ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 01, 2023

BOTTOM-UP METHODS FOR E-COMMERCE PRODUCT MAP ANALYSIS FROM A LARGE GROUP OF SMALL AND MEDIUM-SIZED WEBSITES

Mr A Pullaiah, Assistant Professor, apualliah@rishiubrcollege.ac.in, Rishi UBR Women's College, Kukatpally, Hyderabad 500085

ABSTRACT

Academics and industry professionals have shown considerable interest in the study of product maps in e-commerce because of the insights they give regarding the relationship between items, including their complementarity and rivalry. Existing research, however, has mostly relied on the input of big manufacturers and merchants and data collected from these authoritative sources. In this article, we use a crowdsourced bottom-up methodology, utilising SME e-commerce sites as the primary sources of information. This method permits distributed data processing and the synthesis of viewpoints and information from several distinct sources. In order to quantify product similarities and construct a product map, a graph term frequency-inverse document frequency approach is provided. Using information on more than 90,000 goods from 52 SME sites, a hierarchical community structure was discovered using this technique. The findings revealed a localised distribution pattern for goods produced on the same location. Our research may help online marketplaces increase their product variety via more calculated pricing and inventory management.

Keywords: product map; bottom-up; crowd science; small and medium e-commerce sites

An organization's homogeneity or heterogeneity in terms of the items they offer is shown on a product map. The degree to which items are replacements or complements may be gleaned from an analysis of the product map; this information is crucial for marketing and pricing, particularly in the realm of e-commerce. To achieve the optimal balance between product variety, product depth, and service level, retailers must make decisions about (1) the types of products to stock, (2) the number of products to stock in each category, and (3) the total number of products to stock in each stock keeping unit. They also need to think about things like which items might benefit from being promoted at the same time and how to personalise coupon activities. Consumers' purchasing decisions are influenced by a store's selection, which in turn affects the store's revenue. Clearly, the e-commerce scene has changed in the wake of the worldwide COVID-19 epidemic, with online retail sales rising and the percentage of total retail sales rising from 16% in 2019 to 19% in 2020. Additionally, in 2020, e-commerce sales in Europe increased by 10%. More and more SME sites are springing up as a result of recent studies in international entrepreneurship showing that e-commerce can help small and medium-sized businesses grow by providing access to previously inaccessible markets and customer bases. However, gaining insight into the relationships between goods of SME sites is impossible via analysis of product maps from individual SME sites. Instead, the whole of the SME websites have to be evaluated in reverse.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 01, 2023

Numerous research have looked at product map development to analyse the connection between individual items or classes of products. Elrod et al.'s analysis of customers' reactions to similar and dissimilar items is often cited as the first major research on product maps. Because of the sparseness of the available data, product map studies, according to Elrod et al., tend to zero down on the replacement of narrowly defined product categories. since a result, the definition of product categories is too rigid, since it does not allow for the mixing of most items and does not account for the links among some categories. That is to say, the product map is lost on this category-based approach.

Recent advances in machine learning (ML) and the abundance of high-quality data have opened up a plethora of possibilities for analysing product maps. Manufacturers of consumer durable goods, for instance, may benefit from Kim et al.'s descriptive model of online product searches by obtaining a product-centric visualisation of the industry's competitive structure. Using a text-mining strategy, Netzer et al. showed that it was possible to build a product map using information gleaned from user-generated material. Gabel et al. recently brought out the product map based on shopping baskets, which uses a low-cost data source that all stores have access to via their checkout systems. All of the aforementioned studies are done from the vantage point of manufacturers and major merchants, using their internal data as the basis for their findings. Thus, supermarkets served as models for previous study items. Our study subjects, on the other hand, are a collection of locally owned shops.

Although the volume of a single shop may seem little, when added together it becomes impossible to ignore. Research into "user and product portraits" is also noteworthy. This involves analysing and integrating many types of data, such as user behaviour data, product feature data, etc., to extract information about users' preferences, behaviour patterns, consumption capacities, and so on. Users' personal data and privacy may be at risk with this method, however. Our research relies heavily on information that is both easily accessible and publicly available: the textual product descriptions included on SME sites. That's why we're doing this research in the first place; to draw a product map by looking at how various items relate to one another.

This study presents a decentralised method, inspired by crowd intelligence, in which each SME site is treated as a little and autonomous data source. This implies that all websites may access the same extensive database of product details via their webpages. Insights into specific product maps are gleaned purely from publicly accessible data on web pages, as opposed to depending on predetermined information about item qualities on the sites of major firms. This means it can do analyses of product maps even when access to other data sources is limited or prohibitively costly.

