© 2018 IJSRST | Volume 4 | Issue 5 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Forecasting Product Sale from Twitter using Hidden Markov Model

Dr. Mohd. Abdul Raffey, Mr. Pankaj N. Patil, Dr. Sunil Kawale
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India

ABSTRACT

main aim of this research proposal is to improve the forecasting power of product sales in the different location by considering the hidden issues which affect to make the prediction goes wrong. This work carried with the help of statistical model i.e. Hidden Markov Model (HMM) applied on the extracted dataset of product review from popular micro blogging site "Twitter".

Keywords: Sentiment Analysis, Data Mining, Hidden Markov Model, Twitter, Feature selection, Predictive Analytics.

I. INTRODUCTION

From last decade the machine learning techniques are widely used for every industry to improve their business profits by identifying the failures in advance and to archive this they used predictive Analytics, Data mining and Statistical Techniques to forecast the future. Sale prediction or sale forecasting is having very critical impact on the product sale if the industries have an idea whether the product will going to make good sale in particular location then some advance provision for product supply vice-versa if they know the product is not good then industries make research the causes why and what issue are there and how to overcome them and improve the business.

Data mining helps in discover knowledge from the huge amount of data set to use it Data Mining techniques are helpful for categories data into desired classes. Here we use social media “Twitter” as a platform to collect the data i.e. product reviews of users. Product reviews are collected from all type of tweets posted on Twitter to decide the quality of a product. The review data is extracted with the help of python programming language using “Tweepy” Application Programming Interface (API) such reviews can be used for further analysis.

The Markov Model is the stochastic method mostly used for the system which is randomly changing in behaviour. It will assume that the next state is not dependent on the previous state this model will show all the possible probabilities between them. The Markov model often used to pattern recognition and making the prediction. The Hidden Markov Model (HMM) discovers the hidden stats of the general Markov Model, therefore, the Hidden Markov Model can also use to find an effect on product sales.

II. LITERATURE REVIEW

Wai-Ki CHING et.al (2007), proposed new Multivariate Markov Chain Model which can apply on multiple categorical data sequences and also tested on synthetic data, sales demand data. This model is also useful when the data sequences are short easily. Md. Rafiul Hassan (2009) presents a paper forecasting on the stock market using a combination of Hidden Markov Model and Fuzzy Model, where the HMM is for identifying data pattern and fuzzy Model is used to obtain forecasting value. In research, he concludes

that generated Fuzzy Model is much suitable for best performance. Ivan JANICIJEVIC et.al. (2014) described the Factor Impacting Product Quality (FIPQ) with a hypothetical example for improving product quality and quality management done with Markov chain. Lihong Li et.al. (2014) used Markov Chain Theory on actual market share analysis and also proposed the forecasting model of market share and applied on the municipal automobile sales market forecast, it will improve the market competitiveness of enterprises. Adam Westerski et.al. (2015), explained the characteristics of procurement dataset and also stated the effect on future purchase problem and how to solve them with Markov Chain Model for proper input dataset analysis to understand the horizon of practical deployment possibilities. Hierarchical clustering used for preprocessing it leads to improvement of raw data. Ka Ching Chan (2015), presented four mathematical model as time-varying Markov Model, new extended Time-Varying Markov Model, Novel Markov Model and Homogenous Markov Model, these are illustrated with telecommunication industries to forecasting their market share and show how they all react for the same problem with the different assumption. Samaneh Beheshti-Kashi (2015) has done the survey of forecasting sale in fashion market by using state-of-the-art method, author reviewed different strategies to the predictive value of user-generated content and search queries, from the study author, tell Conventional forecasting methods to face challenges in producing accurate data, uncertain demand, seasonality, product variability and lack of historical data can hardly handle. Pankaj Patil et.al. (2017) experiments on 2000 product review tweets and prove Principle Component Analysis (PCA) is the very useful technique for dimensionality reduction.

III. PROPOSED SYSTEM

Over the traditional model, we have used Hidden Markov Model to forecasting the future with considering the Hidden issues affected to specific product data (reviews about the product) is collected from most popular blogger is "Twitter" as a dataset.

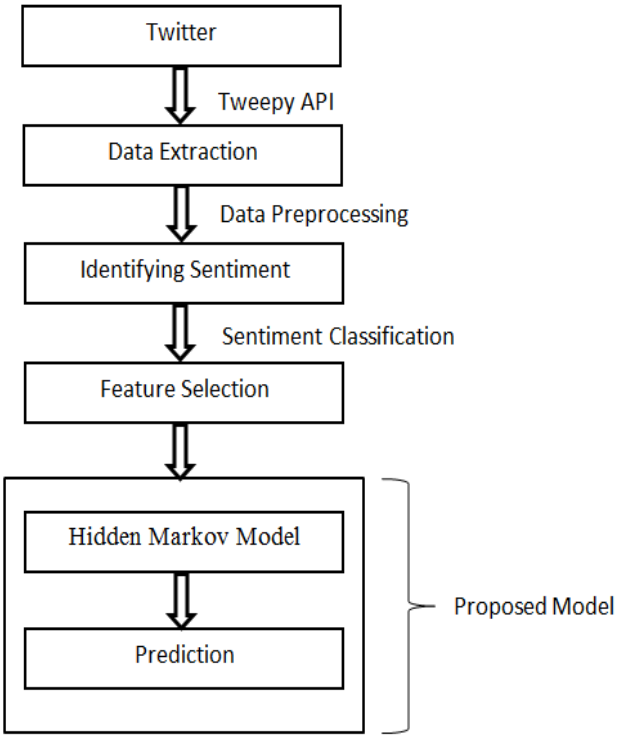


Figure 1. Proposed system

Data Collection

In this step, data is collected from most famous microblogger site i.e. "Twitter". The numbers of tweets are too large so it is not possible to select the tweets manually; hence python is used with "Tweepy.api" as an interface to extract tweets directly from Twitter. To extracting such tweets we need to generate the necessary credential i.e. for example, access_token, access_token_secret, consumer_key & consumer_key_secret, for establishing the connection from the authorized Twitter account. The extracted data is captured to a text file in the format of JSON (JavaScript Object Notation) because it is human-readable format as well as machine also easily parses it.

Pre-processing of dataset

The pre-processing step is involved transferring raw data extracted with the machine to the human understandable format while extraction the data often comes with some unwanted symbols and links so to remove them and resolving such issue the data pre-processing is the important step it prepares the raw data for future processing. It always used in database driven like customer relationship management system and rule-based application like Neutral Network. The extracted data need to go through the series steps of pre-processing like Data Cleaning, Data Integration, Data Transformation, Data Reduction and Data Discretization. These steps are used to clean following contain from data.

- ✓ The raw data contain Escaping HTML characters
- ✓ Decoding data from "utf8" to "ASCII"
- ✓ Apostrophe Look up
- ✓ Stop words removal
- ✓ Punctuation removing
- ✓ Expression removing
- ✓ Standardizing the word
- ✓ URL/link removing

Table 1. Example of Raw and Processed Tweet

Tweet	Battery is woefully inadequate, 2 stars less for the battery. Else a great phone.
Processed Tweet	Woefully, inadequate, less, great

Identifying Sentiments

Third step is of identifying sentiment, to identify the expression of tweets we need to perform sentiment analysis with python programming language to differentiate the extracted tweets in categories like positive, negative & neutral, because of the vast library each extracted word is compared with the popular positive negative word dictionary written by "Bing Liu: Opinion Lexicon" which having 6800 words. After classification stage, the score of tweets defined to depend on this the complete tweets determined as Positive, Negative or Neutral.

There are number of methods to calculate the sentiment score of the sentence but here we are using one of the popular methods.

