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

Sheetal Ambulkar\textsuperscript{1*}, Akshay Saraf\textsuperscript{2}, Neeranjan Chitre\textsuperscript{3}

\textsuperscript{1,2}G H Raisoni University, Saikheda MP
\textsuperscript{3}Professor, GNIT College Nagpur
\textsuperscript{*Corresponding Author E-mail: sSheetalanandkale@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 common 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 banking 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.

![System Components](image)

**Fig. 1: System Components**

The process flow is depicted in figure 2.

![Process Flow](image)

**Fig. 2: Process Flow**

1. **Input data:**
   The source for input data is Twitter. To ensure diverse coverage, Tweets for 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 identifying 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 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 MongoDB 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


[6] Bangalore paper


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