Our goals in this study were to (1) apply our method to the data we gathered from 52 SME e-commerce sites and (2) create a graph term frequency-inverse document frequency (TF-IDF) for analysing product maps tailored to the requirements of SMEs.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 01, 2023

The suggested graph TF-IDF uses a combination of TF-IDF and NLP to create a visual network in the form of a graph reflective of a product map. Gephi, an open-source network analysis and visualisation programme written in Java and based on the NetBeans platform, is responsible for creating this diagram. What makes this paper new are the following. We begin by modifying the TF-IDF model so that it may be used to model data from SME websites. We then utilise this information to construct a network graph in which we can quantitatively depict the degree to which various items compete with one another. In addition, we indirectly assess the competitive nature of SME locations. Third, we construct a visual product map by loading this network structure into Gephi.

This paper's remaining sections are structured as follows. Section 2 describes the methodology and the outcomes, whereas Section 1 explicitly presents the dataset. In Section 3 we provide our final findings.

1 Method and Dataset

1.1 TF-IDF

In natural language processing and data retrieval, the TF-IDF is a standard tool. In information retrieval and text mining, it is used as a weighting factor to determine how significant a given word is in comparison to the whole corpus of texts.

TF-IDF is predicated on the hypothesis that words which appear often in a text are more likely to be significant, and that a word's significance is correlated with how uncommon it is over the full corpus of documents. This method gives more importance to words that are common in one document but not the other than to words that are common in both the document and the corpus.

The frequency with which a word occurs in a document is measured using TF, although it stands to reason that longer papers will have more opportunities to utilise the phrase. Therefore, standardising the "term frequency" is necessary to fix this issue. Term t's TF in document d is calculated using the occurrence count and the total number of terms in the document (as indicated in Eq. (1)).

$$\mathrm{tf}_{(t,d)} = \frac{f_{(t,d)}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

For instance, document "D" has 100 words, and the word "jeans" occurs 5 times; thus, we can calculate the TF value as follows:

$$\mathrm{tf}_{(\text{"jeans"}, D_1)} = \frac{5}{100} = 0.05.$$

In the meanwhile, we employ IDF to filter out certain frequent words. Even if stop words like "a" and "the" are not included in the TF calculation, the algorithm evaluates all terms equally. However, certain frequent words like "and" that cannot be separated in documents get enormous term frequencies, despite the fact that they are not significant terms. This issue is addressed by IDF by giving less importance to frequently used phrases and more importance to seldom used ones.

$$\mathrm{idf}_{(t,D)} = \ln \frac{N}{|d \in D : t \in d|} \tag{2}$$

for some document D, where N is the total number of documents and D is the single document in question.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 01, 2023

In our corpus of 1000 texts, the word "jeans" occurs 101 times thanks to Activat. The IDF may be calculated as follows, using Eq. (2):

$$\mathrm{idf}_{(\text{"jeans"}, D)} = \ln \frac{1000}{101} = 2.293.$$

As indicated in Eq. (3), in practise, the TF-IDF weight of a word in a document is determined by multiplying its TF by its IDF. The TF is the frequency with which the word appears in the document, whereas the IDF is the logarithm of the total documents divided by the frequency with which the term appears in those papers.

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

The document may then be transformed into a word frequency vector using the acquired TF-IDF values of the document's keywords. Document similarity is then determined by calculating the cosine similarity of the vector, as shown in Eq. (4).

$$\mathrm{similarity}(d_i, d_j) = \frac{\sum_{i=1}^{n} (D_i \times D_j)}{\sqrt{\sum_{i=1}^{n} (D_i)^2} \times \sqrt{\sum_{i=1}^{n} (D_j)^2}}$$

For the purposes of this article, we will treat the titles and descriptions of products as documents, with the words inside serving as terms. Then, we can use TF-IDF to determine how significant a given word is within the context of the complete set of product titles and descriptions. The TF-IDF algorithm gives greater weight to words that appear more often in one product's title and description information and less frequently in the information for other items. By comparing it to other goods in the corpus, we can see how it compares to others and learn what makes it special using the similarity measure.

