iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VIII **Month of publication:** August 2020

DOI: https://doi.org/10.22214/ijraset.2020.30950

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

Detection of Fake Online Reviews using ML

Mallikarjuna S B [1], Nithya H L[2]
[1, 2]Department of Computer science and Engineering, BIET Davanagere.

Abstract: Online audits have incredible effect on the present business furthermore, trade. Basic leadership for acquisition of on the web items generally relies upon surveys given by the clients. Thus, shrewd people or gatherings attempt to control item surveys for their own advantages. This paper presents a few semi-supervised and supervised content mining models to recognize counterfeit online audits just as analyses the productivity of both procedures on dataset containing lodging surveys.
Keywords: Semi-supervised, Supervised and Counterfeit.

I. INTRODUCTION

With the hazardous development of data on the web, the web has turned into the best and goliath conveyed registering application today. Billions of site pages are shared by a large number of associations, colleges, scientists, and so forth. Web quest gives extraordinary usefulness to circulating, sharing, sorting out, and recovering the developing measure of data. Accordingly, web indexes have turned out to be increasingly significant and are utilized by a great many individuals to discover essential data. It has turned out to be significant for a website page, to be positioned high in the significant web indexes' outcomes. Accordingly numerous methods are proposed to impact positioning and improve the position of a page. A portion of these methods are legitimate and are called Search Engine Optimization (SEO) systems, yet some are not lawful or moral and attempt to hoodwink positioning calculations. They attempt to rank pages higher than they merit [2].

Web spam alludes to web content that get high position in internet searcher results in spite of uninformed worth. Spamming deceives clients, yet in addition forces existence cost to internet searcher crawlers and indexers. That is the reason crawlers attempt to recognize web spam pages to abstain from handling and ordering them.

The social Web and the expanding notoriety of web based life have prompted the spread of different sorts of substance produced legitimately by clients, the purported User Generated content (UGC). By methods for Web 2.0 innovations, it is workable for each person to diffuse substance via web-based networking media, nearly with no type of confided in outside control. This infers that there are no way to check, from the earlier, the dependability of the sources and the acceptability of the substance created [2].

Technologies are evolving quickly. Old advancements are ceaselessly being supplanted by new and modern ones. These new advancements are empowering individuals to have their work done proficiently. Such a development of innovation is online commercial centre. We can shop and reserve spot utilizing online sites. Nearly, everybody of us looks at reviews prior to acquiring a few items or services. Consequently, on the web audits have turned into an extraordinary wellspring of notoriety for the organizations. Likewise, they have enormous effect on ad also, advancement of items and services. With the spread of online commercial centre, counterfeit online surveys are getting to be incredible matter of concern. Individuals can make false surveys for advancement of their own items that damages the real clients. Additionally, aggressive organizations can attempt to harm every others notoriety by giving phony negative reviews [1].

Analysts have been learning about numerous methodologies for location of these phony online surveys. A few approaches are survey substance put together and some are based with respect to conduct of the client who is posting audits. Substance based investigation centres around what is composed on the survey that is the content of the survey where client conduct put together technique centres with respect to nation, ip-address, number of posts of the analyst and so forth. A large portion of the proposed methodologies are supervised grouping models. Hardly any analysts, likewise have worked with semi-supervised models. Semi-supervised techniques are being presented for absence of dependable marking of the surveys.

In this paper, we make some arrangement draws near for recognizing counterfeit online audits, some of which are semi-supervised what's more, others are regulated. For semi-supervised learning, we use Expectation-amplification calculation. Factual Gullible Bayes classifier and Support Vector Machines (SVM) are utilized as classifiers in our examination work to improve the execution of arrangement. We have for the most part centred around the substance of the survey-based methodologies. As highlight we have utilized word recurrence check, conclusion extremity and length of survey.

A. Objectives

An efficient platform is developed to detect the fake reviews generated by the user in online marketing by the supervised and semi-supervised study. The objective of proposed platform is as follows:

1) To develop reliable and efficient platform for necessary feature extraction from the raw text data.
2) To develop an efficient semi-supervised and supervised text mining techniques for detecting fake online reviews.
3) To do performance analysis of the proposed system.

B. Proposed System

Proposed work, focus on some classification approaches for detecting fake online reviews, some of which are semi-supervised and others are supervised. For semi-supervised learning, Expectation-maximization algorithm is used. Statistical Naive Bayes classifier and Support Vector Machines (SVM) are used as classifiers in our research work to improve the performance of classification. Main focused on the content of the review based approaches. As feature we have used word frequency count, sentiment polarity and length of review.

II. LITERATURE REVIEW

An In 2018, Pankaj Chaudhary, Abhimanyu Tyagi and Santosh Mishra presents article targets giving an examination of the fundamental survey and analyst driven highlights that have been proposed up to presently in the writing to identify phony audits, specifically from those methodologies that utilize directed AI procedures. These arrangements furnish by and large better outcomes as for simply solo approaches, which are regularly in view of diagram based techniques that think about social ties in audit locales. Moreover, this work proposes and assesses some extra new highlights that can be reasonable to order certified and phony audits. For this reason, a regulated classifier dependent on Random Forests have been actualized, by thinking about both surely understood and new highlights, and an enormous scale marked dataset from which every one of these highlights have been extricated.

In 2017, J. K. Rout, A. Dalmia and K.-K. R. Choo clarify how semi-administered learning strategies can be utilized to distinguish spam audits, before exhibiting its utility utilizing a dataset of lodging audits(reviews).

In 2016, Chengai Sun, Qiaolin Du and Gang Tian proposes a novel convolutional neural system model to coordinate the item related audit includes through an item word structure model. To lessen over fitting and high difference, a sacking model is acquainted with pack the neural system model with two productive classifiers. Tests on the genuine Amazon survey dataset exhibit the viability of the proposed methodology.

In 2015, A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari present research around deliberately breaking down and ordering models that recognize survey spam. Next, the examination continues to evaluate them as far as exactness and results.They find that reviews can be arranged into three gatherings that emphasis on techniques to identify spam surveys, singular spammers and gathering spam. Diverse identification methods have various qualities and shortcomings and in this manner support distinctive discovery settings.

In 2014, J. Li, M. Ott, C. Cardie and E. Hovy investigate summed up approaches for distinguishing online misleading assessment spam based on another highest quality level dataset, which is contained information from three distinct areas, every one of which contains three sorts of audits. They Proposed methodology which attempts to catch the general contrast of language utilization among beguiling and honest audits, which they expectation will support clients when settling on buy choices and survey entry administrators.

In 2012, J. Karimpour, A. A. Noroozi and S. Alizadeh propose another strategy to determine this downside by utilizing semi-administered figuring out how to consequently name the preparation information. To do this, they fuse Expectation-Maximization calculation that is a productive and a significant calculation of semi-regulated learning. Trials are completed on the genuine web spam information, which demonstrate the new technique, performs very well.

In 2012, S. Feng, R. Banerjee and Y. Choi made research on syntactic stylometry for double dealing identification, including a to some degree whimsical edge to earlier writing. More than four diverse datasets crossing from the item survey to the article space, they show that highlights driven from Setting Free Grammar (CFG) parse trees reliably improve the recognition execution more than a few baselines that are based just on shallow lexico-syntactic highlights.

In 2011, M. Ott, Y. Choi, C. Cardie and J. T. Hancock studied deceptive feeling spam—invented sentiments that have been purposely composed to sound valid. Incorporating work from brain research furthermore, computational semantics, thay create furthermore, contrast three methodologies with distinguishing misleading feeling spam, and at last build up a classifier that is about 90% exact on our highest quality level feeling spam dataset.

In light of highlight investigation of our scholarly models, they furthermore make a few hypothetical commitments, including uncovering a relationship between misleading feelings and creative composing.
In 2010, E. P. Lim, V.-A. Nguyen, N. Jindal, B. Liu and H. W. Lauw distinguish clients creating spam surveys or then again survey spammers. They recognize a few trademark behaviours of audit spammers and model these practices so as to identify the spammers. Specifically, we look to demonstrate the accompanying practices. To begin with, spammers may target specific items or item bunches so as to boost their impact. Second, they will in general veer off from different commentators in their appraisals of items. They propose scoring techniques to quantify the level of spam for every analyst and apply them on an Amazon audit dataset. They at that point select a sub- set of exceptionally suspicious analysts for further examination by our client evaluators with the assistance of an electronic spammer evaluation programming uncommonly produced for client assessment experiments. They demonstrate that the identified spammers have more significant sway on appraisals contrasted and the unhelpful analysts.
In 2001, J. W. Pennebaker, M. E. Francis and R. J. Booth give a proficient and successful strategy for considering the different passionate, intellectual, and basic segments present in people's verbal and composed discourse tests, they initially built up a book examination application called Linguistic Inquiry and Word Count, or LIWC. The first LIWC application was created as a component of an exploratory investigation of language and divulgence. The program is intended to examine individual or numerous language records rapidly and effectively. Simultaneously, the program endeavours to be straightforward and adaptable in its activity, enabling the client to investigate word use in various manners.

III.SYSTEM DESIGN

System design thought as the application of theory of the systems for the development of the project. System design defines the architecture, data flow, use case, class, sequence and activity diagrams of the project development.

A. System Architecture

This architecture diagram illustrates how the system is built and is the basic construction of the software method. Creations of such structures and documentation of these structures is the main responsible of software architecture.

