Sentiment Analysis:
US Consumer Finance Complaints

MGT 8020: Business Intelligence

Suvaleena Paul | #002638763

Contents

Abstract **3**

1. Introduction **4**

2. Literature Review **5**

3. Text Analytics **6 - 10**

3.1 Dataset **6 - 7**

3.2 Process **6**

Fig 1: Process in RapidMiner **6**

3.3 Data Analysis **7 - 9**

Fig 2: Companies against Count of Complaints **7**

Fig 3: Issue Count per Product and Submission Channel **8**

Fig 4: Response Status of Issues by Issue Count **8**

Fig 5: Word Cloud of Issues **9**

3.4 Text Processing: Process Documents **9 - 10**

Fig 6: Process Documents **9**

3.5 Cross Validation **10**

Fig 7: Model Setup **10**

4. Results and Discussion **11**

Fig 8: SVM Model **11**

5. Conclusion **11**

References **12**

ABSTRACT

This study aims to create a sentiment analysis model on **379,962** Consumer Complaints made for US Financial Institutions using machine learning and sentiment analysis from preliminary positive or negative sentiment. In the study, a vector space was created in the RapidMiner platform, and a classification study was performed on this vector space by Support Vector Machine (SVM) and Performance was checked for Classification model. The classification results for Consumer Complaints came up to 99.99% accurate by SVM. It is seen that the SVM algorithm calculates the best classification results presented in the data set. In the dataset taken, there seems to be **19** False Positives against **94084** True Positives. Hence, we can rightfully say that all complaints lodged are not negative, some are feedbacks and suggestions thus resulting in a neutral or positive sentiment scoring on VADER.

1. Introduction

Thousands of complaints and service request tickets are lodged daily for Financial Institutions and banks by consumers. Most of these complaints are regarding services and products offered by the company. US has a huge dataset for the complaints logged for more than a thousand financial institutions. In this project the top 50 companies are considered based on their Issue Count. A sentiment analysis shows the intensity of like or dislike towards a particular product or service. Some complaints might be concerns or doubts while others might be extreme negative feedback.

Financial Institutions can filter out these complaints based on the sentiment analysis to segregate a real issue or bug in their service from functional feedback. Further Topic Modelling can be run on the real issues to find out the recurring complaints and thus streamlining the service and maintenance unit of the company. A deeper analysis of the complaints by filtering out the company names can give an insight into which bank is preferred most and which company needs more upgradation of services.

The goal of sentiment analysis is to systematically identify, extract, quantify, and study affective states and subjective information using natural language processing, text analysis, computational linguistics, and biometrics. By using sentiment analysis, this paper attempts to predict the different cases within the Consumer Finance Complaints dataset. The complaint sentiments are classified as either positive or negative, and 379,962 complaints are examined to train the model and apply the model to a review dataset for the validation set. The sentiment analysis process in this study includes uploading data via CSV, data cleaning and preparation, set role, Sentiment Extraction using VADER, Text Processing steps include tokenization, word filtering, filtering stop words, followed by nominal to text, process documents, cross-validation, and applying the models.

2. Literature Review

Sentiment analysis is a technique to classify a given dataset or text into positive, negative, and/or neutral categories (Ahmad, Aftab, Bashir, & Hameed, 2018, p. 182). According to Drus & Khalid (2019), one method to conduct sentiment analysis is through machine learning, which applies algorithms to extract and explore sentimental elements of a dataset. Supervised machine learning methods, such as SVM and Naïve Bayes models, require a training dataset in text processing (Drus & Khalid, 2019, p. 5). Supervised learning uses labeled source data to create a training model, which then can be used to predict new, unlabeled data (Ligthart, Catal, & Tekinerdogan, 2021, p. 5004).

Three popular classification algorithms used for machine learning are SVM, Naïve Bayes, and Logistic Regression. Among them, SVM is the preferred algorithm for conducting sentiment classification because researchers have found that SVM has the best performance in classification tasks, while Naïve Bayes has also been shown to have good performance due to its comparative simplicity (Ligthart et al., 2021, p. 5029). Ahmad et al. (2018) state that the detection of sentiment polarity from a text is one popular application of SVM. That is, language is parsed to learn whether it should be categorized as positive or negative. At the same time, researchers have shown that identifying the polarity of language is not always simple (Ligthart et al., 2021, p. 5001). One suggestion to eliminate sentiment ambiguity is to utilize opinion-level context and the Bayesian model (Xia, Cambria, Hussain, & Zhao, 2015).

