JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

VENDOR SELECTION THROUGH REVIEW ANALYSIS

[1]Aliya Abdullah, [2]Anushka Bagde, [3]Aadarsh Kapoor, Dr. Varsha Shah

[1]Designation of 1st Author, [2]Designation of 2nd Author, [3]Designation of 3rd Author
Department of Computer Engineering,
[1]Rizvi College of Engineering, Mumbai, India

Abstract : With the dawn of shopping websites it is much easier for small businesses and individual consumers to buy goods. These websites also have reviews and ratings of the available products. Many people also like to use blogging websites like twitter to give feedback on products. But with the wide range of products that are now available, the decision-making process has become much more difficult. This is why we propose a vendor selection tool that can be used by small businesses or consumers trying to find the perfect product. It aggregates reviews and ratings from e-shops as well as the sentiment people have expressed towards the products on blogging websites such as Twitter. It uses web scraping technique to extract reviews and ratings of the people given under each product. Python's beautifulsoup library is used to scrape reviews and store them in a .csv file. Contents of this .csv file can be accessed and displayed as per requirement. It also uses sentiment analysis technique to extract people's opinions on the microblogging website Twitter. A Support Vector Machine is trained to analyse tweets that have been manually labelled. Then, this SVM is used to classify product sentiment.

***IndexTerms* - Web scraping, Sentiment analysis, Twitter, Amazon, Support Vector Machine.**

I. INTRODUCTION

Big businesses have a structured vendor selection model that they use. They define business requirements, vet current third-party vendors, conduct briefings, schedule demos and complete the vendor selection. A lot of time and investment is put into this process. Obviously, a company needs to have the resources to conduct vendor selection on a large scale.

However, things get much more difficult for small businesses. They do not have sufficient financial resources to conduct extensive research in selecting vendor. Moreover, these businesses are single-handedly managed, or with very few employees. So, it becomes imperative to have an easier and faster vendor selection process. Therefore, there is a need for a system where small businesses can select products.

To vet the different vendors and products, they have to rely on either word-of-mouth or the internet. A person is more likely to know only a few people who have used the product, so the former is not as reliable. The internet, however provides a large sample size with the hundreds of websites with thousands of reviews on it.

Information can be retrieved from the internet in various formats. The most obvious ones are ratings and reviews under the product website or any e-shop. But there are many more sources of information. People regularly give their opinion on products through any personal blogs they may have. They may post their opinions on social media. News regarding the product may also turn out to be a very valuable source of information. As evident, the information is endless. It is not possible for a single person to go through all of this information manually.

This is the motivation behind our proposed model "Vendor Selection through Review Analysis". Currently, there is no one website or article one can read about a product and hope to receive maximum information. The proposed model takes the name of the desired product from the user. It then attempts to gather all information relating to its quality in one place. It accomplishes this by mainly two methods, review scraping and twitter sentiment analysis. It collects reviews and ratings of the products from e-commerce websites. It collects various tweets from the microblogging website, Twitter. It then processes these tweets to extract sentiment information. The results of these two methods are displayed to the user, so that he may make a more well-informed decision about whether or not to buy the product.

II. RELATED WORK

There are several methods available today to perform sentiment classification, and thus it is important to carefully select appropriate methodology. One of the earliest works on sentiment analysis was conducted by Pang et al. [1]. They used movie reviews as a dataset and attempted to find out the best machine learning model (Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM)) suitable for classifying either positive or negative sentiment. To handle negations in word, they added a NOT_ tag to every word between a negation word. The most successful model was a unigram SVM model with an a three-fold cross-validation accuracy of 89.2%.

Sentiment classification specifically on Twitter data was conducted by Go et al. [2]. They classify tweets as either positive or negative, built a machine-learning classifier, and classified positive and negative tweets using SVM, NB and MaxEnt classifiers. They achieved maximum accuracy using Multinomial NB using mutual information.

Classification on financial data from microblogging platform StockTwits was performed by Renault [3] to evaluate the performance of pre-processing methods and machine learning models. For number of n-grams used, adding bigrams improves the accuracy of the classification by 2.2 % compared to a processing with unigrams only. Trigrams and four-gram have no significant impact on the accuracy. For pre-processing, including emojis and punctuation increases classification precision by 0.38% and 0.30 % respectively. Whereas, removing stop words, POS tagging and stemming decrease accuracy. This is due to the nature of stock tweets, where stop words like "up", "down", carry significant sentiment information. Classification was done using Naïve Bayes algorithm (NB), Maximum entropy classifier (MaxEnt), a linear Support Vector Classifier (SVC), a Random Forest classifier (RF) and a MultiLayer Perceptron classifier (MLP). It was found that more complex algorithms (RF and MLP) do not improve the precision of the classification compared to more simple methods such as ME or SVM and have higher time complexity.

