Are fintechs really a hype? A machine learning-based polarity analysis of Brazilian posts on social media

Marina Ponestke Seara
Specialization in Big Data & Analytics
Pontifícia Universidade Católica do Paraná
Curitiba, Brazil
Email: marina.seara@icloud.com

Andreia Malucelli, Altair Olivo Santin, and Jean Paul Barddal
Programa de Pós-Graduação em Informática (PPGIa)
Pontifícia Universidade Católica do Paraná
Curitiba, Brazil
Email: {malu, santin, jean.barddal}@ppgia.pucpr.br

***Abstract*—Fintechs are technology companies that, in contrast to traditional banks, are engaged in digital solutions for payment, money transfers, and real-time notifications. Taking advantage of digital means of communication, most of the service interactions between fintechs and customers occurs via chats or posts in social media. In this work, our goal is to use machine learning to analyze these posts and identify what are the terms used by customers to express positive, neutral and negative customer experiences. During this analysis, we assess the following questions using data from the 3 biggest fintechs in Brazil: (i) what are the most commented topics on social media regarding fintechs, (ii) what are the words more often used by customers to express positive, negative and neutral reactions to the customer service obtained; and (iii) what kind of machine learning model should a fintech use to automatically identify whether a post is positive, negative or neutral.**

I. Introduction

Fintechs are technology companies that act in the financial sector and provide services to customers without the need for physical presence [1]. In contrast to traditional banking systems, fintechs are much more engaged in providing digital payment solutions, contact-less payments, and real-time notification to customers, and are dragging millions of users away from the traditional banking systems.

With this shift of customers, the market share of fintechs has increased rapidly. In the Brazilian context, hundreds of fintechs have been created in the last years, and they offer a diversity of products and services that were previously exclusive to traditional banks, such as credit cards, payment services, investments, and many others. In Brazil, fintechs are often start-ups, and their teams are usually allocated to the development of financial tasks, e.g., transactions, and customer service, e.g., chats and interaction on social media to address customers' needs. As time passes, the amount of data that is collected from different sources about fintechs, i.e., databases, social media, and the web in general, scales up quickly, and their aggregation is an example of *big data*. Naturally, extracting useful information from these massive collections of data is of utmost importance for fintechs, as they can identify which services require special attention and target them accordingly.

In this work, our goal is to use machine learning techniques to extract and analyze the polarity of publicly available posts about the 3 biggest fintechs in Brazil. During this analysis, we wish to identify what drives customers to provide positive, neutral and negative feedbacks w.r.t. the services provided by these companies. More specifically, we intend to tackle the following questions:

- What are the most commented topics on social media regarding fintechs in Brazil?
- What are the words most often used by customers when they wish to express positive, negative or neutral reactions to the experience they have with fintechs?
- What kind of machine learning model should a fintech/bank use to automatically identify whether a post is positive, negative or neutral?

All of the aforementioned questions are of the utmost importance as fintechs wish to discover which of their services are well-rated or not, and target the latter swiftly so that customer experience is improved and customers are retained and do not diverge back to the traditional banking systems.

As contributions of this work, we cite the following:

- A publicly available dataset that contains polarity-labeled posts about Brazilian fintechs,
- The creation of a dictionary of stop-words in Brazilian Portuguese that has been constructed to target fintech-related datasets,
- An analysis of different classifiers on the automatic polarity identification task.

This paper is divided as follows. Section II discusses Fintechs and their impact on the financial sector. Section III briefly surveys related work on polarity inference. Section IV introduces our approach for this problem, which is later assessed in Section V. Finally, Section VI concludes this paper and states envisioned future work.

