IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

Using Product Reviews from Twitter for Mining Insights to Predict Sales

Salama Shaikh[1], Prof. L.M.R.J. Lobo[2]

ME (CSE) Student, Department of Computer Science & Engineering, Walchand Institute of Technology, Solapur, Maharashtra, India [1]

Associate Professor, Department of Computer Science & Engineering, Walchand Institute of Technology, Solapur, Maharashtra, India [2]

Abstract: Tweeter, Facebook, Youtube are common terms now that everyone knows and most people are using them. People have social networking accounts. Social networking activities are a part of social life now days. Human beings interact with each other through social media. Most of the companies who provide services or product have their social accounts on social media to endorse their products or services. They often post online about new product or services. People write reviews about product experiences. These reviews are large in number. We can utilize this data to help a vendor to make intelligent business decisions. This paper presents a system used to mine the predictions of a product. Tweeter data was made use of. A tweeter user posts tweets about a product or service. These tweets were used as database and processed using DynamicLMClassifier algorithm. Based upon the classification, tweets are stored into categories. Categories were negative, positive and neutral. This classification was very helpful to take business decisions.

Keywords: Data mining, Business Intelligence, DynamicLMClassifier, Review Mining.

I. INTRODUCTION

Social media is essential term in most of the people's life. Social media has formal as well as casual reviews. These reviews are enormous in number. Twitter is one of the main social networking services where people who are using this service entitled 'Tweeter users' can send and read short 140-character messages called tweets. These tweets are a very large set of data from all over the world to analyze a particular product and review them for taking business decisions. According to Wikipedia as on March 2016, Twitter has more than 310 million monthly active users [1]. Tweeter users are engaging now in exchanging their ideas, thoughts and views about product, person, event, service and many more. Data Mining is one of the processes in knowledge data discovery. It's all about analyzing data and finding useful information [2] Data mining is the extraction of hidden predictive information from large databases, is a influential new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools are very useful and predict future trends and behaviors, allowing businesses to make practical, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems [3] Data mining is important domain now a days. It has many techniques to find the useful information, mainly hidden patterns from large data sets. Business intelligence is thought of as a technique or procedure that analyses a particular products data and helps the vendor to take decisions.

Business intelligence (BI) is set of practices and tools for the acquisition and transformation of raw data into meaningful and useful information for business analysis purposes [4][5]. Business intelligence is a technique or set of techniques to help vendors for improving retailing and performance of their product or service.

On social media like twitter group of humans debate over a particular product, service, sports person, movie or social event in form of tweets. Human beings share and discuss sentiments on this platform. There are enormous numbers of tweets available, so it is arduous for an individual to find out the utility of tweets in this tremendous cluster of available resources. This massive amount of tweets acts as very useful data sources to discover the review about a particular product or service. People give their likeliness to a particular tweet. This can be useful for analyzing particular products likeliness in people who are using it.

In this paper we propose a system that takes the tweet of a particular product, service, sports person, movie or social event for predicting its sales performance based on classification. In this First multiple tweets are collected then these tweets classify into three categories positive, negative and neutral using dynamic lm classifier with high accuracy. Then according to classification heuristic model for predicting the sales by observing historic data is build. So, propose system that classifies tweets, predict its sales, so that it can have

- The system to estimate economic impact of a product based on tweets.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

The rest of the paper is structured as follows: first section 2 discusses related work. Then section 3 proposes system in that the present system is discussed.
In section 4 there are results and experimental setup of system. Final section concludes this paper with discussing some points.

II. RELATED WORK

The invented system of Tuan-Ann Hoang – Vu hvtuananh [6] predicts the sales performance. A word dictionary is maintained and reviews are classified. There are three modules first one for managing word dictionary, second to fetch and managing reviews, third for analyzing reviews, the algorithm used is Jaccard and cosine similarity.

The system of Anindya Ghose and Panagiotis G. Ipeirotis [7] examines the impact of reviews on economic products data. In this system they performed the econometric analysis and constructed the predictive model by using the Random Forest classifiers. The reviewer characteristics were also studied.

