S. García-Méndez, F. de Arriba-Pérez, Ó. Barba-Seara, M. Fernández-Gavilanes and F. J. González-Castaño, "Demographic Market Segmentation on Short Banking Movement Descriptions Applying Natural Language Processing," 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC), Rome, Italy, 2021, pp. 141-146, doi: 10.1109/ISCSIC54682.2021.00035.

keywords: Text mining; Terminology ; Finance; Process control; Europe; Banking; Natural language processing; Demographic market segmentation; Clustering analysis; Natural Language Processing; short banking movement descriptions; Personal Finance Management

# Demographic market segmentation on short banking movement descriptions applying Natural Language Processing

Silvia García-Méndez
Information Technologies Group,
atlanTTic, University of Vigo
E.I. Telecomunicación, Campus, 36310
Vigo, Spain
sgarcia@gti.uvigo.es

Francisco de Arriba-Pérez
Information Technologies Group,
atlanTTic, University of Vigo
E.I. Telecomunicación, Campus, 36310
Vigo, Spain
farriba@gti.uvigo.es

Óscar Barba-Seara
CoinScrap Finance S.L.
Cobián Roffignac 2, 36002 Pontevedra,
Spain
oscar.barba@coinscrap.com

Milagros Fernández-Gavilanes
Defense University Center, 36920 Marín,
Pontevedra, Spain
mfgavilanes@cud.uvigo.es

Francisco Javier González-Castaño
Information Technologies Group,
atlanTTic, University of Vigo
E.I. Telecomunicación, Campus, 36310
Vigo, Spain
javier@det.uvigo.es

## ABSTRACT

Banking movement descriptions can be a valuable type of short texts for knowledge extraction with application in finance and social studies. Conventional research on text mining has mostly been applied to medium-sized documents. Knowledge extraction from banking movement descriptions is challenging due to the lack of meaningful textual data and their ad-hoc terminology. In this work we present a clustering analysis on short banking movement descriptions based on Natural Language Processing techniques. We exploit the knowledge in an experimental data set composed of almost 20,000 real banking transactions that have been anonymised as required by European data protection regulations. At the end, we were able to extract five distinctive user clusters with similar demographics. Our approach has potential applications in Personal Finance Management.

## CCS Concepts

**• Information systems → Information system applications → Data mining → Clustering**

**• Information systems → Information system applications → Specialized information retrieval → Environment-specific retrieval → Enterprise search**

**• Information systems → Information system applications → Enterprise information systems → Enterprise applications**

**• Social and professional topics → User characteristics → Gender, Geographic characteristics, Age**

## Keywords

Demographic market segmentation, Clustering analysis, Natural Language Processing, short banking movement descriptions, Personal Finance Management

## 1. INTRODUCTION

Modern marketing approaches rely primarily on market segmentation techniques and algorithms to divide the heterogeneous customer mass into homogeneous groups based on common features and patterns. These features may be demographic [1], psychographic (lifestyle variables) [2] or behavioural (usage patterns and benefits sought) [3].

Empirical works have analysed the impact of demographics on banking [1], [4]. Particularly, demographic segmentation is of interest due to its measurable variables (age, gender, region, income, occupation, etc.) that are straightforward to gather. In fact, this continues to be the preferred marketing segmentation in the state of the art [1].

Moreover, short textual banking movement descriptions are characterised by their conciseness [5] (less than 150 words with little meaningful data). Furthermore, their vocabulary is not standardised and it is highly context-specific. This makes traditional Knowledge Extraction (KE) techniques useless. On the other hand, they are much less noisy than long texts.

In this work we seek to discover a small number of groups of clusters representing specific customer needs in the banking sector. We depart from the initial hypothesis that multidimensional socio-demographic features allow to discover cluster membership accurately. Our research is based on real banking movement descriptions of a large population that is

