IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Changing Banking Business Model Using Sentiment Analysis

Shilpa B. L[1*], Shambhavi B. R[2]

[1] Department of Computer Science, VVIET, Mysuru, India
[2]Department of information Science, BMSCE, Bengaluru, India

[*]*Corresponding Author: blshilpa@gmail.com , Tel.: +91-9886893166*

Available online at: www.ijcseonline.org

Accepted: 18/Jan/2019, Published: 31/Jan/2019

***Abstract*—** Social media accounts like blogs, Facebook, Twitter and online discussion sites provide an option for an individual to express his or her opinion. These opinions are usually unstructured data and these are huge in amount. These days a massive number of users collect these recommendations or reviews for products and services, based on which they make their choices. The process of extraction of this insight from unstructured web data can be handled by Natural Language Processing and Big Data Analytics techniques. In this paper, we propose a model to extract this unstructured data from various domains, and then convert it into structured format by using various supervised algorithms. Finally the opinions or sentiments of the users will be presented for further understanding. Based on which the organization can take the necessary step to improve the customer retention.

***Keywords*—**Sentiment Analysis, Natural Language Processing, Unstructured data, Opinion Mining

I. INTRODUCTION

Social media or web based life is amazingly powerful for all businesses to advance and fortify their image. Numerous associations have likewise figured out how to utilize internet based life for significantly improving their services, and the banking industry is no exemption. Banks currently comprehend that internet based life's actual power is found in its capacity to associate with their clients. At the end of the day, banks need to utilize web based social networking to speak with customers, dispatch new items and contributions, show their company's history and show off all that they are doing for betterment of their clients or customers.

Correct implementation of social media offers several benefits to financial service industry. Few important one's among them are:

- In past banks needed to spend such a great amount of cash to discover what clients really need. Extensive surveys and feedback were collected. Internet based life has changed that. Today, web based life can encourage banks and financial service providers to get closer to their customers and know how to improve their products or services in a cost effective way.
- Online networking gives chances to plan client particular offers by collecting and analyzing the data collected on a real time basis.
- Through online networking, banks can offer their client's better client oriented services or offers. Through online networking stages, for example, websites, Facebook page, Twitter, and so forth they can specifically cooperate with their customers and clients and give answers for their grumblings.
- Promoting through web based life can be utilized for brand building. Traditional marketing methods are gradually being replaced by social media marketing. Smart utilization of video and substance can enable banks to reach to millions at one go.

The rest of the paper is organized as follows; Section II contains the related work in the field of Sentiment Analysis. Section III discusses about the various sources of data available to perform Sentiment Analysis. Section IV contains the proposed system. Section V describes the data set and analysis and Section VI concludes the research work with future enhancements.

II. RELATED WORK

In this section we describe the major contributions in the field of Sentiment Analysis.

Hamed M Zolbanin et al. [1] propose a new data processing approach that extracts individual- and database-level historical information from the medical records to improve the performance of readmission analytics. This method was tested and validated using two rather large data sets that belong to chronic diseases with the highest rates of hospital readmissions. This paper shows that Extracting historical information can improve the prediction of hospital readmissions and Comprehensive data processing provides a

competitive advantage to the health care organizations. In this paper large clinical data sets are processed using Big Data technologies and analytics.

This paper [2] reports on an exploratory study utilizing data mining techniques to predict leadership constructs based on game play data. The learning objective of the game is (1) to become aware of devilish dilemmas during crisis situations, and (2) to understand ones‟ leadership style in dealing with these dilemmas. Here authors evaluate several data mining techniques to predict scoring. They considered data set consisting of 21,600 instances. The main aim of this paper is to develop robust predictive models on the basis of which learning instructions could be given to the trainees during game play to increase their learning journey.

Peng, Kunrui, et al.[3] present a study that uses a database encompassing over 53,000 football recruits and over 200 predictive attributes to model the four aspects of collegiate football recruiting, as defined by Virginia's football coaches. Specifically, a desirable athlete is defined as one who 1) would succeed on the field at the collegiate level, 2) will meet Virginia's strict academic standards to achieve four years of playing eligibility, 3) fit Virginia football's "gritty" team culture, which is characterized by players who are resilient and able to overcome challenges, and 4) would commit to Virginia if given an offer.

Dabbas et al. [4] proposed a predictive model to elucidate the forecast performance in large business structures like electric power utility companies. The method uses artificial neural network (ANN) based predictive analytics viewed in data mining contexts. ANN based forecasting suites of data mining are concerned with cases where, the available data for training the ANN is inadequate and refers to a sparsely sampled data set. In other words, the ANN is trained with an ensemble of data whose sparsity is artificially recovered (and scarcity of samples removed) with WKS procedures; as well as, the number of data sets is increased to form an ensemble via re-sampling method of bootstrapping.