TF-IDF does have certain restrictions, however. The first is that it is important to have a high-quality corpus to ensure successful keyword extraction. Increasing the size and variety of the corpus used to train the keyword extraction algorithm is one approach to fixing this problem. Doing so may assist guarantee that the model picks up on significant terms that are reflective of the subject of interest. The quality of the training corpus may also be increased and keyword selection simplified by making use of domain-specific dictionaries and ontologies. The second restriction concerns the potential influence of feature-word placement on the text's degree of distinctiveness. N-grams, which are sequences of n consecutive words, are one such method. By taking into account the context of the feature words, N-grams enhance the precision of text discrimination. Word combinations of two or three words may better express the meaning of the text than single words alone.

1.2 Graph TF-IDF design

In this study, we argue that Eq. (4)-determined commonalities in product descriptions are indicative of importance. We offer the graph TF-IDF network, which illustrates the link between goods and separate SME sites, to give a holistic perspective of the product landscape across these sites.

Where V is the collection of products that serve as nodes and E is the set of

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 01, 2023

edges, we say that the product network is a weighted undirected graph G = (V,E). If the similarity score (specified in Eq. (4)) between two nodes (i.e., two independent products) is greater than zero, then an edge exists between them. Using the similarity (n,, n,) found in Eq. (4), we may determine the edge weight between nodes i and j, denoted by A,.

$$A_{ij} = \begin{cases} \text{similarity}(n_i, n_j), & i \neq j; \\ 1, & i = j \end{cases}$$

Our method uses a hierarchical network topology with two tiers. To begin, using the above technique, we may build a graph network on the product layer. Second, because the goods belong to their individual SME sites, we can regard each site as a huge collection of documents and construct a site graph network using the same strategy. The product layer product map is the subject of analysis in this work.

1.3 Dataset

The data collection is comprised of information from 52 online clothes boutiques. More than 90,000 goods are covered throughout these sites, complete with names, body_html (description data), production kinds, tags, and more.

Reviewing the data set, we find that there are a whopping 83 698 distinct tags, suggesting that almost every product has its own label. In addition, the fact that most items have many tags indicates the tag is somewhat nuanced. Although there are 116 distinct categories of output, the texts within them are brief, and distribution is highly skewed. Table S1 in the ESM of this article's online edition displays the data for each product category, with three significant numbers reserved and in decreasing order. The "Dresses" category has almost twice as many items as the "Knit tops" category, while many other categories have such a low quantity that their share is practically zero. Product data, on the other hand, includes descriptive information in places like the title and body_html. In this work, we primarily employ the product's title and body_html text content to describe its qualities, in light of the aforementioned three considerations.

The product's description, which may be used for profiling data characteristics, is included in body_html. The first step is to convert the content found in body_html into body_text. Second, as discussed in the first section, the data included several meaningless and irrelevant stop words. As a result, the data has to be cleaned up by excluding stop words. We also ran procedures like lowercase conversation and punctuation regularisation. Word length distribution for the product title and body_text after these preparation steps is given in Table 1. Keep in mind that although 0 is the very minimum for body_text, it doesn't always indicate that body_html was initially null. As a further precaution against an empty text feature, we joined the product's title and body_text together.

Two, the Outcome and the Talk

In this part of the installation guide, we cover the first two steps. The first is to use

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 01, 2023

the TF-IDF method to produce a quantifiable measure of product similarity. Second, we build a product network visualisation to show the interconnections between the goods graphically.

1.4 Product relationship quantification

Each product is treated as a separate document for the purposes of the TF-IDF algorithm, which was introduced in Section 1. Each document's terms may be assigned TF-IDF values, and those values can be used to generate a word frequency vector. In this way, we may use the cosine similarity measure to evaluate the correspondence between texts. Because it is based on a corpus, our technique can be used even when fresh papers are added to the mix.

In this work, we illustrate how to quantitatively assess associations between items with the help of an example. The "classic high waist skinny jeans dark denim" product was chosen at random as the test data. The second part of the product is an explanation of the jeans' features, such as the number of pockets, the fabric composition, and so on. Then, we determine the n records that are most like the sample under evaluation.

The eight most comparable samples to "classic high waist skinny jeans dark denim" are included in Table 2, along with their similarity scores and the site indexes from which they were taken. Only the title information is shown here because of the lengthy nature of the content. On the one hand, it's clear that these items are all variations on the same theme—"classic high waist skinny jeans"—with colour serving as the primary differentiator. The fact that all eight samples came from sites 1 and 52 suggests a fierce rivalry between the two. The similarity ratings between the items serve as the basis for establishing such a connection.