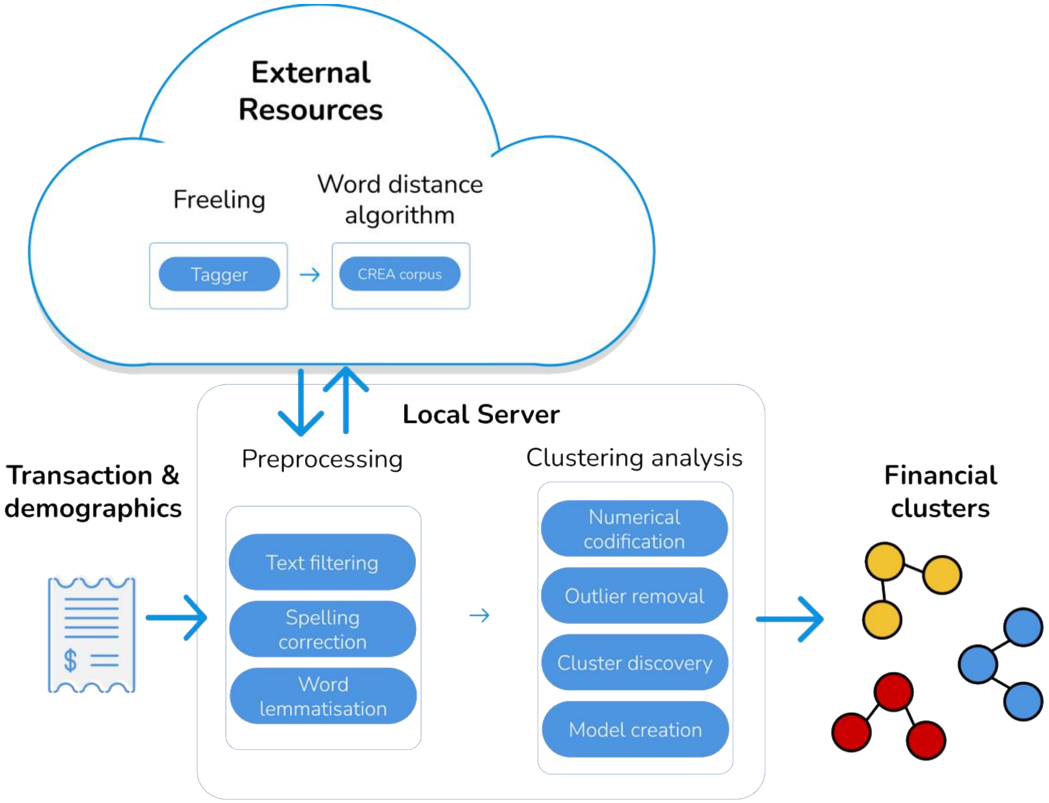

**Figure 1. System scheme.**

representative of banking customers in Spain. It contributes to market segmentation and financial institutions may exploit it to define their strategies.

The rest of this article is organised as follows. Section 2 reviews the state-of-the-art on financial KE systems. Section 3 describes our analysis goals and the clustering proposal. Section 4 presents our results. Finally, Section 5 concludes the paper.

## 2. RELATED WORK

Market segmentation consist in defining groups of customers based on statistical features, by applying cluster analysis [6], [7] or more sophisticated techniques like NLP [8], [9]. There exists a significant body of research on consumer profiling [10], [11], branding [2], risk analysis [12] and customer satisfaction [13], to cite some areas. More in detail, individual customer-level approaches have been followed in a variety of fields, such as stock market investment [1] and risk attitude assessment [14]. Some of these works focus on specific demographic or occupational groups like corporate customers [15].

More closely related to our approach, in which we seek a faithful reflection of the customers of the Spanish banking sector, are works such as [4] on demographic market segmentation. These usually consider age, gender and other demographic features that are easy to identify and measure. Particularly, in [1] the authors linked demographic features to biases in investment behaviour (overconfidence, self-attribution, emotional biases, etc.). They identified age, occupation and investment experience as the most relevant features.

Previous research on short text analysis for financial KE exploited vector representations of the texts [16]. However, short banking movement descriptions are specially challenging due to their representational sparsity [5]. In other words, the traditional Term-Frequency Inverse-Document-Frequency (TF-IDF) approach is not appropriate. It is here where more sophisticated approaches based on NLP techniques can be useful [8]. Thus, inspired by strategies that combine different segmentation techniques [17], the main contribution in this work is a market segmentation approach based on a clustering analysis that not only considers demographic features (such as those in financial research [6], [18]) but also NLP applied to short text banking transaction descriptions.

## 3. SYSTEM DESCRIPTION

In the following sections we will describe our clustering system for short banking transaction movement descriptions. Figure 1 shows its scheme.