As mentioned above, the Naïve Bayes algorithm is well known for the simplicity of technique. It can perform more complex tasks with strong assumptions in high-dimensionality text (Dey et al., 2016, p. 2; Onalaja et al., 2021, p.5). Research shows that Naïve Bayes can reach a high accuracy in sentence polarity studies (Onalaja et al., 2021, p. 4). Another algorithm, Logistic Regression, can also be used to make predictions for a target variable that is either binomial or binary, such as yes or no, 0 or 1 (Kotu & Deshpande, 2015, p. 14). Logistic Regression is commonly used in online written genres that involve customer reviews, such as Amazon and Yelp reviews (Onalaja et al., 2021, p. 5). These authors applied Logistic Regression to movie reviews and found that it performs with a high degree of accuracy in sentiment analysis.

3. Text Analytics

3.1 Dataset:

As stated previously in the paper, the data used in this study are publicly available on Kaggle. The description extracted from the Kaggle platform is:

Each week the CFPB sends thousands of consumers' complaints about financial products and services to companies for response. Those complaints are published here after the company responds or after 15 days, whichever comes first. By adding their voice, consumers help improve the financial marketplace.

3.2 Process:

Figure 2 illustrates the process steps and the workflow diagram created in the RapidMiner. In this step, the data from the file is passed to the text extraction and vector space creation stages. Next, the classification process is conducted, and the classification algorithms' success is compared.

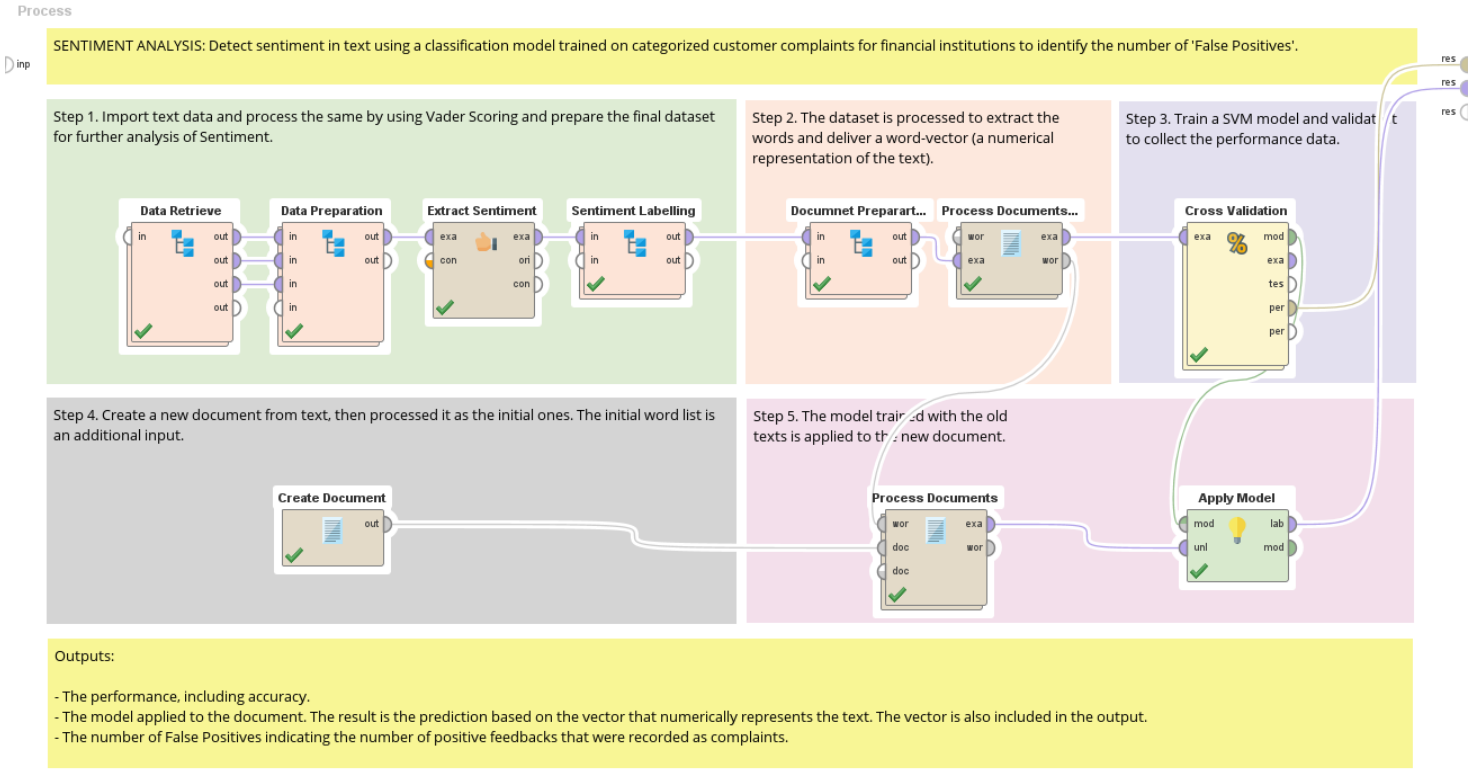

Figure 1: Process in RapidMiner

The Consumer Finance Complaint dataset was downloaded on Kaggle, a platform where users collaborate, find, and publish datasets, use GPU-integrated notebooks, and compete with other data scientists to solve data science challenges. In this workflow model, to add CSV files to the RapidMiner workflow model with the "Read CSV" and "Retrieve" nodes in the program, both the training and example data sets are loaded into the process.

3.3 Data Analysis:

After initial cleaning and pre-processing of data by removing missing values and low quality data, selecting important attributes and eliminating redundant data, the following insights were extracted from the dataset:

- The top three financial institutions are – **Bank of America** followed by **Wells Fargo & Company** and **J.P. Morgan Chase & Co.**

