Innovations in Intelligent Systems and Advanced Engineering
Vol. 1, No. 2, 2025 (August)

Mining Social Network Data for Enhanced Consumer Perception Measurement

D S Deepika
Department of Information Technology, R.M.D. Engineering College, Kavaraipettai, Chennai, Tamil Nadu, India
deepika.it@rmd.ac.in

Raja Thimmarayan
Department of Mechatronics Engineering, KCG College of Technology, Chennai, Tamil Nadu, India
traja70@gmail.com

S. Durga Devi
Department of Computer Science and Engineering, Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, Tamil Nadu, India
sdurgadevi@gmail.com

V. Vidya Lakshmi
Department of Electronics and Communication Engineering, R.M.K. Engineering College, Chennai, Tamil Nadu, India
vidyalakshmi2024@gmail.com

Mohd Nasair Uddin Khan
Department of Computer Science and Engineering, Sri Indu College of Engineering and Technology, Hyderabad, Telangana, India
nasair10@gmail.com

Monisha R
Department of Management Studies, St. Joseph's College of Engineering, Chennai, Tamil Nadu, India
monisharamaraj89@gmail.com

Sathishkumar V E
School of Computing and Artificial Intelligence, Faculty of Engineering and Technology, Sunway University, Bandar Sunway, Selangor Darul Ehsan, Malaysia
sathishv@sunway.edu.my

Hemavathy P
Department of Artificial Intelligence and Data Science, Saveetha Engineering College, Chennai, Tamil Nadu, India
hemamanigandan@gmail.com

Abstract: The Support Vector Machine (SVM) is a prevalent machine learning method which is employed for classification, regression, and many statistical tasks, adept at producing a hyperplane or several hyperplanes in high-dimensional or infinite-dimensional environments. An SVM analyzes datasets and can be trained iteratively to identify patterns of similar behavior or shared features among events within the database. In this study, a new data mining tool was developed to evaluate patients with type 2 diabetes mellitus and their familiarity with Sitagliptin. To achieve this, a two-stage research framework was designed. The first stage involved self-organizing mapping (SOM) for exploratory research, which identified mechanisms based on user preferences expressed in blog entries. Consumer clusters have positive or negative drug views. Network analysis revealed notable forum users in the second round. This modeling provided insights into influential users and their role in shaping drug-related discussions. The results of this study open new avenues for rapid data acquisition, user feedback, and analysis to improve public health outcomes. Furthermore, the insights offer valuable input for healthcare providers and pharmaceutical suppliers.

Keywords: SVM, data acquisition, data mining, feature analysis, predictions

I. INTRODUCTION

From personal interactions to live forums, social media gives patients infinite chances to evaluate prescriptions and discuss medical system experiences. At the same time, it offers businesses unrestricted opportunities to collect feedback on their products and services [1]. For this reason, the social surveillance of online networks has become a top priority for pharmaceutical companies within their IT divisions. Such monitoring enables timely dissemination of products and services, enhances delivery efficiency, increases revenue and profit, and helps minimize operational costs. In recent years, methods for social media aggregation have been explored for bio-monitoring purposes. Social networking fosters collaboration, knowledge sharing, and information exchange in healthcare, effectively creating a virtual environment for collective learning. By applying network modeling and statistical techniques such as network analysis, patterns and insights can be derived from this "information cloud." In such a model, platforms like Facebook, Twitter, and WebMD represent networks composed of nodes (users or alliances) and edges (relationships such as friendships, partnerships, or shared interests) [2].

The most common way to visualize these networks is through graphical representations, which are particularly valuable for simulation. Network modeling can provide deep insights into the dynamics of social networks and their influence on healthcare communication. For example, system models can simulate how information spreads among people (e.g., pandemic news or adverse drug reactions). Similarly, they can illustrate the formation of new connections and demonstrate how specific information shapes consumer groups based on shared concerns about diseases [3]. Input from social platforms, known as sociometric data or adjacency matrices, can be used to construct system representations. Although social systems are often sparse, they remain highly effective for studying network structures. Node degree, network density, and centrality may show how firms, pharmaceutical suppliers, and medication brands rank in these networks. Identifying system communities or clusters is possible. Specialized algorithms help network analysts find communities, which are group nodes that communicate more with one other than with other nodes [4].

