Mining Movie Intention using Bayes and Maximum Entropy Classifiers

Varsha D. Jadhav
Assistant Professor
P.E.S. College of Engineering
Aurangabad, (MS) India

Sachin N. Deshmukh
Professor
Department of CS/IT, Dr. Babasaheb Ambedkar
Marathwada University
Aurangabad, (MS) India

ABSTRACT

Sentiment analysis is becoming one of the most thoughtful research areas for prediction and classification. This paper analyzes and predicts the result for movie reviews. Machine learning techniques Bayes and Maximum entropy for classifying text messages. Movie comments from twitter are retrieved. The two classifiers are analyzed for Hindi movies 'Sultan' and 'Madaari.' Tweets before and after the release of the movie are retrieved. Accuracy is evaluated to compare the Bayes and Maximum entropy methods. R technology is used for the movie review analysis.

General Terms

Intention Mining, machine learning, normalization, Twitter API.

Keywords

Bayes, Maximum Entropy, polarity, emotions, mean absolute error.

1. INTRODUCTION

Social networks today contains enormous amount of text data, which is growing everyday. Intention mining aims to cover the attitude of the author on a particular topic from text data. It is natural language processing and machine learning techniques reveal the attitude. In the recent years it has gained popularity due to its immediate application in business, customer feedback from product reviews, spots reviews and assisting in election campaigns. Movie reviews are a important way to analyze the performance of a movie. Text movie reviews tells us the strong and weak points of the movie which tells us whether the movie in general meets the expectations of the reviewer. Using intention mining, we can find the state of mind of the reviewer and understand the polarity and emotions. Polarity is positive, negative and neutral. Emotions are anger, disgust, fear, joy, sadness, surprise, and unknown. In this paper we use intention mining on a set of movie reviews extracted from twitter and try to understand the overall reaction about the movie. Whether people liked or they disliked it. We analyzed the movie reviews using machine learning methods Bayes and maximum entropy. We compare the two methods for accuracy.

2. RELATED WORK

Minqing Hu and Bing Liu [1], studied the problem of generating *feature-based summaries* of customer reviews of products sold online. Here, *features* broadly mean product features (or attributes) and functions. Given a set of customer reviews of a particular product, the task involves three subtasks: (1) identifying features of the product that customers have expressed their opinions on (called *product features*); (2)

for each feature, identifying review sentences that give positive or negative opinions; and producing a summary using the discovered information. Varsha D. Jadhav and S.N. Deshmukh [2] presented the Bayes model to predict the intentions of the people who tweet on twitter about a specific topic. They predicted the cricket match result which serves as strategic guidance to the captains so as to improve the performance of the team. Kamal Nigam et.al [3] proposes the use of maximum entropy techniques for text classification. In their text classification scenario, maximum entropy estimates the conditional distribution of the class label given a document. A document is represented by a set of word count features. The labeled training data is used to estimate the expected value of these word counts on a class-by-class basis. Improved iterative scaling finds a text classifier of an exponential form that is consistent with the constraints from the labeled data. Kuat Yessenov, et.al [4], presented an empirical study of efficacy of machine learning techniques in classifying text messages by semantic meaning. They used movie review comments from popular social network Digg as the data set and classify text by subjectivity/objectivity and negative/positive attitude. Changlin Ma, et.al [5], proposed a novel topic and sentiment unification maximum entropy LDA model in this paper for fine-grained opinion mining of online reviews. Oaindrila Das, et.al [6] presented a novel approach for classification of online movie reviews using parts of speech and machine learning algorithms. Borislav Kapukaranov and Preslav Nakov[7] presented experiments in predicting fine- rained stars, including halves, for Bulgarian movie reviews. This is a challenging task, that can be seen as (a) multi-way classification, i.e., choosing one out of eleven classes, (b) regression, i.e., predicting a real number, or (c) something in between, namely ordinal regression, i.e., predicting eleven values, but taking ordering into account, e.g., predicting 4 when the actual value is 3.5 would be better than predicting 1.

3. METHODOLOGY

Fig. 1 shows the framework for intention mining system.
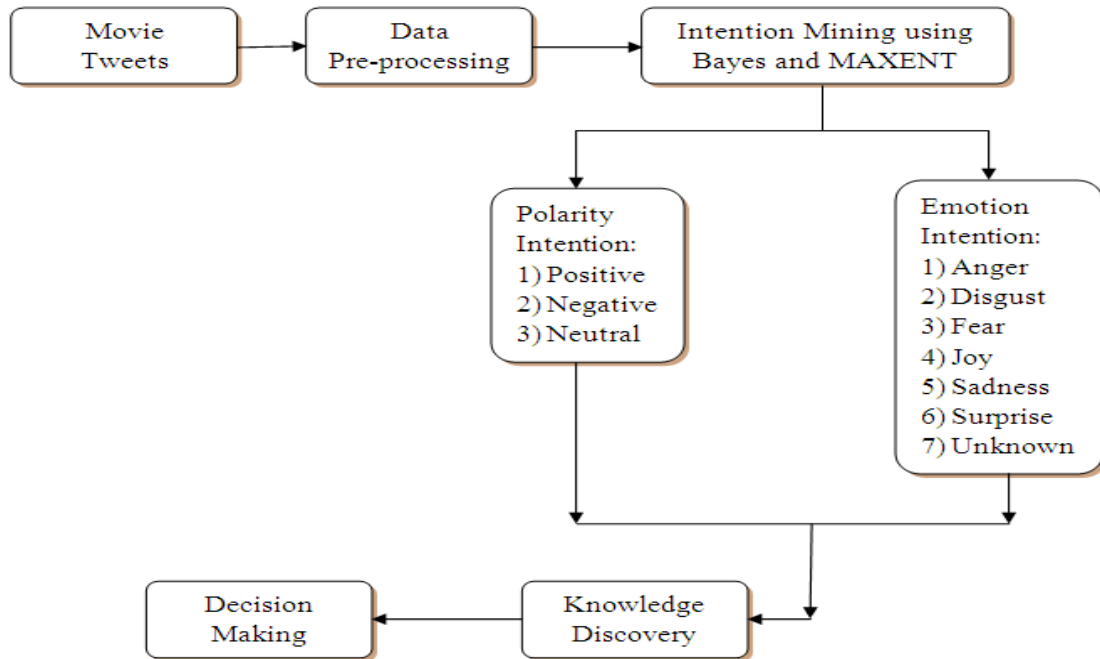
Fig.1 Framework for Intention Mining

The movie tweets are extracted using a keyword. The tweets are pre-processed which are classified for polarity and emotions using Bayes method and maximum entropy method. The intention is mined and decision is made.

4. TWITTER API

Twitter API is a twitter platform which connects the application with the world conversation happening on twitter. Once the twitter API is created we need to obtain credentials such as API keys, API secret, Access token and Access token secrete on the twitter developer site to access the twitter API. Twitter authentication is setup using these credentials. Using a keyword, tweets related to the specified keyword and language are searched. The language parameter that we have specified is English. The twitter library lets the use of Twitter API. The tweets retrieved are saved in .csv (comma separated file).

5. DATA COLLECTION

The tweets for the movie reviews were collected using the keyword 'Sultan movie,' using Twitter API from 5th July 11, 2016 to 9th July 2016. Total 34,694 tweets were retrieved for polarity and emotion intention. The tweets collected each day for the movie Sultan are shown in table 1. The movie Sultan was released on 6th July 2016. Also the tweets for the movie reviews using keyword 'Madaari movie,' was collected from 20th July to 24th July 2016. Total 8033 tweets were retrieved. Table 2 shows the tweets collected on each day for the movie Madaari. The movie Madaari was released on 22th July 2016.

Table 1. Daywise tweet collection for the movie 'Sultan'.

Date	Number of tweets collected
5th July 2016	6723
6th July 2016	6997
7th July 2016	6990
8th July 2016	6993
9th July 2016	6991
Total	**34,694**

Table 2. Daywise tweet collection for the movie 'Madaari'

Date	Number of tweets collected
20th July 2016	514
21th July 2016	736
22th July 2016	1518
23th July 2016	2233
24th July 2016	3032
Total	**8033**

6. DATA PREPROCESSING

The raw data is pre-processed. The pre-processing includes

6.1 Tokenization

Text data is collection of sentences which is split into terms or tokens by removing white spaces, commas and other symbols etc.

6.2 Stopword removal

Removes articles such as (a, an, the).

6.3 Stemming

is the process of reducing related word to the root word. E.g. generalization is generally represented as general.

6.4 Case Normalization

English text contains both higher and lower case characters. This process converts the complete sentences in lowercase or uppercase.

6.5 Usernames

removes @ symbol. Users often include twitter usernames by using @ symbol.

6.6 Usage links

Retweeting is copying another user's tweet and posting to another users account. It is abbreviated with "RT". All tweets with RT are removed.

6.7 Usage of N-gram and part of speech (POS) tags

An n-gram is a N-character slice. N-gram of several different lengths are used simultaneously. POS feature is used because depending on the usage the same word may have different meaning.

7. BAYES METHOD

Whether a review is negative, positive or neutral is predicted using the text. In order to do this, train an algorithm is trained using the reviews and classify to make predictions on the reviews. Error is calculated to see how good the predictions are.

Bayes classifier is used for classification. A Bayes classifier works by figuring out the probability of different attributes of the data being associated with a certain class. This is based on Bayes theorem. The theorem is

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

8. MAXIMUM ENTROPY METHOD

The maximum entropy classifier is a discriminative classifier commonly used in natural processing and information retrieval problems.

The Max Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. Unlike the Naïve Bayes classifier, the Max entropy does not assume that the features are conditionally independent of each other. It is based on the principle of maximum entropy and from all the models that fit the training data, selects the one which has the largest entropy. The Entropy can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more.

Principle of MaxEnt states, take precisely stated prior data or testable information about probability function. Consider the set of all trial probability distribution that would encode the prior data. Of those the one with maximal information entropy is the proper distribution.

The model is represented as

$$P(c \mid d, \lambda) = \frac{1}{Z(d)} \exp[\sum_i \lambda_i f_i(c,d)]$$

P(c|d) is the probability that a class occurs for a given tweet, c is the class, d is the tweet, and λ is a weight vector. The weight vector decide the weight vector decide the significance of the feature in classification. Z(d) is a normalization function, f_i is a feature class function.

9. NORMALIZATION

Data should be normalized to bring all the variables into proportion with one another when doing comparison analysis. The aim of normalization is to make variables comparable to each other. Normalization means to transform observations x into f(x) such that they look normally distributed. It simply means putting different variables on common scale.

As seen from table 1. The number of tweets collected each day is different. The number of positive, negative and neutral polarities each day is different so we need to normalize it.

10. ACCURACY MEASURES

The accuracy measures such as

   i.   Mean error (ME)
  ii.   Root mean square error (RMSE)
 iii.   Mean absolute error (MAE)
 iv.   Mean Percentage Error (MPE)
  v.   Mean Absolute Percentage Error (MAPE)

are used to measure the accuracy. MAE is simply the mean of the absolute errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. MAE tells us how big of an error we can expect from the forecast on average. Cort J. Willmott et.al [8] indicated that MAE is the most natural measure of average error magnitude than RMSE.

11. RESULTS AND DISCUSSION

The results are obtained for Bayes and Maximum Entropy classifiers. Movie reviews of two movies, Sultan and Madaari were retrieved of which polarity and emotion intentions were obtained. The daywise polarity intention and the daywise emotion intention are obtained.

Table 3. shows the daywise intention polarity using Bayes and MAXENT methods for the movie 'Sultan.' The polarity is positive, negative and neutral.

The graph for daywise polarity intention using Bayes and MAXENT is shown in Fig.2. Using the graph the polarity intention can be easily seen and the polarity intention for positive, negative and neutral for both the classifiers, that is Bayes and MAXENT can be compared.
The numbers of positive intentions are more than that of the negative and neutral intentions each day.

Table 3.Daywise intention polarity using Bayes and MAXENT methods for the movie 'Sultan.'

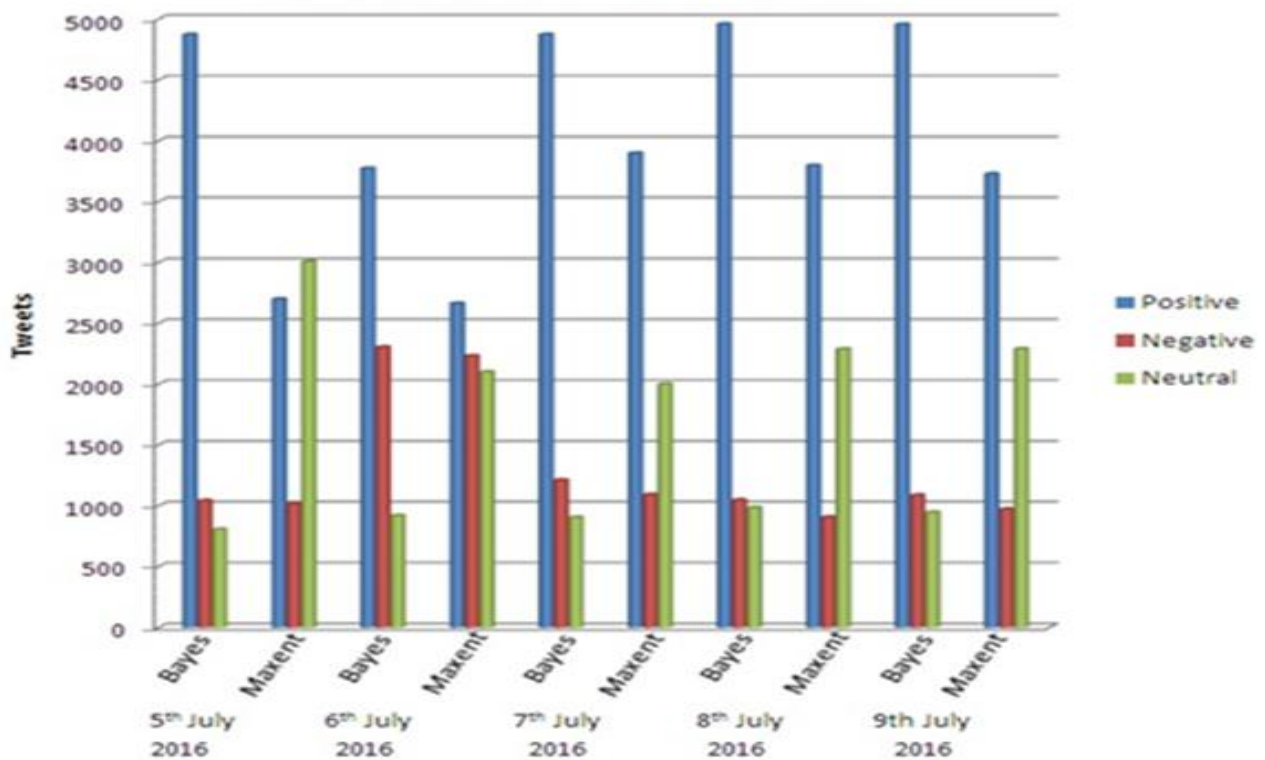| Polarity	5th July, 2016		6th July, 2016		7th July, 2016		8th July, 2016		9th July, 2016	
	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT
Positive	4874	2699	3774	2664	4873	3898	4959	3797	4957	3728
Negative	1043	1017	2304	2233	1212	1093	1048	908	1087	973
Neutral	806	3007	919	2100	905	1999	986	2288	947	2290
Total	**6723**	**6723**	**6997**	**6997**	**6990**	**6990**	**6993**	**6993**	**6991**	**6991**

Fig. 2. Daywise Intention Polarity for the movie Sultan

The positive intention polarity for 'Sultan' for Bayes method is more than the maximum entropy method each day.Table 4. shows the daywise intention emotions using Bayes and MAXENT methods for the movie 'Sultan.'

Using Bayes method we can see that the number of joy emotions is more. Figure 3 shows the intention emotions for the movie Sultan. Unknown emotion is the limitation of our research. The number of joy emotion is the highest.

Table 4.Daywise intention emotions using Bayes and MAXENT methods for the movie 'Sultan'

| Emotions	5th July, 2016		6th July, 2016		7th July, 2016		8th July, 2016		9th July, 2016	
	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT
Anger	75	5347	139	5066	160	4601	60	4515	128	4629
Disgust	9	0	3	0	0	0	3	0	6	0
Fear	44	44	46	46	28	28	13	13	6	6
Joy	1081	1081	1462	1462	1625	1625	1980	1980	1618	1618
Sadness	160	160	114	114	52	52	137	137	251	251
Surprise	82	82	306	306	684	684	345	345	481	481
Unknown	5272	9	4927	3	4441	0	4455	3	4501	6
Total	**6723**	**6723**	**6997**	**6997**	**6990**	**6990**	**6993**	**6993**	**6991**	**6991**
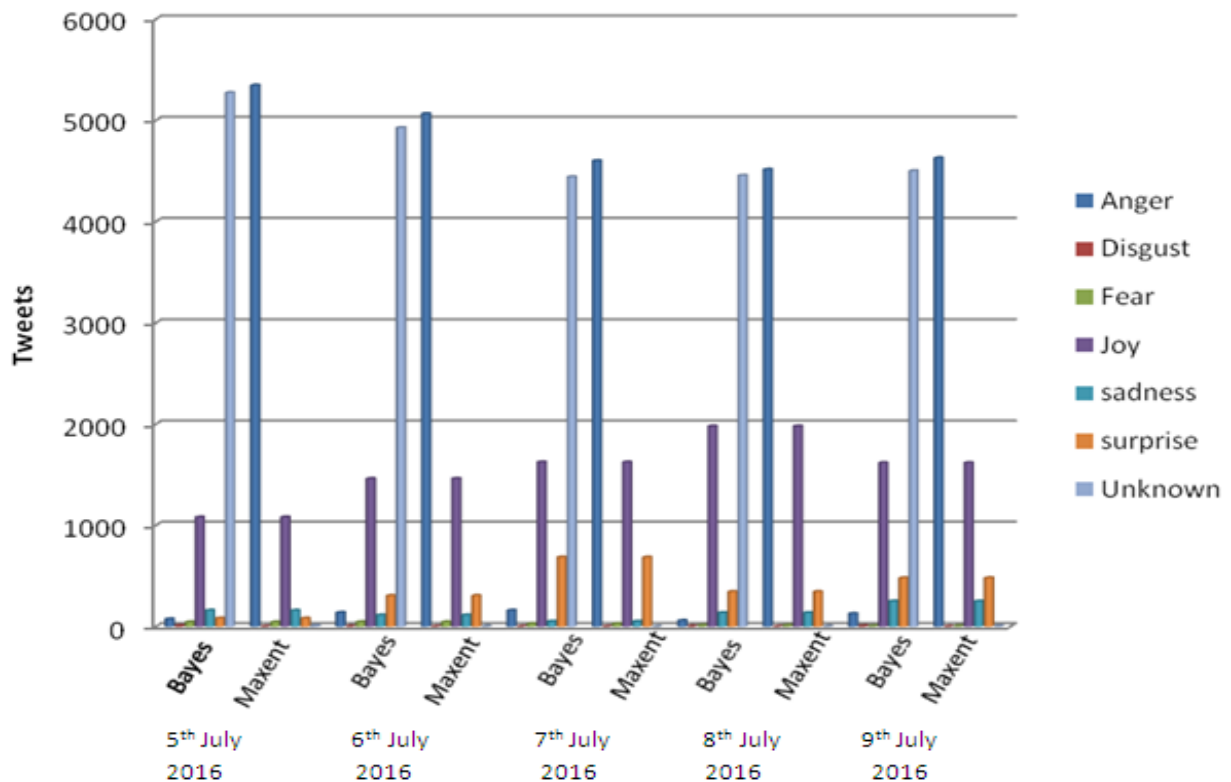
Fig 3. Daywise Intention Emotion for the movie Sultan

Table 5 shows the Mean Absolute Error and the accuracy using Bayes and MAXENT methods.

Figure 4 shows the Accuracy for Bayes and MAXENT methods

Table 5.Mean absolute error and accuracy using Bayes and MAXENT for the movie Sultan

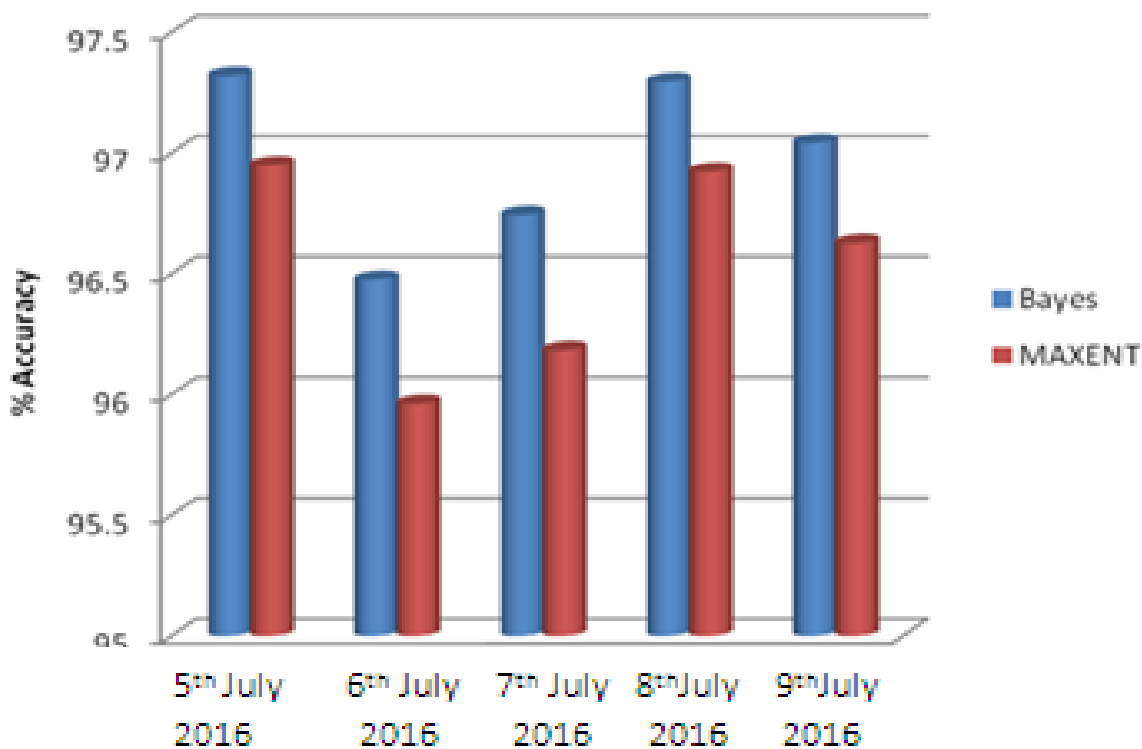| Date	Mean Absolute Error		% Accuracy	
	Bayes	**MAXENT**	**Bayes**	**MAXENT**
5th July 2016	2.6801	3.0515	97.3198	96.9484
6th July 2016	3.5256	4.0404	96.4743	95.9595
7th July 2016	3.2571	3.8182	96.7428	96.1817
8th July 2016	2.7058	3.0779	97.2941	96.9220
9th July 2016	2.9587	3.3732	97.0412	96.6267

