International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-6, Issue-1, September 2016

Sentiment Analysis on Unstructured Social Media Data Compare with Different Classification Algorithms

Vijay Kumar Mishra, Dr. Neelendra Badal

***Abstract*— Presently, analysis of opinions from social media (Twitter) has become very popular merely because such an amount of views is difficult to extract through any other existing custom means of collecting views like surveys, polls etc. The analysis is interesting but at the same time challenging because of a micro blog post generate information on Internet and number of opinions can be expressed which are usually very short and colloquial and choosing the best opinion mining algorithms is very difficult in such type of text. Therefore, we propose a new system which automatically analyzes the sentiments of these types of messages using major classification algorithms. Here we consider Twitter for the task of sentiment analysis which can play significant role on the popularity of the products and services. Hence, an accurate method for predicting sentiments could enable us to understand customers' preferences, their views on the product and services offered by the companies. Companies and organization can use this information to formulate the future planning. These can possibly make positive or negative wave in the business sectors as well as social sectors. The primary center of my proposed work is to analyze the emotions communicated on social networking Twitter so that peoples' feelings, habits and choices are extracted, investigated and used to understand the behavior of the people.**

***Index Terms*— Twitter, Sentiment Analysis, Machine Learning Algorithms, Movies and Songs.**

I. INTRODUCTION

With the rapid growth of mobile information systems and the increased availability of smart phones, social media has become a integral part of daily life in most societies. This improvement has involved the creation of huge amounts of information: information which when analyzed can be used to extract valuable information about a variety of subjects.

People are able to gather the relevant information and are able to share the same on social web. The Web is a virtual environment where people are able to put their experience about the products before buying, the perception of being present rather than actually being there in a real environment. Marketing communication strategies have always been around. It combines the prospect of overcoming public resistance with significantly lower costs and faster delivery. A large no people are now availing access to social media to get and exchange information. Both buyers and sellers can

mutually create and promote 'brands' to benefit one another. Social media has evolved to be the center for instant sharing of information. The information can shared millions of people, who are connected through social media. The power of its reach to large number of people immediately and the openness in sharing experiences, without fear, has emerged the valuable suggestion of business and social entities. Social media has today served as a catalyst for on-line chat where individuals create content, share it, bookmark it and network at a rapid rate. Social media is fast changing the public opinions in society and setting trends and agendas in topics that range from nature and governmental issues to innovation and the diversion business. Since social media can also be inferred as a form of collective wisdom, we decided to inquire its power at predicting real-world outcomes. Amazingly we discovered that the communication of people can indeed be used to make quantitative forecast that beat those of fake markets. These data advertises for the most part include the exchanging of state-unexpected securities, and if sufficiently huge and legitimately planned, they are typically more exact than different systems for removing diffuse data, for example, overviews and suppositions surveys. In particular, the costs in these business sectors have been appeared to have solid connections with watched result frequencies, and in this manner are great pointers of future results. [1].

In the case of social media, the volume and high fluctuation of the data that produce through huge client groups introduces an open door for bridling that information into a structure that consider particular expectations about specific results, without instituting market components. One can also build models to collect the opinions of the aggregate population and gain useful insights into their behavior, while predicting future trends. Moreover, gathering information on how individuals talk with respect to specific products can be useful when designing marketing and promoting efforts [2].

This paper is the classification of sentimental movie review using to experiment using different machine learning algorithms to predict the sentiment of movie and songs reviews. The purpose of this work is to analyze the tweets of movies and songs. This paper presents the results of the research work performed on Twitter to classify Twitter messages of movies and songs as positive sentiment, negative sentiment and neutral statements.

Vijay Kumar Mishra,Assistant Professor, Department of Computer Application, Feroze Gandhi Institute of Engineering And Technology,,RaeBareli , India.

Dr. Neelendra Badal, Associate Professor, Department of Computer Science and Engineering, Kamla Nehru Institute of Technology,Sultanpur , India.