2.2 Product map revealed by communities

In Section 1.2, we introduced a method for building the product network that relies heavily on similarity ratings. Our continued focus on locating and characterising communities, as well as the product map in the network, will allow us to undertake a more thorough examination of the product map throughout the whole product pool.

Table 1 Word length statistics of body_text and title_text.

Text	Minimum length	Maximum length	Average length
Body_text	0	571	24.33
Title_text	2	14	5.98

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 01, 2023

Table 2 Eight samples that are the most similar to the test data.

Series No.	Shop No.	Title	Similarity score
1	1	Classic high waist skinny jeans light blue wash	0.989 217
2	1	Classic high waist skinny jeans white	0.966 868
3	52	Classic high waist skinny jeans light blue	0.942 952
4	52	Classic high waist skinny jeans wine	0.924 600
5	52	Classic high waist skinny jeans khaki	0.923 623
6	52	Classic high waist skinny jeans rust	0.921 935
7	52	Classic high waist skinny jeans light chocolate	0.920 836
8	52	Classic high waist skinny jeans olive	0.915 092

Without any kind of communalization, the original graph network looks like Fig. 1; we picked 18 sites at random from among all the SME sites, sampled 10 samples from each for a grand total of 180 nodes (the graph network with more nodes or sites has similar patterns). Products are represented by the nodes, while SME locations are indicated by colour. Weighted degree, the sum of the weights of the edges incident with the node, is used to quantify the size of a node. In this case, the importance and number of connections of a node are both increased by its size. In addition, the edge's thickness is directly related to its mass. Finding the relationship between the nodes in Fig. 1 is challenging for two basic reasons. To begin, the original graph had a "fully connected" form with an excessively high number of edges, making its structure difficult to decipher. Second, the geographical distribution of nodes is not reflected in the original network, just the connection links between them.

By filtering out edges with weights below 0.14, we are able to create a dense graph network. Here, the weaker the weights, the less connected the nodes are to one another. The Louvain approach for extracting communities from big networks was also used to improve community detection in our product network. Blondel et al. developed this technique on the premise that it is more common for people to know one another within their own communities than to know people from other communities. The primary premise is that individual network nodes would explore their neighbours' community labels before settling on a single community label with the highest modularity increment. Once the modularity has been maximised, each community is treated as a separate node, and the process is repeated until the modularity can grow no further.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 01, 2023

The modularity of our product map network is described by the following equation:

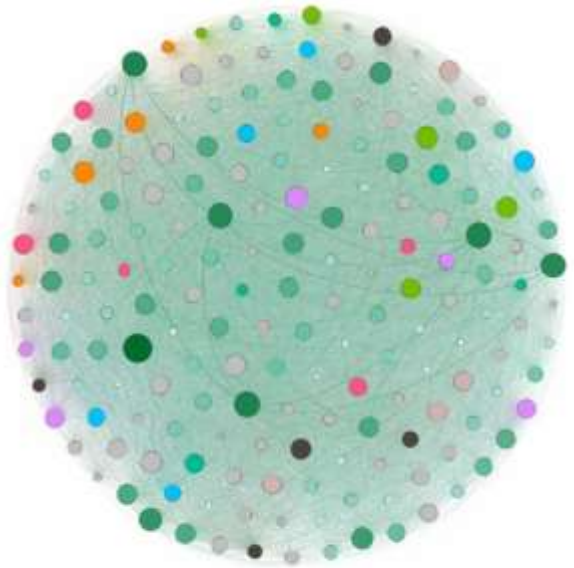


Fig. 1 Original data structure diagram. Each node represents a product, and the color represents the SME site.

$$M = \frac{1}{\sum_{ij} A_{ij}} \sum_{i,j} \left[A_{ij} - \frac{\sum_i A_{ij} \sum_j A_{ij}}{\sum_{ij} A_{ij}} \right] \delta(c_i, c_j) \qquad (6)$$

where c is the group to which product i belongs. To check whether nodes i and j belong to the same group, we utilise the 8 function. The value is 1 if the answer is yes, and 0 otherwise.