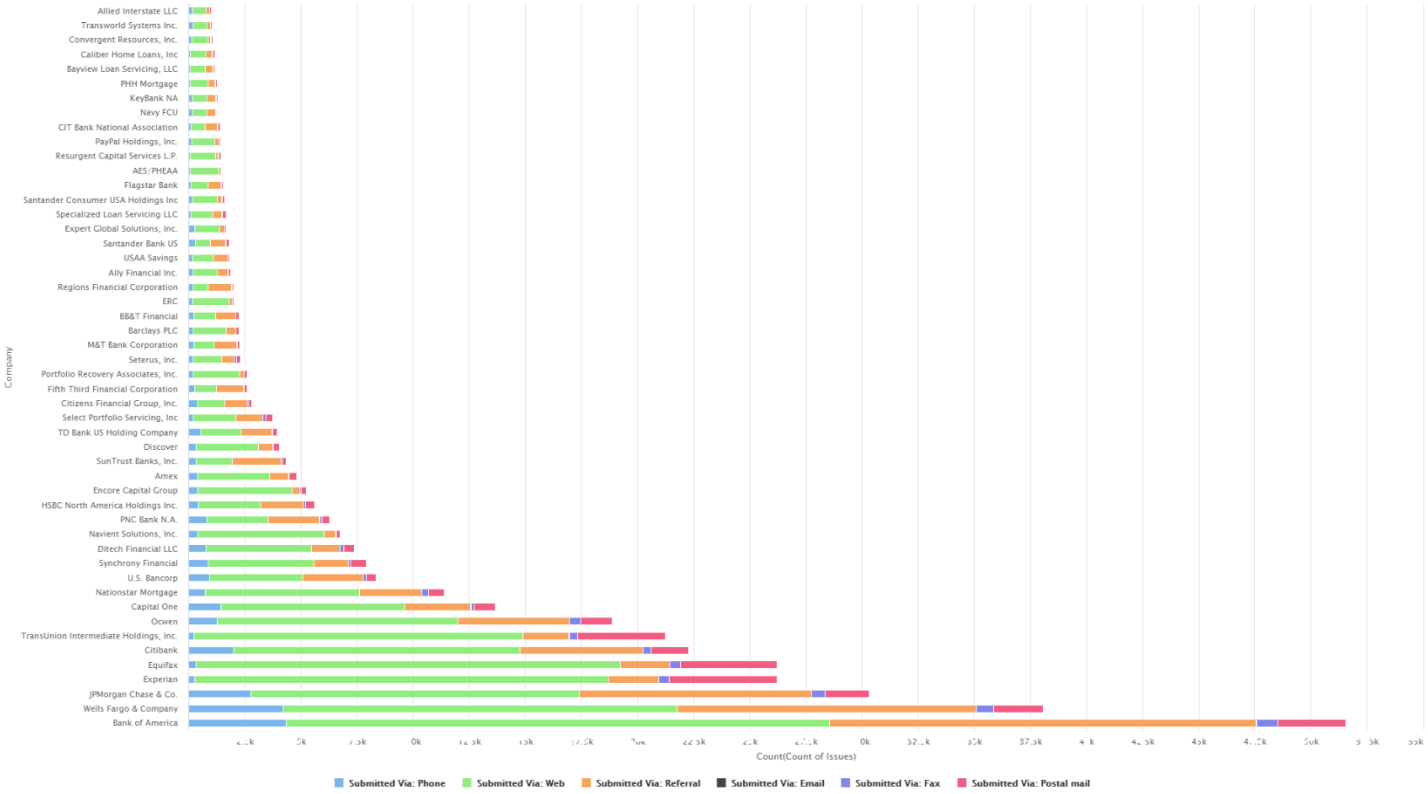

Figure 2: Companies against Count of Complaints

- Maximum complaints are submitted through **Web** channels followed by **Referrals, Postal Mails** and **Phone** and for **Mortgage** followed by **Credit Card** as the product type