II. Literature Review

Although Twitter has been very popular as a web service, there has not been considerable published research on it. Huberman and others [3] studied the social interactions on Twitter to reveal that the driving process for usage is a sparse hidden network underlying the friends and followers, while most of the links represent meaningless interactions. Java et al [4] investigated community structure and isolated different types of user intentions on Twitter. Jansen and others [5] have examined Twitter as a mechanism for word-of-mouth advertising, and considered particular brands and products while examining the structure of the postings and the change in sentiments. However the authors do not perform any analysis on the predictive aspect of Twitter.

There has been some prior work on analyzing the correlation between blog and review mentions and performance. Gruhl and others [6] showed how to generate automated queries for mining blogs in order to predict spikes in book sales. And while there has been research on predicting movie sales, almost all of them have used meta-data information on the movies themselves to perform the forecasting, such as the movies genre, MPAA rating, running time, release date, the number of screens on which the movie debuted, and the presence of particular actors or actresses in the cast. Joshi and others [7] use linear regression from text and metadata features to predict earnings for movies. Mishne and Glance [8] correlate sentiments in blog posts with movie box-office scores. The correlations they observed for positive sentiments are fairly low and not sufficient to use for predictive purposes. Sharda and Delen [9] have treated the prediction problem as a classification problem and used neural networks to classify movies into categories ranging from 'flop' to 'blockbuster'. Apart from the fact that they are predicting ranges over actual numbers, the best accuracy that their model can achieve is fairly low. Zhang and Skiena [10] have used a news aggregation model along with IMDB data to predict movie box-office numbers. We have shown how our model can generate better results when compared to their method.

Akcora et al., [11] in their experiment, try to identify the emotional pattern and the word pattern that claims to change the public opinion, using Twitter data. To identify the breakpoint, researchers use Jaccard's similarity of two successive intervals of words. Sun et al., [12] study fan pages on Facebook to understand diffusion trees. Kwak et al., [13] compare the number of followers, page-ranks and the number of re-tweets as three different measures of influence. Their finding is that the ranking of the most influential users differed depending on the measure.

Movies and songs are released across different parts of the world and Twitter users are also from different parts of the world. The study aims to analyze whether an attitude in the tweets is positive or negative sentiment or a cognitive statement, understand the flow of interpersonal messages across different countries and understand people behavior.

Many previous studies have shaped the goal of our work. Ishii et al. developed a mathematical framework for the spread of popularity of the movie in society. Their model considers the activity level of the bloggers estimated through number of weblog posts on particular movies in the Japanese Blogosphere as a representative parameter for social popularity. Similarly, other researchers have developed models taking the activity level of editors on Wikipedia as a popularity parameter [14].

III. Proposed Methodology

In the case of social media, the outrageousness and high variance of the information that propagates through large user communities presents an interesting opportunity for using that data into a form that allows for specific predictions about particular results, without having to institute market mechanisms. Specifically we consider the task of predicting box-office sentiments for movies and song using the information post on Twitter, one of the fastest growing social media in the Internet. Twitter, a micro-blogging social media site, has experienced a full of popularity in recent years leading to a huge user-base, consisting of several tens of millions of users who actively participate in the creation and propagation of infofmation. Our Methodology divided in several blocks.

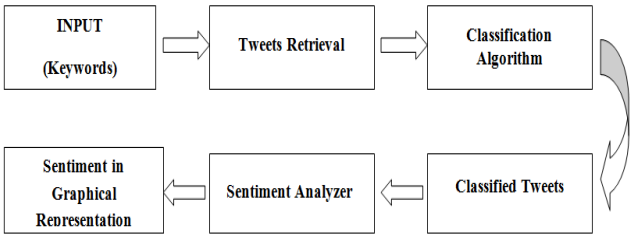

Figure 1 Proposed system block diagram

A. Data Set

The dataset that we used was obtained by crawling hourly feed data from Twitter.com. To ensure that we obtained all tweets referring to a movie and songs, we used keywords present in the movie title and songs title as search arguments. We extracted tweets over frequent intervals using the Twitter Search Api., thereby ensuring we had the timestamp, author and tweet text for our analysis. With an intention to discover the social media signals that potentially possess stronger correlation with the profitability of a film; we identify signals which reflect audience approval from different social media domains. Note that the type of data in each domain may be different, e.g., Twitter is a social stream whereas YouTube is a social video publishing website. Most of social media buzz around a movie that have been captured for this study are before its release or during the first 1-2 weeks.

