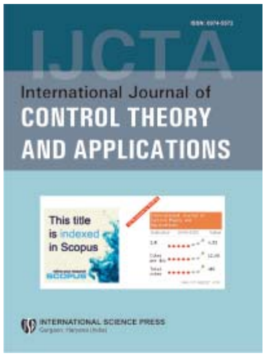

International Journal of Control Theory and Applications

ISSN : 0974–5572

Volume 9 • Number 40 • 2016

Integration of Predictive Analysis Andsentimental Analysis Onabook Forum

SwayamshreePradhan[a] and SrinivasLNB[b]

[a]*M.Tech (Big Data Analytics) Department of Information Technology SRM, University, Kattankulathur, Chennai*

E-mail: swayamshreepradhan1992@gmail.com

[b]*Asst. Professor, Department of Information, Technology, SRM University, Kattankulathur, Chennai*

Abstract: Predictive analytics includes an assortment of statistical systems from machinelearning,data mining and predictive modeling that break down present and authentictruths to make prediction about future or generally obscure eventsthatanalyze current andhistorical facts to makepredictions about futureorunknownevents. It is an ever-growing technology that empowersorganizations to utilize information in both stored and real-timedata, moving from a historical perspective to forward looking predictions. This concept is used in many domains such asonline social media, product design, relationship management etc. As social media speaks to a major approach of how data isbeing created, transferred and consumed, the data can be captured in variousbehavioral cases to estimate the response of the structure under exploration.In thispaper, we are presenting a framework of the predictive model by using social media data.By combining feature selection mechanism and sentimental analysis, we are showingprediction through a variety of interactive visualizations. Keeping in mind the endgoal is to assess how predictions might be performed on online social media data. Here we arepresenting results froma book forumknown asGoodreadsasan indicator for such analysis by using machine learning and sentimental analysis.

Keywords: *Social media analytics, predictive analysis, sentimentalanalysis, visualization, machinelearning.*

1. INTRODUCTION

Inspite of the fact that web-based social media information is unstructured in nature, it gives many chances to scientists for exploring the things. Organizations are relying upon social media to get the colossal measure of data for promoting their product design, marketing and relationship management. The strategies have been produced to catch online talks for model building. Presently the visual analytics group has begun investigating web-based social media information by using different kind of tools. The data is increasing at a quicker rate. The group is concentrating on social media analytics for building up the structures by collecting,monitoring,analyzingandvisualizingthe huge amount of incredible socialmediadata. The visual analytics team can be a combination of information extraction, data analysis andrepresentation of human behavior. This examination has gone from the data extraction to mining of customer's sentiment. Building prescient models, some work has been done already using online social media information. In 2013, VAST(Visual Analytics Science and Technology) manufactured prescient model for the opening end of the week gross of motion pictures and through that, they ran the VAST film industry challenge[2][5]. In 2014, Goodreads piece more particular related work has been directed. They introduced a system for online social media integration and prediction in which it comprised of tools for extracting, examining and demonstrating patterns crosswise over different web-based social media analytics stages. They extracted and coordinated information from Twitter, YouTube and IMDB (internet motion picture information base) [3].In this paper, we are introducing a system in which we are extracting data from Goodreads. This stage comprises of tools for removing, breaking down and demonstrating the feelings over the different web-based social media platforms. This framework takes information from GoodreadsandTwittercharacterizesthe books on the basis of reviews and ratings.

2. BACKGROUND AND RELATED WORK

The goal of the paper is to demonstrate a complete thesis, from data extraction to machine learning analysis and to recognize a few of the dead ends. I am looking at there views and ratings of 15 popular books.

2.1. Seb scraping

Web scraping is utilized for extracting the information from different sites. To scrape data from websites, we identify the DOMs (Document object model) from which data is to be extracted and then using programming tools automate the process on multiple web pages.This data is then stored on a local PC or a databases so that it can be easily accessed.Pre-existing web scrapingsoftwareusually simplifies this task based on the user's requirements. It is either custom worked for a particular site or is one which can be designed to work with any website. Here we are scraping the Goodreads reviews which are in text format with a non-text variable attached that is left by the reviewers. For this procedure we are utilizing the RSelenium package by starting an RSelenium server. For extracting the text which we require in our project, we looped through 100 pages or so of reviews for each of the books. At first we need to identify theplace of occurrence of the reviews appear in the page code. This is done by utilizing Selector Gadget, a Chrome extension that permits to recognize CSS selectors. When we have identified the correct CSS selector [18], we simply pass its name to RSelenium server. We for the most part extracted the html code for the reviews and after that we will get the text format which we want for further analysis [19].