Fig. 4. Accuracy for Bayes and MAXENT for the movie Sultan

The accuracy for Bayes method is more on each day as compared to maximum entropy. But the number of tweets collected each day is different. So the data needs to be normalized. The positive, negative and neutral polarities obtained needs to be normalized. Normalization is performed by adding all the positive polarities obtained on each day and divide it by the total number of tweets retrieved.

Normalization is performed for Bayes and MAXENT methods and then compared to get the best results.

11.1 Normalization of data for the movie Sultan

11.1.1 For Bayes method

Let P_1 be the sum of all positive intentions.

$$P_1 = 23437$$

Normalization factor for positive intention

$$= \frac{23437}{34697} = 0.6754 \qquad (1)$$

Let N_1 be the sum of all Negative intentions.

$$N_1 = 6694$$

Normalization factor for Negative intention

$$= \frac{6694}{34697} = 0.1929 \qquad (2)$$

Let NE_1 be the sum of all Neutral intentions.

$$NE_1 = 4563$$

Normalization factor for Negative intention

$$= \frac{4563}{34697} = 0.1315 \qquad (3)$$

11.1.2 For MAXENT method

Let P_2 be the sum of all positive intentions.

$$P_2 = 16786$$

Normalization factor for positive intention

$$= \frac{16786}{34697} = 0.4837 \qquad (4)$$

Let N_2 be the sum of all Negative intentions.

$$N_2 = 6224$$

Normalization factor for Negative intention

$$= \frac{6224}{34697} = 0.1793 \qquad (5)$$

Let NE_2 be the sum of all Neutral intentions.

$$NE_2 = 11684$$

Normalization factor for Negative intention

$$= \frac{11684}{34697} = 0.3367 \qquad (6)$$

The normalization factor for positive intention of Bayes method is more than the normalization factors of negative and neutral intention of the Bayes method. Also the normalization factor of positive intention of the MAXENT method is more than the normalization factor of negative and neutral intention of the MAXENT method. When comparing Bayes and MAXENT the normalization factor of positive intention of Bayes method is highest than that of the positive intention of MAXENT method.Table 6 shows the average of positive tweets for the movie Sultan

Table 6. Average for positive tweets for Bayes and MAXENT methods for Sultan

Date	Bayes		MAXENT	
	No. of Positive tweets	Average	No. of Positive tweets	Average
5th July 2016	4874	72.4973	2699	40.1457
6th July 2016	3774	53.9374	2664	38.0734
7th July 2016	4873	69.7138	3898	55.7653
8th July 2016	4959	70.9137	3797	54.2971
9th July 2016	4957	70.9054	3728	53.3257
Total	**23,437**	**337.9678**	**16,786**	**241.6074**

The Accuracy for all the 5 days is more for Bayes method. Considering the Bayes method alone, the numbers of positive tweets are more than negative and neutral tweets. Also, considering MAXENT method alone the numbers of positive tweets are more than that of negative and neutral tweets. Making comparison of Bayes and MAXENT the number of positive tweets for Bayes method are more than the number of positive tweets of MAXENT method every. Also the average of positive for Bayes and MAXENT methods for each day is calculated, as shown in table 6. Again calculating the average of total average for both Bayes and MAXENT it gives

67.5935% for Bayes method and 48.3214% for MAXENT method.

Table 7 shows the daywise intention polarity for the movie 'Madaari.' Figure 5 shows the graph for the daywise polarity. As seen from the graph the numbers of positive intentions are more than that of negative and neutral intentions for Bayes approach. For the maximum entropy method the numbers of neutral intentions were more before the movie was released. The movie was released on 22nd July 2016. On 22nd,23rd and24th July 2016 the number of positive intentions are more. While comparing Bayes and maximum entropy methods the

number of positive intentions obtained by Bayes method is more than that of the positive intentions obtained by maximum entropy method. Table 8 shows the intention emotions for the movie Madaari, and figure 6 shows the graph for the intention emotions. For Bayes method the numbers of joy emotions are highest. Unknown is the limitation of our research. For MAXENT method the numbers of anger intentions are more. But on the contrary the positive intention polarity if more, so we neglect it and consider other emotions. Making this assumption, the numbers of joy emotions are highest.

Table 7.Daywise intention polarity using Bayes and MAXENT methods for the movie 'Madaari.'

