ISSN 2278-3091

Volume 10, No.1, January - February 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse051012021.pdf

https://doi.org/10.30534/ijatcse/2021/051012021

Analyzing the Voice of Customer through online user reviews using LDA: Case of Moroccan mobile banking applications

Meriem Tabiaa[1], Abdellah Madani[2]

[1]LAROSERI Laboratory, Computer Science Departement, Faculty of Sciences, Chouaib Doukkali University,
Morocco, meriemtabiaa@gmail.com

[2]LAROSERI Laboratory, Computer Science Departement, Faculty of Sciences, Chouaib Doukkali University,
Morocco, madaniabdellah@gmail.com

ABSTRACT

Nowadays, most of the organizations make their mobile applications available through different stores, such as Google Play Store, Apple App Store, and Windows Phone Store. Banks and financial institutions have also provided mobile applications for their customers. These app stores not only allow applications to be downloaded, but they also permit users to leave comments and reviews. In this paper, we will start first by looking at eight Moroccan mobile banking applications in the Google Play Store. Data that hasn't been exploited by Moroccan banks yet. Once the preprocessing phase is complete, we will examine and analyze user reviews using Latent Dirichlet allocation (LDA) to extract and identify topics. Topics discovered focus mainly on Security, services, quality, and interface. While customer reviews can influence future demand, they can also be used by managers to improve their services and customer experience.

Key words:. Latent Dirichlet Allocation, Topic Model, user reviews, e-banking, Consumer feedback.

1. INTRODUCTION

E-Banking is turning into a competitive industry where innovation thrives, due to the importance of this sector for the economy, making it an alluring field for researchers. E-Banking is a domain that creates a huge amount of data, where AI applications can be potentially profitable for businesses by increasing the visibility and acknowledgment of research achievements [1] [2].

With the rise in the service economy, the success of service providers now depends on their ability to provide customer-centric services [3]. The importance of perceived service quality and customer experience is widely acknowledged [4].

In fact, an organization does not work to produce, but above all to satisfy the needs of its customers and even to anticipate their expectations. The customer is the reason for the organization's existence and is, therefore, involved in all its processes.

The voice of the customer plays a crucial role in companies. Classical methodologies in this field have focused mainly on market surveys and questionnaires to obtain customer preferences. Listening to the customer is only important if the feedback is analyzed and if an effort is made to make corrections. It is important to invite the customer to describe their experience in their own words, but it is also essential to act through customer feedback. In most cases, companies are constantly collecting feedback from consumers through satisfaction surveys, online evaluations, e-mails, as well as in person, to support and improve grievances [5].

Meanwhile, online reviews have proven to be an excellent and reliable channel for not only understanding customer needs for a product or service, but also for analyzing product competitiveness in the marketplace.

Previously we have conducted an empirical study on a small sample of the population limited by a geographical area, and we did it through a field survey. The study aimed to acceptance and use of e-banking by Moroccan customers [6]. The results obtained have motivated us to move to another level.

Conclusively, we argue that online reviews are more beneficial source of data than traditional survey methods. Through this methodology, it is easier to understand the entirety of the customer's experience. As is known, while online reviews have a lot of influence on tourist establishments such as restaurants or hotels. They also have an impact on the industry and financial organizations, the online consumer reviews have a particularly important impact on the search of information, the evaluation and the

decision-making of the consumers in many fields [7]. User reviews are an important part of the user experience, particularly since subject satisfaction plays a huge role in acceptance. it is useful to better understand what precisely users state in their reviews and if any particular aspects would impact the depth of feedback that a user will leave [8].
Hence, online user review is an important tool for gathering information about consumers and thus for creating opportunities. For this purpose, online banks must analyze their feedback to improve their services and products. So, finding and collecting data is no longer the impossible task, since customers comment day after day in forums, social networks. But the complexity lies in how these data must be used to set up strategic actions.

The goal of our research is to develop an approach based on topic modeling using information from online reviews on Google Play Store and using a dataset that was missed by decision-makers. However, this research focuses on one type of topic model, utilizing a topic modeling technique called latent Dirichlet allocation (LDA) [9].

