Text Analytics and Big Data in the Financial domain

Jean-Pierre Kuilboer
Information Systems and Management Science
University of Massachusetts Boston
Boston, USA
jeanpierre.kuilboer@umb.edu

Tristan Stull
Information Systems and Management Science
University of Massachusetts Boston
Boston, USA
tristan.stull001@umb.edu

Abstract — **this research attempts to provide some insights on the application of text mining and Natural Language Processing (NLP). The application domain is consumer complaints about financial institutions in the USA. As an advanced analytics discipline embedded within the Big Data paradigm, the practice of text analytics contains elements of emergent knowledge processes. Since our experiment should be able to scale up we make use of a pipeline based on Spark-NLP. The usage scenario is adapting the model to a specific industrial context and using the dataset offered by the "Consumer Financial Protection Bureau" to illustrate the application.**

Keywords - Analytics, Topic modeling, CFPB, Decision support system, consumer complaint narratives

I. Background and Literature Review

This research aims at investigating text analytics in the context of big data from an information-systems and business perspective. Academic research in information systems seeks to reason about and deliver effective theory and conceptual mechanisms that are relevant to practice [1][2]. Scholars argue that data science is becoming *the* practice area in information systems. In this sense, this work is practice-oriented and aims at using some of the latest approaches in the field.

Significantly, much of the information that matters most to businesses and other organizations is to be derived from text and unstructured data [3]. Text Analytics (TA) solutions support making sense of large volumes of text by leveraging methods and techniques from several contributing disciplines. Text analysis (or text mining) methods, often termed tasks, apply natural language processing (NLP) research to extract knowledge from text. Examples of tasks include information extraction, opinion mining, classification, summarization, and document or topic modeling [4][5]. Per David Blei, text analysis may be suitable for any field where "texts are a primary object of study" [6]. Considering it as an analytics discipline within the practice of information systems means looking at it from a managerial perspective: asking how various techniques as well as enabling tools and practices are used to "drive decisions and actions" [7]. Since we are not aiming at developing new algorithms, we use a mix of existing software modules and techniques applied to a large dataset {slightly below 2 million complaints}. The dataset covers complaints against financial institutions from 2011 to December 2020. To analyze the data we use Python, Pyspark, and import the Spark-NLP module for text mining. Spark-NLP [8] is currently widely used in natural language processing and is on par with the cluster of modern solutions to NLP (e.g. spaCY, Hugging Face, NLTK, Stanford CoreNLP, Gensim, Allen NLP, Rasa NLU, and others). Since it is based on Spark, it can accommodate very large datasets efficiently.

The Consumer Financial Protection Bureau (CFPB) regulates the offering and provision of consumer financial products or services under the federal consumer financial laws and educates and empowers consumers to make better informed financial decisions. In existence since 2011, it aims at collecting financial consumers' complaints and offers mitigation between them and the institutions. Since CFPB is part of open government, the data is made available to download and analysis for both industry and academic use. While the US-centered dataset could be seen as a weakness, the analysis can easily be extended to other regions or formats. Since 2011, a few researchers have published their insights but much more could be extracted from this moving target given the explosion of new ways to process big data and emerging NLP algorithms. From the early stages, the data has proven useful for longitudinal analysis as it provides both the text of the complaints and rich contextual metadata.

Ayres et al. [9] investigated the emerging CFPB consumer complaints at a company-specific level as well as zip code demographics level, and their analysis could be easily replicated given the growing dataset to provide a longitudinal or location-based analysis. They concluded that banks such as Bank of America, or Citibank had a slower than average response to complaints.

Littwin [10] took a different slant from Ayres et al. (2013) and explored the reasons why government agencies should process the CFPB consumer complaints and whether these reasons justify the resources that complaint processing requires.

In 2017 the CFPB dropped one of the parameters (i.e. note whether the consumer did not accept the response from institutions) as it was deemed difficult to standardize.