Ruz et al.[4] attempt to find out sentiments during the 2010 Chilean earthquake and the 2017 Catalan independence
Referendum through tweets. For pre-processing, they removed URLs, hashtags, targets, punctuations, symbols, numbers, repeated characters, and converted tweets to lower case. They used, bag-of-words (BOW) technique to convert training tweets into a numeric representation resulting in a term document matrix (TDM). For dataset 1, the best performance was obtained by the SVM classifier followed by the BF Tree Augmented Naive Bayes (TAN) model with 19 edges. For dataset 2, RF obtained the best accuracy, and the TAN model showed better results (compared to RF and SVM) by passing the 80% accuracy threshold.

Back et al. [5] propose a system to extract sentiment data from Social Network Services (SNS) big data. This data is vastly unstructured. They use two models, the Naïve Bayes algorithm and natural language processing (NLP). Experimental results showed that NB algorithm gave a 63.5% accuracy, much lower than NLP method. However, based on data processing speed analysis, the NB algorithm's processing speed was approximately 5.4 times higher than NLP method's.

Zvonarev et al. [6] Use tweets in the Russian Language as a dataset. They pre-processed tweets by lemmatization, punctuation removal, capitalisation and stop-word removal. To handle negation, they concatenated "not" with the following word. They used three models, logistic regression, XGboost and Convolution Neural Networks. While logistic regression produced an accuracy of 76.7% in 45.2 seconds, CNN produced an accuracy of 79.5% in 6 hours and 11 minutes. It is quite clear that CNN performs better. However, training such a model needs much more time than LR. Hence, depending on available time and computing power for modelling LR may be preferred.

The digital world is growing at a pace that exceeds the speed of any man made fastest prime movers. Here the term growing is used in context to the size of data. Farley, E.J., & Pierotte L [7] have explained that at 487bn gigabytes (GB), if the world's rapidly expanding digital content were printed and bound into books it would form a stack that would stretch from Earth to Pluto 10 times. The main contributors to this digital warehouse are social media, government surveillance cameras and plenty of other independent websites which are updated on daily basis such as inventories system of companies, their daily revenues as well as E-Commerce websites that come up with FMCG's on daily basis. In this digital age, this web data is the most essential resource for any business. The main focus of this paper is to highlight the collection of data through scraping as API's are not available for each and every data source. Web Monitoring, Scraping and digital forensic is one of the prominent areas in the domain of Big Data and Sentiment Analysis. A number of software products and tools are available in the technology market which are used to guard the network infrastructure and confidential data against cyber threats and attacks. For a long time, the monitoring of servers and forensic analysis of network infrastructure has been done using packet capturing (PCAP) tools. These activities are performed using PCAP and related tools available in the market which includes open source software as well as commercial products. As far as the fame and usage of the software suites is concerned, the open source market is getting popular because of the scope of customization and organization specific personalization of the software products. In this research paper, an approach is depicted for the fetching and analysis of live data from social media portals and using such approaches the sentiment data analysis can be implemented effectively.

P. Ashiwal et al. [8] said website creators also use the web scraping where collecting data from the different social media websites, what is trending and what is in etc. Web scraping is used in one of the cases in which it is used to scrape the content of a particular category of book in the Amazon store. In another case web scraping is used to scrape the contents from Twitter on the basis of hash tags or by searching the keywords in the twitter. In the field of machine learning web scraping is used in sentiment analysis, where the data is scrapped from the websites.
Sameer Padghan et. al., [9] projected an approach where data extraction is done from web pages in assistance with web scraping easily. This method would enable the data to be scrapped from numerous websites that will minimize human intervention, save time and also enhance the quality of data relevance.
Anand Saurkar et. al., [10] discovered the latest technique named Web Scraping. Web scraping is a quite important methodology used to produce structured data based on the unstructured data available on the internet. Scraping formed structured data, subsequently collected and evaluated in spreadsheets in the central database. This research focuses on a summary of the data extraction process of web scraping, various web scraping strategies and most of the latest tools utilized to scrape the web.
Federico Polidoro et. al., [11] concentrated on the outcomes of web scraping evaluation strategies with particular orientation to user electronics services and goods throughout the sector of commodity price studies. Although the research done has so far been

performed in a small amount of time, that you can see in whatever followed, it has enabled to attain important, but not conclusive, novel efficiencies results.

Web scraping is used in technologies such as Market research using web data in any of the industries. Even web scraping technology is used in price comparing sites where it compares the price of an item or room from different websites. In advance these applications use the web scraping to scrape the content from the dependent websites. Various government and private watch dogs use the web scraping to monitor the malicious activities going on the internet. Shreesha M et al. [12].