B. Classification of Data

Data can be classified into different categories depending on the domain i.e.

- Political, Business, Sports etc.
- Recommendations, Complaints
- Positive, Negative

Lexicon based classification

- Requires a dictionary of words and their polarity scores.

Supervised Learning classification

- Requires training data to create a classification model.

Various visualization techniques

- WEKA can be applied to classification of data.

C. System Process

Opinion mining or sentiment analysis is the process of determining the feelings expressed by an individual in his writing. There are two basic methods that exist; the first is the document level and the second is the sentence level. In the document level, the analysis is based on the complete document, where as in the sentence level, the analysis is performed at the sentence level. Since tweets comprise only 140 characters, we have used the methods that are related to the sentence level. The following tools and methods will be developed to address the objective:

- An in-house tool using Twitter API will be developed to download related tweets from Twitter database in an automated manner.
- A Sentiment Analyzer tool will be developed using python/java and natural language tool kit libraries by trying different supervised machine learning algorithms. The best classifier model will be chosen for the final classification.
- The collected tweets are classified, using the above classifier, into three different classes – positive sentiment, negative sentiment and natural statement.
- Unwanted tweets that fell into none of the above mentioned classes are classified using any filter.

This project is the classification of sentimental movie review using to experiment using machine learning algorithms to predict the sentiment of movie reviews. This prediction was performed on the movie review dataset [15].

IV. Machine Learning Algorithms

D. Support Vector Machine

Support vector machine is a supervised learning model, very similar to linear regression, which analyzes data, recognizes patters and uses these results for predictions. Here we used the SVM library to use the complete functionality of the algorithms. The advantage of Support Vector Machines is that they can make use of certain kernels in order to transform the problem, such that we can apply linear classification techniques to non-linear data.

E. J48 Decision Tree

J48 Decision tree is a predictive machine-learning model that decides the target value of a new sample based on various attribute values of the available data.

F. Logistic Regression

Logistic Regression is one of the best probabilistic classifiers, measured in both log loss and first-best classification accuracy across a number of tasks. The dimensions of the input vectors being classified are called "features" and there is no restriction against them being correlated.

II. Training and Testing the Classifiers

Consumer behavior involves the thoughts and feelings people experience and the actions they perform in the consumption and usage of a product. Consumer thinking, feelings, actions are constantly changing. The Wheel of Consumer Behavior is critical for developing a complete understanding of consumers and selecting strategies to influence them. In order to understand how consumers talk about the movies, sentiments are measured based on the following classification: Positive Sentiment, Negative Sentiment and Cognitive Statement.

Here for preprocessing, filtering, visualization Weka API is used, it is a popular suite of machine learning software written in Java. All algorithms that software provides can be used directly from code by importing *weka.jar* file. Weka contains tools for data preprocessing, classification, regression, clustering and visualization.

Precision represent proportion of samples that are truly of a class divided by the total sample classified as that class.

Recall represents proportion of samples classified as a given class divided by the actual total in that class.

F-Measure which is a combined measure for precision and recall calculated as

F-measure = 2 * Precision * Recall / (Precision + Recall)

III. Experimental Analysis and Results

The results are used to compare with machine learning algorithms of sentimental analysis of social data of movies and songs. For the better results, we have to measure its performance by applying suitable measures.

This work, we have used to different machine algorithms. These algorithms are used to sentiment analysis of tweeted data of movies and songs. For sentiment analysis, we search the movie name as like *X-man*, and then we recolonized the tweet data about movie and song.