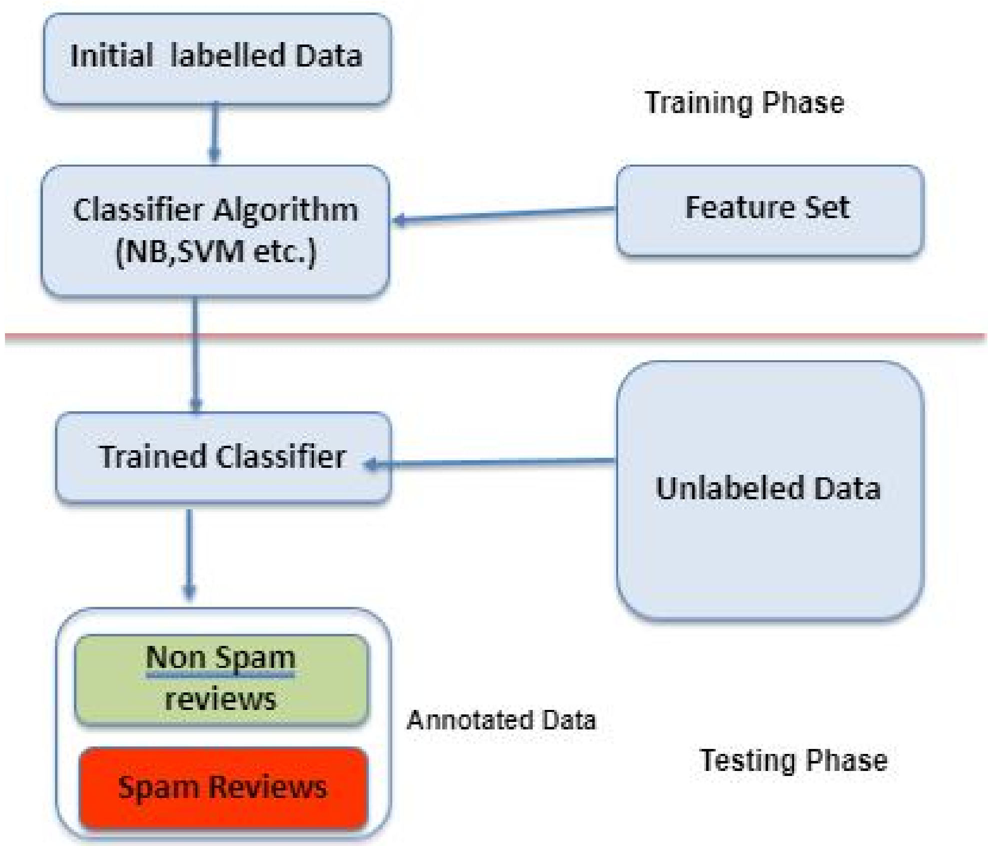


Fig. 1 Architecture Diagram.

B. Data Flow Diagram

Data flow diagram also referred as bubble graph. This diagram is useful for representing the system for all degree of constructions. The figure is differentiated into parts which show maximizing data path & practical aspect.

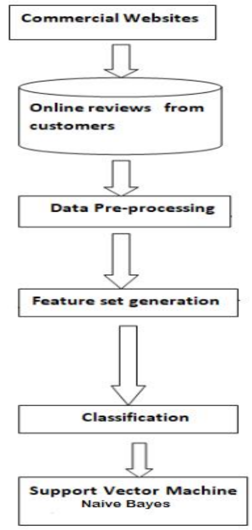


Fig.2 Data Flow Diagram

The Fig. 2 is the dataflow diagram of proposed system. Firstly, we select customer reviews as datasets from commercial website where we get both genuine and fake reviews done for marketing. Than the dataset is undergoes preprocessing stage where null data, stop words and unnecessary data is removed and dataset is transformed into particular format we say CSV in this project. Next dataset in fed into feature selection process based on this feature classification is done by training classifier using these features.

IV. IMPLEMENTATION

A. Supervised

We employ two different supervised machine learning methods, namely Multinomial Naive Bayes and Support Vector Machine. Both these techniques require labeled data to build classifiers that annotate data as spam or non-spam. Manual labeling of data is very resource intensive, requires a lot of time and training and still the authenticity of the labeled data cannot be determined with certainty. This is one of the major drawbacks of the supervised learning methods and to mitigate the effects of these weaknesses we explore a semi supervised approach for spam detection based on a co-training algorithm.

Algorithm 1 Algorithm for detecting spam reviews using SVM

```
documentsList= input dataset
for i = 1 to 5 do
    training_data, testing_data = Split documentsList
    tokenize = CountVectorizer(training_data).fit_transform(training_data)
    tfidf = TfidfTransformer().fit_transform(tokenize)
    classifier = SGDClassifier().fit(tfidf)
    prediction = classifier.predict(testing_data)
end for
```

B. Semi Supervised

We use a semi supervised two view co-training algorithm to annotate the large set of unlabeled data from a small labeled data set. The motivation behind implementing this technique is that co-training algorithm takes advantage of the feature split when learning from labeled and unlabeled data. The feature sets we presented are independent of each other as review features are more focused on content and text of reviews. While reviewer features focus on friend count, rating deviation and review count of each reviewer.

Another motivation is that manual labeling of data is labor intensive and resource consuming. Labeling even a small set of data set requires a lot of effort and still we are left with a large set of unlabeled data. Co training aims at utilizing this small set of labeled data to annotate the unlabeled reviews. Its approach is to incrementally build classifiers over each feature sets.

It is a two view algorithm, where the first view is to directly detect if the review is spam; the other view is to detect if the author of the review is spammer. The major steps of the algorithm are:

1) For each review it uses two views of feature sets. Review features ‘Fr’ and reviewer features ‘Fu’. ‘L’ is a small set of labeled reviews and ‘U’ is the large set of unlabeled reviews.
2) We learn two classifiers, C_r based on review features and C_u based on reviewer features.
3) C_r labels reviews from U based on Fr, p positive and n negative reviews are from U, $T_{reviews}$.
4) Cu labels reviews from U based on Fu, p positive and n negative reviews are from U, $T'_{reviewers}$.
5) Extract reviews $T'_{reviews}$ authored by $T'_{reviewers}$
6) Move $T_{reviews}$ U T'_{review}s from U to L.

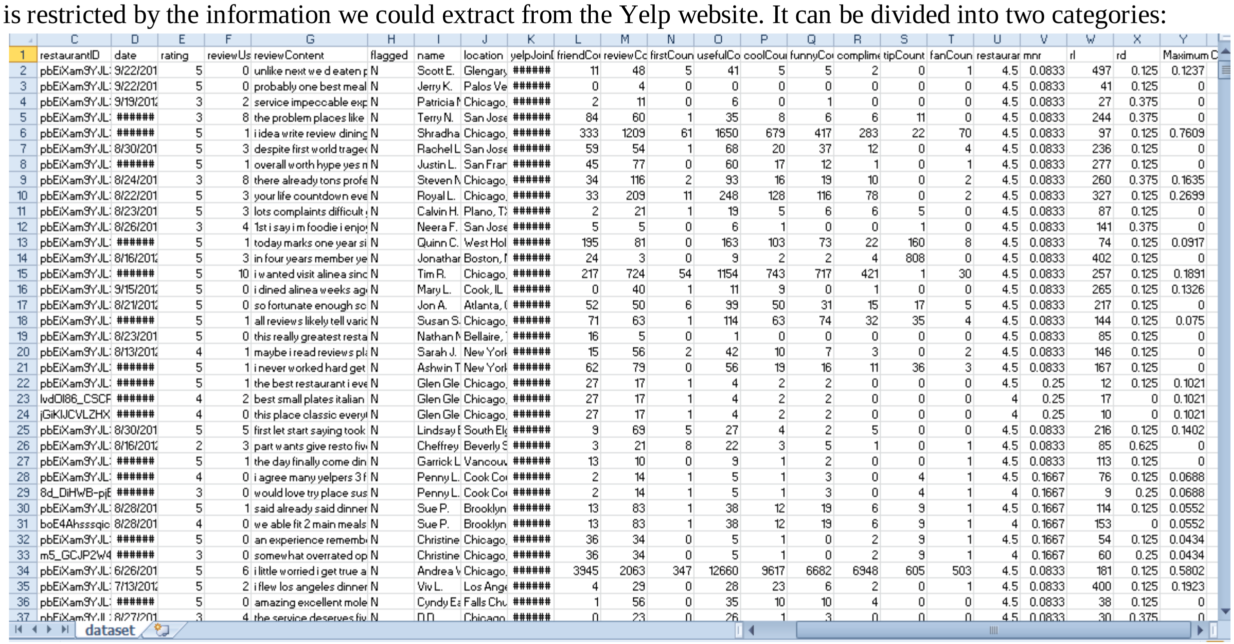
C. Data Collection Process

The data set utilized by our project primarily comes from the Yelp website. It contains data in JSON format and the data which was of particular use to us contains the following entities: Review Object and User (Reviewer) Object. We have only considered data related to restaurants from yelp dataset ignoring the other ones. For this we parsed out records corresponding to restaurants only. The following steps were involved in the data collection process

1) Filter out all business objects in the business data file having the category “Restaurant” in the records
2) Filter out all the reviews having the filtered business IDs in them.
3) Filter out all user IDs from the review data file and use that to fetch details of every user from user data file who had reviewed at least one of the restaurants.

The figure 3 shows the dataset content which is in CSV file.

By combining the both supervised and semi supervised method, we introduce the feature set that we used for spam detection. Based on previous works, we selected a set of features that play a key role in determining whether a review is spam or not. This set of features is restricted by the information we could extract from the Yelp website. It can be divided into two categories:



	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	restaurantID	date	rating	reviewUs	reviewContent	flagged	name	location	yelpJoin[	friendCo	reviewCc	firstCoun	usefulCo	coolCour	funnyCo	complim	tipCount	fanCoun	restaurar	mnr	rl	rd	Maximum C
2	pbEiXam9YJL:	9/22/201	5	0	unlike next we d eaten p	N	Scott E.	Glengary	######	11	48	5	41	5	5	2	0	1	4.5	0.0833	497	0.125	0.1237
3	pbEiXam9YJL:	9/22/201	5	0	probably one best meal	N	Jerry K.	Palos Ve	######	0	4	0	0	0	0	0	0	0	4.5	0.0833	41	0.125	0
4	pbEiXam9YJL:	9/19/2012	3	2	service impeccable exp	N	Patricia I	Chicago,	######	2	11	0	6	0	1	0	0	0	4.5	0.0833	27	0.375	0
5	pbEiXam9YJL:	######	3	8	the problem places like	N	Terry N.	San Jose	######	84	60	1	35	8	6	6	11	0	4.5	0.0833	244	0.375	0
6	pbEiXam9YJL:	######	5	1	i idea write review dining	N	Shradha	Chicago,	######	333	1209	61	1650	679	417	283	22	70	4.5	0.0833	97	0.125	0.7609
7	pbEiXam9YJL:	8/30/201	5	3	despite first world trage	N	Rachel L	San Jose	######	59	54	1	68	20	37	12	0	4	4.5	0.0833	236	0.125	0
8	pbEiXam9YJL:	######	5	1	overall worth hype yes n	N	Justin L.	San Frar	######	45	77	0	60	17	12	1	0	1	4.5	0.0833	277	0.125	0
9	pbEiXam9YJL:	8/24/201	3	8	there already tons profe	N	Steven N	Chicago,	######	34	116	2	93	16	19	10	0	2	4.5	0.0833	260	0.375	0.1635
10	pbEiXam9YJL:	8/22/201	5	3	your life countdown eve	N	Royal L.	Chicago,	######	33	209	11	248	128	116	78	0	2	4.5	0.0833	327	0.125	0.2699
11	pbEiXam9YJL:	8/23/201	5	3	lots complaints difficult	N	Calvin H.	Plano, T	######	2	21	1	19	5	6	6	5	0	4.5	0.0833	87	0.125	0
12	pbEiXam9YJL:	8/26/201	3	4	1st i say i m foodie i enjo	N	Neera F.	San Jose	######	5	5	0	6	1	0	0	1	0	4.5	0.0833	141	0.375	0
13	pbEiXam9YJL:	######	5	1	today marks one year si	N	Quinn C.	West Hol	######	195	81	0	163	103	73	22	160	8	4.5	0.0833	74	0.125	0.0917
14	pbEiXam9YJL:	8/16/2012	5	3	in four years member ye	N	Jonathar	Boston, I	######	24	3	0	9	2	2	4	808	0	4.5	0.0833	402	0.125	0
15	pbEiXam9YJL:	######	5	10	i wanted visit alinea sinc	N	Tim R.	Chicago,	######	217	724	54	1154	743	717	421	1	30	4.5	0.0833	257	0.125	0.1891
16	pbEiXam9YJL:	9/15/2012	5	0	i dined alinea weeks ag	N	Mary L.	Cook, IL	######	0	40	1	11	9	0	1	0	0	4.5	0.0833	265	0.125	0.1326
17	pbEiXam9YJL:	8/21/2012	5	0	so fortunate enough sc	N	Jon A.	Atlanta, (	######	52	50	6	99	50	31	15	17	5	4.5	0.0833	217	0.125	0
18	pbEiXam9YJL:	######	5	1	all reviews likely tell vari	N	Susan S.	Chicago,	######	71	63	1	114	63	74	32	35	4	4.5	0.0833	144	0.125	0.075
19	pbEiXam9YJL:	8/23/201	5	0	this really greatest resta	N	Nathan N	Bellaire,	######	16	5	0	1	0	0	0	0	0	4.5	0.0833	85	0.125	0
20	pbEiXam9YJL:	8/13/2012	4	1	maybe i read reviews pl	N	Sarah J.	New York	######	15	56	2	42	10	7	3	0	2	4.5	0.0833	146	0.125	0
21	pbEiXam9YJL:	######	5	1	i never worked hard get	N	Ashwin T	New York	######	62	79	0	56	19	16	11	36	3	4.5	0.0833	167	0.125	0
22	pbEiXam9YJL:	######	5	1	the best restaurant i eve	N	Glen Gle	Chicago,	######	27	17	1	4	2	2	0	0	0	4.5	0.25	12	0.125	0.1021
23	lvdOI86_CSCF	######	4	2	best small plates italian	N	Glen Gle	Chicago,	######	27	17	1	4	2	2	0	0	0	4	0.25	17	0	0.1021
24	jGiKIJCVLZHX	######	4	0	this place classic every	N	Glen Gle	Chicago,	######	27	17	1	4	2	2	0	0	0	4	0.25	10	0	0.1021
25	pbEiXam9YJL:	8/30/201	5	5	first let start saying took	N	Lindsay I	South El	######	9	69	5	27	4	2	5	0	0	4.5	0.0833	216	0.125	0.1402
26	pbEiXam9YJL:	8/16/2012	2	3	part wants give resto fiv	N	Cheffrey	Beverly S	######	3	21	8	22	3	5	1	0	1	4.5	0.0833	85	0.625	0
27	pbEiXam9YJL:	######	5	1	the day finally come din	N	Garrick L	Vancouv	######	13	10	0	9	1	2	0	0	1	4.5	0.0833	113	0.125	0
28	pbEiXam9YJL:	######	4	0	i agree many yelpers 3 f	N	Penny L.	Cook Co	######	2	14	1	5	1	3	0	4	1	4.5	0.1667	76	0.125	0.0688
29	8d_DiHWB-pjE	######	3	0	would love try place sus	N	Penny L.	Cook Co	######	2	14	1	5	1	3	0	4	1	4	0.1667	9	0.25	0.0688
30	pbEiXam9YJL:	8/28/201	5	1	said already said dinner	N	Sue P.	Brooklyn	######	13	83	1	38	12	19	6	9	1	4.5	0.1667	114	0.125	0.0552
31	boE4Ahsssqic	8/28/201	4	0	we able fit 2 main meals	N	Sue P.	Brooklyn	######	13	83	1	38	12	19	6	9	1	4	0.1667	153	0	0.0552
32	pbEiXam9YJL:	######	5	0	an experience rememb	N	Christine	Chicago,	######	36	34	0	5	1	0	2	9	1	4.5	0.1667	54	0.125	0.0434
33	m5_GCJP2W4	######	3	0	somewhat overrated op	N	Christine	Chicago,	######	36	34	0	5	1	0	2	9	1	4	0.1667	60	0.25	0.0434
34	pbEiXam9YJL:	6/26/201	5	6	i little worried i get true a	N	Andrea V	Chicago,	######	3945	2063	347	12660	9617	6682	6948	605	503	4.5	0.0833	181	0.125	0.5802
35	pbEiXam9YJL:	7/13/2012	5	2	i flew los angeles dinner	N	Viv L.	Los Ange	######	4	29	0	28	23	6	2	0	1	4.5	0.0833	400	0.125	0.1923
36	pbEiXam9YJL:	######	5	0	amazing excellent mole	N	Cyndy Ea	Falls Chu	######	1	56	0	35	10	10	4	0	0	4.5	0.0833	38	0.125	0
37	pbEiXam9YJL:	8/27/201	3	4	the service deserves fiv	N	D.D.	Chicago	######	0	23	0	26	1	3	0	0	0	4.5	0.0833	30	0.375	0

dataset

Fig. 3 Dataset CSV file

As classifier, we have used Support Vector machines(SVM) and Naive Bayes(NB) classifier with co-training algorithm. Scikit Learn package of Python programming language provides sophisticated library of these classifiers. Hence for our research work, we have used Python with scikit-learn and numpy packages. We have tuned the parameters of the SVM for better results. For supervised classification, we have used Multinomial Naive Bayes and SVM classifiers. We know, Naive Bayes classifier can be implemented where conditional independence property is maintained. As, text comes randomly from user mind, we can’t know what the next line and word is going to be. Hence, Naive Bayes classifier is popularly used in text mining. It is probabilistic method hence it can be used both for classification and regression. It is also very fast to calculate.

V. PERFORMANCE ANALYSIS

We run the co training algorithm using two different classifiers based on Naive Bayes method and SVM and analyze the results obtained using each method. The results obtained from the co-training method are evaluated based on the evaluation metrics of precision, recall and F score.

Precision = Sp∩Sc/Sp ,

Recall = Sp∩Sc/Sc ,

F = 2∗Precision∗Recall/(Precision + Recall)

where, Sc is the set of true review spams, Sp is the set of predicted review spams. The co-training algorithm is simple to implement and the only mathematical background required would be an understanding of the Naive Bayes algorithm and SVM algorithm. As we will observe through the results the co-training algorithm takes advantage of the feature split that is not considered by either Linear SVM or Multinomial Naive Bayes methods and produces superior results.

We evaluate the performance using Precision, Recall and F-Score. Fig 4 and 5 shows the result of MNB and SVM with different feature sets. We observe little variation in the performance of NB method as we increase feature sets, which indicates that the other features are dominated by our four best features.. When we include only review features and exclude the reviewer features the F-Score drops the most in both NB and SVM method.

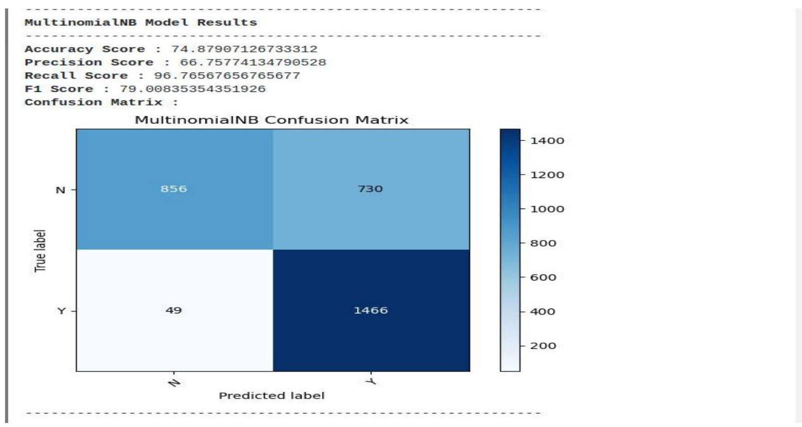


Fig. 4 Performance Results of MNB method with feature set

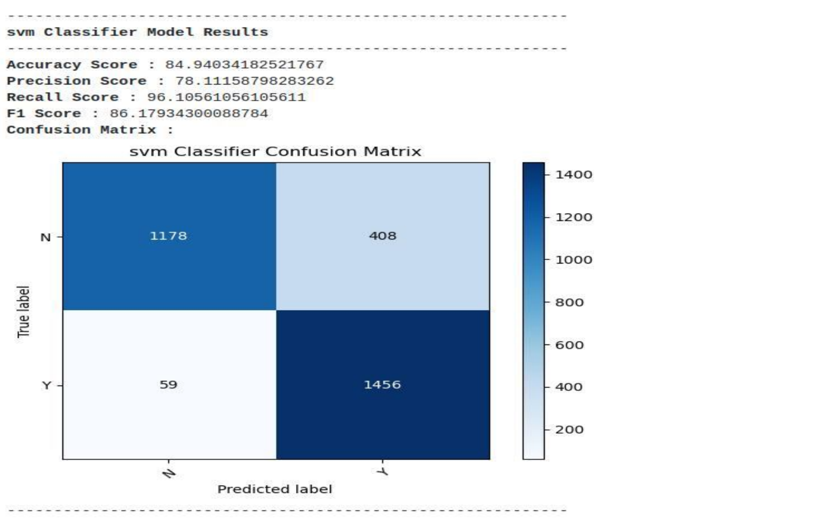


Fig. 5 Performance results of SVM method with feature set

Version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

VI. RESULTS AND OUTPUT SCREENSHOTS

The results of supervised methods are shown in the graph in Figure 7.1. The machine learning method Multinomial Naïve Bayes (NB) performs significantly better as compared with the Linear SVM. We used 26900 reviews as training data set for NB and liner SVM method to build classifiers. We initially selected the four best features by implementing feature selection and observe that the Naive Bayes method clearly outperforms SVM. The four best features suggested by feature selection are Reviewer friend count, Reviewer review count, Bigram measures and Length of the review text.

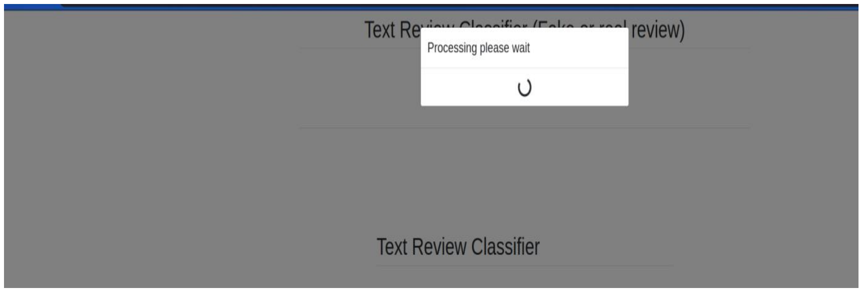


Fig. 6 CSV datset file is processing

Figure 5 depict the processing of the proposed sytem after uploading the dataset which is in CSV file format.

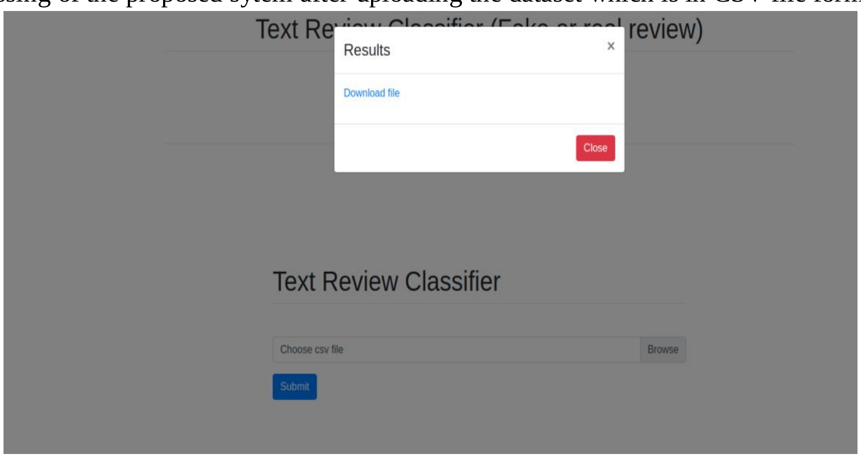
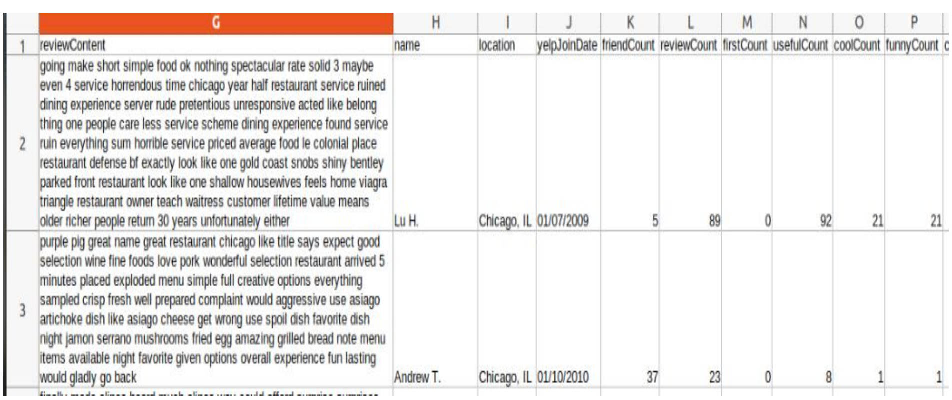


Fig. 7 Predicted file is ready to download

The figure 6 shows the GUI which shows the small window is pop-up in which we have a interface to download the review predicted file.

The predicted result file which is downloaded is shown in below figure 7 and 8. In fig 7 we have Review content, name, location, date and feature sets counts of particular reviews respectively.



	G	H	I	J	K	L	M	N	O	P
1	reviewContent	name	location	yelpJoinDate	friendCount	reviewCount	firstCount	usefulCount	coolCount	funnyCount
2	going make short simple food ok nothing spectacular rate solid 3 maybe even 4 service horrendous time chicago year half restaurant service ruined dining experience server rude pretentious unresponsive acted like belong thing one people care less service scheme dining experience found service ruin everything sum horrible service priced average food le colonial place restaurant defense bf exactly look like one gold coast snobs shiny bentley parked front restaurant look like one shallow housewives feels home viagra triangle restaurant owner teach waitress customer lifetime value means older richer people return 30 years unfortunately either	Lu H.	Chicago, IL	01/07/2009	5	89	0	92	21	21
3	purple pig great name great restaurant chicago like title says expect good selection wine fine foods love pork wonderful selection restaurant arrived 5 minutes placed exploded menu simple full creative options everything sampled crisp fresh well prepared complaint would aggressive use asiago artichoke dish like asiago cheese get wrong use spoil dish favorite dish night jamon serrano mushrooms fried egg amazing grilled bread note menu items available night favorite given options overall experience fun lasting would gladly go back	Andrew T.	Chicago, IL	01/10/2010	37	23	0	8	1	1

Fig.8 Part of Predicted Result along with some feature sets count

In the below fig 8 we have predicted result which is in red color column and actual review is in green color column.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

Q	R	S	T	U	V	W	X	Y	Z	AA	AB
entCount	tipCount	fanCount	restaurantRating	mnr_x	rl	rd	Maximum Content Similarity_x	mnr_y	Maximum Content Similarity_y	IS FAKE(ACTUAL)	IS FAKE(PREDICTED)
3	6	4	4	0.25	94	0.75	0.182124532832036	0.25	0.089568528548493	N	N
1	0	1	4	0.083333333333333	79	0	0.131602026030215	0.083333333333333	0	N	Y
0	0	0	4.5	0.083333333333333	96	0.125	0.134237118813804	0.083333333333333	0	Y	Y
0	0	0	4.5	0.166666666666667	16	0.125	0.026265412280896	0.166666666666667	0.030726286001428	Y	Y

Fig 9 Predicted Result along with some feature sets count

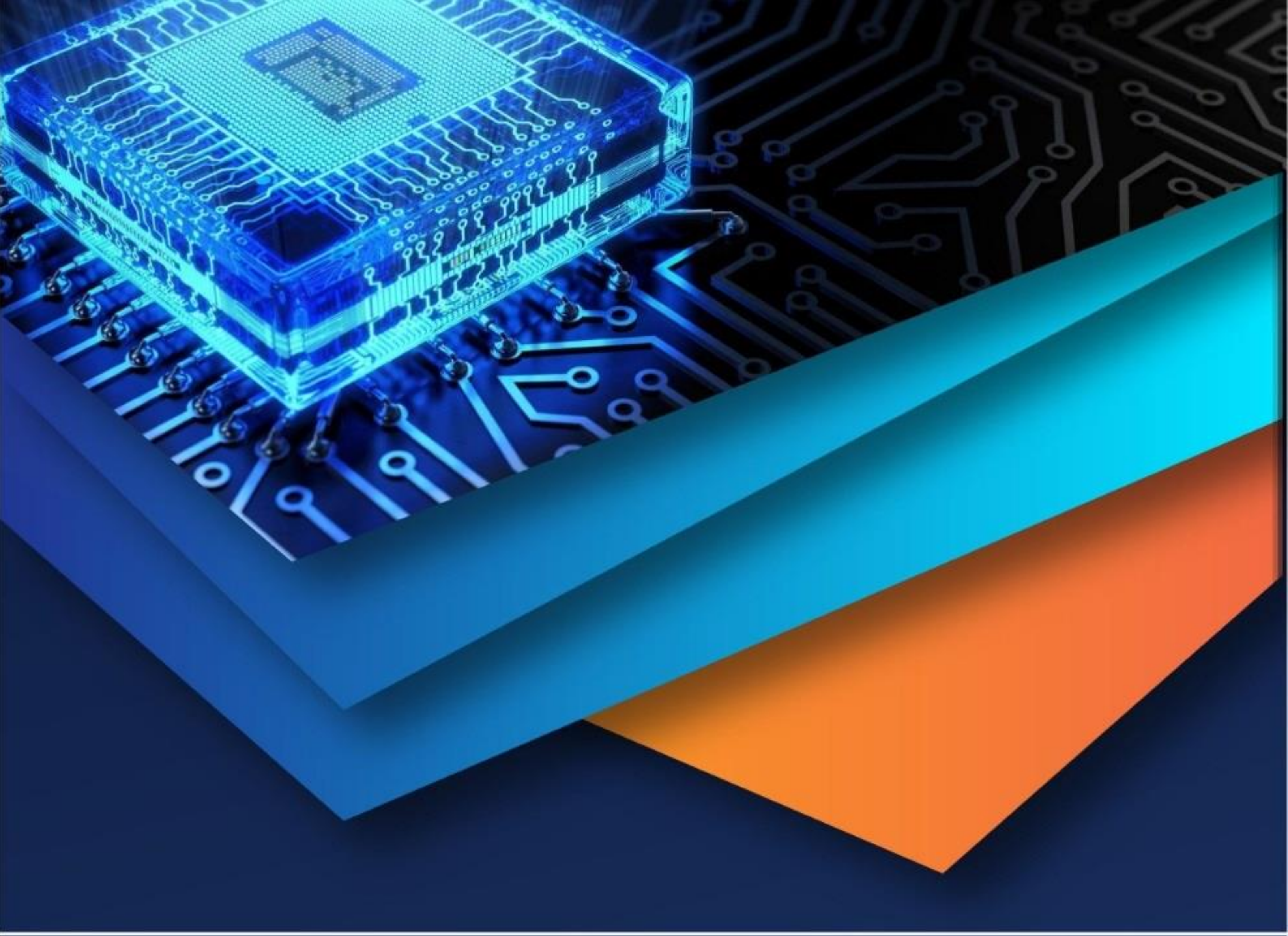
VII. CONCLUSION

Several semi-supervised and supervised text mining techniques are showed for detecting fake online reviews in this research. We have combined features from several research works to create a better feature set. Also we have tried some other classifier that was not used on the previous work. Thus, it has been able to increase the accuracy of previous semi-supervised techniques done. Here it is also found out that supervised Naive Bayes classifier gives the highest accuracy. This ensures that our dataset is labelled well as we know semi-supervised model works well when reliable labelling is not available.

In future, user behaviours can be combined with texts to construct a better model for classification. Advanced pre-processing tools for tokenization can be used to make the dataset more precise. Evaluation of the effectiveness of the proposed methodology can be done for a larger data set. This research work is being done only for English reviews. It can be done for Bangla and several other languages.

REFERENCES

[1] Pankag Chaudhary, Abhimanyu Tyagi and Santosh Mishra, “Fake Review Detection through Supervised Classification”, Conference on Recent Innovations in Emerging Technology & Science, April 6-7, 2018
[2] J. K. Rout, A. Dalmia, and K.-K. R. Choo, “Revisiting semi-supervised learning for online deceptive review detection,” IEEE Access, Vol. 5, pp. 1319–1327, 2017.
[3] Chengai Sun, Qiaolin Du and Gang Tian, “Exploiting Product Related Review Features for Fake Review Detection,” Mathematical Problems in Engineering, 2016.
[4] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, ”Detection of review spam: a survey”, Expert Systems with Applications, vol. 42, no.7, pp. 3634–3642, 2015.
[5] J. Li, M. Ott, C. Cardie, and E. Hovy, “Towards a general rule for identifying deceptive opinion spam,” in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), 2014.
[6] J. Karimpour, A. A. Noroozi, and S. Alizadeh, “Web spam detection by learning from small labeled samples,” International Journal of Computer Applications, vol. 50, no. 21, pp. 1–5, July 2012.
[7] S. Feng, R. Banerjee, and Y. Choi, “Syntactic stylometry for deception detection,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, Vol. 2, 2012.
[8] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies (ACL-HLT), vol. 1, pp. 309–319, Association for Computational Linguistics, Portland, Ore, USA, June 2011.
[9] E. P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, “Detecting product review spammers using rating behaviors,” in Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), 2010.
[10] J. W. Pennebaker, M. E. Francis, and R. J. Booth, ”Linguistic Inquiry and Word Count: Liwc,” vol. 71, 2001.



10.22214/IJRASET

INDEX COPERNICUS
45.98

ISRA JIF
IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY