Data mining in social networking sites: A social media mining approach to generate effective business strategies

Sanjeev Pippal¹, Lakshay Batra², Akhila Krishna³, Hina Gupta⁴, Kunal Arora⁵

Computer Science and Engineering Department, JRE Group of Institutions
Greater Noida, India
¹sanjeev.pippal@jre.edu.in
{²lakshay.batra.29, ³akhilakrishna, ⁴guptahina189}@gmail.com
⁵kunal_arora93@yahoo.in

Abstract—Mining social media is a new plan to boom business. The Social media houses vast amount of user-generated data which can be used for data mining. Marketing enthusiasts are searching for means to utilize these mined business information for the intake of their sales/marketing and advertising teams. The mined information from social platforms can considerably impact business strategy of any business enterprise. In this paper we have attempted to quantify and classify such approaches by different authors. Further we suggest how this survey and study of the data mining approaches can benefit the understanding the importance of social media mining.

Keywords - Data Mining; Social Media; Business Intelligence; PCA; EFA

I. INTRODUCTION, MOTIVATION & GOAL

The process of collecting, searching through and analyzing a large amount of data in a database, so as to discover patterns or relationships is the use of data mining to detect fraud. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information, information that can be used to increase revenue, cuts costs or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining also known as data or knowledge discovery is the course of action of analyzing data from different viewpoints and shortening it into information that can be used to determine patterns and relationships in data in order to help make better business decisions and thus increase revenue, project trend, interactive marketing, cuts costs. Its application domains include bioinformatics, business intelligence, retail industry, telecommunications, predictive analytics, decision support systems and social media.

Social media has vast amount of user-generated data which can be utilized for data mining. Data mining of social media can amplify use of social media and perk up commercial intelligence to transport enhanced services. For example, data mining techniques can identify user sentiments for anticipatory preparation to develop suggestion systems for business of specific products and even to build new friendships or connect certain interest groups. (Facebook uses likes, groups as well as posts of users to recommend users specific ads as well as new pages and groups.)

Marketing experts are searching for means to utilize them for their sales and advertising teams. Quite a lot of type of pattern detection that are normally used in social media data mining include: Sequential Patterns, Association learning, Clustering, Classification, Regression, Deviation detection, and Summarization.
A survey was done among a group of people to predict the working of these patterns and is included in the paper. Challenges comprise of judgment of the right stability between customer and network information. An optimistic message that goes viral can open doors to enormous sales explosion whereas it’s contradictory will wreck that merchandise or even a company instantly.

II. RELATED WORK
In [1] the author has studied the issues related to analysis of social networks using web mining techniques. Apart from this the application of web mining techniques and a general process for social networks analysis has been discussed. The paper also highlights the difficulties in selecting data samples, finding communities and patterns in social networks and analyzing overlapping communities.

In [2] the author showcases a systematical data mining approach to mine intellectual knowledge from social data. The author took Facebook as a primary data source and proposes to use different data mining techniques to analyze this social networking site and other sites too. One algorithm that the author had discussed was K-nearest neighbor (K-NN). This algorithm classifies objects based on samples.

In [3], the deliberative aspect of social media governance was studied by the authors. Main focus was on a setting which could trace the decision-making by a site’s elite. The primary setting was the Wikipedia promotion process in which users of Wikipedia could be nominated to become admins. The authors investigated on the temporal dynamics of the elections, identified ordering effects that contrasted with standard theories of herding and information cascades. They also found that several forms of assessment played a crucial role in how voters made decisions. However the author did not consider recursive voting i.e. for a given voter, the timing of one’s arrival affects how much information one has about earlier votes, which may in turn affect one’s own vote.

In [4] the author discusses the application of correlation, clustering, and association analyses to social media. The main purpose of this paper was to describe how data mining and text analytics can be applied to social media in order to identify key themes in the data. To be more specific the author described the analysis of Twitter posts. Certain issues in terms of accuracy while collecting the data from social media were also highlighted.

In [5] the authors have studied on the techniques that are currently used to analyse SM. Also, they have accounted those techniques that can be considered further in this field. In this paper the analysis of SM data has proved to be effective, this is so because of the capacity data mining possess in handling noisy, large and dynamic data. According to the authors, in future to mine the data generated on SM currently used and yet-to-be-explored data mining techniques will be used.

In [6] the authors explain how social networking site helps a company and its brand to convey messages to massive amount of people very easily. The aim of the authors was to cover up protection strategies for business infrastructure. The authors also discussed about few proactive defensive strategies against the malicious users who could leak the information of a company and get access to the messages that were to be conveyed. The strategies that were mainly discussed included awareness, monitoring and logging, acceptable use policy (AUP) and enforcement.