Bastani et. al. [11] used a text mining approach based on latent Dirichlet allocation (LDA) to analyze the CFPB consumer complaints. They aimed to extract latent topics in the CFPB complaint narratives and explores their associated trends over time. They also published a visualization on Tableau Public visually demonstrating some of their results.

II. Business Problem and unit of Analysis

The consumer complaints indicate the consumers' dissatisfaction with products and services provided by financial institutions. The data expresses the business problem in terms that make sense for text analytics. The significant fields we use are the data received, the product, the company, the State and zip

2021 16th Iberian Conference on Information Systems and Technologies (CISTI)
23 – 26 June 2021, Chaves, Portugal
ISBN: 978-989-54659-1-0

code, the consumer complaint narrative, date sent to the company, date of the answer, and a few more metadata. The consumer has the opt-out option not to have their complaint published.

The data approach determines the base text corpus including the feasibility and acquisition steps. The data on the CFPB website is relatively clean since the user uses a template with clear attributes to submit the complaints. The loading process only needs rudimentary steps to prepare the dataset for use. The dataset is available in both CSV and JSON formats. The unfiltered set involves about 2 million consumer complaints collected from 2011 to 2020. For efficiency purposes, we transformed the dataset from a CSV format to be stored by Spark in parquet columnar format, to be more amenable to Spark and big data processing. We also cast the various date from a string to a date format to enable date operations such as grouping by day, month, year, or evaluate a period for company responses.

The processing model includes determining the workflow/pipeline model, the user roles, and computational implications. The first issue to be tackled is to match the text analysis method with the task in the context of the problem. Next, a pipeline might be considered in terms of how the workflow should be set up, what pre-processing methods apply, and whether interactivity or iterative behaviors are indicated. We used a snapshot of the dataset as of December 2020 but the collection is ongoing.

Finally, the processing model's implications for computation are addressed. Blei et al. [12] developed the well-known latent Dirichlet allocation model (we did a simple topic modeling [using LBA] in our analysis of 20 companies having the most complaints).

The User-Interface and Pipelined approach entails user interface design for processing. McKelvey et al. leverage the dashboard metaphor to combine key interface elements in their analysis of text analytics use cases across a range of disciplines with a common theme of 'investigation' [13]. While not an application example, Oslhannikova et al. help make the case for visualization's place of prominence in big data analytics by charting the field and focusing on the problem of human cognition in problems addressing large data volumes [14]. We used visual topic modeling and wordmap in our supplementary analysis of companies' SEC reporting 10Q/K.

We used an Anaconda virtual environment platform and Python notebooks, Pyspark 2.4.7, the John Snow Spark-NLP, and some Matplotlib to graph some charts of complaints per day, month, and year.

The next section describes how text analytics relates to big data.

III. Text Analytics and Big Data

Text Analytics is related to big data in several ways. Firstly, it is embedded with the larger context of big data analytics. Text analytics' raison d'etre stems from the volume: situations in which the amount of text is too large to be processed usefully without automation support. Typical text analytics problems involve corpora in the order of millions, hundreds of millions, or even billions of text documents [15][16][17]. Text analytics is also considered a sub-practice of big data analytics. It is sometimes referred to as a 'method' of big data analytics [18], or a 'technique' [19] [20]. Data mining methods such as classification, clustering, and information retrieval have their versions in text analytics. Big data processing techniques such as MapReduce and parallel processing could also leverage text analytics applications. Text analytics may, therefore, be considered a sub-discipline of both data mining and big data analytics, respectively.

Text Analytics differs from non-text big data analytics in two important ways. Firstly, NLP itself is an evolving scientific discipline. Its core, foundational knowledge is actively being determined. At the risk of overgeneralizing, we can say that the science of how to make sense of large volumes of text is far from being a 'solved' problem. The techniques for processing structured data are much better understood than for text. There are many open issues concerning fundamental methodological and technical matters such as what dimensionality reduction technique is appropriate, and different algorithmic approaches. Secondly, there is no comprehensive guideline available for how to compose a text analytics solution. The john Snow environment itself includes more than 700 pre-trained models and more than 450 pipelines.