Fig. 2 is a visual depiction of the product map of SME locations. Specifically, Fig. 2a displays the geographical properties of the product network as well as the interproduct linkages. There is a SME sites index and the product type making up the node label for each product. The community to which a node belongs may be inferred from the fact that its nodes and edges are of the same colour. The amount of connections connected to a node is represented by its degree, the community to which a node belongs is shown by its modularity class, and the weighted edges record the association between products. The more interconnected section may be modelled as a community where communication is frequent within a small group but infrequent among larger groups. There is a positive correlation between product similarity and edge weight within this group. A degree distribution with a mean of 9.41 relative products and a maximum of 24 is shown in Figs. 2b and 2c, while the distribution of modularity classes has the same shape. When Eq. (6) is used to determine modularity, the resulting value is 0.682. There are 6 major hidden neighbourhoods (with more than 15 nodes) represented by the following colours inside the community network: red for decorations, blue for knits and bodysuits, dark green for trousers, orange for kids' clothes, purple for dresses, and green for sportswear. According to the densities of node labels and edges, products from the same SME locations are often dispersed in the same neighbourhood. In addition, real-world findings corroborate the notion that there is more product similarity inside a group than there is across communities.

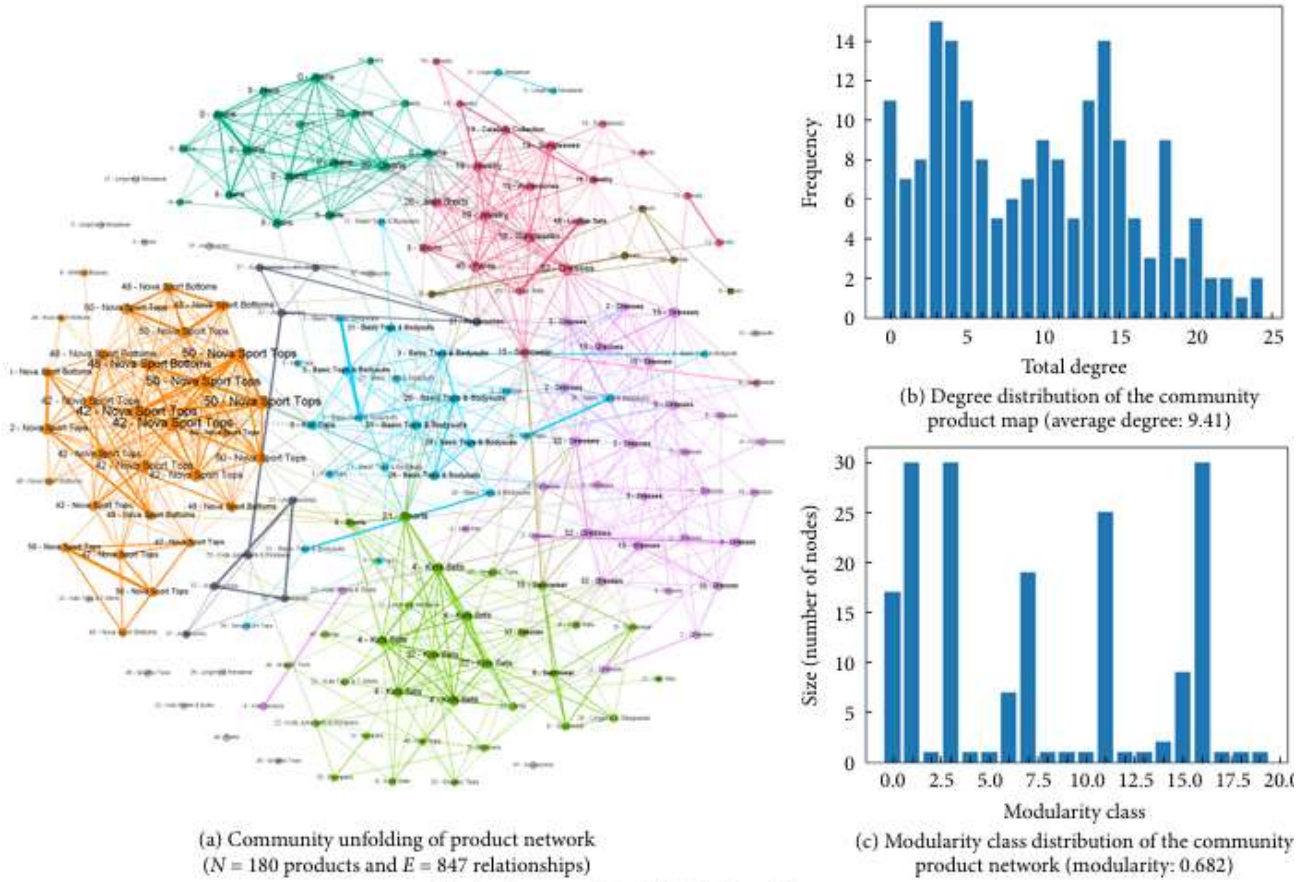
3 Conclusion

Overall, the suggested decentralised method for researching product maps in online commerce may aid small and medium-sized enterprise online stores immensely. This bottom-up, crowd-sourced approach has the potential to deepen our comprehension of the market's competitive dynamics. It may also be of great use in directing e-commerce sites towards optimal choices about price, stock levels, and promotion.

Academics and professionals alike might benefit greatly from more study of this topic. Additional e-commerce sites might be

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 01, 2023

included in a future research to fine-tune the product map analysis and improve the reliability of the findings. The benefits and drawbacks of a widespread implementation of the suggested decentralised strategy may also be investigated in future studies.



(a) Community unfolding of product network (N = 180 products and E = 847 relationships)

(b) Degree distribution of the community product map (average degree: 9.41)

(c) Modularity class distribution of the community product network (modularity: 0.682)

Fig. 2 Representation of SME sites' product map.

Additional data sources, such as customer reviews and ratings, might be used in future works to acquire a more in-depth knowledge of product interactions within a market. E-commerce companies might exploit this information to their advantage by tailoring their product selections and advertising campaigns to their customers' tastes and habits.

Prospective applications of the suggested bottom-up approach to the analysis of product maps in other sectors may also be investigated in further research. Financial services, healthcare, and transportation are just a few examples of sectors where this methodology might be utilised to better comprehend the interdependencies between goods and services.

In addition to these possible lines of inquiry, future study may refine and perfect algorithms and methods for analysing and displaying product map data. Foreseeing future market circumstances may need the use of ML and other cutting-edge data analysis approaches, such as the identification of patterns and trends within the data and the creation of predictive models.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 01, 2023

Acknowledgment

This work would not have been possible without the guidance of the many excellent instructors who assisted with the research, development, and writing.

Electronic Supplementary Material

Supplementary materials including

Table S1: Proportion of production type.

All the supplementary materials are available are available in the online version of this artide at https://doi.org/10.26599/ IJCS.2023.9100006.

Dates

Received: 16 February 2023; Revised: 28 March 2023; Accepted: 3 April 2023

References

[1] T. Elrod, G. J. Russell, A. D. Shocker, R. L. Andrews, L. Bacon, B. L. Bayus, J. D. Carroll, R. M. Johnson, W. A. Kamakura, P. Lenk, et al, Inferring market structure from customer response to competing and complementary products, Mark Lett, val 13, no. 3, pp. 221-232, 2002.

[2] R. Venkatesan and P W. Farris, Measuring and managing returns from retailer-customized coupon campaigns, J. Mark, vol. 76, no. 1, pp. 76-94, 2012.

[3] LK Mensah and D. S. Mwakapesa, Cross-border e-commerce diffusion and usage during the period of the COVID-19 pandemic. A literature review, in Proc. 3 Africa-Asia Dialogue Network (ADN) Int. Conf. Advances in Business Management and Electronic Commerce Research, Ganzhou, China, 2021, pp. 59-65.

[4] D. Tolstoy, E R Nordman, S. M. Hänell, and N. Ozbek, The development of international e-commerce in retail SMEs An effectuatial perspective, J World Bus, vol. 56, no. 3, p. 101165, 2021

[S] S. Gabel, D Guhl, and D. Klapper, P2V-MAP Mapping market structures for large retail assortments, Journal of Marketing Research, vol. 56, no. 4, pp. 557-580, 2019.

[6] J. B. Kim, P. Albuquerque, and B. J. Bronnenberg, Mapping online consumer search, J. Mark Res., vol. 48, no. 1, pp. 13-27, 2011,

7] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko, Mine your own business: Market-structure surveillance through text mining, Mark Sci., vol. 31, no. 3, pp. 521-543, 2012

[8] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, A survey of user profiling State-of-the-art, challenges, and solutions, IEEE Access, vol. 7, pp. 144907-144924, 2019.

[9] S. Qaiser and R. Ali, Text mining Use of TF-IDF to examine the relevance of words to documents, Int. J. Comput. Appl., vol. 181, no. 1, pp. 25-29, 2018.

[10] L. Yao, C. Mao, and Y. Luo, Graph convolutional networks for text classification Proc. AAAI Conf. Artif Intell., vol. 33, no. 1, pp. 7370-7377, 2019.