III. EMPIRICAL SEARCH

Fig. 2 Datamining tasks
Data mining tasks include predictive tasks and descriptive tasks. Predictive tasks make predictions based on user data collected whereas the descriptive tasks determine relationships among data, formulate patterns to depict relationship. Under predictive tasks we have classification, regression and deviation detection. Classification function is created by analyzing the relationship between attributes and data objects whose class labels are well known. Eg: Collect various demographic, lifestyle, education related information about customers to target specific group. Regression is a function that maps a data item to a real valued prediction variable. Deviation detection detects deviation from normal behavior. Eg: Change of search history of user.

Under descriptive tasks we have clustering, summarization, association learning and sequential patterns. Clustering is the process of making group of abstract objects into similar subclasses. Eg: find groups of posts that are analogous to each other based on the significant terms in them. Summarization is simple descriptions about subset of data. Association learning is if/then statements that identify associations/relationships between unrelated data. Sequential Patterns finds the complete set of frequent subsequences from a given set of sequences.

Data Mining in business strategy include Cart Analysis: It is based on future prediction of customer behavior by past performance, including purchases and preferences. Future sales: Predicting future sales plays vital role for the retail company. Database Marketing: Database marketing is one of the most successful business application of data mining. Mining of historical customer data helps determine patterns and trends to build customer profile for effective marketing. Merchandise Planning: This is helpful for offline or online companies. For the offline, a company looking to grow by adding stores can evaluate the amount of merchandise they will need by looking at the exact layout of a current store. For an online business, merchandise planning can help you determine stocking options and inventory warehousing. Card Marketing: If your business involves issuing credit cards, you can collect the information from usage, identify customer segments and then based on information on these segments boost acquisition, target products to develop and design prices. Customer Loyalty: Retailers maintain loyalty management system to provide data for customer driven marketing. Most retailers today have adopted such a system, which is extensively used for reward point accumulation & redemption. Eg: Stanmax, Big bazaar. Market Segmentation: One of the best uses of data mining is to segment your customers. Identify the common characteristics of customers who buy the same products from your company.

As an insight to social media we have briefly described few types of social media. We start with social networking sites. These sites connect people with similar interests and background. Eg: Facebook and LinkedIn. Next we move on to bookmarking sites. These sites act as a normal bookmark. It allows saving, organizing and managing links to various websites and resources around the internet. Eg: Stumble Upon. Next we have social news. These are services that allow people to post and share various news or links to outside articles. It also enables people to cast their vote against the post. Eg: Digg and Reddit. Media sharing is another type of social media which helps in uploading and sharing various media such as pictures and video. Eg: YouTube, Instagram, Flicker. Yet another type of social media is micro blogging. It focuses on short updates that are circulated to the ones who subscribe to receive the updates. Eg: Twitter. Finally we have blog comments and forums. - Blog comments are attached to blogs. The discussion takes place with reference to the blog post. Online forums allow members to hold conversations by posting messages.
Coming on to the parameters of social media marketing we have ten of them. Parameters are as follows goal, specific, what, how, where, timed, consumer mindset, assessment, observe your competitor and customer characteristics. Description for each parameter is as follows: Goal describes what are the goals for social media campaign? Specific explains whom do we want to engage in? What states the needs of the audience? How defines how do you wish to engage the audience? Where suggests what type or platform of social media are we going to use? Timed tells what is the best time to post according to audience usage? Consumer-mindset suggests what kind of content does the consumer relates to? What are the views of consumer about the branded usage of platform. Where the campaign is carried out? Assessment shows how are we going to measure results of the campaign? Eg: use of Google analytics to observe the traffic on your website. Observe your competitor states what are the strategy of your competitor both in online as well as offline domain. Customer characteristics include location, gender, age, likes and interests, relationship status, workplace and education.

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored user data based on open-ended user queries.

We start with LinkedIn’s recruiter program which analyses member’s profile and site behavior to steer them to job opportunities that might be interesting. Next is Twitter. The site is an endless stream of experiences, opinions and sentiments of consumer about everything from computers to movies is used by companies. After Twitter we have Pinterest that captures and examines data related to the network of “pins” and “re-pins”. Aim at retailers seeking to boost brand enhancement and customer engagement. Then we have Facebook. While putting an ad on facebook you can consider the following parameters: location, gender, age, likes and interests, relationship status, workplace and education of your target audience.

And finally we take Google. Google search (web, images, news, blogs, etc.) keeps a track of whatever we search. Web crawling – Googlebot performs the task of collecting document from the web thereafter creating a searchable index for the Google search engine. Website analytics – Google Analytics tracks the traffic of a website. Google Finance – finance data. YouTube – viewing behaviors of users. Google Books – With this feature Google can figure out what people are reading and what they want to read. Google Maps and Google Earth – Google also provide a facility to track parts of the world which you interested in. Ad serving – ads that get maximum click. Email – Gmail content, both sent and received emails are parsed and analyzed. Your contact network – Your contacts in Google Talk, Gmail, Android are tracked by Google. Twitter – Google has direct access to all tweets that pass through Twitter. Google Apps (Docs, Spreadsheets, Calendar, etc.) – valuable data/documents of users. Google Plus – your public profile info.

Google Public DNS – what websites people access most frequently? The Google Chrome browser – What is your web browsing behavior? What sites do you visit? What are your bookmarks? Feedburner – With the help of Feedburner a blogger can analyze who has subscribed to his/her blog. Every Feedburner link is tracked by Google.

IV. PROPOSED APPROACH

Methods: We propose to collect the data available in social media profiles. As the data is available across a number of dimensions or characteristics, to find any meaningful applications of these data we must reduce them to a manageable numbers. We propose to do an EFA (exploratory factor analysis) using Principal Component Extraction method in SPSS 16.0 to have an initial idea of probable factors to reduce the data. After that we club together the various parameters into a few numbers of manageable factors with the help of similar conceptual bases. Now the stratified or classified data can be used for business analysis. Based on existing literature correlations between various factors can be predicted or inferred.

As proposed we conducted a survey. Before going into details, there are few terms which one should be able to understand. They are PCA and EFA. PCA stands for Principal Component Analysis. PCA is a technique used for the classification and compression of data. Its purpose is to reduce the dimension of a data sample by finding a new set of variables from the original set. Application of PCA includes data compression, practical computation of PCs, probability distribution for
The survey consisted of nine random questions which were asked to a group of people. Their answers were recorded in a string of letters. Each reply was converted to a numeric value. These numeric values where the input to the tool (a software which does factor analysis). EFA with a varimax rotation was done in order to reduce the data. Once EFA is done we get a rotated component matrix as shown in fig. 3, in order categorize the questions into groups such that we get the hidden factors.

V. OBSERVATION & RESULT

<table>
<thead>
<tr>
<th>Rotated Component Matrix</th>
<th>Component 2</th>
<th>Component 3</th>
<th>Component 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAR00001</td>
<td>0.74</td>
<td></td>
<td>0.422</td>
</tr>
<tr>
<td>VAR00002</td>
<td></td>
<td>0.696</td>
<td>-0.170</td>
</tr>
<tr>
<td>VAR00003</td>
<td></td>
<td></td>
<td>0.674</td>
</tr>
<tr>
<td>VAR00004</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAR00005</td>
<td></td>
<td>0.711</td>
<td></td>
</tr>
<tr>
<td>VAR00006</td>
<td></td>
<td>0.016</td>
<td>-0.466</td>
</tr>
<tr>
<td>VAR00007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAR00008</td>
<td>0.568</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAR00009</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3 Factor analysis result

The figure explains us that from the survey the nine questions could be grouped into four such that it gives us the hidden factors. It was observed that in group 1, the two following questions were clubbed How often do you click on an ad because your friend liked it? and have you ever posted an ad on Facebook? This gave us our first hidden factor i.e. Business Use of Facebook. The next group consisted of the following questions Are the recommendation to join new groups or like new pages in accordance to your interest? And Do you see ads on Facebook in accordance with your interest or searches on search engines? This gave us our second hidden factor i.e. Interest /History. The third group had the following questions Are you satisfied with your Facebook experience? Are you satisfied with variety of content in your News Feed? And How often do you Facebook? This resulted in User experience, which is our third hidden factor. And finally our last group had questions like Are Facebook ads successful according to you? And Are parameters asked while posting an ad relevant? Hidden factors about ads were found.

VI. CONCLUSION & FUTURE WORK

It can be concluded that social media mining is a new initiative to build new business strategies. The Social media houses vast amount of user-generated data which can be used for data mining, therefore guarantee a huge potential in terms of knowledge. Market venturist’s are using various mining techniques to gain insight into business information for the intake of their sales/marketing and advertising teams. The mined information from social platforms can significantly impact business strategy of any business enterprise. An attempt was made to quantify and classify such approaches by different authors and compare such methods within a framework of mining parameters. Further we suggest how this survey and study of the data mining approaches can benefit the understanding the importance of social media mining.

While designing social media strategy for a company/ product several factors arise. But after factor analysis of these factors is conducted reduced hidden factors come up. These factors can be utilized to segment or group the audience. Using these hidden factors we can implement data mining in social media to generate effective business strategy.

References


[16] V. Sathiya Sentharam & Dr. Sai Satya Narayana Reddy. Data Mining Tasks Performed By Temporal Sequential Pattern


[18] Neelamadhab Padhy, Dr. Pragnyaban Mishra & Rasmita Panigrahi. The Survey of Data Mining Applications And Feature Scope