The data was collected from Amazon.com of 411 products over a period of 15 month. A study was conducted in movie domain by Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang, and Aijun [8]. Their analysis showed that sentiments and quality of product has considerable impact on sales performance. They proposed Sentiment Probabilistic Latent Semantic Analysis (SPLSA) Autoregressive Sentiment and Quality Aware model (ARSQA).

Twitter reviews are used to do opinion mining by using Support Vector Machine Algorithm. Abd. Samad Hasan Basari, Burairah Hussin, I. Gede Pramudya Ananta and Junta Zeniarja [9] showed Support Vector Machine Algorithm has 70% accuracy but using hybrid Particle Swarm Optimization (PSO) improved the election of best parameter in order to solve the dual optimization problem with improved accuracy 77%.

Dipak Gaikar, Bijith Marakarkandy [10] surveyed movie sales in prediction. They analysed, impact of the positive, negative, strongly positive and strongly negative online reviews of movies on the audience. They have used sampling and regression techniques.

The objective of the research was to understand the effect of valence and volume of online reviews using sentiment analysis on attitude and purchase intention of a product. The dataset used is Twitter API (twitter 4j).

III.METHODOLOGY

This paper presents a system that estimates sales performance of a product or service. The system's overall architecture is shown in figure 1. It shows basic architecture for the system. The system first fetches tweets from Twitter using Twitter API library.

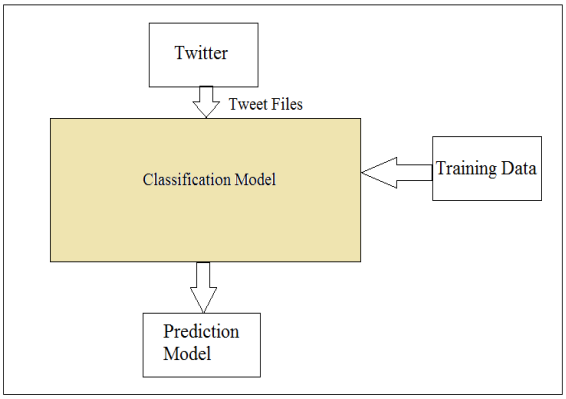

Figure 1 System Architecture

A. Twitter
The system fetches tweets from twitter. In this system we uses Twitter4j library for Twitter API.

B. Classification Model
The classification model classify tweets into three categories (positive, negative and neutral), supervised learning algorithm 'Dynamic LM Classifier' is used. The algorithm works on semantic model. This algorithm implements training and classification, it may be used in tag-a-little, learn-a-little supervised learning without retraining epochs. This makes it ideal for active learning applications.[11]

C. Training Data
Training data is number of tweet files that are going to be used by dynamic lm classifier for training. Training data is provided in tweet files and classified in to the negative, positive & neutral.

D. Prediction Model
This system predicts the sales performance of the particular product for which classification is done.
The Positive to negative tweets ratio is considered for predicting the sales performance.
This prediction model is a heuristic model. It predicts sales performance by observing historic data.

IV.EXPERIMENTAL SETUP AND RESULTS

The system has the flow as outlined in figure 2. The developer account is needed to use twitter data. By using twitter4j we are fetching tweets of a particular product from Tweeter. Then stream of tweets is given to Dynamic LM classifier, the algorithm uses training data to classify tweets in to the positive, negative and neutral. A Dynamic LM Classifier is a language model classifier. Training is based on a multivariate estimator for the category distribution and dynamic language models for the per-category character sequence estimators [12] Tweets are classifying as positive, negative and neutral. Prediction model uses heuristic data based upon observing some historic data. Results of classification are in the form of negative, positive and neutral.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