### 3.1 PREPROCESSING

Prepossessing is needed to discard irrelevant and noisy data to deliver good quality data to the clustering analysis. It comprises the following steps:

- Text filtering. We apply regular expressions to the movement descriptions to detect and remove numerical identifiers, URLs and symbols as well as non-Unicode characters. We removed stop words like prepositions, conjunctions and determiners. To discard words without a significant semantic load, we keep only those with a minimum length of 4 characters. Finally, we discard non-Spanish words.
- Spelling correction. The terms in banking movement descriptions are usually shortened due to length constraints. Thus, we correct and complete them using the word distance algorithm and the Spanish frequency

reference corpus (CREA) by *Real Academia Española de la Lengua*[1].

- Word lemmatisation. We tokenise the short text banking movement descriptions and then lemmatise each element.

## 3.2 CLUSTERING

Our ultimate objective is to identify homogeneous clusters of users based on patterns in short text banking movement descriptions and demographic data about the customers. We perform a non-hierarchical K-means cluster analysis using all features, both demographic and NLP-derived.

Textual demographic features (type of payment, banking institution, gender and employment) are translated to a numerical space. The movement description itself is expanded as word-grams composed of $n$ adjacent words of the banking movement descriptions.

Next, we remove the outliers from the sample by inspecting the distribution of the entries in the data set according to the different features. We discard the entries with upper and lower extremes beyond the respective 1% quartiles.

To compute the optimal $K$ value, we use the mean distance between the data points and their centroids.

# 4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we analyse the demographic clusters discovered by our analysis of banking movement descriptions and their metadata. First, we study the relevance of the features for the clustering model. Then, we present insightful information about the clusters we identified.

## 4.1 EXPERIMENTAL SETTING

The analysis was performed with a computer with the following characteristics:

- Ubuntu 18.04.2 LTS 64 bits
- Intel@Core i9-9900K 3.60 GHz
- 32GB DDR4 RAM
- 500 GB (7200 rpm SATA) hard disk + 256 GB SSD ROM

We used the following software implementations:

- Text filtering: we removed Spanish stop words with NLTK[2], and we discarded non-Spanish words with the Enchant Python module[3].
- Spelling correction: we apply the word distance algorithm of the Enchant Python module.
- Word lemmatisation: we employ the Freeling [19] library[4].
- Translation of textual demographic features to a numerical space: OrdinalEncoder from the Scikit-learn Python library[5]
- K-means clustering: Scikit-learn Python library[6].

## 4.2 DATA SET

The fin-tech startup CoinScrap Finance S.L. provided us with a data set[7] composed of 19,243 real banking movement descriptions plus associated demographic metadata of the users who made the transactions. Each entry in the experimental data set is composed of:

- Transaction ID: unique transaction identifier (not used for clustering).
- User ID: unique user identifier (not used for clustering)
- Description: short text banking movement description.
- Demographic features:
    - Type of payment: indicates if the user paid with cash, debit or credit card.
    - Transaction amount: monetary amount of the transaction.
    - Banking institution: institution from which the transaction was issued (Abanca, Banco Sabadell, Banco Santander, BBVA, Bankia, Caixabank or ING Direct).
    - Gender: user gender.
    - Age: user age.
    - Occupation: user occupation.
    - Employment: active on behalf of others, active on own account, retired or unemployed.
    - Income: monthly user income.
    - Savings: user savings.
    - Type of location: rural or urban.

Feature type of location was not available in principle. We generated it ourselves from the user post codes using the information in the Spanish National Statistics Institute (INE[8]). We considered that the areas with more than 50,000 citizens were urban. Otherwise, we considered them rural.

Once the outliers were removed as described in Section 3.2, by considering the transaction amount, income and savings features, 15,863 valid entries remained. Each entry was then expanded with 1,000 extra word-gram features each. To generate the word-grams matrices we used CountVectorizer[9] from the Scikit-Learn Python library, with the configuration parameters in Listing 1.

[1] Available at http://corpus.rae.es/lfrecuencias.html, June 2021.

[2] Available at https://www.nltk.org, June 2021.

[3] Available at https://pypi.org/project/pyenchant, June 2021.