Identifying such communities enables the collection of actionable knowledge for the entire healthcare ecosystem. Pharmaceutical firms, for instance, can use this information to better target marketing investments. Healthcare providers can use these insights to enhance patient satisfaction and reduce adverse treatment outcomes. Medical professionals can also gain valuable feedback from colleagues and patients to support clinical decision-making and improve treatment outcomes. Finally, patients themselves can make more informed healthcare decisions by leveraging the shared expertise of others within these networks [5]. Figure 1 shows the data mining approach overview.

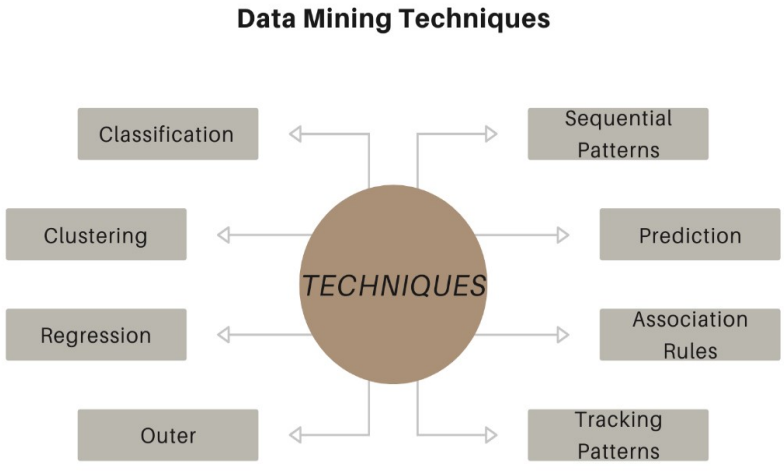

Figure 1: Data Mining Approach Overview

Social networks are non-structured, semi-structured, and heterogeneous, which makes capturing and analyzing their data more complex. The link mining system uses social networks, link analysis, hyperlink and site mining, graph mining methods, connection learning, and inductive logic programming to mine social media. Link mining includes link-based object categorisation, object type prediction, link type prediction, link life prediction, cardinal link prediction, and relation prediction [6]. Relation prediction leverages social network models' intrinsic properties to detect probable network relationships. Viral marketing analyses customer experiences and targets the most socially connected to boost word-of-mouth. Similarly, discussions in newsgroups benefit from "response" relationships, where graph partition algorithms classify responses based on the extent of agreement or disagreement among users. In addition, multi-link system analysis extracts relationships, calculates connections, and classifies them according to user-acquired knowledge [7].

Traditional social science studies often rely on surveys that focus on specific information-gathering topics. However, such surveys typically collect limited datasets, often no more than a few hundred responses, which restricts the depth of analysis. By contrast, social media generates disproportionately large datasets, enriched with user experiences contributed by thousands of individuals. Data collection from social networks can be achieved in two main ways:

1. Crawling APIs offered by websites such as Twitter, YouTube, Flickr, and Email services.

2. Scraping content from rendered HTML pages.

These approaches allow researchers to track evolving conversations and conduct important public health studies [8]. In previous experiments, algorithmic approaches have been widely used to extract consumer sentiment and viewpoints. For example, how social media users associated influenza with patient-reported information are studied. A sentiment analysis of message boards related to technology stock prices is considered, while concern words are explored, including noun forms, verb types, negations, and prepositions. A method for intelligently scanning online marketplaces to detect customer discounts is discussed. An emotion classification system is discussed which uses three classifiers: naïve Bayes, maximum entropy, and support vector machines [9].

Other studies also leveraged specialized data sources: FDA's Adverse Event Monitoring System and genetic programming to map customer satisfaction, post polarity examination in news sources using both word content and context [14]. Overall, literature surveys indicate that businesses have widely used social media to analyze consumer emotions. Few studies have identified famous users or explored how forums influence others' beliefs and activities like the data mining literature [12]. The novelty of the proposed methodology is threefold. First, it identifies prominent users, unlike most previous research. To achieve this, the approach examines how platform relationships influence user opinions and behaviors. Second, it automatically tags messages with positive or negative sentiment terms [10]. Finally, it applies word frequency statistics in combination with self-organizing map analysis, followed by network analysis using modified graph theory techniques. This allows for the classification of user groups and the identification of potential future users.