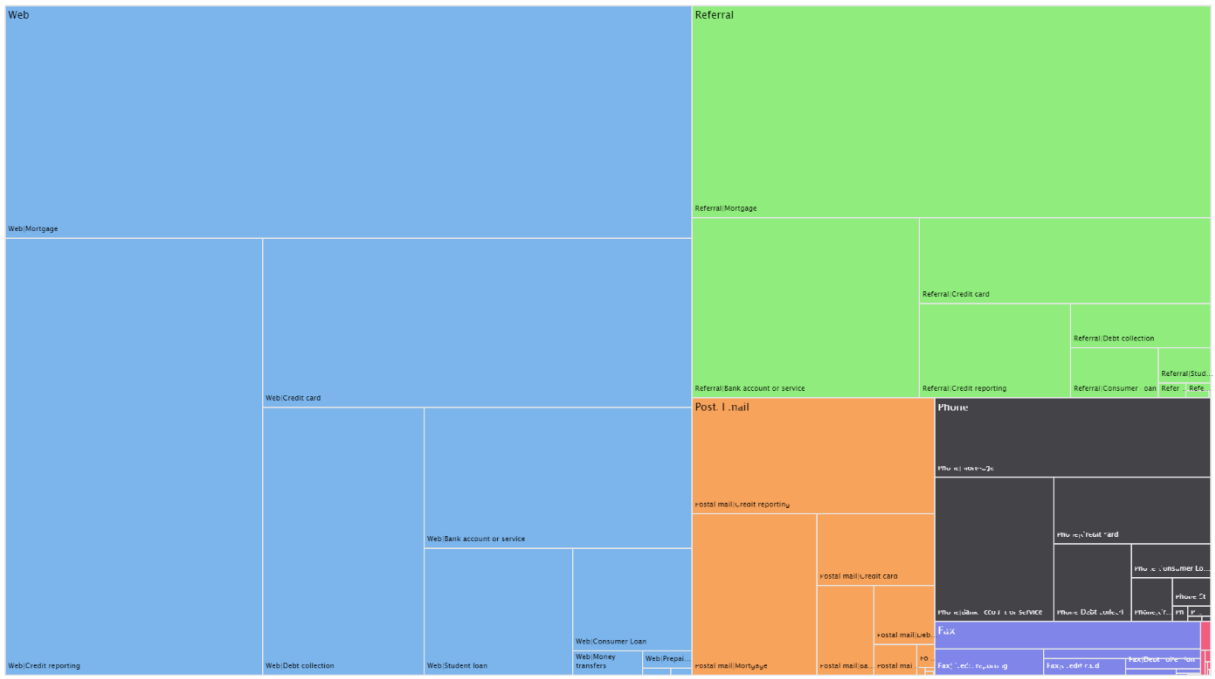

Figure 3: Issue Count per Product and Submission Channel

- More than **260000** complaints have been **Closed with Explanation**

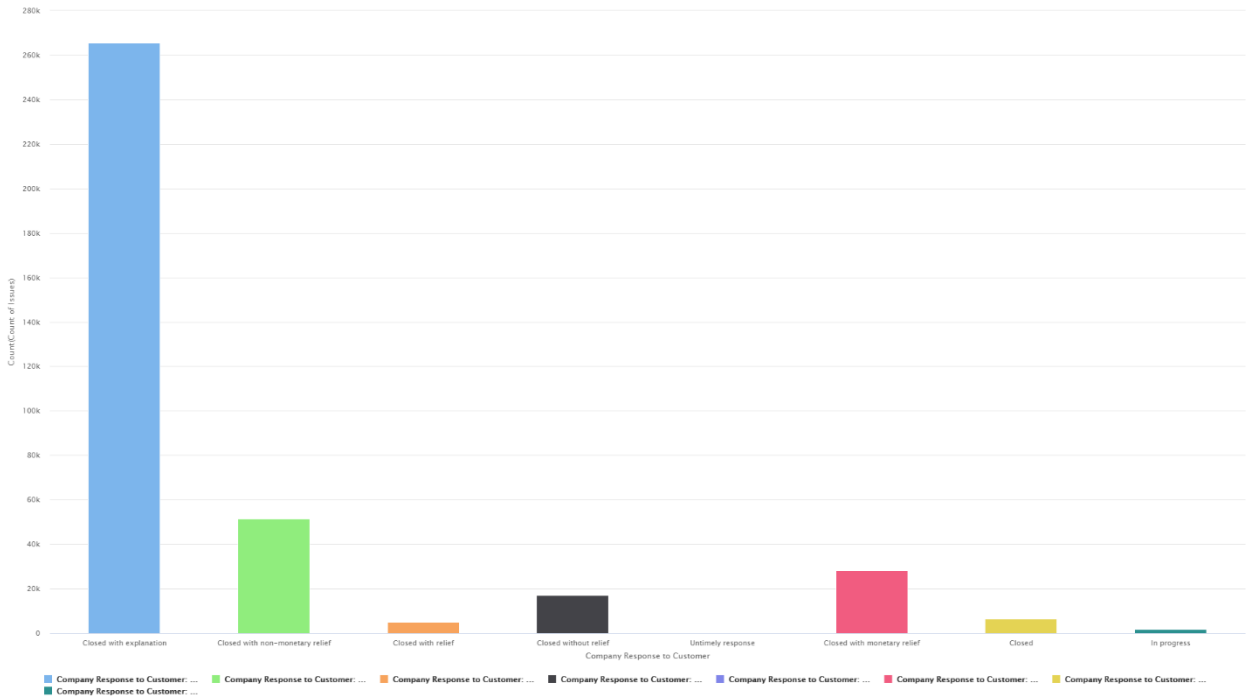

Figure 4: Response Status of Issues by Issue Count

- The most complained issue is concerning **Loan Modification, collection, and foreclosure**

Figure 5: Word Cloud of Issues

3.4 Text Preprocessing: Process Documents:

In this step, filters are added to the prepared data based on some rules. Pre-processing of the document is the preparation of the dataset before applying any algorithm on it (Sahu & Ahuja, 2015, p. 3). As a result, the polarity of the document can be labeled more quickly. The document is preprocessed as follows:

1. **Tokenize:** This operator splits the text of a document into a sequence of tokens. Tokens are sequences of characters in a document that are grouped together into a useful semantic unit.
2. **Transform Cases:** This operator transforms all characters in a document to either lowercase or uppercase. For this paper, lowercase was chosen.
3. **Filter Stopwords (English):** This operator filters English stopwords from a document by removing every token which equals a stopwords from the built-in stopwords list.

Figure 6: Process Documents

As shown in Figure 6, tokenization, transform cases, and filter stopwords (English) were used on the training and test data sets.

3.5 Cross Validation:

In this step, the SVM, Naïve Bayes, and Logistic Regression models are trained and are then validated to the example set to collect the performance data. In statistical analysis, cross-validation refers to determining how well the results will generalize to another independent data set. The model training is done as follows:

1. **Modeling operators:**
    a. **SVM:** This operator uses Stefan Reupping's internal Java implementation of the SVM learner (Support Vector Machine).
    b. **Naïve Bayes:** This type of classifier builds a good model even with a small amount of data. It has a high level of bias, but low level of variance.
    c. **Logistic Regression:** This operator is a Logistic Regression learner and is based on the myKLR by Stefan Reupping as well.
2. **Apply Model:** An example dataset is applied to a model using this operator.
3. **Performance:** A list of performance criteria values is returned by this operator for performance evaluation. Based on the type of learning task, these performance criteria are automatically determined.