2.2. Sentimental analysis on socialmedia data

Sentimental analysis is more on opinion mining. When we perform sentiment analysis on some content, we basically look for the opinions in that content and pick the sentiment withinthe opinions. An opinion is an expression consisting of two components known as a target and a sentiment on that target. It is the procedure of computationally distinguishing and sorting opinions which is communicated in text format, particularly to decide the author's disposition towards a specific point, item, that is whether it is surei.e. positive, negative, or nonpartisani.e. neutral. It passes on the author's opinion in a post .Social media is a brilliant wellspring of data. In the time of the Social Web, client contributed substance have turned into the standard. The measures of information created by people, business, government, and research specialists have been experiencing a development.2013, Facebook and Twitter had 1.23 and 0.64 billion dynamic clients [12], separately. The quantity of fellowship edges of Facebook is evaluated to be more than 100 billion. The surge of measures of client contributed substance, for example, online customer reviews, online news, personal dialogs,searchquerieshave required the innovative work of another era of investigation strategies and devices to adequately prepare them, ideally progressively or close ongoing[14][15]. Big data is regularly portrayed by three measurements named the 3 V's: Volume, Velocity, and Variety Social Media Analytics empowers associations to use online networking information for business bits of knowledge and measure the impact of web-based social networking activities[4]. Different advantages of social media analytics are –to recognize and deliver customer satisfaction to hold customer belief, understand new prerequisites and planned customers, and get client input for proactive basic leadership. Different phases of web-based social media examination are: information extraction and collect the data, estimation - showcase conclusion examination of items, social media integration and representation - for better bits of knowledge, imaginative and vital business choices. Different procedures utilized as a part of web-based social networking examination are clustering, classification, information extraction, and correspondence visualizations. Social media Analytics empowers associations to use web-based social networking information for business bits of knowledge and measure the impact of web-based social networking activities [13]. Different advantages of online networking investigation are - distinguish and deliver client worries to hold client reliability to items, see new prerequisites and imminent clients, address rivalry and acquire client criticism for proactive basic leadership. Different phases of social media analytics are: information accumulation, estimation – showcase notion examination of items, data analysis and perception through proper visualizations for better experiences, inventive and key business choices.

2.3. Predictiveanalysis

Predictive analytics incorporates an assortment of statistical techniques from data mining, predictive modeling and machinelearning,which analyze current and historical facts to make predictions. In business, predictivemodels exploitdesigns found in chronicled and value-based information to distinguish opportunities andrisks to permit evaluation of hazard or potential related with a specific arrangement of conditions. The defining functional impact of these specialized methodologies is thatpredictive analytics provides apredictive score (probability) for every person (customer, employee, healthcarepatient,product SKU,vehicle, component, machine,or other organizational unit) to decide, illuminate, or impact hierarchical procedures that relate crosswise over extensive quantities of persons, Note that our proposed framework is not the first to address the determined prediction. A large number of tools exist for both customers and experts (*e.g.*,R Shiny). Thesesoftware packagesandtoolsprovideavarietyofmachinelearning algorithmsthat can be utilized for predictiveanalytics tasks,such as feature selection, parameteroptimization and result validation.A number of these frameworks offer fundamental visualizations including residualplots, scatterplotsand linecharts. However,mostof these visualizationsare just used to show the final outcomes rather, these frameworks frequently select to show benchmark models or straightforward factual measures for result approval, functioning as a greater amount of a black-box system. The goal of our framework istoo straightforwardly incorporate the analystinto themodelbuilding loopby enabling feature selection formodelbuilding and comparison. Weinclude toolssuch ascorrelationrankings and Parallel Coordinate Plotsforquick comparison [1] [17]. In addition, we have additionally made an assortment of tools for consequently recommending comparative examples inside a dataset to empower the examiner to distinguish anomalies and approve models in view of the exactness of expectation with respect to comparative cases. As of late, specialists in the visual examination group have been creating strategies for enhancing the model building and prescient analysis. We concentrate on how much data and control ought to be interested in the client. Most firmly identified with our work is what built up an intuitive visual structure for choosing subset components to enhance relapse models.

2.4. Machine learning

Machine learningis a procedure that is proportionate to datamining.This is aprocess in which data looks for patterns. Machine learning uses the information to identify designs in information and change program activities in like manner. Machine learning calculations are either regulated or unsupervised. In supervised learning algorithm, we have information and yield factors and we utilize calculation to take in the mapping procedure from contribution to results. Unsupervised learning is where we have input data and no corresponding output variable.