[4] Available at http://nlp.lsi.upc.edu/freeling/node/1, June 2021.

[5] Available at https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder.html, June 2021.

[6] Available at https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html, June 2021.

[7] We will make this data set available to other researchers on request.

[8] Available at https://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254734710984, June 2021.

[9] Available at https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html, June 2021.

**Listing 1. Configuration parameters of the *CountVectorizer* function for the generation of word-grams.**

```
analyzer = `word',
tokenizer = basic_tokenize, #This method split words
lowercase = True,
max_df = 0.3,
ngram_range=(1,1),
max_features=1,000,
strip_accents= `ascii'
```

For the clustering analysis, we trained a K-means model by varying K between 1 to 20. We observed in Figure 2 that for K = 5, the curve of the quality criterion tends to a constant value.

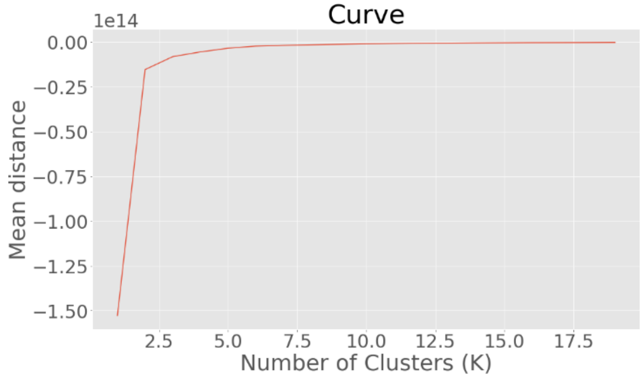

**Figure 2. K-means mean distance deviation.**

## 4.3 DISCUSSION OF DATA SET COHERENCE

We analysed the correlation between numerical (age, income, and savings) and categorical demographic features (type of payment, employment and banking institution). Note that for the sake of clarity and conciseness we only included the most representative ones.

First, figures 3 and 4, show the relation between type of payment and the numerical demographic variables age and income. In Figure 3 we can observe that debit cards are extensively used by people between 20 and 35 years old, while other methods of payment (cash and credit card) are more common in older people. Figure 4 suggests that cash and credit card are the preferred methods of payment for people with lower income, while debit cards are uniformly used.

Second, in figures 5 and 6, the relation between employment feature and the demographic age and savings features is more complex. In Figure 5 unemployed and retired users are dominant as age decreases and increases, respectively. This coherence simply endorses the representativeness of the experimental data set. Furthermore, in light of Figure 6, we notice that retired and active people working on their own account have the higher savings, as expected.

Finally, given the varied banking institutions in the data set, Figure 7 is interesting because it reflects the relation between them and the age feature of their customers. Note that ING Direct and BBVA are the most popular banks among young people, while Abanca, Bankinter and Caixabank are preferred by seniors, as it could also be expected from the marketing strategies of these companies.

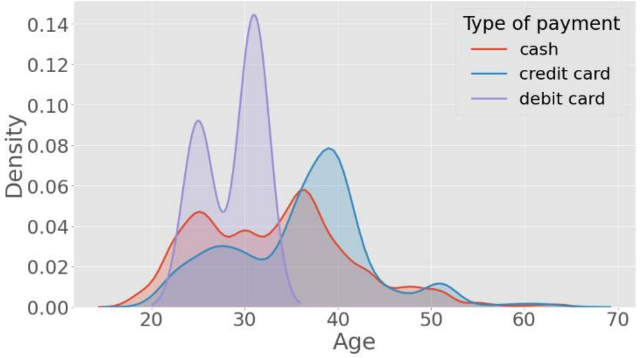

**Figure 3. Type of payment, versus age.**

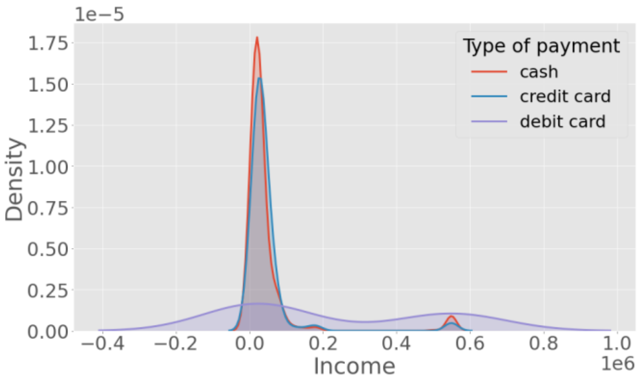

**Figure 4. Type of payment, versus income.**

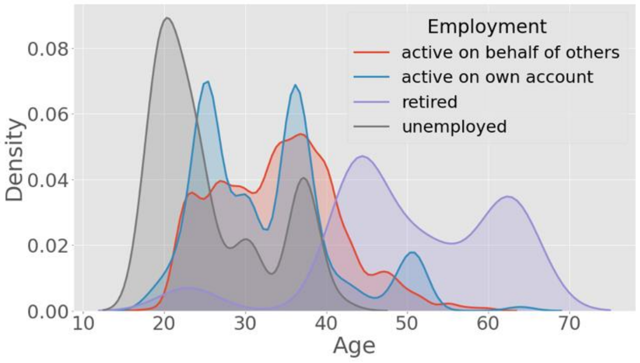

**Figure 5. Employment, versus age.**

**Table 1. Clustering analysis based on financial and demographic features.**

| | Payment | | Gender | | Employment | | | Location | |
|---|---|---|---|---|---|---|---|---|---|
| | Debit | Credit | Female | Male | Non-Freelance | Freelance | Non-active | Urban | Rural |
| C1 | | X | X | | | | X | | X |
| C2 | X | | | X | | X | | X | |
| C3 | | X | | X | X | | | X | |
| C4 | | | | X | X | | | | X |
| C5 | | | | X | | X | | | X |

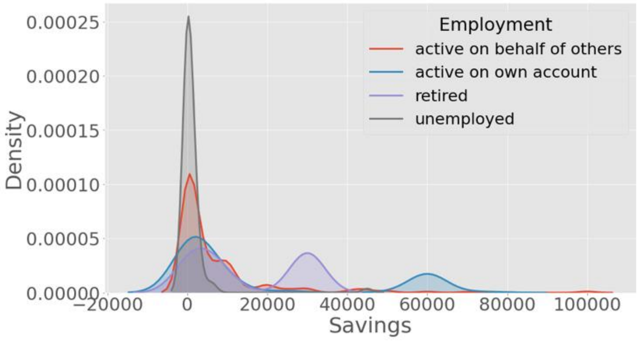

**Figure 6. Employment, versus savings.**

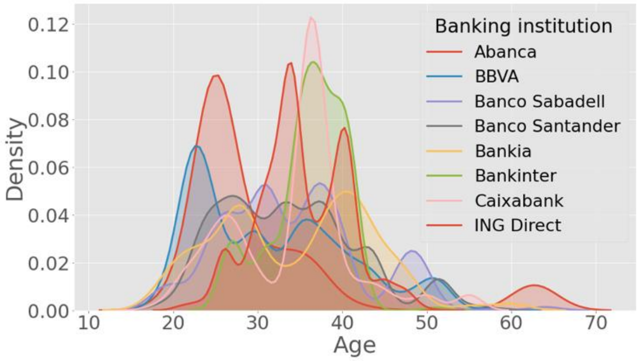

**Figure 7. Banking institution versus age.**

## 4.4 DISCUSSION OF CLUSTERING RESULTS

A differential aspect of this work is the combined application of demographic segmentation and NLP of movement descriptions to perform the clustering of banking movements. That is, word-grams contribute along with the demographic and numerical features to the discover of clusters. As a result, the analysis produces five well differentiated clusters in the experimental data set.

Let us first discuss them from the perspective of the demographic features involved. As we can observe in Table 1, cluster 1 is composed of women who are either retired or unemployed and predominantly use debit cards as preferred method of payment. The rest of clusters correspond mostly to male population. Particularly, cluster 2 groups users of debit cards who work on their own account. Conversely, users in cluster 3 work for an employer and prefer credit cards. Note that both clusters 2 and 3 belong to urban populations. Finally, clusters 4 and 5 correspond to rural areas, respectively to employees and freelances.

Table 2 shows the intra-cluster averages and standard deviations of demographic features age, income and savings. Cluster 1 contains the customers with lower income and savings levels. Clusters 2 and 3 contain the youngest and oldest users in the study, respectively. Surprisingly, the cluster 2 has the highest level of income and the cluster 3 the highest savings. Clusters 4 and 5 represent a mix of all demographic features compared to the other three clusters.

**Table 2. Intra-cluster averages and standard deviations of the demographic features.**

| | Age | Income | Savings |
|---|---|---|---|
| C1 | 32.4±8.1 | 19,866.6±11,171.5 | 3,657.3±5,815.0 |
| C2 | 26.1±5.0 | 546,498.7±10,550.2 | 2,991.2±455.8 |
| C3 | 43.5±5.8 | 166,789.0±16,171.9 | 57,009.2±30,445.7 |
| C4 | 38.4±6.9 | 69,747.9±18,632.7 | 8,033.8±10,662.8 |
| C5 | 37.5±5.0 | 33,126.7±14,382.3 | 54,205.1±12,231.1 |

The effect of the descriptions to movement clustering can be evaluated a posteriori by inspecting each resulting cluster with a novel system that combines NLP with Machine Learning to classify transaction descriptions [5]. Table 3 shows the dominant classes per cluster in decreasing relevance order. We can observe that the class patterns of the clusters are clearly different.

## 5. CONCLUSIONS

In this article paper we describe a clustering scheme to analyse banking movement descriptions. As far as we know this is the first study of these short texts combining demographic data with NLP techniques. Experimental results on a real data set monitoring the activity of the customers from Spanish banks show coherent behaviours and the appearance of informative user groupings.

Based on these promising results obtained, we plan to apply this analysis to banking transactions of other countries to study regional peculiarities.

As future work, we will analyse the impact of the date feature, whose periodicity is correlated with certain banking transactions and perform finer analyses based on the occupation feature (e.g. professional fields).

## ACKNOWLEDGMENTS

This work was supported by Xunta de Galicia grant GRC 2018/053, Spain. The authors are indebted to CoinScrap Finance S.L. for providing us with the experimental data set.

**Table 3. Intra-cluster five most dominant transaction classes.**

| | Transaction class | Cluster share (%) |
|---|---|---|
| C1 | Shopping | 31.2 |
| | Leisure | 16.5 |
| | Transfers | 8.9 |
| | Means of transport | 6.9 |
| | Financial expenses | 5.6 |
| C2 | Means of transport | 19.6 |
| | Leisure | 18.2 |
| | Shopping | 10.2 |
| | Financial expenses | 7.9 |
| | Transfers | 6.4 |
| C3 | Shopping | 34.9 |
| | Leisure | 11.5 |
| | Transfers | 10.1 |
| | Financial expenses | 8.3 |
| | Household expenses | 5.1 |
| C4 | Shopping | 40.3 |
| | Leisure | 12.4 |
| | Means of transport | 8.3 |
| | Transfers | 6.6 |
| | Financial expenses | 4.7 |
| C5 | Shopping | 24.8 |
| | Means of transport | 11.5 |
| | Leisure | 9.8 |
| | Financial expenses | 7.1 |
| | Transfers | 6.6 |

# REFERENCES

[1] H. K. Baker, S. Kumar, N. Goyal, and V. Gaur, "How financial literacy and demographic variables relate to behavioral biases," *Manag. Financ.*, vol. 45, no. 1, pp. 124–146, Jan. 2019, doi: 10.1108/MF-01-2018-0003.

[2] L. Gajanova, M. Nadanyiova, and G. Lazaroiu, "Specifics in Brand Value Sources of Customers in the Banking Industry from the Psychographic Point of View," *Cent. Eur. Bus. Rev.*, vol. 9, no. 2, pp. 1–18, 2020, doi: 10.18267/j.cebr.232.

[3] T. C. Phan, M. O. Rieger, and M. Wang, "Segmentation of financial clients by attitudes and behavior," *Int. J. Bank Mark.*, vol. 37, no. 1, pp. 44–68, Feb. 2019, doi: 10.1108/IJBM-07-2017-0141.

[4] S. Varma and R. Gupta, "Impact of demographic variables on factors of customer satisfaction in banking industry using confirmatory factor analysis," *Int. J. Electron. Bank.*, vol. 1, no. 4, p. 283, 2019, doi: 10.1504/IJEBANK.2019.10022902.