In [5] the author examines the use of massive, fine-grained data on consumer behaviour. In this paper one of the striking results shown is that there is no appreciable improvement from moving to big data when compared to using traditional structured data. However, in contrast, when using fine-grained behaviour data, there continues to be substantial value to increasing the data size across the entire range of the analyses. This suggests that larger firms may have substantially more valuable data assets.

Huge amount of data is not only generated in large firms or industries but also health care industry is producing massive amount of data. In [6] an EHR data management system is provided to process the massive amount of health care data. The system is built on Hive. Patient data is uploaded from variety of sources like flat files, web pages to Hive. This data is easily sent to reports application for generation of reports and charts. These charts are useful for doctors and researchers to understand and propose medications based on evidence from a large number of past patient records.

In paper [7] the authors describe a methodology and architecture to support the development of games in a predictive analytics context. These games serve as part of an overall family of systems designed to gather input knowledge, calculate results of complex predictive technical and social models, and explore those results in an engaging fashion.

Paper [8] applies predictive analytics techniques to establish a decision support system for complex network operation management and help operators predict potential network failures and adapt the network in response to adverse situations. The resultant decision support system enables continuous monitoring of network performance and turns large amounts of data into actionable information. This paper provides examples of actual power grid data to demonstrate the capability of decision support system.

III. SOURCES OF DATA

User opinions or reviews play an important role in the continuous improvement of services provided by the banks. With the growing use of internet, it is very important to understand the views of products and services provided to the customers. Some of the important sources of data are

A. Blogs

Blog is an internet platform which is similar to a website. A blog often publishes information which is not formal. Till 2009 blog was designed to be operated by a single user. But from 2010 onwards, the design was changed to accommodate multiple users. Blog usually allows the author to express ones personal opinion on a particular subject and doesn't need one to have knowledge about HTML. Blog was the first step in the field of social networking. It allows people to discuss and interact with other users on any particular topic.

B. Review Sites

Whenever a user wants to buy or take any service more than self, others opinions plays an important role. With the increase use of web these reviews or opinions are easily available over the web sites. These reviews are available over the web in unstructured format. Hence mostly they are not considered for any analysis. But these data can be used in sentiment classification.

C. Micro- blogging

Micro-blogging is the outstanding communication tool for web based users. A large number of messages appear daily in web-sites for micro-blogging such as Twitter, Tumblr and

Facebook. Twitter is the most popular among them, where users express their messages called "tweets". These tweets are used to express their own feelings or recommendations about various points. These tweets may sometime play an important role in Opinion Classification.

IV. PROPOSED SYSTEM

The proposed system collects the unstructured data from various sources like blogs, consumer complaints forum and Twitter. This task is performed by a crawler. Based on the keyword given the crawler crawls the data. The crawler is designed in such a way that it extracts only the text data and does not extract images or videos present the web page. The crawler4j API is used to perform this task.

Once the data is crawled from different sources next step is to pre-process it and integrate it into a single repository. In this step there were many challenges that were encountered. Few important ones among them are:

- The feedback or comments or reviews were not only given in English language but they were in other languages as well. Hence it was very important to identify the same and consider only those comments which are in English language.
- If a user one comments "the credit limit given on my card is excellent" and user two comments "the credit restrain is very good on my card". In this case both are talking about the same feature group but with different wordings. Hence the two needs to be grouped to same feature group.
- As the users are free to comment in any form in the forums. The user or customer can use abbreviations or short words in their comment. For example CC no for credit card number, gud for good, v for we etc. These words need to be annotated to their root word by using various methods.
- Need for identification of spam and fake reviews, which can be done mainly by identifying duplicates.

The process of integrating data set of similar classes is called as clustering. In this work we are using k-means clustering approach. The Algorithm for k-means is as follows,

1. Randomly select *c* centroids.
2. Calculate distance from each centroid *vi* to all data points *xi*.
3. Repeat
4. Assign each data point xi to the centroid with minimum distance.
5. Recalculate new centroids from each new
6. Recalculate distance from each new centroid to all data points.
7. Until no data point was reassigned a new centroid.

It concerns with two steps. First step is to determine 'k' centres randomly for each cluster; the Second step is about determining the distance between data points in dataset and finding the centroid by assigning the data point to its nearest cluster [9, 10].

The two steps are done by using Square Error Criterion Method and Euclidian distance method. The Square Error Criterion for k-means is defined by equation (1).

$$E = \sum_{i=1}^{k} \sum_{peci} |p - mi| \rightarrow \quad (1)$$

P is the data point, Mi is the Centroid for Cluster Ci and E is the sum of squared error of all points in dataset.