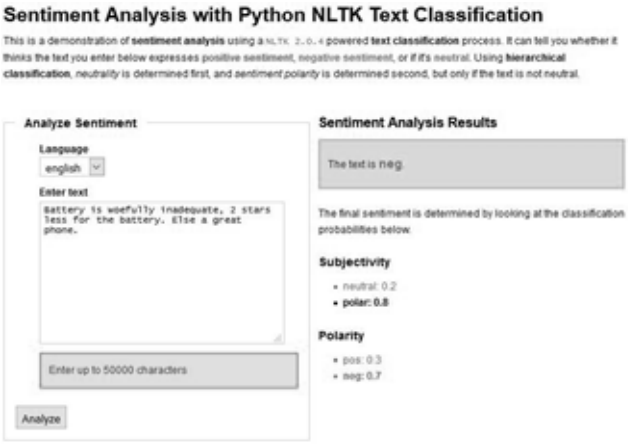
$$\text{Sentiment} = (P - N) / (P + N + O)$$

Where, P=Positive Word(s)

N=Negative Word(s)

O= Total Word(s)

So, those total numbers of words are (P+N+O)



(Source: http://text-processing.com/demo/sentiment/)

Figure 2. Demo of Sentiment Analysis

The following scale shows how the sentence is measured by their score

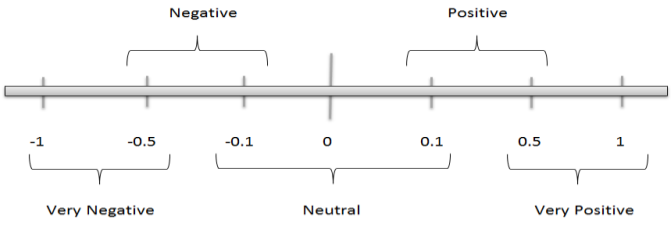


Figure 2. Sentiment Score Range

i.e. if the score is in range 0.5 to 1 then the sentence or word considered as Very Positive, 0.1 to 0.5 as Positive, -0.1 to 0.1 Neutral, -0.1 to -0.5 as Negative and for -0.5 to -1 is Very Negative.

Feature Selection:

The feature selection step is the data reduction. It is also called as Variable Selection or Attribute Selection at this stage we got the idea to decide to select variable under consideration for decision making by finding the correlation of all variable with each other and percentage of extraction for the understanding of which variables are extracting the maximum meaning from the whole dataset. These are the final variable which helps you to make the correct decision with utilizing the processing and CPU memory.

Hidden Markov Model (HMM)

We used Hidden Markov Model to find what will independently happen next from the past collected data from a source.

Here is the practical scenario that explains how it works, supposes we want to forecast what will be the user's emotions about the product for this we have to decide how many tweets to be considered in the study. It is a very hectic task to collect the data but the Hidden Markov Model will provide the facility to make the decision from last past data.

$P\ (tweetEmotion_t \mid tweetEmotion_{t-1}, tweetEmotion_{t-2}, \ldots, tweetEmotion_{t-n})$

Here 2000 past tweets are taken for proposed model and go through the following steps.

(1)Calculate probabilities based on the past product reviews

- For example how many users are taking positive, how many will be negative about the product and some as the neutral.

(2)Used Naïve Bayes Probability Equation for calculating probabilities.

$$P(\text{tweetEmotion X given Y}) = \frac{P\ (\text{tweetEmotion X and Y occouring})}{P(\text{TweetEmotion Y})}$$

- The probability of Product X will Positive, Given that Product is Negative in last time
- The probability of Product X will Negative, Given that Product is Positive in last time

(3) Calculate the probability of each state (Positive, Negative & Neutral) as follows

- P (Positive | Negative) is Product X will good today but bad last time.
- P (Positive | Neutral) is Product X good today but no tweets last time.
- P (Positive | Positive) is Product X good for today and also good for last time.

IV. ANALYSIS

By the proposed model above, following probabilities calculated for the transition probability matrix of hidden Markov model i.e.

$P\{ X_t = Positive / X_{t-1}=Positive \} = 0.61$
$P\{ X_t = Positive / X_{t-1}=Negative \} = 0.21$
$P\{ X_t = Positive / X_{t-1}=Neutral \} = 0.18$
$P\{ X_t = Negative / X_{t-1}= Negative \} = 0.32$
$P\{ X_t = Negative / X_{t-1}=Positive \} = 0.44$
$P\{ X_t = Negative / X_{t-1}= Neutral \} = 0.24$
$P\{ X_t = Neutral / X_{t-1}= Neutral \} = 0.7$
$P\{ X_t = Neutral / X_{t-1}=Positive \} = 0.41$
$P\{ X_t = Neutral / X_{t-1}= Negative \} = 0.52$

The following chart depicts transition probabilities, the circles in the chart are denotes all possible states.