From the results, each communication board produced a list of medicines and treatment strategies discussed by patients. The goal was to determine which medications or devices patients most frequently mentioned [11]. Across multiple pharmacology forums, Sitagliptin emerged as the most widely discussed medication, especially on the DiabetesDaily.com message board, compared with four other forums [15]. Using the above methods, the wordlist was refined, and the corresponding graph was generated. The diagram classifies words into positive and negative categories, where terms contributing to either sentiment were identified. This classification was based not only on word frequency but also on contextual meaning within forum posts and discussions [13].

II. PROPOSED SYSTEM

Self-organizing maps (SOMs) are a machine learning approach often used for grouping extensive datasets. A SOM comprises a neural network layer where input data are mapped onto a set of nodes, each linked to a certain weight vector that denotes the output space. The values of these weight vectors delineate the attributes of the respective clusters. SOMs systematically arrange data into a structured map, whereby neurones linked by

like data vectors cluster together. Due to its visualization skills and proficiency in managing high-dimensional data, SOMs are especially effective for streamlining intricate investigations. Bonato et al. revealed that wavelet packet techniques may augment SOM performance by decreasing the dimensionality of the data space while preserving essential information, thereby enhancing clustering. Figure 2 depicts the comprehensive system architecture.

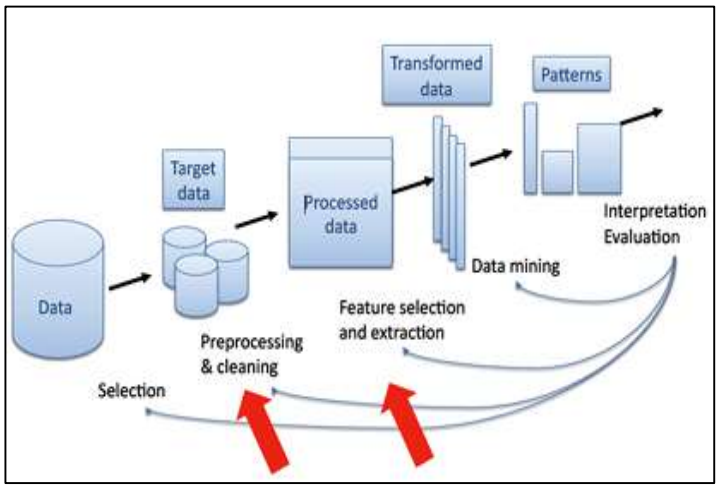

Figure 2: SOM Architectural Proposed System

The training phase introduces new input information to the weight-vector system, which assigns each data vector to its corresponding best-matching neuron (BMN). These neurons are adjusted according to the new input, and their neighboring neurons are also updated, albeit to a lesser extent. Neurons farther from the BMN experience smaller adjustments to their weight vectors. This iterative process continues until convergence conditions are satisfied for all input vectors, resulting in a finalized nine-by-nine SOM. To examine how the vectors clustered according to specified words, the updated wordlist was integrated into the MATLAB SOM Toolbox. The Self-organising Map was trained with several map dimensions, and internal validation was conducted using quantisation and topographic errors. The quantisation error quantifies the mean distance between each input vector and its corresponding neurone, while the topographic error evaluates the preservation of the map's topology by determining the fraction of input vectors whose Best Matching Neurones (BMNs) are not contiguous. The ideal map size was established based on minimum quantisation values and a topographic error of 0.1257×10^{-7}.