2.5. Proposed methodology

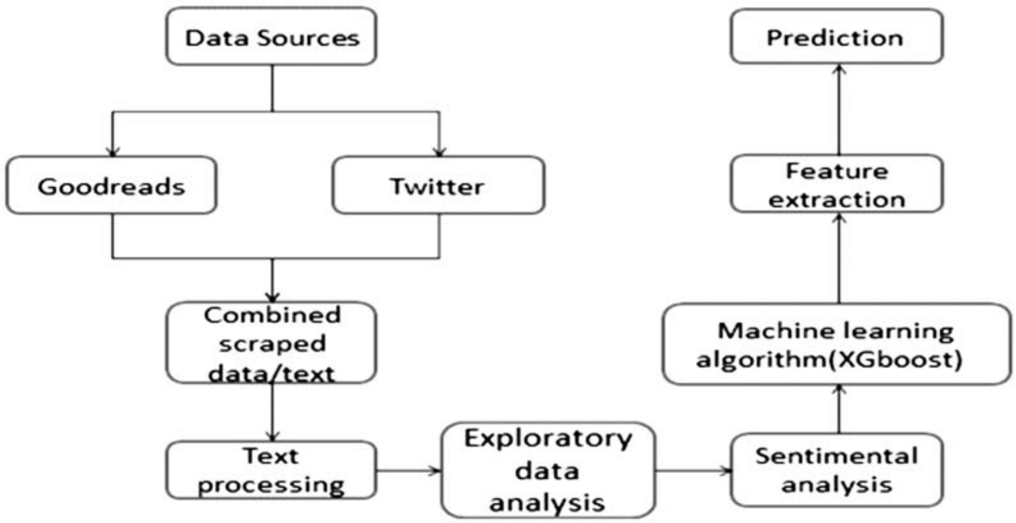

Figure 1

3. DATA DESCRIPTION

3.1. Goodreads

Goodreads permits itsusers to searchthrough a database of hundreds and thousands of books and their reviews. Users can register on the website to access the reading lists created by the strong community on the website as well as create their own suggestion lists.The website, owned by Amazon was launched in January 2007 and has garnered a total of 55 million site users to data. For our study, we gathered information for 15 books along with their reviews and ratings.

3.2. Twitter

We collected tweets for 15 books and combined it with Goodreads reviews and did the analysis on the basis of those combined data.

3.3. Framework for the analytics

Our system concentrates on multisource information from Twitter and Goodreads. We combine web scraping, exploratory data analysis and machine learning. We join trend analytics, sentimental analysis and similarity metrics and feature selection for model building, evaluation and forecast/prediction. For evaluation of this system,we are joining data from Goodreads and Twitter and attempting to analysisthis data over an assortment of different visual analytics platform.The framework was built under JSON, R and Shiny. Preprocessing has been accomplished for exploratory analysis and estimation and word frequency numbers and about intuitive rates are gotten for visualizing the information/data. The machine learning part is used to describe the feature selection. We perform the machine learning thing by using the XGBOOST algorithm.

4. ALGORITHIM USED FOR MACHINE LEARNING

4.1. Machine learning

Machine learning assumes a key part in an extensive variety of basic applications, for example, data mining, natural language processing, and imageprocessing [9]. The program is “trained” on a pre-characterized set of “training cases”, which then encourage its capacity to achieve a precise conclusion when given new information *i.e*. called as managed machine learning or supervised machine learning. The program is given a cluster of information and must discover patterns and connections in it. This process isknown as unsupervised machine learning. In this project weare using amachine learning algorithm (extreme gradient boosting) for extracting the features from the reviews [8].

4.2. Xtreame gradient boosting

Xtreame gradient boosting *i.e*. XGBoost algorithm is an application or implementation of gradient boosted decision trees designed for speed and performance. It more often than not pushes the limit of the computations resources for boosted tree algorithms. That is the reason in many approaches, this algorithm is used. It is an implementation of gradient boosting machine. It is an open sourced instrument computed in C++, Rand python. It is a tree based algorithmwhich is easy to use and is more efficient than other tree based algorithms and can keep running on multiple clusters in parallel.The accuracy of this algorithm is better than other machine learning algorithms. The implementation of the algorithm was engineered for an efficiency of computing time and memory resources. A design objective was to make the best use of available resources to train the model. Some key algorithm implementation features include: Sparse Aware implementation with automatic handling of missing information values. Block Structure to support the parallelization of tree construction. Continued Training with the goal that you can further lift an already fitted model on new information. This algorithm goes by loads of different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines. The XGBoost library implements the gradient boosting decision tree algorithm. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the misfortune when adding new models. This approach supports both regression and classification predictive modeling problems.

5. IMPLEMENTATION AND EXPECTED RESULTS

5.1. Overview

The purpose of this project is to demonstrate a complete example, going from data collection to machine learning analysis and to show the sentimental analysis on them. We are looking at the reviews for fifteen popular books. I have chosen books in the different genres in order to make the comments more heterogeneous.These fifteen books are popular enough that I could easily pull a few thousands reviews fo reach, yielding a significant corpus with minimum effort. The final dataset consisted of 30 reviews, ratings and reviewer names per book.

5.2. Web scraping

The reviews of Good reads are in text format and ratings are in non-text variable which is attached to there views. The ratings are left by there viewers. We scrapped all these by the use of the RSelenium package.We worked on a book by book basis. We had to change the name of the book manually in the script and ran the code for each book. We connect Firefox through the RSelenium server which navigates to the URL of the book if we want to interact directly with the browser. For extracting the content that we need, we will be looping through pages and comments of each book. First we will identify where the reviews appears in the script. We do this by using “Select or Gadget” which is a chrome extension. After getting there views in the text format,wedo thecleaning processby the regularexpressions.Weidentified thevariousexpressionsthatcanbe appeared between the reviewer’s name and therating, and put the min a regular expression to determine the ending point of the name, then we try to extract the name. After doing these processes we put our clean data into a data frame or a tabular format. That will be looking like:

	book	reviewer	rating	review
1	Little_Wo	Susan	it was ama	Someone I know claimed this no longer has value that she would never recommend it because it s saccharine has a religious agenda and sends a bad
2	Little_Wo	Barry Pier	liked it	Okay I m just gonna say this I liked Little Women I m an -year-old guy and I liked Little Women What It s quaint It s quaint as fuck I m such a Jo
3	Little_Wo	Fabian	really like	Yes yes I m a grown ass man reading this but I m not ashamed I also read the Twilight sa-ha ha -ga a bunch of Charlaine Harris as well remember Som
4	Little_Wo	Rory	did not lik	This review has been hidden because it contains spoilers To view it click here I hated this book I can t even begin to go into all the reasons I dislike t
5	Little_Wo	Corrie	it was ama	The book begins Christmas won t be Christmas without any presents grumbled Jo lying on the rug It s so dreadful to be poor sighed Meg looking dow
6	Little_Wo	Dottie	it was ama	My copy of this is probably years old -- I ve probably read it at least twenty-five times One of my all-time favorite books One of my favorite authors
7	Little_Wo	AMEERA	it was ama	i can tell this become my favorite classic book besides all classics books of the queen of classics books Jane Austin and u can see a lot of classic word
8	Little_Wo	Annalisa	liked it	I m definitely a victim of modern society when I find this book slow Had I read it in its day or even as a youth it would probably be fantastic but as it i
9	Little_Wo	Huda Yah	liked it	https www youtube com watch v rFkXV https www youtube com watch v rFkXV - D more
10	Little_Wo	Emer ALit	it was ama	So in keeping with my recent attempts to write reviews for all my five star reads here s one for my absolute favourite book from my childhood Little
11	Little_Wo	Jonathan	it was ama	Little Women remains to this day one of the books I have curiously read the most And I m not ashamed to state this Why should I be The notion that
12	Little_Wo	helen the	really like	Two years ago I read the first part of this novel and quite liked it The March family consists of the most endearing characters and I had fun reading ab
13	Little_Wo	Pooja	it was ama	When I was years old I used to watch its anime show on a channel that time I didn t know its name I was merely interested in the show But thankfull
14	Little_Wo	Shovelmo	did not lik	To me this book is just a big neon highlighted literary exclamation mark defining how incredibly different I am from my mother She loves this book I
15	Little_Wo	Matthew	liked it	Updated - Update at endSo this is going to be my most confusing review to date and I am going to need some help from people who read this so ple
16	Little_Wo	Zo	really like	Book for I had to read this book for my Children s Lit class and I loved it We ve done a lot of discussion which has really opened my mind to new thin
17	Little_Wo	Sherwood Smith		There will be spoilers Now if she had been the heroine of a moral story-book she ought at this period of her life to have become quite saintly renou
18	Little_Wo	Tea Jovar	it was ama	Knjiga moje mladosti Ah kako smo je svi gutali
19	Little_Women		it was ama	This book was really good You know those books that you aren t sure what made them your favorite this is one of those books but i will try For one r
20	Little_Wo	Mandy Cr	really like	A classic The book and movie both did me in Tears
21	Little_Wo	Helle	really like	My heart is melting a little as I close this book at long last my mind contented It is a sweet book occasionally a bit too sweet but there were many pl
22	Little_Wo	Antof	it was ama	I have said for years and years how much I like this book but I realized when I started reading it on Sunday that I might not have picked it up since th
23	Little_Wo	Xime Gar	it was ama	Me encant en serio Este libro y yo tenemos una larga historia Es el libro favorito de mi mam y como tal ella habr a preferido que este fuera mi prime
24	Little_Wo	Jo Woolf	did not lik	Read as part of the Infinite Variety Reading Challenge based on the BBC s Big Read poll The one thing I m not going to do is apologise for not liking th

Figure 2

5.3. Exploratory data analysis

For doing exploratory data analysis, we need to prepare the data first. Some of there views are not in English. We will be putting language detection algorithm to reclassify there views. Then we exclude all ratings that do not correspond to a1 to 5star rating. Then we will store the ratings in numerical format.

	A	B
1	book	rating
2	Little_Wor	5
3	Little_Wor	3
4	Little_Wor	4
5	Little_Wor	1
6	Little_Wor	5
7	Little_Wor	5
8	Little_Wor	5
9	Little_Wor	3
10	Little_Wor	5
11	Little_Wor	5
12	Little_Wor	4
13	Little_Wor	5
14	Little_Wor	1
15	Little_Wor	3
16	Little_Wor	4

Figure 3

Then we distribute the ratings for our future analysis.