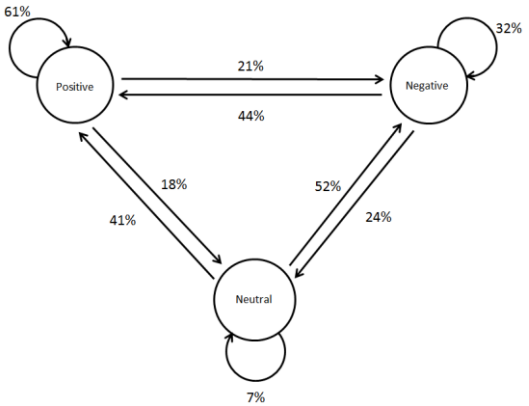


Chart1. Transition Probabilities

However, using the above chart and Markov Assumption, we can easily forecast whether the next tweet will be positive or negative.

Suppose, a current tweet is positive (i.e. 61%) then maximum chances of tweets are positive (i.e.60%) then next is Negative (i.e. 21%) then the calculation is as 61 % *61 % *21 % equals to 78 % and so on.

V. CONCLUSION

In this research paper, we have analyzed 2000 tweets as a dataset; these tweets are identified from #productreview + #iphone7 hashtags for 4 weeks. We went through the relevant data mining technique for Knowledge Discovery and Principle Component Analysis (PCA) is used to reduce the dimensionality by focusing on the features which are extracting more meaning. In addition, we have used Hidden Markov Model to consider the hidden states which can be affecting on accurate forecasting. The proposed model is more accurate to predict features of any product than the conventional prediction technologies. This also helps to industries to promote their product based on the reviews so that they could improve sells of the item or they can manage wide publicity to promote the product.

VI. REFERENCES

[1]. Adam Westerski, Rajaraman Kanagasabai, Jiayu Wong, and Henry Chang "Prediction of Enterprise Purchases using Markov models in Procurement Analytics Applications" in 19th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems pp 1357 – 1366

[2]. Ivan JANICIJEVIC, Jovan FILIPOVIC, Jasmina MISCEVIC "USING A MARKOV CHAIN FOR PRODUCT QUALITY IMPROVEMENT SIMULATION" U.P.B. Sci. Bull., Series D, Vol. 76, Iss. 1, 2014 ISSN 1454-2358 pp227-242

[3]. Ka Ching Chan "Market Share Modeling and Forecasting using Markov Chain and Alternative Models" in International Journal of Innovative Computing Information and Control Volume 11 Number 14 August 2015 ISSN 1349-4198 pp1205 - 1218

[4]. Lihong Li, Jie Sun*, Yan Li and Hai Xuan (2014) "Mathematical model based on the product sales market forecast of Markov forecasting and application" in Journal of Chemical and Pharmaceutical Research, 2014, 6(6)p.p. 1359-1365

[5]. Md. Rafiul Hassan (2009), A combination of hidden Markov model and fuzzy model for stock market forecasting Neurocomputing 72 pp3439-3446

[6]. Dr. Mohd. Abdul Raffey, Mr. Pankaj N. Patil, Dr. Sunil Kawale (2017) "Sentimental Analysis of Tweets using Principle Component Analysis Technique" International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 5 Issue XI November 2017 pp 1601-1604

[7]. Samaneh Beheshti-Kashi, Hamid Reza Karimi, Klaus-Dieter Thoben, Michael Lütjenb & Michael Teucke, "A survey on retail sales forecasting and prediction in fashion markets"Systems Science & Control Engineering: An Open Access Journal, 2015 Vol. 3, 154–161

[8]. Wai-Ki CHING, Li-Min LI , Tang LI , Shu-Qin ZHANG (2007) "A New Multivariate Markov Chain Model with Applications to Sales Demand Forecasting" in International Conference on Industrial Engineering and Systems Management IESM 2007.

[9]. https://stackoverflow.com/questions/33543446/what-is-the-formula-of-sentiment-calculation

[10]. http://www.dummies.com/programming/big-data/data-science/how-to-utilize-the-markov-model-in-predictive-analytics/