International Journal of Engineering & Technology, 7 (4.19) (2018) 366-368

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Twitter based Analytics for Business Footprints of the Banking Sector in India

Sheetal Ambulkar[1*], Akshay Saraf[2], NeeranjanChitre[3]

[1,2]G H Raisoni University, Saikheda MP
[3]Professor, GNIT College Nagpur
**Corresponding Author E-mail: [1]Sheetalanandkale@gmail.com*

Abstract

Twitter is one of the world's most popular social media platforms with over 330 million users. Many businesses use Twitter to reach and connect with their customers. There are a number of advantages that using twitter can bring to a business. Some of the prominent advantages of tweeter platform to the businesses are the reach to broad spectrum of customers worldwide, delivery of customer services, establishment of brand identity, gathering customer feedback etc. The best part of it is that these advantages are at the free of cost. This paper proposes a system, that can help the banking sector in India, to compare their impression on the general customer by analyzing the tweets by the banking organizations and the replies by their customers The analysis includes the tweet handles by two nationalized banks (State Bank of India and Punjab National Bank) and two private sector banks (ICICI and HDFC).

Keywords: *Big data applications, Data mining, Data visualization, Machine learning algorithms, Natural language processing, Sentiment Analysis, Social computing, Social Intelligence, Twitter.*

1. Introduction

Twitter is one of the biggest marketing platforms for any business in the world. With faster digitization in the banking sector of India, twitter has become one of the come popular mechanisms for banks to promote their products, offers and services to the customers. For customers the twitter has become an easy and quick way to share their feedback and comments from anywhere in the world.

Although the mechanism to interconnect between banking sector and customer has become easy, to extract the sentiment of the customers and their perspective towards the products is still a challenge to the banking businesses. Often in the business, the offerings and responses to the competitor organization is very crucial for further planning.

This ideation paper proposes a system that provides a twitter based approach towards understanding the baking market from customers view point and improve the bank's brand, design future marketing strategies and campaigns format.

- Abbreviations and Acronyms

API – Application Programming Interface
DB – Database
NLP – Natural Language Processing

2. Related Work

Strategic use of social media data not only impacts the way in which the financial institutions market their product and services, but also on how they conduct competitive analysis for product and service design. Banks have established their presence on social media like Facebook, Twitter, and LinkedIn. Twitter is a massive social media which enables microblogging through tweets which are public. Every word, photo, video and follower can have an impact. There is huge volume and variety of data on twitter which can be analyzed using the big data approaches suited for financial sector. [1,2]

Additionally, Twitter API's are available for publically practitioners and researchers which can aid in data analysis of twitter data. [1]. each account on Twitter is associated with a unique id and a unique Twitter handle which can be used to retrieve the profile and tweets for data analysis.

Twitter data analytics was researched in the past on a variety of domain like Stock market [4], supply Chain [5], Hospitality [6] etc. Tweets extraction was done for a span varying from 3 to 12 months by different authors. Tweet is an unstructured data, which needs to be filtered by using various Natural language processing techniques like stemming, stop word and proper noun removal [7] to obtain the useful data.

The most important phase of this research lies in mining the data for extracting knowledge for gaining deep insights into the tweets for customer behavior, feedback on products and complaints.

Sentiment analysis will find out the sentiments of the costumer to be positive, negative or neutral. Most of the literature uses lexicon based approach, but this requires a good and powerful dictionary which is not always available. Semantics of the text also plays a major role when performing sentiment analysis, which is usually ignored. Researchers have found that increased accuracy can be achieved if semantics are incorporated. [8]. A bunch of research have been done on prediction of the stock market data based on the sentiment analysis and other algorithms like - SOFNN (Self Organizing Fuzzy Neural Networks) [10]

A variety of NLP algorithms are used for clustering like K-mean, hash tagging, TF-IDF [11] and then context analysis should be carried out on the tweets for understanding the correlation, aggregation and association form the tweets. Recent focus is on the topic Spatio-temporal clustering of social media data[12] .

3. System Design Units

This section presents the different modules of our system. Overall, the system has five main process components (refer figure 1) viz Data extraction, tweets pre-processing, data mining, Data benchmarking and data visualization as depicted in figure 1.

Fig. 1: System Components

The process follow is depicted in figure 2

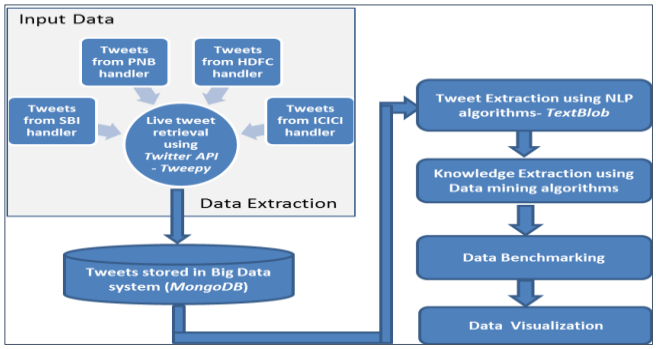

Fig. 2: Process Flow

1. Input data:
The source for input data is Twitter. To ensure diverse coverage, Tweets for the four banks – two private –ICICI, HDFC and two nationalized banks – SBI, PNB are considered in scope for the analysis. To acquire a sufficient volume of data tweets for a year will be under consideration. Further to uncover the association trends in the products bought by customers, multiple twitter handles for the same banking institutions will be used for extracting the data.

2. Data Acquisition and Filtering
This module extracts the domain specific data from Twitter for the four banks in scope, using the Twitter API – Tweepy. The domain specific tweets are store in a NoSQL database – MongoDB for further processing. NoSQL database was a good choice to store the unstructured tweets and also has a good range of aggregation functions for querying unstructured data.

3. Tweets Pre-processing and Extraction
This tweets dataset composed for short messages (tweets) require a series of customized preprocessing to extract the keyword for quality analysis.
As a first step, all proper nouns are removed, this ensure that person and place names are not inferred as keywords. Unlike this research where proper nouns have no weightage, there could be some scenarios where proper nouns are important keywords like for election campaign analysis person name has to be categorized as per the political party, then the proper noun elimination should be skipped. Further in the tokenizing step the URL and special characters are discarded.
Tweets are normalized by correcting common English words used on social media like plssss is replaced with please, b'tween with between. This process uses dictionary for common social media representations,
Stop words like is, an, and, which etc that don't add value to the domain specific words are removed. The stop words reference dictionary is customizable.
Next stage words are stemmed, works like "complain" and "complaint" have the same root. Hence in the next phase of filtering the root is retained and remaining words are discarded.
Length based filtering is done on the domain specific words. Words with length less than 3 and more than 15 are removed.

4. Knowledge Extraction Using Data Mining :
Until this step, the input data is mere individual text entities, without any categorization, association and no inference drawn from the data. Knowledge extraction module does the following using the natural language processing algorithms:

- Clustering the topics by using the Hash tags used in the tweets, replies and retweets, which aids in understanding the widely discussed topic by the customers and can be an area of focus.
- Sentiment analysis to determine customer's perception about the products which is be a vital input for the future strategy design. Sentiment analysis is well-known and the widely used technique, which helps in identifying public sentiment from underlying text. A lexicon based algorithm is used, which is proven to the best accuracy, with support of a well-constructed dictionary.
- K-means clustering algorithm is further used to categorize the data based on the products that are under consideration.
- Association in the products can be uncovered using the Mongo DB association and aggregation techniques. The association input will be of great importance to deciding the offers on the products.

5. Data Benchmarking: Social media analytics of your account against the industry peers' social media presence would reveal key metrics on positioning your products.

6. Data Visualization: In addition to the above steps of processing and data mining, it is equally important to presents the insights drawn through various visual techniques in a comprehensible manner. Usage of appropriate visualization technique said in ease of understanding and hence increases the probability of usage of this metrics into actions.

3. Tools and Technology to be used

The following open source languages, tools and libraries would be used for implementing the proposed system:

- Python:

Python is an open-source and object-oriented programming language. As it is open source, there is an availability of many libraries and APIs to perform one function. Extraction of Twitter APIs is easily facilitated by the usage of Python. It can be utilized for a wide range of applications like scripting, developing and testing.It is majorly preferred over the other scripting languages, because of its elegance and simplicity.