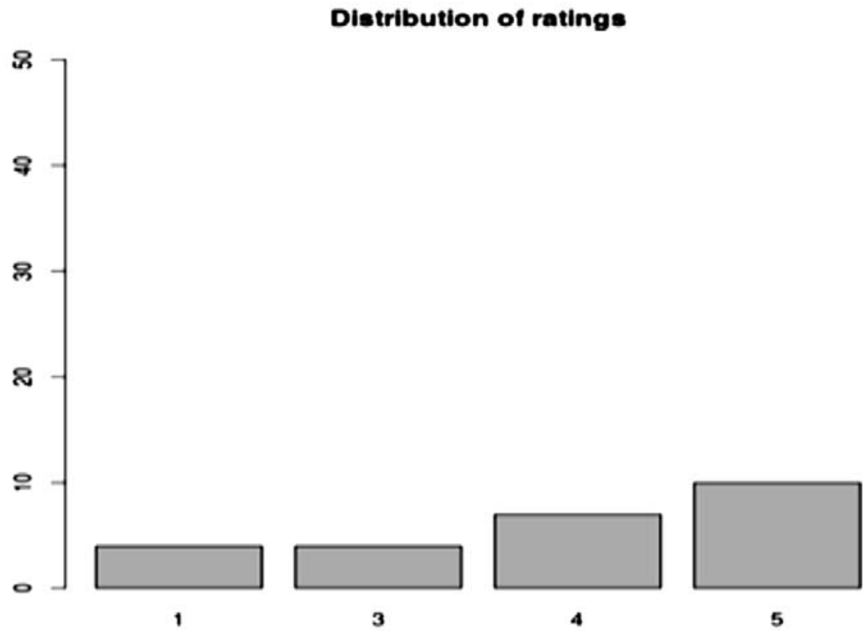

Figure 4

We look into the distribution of review length.

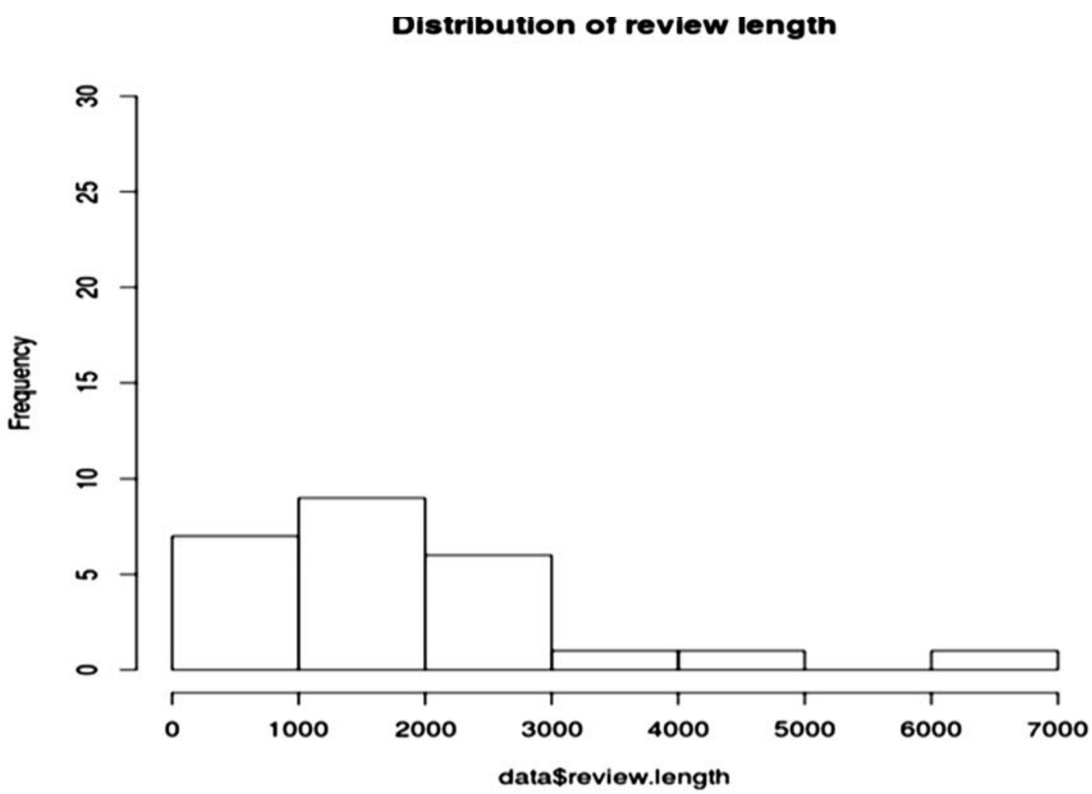

Figure 5

Finally we will show the plot of review length by positive/negative score.

The diagram below shows the plot of review length by ratings.

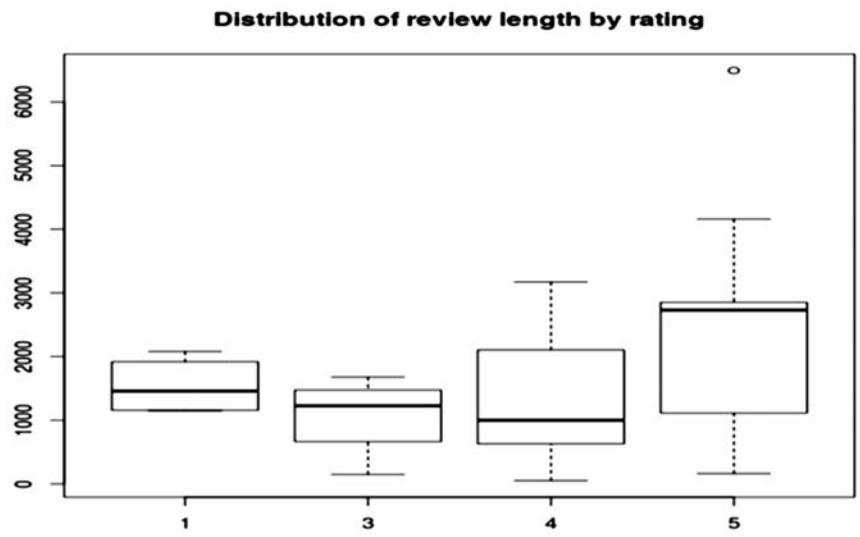

Figure 6

5.4. Sentimental analysis

In this section we are going to show the "positive" or"negative" aspect of the words to see if it correlates with the ratings. We will be establishing lexicons of words with a positive/negative score.

ating	review.id	review	count.afir	count.afir	count.bin	count.bing.positive
5	1	Someone	10	22	13	23
3	2	Okay I m	1	2	1	4
4	3	Yes yes I	15	22	9	18
1	4	This revie	18	14	19	13
5	5	The book	21	20	17	31
5	6	My copy c	NA	9	NA	8
5	7	i can tell t	NA	1	NA	3
3	8	I m defini	7	10	10	12
5	9	So in kee	5	29	7	31
5	10	Little Wo	9	20	14	27
4	11	Two year	1	12	1	9
5	12	When I w	2	3	2	2
1	13	To me thi	8	6	9	8
3	14	Updated -	3	13	2	16
4	15	Book for I	1	5	1	5
5	16	This book	1	17	2	17
4	17	A classic 1	1	NA	NA	1
4	18	My heart	6	34	13	36
5	19	I have sai	14	38	13	41
1	20	Read as p	8	16	6	16
3	21	This mini	3	9	5	10
5	22	I believe	17	35	17	41
1	23	No wond	17	12	13	15
4	24	I once dic	6	19	3	21

Figure 7

We assigned a positivity/negativity score to each review by calculating the average score of allthe words in there view.

rating	review.id	median_s	mean_sentiment
5	1	1.5	0.8125
3	2	2	0
4	3	1	0.675676
1	4	-1	-0.25
5	5	-1	-0.04878
5	6	2	1.666667
5	7	2	2
3	8	2	0.588235
5	9	2.5	1.882353
5	10	2	0.896552
4	11	2	2
5	12	1	0
1	13	-1	-0.35714
3	14	2	1.6875
4	15	2.5	1.666667
5	16	2.5	2.277778
4	17	-2	-2
4	18	2	1.625
5	19	2	1.153846
1	20	1	0.708333
3	21	2	1.333333

Figure 8

The plot looks like:

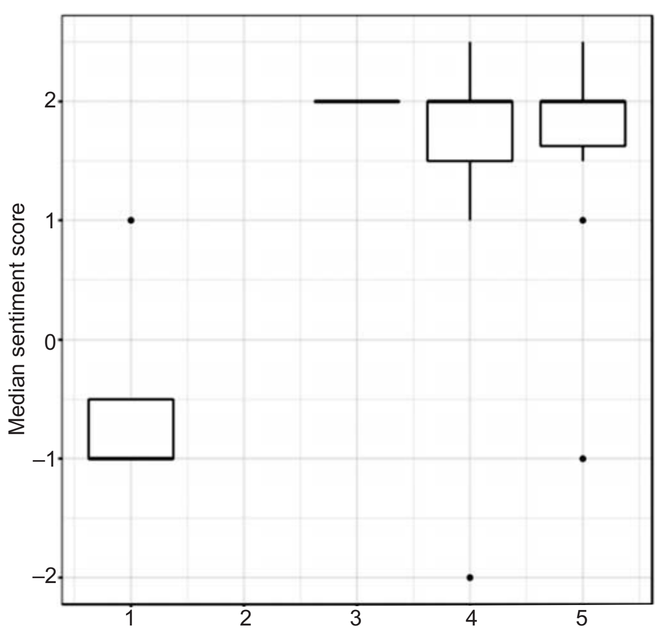

Figure 9: (*a*)

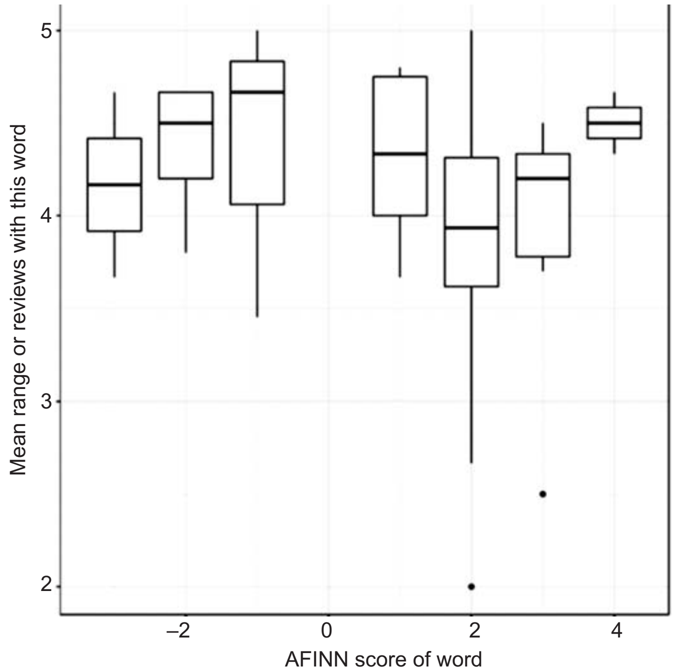

Figure 9: (*b*)

We are going to count the number of negative and positive words in each review, according to the two lexicons, for use in the machine learning algorithm.

5.5. Machine learning

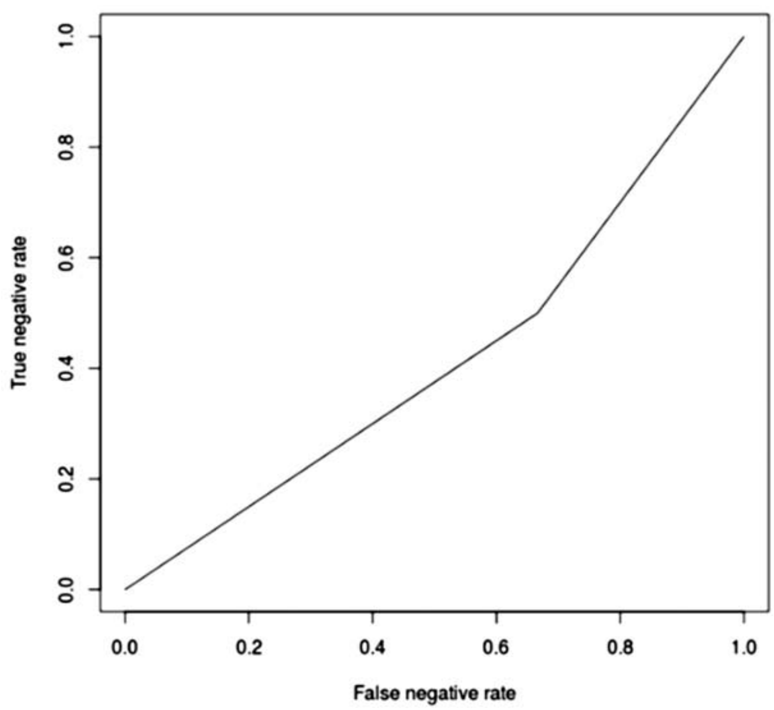

Figure 10

We are using the XGBOOST algorithm here as it gives the best results as compared to support vector machine and random forest algorithm. We started by calculating our base line accuracy, what always predicts the most frequent category, and then we calibrate our model. The XGBoost algorithm produces a probabilistic prediction, that's why we need to determine a threshold value over which we will be classifying are view as good. For this purpose, we plot the Receiver Operating Characteristic curve for the true negative rate against the false negative rate. By using at hreshold of about 0.8 (where the curve becomes red), we can correctly classify more than 50%of the negative reviews (the true negative rate)while misclassifying as negative reviews less than 10% of the positive reviews(the false negative rate).

Our over all accuracy is 82%, so we beat the bench mark of alway spredicting thata review is positive. While catching 53.5% of the negative reviews. If we consider the relative importance of features in XGBOOST algorithm, then the plot will be like:

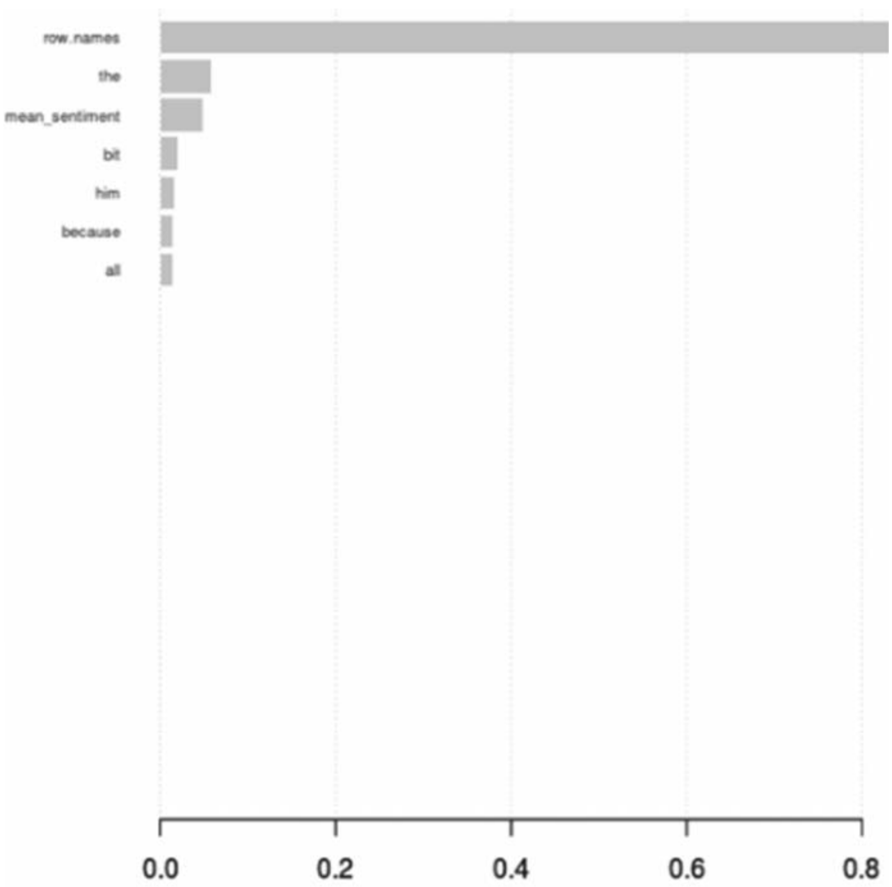

Figure 11

6. CONCLUSION

Here in this projectwe haveevaluated the things by starting the thingsfrom webscraping to sentimentanalysis to predictive analytics with machinelearning. Wehaveusedthe XGBOOST algorithm inthisprospectivetobuildour model.in orderto predict the things more accurately ortogetmoreaccuracy,other respectivemachinelearning algorithmscanbe used.Thereareseveralwayswecouldimprove ontheanalysissuchasif we investigate reviews, we can conclude that they have been misclassified. We can do further analysis in order to recognize what are the features we can add to it and can build a meaningful analytical pipeline.

REFERENCES

[1] W Berger, H. Piringer, P. Filzmoser, and E. Groller. Uncertainty-aware exploration of continuous parameter spaces usingmultivariate prediction.In Computer Graphics Forum, volume 30, pages 911–920. Wiley OnlineLibrary, 2011.

[2] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guidingfeature subset selection with an interactive visualization. In IEEE Conferenceon Visual Analytics Science and Technology, pages 111–120. IEEE,2011.

[3] Z. Zhao, L. Wang, and H. Liu.Efficient spectral feature selection withminimum redundancy. In AAAI Conference on Artificial Intelligence,2010.

[4] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of googleflu: Traps in big data analysis. Science, 343(6176):1203–1205, 2014.

[5] Boureau, Y-L, Bach, Francis, LeCun, Yann, and Ponce,Jean. Learning mid-level features for recognition. InComputer Vision and Pattern Recognition (CVPR), 2010IEEE Conference on, pp. 2559–2566. IEEE, 2010a.

[6] Y. Lu, F.Wang, and R. Maciejewski. VAST 2013 Mini-Challenge 1: BoxOffice VAST-Team VADER. In IEEE Conference on Visual AnalyticsScience and Technology, 2013.

[7] Girshick, Ross B., Donahue, Jeff, Darrell, Trevor, andMalik, Jitendra. Rich feature hierarchies for accurateobject detection,2013

[8] LeCun, L. Bottou, Y. Bengio, P. Haffner. 1998.Gradient-based learning applied to document recognition. In Proceedings of the IEEE, 86(11):22782324, November.

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K.Kavukcuglu, P. Kuksa. 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 12:2493–2537.

[10] Rianne Dekker and GodfriedEngbersen, “How social media transform migrant networks and facilitatemigration.”,Novermber 2012, IMI, University of Oxford.

[11] Aron Culotta ,Nirmal Kumar Ravi, Jennifer Cutler, “Predicting the Demographics of Twitter Users from Website Traffic Data”, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[12] Marco Pennacchiotti , Ana-Maria Popescu, “A Machine Learning Approach to Twitter User Classification”, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

[13] Huiji Gao, GeoffrayBarbiar and RobeccaGoolsby. “Harnessing the crowdsourcing power of social media for disaster relief”, IEEE Intelligent Systems, 2011.

[14] J. S. Simonoff and I. R. Sparrow.Predictingmovie grosses: Winners and losers, blockbusters and sleepers. Chance, 13(3):15–24, 2000

[15] J. Ulbts. JMDB: Java Movie Database. http://www.jmdb.de/, 2014.

[16] SAS Institute Inc. SAS/STAT Software, Version 9.3. Cary, NC, 2011.

[17] RCoreTeam. R:ALanguageandEnvironmentforStatisticalComputing. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[18] M. Harrower and C. A. Brewer.Colorbrewer.org: An online tool for selectingcolourschemesformaps. The Cartographic Journal, 40(1):27– 37, 2003.

[19] K. Ericsson and H. Simon. Protocol Analysis: Verbal Reports as Data. MIT Press, Cambridge, MA, 1993