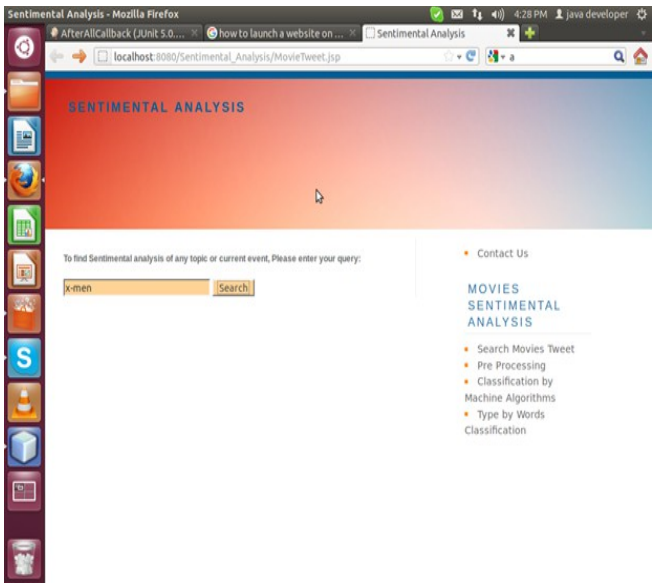

Figure 2 Entered Movies Name

After collection of tweet data, we applied machine algorithms for analyzing popularity.

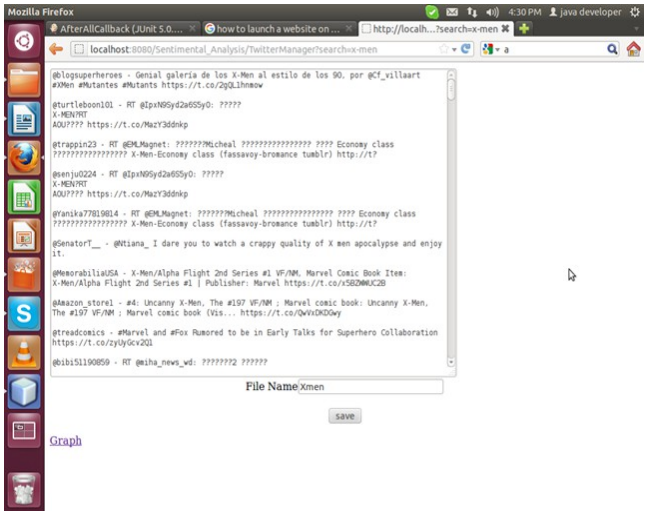

Figure 3 Output of Tweet data

This work applied as different sentiment algorithms for analysis. For analytical comparisons, we make use of Pie Chart and Bar Graph.