978-1-5386-4829-2/18/$31.00 ©2018 IEEE

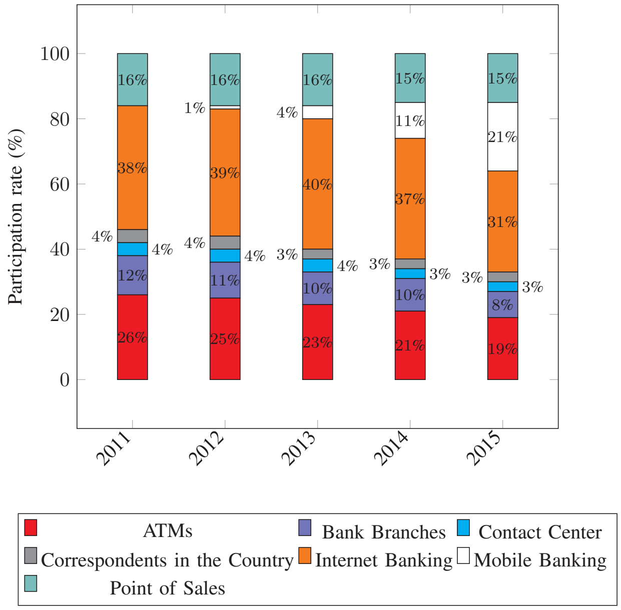

Fig. 1. Participation rate of different banking channels in Brazil. Adapted from https://rct.dieese.org.br/index.php/rct/article/view/143/pdf_1.

II. Fintechs

Fintechs are technology companies that provide financial-related services to customers without the need of physical presence[1], e.g., branches and ATMs; as interactions are made via digital channels, such as social networks, websites or mobile applications [1].

According to the latest reports on fintechs[2,3], digital payment solutions, contactless payments, digital money transfers and real-time notifications are the main advantages fintechs have in comparison to traditional financial companies.

In the Brazilian context, the impact of fintechs is easily noticeable. For instance, a recent survey published by the Brazilian Bank Association (FEBRABAN)[4] showed that approximately 400 fintechs are now working in Brazil, and these offer a diversity of products and services that were previously exclusive to traditional banks, such as credit cards, payment services, investments, and many others.

Figure 1 depicts the participation rates of different financial channels in Brazil between the years of 2011 and 2015. In the participation rates, it is clear that fintech-related channels, such as internet and mobile banking rates are swiftly increasing, where the latter increased from 1% to 21% in a short timespan of 4 years. On the other hand, traditional channels like ATMs and branches observe decreases in their participation rates. Together, these results show that the market shifts towards fintechs and their services in an increasing rate.

Another important trait of fintechs is that they use social networks to both advertise and obtain feedback from customers. Therefore, it is of the utmost importance for these companies to swiftly and accurately identify positive, neutral, and negative posts so that these can be analyzed to identify which parts of the customer service are successful and which deserve further attention. To achieve this goal, this study conducts an initial analysis of the polarity of posts of the 3 biggest fintechs in Brazil. In practice, we will show how data acquisition, wrangling, and text mining techniques were used to determine how the polarity of social media posts can be automatically extracted with reasonable accuracy. As a result, both fintechs and banks will be able to analyze what drives customers to have positive, negative and neutral considerations about their financial services.

III. Related Works

The number of works that combine polarity analysis with machine learning and fintechs is scarce. Therefore, in this section, we survey existing works that propose the use of big data and machine learning techniques in traditional banking systems while highlighting those that tackle polarity and emotion identification in social media.

First, authors in [2] survey the relevance of different big data approaches to the financial sector, while also outlining the current challenges for adoption and gaps that are yet to be fulfilled by the technology companies and researchers.

Regarding sentiment and polarity analysis, authors in [3] survey the main approaches that use machine learning and highlight that this is still an open field for research. Related to the banking sector, the work of [4] proposes an approach to assess the polarity of text documents from the Central Bank of Italy by combining text mining techniques and Support Vector Machine (SVM) classifiers. Similarly, artificial neural networks have been used in [5] to quantify and predict customer satisfaction and credit scoring in a German bank.

Also, a considerable amount of effort has been put on the usage of classification systems to extract sentiments and polarities from different social media, including Twitter [6], Facebook [7], [8] and web pages in general [9], [10], [11].

Finally, it is important to emphasize that even though a fair amount of research has been put on polarity analysis from social media, they often do not target financial services or investigate posts written in languages other than English. Therefore, our paper tackles these gaps by analyzing posts on social media written in Portuguese on the scope of financial services provided by fintechs.

IV. Proposed Method

In this section, we detail the steps taken during the processes of (i) data acquisition, (ii) its preprocessing, and (iii) the creation of the machine learning model for automatic polarity identification. The code representing these steps and dataset constructed can be found at https://github.com/jpbarddal/fintechs-polarity-analysis.

[1] https://rct.dieese.org.br/index.php/rct/article/view/143/pdf_1

[2] https://www.efma.com/study/detail/26003

[3] https://www.capgemini.com/service/the-world-fintech-report-2017-key-topics/

[4] https://issuu.com/revistaciab/docs/pesquisa_febraban_de_tecnologia_ban_547cac9b8acb07

TABLE I
DETAILS OF THE DATASET AFTER THE PREPROCESSING STEP. THE ACTUAL NAMES OF FINTECHS WERE REDACTED TO PRESERVE THE ANONYMITY OF THE COMPANIES.

Fintech	Period	# of positive posts	# of neutral posts	# of negative posts	# of unlabeled posts
Fintech A	01-Mar-2016 to 30-Oct-2017	143	250	276	6258
Fintech B	01-Mar-2016 to 30-Oct-2017	255	129	189	3466
Fintech C	01-Jan-2016 to 30-Mar-2017	294	235	171	11574
Total	*01-Jan-2016 to 30-Oct-2017*	*592*	*614*	*636*	*21298*

Data acquisition. To gather posts about fintechs, we used NetVizz [12] to extract all the public posts available in Facebook pages of the fintechs between January 1st of 2016 and October 30th, 2017 from the pages of the 3 major fintechs in Brazil. For the sake of anonymity, all users' names and names of the fintechs have been removed from the text corpus and are redacted in this work. Therefore, these fintechs are hereafter referred to as Fintechs A, B, and C. As a result, a total number of 38,513 posts were obtained, yet, only 1,842 have been labeled as positive, negative and neutral w.r.t. customer satisfaction. Here, we clarify that instances have been manually labeled by a banking expert that has Brazilian Portuguese as the first language. Even though having the entire dataset labeled was desired, the corpus was too massive to be analyzed by a single person in a feasible time.

Preprocessing. During this step, the dataset was analyzed and cleansed. First, API-related variables were removed, and these include: post (*post_id*) and user (*post_by*) indices, timestamp of the post (*post_published*), a flag that determines whether the post is a reply to another publication (*is_reply*), and the number of likes that post received (*comment_like_post*). Next, all special characters, accentuation, numbers, URLs, emoticons were dropped. Finally, both blank posts and texts created by the page owner (the fintech), were also removed, as the latter would be biased towards the positive polarity, while the former would not bring any insights to the analysis.

After the cleansing of the dataset, the Natural Language Toolkit (NLTK) [13] was used to allow the removal of Portuguese stop-words. At this point, it is important to emphasize that the number of stop-words available in NLTK is scarce compared to other languages, i.e., English, and thus, all the labeled posts have been manually analyzed to identify new stop-words which should be removed in conjunction with the default ones provided by NLTK. The actual stop-words can be found in the project repository, but some examples include slangs and acronyms common in posts from Brazilians in social networks, e.g. "vc", "tb", "tbm", "kkkk", "hahaha" and "nuss". Next, a traditional stemming process that has been performed with the *RLPSStemmer* method and the corpus was tokenized with the bag-of-words strategy. Table I details the main characteristics of this dataset after the preprocessing step and shows the number of positive, negative and neutral posts per fintech.

Classifiers. Given the cleansed dataset obtained during the previous steps, three classifiers were tested with the goal of learning a predictive model that can determine whether a post is positive, neutral or negative w.r.t. customer satisfaction. Precisely, the Multinomial Naive Bayes (NB), Decision Tree (DT) and Random Forest [14] (RF) classifiers provided by NLTK and scikit-learn [15] were tested with their default configurations.

Evaluation and Validation. Finally, the classifiers mentioned above were validated in a 10-fold stratified cross-validation scheme provided by scikit-learn. Since the overall number of positive, neutral and negative posts is similar to the entire dataset and folds due to stratification, classification rates would not be biased towards any specific class, and thus, evaluation has been conducted using accuracy. Results are reported using a box-plot to allow the visualization of the variance of the classification rates per classifier. Lastly, the results obtained are compared using a combination of Friedman and Nemenyi's statistical tests following the protocol proposed in [16].