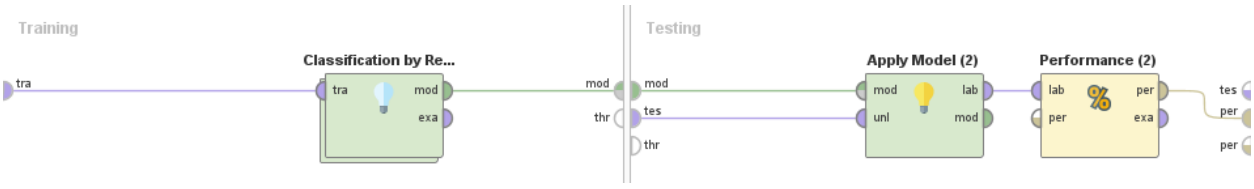

Figure 7: Model Setup

4. Results and Discussions

The classification results of applying machine learning model SVM to the US Consumer Finance Complaints dataset, as shown in figure 8 is 99.99%. The SVM model performs the classification task with the highest accuracy rate.

Although the SVM model and the Logistic Regression model both can predict relatively better results, and the SVM model overall is slightly more accurate than the Logistic Regression model, the precision and recall rates reflect the opposite results. To be specific, the class recall rate of true positive in the SVM model is better than that in the Logistic Regression model. Meanwhile, the class precision rate of predict negative in the SVM model is lower than that in the Logistic Regression model.

accuracy: 99.99% +/- 0.01% (micro average: 99.99%)

	true negative	true positive	class precision
pred. negative	285854	19	99.99%
pred. positive	5	94084	99.99%
class recall	100.00%	99.98%	

Figure 8: SVM Model

5. Conclusion

As a result of this sentiment analysis of **379,962** US Consumer Finance Complaints with RapidMiner, it is shown that the SVM model can produce the best classification. To better understand the sentimental polarity of consumer complaints, it is necessary to add different filters based on some rules. Among them, the text processing procedure is important because it prepares a pre-processed dataset to train, validate, and apply models. In addition, a pre-processed dataset also helps with eliminating unnecessary information and sentiment ambiguity that may impact the accuracy of results. This binary classification study analyzes people's positive and negative feedback on services and tells a clear difference of people's preference of one company over another depending on the products offered.

References

- Ahmad, M., Aftab, S., Bashir, M.S., & Hameed, N. (2018). Sentiment Analysis using SVM: A Systematic Literature Review. *International Journal of Advanced Computer Science and Applications, 9*(2), 182-188. https://doi.org/10.14569/IJACSA.2018.090226
- Amulya, K., Swathi, S. B., Kamakshi, P., & Bhavani, Y. (2022). Sentiment analysis on IMDB movie reviews using machine learning and deep learning algorithms. *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. https://doi.org/10.1109/icssit53264.2022.9716550
- Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using Naive Bayes and K-NN classifier. https://doi.org/10.48550/arXiv.1610.09982
- Drus, Z., & Khalid, H. (2019). Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. *Procedia Computer Science, 161,* 707-714. https://doi.org/10.1016/j.procs.2019.11.174
- Kotu, V., & Deshpande, B. (2015). *Predictive analytics and data mining.* ELSEVIER.
- Lakshmipathi, N. (2019, March 9). *IMDB dataset of 50K movie reviews*. Kaggle. Retrieved July 11, 2022, from https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
- Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review, 54,* 4997-5053. https://doi.org/10.1007/s12559-014-9298-4
- Mierswa, I. (2021). *Machine learning and RapidMiner tutorials: RapidMiner Academy*. Machine Learning and RapidMiner Tutorials | RapidMiner Academy. Retrieved July 17, 2022, from https://academy.rapidminer.com/
- Onalaja, S., Romero, E., & Yun, B. (2021). Aspect-based sentiment analysis of movie reviews. *SMU Data Science Review, 5*(3), 1-22.
- Sahu, T., & Ahuja, S. (2016). Sentiment Analysis of Movie Reviews: A Study on feature selection & classification algorithms. *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*. https://doi.org/10.1109/microcom.2016.7522583