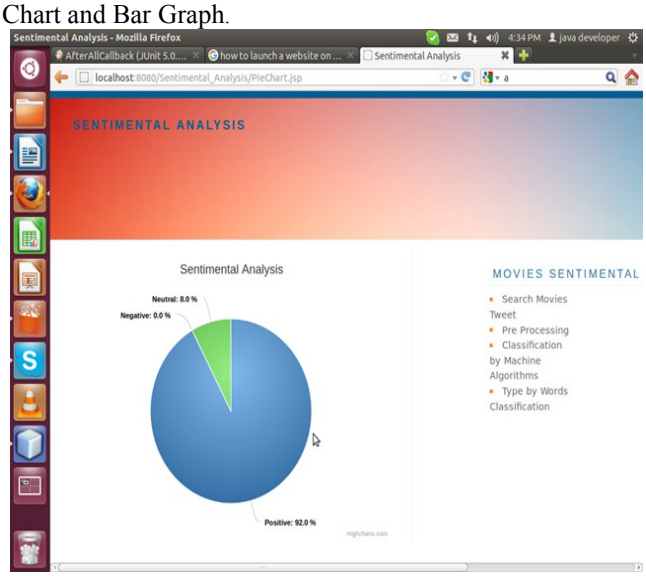

Figure 4 Basic Pie-Chart without apply any algorithm

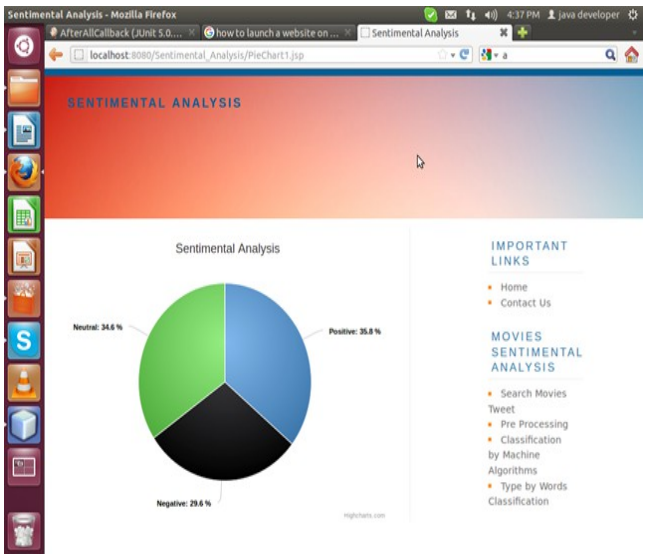

Figure 5 Pie-Chart after applying J48 algorithm

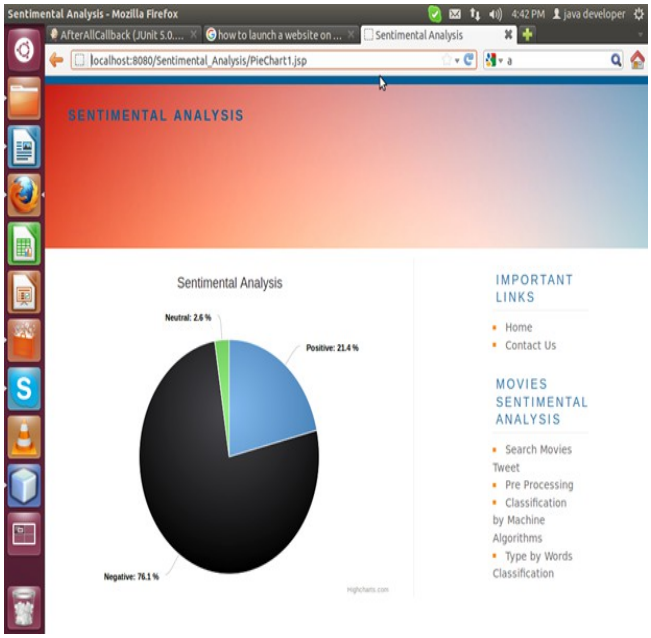

Figure 6 Pie-Chart after applying Logicstic reression

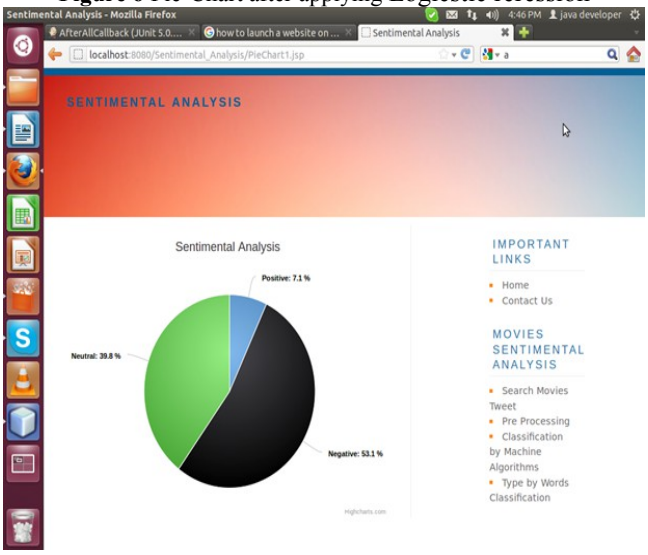

Figure 7 Pie-Chart after applying SVM Algorithm

Performance compression of SVN, Logistic Regression and J48 Decision tree algorithms using WEKA:

Table 1 Performance compression of algorithms with different parameters

Parameters	SVM	Logistic Regression	J48
TP Rate	0.501	0.752	0.372
FP Rate	0.11	0.350	0.262
Precision	0.82	0.752	0.362
Recall	0.501	0.752	0.372
F- Measure	0.622	0.752	0.37
ROC Area	0.696	0.807	0.552

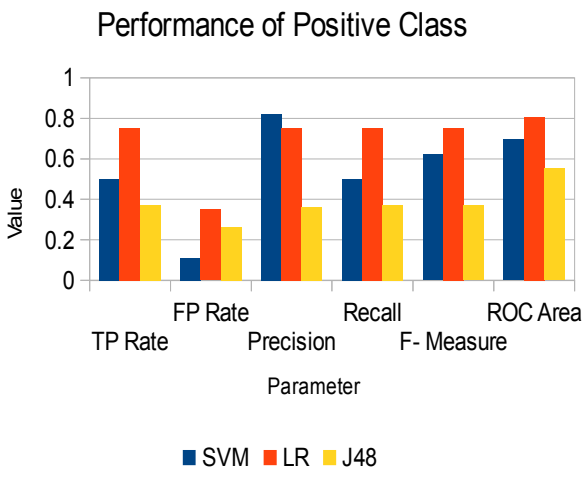

Figure 8 Performance comparisons of SVN, Logistic Regression and J48 Decision tree algorithms with different parameters