- Twitter API (Tweepy & TextBlob):

Tweepy is a library in Python, for accessing the Twitter APIs.These APIs are used to extract and download the messages and tweets in real time. High volume of tweets and creating a live feed is facilitated only using this library.
TextBlob is another library provided by Python, which is used to carry out processing on textual data, extracted through the Tweepy library. It provides a simple API to carry out NLP tasks. It can efficiently carry out NLP tasks like noun phrase extraction, translation and analysis.

- MongoDB

MongoDB is an open-source, platform independent document-oriented database. It can be easily paired with Python using the

official connectors provided by MongoDB. MongoDB is mainly used with Python because of the support in field, range query, and regular expression searches. Besides just a database program, MongoDB can also be used as a file system.

- PyQT

PyQt is an open source plugin in Python to bind Python and GUI toolkit Qt for visual representation of the findings.

4. Conclusion

With exponential increase in social media usage in recent years by various banks primarily to market their products. This research work will exploit social media data analytics using twitter data to produce some customized metrics, which would be inputs to the bank for design of new products, revising the social media strategy etc. The data analysis would cover multiple facets – sentimental analysis, domain centric approach and customized visualization. Also factoring some descriptive context analysis with the clustered tweet data, will aid in understanding the exact customer perspective on the product along with the emotion.

References

[1] Shamanth Kumar, Fred Morstatter, HuanLiu, "Twitter Data Analytics" in Springer, Aug 2013, Pages –1-5

[2] Jennifer. Q. Trelewicz, Big Data and Big Money: The Role of Data in the Financial Sector , Accession Number: 16948585 , DOI: 10.1109/MITP.2017.45, Publisher: IEEE

[3] Anber, H., A. Salah, and A. A. Abd El-Aziz, "A Literature Review on Twitter Data Analysis", International Journal of Computer and Electrical Engineering, vol 8(3), 2016: Journal of Computers, 2016..

[4] Spin-offs in Indian Stock Market owing to Twitter Sentiments, Commodity Prices and Analyst Recommendations Voice of Customers: Text Analysis of Hotel Customer Reviews (Cleanliness, Overall Environment & Value for Money)

[5] Bongsug (Kevin)Chae, Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research, International Journal of Production Economics Volume 165, July 2015,

[6] Bangalore paper

[7] Anber, H., A. Salah, and A. A. Abd El-Aziz, "A Literature Review on Twitter Data Analysis", International Journal of Computer and Electrical Engineering, vol 8(3), 2016: Journal of Computers, 2016..

[8] Crockett, KA and Mclean, D and Latham, A and Alnajran, N (2017) Cluster Analysis of Twitter Data: A Review of Algorithms. In: 9th International Conference on Agents and Artificial Intelligence (ICAART), 24 February 2017 - 26 February 2017, Portugal.

[9] Anshul Mittal , ArpitGoel , "Stock Prediction Using Twitter Sentiment Analysis", Stanford University

[10] Zhichao Han, DATA AND TEXT MINING OF FINANCIAL MARKETS USING NEWS AND SOCIAL MEDIA, A Dissertation submitted to the University of Manchester

[11] M. Budde, J. De Melo Borges, S. Tomov, T. Riedel,and M. Beigl. Leveraging spatio-temporal clustering for participatory urban infrastructure monitoring. In Proceedings of the First International Conference onIoT in Urban Space, pages 32{37. ICST (Institute for Computer Sciences, Social-Informatics and Tele-communications Engineering), 2014.

[12] Tarmazakov, E.I., Silnov, D.S. Modern approaches to prevent fraud in mobile communications networks (2018) Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018, 2018-January, pp. 379-381. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048047820&doi=10.1109%2fEIConRus.2018.8317111&partnerID=40&md5=fd2de2dd9837e66316a7f042763b9927 DOI: 10.1109/EIConRus.2018.8317111

[13] Balanyuk, Y.B., Silnov, D.S., Goncharov, D.E. Applying memshift technology to increase GPU performance (2018) Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018, 2018-January, pp. 275-276. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048035288&doi=10.1109%2fEIConRus.2018.8317084&partnerID=40&md5=b8306ca07043e666227858a7e4ef6cb1 DOI: 10.1109/EIConRus.2018.8317084

[14] Goncharov, D.E., Zareshin, S.V., Bulychev, R.V., Silnov, D.S. Vulnerability analysis of the Wifi spots using WPS by modified scanner vistumbler (2018) Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018, 2018-January, pp. 48-51. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048157038&doi=10.1109%2fEIConRus.2018.8317027&partnerID=40&md5=bcc54f5086136ebe0d68e414ef303a46 DOI: 10.1109/EIConRus.2018.8317027

[15] Mushtakov, R.E., Silnov, D.S., Tarakanov, O.V., Bukharov, V.A. Investigation of modern attacks using proxy honeypot (2018) Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018, 2018-January, pp. 86-89. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048016646&doi=10.1109%2fEIConRus.2018.8317036&partnerID=40&md5=9a6ad869d7cf18b4bf8f26edfc218093 DOI: 10.1109/EIConRus.2018.8317036

[16] Frolov, A.A., Silnov, D.S., Geraschencko, Y.Y., Sadretdinov, A.M., Kiamov, A.A. Research of mechanisms counteracting the distribution of prohibited content on the Internet (2018) Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018, 2018- January, pp. 298-302. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047985123&doi=10.1109%2fEIConRus.2018.8317092&partnerID=40&md5=0dc4ec2fb660a7ed2d24babd6ea93734 DOI: 10.1109/EIConRus.2018.8317092

[17] Prokofiev, A.O., Smirnova, Y.S., Silnov, D.S. Examination of cybercriminal behaviour while interacting with the RTSP-Server(2017) International Conference on Industrial Engineering, Applications and Manufacturing, ICIEAM 2017 - Proceedings, art. no. 8076437. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85039927486&doi=10.1109%2fICIEAM.2017.8076437&partnerID=40&md5=ad99910201c39909410085d73349a039 DOI: 10.1109/ICIEAM.2017.8076437

[18] Prokofiev, A.O., Smirnova, Y.S., Silnov, D.S. The Internet of Things cybersecurity examination (2017) Proceedings - 2017 Siberian Symposium on Data Science and Engineering, SSDSE 2017, art. no. 8071962, pp. 44-48. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85040368095&doi=10.1109%2fSSDSE.2017.8071962&partnerID=40&md5=d949f724d5786343634f9e49ba0d65a2 DOI: 10.1109/SSDSE.2017.8071962

[19] Arzhakov, A.V., Silnov, D.S. Architecture of multithreaded network scanner (2017) International Conference of Young Specialists onMicro/Nanotechnologies and Electron Devices, EDM, art. no. 7981704, pp. 43-45. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85027159579&doi=10.1109%2fEDM.2017.7981704&partnerID=40&md5=b8dda0ef229557a5383568bbb688dfbe DOI: 10.1109/EDM.2017.7981704

[20] Mushtakov, R.E., Silnov, D.S. New approach to detect suspicious activity using HTTP-proxy honeypots (2017) Proceedings of the 2017 IEEE Russia Section Young Researchers in Electrical and Electronic Engineering Conference, ElConRus 2017, art. no. 7910525, pp. 187-189. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85019479586&doi=10.1109%2fEIConRus.2017.7910525&partnerID=40&md5=ce6f84a452a95b67057221e22900a6ce DOI: 10.1109/EIConRus.2017.7910525

[21] Taran, A., Silnov, D.S. Research of attacks on MySQL servers using HoneyPot technology (2017) Proceedings of the 2017 IEEE Russia Section Young Researchers in Electrical and Electronic Engineering Conference, ElConRus 2017, art. no. 7910533, pp. 224-226. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85019476121&doi=10.1109%2fEIConRus.2017.7910533&partnerID=40&md5=066268db0f19141b80a6be0edc0ee8c1 DOI: 10.1109/EIConRus.2017.7910533.