V. ANALYSIS

In this section, we answer the main questions brought up in the introduction. First, we analyze the results obtained by the classifiers tested, and highlight which one should be used for the polarity prediction task in the fintech domain. Next, we discuss which are the most common terms (words and sentences) used by customers for each of the fintechs. Finally, we highlight which terms are most often used when customers wish to express positive, negative and neutral reactions w.r.t. their customer experience with any of the fintechs studied.

Classification results. In Figure 2, we can see that NB (74.32%) is, in average, the best performing method when compared to DT (69.49%) and RF (68.57%) classifiers. To determine whether the difference amongst these methods is significant, a combination of Friedman and Nemenyi statistical tests was used, and as a result, NB was found to be superior to the others with a 95% confidence level (Figure 3). These results, when associated with the smaller computational resources consumption and the ability to pinpoint which words (terms) are essential for polarity prediction, make NB the most appropriate classifier for this task.

Now, working under the assumption that NB is the most accurate classifier, we use it to obtain estimates of the polarities of posts for each fintech. The results are reported in Figure 4. First, it is important to highlight the number of neutral posts across all fintechs, which go from 28% (Fintech C) to nearly 70% (Fintech A). Next, analyzing the results for Fintech A, we can see that the number of posts that were labeled as positive

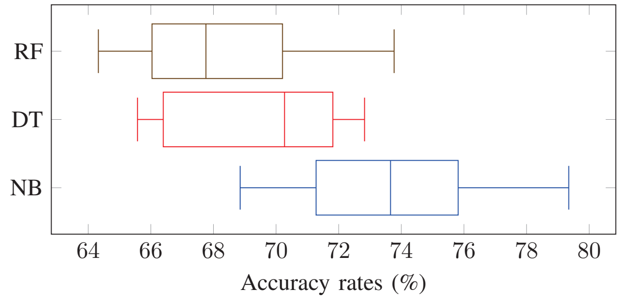

Fig. 2. Comparison of the accuracy rate (%) obtained for each of the tested classifiers (NB = Multinomial Naive Bayes, DT = Decision Tree, and RF = Random Forest).

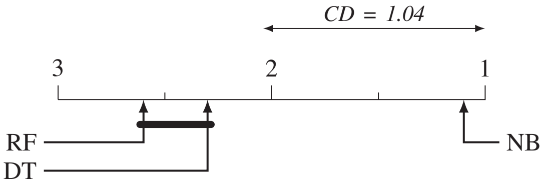

Fig. 3. Results of Friedman and Nemenyi tests. CD stands for the "critical distance" calculated following Nemenyi's statistic. Bars between classifiers highlight pair of classifiers that behave similarly, it is, without significant statistical difference.

and negatives is similar. In contrast, the results for Fintech B show that we have many more negative posts rather than positive ones, and this becomes even more evident when one checks the results for Fintech C.

Furthermore, given the percentages of neutral (39.55%) and negative (34.14%) posts, we argue that despite the interest of Brazilians customers to use the new digital banking processes provided by fintechs, there is still the need for fintechs to double-check their processes as customers are not, in majority, happy with the services provided.

On the other hand, fintechs should continue to obtain and extract information from posts from social networks, as they are widely used by customers as a channel to interact with each other, and more importantly, to provide feedback to the company publicly.

Most common terms. The results for the most common terms (words and sentences) used by customers (Figures 5 through 7) show that the services offered by fintechs are in an early maturity stage. For Fintech A (Figure 5), the volume of posts classified as positive (28.83%) and negative (27.81%) are close. This indicates that their customers express that they enjoy having banking services being provided by fintechs ('aprovadas contas parabens' - '*approved accounts congratulations*', 'aprovadas contas surpresa' - '*approved accounts surprise*'). The main examples are: (i) the waiting times for account openings ('favor aprovar' - '*please approve*', 'liberar contas' - '*please unlock accounts*'), (ii) doubts about how to find out the current balance and limits for purchases ('aprovadas hoje saber limite cartao compras saques' - '*approved today know card limit purchases withdrawals*'), and

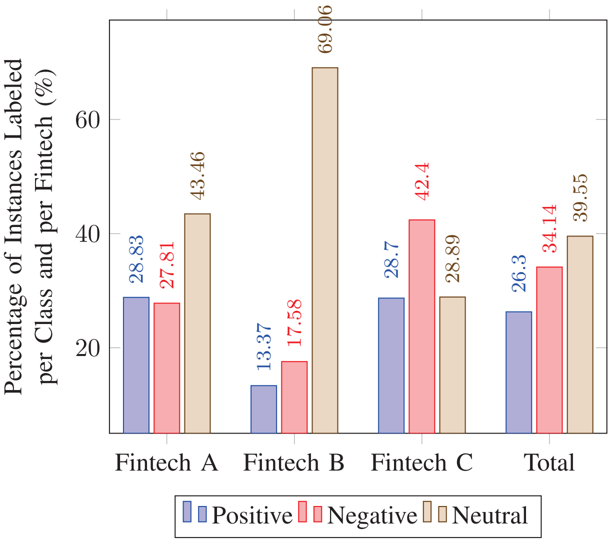

Fig. 4. Prediction rates (%) for Naive Bayes predictions on unlabeled data.

(iii) requests about new services as pre-paid mobile phone top-ups ('aprovadas hoje falta recarga celular' - '*approved today lacks mobile phone top-up*'). In addition to these, it is also clear that the longer waiting times in customer service and lack of response by the fintech ('entrar contato comigo inbox' - '*reach out to me inbox*') are terms widely found inside the corpus.

Regarding the results for Fintech B, reported in Figure 6, we see that 69.06% of the posts were classified as neutral, another 17.85% as negative, and only 13.37% as positive. With almost 87% of the total number of posts expressing negative and neutral feedbacks, it becomes clear that the products and services provided by this fintech are in an embryonic stage, which indicates the importance of textual analysis of this fintech to the greatest points of pain felt by customers. Focusing on the actual terms, the biggest complaints about the services of this fintech are on attempts to contact the fintech in many ways but not receiving the expected feedback ('nao tive nenhuma assistencia' - '*no assistance*', 'quero resposta esperando semanas horas demoradas' - '*need answer waiting weeks hours long*', 'quero convite responde' - '*want invitation answer*'). For instance, (i) problems in the registration and PIN code definition ('esqueci senha' - '*forgot password*'), (ii) issues when downloading mobile applications ('recebi convite nao consigo abaixar aplicativo celular ajuda' - '*received invitation cannot download app help*'), and (iii) problems when completing the enrollment process for an account ('mandei mensagem tentando resolver baixo app tento colocar' - '*sent message trying solve download app try use*') are the most cited topics for this fintech.

Finally, Figure 7 shows the impressive number of negative posts (42%) compared to positive and neutral posts (approximately 29% for both). Term-wise, the biggest complains target: (i) delays in the account opening processes ('nao consigo re-

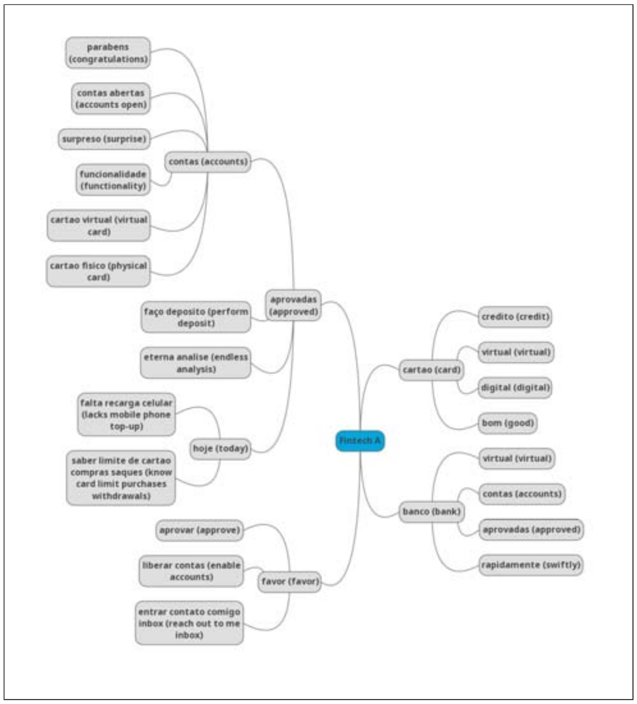
parabens (congratulations)
contas abertas (accounts open)
surpreso (surprise)
contas (accounts)
funcionalidade (functionality)
cartao virtual (virtual card)
cartao fisico (physical card)
aprovadas (approved)
faço deposito (perform deposit)
eterna analise (endless analysis)
falta recarga celular (lacks mobile phone top-up)
hoje (today)
saber limite de cartao compras saques (know card limit purchases withdrawals)
aprovar (approve)
liberar contas (enable accounts)
favor (favor)
entrar contato comigo inbox (reach out to me inbox)
Fintech A
credito (credit)
virtual (virtual)
cartao (card)
digital (digital)
bom (good)
virtual (virtual)
contas (accounts)
banco (bank)
aprovadas (approved)
rapidamente (swiftly)

Fig. 5. Most used terms for Fintech A. Terms are reported in Portuguese along with their translation to English in parentheses.

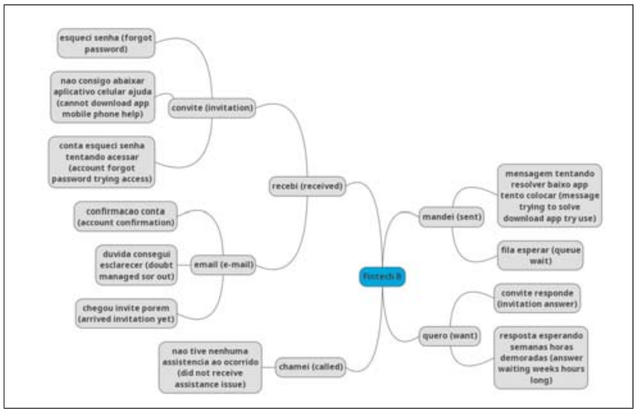
esqueci senha (forgot password)
nao consigo abaixar aplicativo celular ajuda (cannot download app mobile phone help)
convite (invitation)
conta esqueci senha tentando acessar (account forgot password trying access)
recebi (received)
confirmacao conta (account confirmation)
duvida consegui esclarecer (doubt managed sor out)
email (e-mail)
chegou invite porem (arrived invitation yet)
nao tive nenhuma assistencia ao ocorrido (did not receive assistance issue)
chamei (called)
Fintech B
mensagem tentando resolver baixo app tento colocar (message trying to solve download app try use)
mandei (sent)
fila esperar (queue wait)
convite responde (invitation answer)
quero (want)
resposta esperando semanas horas demoradas (answer waiting weeks hours long)

Fig. 6. Most used terms for Fintech B. Terms are reported in Portuguese along with their translation to English in parentheses.

alizar processo acesso conta' - '*cannot perform process access account*', 'nao chegou' - '*did not arrive*'), (ii) difficulties on unlocking credit cards ('nao consigo desbloquear' - '*cannot unlock*'), (iii) bugs in the mobile apps ('nao consigo finalizar cadastro' - '*cannot finish enrollment*', 'nao consigo entrar conta' - '*cannot access account*', 'nao consigo abrir nada pagina' - '*cannot open page*'), and (iv) long waiting times when reaching out to the customer services of the fintech ('nao recebi email banco' - '*did not receive email bank*', 'nao recebi resposta telefonei' - '*did not receive answered I called*').

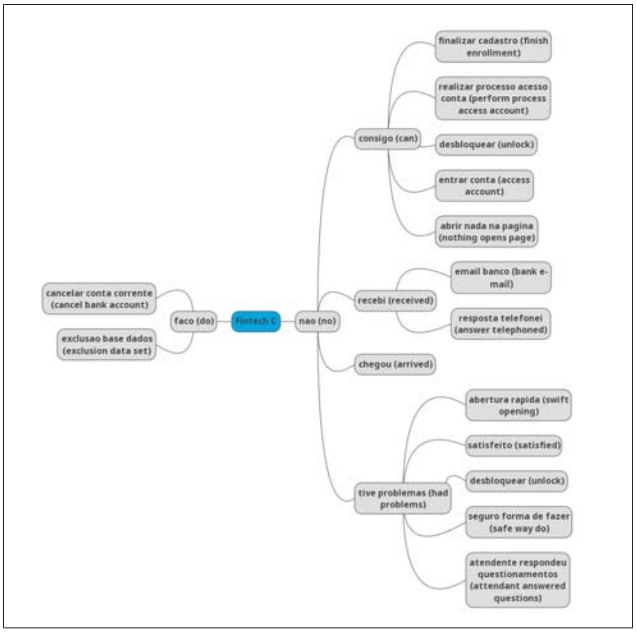
finalizar cadastro (finish enrollment)
realizar processo acesso conta (perform process access account)
consigo (can)
desbloquear (unlock)
entrar conta (access account)
abrir nada na pagina (nothing opens page)
cancelar conta corrente (cancel bank account)
faco (do)
Fintech C
nao (no)
exclusao base dados (exclusion data set)
email banco (bank e-mail)
recebi (received)
resposta telefonei (answer telephoned)
chegou (arrived)
abertura rapida (swift opening)
satisfeito (satisfied)
desbloquear (unlock)
tive problemas (had problems)
seguro forma de fazer (safe way do)
atendente respondeu questionamentos (attendant answered questions)

Fig. 7. Most used terms for Fintech C. Terms are reported in Portuguese along with their translation to English in parentheses.

Main terms and polarity. During the Naive Bayes's (NB) training step, the most informative features (words) were extracted and these are presented in Table II. The words 'nothing', 'open', 'can', 'client' and 'no' were found by NB to be terms with a negative connotation, while 'beautiful', 'Fintech C', and 'congratulations' were highlighted as positive ones. For the words 'no', followed by the terms 'can' and 'open', which are used by consumers to express dissatisfaction w.r.t. the account opening processes.

Another point that requires attention is the name of 'Fintech C' that is associated with a positive polarity, followed by the 'beautiful' and 'congratulations' terms. This shows that customers, when using the name of this specific fintech, tend to express positive feedback, which is an interesting behavior.

VI. Conclusion

In this work, we tackled the problems of scraping, wrangling and mining fintech-related customer posts from social networks. These posts were then processed to construct a corpus in Brazilian Portuguese, which was used to induce machine learning models for polarity identification. The rationale behind all of the latter steps is that fintechs and banks would be able to automatically analyze the posts on social media, allowing these companies to have insights about the acceptance or rejection of newly created products and solutions as they are offered to customers.

The accuracy rate of 74% obtained during this analysis shows that the proposed method is feasible, and would be further improved as fintechs receive feedback from customers every day, thus enriching their dataset for future and more complex analyses.

TABLE II

MAIN TERMS AND THEIR RELATION TO POSITIVE (P), NEGATIVE (N) AND NEUTRAL (X) POLARITIES. TERMS ARE PRESENTED IN PORTUGUESE, AND THEIR TRANSLATION IN ENGLISH IS PROVIDED IN PARENTHESES. (*) REPRESENTS THE CASE WHERE THE NAME OF FINTECH C HAS BEEN HIGHLIGHTED BY THE ALGORITHM, YET, THE ACTUAL NAME HAS BEEN OMITTED FOR THE SAKE OF ANONYMITY.

Words	Percentages	Meaning
Nada (nothing)	(n) : (x) = 57%	The word “nada” has 57% chance of implying (n) rather than (x)
Abrir (open)	(n) : (x) = 33%	The word “abrir” has 33% chance of implying (n) rather than (x)
Consigo (can)	(n) : (x) = 29%	The word “consigo” has 29% chance of implying (n) rather than (x)
Lindo (beautiful)	(p) : (x) = 24%	The word “lindo” has 24% chance of implying (p) rather than (x)
Fintech C (Fintech C) - **(*)**	(p) : (x) = 22%	The word “Fintech C” has 22% chance of implying (p) rather than (x)
Parabéns (congratulations)	(p) : (n) = 20%	The word “parabéns” has 20% chance of implying (p) rather than (n)
Cliente (client)	(n) : (x) = 19%	The word “cliente” has 19% chance of implying (n) rather than (x)
Não (no)	(n) : (x) = 19%	The word “não” has 19% chance of implying (n) rather than (x)