After training, word vectors were mapped into the SOM and emerging clusters were analysed for negative vector variable connections. Clusters containing three or fewer posts were discarded to reduce noise. Within the remaining clusters, occurrences of terms were counted. The SOM map visually distinguished between "positive" and "negative" words, allowing identification of subgroups and the posts associated with them. This enabled the evaluation of customer sentiment whether positive or negative. Subsequently, forum posts were analyzed to identify prominent participants and their interactions. Networks were constructed where nodes represent participants and links represent interactions. Networks were analyzed both in non-directional and directional forms: non-directional nodes reflect the number of connections, while directional nodes capture the flow of information between participants. Four distinct types of nodes were identified following the Wasserman framework: isolated nodes, senders, receivers, and brokers. Network density was calculated as the number of links present at a given time. A provider-based theoretical methodology was employed for its broad applicability in social network research and its ability to predict user behaviour and relationships. This approach is particularly suitable for network-based analysis of forums, capturing the internal dynamics among participants. Bidirectional edges were established between original posts and replies to represent the flow of knowledge from the poster to the commentator. Additional annotations were then linked to the posts for further analysis.

III. RESULTS AND DISCUSSION

The SOM-based graphic representation of the Diabetes Daily forum community, highlighting positive and negative terms, is presented. The SOM was trained on a subset of the data before being applied to the whole dataset. This preliminary training was necessary to ensure that the SOM could effectively model the patterns in the information. For training purposes, 30% of the data was selected for SOM preparation. To evaluate the weight of terms associated with sentiments toward Sitagliptin, a 13 × 13 map size was employed, and 28 terms from the updated word list were incorporated during the training process. One criterion for variable selection was that each term had to occur at least ten times in the dataset. This ensured a reliable calculation while excluding statistically negligible outliers. Most posts aligned with either positive or negative terms, represented at four points with corresponding weight vector values, which defined the positive and negative regions of the SOM plot. This visualization revealed an emerging picture of consumer sentiment, approximately divided between satisfaction and dissatisfaction with Sitagliptin. Negative perceptions primarily arose from treatment side effects, a finding corroborated by medical literature reporting similar symptoms. Other sources of negative sentiment included frustrations related to drug costs and dissatisfaction with aspects of the medical community. Conversely, positive sentiment was largely linked to customer satisfaction and trust in physician recommendations. The SOM values corresponding to positive terms from the expression list are represented by *P*, a sequence indicating the magnitude of optimistic sentiment for each term. Figure 3 illustrates the various parameters used for performance analysis, providing a detailed overview of the mapping and clustering results.

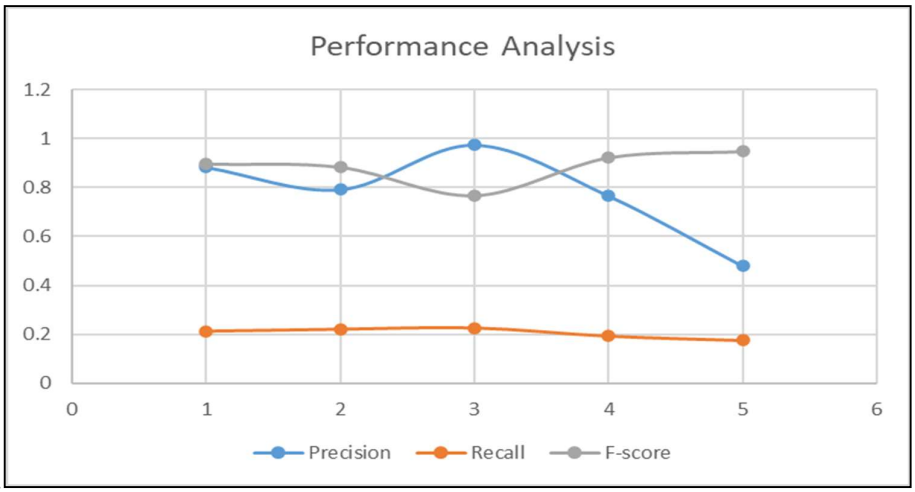

Figure 3: Performance Analysis

The analysis focused on the factors described above to achieve a preliminary separation of consumer sentiment regarding Sitagliptin on the forum. This study aimed to transform forum posts on type 2 diabetes mellitus into vector representations to facilitate an intelligent understanding of Sitagliptin usage. The findings highlight new challenges and opportunities for developing more systematic strategies in this area. Cluster efficiency analysis results are presented in Table 1. A fragmented consensus on Sitagliptin depends on individual patient circumstances and specific variable factors. Given the diversity of user experiences, the presence of a social media forum provides insights from individuals with varying outcomes. Despite these variations, the analysis was able to capture both positive and negative sentiments, which were subsequently corroborated by studies on the efficacy and side effects of Sitagliptin. A more comprehensive understanding of patient feedback on medications and healthcare services could be achieved by incorporating up-to-date information in future analyses. Figure 4 presents the results of the efficiency analysis.

