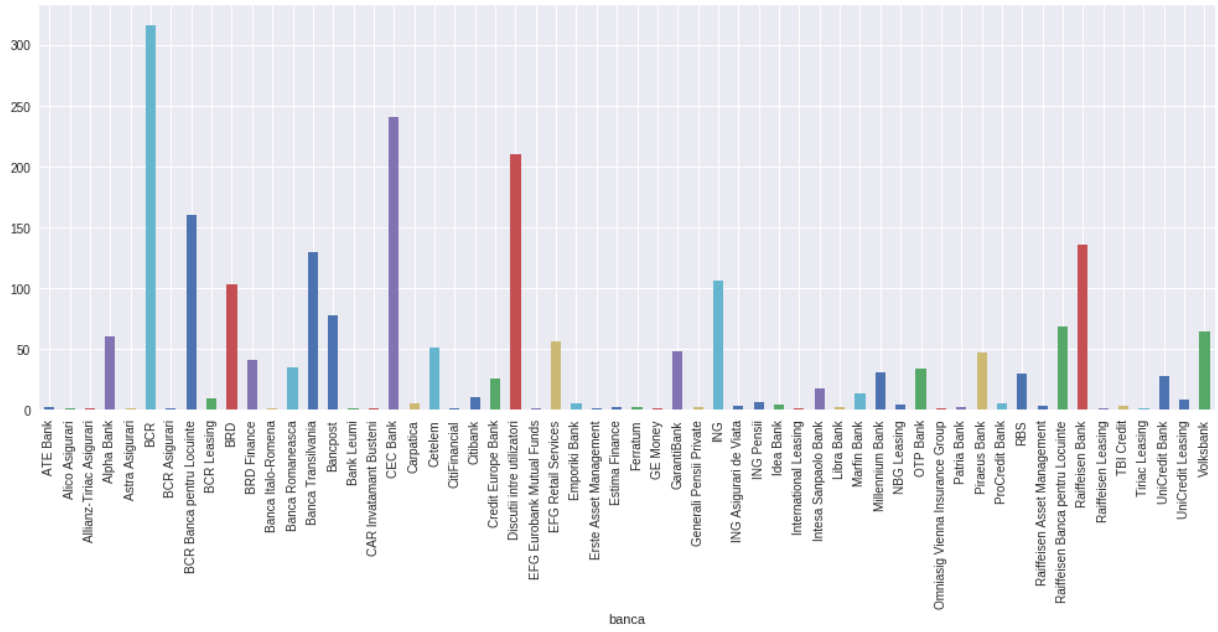

Fig. 6. Distribution of financial banking institutions

3. Conclusions and Further Research

Nowadays, with the proliferation of financial banking reviews, microblogs, forum discussions, blogs, social networks and other forms of expression, more and more customers become aware growth of web's popularity. For instance, before making a decision regarding a financial banking product or a service, customers explore other opinions regarding that

DOI: 10.12948/issn14531305/23.1.2019.04

product or service. [17] Sentiment Classification is successfully used in identification of customer feelings regarding of a product or a service. [18-32] Analyzing the customers feedback's and expectations is a major aid in measuring overall performance, sales and improving financial banking institutions marketing strategies, especially on their online presence.

This paper introduces the research framework for a dataset creation in a financial banking domain in order to be further used in a Supervised Machine Learning context.

To the best of our knowledge, there is no available dataset for sentiment classification of Romanian financial banking customer reviews. The dataset contains 2234 financial banking posts in Romanian language from Conso portal between June 2009 and April 2018. To build the dataset, web scraping technique based on Scrapy framework is used. The extracted information is so vast and it contains not only customers' opinions, but also experiences and facts (e.g. discussions about petitions, legislation modifications, etc.). Therefore, several posts are subjective texts (i.e. "opinionated information") whilst others are objective texts (i.e. "factual information"). Each post in the dataset is annotated as: "Positive", "Negative" or "Neutral"
and "Opinion, Experience or Fact" in order to capture both subjectivity and objectivity of customers' reviews.

As further research, the main objective is to build supervised machine learning classification models for Sentiment Analysis in order to explore opinions insights from financial banking customers. This is particular challenging because the models have to deal with classification of reviews in Romanian language and imbalanced class distribution.

References

[1] Danah M. Boyd, Nicole B. Ellison, "Social Network Sites: Definition, History, and Scholarship", *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, October 2007

[2] Ye, Q., Law, R., Gu, B., & Chen, W., "The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings*", Journal of Computers in Human Behavior*, vol. 27, no. 2, pp. 634-639, March 2011

[3] Cantallops, A. S., and Salvi, F., "New consumer behavior: A review of research on eWOM and hotels", *International Journal of Hospitality Management*, vol. 36, pp. 41-51, January 2014

[4] T. Matthew, *Guide to Web scraping with PHP*. Marco Tabini & Associates Inc., pp. 66-68, 2010

[5] K. M. A. Khan and D. K. Sharma, "Self-adaptive ontology-based focused crawling: A literature survey," in Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 5th International Conference on (pp. 595-601). IEEE, pp. 595-601, 2016.

[6] S. Batsakis, E. G. Petrakis, and E. Milios, "Improving the performance of focused web crawlers," Data & Knowledge Engineering, vol. 68, no. 10, pp. 1001–1013, 2009

[7] Manning, C., Raghavan, P., & Schütze, H., Web crawling and indexes. In Introduction to Information Retrieval (pp. 405-420). Cambridge: Cambridge University Press, 2008

[8] Dimitrios Kouzis-Loukas, *Learning Scrapy - Learn the art of efficient web scraping and crawling with Python*, Packt Publishing, pp. 64-65, 2016

[9] Scrapy official documentation [Online]. Available: https://docs.scrapy.org/en/latest/

[10] Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B., A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. Psychological Methods, vol. 21, no. 4, pp. 475-492, 2016

[11] Erin J. Farley, Lisa Pierotte, An Emerging Data Collection Method for Criminal Justice Researchers [Online]. Available:

DOI: 10.12948/issn14531305/23.1.2019.04

https://www.jrsa.org/jrsa-documents/webscraping.pdf
[12] Wang, J., & Guo, Y., Scrapy-Based Crawling and UserBehavior Characteristics Analysis on Taobao. 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, 2012
[13] D. Kurniawati and D. Triawan, "Increased information retrieval capabilities on e-commerce websites using scraping techniques," 2017 International Conference on Sustainable Information Engineering and Technology (SIET), Malang, pp. 226-229, 2017
[14] Bassam Farooq, Mohd. Shahid Husain, Mohammad Suaib, New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings, International Journal of Advanced Research in Computer Science, vol. 9, no. 2, pp. 64-67, April 2018
[15] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD'04, pp. 168–177, New York
[16] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.," in LREC, 2010, vol. 10, pp. 2200–2204.
[17] I. Raicu, M.C. Turkes, An opinion mining and sentiment analysis approach for evaluating customer satisfaction in a digital banking environment, Annales Universitatis Apulensis Series Oeconomica, 2 (18), 2016
[18] L. A. Cotfas, C. Delcea, I. Raicu, I. A. Bradea and E. Scarlat, "Grey sentiment analysis using SentiWordNet," 2017 International Conference on Grey Systems and Intelligent Services (GSIS), Stockholm, 2017, pp. 284-288.
[19] K. D. Rosa and J. Ellen, "Text Classification Methodologies Applied to Micro-Text in Military Chat," 2009 International Conference on Machine Learning and Applications, Miami Beach, FL, 2009, pp. 710-714.
[20] C. Liu, Y. Sheng, Z. Wei and Y. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), Lanzhou, 2018, pp. 218-222.
[21] Z. Li, W. Shang and M. Yan, "News text classification model based on topic model," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, 2016, pp. 1-5.
[22] F. Miao, P. Zhang, L. Jin and H. Wu, "Chinese News Text Classification Based on Machine Learning Algorithm," 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, 2018, pp. 48-51.
[23] M. H. Moattar, M. M. Homayounpour and D. Zabihzadeh, "Persian Text Normalization using Classification Tree and Support Vector Machine," 2006 2nd International Conference on Information & Communication Technologies, Damascus, 2006, pp. 1308-1311.
[24] Q. Xu, Z. Liu, "Automatic Chinese Text Classification Based on NSVMDT-KNN", 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, pp. 410-414, 2008.
[25] H. Zhuang, C. Wang, C. Li, Q. Wang, X. Zhou, "Natural Language Processing Service Based on Stroke-Level Convolutional Networks for Chinese Text Classification", 2017 IEEE International Conference on Web Services (ICWS), pp. 404-411, 2017.
[26] S. Das and A. Das, "Fusion with sentiment scores for market research," 2016 19th International Conference on Information Fusion (FUSION), Heidelberg, 2016, pp. 1003-1010.
[27] E. Birbeck and D. Cliff, "Using Stock Prices as Ground Truth in Sentiment Analysis to Generate Profitable Trading Signals," 2018 IEEE Symposium Series on

DOI: 10.12948/issn14531305/23.1.2019.04

Computational Intelligence (SSCI), Bangalore, India, 2018, pp. 1868-1875. [28] J. Bollen, H. Mao, X. Zeng, "Twitter Mood Predicts the Stock Market", Journal of Computational Science, vol. 2, no. 1, pp. 1-8, 2011.
[29] N. Oliveira, P. Cortez, N. Areal, "On the Predictability of Stock Market Behavior Using StockTwits Sentiment and Posting Volume", Portuguese Conference on Artificial Intelligence, pp. 355-365, 2013.
[30] X. Zhang, H. Fuehres, P.A. Gloor, "Predicting Stock Market Indicators Through Twitter I hope it is not as bad as I fear", Procedia-Social and Behavioral Sciences, vol. 26, pp. 55-62, 2011.
[31] T. Rao, S. Srivastava, "Analyzing Stock Market Movements Using Twitter Sentiment Analysis", International Conference on Advances in Social Networks Analysis and Mining., 2012.
[32] Ion SMEUREANU, Cristian BUCUR, Applying Supervised Opinion Mining Techniques on Online User Reviews, Informatica Economică vol. 16, no. 2/2012

Irina RAICU has a strong background in computer science. She received her B.S. and M.S. in Business Informatics from the Bucharest University of Economic Studies. Always passionate about research and innovation, in 2013 Irina's research activity was appreciated with a Performance Fellowship from Bucharest University of Economic Studies. At the moment, Irina is a PhD candidate in Business Informatics at Bucharest University of Economic Studies. She has a strong collaboration with Center of Research in Informatics from Paris 1 Pantheon-Sorbonne University. Also, Irina has been involved into several research activities to French Institute for Research in Computer Science and Automation. Along with her research activity, Irina successfully journeyed through various roles in dynamic and most demanding business environments. Currently, she is an Artificial Intelligence Technical Manager in an international company. Her main research interests are Sentiment Analysis, Text Classification, Natural Language Processing and Machine Learning.

DOI: 10.12948/issn14531305/23.1.2019.04

© 2019. This work is published under https://creativecommons.org/licenses/by/4.0/(the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.