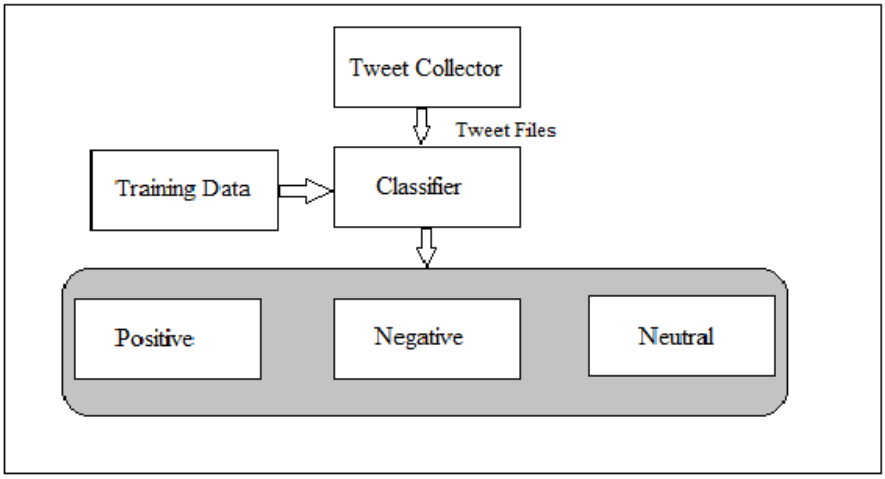

Figure 2 Flow of the System

Sr No.	Manufacturer Goods	Training Data		Testing Data		Classification Results Time in Minutes
		Count of Tweets	Percentage	Count of Tweets	Percentage	
1	Iphone	450	0.95%	47348	99.05%	2.08
2	Windows 10	450	0.91%	49231	99.09%	2.12
3	Samsung Phone	450	1%	44982	99%	2.15

Table 1- Result Table

Table 1 shows the final results of the system, here we are taking three sample products namely Iphone, Windows and Samsung. In the first column there are training tweets, second column contains the product tweets fetched from twitter. Third column contains the time that have taken by system to generate result.

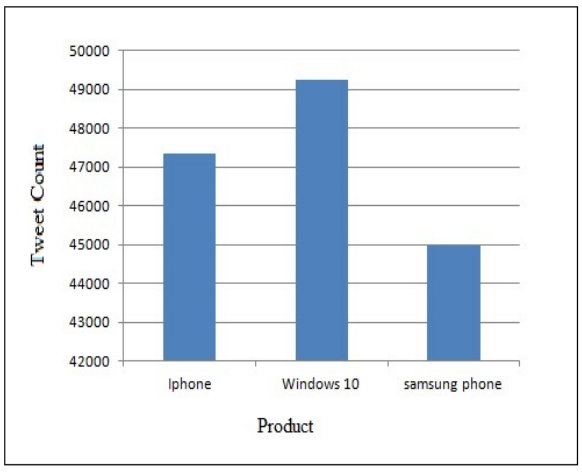

Figure 3 Total Product Tweets Count

Figure 3 represents Product Tweets Count. System fetches number of tweets with respect to product. Here Iphone, Windows and Samsung phone products fetches 47348, 49231 and 44982 tweets respectively.

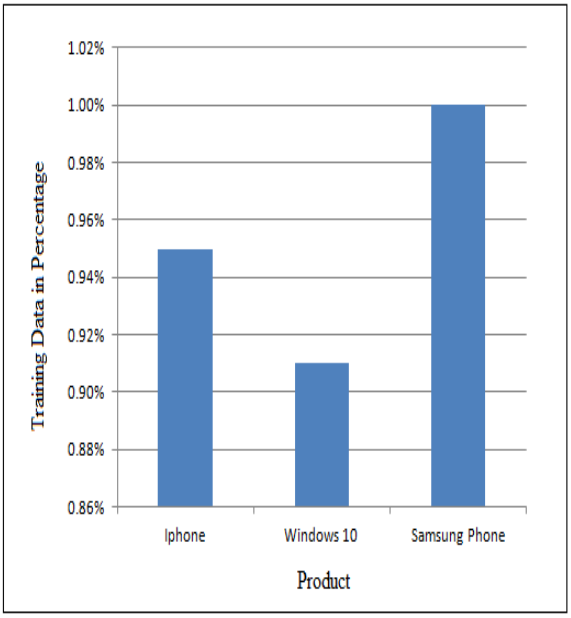

Figure 4 Traing Data wrt Products

Figure 4 shows the graph of training data we provided to products.