As future works, we envision the following: (i) an analysis of emoticons, emojis, and hashtags as these may provide important insights and correlations to the polarity of post, (ii) a more in-depth treatment of the vocabulary available in these posts, as they are quite informal and replete with slangs that differ from different regions in the country, (iii) introduce an automatic grammar and spell checks in the process, (iv) the handling of names, as they appear very often in posts; (v) the testing of different feature extraction techniques for text, such as deep learning convolutions, Word2Vec [17] and Hashing Tricks [18], (vi) verifying if automatic approaches like autoML [19] can significantly improve the results obtained here; and finally (vii) the implementation of the proposed flow in a big data setting so that the massive amount of data available in social media could be mined.

REFERENCES

[1] E. B. Liarte, “Radiografía del fintech - clasificación, recopilación y análisis de las principales startups,” Master’s thesis, Universitat Politècnica de Catalunya, 2016.

[2] J. Q. Trelewicz, “Big data and big money: The role of data in the financial sector,” *IT Professional*, vol. 19, no. 3, pp. 8–10, 2017.

[3] H. Kaur, V. Mangat, and Nidhi, “A survey of sentiment analysis techniques,” in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Feb 2017, pp. 921–925.

[4] G. Bruno, “Text mining and sentiment extraction in central bank documents,” in *2016 IEEE International Conference on Big Data (Big Data)*, Dec 2016, pp. 1700–1708.

[5] P. S. Patil and N. V. Dharwadkar, “Analysis of banking data using machine learning,” in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Feb 2017, pp. 876–881.

[6] F. Bravo-Marquez, E. Frank, and B. Pfahringer, “Building a twitter opinion lexicon from automatically-annotated tweets,” *Know.-Based Syst.*, vol. 108, no. C, pp. 65–78, Sep. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2016.05.018

[7] V. Franzoni, Y. Li, and P. Mengoni, “A path-based model for emotion abstraction on facebook using sentiment analysis and taxonomy knowledge,” in *Proceedings of the International Conference on Web Intelligence*, ser. WI ’17. New York, NY, USA: ACM, 2017, pp. 947–952. [Online]. Available: http://doi.acm.org/10.1145/3106426.3109420

[8] A. hwee Tan, “Text mining: The state of the art and the challenges,” in *In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, 1999, pp. 65–70.

[9] S. Canuto, M. A. Gonçalves, and F. Benevenuto, “Exploiting new sentiment-based meta-level features for effective sentiment analysis,” in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ser. WSDM ’16. New York, NY, USA: ACM, 2016, pp. 53–62. [Online]. Available: http://doi.acm.org/10.1145/2835776.2835821

[10] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, “Current state of text sentiment analysis from opinion to emotion mining,” *ACM Comput. Surv.*, vol. 50, no. 2, pp. 25:1–25:33, May 2017. [Online]. Available: http://doi.acm.org/10.1145/3057270

[11] Q. You, “Sentiment and emotion analysis for social multimedia: Methodologies and applications,” in *Proceedings of the 2016 ACM on Multimedia Conference*, ser. MM ’16. New York, NY, USA: ACM, 2016, pp. 1445–1449. [Online]. Available: http://doi.acm.org/10.1145/2964284.2971475

[12] B. Rieder, “Studying facebook via data extraction: The netvizz application,” in *Proceedings of the 5th Annual ACM Web Science Conference*, ser. WebSci ’13. New York, NY, USA: ACM, 2013, pp. 346–355. [Online]. Available: http://doi.acm.org/10.1145/2464464.2464475

[13] E. Loper and S. Bird, “Nltk: The natural language toolkit,” in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ser. ETMTNLP ’02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 63–70. [Online]. Available: https://doi.org/10.3115/1118108.1118117

[14] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

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

[16] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006. [Online]. Available: http://dl.acm.org/citation.cfm?id=1248547.1248548

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[18] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg, “Feature hashing for large scale multitask learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML ’09. New York, NY, USA: ACM, 2009, pp. 1113–1120. [Online]. Available: http://doi.acm.org/10.1145/1553374.1553516

[19] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, “Efficient and robust automated machine learning,” in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2962–2970. [Online]. Available: http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf