Customer Sentiments Toward Saudi Banks During the Covid-19 Pandemic

Dhuha Alqahtani, Lama Alzahrani, Maram Bahareth, Nora Alshameri, Hend Al-Khalifa, Luluh Aldhubayi
Information Technology Department, College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia
{441203741, 442203072, 439203913, 442204277}@student.ksu.edu.sa,
{hendk, laldubaie}@ksu.edu.sa

Abstract

In view of the recent interest of Saudi banks in customers' opinions through social media, our research aims to capture the sentiments of bank users on Twitter. Thus, we collected and manually annotated more than 12,000 Saudi dialect tweets, and then we conducted experiments on machine learning models including: Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (RL) as well as state-of-the-art language models (i.e. MarBERT) to provide baselines. Results show that the accuracy in SVM, LR, RF, and MarBERT achieved 82.4%, 82%,81%, and 82.1% respectively. Our models code and dataset will be made publicly available on GitHub.

1 Introduction

Over the last decade, social media sites generated enormous content online which conducts challenges to decision making and manual content analysis (Almuqren & Cristea, 2021). The Saudi banks sector has undergone essential changes over the decade. They have taken advantage of expanding their operations of product diversification as well as the features of scale and scope economies. The changes affected the feelings and sentiments of customers and their dealings with banks as well as the choice of the bank based on the features and services they provide. In addition, appropriate treatment of the customer is a milestone in the selection of a bank, and good treatment is characterized by quick interaction, offers and satisfactory customer service.

Social media sites are one of the platforms where customers' opinions can be collected and analyzed. Our research aims at capturing customers' sentiments of Saudi banks, by analyzing their feelings and revealing their opinions.

Although there are several banks in Saudi Arabia, yet, for the purpose of this study, only four Saudi banks were selected namely (AlRajhi bank, Alinma bank, Saudi National Bank (SNB), and Saudi Investment Bank (SAIB)) as they are considered the top Saudi banks according to Semrush website[1].

We collected a corpus of more than 12,000 Arabic tweets and manually labeled them with three sentiments (positive, negative, and neutral). Then we applied three machine learning models on the corpus, namely: Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) as well as MarBERT pretrained model for sentiment analysis.

The rest of this paper is organized as follows. Section 2 presents the state-of-the-art and related work on banking sentiment analysis. Section 3 demonstrates the corpus construction steps and the annotation process. Section 4 presents the analysis results which include model selection and corpus evaluation. Section 5 concludes the paper with discussion of the results, limitations and future work.

[1] https://www.semrush.com/website/top/saudi-arabia/banking/

2 Related work

One of the research areas in sentiment analysis is to capture customer sentiments regarding vital services such as Banking, by knowing and classifying customer opinions, this will help improve the titled services.

In this section, we will highlight previous work in sentiment analysis for banks. The study by Kazmaier and Vuuren (2020) conducted a sentiment analysis as unstructured customer reviews that related to services and products for a retail bank in South Africa. They used a machine learning model to detect sentiment with a high level of qualified performance. The result shows that custom learning-based models are better than previous models that used commercial tools and were pre-trained for sentiment classification.

Eksa Permana et al. (2020) present a study to determine the customer sentiment on mobile banking applications to specify the aspects that need to be maintained or improved
in the application. They used Naive Bayes models to discover the sentiment analysis. The results displayed high accuracy at the value of k=5, which is accuracy with a value of 86.762% and precision with 92.482% also 93.474% for recall.

(Gavval et al., 2019) proposed a visual sentiment analysis for customer complaints related to services and products of four superior Indian banks. The author leverages the available bank's compliant responses dataset which consists of 749k consumer opinion on four Indian banks namely: Axis Bank, HDFC Bank, ICICI Bank, and SBI. They used a Self-organizing feature map (SOM) and CUDA-based Self Organizing Feature Map (CUDASOM) algorithm. They mentioned that the performance of CUDASOM algorithm increased the speed of CPU up to 44 times which improved the result of experiment.

Krishna et al. (2019) applied sentiment analysis on Indian bank's customer complaints. They used machine learning techniques such as Support vector machines (SVM), Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) towards applying preprocessing the raw textual data. The findings showed that the ML models with three banks dataset performed the best result where LR obtained (77%), RF (77%), SVM (75%), and NB (74%).

From the previous work, we noticed that several research used machine learning methods such as (Kazmaier and Vuuren, 2020), (Eksa Permana et al.,2020), (Gavval et al., 2019), and (Krishna et al., 2019).

As shown in Table 1, we present some of studies with the name of the bank or region, the applied model and results. In our research, the main contribution is to construct a gold standard dataset for the Saudi banks customers' sentiment.

Finally, the corpus will be evaluated using machine learning and pretrained models.

Study	Bank	Method	Result
Kazmaier and Vuuren (2020)	Retail bank in South African	Naive Bayes, SVM, CNN, ANN, and LR	LR: 84.00%
Eksa Permana et al. (2020)	Mobile banking applications	Naive Bayes	NB:86.76%
Gavval et al. (2019)	Four superior Indian banks	CUDASOM algorithm	Speed up the CPU performance 44 times
Krishna et al. (2019)	Indian bank	SVM, NB, LR, DT, RF	LR: (77%), RF: (77%), SVM:(75%), NB: (74%)

Table 1: Summary of relevant studies

3 Dataset

This section contains two parts related to constructing the Saudi Bank Corpus. The first part is to collect the data and a description of the corpus. While the second one represents the corpus annotation phase and statistics about the corpus.

3.1 Data Collection

The Saudi banks' corpus data were collected from Twitter using Tweepy python library. We used the banks' Twitter account mentions of the four Saudi banks to retrieve the tweets.

The data were collected for a whole month, starting from the 1st of September 2021 until the 30th of September 2021, throughout the COVID-19 outbreak when all bank institutions switched their services online. During that period, most people encountered online issues when using their bank services. Therefore, due to the COVID-19 constraints, social media platform such as Twitter was the main channel to communicate with the bank's organizations. The retrieved tweets were focusing on getting customers' feedback or opinion on the banks. A total of 52,254 tweets were collected during the whole month.

As we scrapped the tweets using the official bank account (@bank_account) on Twitter, as a result, the number of tweets crawled was depend on the number of bank customers, whether the Twitter account is active or not, and if there is another account for customer complaints. Moreover, the bank that has a separate customer care account received too many tweets from their customers than the bank that has one official account. Moreover, the bank that is 24/7 active and responds quickly has many user replies and comments. The special occasions such as National day and the recent collaboration between banks plays a crucial role in increasing customer response accordingly.

3.2 Data Preprocessing

Since the data is collected from Twitter, it could contain noises and unwanted symbols, which could eventually affect the model's performance. In this section, several normalization and cleaning steps were applied on the collected data using regular expression patterns. The pre-processing steps are listed below:

- **Normalizing Letters:** Some of the letters got changed to have fixed shape for example "إأآ" would be "ا".
- **Normalizing numbers:** All the numbers have standard representation in our case, we used Arabic numbers for example "٩٩" would be "9".
- **Remove duplicate letters:** All the letters that occur for more than two times get limited to two times since some words could have the same letter twice, this process has a special case where the laugh such as "هههههه" replaced by "ضحك"
- **Remove duplicate tweets:** Any duplicated tweet is kept only once to make sure we don't annotate the same sentence more than once.
- **Removing punctuation marks:** All the punctuation marks got deleted except the "!?." Since it could have meaning for the sentiment.
- **Removing duplicate whitespace:** White spaces between words were eliminated to one space only.
- **Remove hyperlinks or URLs:** This process consists of removing all the URLs (HTTP or HTTPS).
- **Remove hashtags and mentions:** For example, #something" and "@someone" will be deleted.
- **Remove special characters:** Special characters include ".,$,%,&,*, etc." were deleted.
- **Remove Arabic diacritics:** Deleting Tashkeel including "ً"
- **Remove Tatweel words:** Refers to removing the stretching word space that is represented as (-) symbol.
- **Remove non-Arabic word**s: Any non-Arabic words got deleted, for example, English words will get deleted.
- **Remove sentences with less than five words:** Sentences with less than five words could be too short to represent any type of sentiment so, we deleted them.
- **Eliminate tweets with neutral sentiment:** we used Camel and Mazajak tools that is designed for Arabic SA. We used it to eliminate the neutral tweets, since we wanted to make sure that the tweets would contain positive or negative sentiment.

3.3 Data annotation

Based on the type of analysis that will be performed on the corpus "sentiment analysis," it was decided that the annotation will be a single level annotation, where each sentence will be annotated into either positive, negative, or neutral. In order to annotate the data, the data was split equally into two divisions. Specifically, each division was annotated by two annotators from the authors. Furthermore, the annotation was made blindly, which means that each annotator did not have access to the other annotators' annotation. At the end, the opinion of the two annotators got compared. If there was no agreement on the label, the tweet was eliminated.

The manual annotation process required quality assurance since human opinion varied in nature according to several aspects such as education, age,

culture, and more. Therefore, it is important to measure the level of agreement between different annotators. The inter-annotator agreement (IAA) is a measure that is used if there were multiple annotators who selected the same decision for a particular category. The IAA is used to validate and interpret the labeling result. Three common metrics are used the percent agreement, Cohen kappa, Fleiss' kappa to compute the IAA for the annotation process. Percent agreement is the simplest method, but it doesn't consider the possibility of random guesses annotation. Cohen kappa and Fleiss' kappa solve the drawbacks of Percent agreement. Cohen kappa is more suitable for two annotators while Fleiss' kappa is suitable for more than two annotators. So, the selection of the metrics is according to the number of individuals who annotate the corpus. Therefore, We used the Cohn kappa metric in this research since two annotators annotated each record. The metric obtains a result above 0.65, which consider as a substantial agreement and it means that the result of the annotation is suitable to train the classification models.

The final corpus was saved as a CSV format that contains five columns as shown in figure 1. the first column shows the cleaned text. the second column contain the name of the bank either AlRajhi, Alinma, Saudi Investment bank (SAIB) or the Saudi National banks (SNB). The third and fourth columns contain the tokenized words of each sentence, one with stop words and the other without stop words. The last column has the annotation label made by the annotators of the tweet. After the corpus pre-processing and annotation, we performed some statistical analysis on the corpus. The analysis shows that our corpus contains 1,567,628 tokens and 82,015 word types. In addition, according to the bank names: we gathered 7,775 tweets for AlRajhi bank, 2,705 for SNB, 920 for Alinma bank, and 648 for SAIB. Moreover, the results of the corpus annotations show that among the gathered tweets, 8,669 of them were labeled as "Negative", 2,143 were labeled as "Positive", and 1,236 tweets were labeled as "Neutral".

	A	B	C	D	E	F	G
1	Tweet	Bank	Tokens	Tokens without stop words	Annotator 1 & 3	Annotator 2 & 4	Final annotation
2	الله يقلعك من بنك دمر مستقبلنا الله حسبنا	SAIB	['بنك', 'من', '	['بنك', 'فيك', 'الوكيل', 'ونعم', 'ه	NEG	NEG	NEG
3	صار لي يومين يحاول اوصل لاي شخص من	SAIB	['وهمي', 'شغل	['وهمي', 'شغل', 'بنك', 'انتم', 'ملا	NEG	NEG	NEG
4	صادق كثير من الزملاء يشتكون منه الصراحه	SAIB	['تهاءيا', 'عنه	['تهاءيا', 'عنه', 'نظر', 'صرفت', 'ه	NEG	NEG	NEG
5	ماقدرنا ناخذ بطاقه مدي وحسابي ذهبي ومو	SAIB	['عندكم', 'كبير	['عندكم', 'كبير', 'خلل', 'الافضل'	NEG	NEG	NEG
6	رواتبهم نصها اقساط وش يدور من عروض	SAIB	['بعد', 'عروض	['عروض', 'يدور', 'وش', 'اقساط'	NEG	NEG	NEG
7	اما كذا وانا اتضجر علي دقيقه وانت دقيقه	SAIB	['للتمديد', 'بله	['للتمديد', 'قابله', 'سنه', 'وانا', 'ه	NEG	NEG	NEG
8	لوسمحتوا قاعدين نتصل عليكم محد يرد	SAIB	['استفسار', 'ي	['استفسار', 'عندي', 'يرد', 'محد'	NEG	NEG	NEG
9	لانورطون مثلي البنوك الثانيه افضل من واقع	SAIB	['خبره', 'واقع	['خبره', 'واقع', 'افضل', 'الثانيه', '	NEG	NEG	NEG
10	انا منهم ولازلت اسدد طيله الست سنوات	SAIB	['ماتغير', 'فسه	['ماتغير', 'القرض', 'مبلغ', 'ولازال'	NEG	NEG	NEG
11	احسنت ربي يبيلك الخير انا ماخذ من عندهم	SAIB	['وحط', 'شيل	['وحط', 'شيل', 'معهم', 'وانا', 'لك	NEG	NEG	NEG
12	الشركات الكبيره والي العملاء طق مطاقق	SAIB	['كذا', 'ظارهم	['كذا', 'انتظارهم', 'ماكان', 'مطاقق	NEG	NEG	NEG
13	ما انسي تجربه اخوي معاكم ووصفكم بانكم	SAIB	['البنك', 'هذا	['البنك', 'عمليه', 'جيد', 'التمويل'	NEG	NEG	NEG
14	اي والله ومشكله التعامل كذا	SAIB	['كذا', 'لتعامل	['كذا', 'التعامل', 'ومشكله', 'والله']	NEG	NEG	NEG
15	هذا البنك علي اساس من الزحمه الي عندهم	SAIB	['عندهم', 'لي	['عندهم', 'الي', 'الزحمه', 'اساس'	NEG	NEG	NEG
16	ما ادري يتردون بكره ولا كيف	SAIB	['كيف', 'ولا	['كيف', 'بكره', 'يتردون', 'ادري']	NEG	NEG	NEG
17	مو معقوله انتظر دقيقه علي الجوال والي	SAIB	['مشكلتي', 'حل	['مشكلتي', 'احل', 'علي', 'يرد', 'ي	NEG	NEG	NEG
18	كلمه حق بهذ البنك رقي في التعامل متفهمين	SAIB	['منه', 'حاجتك	['منه', 'حاجتك', 'قدر', 'خذ', 'ط	POS	POS	POS
19	للاسف تم احتسابها لي بطريقه خاطءه كيف	SAIB	['الاحتساب', '	['الاحتساب', 'علي', 'الاعتراض', 'م	NEG	NEG	NEG
20	مو طبيعي الانتظار في فلكس كول جلست	SAIB	['الخدمه', 'ذا	['الخدمه', 'ذا', 'وش', 'موضوعي',	NEG	NEG	NEG

Figure 1: Sample of the dataset file

4 Experiments and Results

4.1 Model Selection

The second stage in conducting the sentiment analysis on Saudi banks was the model selection. We had many experiments in order to find the best model that gains the highest accuracy in detecting sentiments. These experiments involved implementing three machine learning techniques which are: Support vector machine (SVM), Logistic regression (LR), Random Forest (RF). In addition, we implemented a fourth model using MarBERT model. The feature representation was implemented using the term frequency-inverse document frequency (TF-IDF). According to (Al-Twairesh et al., 2017), the TF-IDF representation shows a good performance on Arabic text. To train the model, we split the corpus into 80% for training and 20%for testing. The decision for selecting these models was based on the Arabic research community in which the result of the selected model showed good predictions especially for sentiment analysis. Different appropriate parameters were used for each model to obtain the desired performance.

The first machine learning model is the Support Vector Machine (SVM) which is a supervised machine learning algorithm. The second machine learning model is the logistic regression (LR) which is a supervised machine learning algorithm and used widely for binary classification problems. However, LR can be used to perform a multilabel classification task such as sentiment analysis. In the current work, we evaluated the corpus with multilabel logistic regression classification for sentiment analysis namely positive, negative, and neutral. The logistic regression has shown a successful result in sentiment analysis in diverse languages such as English and Arabic (Bessou and Aberkane, 2019). Thus, motivated by the promised result achieved in the previous study, we will

conduct a corpus evaluation using logistic regression and perform fine-tuning with regularization parameters using a grid search to obtain the best result. The model has been evaluated with the predefined parameters using 5-fold cross-validation. The logistic regression model has shown improvement in performance with 82% accuracy.

The third machine learning model is the Random forests algorithm. The Random forests algorithm is a supervised machine learning algorithm that builds many individual decision tree classifiers. The main advantage of using the Random forests algorithm is that the generated decision trees are not correlated, therefore the classification error made by one tree is not seen by the other decision trees. On the other hand, random forest algorithm suffers from the slowness disadvantage. We apply this algorithm by using Random Forest Classifier module, Randomized SearchCV() method, and 5-fold cross-validation. Randomized SearchCV() apply tunning on the parameters to choose randomly the optimal parameters.

The last experiment was conducted using the MarBERT model, which is a pre-trained language model based on the Arabic dialectical data from social media. We fine-tuned the model based on the model provided by Abdul-Mageed et al. on GitHub page[2]. The model hyperparameters were selected based on the original model where the learning rate equals to 2×10-6, the batch size equals to 32, the maximum sequence length equals to 128, and the epoch equals to 5.

4.2 Model Evaluation

For evaluating the models' performance, the following metrics were used: Accuracy, Precision, Recall, and F-measure.

For all classifiers, we have computed the TF-IDF method which plays a crucial role in the training stage to select the most important words among the dataset.

Table2 shows the performance of all models using the metrics: accuracy, precision, recall, and F1 score. Overall, the evaluation results were quite close among all models. However, the best performing result was achieved by SVM with 82.4% accuracy and an F1 score of 79%. On the contrary, logistic regression reported the most stable results in accuracy and F1 score with 82% and 81% which represent the highest result compared with other classifiers.

Model name	Accuracy	Precision	Re-call	F-measure
MarBERT	82.1	70.2	66.3	68.0
Logistic regression (LR)	82.0	**80.0**	**82.0**	**81.0**
Random Forest (RF)	81.0	78.0	81.0	77.0
Support Vector Machine (SVM)	**82.4**	**80.0**	**82.0**	79.0

Table 2: models' results

4.3 Error analysis

To demonstrate the model accuracy and justify the performance of the results. An error analysis process was performed on our best used model which is LR model; to show where the model succeeds and failed to classify the data and set observation for the model limitations.

Figure 2 shows the confusion matrix of the classification error rate for each category (sentiment labels) for the LR model. After walking through the confusion matrix for the model, it is clearly noticeable what categories were misclassified and identified errors made by the classifiers. To demonstrate more, we grouped all these errors and created an observation for common errors in the prediction against the original dataset. We observed that the most misclassified class is NEU label, it is mostly misclassified as NEG. We found this case 158 times out of 244 NEU label tweets. This represent 64% of the actual NEU data that means that most of the NEU data has been misclassified, we attributed the reason of the misclassification to the imbalanced dataset. The second most error was POS label misclassified as NEG; it occurred 133 times which represent 29.7% of the POS tweets. The least error cases were for the NEG where nit

[2] GitHub - UBC-NLP/marbert: UBC ARBERT and MarBERT Deep Bidirectional Transformers for Arabic

was misclassified as NEU; it appeared only 50 times which represent 2.9% of NEG tweets.

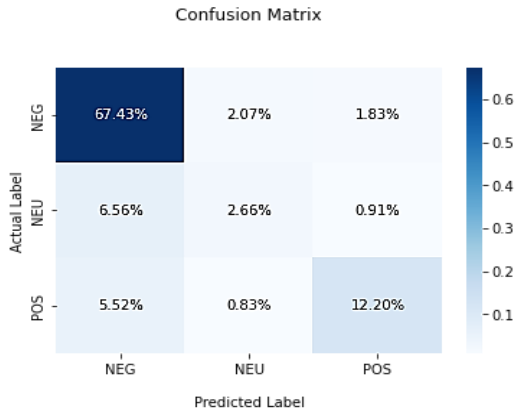

Figure 2: Confusion Matrix of LR model with percentage.

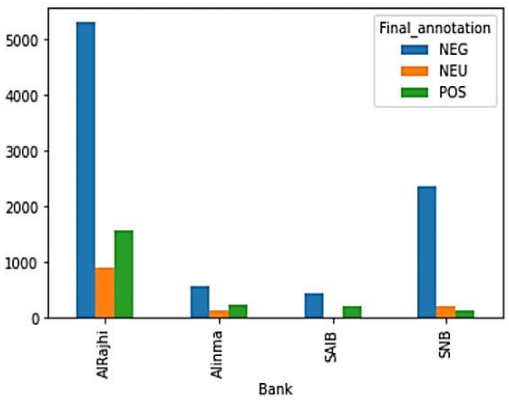

Figure 3: Distribution of tweets according to polarity: Positive, Negative, Neutral per bank.

We conclude that the misclassification errors occur as a result of imbalanced dataset. We noticed that because the NEG class represents about more than 70% of the dataset, that leads to misclassification of POS and NEU as NEG label. To deal with this issue we are using F-measure for evaluation. As a future work, this issue should be solved by increasing the size of the dataset to avoid dataset imbalance.

5 Discussion and Conclusion

In this study, we analyzed customers' sentiments on Twitter toward four Saudi Banks. A total of 50K raw tweets were scrapped using Twitter web scrapping tool. After data cleaning and preprocessing we ended up with 12k tweets. The dataset was manually annotated into three categories positive, negative, and neutral where we observed that among all classes the dominant tweets were pertained to negative sentiment. Additionally, as shown in Figure 3 we found that AlRajhi bank has the highest number of negative tweets followed by Saudi National Bank (SNB). Likewise, the positive and neutral tweets were the lower one.

We have utilized the annotated data to train three baseline classifiers namely Support vector machine (SVM), Random Forest (RF) and Logistic regression (LR) and fine-tuned one pretrained language model MarBERT. The baseline models were trained and tested on the same dataset which has been evaluated using 5-fold validation. Whereas MarBERT transformer model was trained and evaluated using the same evaluation method. The evaluation results show that the best accuracy result was achieved by logistic regression (LR) with 81.0 F1 score which outperforms the pre-trained model MarBERT that had achieved F1 score of 68%. Technically speaking, the accuracy result for all observed classifiers were very close to 82%. Other metrics have large differences between classifiers performance where the highest precision and recall results were 80% and 82% respectively which was achieved by two classifiers LR and SVM. Furthermore, we can conclude that overall SVM and LR models outperform the pre-trained model MarBERT in precision and recall as well as F1 score.

In the future, we plan to increase the dataset and try to prevent the overfitting problem using some techniques such as data augmentation as well as investigate the performance of deep learning models (e.g. BiLSTM and CNN) on the proposed dataset. We might also try to use ensemble models to improve the experiments results.

References

Almuqren, L., and Cristea, A. (2021). AraCust: A Saudi Telecom Tweets corpus for sentiment analysis. PeerJ Computer Science, 7, e510. https://doi.org/10.7717/peerj-cs.510

Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., and Al-Ohali, Y. (2017). AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets. Procedia Computer Science, 117, 63–72. https://doi.org/10.1016/j.procs.2017.10.094

Azunre, P. (2021, July). Transfer Learning for Natural Language Processing. Manning Publications. https://www.manning.com/books/transfer-learning-for-natural-language-processing

Bessou, S., and Aberkane, R. (2019). Subjective Sentiment Analysis for Arabic Newswire Comments. Journal of Digital Information Management, 17(5), 289. https://doi.org/10.6025/jdim/2019/17/5/289-295

Di, L., Shaiban, M. S., and Hasanov, A. S. (2021). The power of investor sentiment in explaining bank stock performance: Listed conventional vs. Islamic banks. Pacific-Basin Finance Journal, 66, 101509. https://doi.org/10.1016/j.pacfin.2021.101509

Eksa Permana, M., Ramadhan, H., Budi, I., Budi Santoso, A., and Kresna Putra, P. (2020a). Sentiment Analysis and Topic Detection of Mobile Banking Application Review. 2020 Fifth International Conference on Informatics and Computing (ICIC), 1–6. https://doi.org/10.1109/ICIC50835.2020.9288616

Eksa Permana, M., Ramadhan, H., Budi, I., Budi Santoso, A., and Kresna Putra, P. (2020b). Sentiment Analysis and Topic Detection of Mobile Banking Application Review. 2020 Fifth International Conference on Informatics and Computing (ICIC), 1–6. https://doi.org/10.1109/ICIC50835.2020.9288616

Elamir, E. A. H., and Mousa, G. A. (2020). Sentiment Analysis of Banks' Annual Reports and Bank Features: LASSO Approach. 2020 International Conference on Decision Aid Sciences and Application (DASA), 42–48. https://doi.org/10.1109/DASA51403.2020.9317075

Gavval, R., Ravi, V., Harshal, K. R., Gangwar, A., and Ravi, K. (2019). CUDA-Self-Organizing feature map based visual sentiment analysis of bank customer complaints for Analytical CRM. ArXiv:1905.09598 [Cs]. http://arxiv.org/abs/1905.09598

Kazmaier, J., and Vuuren, J. van. (2020). Sentiment analysis of unstructured customer feedback for a retail bank. ORiON, 36(1), 35–71. https://doi.org/10.5784/36-1-668

Krishna, G. J., Ravi, V., Reddy, B. V., Zaheeruddin, M., Jaiswal, H., Teja, P. S. R., and Gavval, R. (2019). Sentiment Classification of Indian Banks' Customer Complaints. TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 429–434. https://doi.org/10.1109/TENCON.2019.8929703

Olson, D., and Delen, D. (2008). Advanced Data Mining Techniques. In Springer. USA. https://doi.org/10.1007/978-3-540-76917-0