While every implementation follows these three key steps, how they achieve each step can differ, altering the scope and ability of the web scraper. The following are a few techniques used in implementing a Web Scraper:

- HTML Parsing
- DOM Parsing
- Text Pattern Matching.

Miller [13] infers that each technique has its strengths and reasons for use. HTML parsing can quickly extract basic information from a site. DOM parsing uses embedded browsers to parse the Document Object Model for a page, giving greater insight to the page and any generated content from client-side scripts such as JS.

III. Research Methodology

3.1 Sentiment Analysis

3.1.1 Data Collection

The dataset used for training the model is Sentiment140, made by Alec Go, Richa Bhayani, and Lei Huang. The data was collected automatically by assuming that any tweet with positive emoticons, like :), were positive, and tweets with negative emoticons, like :(, were negative. The Twitter Search API was used to collect these tweets by using keyword search. Tweets with positive sentiment are numbered 4, with negative sentiment are numbered 0, and neutral tweets are numbered 2.

The same Twitter Search API is used to collect tweets about the vendor, and sentiment analysis is performed on that.

3.1.2 Data Pre-processing

Due to the informal nature of tweets, they may consist of a lot of inconsistencies. Thus is is important to pre-process the data before passing it on to the model. The following preprocessing is done:

1) Stop word removal: The NLTK (Natural Language Toolkit) English stop word corpus is used to remove all stop words like "I", "me", "our", "we", etc.
2) Capitalization: The tweets are converted to lowercase.
3) Removal of mentions: Any users mentioned / tagged in the tweet are removed.

3.1.3 Feature Extraction

We have coded the text as Bag of Words and applied an SVM model.

After pre-processing, we have to extract features from the tweets. In our proposed model, we use the Bag-Of-Words (BOW) method using bigrams to handle negation. This is because "good" and "not good" have opposite sentiments and the model should be able to handle this. The vocabulary consists of all words in the training tweets. This method represents whether a word is present in a word or not using a vector. We use Sci-Kit Learn's CountVectoriser library to achieve this.

3.1.4 Sentiment Analysis Model

The Sentiment Analysis model we have selected is the Support Vector Machine. As evident from our literature survey, more complex models like convolution neural networks have lengthy processing times and other machine learning models like Naïve Bayes algorithms don't give accuracy as high as SVMs. Therefore, SVM is the model selected. We use Sci-Kit Learn's SVC class to implement the model. We use cross validation and grid search to find good hyperparameters for our SVM model.

The model classifies the data into three classes, namely positive, neutral, and negative. This can be use to calculate the overall user opinion about any particular product.

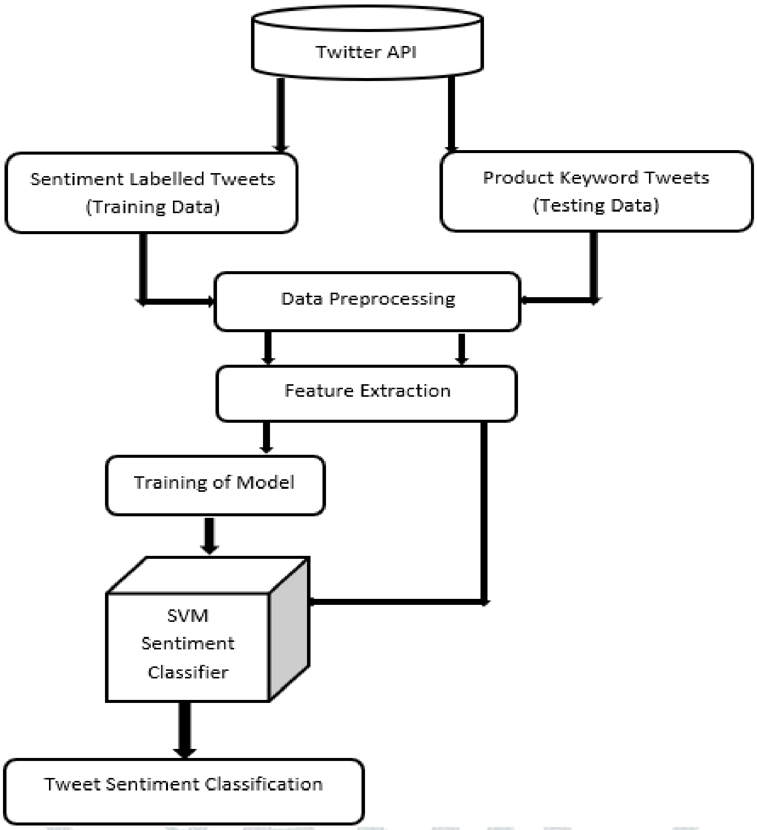
Twitter API
Sentiment Labelled Tweets (Training Data)
Product Keyword Tweets (Testing Data)
Data Preprocessing
Feature Extraction
Training of Model
SVM Sentiment Classifier
Tweet Sentiment Classification