[5] S. García-Méndez, M. Fernández-Gavilanes, J. Juncal-Martínez, F. J. Gonzalez-Castaño, and O. Barba Seara, "Identifying Banking Transaction Descriptions via Support Vector Machine Short-Text Classification Based on a Specialized Labelled Corpus," *IEEE Access*, vol. 8, pp.61642–61655,2020,doi:10.1109/ACCESS.2020. 2983584.

[6] D. Kamthania, A. Pawa, and S. Madhavan, "Market Segmentation Analysis and Visualization using K-Mode Clustering Algorithm for E-Commerce Business," *J. Comput. Inf. Technol.*, vol. 26, no. 1, pp. 57–68, 2018, doi: 10.20532/cit.2018.1003863.

[7] D. Arunachalam and N. Kumar, "Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making," *Expert Syst. Appl.*, vol. 111, pp. 11–34, 2018, doi: 10.1016/j.eswa.2018.03.007.

[8] A. Ahani, M. Nilashi, O. Ibrahim, L. Sanzogni, and S. Weaven, "Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews," *Int. J. Hosp. Manag.*, vol. 80, pp. 52–77, Jul. 2019, doi: 10.1016/j.ijhm.2019.01.003.

[9] H. Liu, Y. Huang, Z. Wang, K. Liu, X. Hu, and W. Wang, "Personality or Value: A Comparative Study of Psychographic Segmentation Based on an Online Review Enhanced Recommender System," *Appl. Sci.*, vol. 9, no. 10, p. 1992, 2019, doi: 10.3390/app9101992.

[10] B. Sluban, M. Mikac, P. Kralj Novak, S. Battiston, and I. Mozetič, "Profiling the EU lobby organizations in Banking and Finance," *Appl. Netw. Sci.*, vol. 3, no. 1, p. 44, Dec. 2018, doi: 10.1007/s41109-018-0099-7.

[11] J. Cui, C. Yan, and C. Wang, "ReMEMBeR: Ranking Metric Embedding-Based Multicontextual Behavior Profiling for Online Banking Fraud Detection," *IEEE Trans. Comput. Soc. Syst.*, pp. 1–12, 2021, doi: 10.1109/TCSS.2021.3052950.

[12] Y. Pan, L. Zhang, X. Wu, and M. J. Skibniewski, "Multi-classifier information fusion in risk analysis," *Inf. Fusion*, vol. 60, pp. 121–136, 2020, doi: 10.1016/j.inffus.2020.02.003.

[13] J. Zhou, L. Zhai, and A. A. Pantelous, "Market segmentation using high-dimensional sparse consumers data," *Expert Syst. Appl.*, vol. 145, p. 113136, May 2020, doi: 10.1016/j.eswa.2019.113136.

[14] J. Belás, J. Dvorský, J. Kubálek, and L. Smrčka, "Important factors of financial risk in the SME segment," *J. Int. Stud.*, vol. 11, no. 1, pp. 80–92, Mar. 2018, doi: 10.14254/2071-8330.2018/11-1/6.

[15] L. Ronda, C. Valor, and C. Abril, "Are they willing to work for you? An employee-centric view to employer brand attractiveness," *J. Prod. Brand Manag.*, vol. 27, no. 5, pp. 573–596, 2018, doi: 10.1108/JPBM-07-2017-1522.

[16] M. Bounabi, K. El Moutaouakil, and K. Satori, "A probabilistic vector representation and neural network for text classification," in *Commun. in Comput. and Inf. Sci.*, vol. 872, Springer, 2018, pp. 343–355.

[17] J. An, H. Kwak, S. Jung, J. Salminen, and B. J. Jansen, "Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data," *Soc. Netw. Anal. Min.*, vol. 8, no. 1, p. 54, 2018, doi: 10.1007/s13278-018-0531-0.

[18] C. Wang and D. Han, "Credit card fraud forecasting model based on clustering analysis and integrated support vector machine," *Cluster Comput.*, vol. 22, no. S6, pp. 13861–13866, Nov. 2019, doi: 10.1007/s10586-018-2118-y.

[19] L. Padró and E. Stanilovsky, "FreeLing 3.0 : Towards Wider Multilinguality," *Proc. Lang. Resour. Eval. Conf. ELRA*, pp. 2473–2479, 2012.