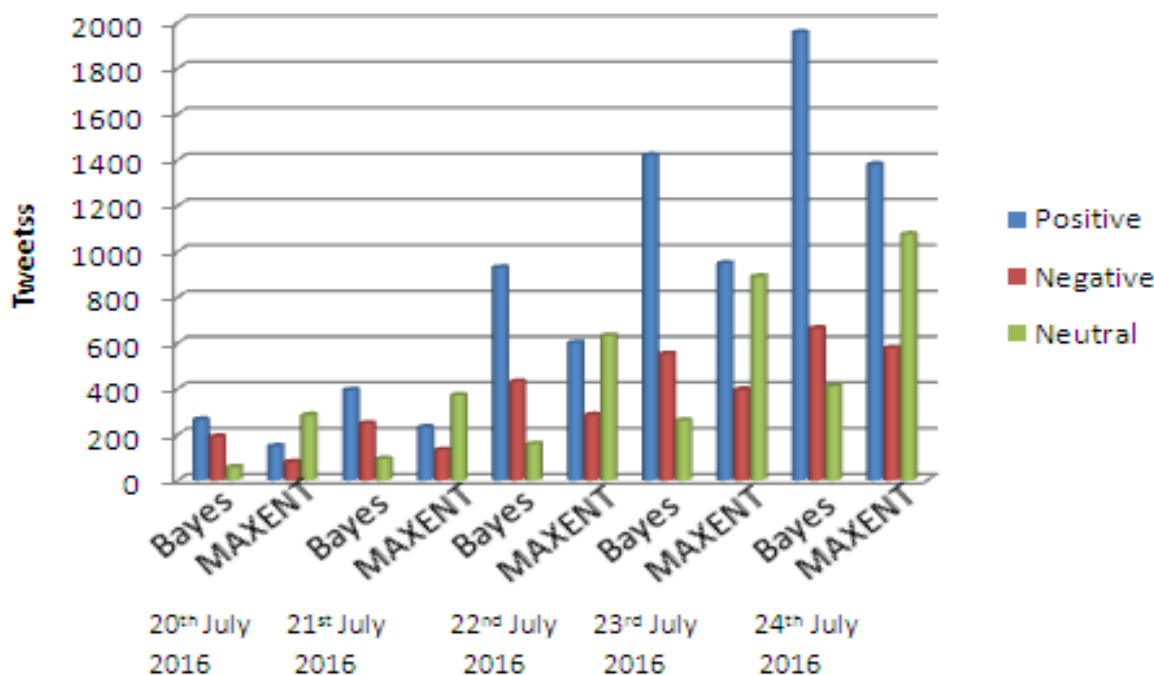| Polarity	20th July, 2016		21th July, 2016		22th July, 2016		23th July, 2016		24th July, 2016	
	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT
Positive	265	149	394	232	929	601	1420	947	1956	1379
Negative	191	80	248	132	431	285	552	396	664	578
Neutral	58	285	94	372	158	632	261	890	412	1075
Total	**514**	**514**	**736**	**736**	**1518**	**1518**	**2233**	**2233**	**3032**	**3032**

Fig. 5. Twitter intention polarity for the movie 'Madaari'

Table. 8. Daywise Intention Emotions for the movie Madaari

| Emotions	20th July, 2016		21th July, 2016		22th July, 2016		23th July, 2016		24th July, 2016	
	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT	Bayes	MAXENT
Anger	13	343	13	534	19	1140	32	1680	46	3217
Disgust	18	0	0	0	0	0	0	0	0	0
Fear	0	0	0	0	1	1	33	33	37	37
Joy	0	18	33	33	145	145	233	233	343	343
Sadness	152	152	162	162	178	178	197	197	187	187
Surprise	1	1	7	7	54	54	90	90	148	148
Unknown	330	0	521	0	1121	0	16487	0	2271	0
Total	**514**	**514**	**736**	**736**	**1518**	**1518**	**2233**	**2233**	**3032**	**3032**
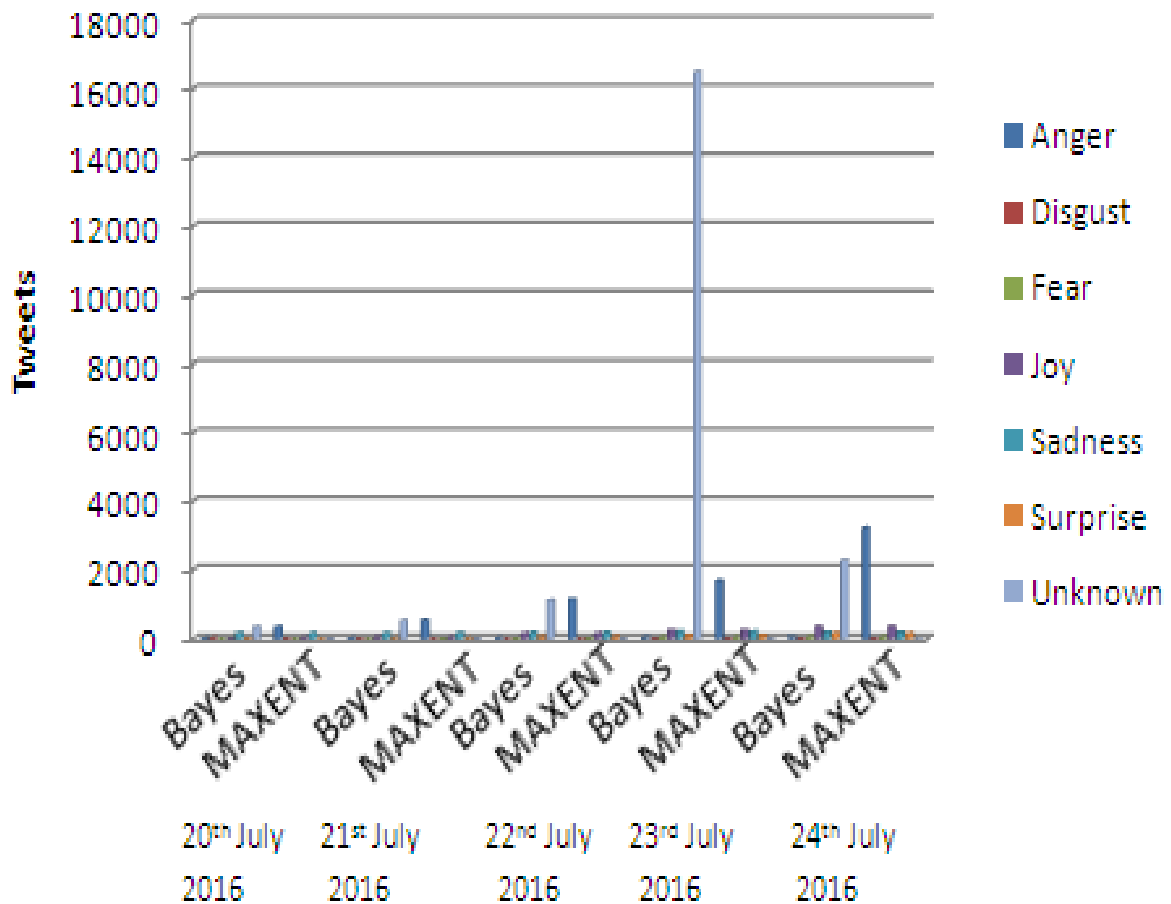
Fig.6. Daywise Intention Emotion for the movie Madaari

Table 9 shows the Mean Absolute Error and the accuracy using Bayes and MAXENT methods for the movie 'Madaari.'

Figure 7 shows the graph of Accuracy for Bayes and MAXENT methods.

Table 9.Mean absolute error and accuracy using Bayes and MAXENT for the movie Madaari

Date	Mean Absolute Error		% Accuracy	
	Bayes	**MAXENT**	**Bayes**	**MAXENT**
20th July 2016	2.0381	2.4320	97.961	97.5679
21th July 2016	1.9463	2.3186	98.0536	97.6813
22th July 2016	1.9486	2.2427	98.0513	97.5727
23th July 2016	1.9440	2.2383 98.0559	98.0559	97.7616
24th July 2016	1.9496	2.2469	98.0503	97.7530

The accuracy for Bayes method is more each day as compared to that of the Maximum Entropy method.
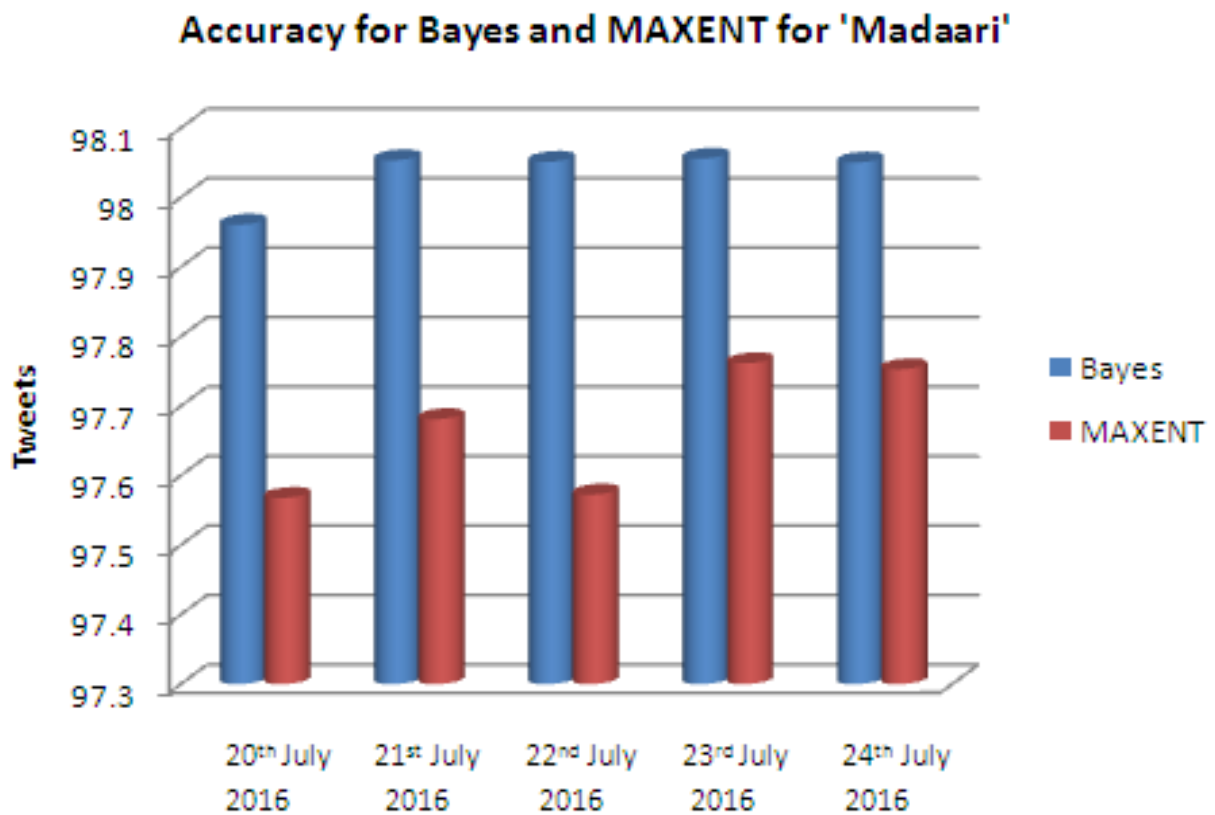

Fig. 7. Accuracy for Bayes and MAXENT for the movie Madaari

The number of tweets collected each day is different. The data needs to be normalized for each day as well as for the

positive, negative and neutral polarities. The total numbers of tweets collected for the movie Madaari are 8033.

11.2 Normalization of data for the movie Madaari

11.1.1 For Bayes method

Let P_3 be the sum of all positive intentions.

$$P_3 = 4955$$

Normalization factor for positive intention

$$= \frac{4955}{8033} = 0.6168 \qquad (7)$$

Let N_3 be the sum of all Negative intentions.

$$N_3 = 2068$$

Normalization factor for Negative intention

$$= \frac{2068}{8033} = 0.2574 \qquad (8)$$

Let NE_1 be the sum of all Neutral intentions.

$$NE_3 = 683$$

Normalization factor for Negative intention

$$= \frac{683}{8033} = 0.0850 \qquad (9)$$

11.1.2 For MAXENT method

Let P_2 be the sum of all positive intentions.

$$P_4 = 1379$$

Normalization factor for positive intention

$$= \frac{1379}{8033} = 0.1716 \qquad (10)$$

Let N_2 be the sum of all Negative intentions.

$$N_4 = 578$$

Normalization factor for Negative intention

$$= \frac{578}{8033} = 0.0719 \qquad (11)$$

Let NE_2 be the sum of all Neutral intentions.

$$NE_4 = 1075$$

Normalization factor for Negative intention

$$= \frac{1075}{8033} = 0.1338 \qquad (12)$$

Considering Bayes method, the normalization factor for positive intention is more than the negative and neutral intentions. Also, for MAXENT method, the normalization factor for positive intention is more than that of the negative and neutral intentions. Comparing Bayes and MAXENT the normalization factor of positive intention of bayes method is highest than the normalization factor of positive intention of MAXENT method.

Table 10 shows the average of positive tweets for the movie Sultan

Table 10. Average for positive tweets for Bayes and MAXENT methods for Madaari

Date	Bayes		MAXENT	
	No. of Positive tweets	Average	No. of Positive tweets	Average
20th July 2016	256	49.8054	149	28.9883
21th July 2016	394	53.5326	232	31.5217
22th July 2016	929	61.1989	601	39.5915
23th July 2016	1420	63.5915	947	42.4093
24th July 2016	1956	64.5118	1379	45.4815
Total	**4955**	**292.6402**	**3308**	**187.9923**

The accuracy for Bayes method is more than the accuracy of MAXENT method. Again calculating average of the total average of both Bayes and MAXENT methods, we get 67.5534% for Bayes method and 48.3830% for MAXENT method.

12. CONCLUSIONS

In comparison of machine learning classifiers, Bayes and Maximum Entropy, mining intention for movie review prediction, the Bayes classifier is more efficient than the MAXENT. 67.5935% accuracy is obtained for movie 'Sultan'

using Bayes method and 48.3214% using MAXENT method. The movie Sultan was released on 6th July 2016. According to the Hindi movie reviews of Times of India dated 6th July 2016 the review rating is 3.5 out of 5, which is 70%. Prediction using Bayes and MAXENT is close to the Hindi movie reviews of Times of India. For the movie Madaari Bayes classifier is giving 67.5935% and MAXENT classifier is giving 48.3830% accuracy. The movie Madaari was released on 22nd July 2016. According to the Hindi movie reviews of Times of India dated 22nd July 2016 the review rating is 3.5 out of 5, which is 70%. Prediction using Bayes

classifier for the movie Madaari is close to the Hindi movie reviews of Times of India. Also the normalization factor for the positive intention for Bayes method is highest than that of MAXENT method for both Sultan and Madaari movies. It is 0.6754 for Sultan movie and 0.6168 for Madaari movie. Hence, Bayes method is efficient than Maximum entropy method for text classification.

13. FUTURE WORK
More machine learning classifiers can be used to analyze text data. Also anger and unknown emotions can be further analyzed to get more accurate results.

14. ACKNOWLEDGMENTS

15. REFERENCES
[1] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews," *DD'04*, August 22–25, 2004, Seattle, Washington, USA.

[2] Varsha D Jadhav, S.N. Deshmukh, "Twitter Intention Classification Using Bayes Approach for Cricket Test Match Played Between India and South Africa 2015," Proceeding of International Conference on Internet of Things, Next Generation Network and Cloud Computing 2016. ISSN: *0975 – 8887,*

[3] Kamal Nigam,John Lafferty, Andrew McCallum, "Using maximum entropy for text classification," In IJCAI-99 Workshop on Machine Learning for Information Filtering.

[4] Kuat Yessenov, Sasa Misailovic, Sentiment Analysis of Movie Review Comments," 6.863 Spring 2009 final project, May 17, 2009.

[5] Changlin Ma, Meng Wang, Xuewen Chen, "Topic and Sentiment Unification Maximum Entropy Model for Online Review Analysis," *WWW 2015 Companion,* May 18–22, 2015, Florence, Italy.ACM 978-1-4503-3473-0/15/05.

[6] Oaindrila Das, Rakesh Chandra Balabantaray, "Sentiment Analysis of Movie Reviews using POS tags and Term Frequencies," *International Journal of Computer Applications (0975 – 8887) Volume 96– No.25, June 2014*.

[7] Borislav Kapukaranov, Preslav Nakov, "Fine-Grained Sentiment Analysis for Movie Reviews in Bulgarian," Proceedings of Recent Advances in Natural Language Processing, pages 266–274,Hissar, Bulgaria, Sep 7–9 2015.

[8] Cort J. Willmott, Kenji Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error RMSE) in assessing average model Performance, *Climate Research, Vol. 30: 79–82.*

[9] http://timesofindia.indiatimes.com, Hindi movie reviews.