This paper is organized as follows: in the next section, we present the context of topic modeling within the banking sector, and online user reviews. Thereafter, we outline the topic modeling technique and the LDA. This is followed by the methodology, the discussion of the research results. Finally, based on all these findings could help decision-makers and managers design better launch strategies.

2. LITERATURE REVIEW

Banks get the most benefit from big data since they can now easily and quickly extract relevant information from their data and convert it into benefits for their own and their customers [10]. Social media data can be used to identify topics for discussion at a specified instant. Previous research indicates that these data “can be a good source of entity-oriented topics that have low coverage in traditional media news” [11].

Prior to this, research on user review generated by Social Media has focused on the tourism and hotels sector, e-commerce [5]. Only a limited amount of research currently focuses on consumer feedback in the banking context.
It is noted that most of the studies reviewed used external data. The reasons for this can be twofold; authors would prefer to use the external data since they are public and free to use, while the internal data are the ownership of the company and numerous restrictions can apply in using them as a data source for text mining [12].

In Indonesia, research was conducted in purpose to investigate how bank customer conversation network evolves and dynamically spread the information for the purpose of SCRM (social customer relationship management) by calculating its network properties and compare the result between three Indonesians banks. Also, based on Twitter discussions related to the banks, they are conducting a sentiment analysis to find out which of the three banks has the most positive sentiment towards the banks. In this research, they adopt the Appraisal theory method used to classify sentiment because the method is proven applicable to social media conversation. Appraisal theory describes how authors use language to communicate with others and it shifts sentiment classification further and considers the appraisal expression [13].

In a further study, researchers experimented models of bi-directional relations patterns among eWOM metrics and bank profitability along time. The results showed that both the ranking by stars and the verbal emotions expressed by consumers in eWOM significantly predicted an improvement in the future profitability of companies. In that case, for several reasons, they used Linguistic Inquiry and Word Count LIWC, a popular text analysis software: LIWC extracts quantitative data, making the resulting data amenable to statistical analysis. Besides, LIWC not only evaluates fundamental grammatical features of texts but also offers insights regarding significant psychological processes; LIWC performs this function automatically and avoids the chore of hand-encoding. LIWC is one of the most widely-used and validated ATA tools used in academic research currently. Despite the fact that this research shows that the LIWC is an effective tool for studying the eWOM, its strategy for counting words has its limitations [14]. Also this study focused on two text analysis indicators including Positive Feelings and Anger [15].

Another approach proposed; this one is based on rules and classification to analyze the sentiment of Chinese microblogs related to finance. At first they employed an Improved Label Propagation Algorithm (I-LPA) in order to build the lexicon of sentiments systematically. Afterwards, based on the microblog topics, they split the microblogs into multi-topic microblogs and mono-topic microblogs by topical classification. Concerning the multi-subject microblogs, they conduct a rule-based sentiment analysis. Three layers of filtering rule are used to identify the emotion agents of a specified topic.Next, the sentiment is calculated according to the syntactical dependence relationship among the words of sentiment and the agents of emotion. On the other hand, single-topic microblogs exploit classification based on SVM to compute emotion. The results show that I-LPA is effective and the method of sentiment analysis is promising and outstanding for not only single-topic microblogs but also the multi-topic [16].

2.1 Analysis of online user reviews

The second related line of research is the analysis of online reviews. User reviews can behave like online Word-Of-Mouth (WOM). WOM is recognized as influential in information transmission, particularly with good experiences [17] [18]. User-generated online reviews implicitly communicate user-perceived quality, from which “perceived ease of use” and “perceived usefulness” may be inferred [19]. This creates a feedback loop granting a focused opportunity for refinement in subsequent iterations.

A user review usually consists of a numerical rating showing the general opinion of the reviewers, and a text describing the reviewers' evaluation in detail. In previous decades, studies of user reviews have mainly focused on rating information, but have barely explored the content of the text. Lately, the text analysis of user reviews has received significant attention. However, since user reviews of products are often multidimensional, which cannot simply be captured by a single numerical rating, textual analysis of reviews provides richer information for researchers and professionals to understand consumers [20].

In fact, the pertinence of reviews depends on the perspective. Developers get value from positive reviews, while negative reviews provide more value to the upcoming community of potential users. It is common for unsatisfied consumers to express their experience of an unfulfilling purchase from a company by manifesting their intention not to continue shopping, warning other consumers not to buy from that company as well. User review can also consider as a benchmark, internal audit for a firm [5].

Another study also states that users will leave reviews of varying length (often domain-specific) and can possess abbreviations, colloquial expressions, and non-standard spelling [21]. It is also acknowledged that reviews often addressed myriad aspects within the context of the domain and the object under review [22] [23].

2.2 The application of topic modeling and LDA to the user reviews analysis

Research on topic models has recently picked up the pace, especially in the field of generative topic models such as LDA [9], their hierarchical extensions [24], topic quality assessment and visualization [25] [26].

Over the last couple years, techniques of topic modeling using probabilistic latent semantic analysis PLSA and LDA have been increasingly used for various purposes, from simple applications to unsupervised analysis including many extensions to add-on items. and structure that may be required in specific settings, supervised versions to be employed as text classifiers, an analysis of the topic's evolution in a set of documents along time, image recognition and classification, and so on… [27].

Previous researches used LDA-based approaches and word similarity to determine products' attributes from online user reviews. LDA-based methods employ the tf (frequency of terms) and df (frequency of documents) to eliminate common and local words that appear in specific review documents. However, due to the mixed-part-of -speech (POS), which includes sentimental words, this means that we cannot easily identify product attributes in LDA [28].
LDA turns into a mainstream approach to topic modeling. However, none studies have been conducted to analyze online user reviews on the banking sector, using the LDA method. This present research aims to fill this gap in LDA applications.

3. METHODOLOGY: EXPERIMENTS, RESULTS AND DISCUSSION

3.1 Experiments and results

In order to apply LDA on online user reviews and extract topics, we have designed an approach based on 7 main steps: The first step is data collection from Google Play Store. The second step is the English translation of text review. The third step is to augment our dataset, then we proceed to the data pre-processing step, afterwards to the transformation before the model training and we finish our approach with the topics nomination. You will find below our overall framework (Figure 1), detailed in the following sub-sections.

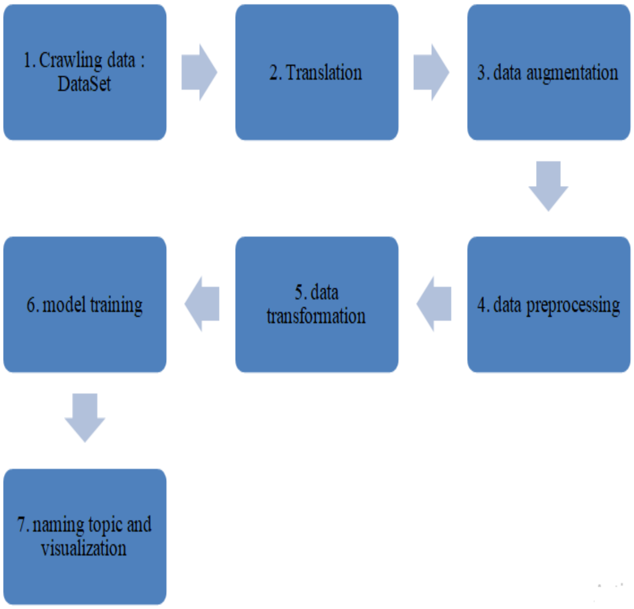
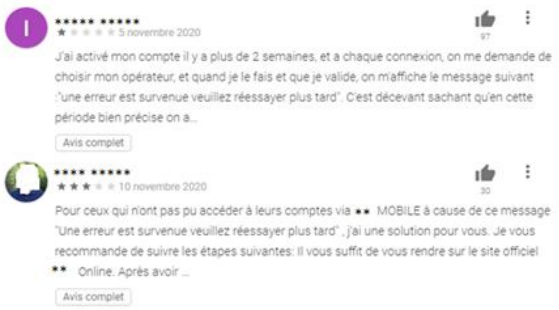


Figure 1: Our proposed Framework.

3.1.1 Crawling Data: Data sets

The first step of our approach is to collect online customer reviews. For this purpose, various techniques can be used to collect social media data, such as web crawling and application programming interfaces (APIs) provided by Twitter, Facebook or blog services. For this study, we developed a script in JavaScript that allows scraping and downloading all comments from the Google Play Store (Figure 2).



***** *****
★ 5 novembre 2020
97

J'ai activé mon compte il y a plus de 2 semaines, et a chaque connexion, on me demande de choisir mon opérateur, et quand je le fais et que je valide, on m'affiche le message suivant :"une erreur est survenue veuillez réessayer plus tard". C'est décevant sachant qu'en cette période bien précise on a...

**** *****
★★★ 10 novembre 2020
30

Pour ceux qui n'ont pas pu accéder à leurs comptes via ** MOBILE à cause de ce message "Une erreur est survenue veuillez réessayer plus tard" , j'ai une solution pour vous. Je vous recommande de suivre les étapes suivantes: Il vous suffit de vous rendre sur le site officiel ** Online. Après avoir ...

Figure 2: Examples of online comments and customer reviews on the Google Play store.

More precisely, the comments collected refer to the time period from January 2013 to December 2019.

The initial dataset contained: Total number of words equals 48493 words grouped in 7832 sentences, with an average of 6 words per sentence. Notably, the lexical density is defined as the number of lexical words (or content words) divided by the total number of words. The lexical words give meaning to a text and provide information about its content) is very small, so we can conclude that there is a lot of noise that requires pre-processing of this set of data. Each review on the Google Play Store is composed of a rate from 1 to 5, a name of user and a body of review. Our scripts download all the information available for a review and preserve the underlying data relationships. For our data sets, the data collected is composed of: date of publication, customer review (text format).

3.1.2 Translation

Prior to augmenting the dataset, we decided that it was best to translate it into English in order to use the effective language tools. Many APIs are used for translation, including paid and free ones. The most important objective to achieve with the translation is to avoid any loss of meaning. For several reasons, we decided to use the Google-trans API. Googletrans is a free and illimited python library that has implemented the Google Translate API.

3.1.3 Data augmentation

Concerning ML Algorithms, dataset is a key factor in the usefulness of these models. Many problems arise when datasets are mismatched, when data is sparse, or when data becomes hard to collect, or difficult to label. One of the methods which handles these shortcomings is data augmentation [29].

For our dataset we use EDA (easy data augmentation techniques) for boosting performance. EDA consists of four simple but effective actions: changing synonyms, arbitrary insertions, random permutations and eliminations. EDA substantially boosts performance and reduces over-fitting when training on smaller datasets [30].

We augmented the data twice more. We notice an augmentation in terms of words and also a gain in lexical density, this is due to the replacement and insertion of synonyms that does not contain noise.

Pre-processing is an important step in a study that uses text as a source material, but it decreases the number of the dataset because of the noise removed. Therefore, increasing the dataset is still the best solution to solve this problem, hence the use of EDA, which has served us well (Table 1).

Table 1: Example of data augmentation

Review	Augmentation of review
code document username gives identifier wrong activation message mention application error photograph incorrect bank suddenly	codification papers username output mistake yield identifier code short_circuit catch askew application awry activating defective content cite practical_application samarium sedimentation fault message citation energizing snap faulty document deposit short
flexible install time application new try method needs find agency require password simpler go time	flexible install time meter natura covering New seek install simpleton flexile method acting need recover agency password regain federal agency new call fo die demand search request countersign simple method give-up_the_ghost metre clock

3.1.4 Data Preprocessing

Real-world data is generated from many sources and processes. They may contain anomalies or incorrect values that compromise the quality of the dataset: it has not been transformed, cleansed or changed at all [31]. For Our Text pre-processing , we have 4 sub- steps, including capitalization [32], elimination stop words [33], word text tokenization [34], and word lemmatization [35].

Table 2: A sample of a text review before / after preprocessing.

Before preprocessing	Pay card really work properly really nil last web load procedure sense day also last make application care message unload anything day
After preprocessing	[‘card’, ’work’, ’load’, ’procedure’, ’sense’, ‘application’, ’care’, ’message’ ’ unload’]

3.1.5 Data transformation

- TF-IDF (Term Frequency - Inverse Document Frequency) is a similarity measure widely used in information retrieval. TF-IDF's main idea is to correspond each document into a vector having the size of the common dictionary [36]. This N-Grams approach is commonly used in the statistical natural language processing. It is language-dependent and works well in the case of noisy text [37].
- Corpus and Dictionary: The two main inputs to the LDA topic model are the dictionary (id2word) and the corpus. Gensim generates a unique ID for each word in the document. The above produced corpus is a mapping of (word_id, word_frequency) (Figure 3). The original data has 4 columns (name of the bank, the date, the name of the user, the user review).
-

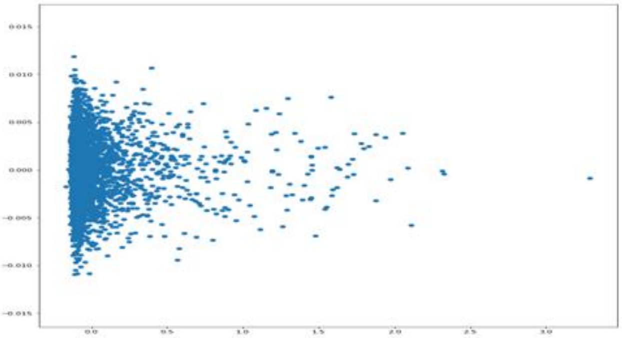


Figure 3: Creating the dictionary.

- Determinate the number K of topics by using the Elbow method: In this research, we deployed a method based on the Elbow method [38]., to estimate the optimal number of topics. To use this parameter as an LDA model input, after applying the Elbow method to all datasets, the model specified that the optimal number of the topic is 4 topics (Figure 4).

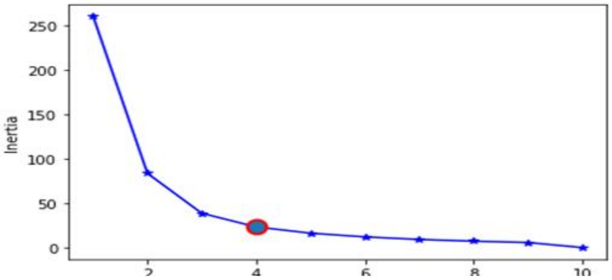


Figure 4: Elbow Graph determinate the number of topic K.

3.1.6 LDA Model training

To implement LDA, we used tools from the Python Libraries, which provide functionalities to analyze semantic structure in texts [39]. Based on the results of Elbow methods, the model specified that the optimal number of topics K= 4 as the number of topics to extract. Apart from that, we used hyper-parameters α and β that affect sparsity of the topics with symmetric 1.0/K priors (we’ll use the default for the base model).

The above LDA model is built with 4 topics where each topic is a combination of keywords and each keyword contributes a certain weightage to the topic. For each topic we can visualize key-words and the weightage of each one. The table (Table 3) below represents each topic with the top of 20 words.

Since the LDA works like a blending model, the same word can appear in more than one topic. Below, we briefly introduce each topic.

Table 3: Top 20 words for each topic.

Topic #k	Top Words of each topic
Topic #0	application bank update good problem time transfer work app operation agency account code bug card mobile cih service pay open
Topic #1	account application day enter history work give_thanks customer connect problem phone job number service piece_of_work access message unable block impossible
Topic #2	payment application morocco invoice recharge defrayal add lack receive internet iam information charge telecom pay electricity water net option service
Topic #3	version update password connection answer bac convenient fingerprint service instal fresh bank app android work install contact watchword commodious connexion

Our LDA model produced 4 topics. Each topic is a collection of word-weight couples. Words with high corresponding weight values are most representative of the topic (Table 4).

Table 4: Words Distribution.

Topic 0		Topic 1		Topic 2		Topic 3	
0,075	update	0,052	Application	0,046	Service	0,067	Bank
0,072	application	0,036	Bill	0,041	Customer	0,049	Problem
0,046	time	0,029	Payment	0,035	Account	0,018	Bad
0,024	bug	0,025	Transfer	0,029	Application	0,013	Implementation
0,024	block	0,022	Access	0,019	Enter	0,014	Job
0,023	day	0.021	Good	0,018	Password	0,013	Unavailable
0,019	work	0,018	Pay	0,015	History	0,013	Server
0,014	connect	0,018	App	0,014	Call	0,011	Attempt
0,012	agency	0,013	Lack	0,013	Client	0,010	Month
0,009	clip	0,013	Change	0,012	Contact	0,010	Wait

In the LDA model, each document is composed of several topics. However, usually one of these topics is dominant. The table below shown this dominant topic for each sentence and indicates the weight of the topic and the keywords in a well-formatted output. This way, you will know which document belongs primarily to which topic.

Table 5: Words Distribution.

Document Nember	Dominant Topic	Topic_Perc_Contrib	Keywords	Text
0	0.0	0.7207	update, application, time, bug, block, day, wo...	application work web procedure really nil load...
1	0.0	0.3842	update, application, time, bug, block, day, wo...	loading work really wage nil load nothing trul...
2	0.0	0.4566	update, application, time, bug, block, day, wo...	application work well overall real problem con...
3	0.0	0.5100	update, application, time, bug, block, day, wo...	drop hush real_number work refer well real per...
4	1.0	0.6720	application, bill, payment, transfer, access, ...	really disappointed nothing change receive por...
5	1.0	0.9333	application, bill, payment, transfer, access, ...	change pack manipulator practical_application ...
6	1.0	0.5597	application, bill, payment, transfer, access, ...	application month announce still feed account ...
7	1.0	0.5470	application, bill, payment, transfer, access, ...	possibility announce tranquilize provender fee...
8	3.0	0.8728	bank, problem, bad, implementation, job, unava...	big webpay refill example card cancel problem
9	3.0	0.8467	bank, problem, bad, implementation, job, unava...	problem identity_card card beginning offset tr...

3.1.7 Naming topics

For our Research, We use WordNet as our lexical resource of choice [40]. WordNet unequivocally models synonymy by connecting synonyms to a similar sense. In WordNet, each sense has a related definition. WordNet senses are associated with relations such as synonymy, hypernymy, similar attributes, etc.

Instead of showing a list of the top 20 words, we show a word cloud of the top 10. Which the size of each word is measured by the probability of having that word in the topic. As shown in Figure 5.

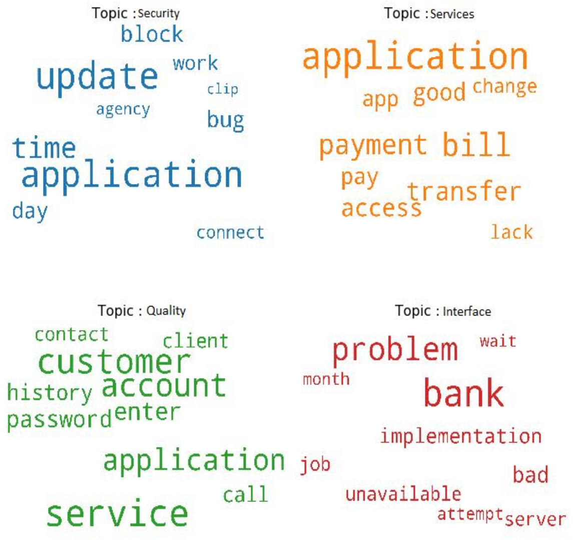


Figure 5: Naming topics by Wordnet.

3.2 Discussion and limitations.

Applying LDA's topic model; designed mainly for short texts, 4 topics were identified. Based user reviews, banks will be able to improve their application design, security, system, and customer relationship management.

Analyzing user opinions is a challenging exercise. For our research, we examined reviews of 8 mobile banking applications. A further challenge is the small number of relevant reviews. Before our pre-processing phase, we augment the dataset to extend it. In the topic modeling step, the LDA model and the ElBow model obtained the best results.

We can visualize that topic 0 (security) reported updates, bugs and application problems. Developers should devote extra attention to those issues, While reading through the reviews, we notice that for many of the complaints, users also report that they recently updated their app and have been experiencing authentication problems ever since (Figure 5).

Moreover, in topic 1 (services) the users repeat often words payment, application, bills, users would like that their application could make more operations and get more services (Figure 5).

Since users review apps as a whole, they often raise issues that are not directly the responsibility of the developers; some reviews are directed towards product managers or other team members. Which is the case of topic 2 (quality) we notice the words: client, account, history and service. Users claim the quality of the service (Figure 5).

Concerning the last topic (the interface), we can deduct the fact that users are experiencing problems with the interface. Users report frequent problems with recent updates due to design modifications. This is an important area for improvement, especially the recurrence of the words “problem” and “bad”.

Our study experiences limitations.

- Data Anomalies: we noticed numerous anomalies in the scraped data in the Google game store. For example: some users wrote their reviews by combining two languages. Others preferred to write their reviews in Moroccan Dialect using Latin letters. As well as some of the reviews, only contain a single word, such as "good" or "satisfied".
- Access to the internet and Smartphone: A further important limitation is that online customer reviews cannot be generalized, as they assume that all consumers have installed the applications and have access to the internet. However, those who do not have access may not be able to rate and comment.
- Generalization of findings: our study used consumer feedback from eight bank applications. However, the selection of more than one mobile application allows for a generalization, and therefore input can affect the findings, so each bank needs to analyze customer feedback in a separate way.
- Single data source: we only used scrapped data from a single plate-form (google play) for the analysis. In the future, we may look to other plateforms; and even on separate dates to study consumer feedback.
- Human interpretation: The LDA model required a human interpretation based on the probabilities assigned to the words
- Fine-tuning: The LDA model needs a lot of fine-tuning, which involves a significant time and energy effort.

4. CONCLUSION AND FUTURE WORK

The current worldwide economic crises as well as in Morocco, due to the propagation of a new generation of the corona virus. Financial organizations are obliged to digitize all their services; it is no longer a luxury but a requirement. The customer relationship management will mainly be done online, by messaging, telephone, social networks, etc... Banks must exploit online reviews to listen better to customers, to anticipate and develop strategies and operational plans of action. Our findings can help developers and managers better anticipate the users-reviews and prioritize their limited QA resources towards the most impactful reviews.
During the conduct of this research, several points have been identified that deserve further study:

i. It would be interesting to compare the text reviews for each bank separately.
ii. For this research, the text corpus was extracted from Google Play Store. It could be interesting to explore other data sources such as social networks and forums for example.
iii. Another interesting area of research will be exploring the model by applying it to private reviews, such as instant messaging on messenger.
iv. Future studies can use more data with the proposed approach, and it can apply sentiment analysis and deep learning
v. Finally, we can envisage a comparative study of Moroccan banking applications versus other applications in other countries.

REFERENCES

1. Tabiaa M, Madani A, El Kamoun N (2017) E-Banking: Security risks, previsions and recommendations. Int J Comput Sci Netw Secur 17:189–196
2. Tabiaa M, Madani A (2019) The deployment of Machine Learning in eBanking: A Survey. In: 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS). IEEE, pp 1–7

3. Gustafsson A, Nilsson L, Johnson MD (2003) The role of quality practices in service organizations. Int J Serv Ind Manag

4. Halvorsrud R, Kvale K, Følstad A (2016) Improving service quality through customer journey analysis. J Serv Theory Pract 26:840–867. https://doi.org/10.1108/JSTP-05-2015-0111

5. Mou J, Ren G, Qin C, Kurcz K (2019) Understanding the topics of export cross-border e-commerce consumers feedback: an LDA approach. Electron Commer Res 19:749–777. https://doi.org/10.1007/s10660-019-09338-7

6. Tabiaa M, Madani A, EL KAMOUN N (2018) L'ADOPTION DE L'E-BANKING CHEZ LES CLIENTS : ETUDE EMPIRIQUE AU MAROC

7. Pekarskaia Dauxert T (2019) Online Consumer Reviews in the Hotel Sector: Why and How to Answer? In: Baghdadi Y, Harfouche A (eds) ICT for a Better Life and a Better World. Springer International Publishing, Cham, pp 313–322

8. Vasa R, Hoon L, Mouzakis K, Noguchi A (2012) A preliminary analysis of mobile app user reviews. In: Proceedings of the 24th Australian Computer-Human Interaction Conference on - OzCHI '12. ACM Press, Melbourne, Australia, pp 241–244

9. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

10. Srivastava U, Gopalkrishnan S (2015) Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks. Procedia Comput Sci 50:643–652. https://doi.org/10.1016/j.procs.2015.04.098

11. Schumaker RP, Zhang Y, Huang C-N, Chen H (2012) Evaluating sentiment in financial news articles. Decis Support Syst 53:458–464

12. Pejić Bach M, Krstić Ž, Seljan S, Turulja L (2019) Text Mining for Big Data Analysis in Financial Sector: A Literature Review. Sustainability 11:1277. https://doi.org/10.3390/su11051277

13. Alamsyah A, Indraswari AA (2017) Social Network and Sentiment Analysis for Social Customer Relationship Management in Indonesia Banking Sector. Adv Sci Lett 23:3808–3812. https://doi.org/10.1166/asl.2017.9279

14. Mehl MR, Gill AJ (2010) Automatic text analysis.

15. Tang C, Mehl MR, Eastlick MA, et al (2016) A longitudinal exploration of the relations between electronic word-of-mouth indicators and firms' profitability: Findings from the banking industry. Int J Inf Manag 36:1124–1132. https://doi.org/10.1016/j.ijinfomgt.2016.03.015

16. Yan D, Hu B, Qin J (2018) Sentiment Analysis for Microblog Related to Finance Based on Rules and Classification. In: 2018 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, Shanghai, pp 119–126

17. Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. Mark Sci 23:545–560

18. Granovetter MS (1977) The strength of weak ties. In: Social networks. Elsevier, pp 347–367

19. Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 319–340

20. Wang F, Yang Y, Tso GKF, Li Y (2019) Analysis of launch strategy in cross-border e-Commerce market via topic modeling of consumer reviews. Electron Commer Res 19:863–884. https://doi.org/10.1007/s10660-019-09368-1

21. Platzer E (2011) Opportunities of automated motive-based user review analysis in the context of mobile app acceptance. Proc CECIIS 309–316

22. Gebauer J, Tang Y, Baimai C (2008) User requirements of mobile technology: results from a content analysis of user reviews. Inf Syst E-Bus Manag 6:361–384

23. Duan W, Gu B, Whinston AB (2008) Do online reviews matter?—An empirical investigation of panel data. Decis Support Syst 45:1007–1016

24. Teh YW (2006) A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp 985–992

25. Chang J, Gerrish S, Wang C, et al (2009) Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems. pp 288–296

26. Blei DM, Lafferty JD (2009) Topic models. In: Text mining. Chapman and Hall/CRC, pp 101–124

27. Nikolenko SI, Koltcov S, Koltsova O (2017) Topic modelling for qualitative studies. J Inf Sci 43:88–102. https://doi.org/10.1177/0165551515617393

28. Joung J, Kim HM An LDA-based Approach for Product Attribute Identification from Online Customer Reviews. 8

29. Sharifirad S, Jafarpour B, Matwin S (2018) Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Association for Computational Linguistics, Brussels, Belgium, pp 107–114

30. Wei J, Zou K (2019) EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. ArXiv190111196 Cs

31. García S, Luengo J, Herrera F (2015) Data preprocessing in data mining. Springer

32. Wichaiwong T, Koonsanit K, Juruskulchai C (2008) A Simple Approach to Optimized Text Compression's Performance. In: 2008 4th International Conference on Next Generation Web Services Practices. IEEE, Seoul, Korea, pp 66–70

33. Porter MF (1980) An algorithm for suffix stripping. Program 14:130–137

34. Palmer DD Chapter 2: Tokenisation and Sentence Segmentation. 23

35. Airio E (2006) Word normalization and decompounding in mono- and bilingual IR. Inf Retr 9:249–271. https://doi.org/10.1007/s10791-006-0884-2

36. Eck M, Vogel S, Waibel A Low cost portability for statistical machine translation based on n-gram frequency and TF-IDF. 7

37. Khreisat L (2006) Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. DMIN 2006:78–82

38. Thorndike RL (1953) Who belongs in the family. In: Psychometrika. Citeseer

39. Suresh H, Locascio N (2015) Autodetection and Classification of Hidden Cultural City Districts from Yelp Reviews. ArXiv150102527 Cs

40. Fellbaum C (2010) WordNet. In: Theory and applications of ontology: computer applications. Springer, pp 231–243