IV. Conclusion

The study attempts to examine the use of micro-blogging as a communication channel. The messages expressed in Twitter micro-blogging can be related to human behavior. People can express either positive or negative sentiments and also information can be neutral in nature. Different regions express different sentiments depending on the nature of the movie and how the movies impact cultural sentiments. After applying different machine algorithms, we analyze the sentiment behavior expression though the overall sentiment behavior expressed is positive, people have also expressed negative sentiments, which cannot be ignored. Classification of the tweets into positive sentiments, negative sentiments and neutral statements specify the extent of varied consumer view on any specific aspect which provide important and useful input to various enterprises in formalizing their strategies and carrying out course corrective actions with respect to their strategies. The extents of the inputs / feedback through Tweets are far larger than any other normal means of collecting customer feedback.

References

[1] Bruns, A., The Active Audience: Transforming Journalism from Gatekeeping to Gatewatching. In *Making Online News: The Ethnography of New Media Production*. Eds. Chris Paterson and David Domingo. New York: Peter Lang, 2008.

[2] Boyd, Danahd and Ellison, N.. Social Network Sites: Definition, History and Scholarship. *Journal of Computer-Mediated Communication*, 13(1), 1, 210-230, 2007.

[3] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. First Monday, 14(1), Jan 2009.

[4] Akshay Java, Xiaodan Song, Tim Finin and Belle Tseng. Why we twitter: understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56–65, 2007.

[5] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power:Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 2009.

[6] Karniouchina, Ekaterina, "Impact of star and movie buzz on motion picture distribution and box office revenue*", Intern. J. of Research in Marketing*, vol. 28, pp. 62-74, 2011.

[7] Spann, Martin & Bernd Skiera," InternetBased Virtual Stock Markets for Business Forecasting", *Management Science* , vol. 49, no. 10, pp. 1310-1326, 2003.

[8] http://www.stateofdigital.com/how-to-recognize-twitter-bots-6-signals-to-look-out-for

[9] Asur S, Huberman BA,"Predicting the future with social media" in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, DC, pp. 492-499, 2010.

[10] Joshi M, Das D, Gimpel K, Smith N,"Movie reviews and revenues: An experiment in text regression", in *Proceedings of NAACL-HLT*, PA, pp. 293-296, 2010.

[11] Akcora, C. G., Bayir, M. A., Demirbas, M., and Ferhatosmanoglu, H., "Identifying breakpoints in public opinion", In ACM Proceedings of the First Workshop on Social Media Analytics, July, pp. 62-66. 2010.

[12] Sun, E., Rosenn, I., Marlow, C. and Lento, T. (2009), "Gesundheit! modeling contagion through Facebook news feed", In Proc. Of International AAAI Conference on Weblogs and Social Media, May, pp. 22.

[13] Kwak, H., Lee, C., Park, H. and Moon, S. (2010), "What is twitter, a social network or a news media?", *In ACM Proceedings of the 19th International Conference on World Wide Web*, April, pp. 591-600.

[14] Mestyán, M., Yasseri, T., and Kertész, J.,"Early Prediction of Movie Box Office Success based on Wikipedia Activity Big Data", arXiv preprint arXiv:1211.0970, 2012. Science and Technology, Vol. 3, issue 3, pp. 1878-1884, March 2013.

[15] http://www.cs.cornell.edu/people/pabo/movie-review-data/review polarity.tar.gz

Vijay Kumar Mishra Assistant Professor of Department of Computer Application, Feroze Gandhi Institute of Engineering and Technology, Raebareli, India.

Dr. Neelendra Badal is a working as Associate Professor of Department of Computer Science and Engineering, Kamla Nehru Institute of Technology, Sultanpur, India.