Table 1: Cluster Efficiency Analysis

#Cluster	1	2	4	8
Efficiency (n=0.5)	0.768	0.656	0.533	0.492
Efficiency (n=0.75)	0.796	0.679	0.552	0.509
Efficiency (n=1)	0.825	0.704	0.572	0.528

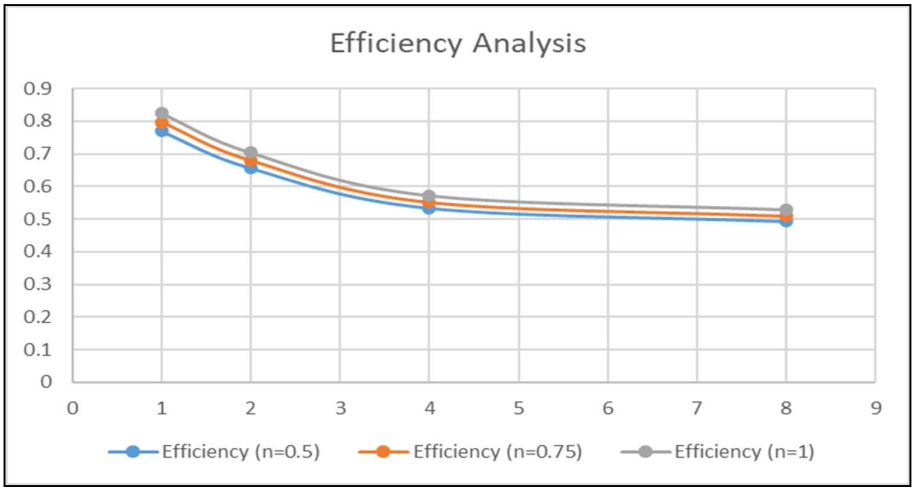

Figure 4: Efficiency Analysis

An integrated approach would need further growth through user impact websites and user engagement. It will take other participants' user experiences research, friendships, and rankings on social media sites. The similarity measure analysis values are Shown in Figure 5.

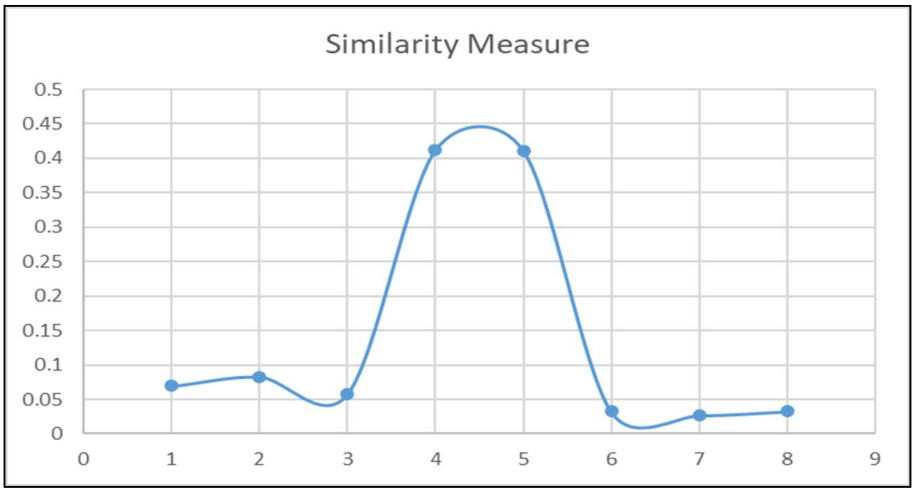

Figure 5: Similarity Measure Analysis

More specialized text mining methods incorporate advanced lexical dictionaries and processing techniques that analyze both formal terminology—such as disease names, treatments, side effects and informal language, including colloquial expressions for diseases, medications, and withdrawal symptoms commonly used

on social media platforms. Future investigations of the drug across multiple social media channels could provide updated insights, particularly if certain networks avoid discussions about the medication.

IV. CONCLUSION

Social media is increasingly used as a platform for sharing personal experiences and opinions. This presents an opportunity for businesses to optimize healthcare delivery, reduce costs, and enhance product development. A detailed analysis of posts from six key users revealed that their knowledge of Vildagliptin and Sitagliptin, derived from both online sources and personal experiences, was largely informative and diverse. These users were frequently sought out by other participants for guidance and insights regarding Sitagliptin, demonstrating engagement with both novice and long-term members of the forum. Behavioral analysis indicated that these individuals acted as primary information brokers for Sitagliptin on the Diabetes Daily forum. This approach could potentially be integrated into smartphone applications or physiological tracking systems as part of a software toolkit. Both patients and organizations would benefit patients could provide direct feedback on products and services, while organizations could use timely, structured reviews to generate comprehensive evaluations of product performance, identify areas needing improvement, and guide future development strategies.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1]. D. Liben-Nowell, and J.M. Kleinberg, “The link prediction problem for social networks,” Journal of the American Society for Information Science and Technology, vol. 57, pp. 556-559, 2007.
[2]. Z. Lacroix, H. Murthy, F. Naumann, and L. Raschid, “Links and Paths through Life Sciences data sources,” in Proceedings of the 1st International Workshop on Data Integration in the Life Sciences, pp. 203-211, 2004.
[3]. J. Noessner, M. Niepert, C. Meilicke, and H. Stuckenschmidt, “Leveraging Terminological Structure for Object Reconciliation” in The Semantic Web: Research and Applications, Heidelberg, Berlin: Springer, pp.334-348, 2010.
[4]. M.E.J. Newman, “Detecting community structure in networks,” European Physical Journal, vol. 38, pp. 321-330, 2004.
[5]. J. Huan and J. Prins, “Efficient Mining of Frequent Subgraphs in the Presence of Isomorphism,” in Proceedings Of the 3rd IEEE International Conference on Data Mining, pp. 549-552, 2003.
[6]. D. Hand, “Principles of Data Mining,” Drug Safety, vol. 30, pp. 621-622, 2007
[7]. J. Hans, and M. Kamber. Data Mining: Concepts and Techniques ed. Burlington, Mass: Morgan Kaufmann, 2006
[8]. P. Soucy and G.W. Mineau, “Beyond TFIDF Weighting for Text Categorization in the Vector Space Model,” IJCAI'05 Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp. 1130-1135, 2005.
[9]. S. Bird, E. Klein, and E. Loper, “Natural Language Processing with Python,” Sebastopol, CA: O’Reilly Media, 2009.
[10]. S.R. Das, and M.Y. Chen, “Yahoo! for Amazon: Sentiment extraction from small talk on the Web,” Management Science, vol. 53, pp.1375-1388, 2007.
[11]. T. Kohonen. Self-Organizing Maps, 3rd ed. Heidelberg-Berlin: Springer, Dec. 2000
[12]. P. Bonato, P.J. Mork, D.M. Sherill, and R.H. Westgaard, “Data mining of motor patterns recorded with wearable technology,” IEEE Engineering in Medicine and Biology Magazine, vol. 22, pp. 110-119, 2003.
[13]. J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, “Self- Organizing Map in MATLAB: The SOM Toolbox,” In Proceedings of the Matlab DSP Conference, vol. 99, pp. 35-40, 1999.
[14]. W. Liu, J. Zhao, and D. Wang, “Data mining for energy systems: Review and prospect,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 11, no. 4, pp. 1-9, 2021.

[15]. V. Rajkumar, “A methodology for direct and indirect discrimination prevention in data mining,” International Journal of MC Square Scientific Research, vol. 5, no. 1, pp. 28-36, 2013.