The Euclidian distance is generally used to calculate the distance between data points and centroid of a cluster. The distance between two vectors can be calculated by using equation (2).

$$X = (X1, X2, X3, \ldots . Xn) \;\&\; Y = (Y1, Y2, Y3, \ldots . Yn)$$

$$d(Xi, Yi) = \sqrt{\sum_{i=1}^{n} (Xi - Yi)^2} \rightarrow \quad (2)$$

The figure 1 below depicts an overview of the system proposed.

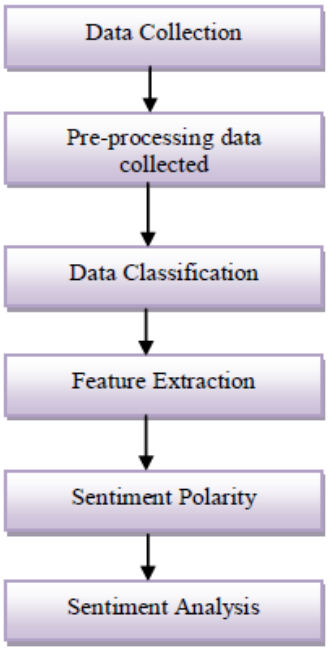

Figure 1: System Overview

Next step is to classify the data into different sentiment classes such as Positive, Negative and neutral. To do this we have used Python's Textblob module. This module in turn

uses Naive Bayes and Decision tree classifier to classify the pre-processed data. Naive Bayes is probabilistic classifier that belongs to machine learning family that assumes feature independence. It is simple and most commonly used classifiers. It is based on Baye's theorem and is preferable for the input with high dimensions.
Naive Bayes is a probabilistic classifier, meaning that for a document d, out of all classes c 2 C the classifier returns the class ˆc which has the maximum posterior probability given the document. In equation (3) we use the hat notation ˆ to mean "our estimate of the correct class".

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) \quad \rightarrow (3)$$

The intuition of Bayesian classification is to use Bayes" rule to transform equation (3) into other probabilities that have some useful properties. Bayes" rule is presented in equation (4); it gives us a way to break down any conditional probability P(x|y) into three other probabilities:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \quad \rightarrow (4)$$

We can substitute equation (4) in (1) to get equation (5)

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) = \underset{c \in C}{\operatorname{argmax}} \frac{P(d|c)P(c)}{P(d)} \quad \rightarrow (5)$$

We can easily simplify the equation (5) by dropping the denominator *P(d).*

Finally based on the polarity an opinion model is presented. This model provides many advantages like

- It is more affordable in contrast to the traditional approach of obtaining customer feedback.
- It gives deft approach to pick up client bits of knowledge.
- Helps in knowing the industry's strengths and weakness.
- Gives capacity to follow up on client proposal
- Provides more exact and wise understanding of customer's perceptions and reviews.

V. DATA SET AND ANALYSIS

The Proposed method developed is implemented using Python. Python is a high level, interpreted programming language, created by Guido van Rossum. We have used Textblob a text processing library file in Python. It permits to perform different natural language processing (NLP) task such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. It embodies some of the best classifiers such as Naïve Bayes, decision tress etc hence an efficient tool for sentiment analysis.

The data set consisted of 5621 opinions of users regarding banking domain, among which 2821 were from customer complaints website, 1800 were taken from twitter, 126 were taken from blogs, 104 were taken from different discussion forum and remaining 770 was taken from Facebook.

The steps followed to obtain the unstructured text data is presented in the previous section. This data is initially analysed without performing any pre-processing. When the data was considered without pre-processing, in this case though many of the text data contained useful information in it, it was ignored. Table I shows the sample data set being analysed without performing any pre-processing of the data. But it shows that most of them are considered as neutral hence the same sample was pre-processed and analysed.

Table 1. Performance of Combined Data Sources

Data Set	Positive	Negative	Neutral
Without Pre-Processing	39.2	31.6	29.2
With Pre-Processing	42.35	32.67	24.98

Clearly there is drastically difference in examination of the same data set before and after pre-processing.
Figure 2 below shows the same when applied separately on each data set.

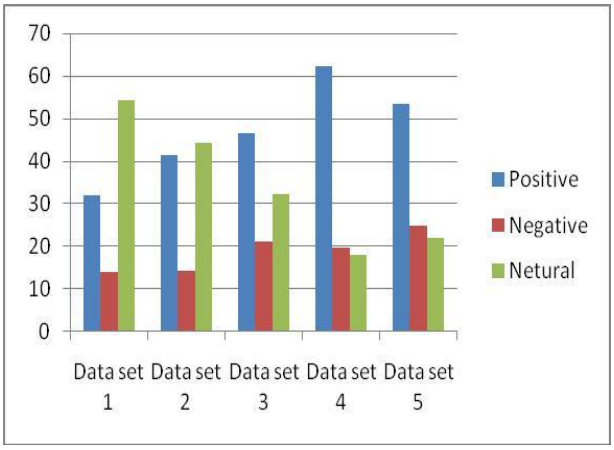

Figure 2: Performance on individual data set

VI. CONCLUSION AND FUTURE WORK

The advanced usage of web based life has urged for a system to analyze this data. This paper presents the proposed model for predicting the opinions of customers about various products or services. Also it clearly highlights the importance of pre-processing the extracted data. Consequently opinion examination has unmistakable standpoint for future improvement.

The proposed technique is utilized in all for all the unstructured information that is gathered for managing a banking industry. This framework does not arrange for the separation of information into various classifications like credit card remarks are unique in relation to loan related and they should be skewered. This should be possible sooner rather than later for more exact outcomes.

REFERENCES

[1] Zolbanin, Hamed M., and Dursun Delen. "Processing electronic medical records to improve predictive analytics outcomes for hospital readmissions." *Decision Support Systems* (2018).

[2] De Heer J., Porskamp P. (2019) Predictive Analytics for Leadership Assessment. In: Kantola J., Nazir S., Barath T. (eds) Advances in Human Factors, Business Management and Society. AHFE 2018. Advances in Intelligent Systems and Computing, vol 783. Springer, Cham

[3] Peng K, Cooke J, Crockett A, Shin D, Foster A, Rue J, Williams R, Valeiras J, Scherer W, Tuttle C, Adams S. Predictive analytics for University of Virginia football recruiting. InSystems and Information Engineering Design Symposium (SIEDS), 2018 2018 Apr 27 (pp. 243-248). IEEE.

[4] Dabbas, Mohammad, Perambur S. Neelakanta, and Dolores DeGroff. "ANN-based predictive analytics of forecasting with sparse data: Applications in data mining contexts." In *Recent Trends in Information Technology (ICRTIT), 2013 International Conference on*, pp. 62-67. IEEE, 2013.

[5] Martens, David, Foster Provost, Jessica Clark, and Enric Junqué de Fortuny. "Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics." *MIS quarterly* 40, no. 4 (2016).

[6] Chennamsetty, Haritha, Suresh Chalasani, and Derek Riley. "Predictive analytics on Electronic Health Records (EHRs) using Hadoop and Hive." In *Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on*, pp. 1-5. IEEE, 2015.

[7] Riensche, Roderick M., Patrick R. Paulson, Gary Danielson, Stephen D. Unwin, Scott Butner, Sarah Miller, Lyndsey Franklin, and Nino Zuljevic. "Serious Gaming for Predictive Analytics." In *AAAI Spring Symposium: Technosocial Predictive Analytics*, pp. 108-113. 2009.

[8] Huang, Zhenyu, Pak Chung Wong, Patrick Mackey, Yousu Chen, Jian Ma, Kevin Schneider, and Frank L. Greitzer. "Managing Complex Network Operation with Predictive Analytics." In *AAAI Spring Symposium: Technosocial Predictive Analytics*, pp. 59-65. 2009.

[9] M.P.S. Bhatia and Deepika Khurana, " Experimental Study of Data Clustering using k-means and Modified Algorithms", International Journal of Data Mining and Knowledge Management Process(IJDKP), Volume 3, No. 3, May-2013.

[10] Srihari A.Hudli, Aditi A.Hudli, Ananad V.Hudli, "Identifying Online Opinion Leaders Using K-means Clustering", 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA).

[11] J.V.N. Lakshmi, Ananthi Sheshasaayee, "*A Big Data Analytical Approach for Analyzing Temperature Dataset using Machine Learning Techniques*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.3, pp.92-97, 2017

[12] R.Anupriya, P.Saranya, R.Deepika, "*Mining Health Data in Multimodal Data Series for Disease Prediction*", International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.2, pp.96-99, 2018

Authors Profile

Mrs. Shilpa B L pursed Bachelor of Engineering from Visvesvaraya Technological University, India in 2009 and Master of Technology from Visvesvaraya Technological University, India in year 2011. She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science and Engineering, VVIET, Mysuru, Affiliated to Visvesvaraya Technological University, India. She is a life member of the IAENG since 2013, ICSES since 2017. Her main research work focuses on Predictive Analytics, Big Data analytics, Natural Language Processing, Data mining and Machine Learning. She has 7 years of teaching experience and 2 years of Industry Experience.

Dr. Shambhavi B R completed her Ph.D. from Visvesvaraya Technological University, India in the area of Natural Language Processing. She is a life member of Indian Society for Technical Education (ISTE) and IEEE. He has published more than 20 research papers in reputed international journals. Her main research work focuses on Natural Language Processing, Big Data Analytics and Mining and Analysis of Data Patterns. She has 12 years of teaching experience and 3 years of Industry Experience.