Fig 1. Sentiment Analysis

3.2 Review Scraping

3.2.1 Data Extraction

The data will be directly extracted from the website on the selected product. Python's Beautiful Soup library functions will be used to read the HTML document and find the specified data using the tags they are displayed under in the HTML page. We use Beautiful Soup's find_All function to extract information displayed under the specified tag for example: finding the reviewing customer's name under the <p> tag having class and id=name. The find_All function returns all customer names separated by commas. We then store this extracted data into a csv (Comma Separated Values) file which can be accessed by the backend of the website we create to display reviews.

3.2.2 Data Pre-Processing:

The extracted data may contain reviews from anonymous customers. For such customers, instead of having a null name, we replace them with anonymous in our data. We take 5-star, 4-star, 3-star rating percentages under 'good' ratings and 2-star, 1-star rating as bad rating percentages. This will be done be Beautiful Soup using split function.

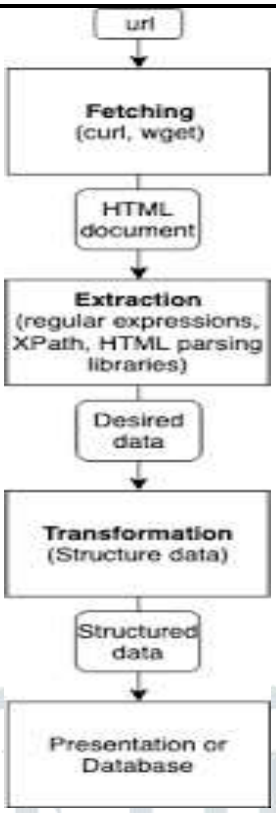

Fig 2. Review Scrapping

IV. Conclusion

Here, we have proposed a working model of Vendor Selection using Sentiment Analysis and web scraping techniques which can help small businesses find trusted vendors for buying their products. The proposed model on implementation will suggest vendors based on the reviews and reactions left by tried and tested customers.

V. REFERNCES

[1] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques." *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. 2002.

[2] Go, Alec, Lei Huang, and Richa Bhayani. "Twitter sentiment analysis." *Entropy* 17 (2009): 252.

[3] Renault, Thomas. "Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages." *Digital Finance* 2.1 (2020): 1-13.

[4] Ruz, Gonzalo A., Pablo A. Henríquez, and Aldo Mascareño. "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers." *Future Generation Computer Systems* 106 (2020): 92-104.

[5] Back, Bong-Hyun, and Il-Kyu Ha. "Comparison of sentiment analysis from large Twitter datasets by Naïve Bayes and natural language processing methods." *Journal of information and communication convergence engineering* 17.4 (2019): 239-245.

[6] Zvonarev, Andrey, and A. Bilyi. "A comparison of machine learning methods of sentiment analysis based on Russian language Twitter data." *Proceedings of the 11th Majorov International Conference on Software Engineering and Computer Systems (MICSECS)*. 2019.

[7] Farley, E.J., & Pierotte L. (December 2017). *Web Scraping: An Emerging Data Collection Method for Criminal Justice Researchers.* Washington, DC: Justice Research and Statistics Association.

[8] P. Ashiwal, P. Tripathi and R. Miri, *Web Information Retrieval Using Python and BeautifulSoup*, vol. 4, no. Vi, pp. 335-339, 2016.

[9] Anand V. Saurkar, Kedar G. Pathare and Shweta A. Gode, "An Overview On Web Scraping Techniques And Tools," International Journal on Future Revolution in Computer Science & Communication Engineering, pp. 363-367, Vol. 4, 2018.

[10] Sameer Padghan, Satish Chigle and Rahul Handoo, "Web Scraping-Data Extraction Using Java Application and Visual Basics Macros," Journal of Advances and Scholarly Researches in Allied Education, pp. 691-695, Vol.15, 2018.

[11] Federico Polidoro, Riccardo Giannini, Rosanna Lo Conte, Stefano Mosca and Francesca Rossetti, "Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation," Statistical Journal of the IAOS, pp. 165-176, 2015.

[12] Shreesha, M., S. B. Srikara, and R. Manjesh. "A Novel Approach for News Extraction Using Webscraping." *3rd National Conference on Image Processing, Computing, Communication, Networking and Data Analytics*. 2018.

[13] Miller, Jared. "Comparative Study of Web Scraping Tools and Implementations."