Complexity put the development of new text analytics solutions outside the reach of many, if not most companies. While large organizations in the finance industry are retooling along the line of the most advanced technology companies, most normal companies do not have the capabilities to developed advanced text mining, AI, and such applications. To make text-mining analytics more accessible to the masses academic research has a role to play. On one hand, many universities are developing analytics and data science programs to educate new cohorts of students. Along with research, academia can publish research outside the realm of the restricted internal groups of the financial industry. Examples of advanced closed systems in banking include projects at JP Morgan, Morgan Stanley, or Fidelity but the canvas of tools and techniques used by these teams, although available, is unlikely to be known outside the firms. Academia can mimic such workflows using open source tools and libraries and reach a good level of understanding.

IV. Some of the Results

Since the corpus of the complaints is in English, we borrowed the stopwords from the common NLTK library. From the John Snow repository, we created a pipeline using some of the Spark-NLP libraries including documentAssembler, tokenizer, normalizer, lemmatizer, stopwords_cleaner, finisher.

The DocumentAssembler is a transformer from the spark-nlp base. Once we pass the data through the following processing steps of tokenization, normalization, lemmatization, and English stopwords cleaner, the documents are in a relatively clean form. The last steps is to use a finisher annotator. Once we have our NLP pipeline ready to go, if we want to use our annotation results somewhere else where it is easy to use, the Finisher outputs annotation(s) values into string that can be easy to input to other processes. Fig. 1 illustrates our pipeline.

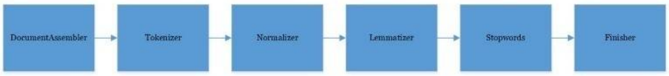

Figure 1. NLP Processing Pipeline

From our data cleaning preliminary process, we loaded the complaints about the 20 largest offenders in the domain and run text analysis for a few of the largest. Among the largest, some of the names are well-recognized organizations in the field such as Equifax, Experian, Transunion, followed by some front-runners of the banking industry such as Bank of America, Wells Fargo, JP Morgan, Citibank, Capital One, PNC Bank, or American Express.

In addition, to be noticed is that the top three in terms of the number of complaints are the three main credit reporting agencies in the USA. Many of the complaints originate from discrepancies about the perceived views on credit, consumers being refused credit, length of bad credit lingering in consumer files. Some of second-tier on the top 20 most complained are very large banks. Since we did not normalize for bank size nor geographic coverage, the large banks could be expected to have a larger number of complaints. On the other hand, they should also have better processes and should be expected to have fewer complaints if they adhered to their stated mission. The third group of financial institutions on the top 20 were non-bank organizations involved in the financial market such as Navient, which specialized in students' debt and the common problem in the USA for students carrying large debt for a long time. Another group is exemplified by mortgage and refinancing companies such as Ocwen Financial that deals with lots of problem mortgages or foreclosures during financial downturns.

When running tf-idf or trying to find original wording in the complaint corpus some of the results were tainted by wrong spelling on many of the complaints. For example, consumers may have mistyped bac for bank or Wel for Wells, etc...

Overall a longitudinal analysis of the complaints still uncovers a significant rise in complaints during the recent years [but this could be tainted by the fact that more consumers may be aware of available reporting avenues, rise in Internet search, etc].

Aside, we also run Machine learning, document summarization, and wordcloud experiments [21] on the SEC filing for the same firms subject to most complaints available in machine-readable format {e.g. PDF}. We found consistent incongruence between the stated values such as respect, customer focus, quality of life, diversity, inclusion, or integrity and the assertions in the complaints. As expected the 10K or 10Q filings targeted a projection of positive images aimed at investors and the financial market and in sharp contrast to the image projected by the complaints filed against them. Some of the hurdles encountered in this part of our experiments were for firms not filing on the SEC market, based-offshore, or having documents not amenable to analysis.

We then run TF-idf for some of the firms to see what would distinguish the top firm in the sample in terms of vocabulary used in the complaints and classified them by theme.

V. Conclusion and Limitations

This short research aims to show that the text analytics paradigm can be employed to address the problem of consumer complaints in the financial sector and provides an outline of how it could be applied. Additional artifacts could be needed to carry out the research. Artifacts that may be created, assembled, or accessed in the course of the research include template workflows, best practices libraries, and IT-specific elements such as development methods, toolsets, and product descriptions. During the experimentation, some of the constraints emerged such as software run being tied to a particular version of the software in the modules {e.g. Python 3.6, Spark 2.4.7, Spark-NLP 2.7.3 } while not limiting at a point in time it can restrain the ability of other researchers to replicate the results. Early April 2021, we updated the various components t {Python 3.8, Spark 3.1.1, and Spark-NLP 3.0.1 } and we are in the process of regression testing the various portion of our code. Another bottleneck could arise if one module along the pipeline does not evolve in harmony with the others or some dataset is discontinued. While Spark-NLP was state of the art by mid-2020 it could be superseded by a better solution soon thereafter.

References

[1] Benbasat, I., & Zmud, R. W. (1999). Empirical research in information systems: The practice or relevance. MIS Quarterly, 23(1), 3–16.

[2] Davenport, T. H., & Markus, M. L. (1999). RIGOR vs. RELEVANCE REVISITED: RESPONSE TO. MIS Quarterly, 23(1), 19–23.

[3] Sabherwal, R., & Becerra-Fernandez, I. (2011). Business intelligence: Practices, technologies, and management. John Wiley & Sons.

[4] Jiang, H., Lin, P., & Qiang, M. (2016). Public-Opinion Sentiment Analysis for Large Hydro Projects. Journal of Construction Engineering and Management, 142(2).

[5] Aggarwal, C. C., & Zhai, C. (2012). Mining text data. Springer Science & Business Media.

[6] Blei, D. M. (2012). Probabilistic topic models. Commun. {ACM}, 55(4), 77–84.

[7] Liberatore, M. J., & Luo, W. (2010). The analytics movement: Implications for operations research. Interfaces, 40(4), 313–324.

[8] John Snow Labs - https://www.johnsnowlabs.com/ accessed February 19 - 2021.

[9] Ayres, I., Lingwall, J., and Steinway, S. (2013). Skeletons in the Database: An Early Analysis of the CFPB's Consumer Complaints. Fordham J. Corp. & Fin. L.,19,343.

[10] Littwin, A. K. (2015). Examination as a Method of Consumer Protection. Temple Law Review, 87(807).

[11] Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems with Applications, 127, 256-271.

[12] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993–1022

[13] McKelvey, K., Rudnick, A., Conover, M. D., & Menczer, F. (2012). Visualizing communication on social media: Making big data accessible. ArXiv Preprint ArXiv:1202.1367.

[14] Olshannikova, E., Ometov, A., Koucheryavy, Y., & Olsson, T. (2015). Visualizing Big Data with augmented and virtual reality: challenges and research agenda. Journal of Big Data, 2(1), 22.

[15] Sukhija, N., Tatineni, M., Brown, N., Van Moer, M., Rodriguez, P., & Callicott, S. (2016). Topic modeling and visualization for big data in

social sciences. In 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (pp. 1198–1205).

[16] Newman, D., Smyth, P., & Steyvers, M. (2006). Scalable parallel topic models. Journal of Intelligence Community Research and Development, 5.

[17] Gruhl, D., Chavet, L., Gibson, D., Meyer, J., Pattanayak, P., Tomkins, A., & Zien, J. (2004). How to build a WebFountain: An architecture for very large-scale text analytics. IBM Systems Journal, 43(1), 64–77.

[18] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144.

[19] Choi, T.-M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. Production and Operations Management, 27(10), 1868–1883.

[20] Chiang, R. H. L., Grover, V., Liang, T.-P., & Zhang, D. (2018). Strategic Value of Big Data and Business Analytics. Taylor & Francis.

[21] Lookabaugh, B., Puri, S., Tatsat, H. (2020). Machine Learning and Data Science Blueprints for Finance. United States: O'Reilly Media