Training Data is essential part of the Dynamic LM Classifier as it learns pattern from it to efficiently classify tweets. Training data is provided manually as training tweets.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

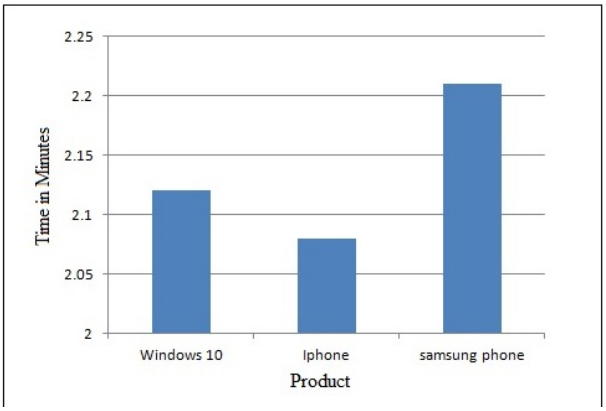

Figure 5 Results Time in Minutes

As outline in Figure 5.The graph of time required by system to generate result. The first bar is for windows , system takes 2.12 minutes, for Iphone 2.08 minutes and for Samsung phones it takes 2.15 minutes.

V. CONCLUSION

The designed a dynamic system which takes historic as well as real time data. The Dynamic LM Classifier Algorithm used for classifying tweets which is of supervised learning method. The system is useful to vendors of all kinds of manufacture goods and services. Our classification model is beneficial to know the current trend of product in twitter, while knowing the sales performance vendor may take necessary steps afterwards.
The system summons the product related tweets using Twitter4j and stores tweets. Classification algorithm Dynamic LM Classifier classifies tweets into positive, negative and neutral categories.

REFERENCES

[1] "TwitterWikipedia","https://en.wikipedia.org/wiki/Twitter
[2] "Data Mining: What is Data Mining?", http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[3] "An Introduction to Data Mining", http://www.thearling.com/text/dmwhite/dmwhite.htm
[4] "Business intelligence – Wikipedia", https://en.wikipedia.org/wiki/Business_intelligence#References
[5] Turner, Dawn M. "What is Venture Management?", www.VentureSkies.com. VentureSkies. Retrieved 24 February 2016
[6] Anindya Ghose and Panagiotis G. Ipeirotis "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics" Ieee Transactions On Knowlegde And Data Engineering, VOL. 23, NO. 10, October 2011
[7] Mohsin Nadaf, Akshay Deshpande, Sneha Tirth "Using Business Intelligence for Mining Online Reviews for Predicting Sales Performance" IJECS Volume 4 Issue 5 May, 2015 {https://www.ijecs.in/issue/v4-i5/1%20ijecs.pdf}
[8] Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang, and AijunAn "Mining Online Reviews For Predicting Sales Performance: A case Study in the Movie Domain" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 4, APRIL2012{http://www.computer.org/csdl/trans/tk/2012/04/ttk2012040720-abs.html}
[9] Abd. Samad Hasan Basari, Burairah Hussin, I. Gede Pramudya Ananta, Junta Zeniarja, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization" ,Malaysian Technical Universities Conference on Engineering Technology 2012, MUCET 2012 Part 4 - Information And Communication Technology, Procedia Engineering 53 (2013) 453 -462
[10] Dipak Gaikar, Bijith Marakarkandy, "Product Sales Prediction Based on Sentiment Analysis Using Twitter Data", www.ijcsit.com,(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2303-2313
[11] http://aliasi.com/lingpipe/docs/api/com/aliasi/classify/DynamicLMClassifier.html
[12] http://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm
[13] "J48 – Wikipedia" https://en.wikipedia.org/wiki/J48
[14] https://github.com/hvtuananh/lingpipe/blob/master/src/com/aliasi/classify/DynamicLMClassifier.java

BIOGRAPHIES

Salama Hasan Shaikh is pursuing her Master Degree of Computer Science & Engineering from Walchand Institute of Technology, Solapur, and Maharashtra, India. She has received Bachelor degree in Computer Engineering from Pune University. Her research interests include Data Mining, Big Data & Business Intelligence.

Prof. L.M.R.J. Lobo is an Associate Professor in Department of Computer Science and Engineering & HOD in Department of Information Technology, Walchand Institute of Technology, Solapur, and Maharashtra, India. He is pursuing his Ph.D in Genetic Algorithm based approach to Data Mining. He has twenty-six years of teaching experience. His areas of interest include Data mining, Genetic Algorithm & Artificial Intelligence.