IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

Sentiment Analysis using Machine Learning Techniques

Jami Ruthvik[1], Challa Akhil[2], Ampolu Vasudeva Rao[3], Jonnakutti Rakesh[4], B Avinash Kumar[5]
Students, Department of Computer Science and Engineering[1,2,3,4]
Assistant Professor, Department of Computer Science and Engineering[5]
Raghu Institute of Technology, Visakhapatnam, AP, India

Abstract: *Analytics research includes the field of sentiment analysis. To make sense of this, computational methods can be used to read raw data. Analysis is what this is. Written expression that is either positive, negative, or neutral can be assessed using sentiment analysis. People use a variety of social media platforms, including Facebook, Twitter, etc. Machine learning algorithms can be effectively used to ascertain people's sentiments. Sentiment analysis is a field that has developed to automate the study of such data. Sentiment analysis aims to identify and extract human emotions from text. It seeks to find opinionated information on the Web and categorise it based on its polarity, or whether it has a positive or bad meaning. In contrast to conventional text-based analysis, it is a text-based analysis that helps to swiftly determine the customer's reaction.*

Keywords: Naive bayes, Sentiment analysis, support vector machine, Natural language processing, machine Learning, Natural language toolkit, Data pre-processing, Artificial Intelligence, comma separated values, deep learning. Analysis, python, Vectorization

I. INTRODUCTION

"Sentiment Analysis" is a natural language processing technique that identifies the tone behind the body of the text. Analysis of the customer's positive or negative attitude in text is known as sentiment analysis. Businesses can monitor online chats to use contextual analysis of identifying information to better understand the social attitude of their customers.

As consumers share their ideas and opinions about the brand more freely than ever before, sentiment analysis has become a potent tool for monitoring and interpreting online dialogue. By automatically analysing feedback and reviews from surveys or social media interactions, you may learn what makes a customer pleased or unhappy. You may also use this information to tailor your products and services to your customers' demands and build your brand. Recent advancements in machine learning and deep learning have increased the efficacy of sentiment analysis systems. Research and analysis can be carried out using cutting-edge artificial intelligence and machine learning methods.

We employ the Multinomial Naive Bayes technique in this project. It applies the naive Bayes technique to distributed multinomial data. Natural Language Processing uses the Multinomial Naive Bayes algorithm as a probabilistic learning technique most frequently (NLP). The method, which guesses the tag of a text such as an email or newspaper article, is based on the Bayes theorem. For a given sample, it determines the probabilities of each tag, and then outputs the tag with the highest probability.

In this study, the text was pre-processed using the Natural Language Processing (NLP) technique. The study of natural language processing, or NLP, is a branch of computer science and artificial intelligence that focuses on how computers interact with human (natural) languages, particularly how to teach computers to process and analyse massive volumes of natural language data. It is the area of machine learning that deals with managing predictive analysis and text analysis in general.

A free machine learning library for the Python programming language is called Scikit-learn. With some essential algorithms implemented in C Python for efficiency, Scikit-learn is primarily developed in Python. Here, our attention is on the problem of sentiment categorization, which involves taking a section of unlabelled text and attempting to categorise it in accordance with its overall sentiment.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

II. LITERATURE SURVEY

As researcher [1] have been using Twitter for various purposes to locate prominent users Sentiment analysis is used to learn about the customers' opinions about restaurant's services through their Twitter discussions. Naïve Bayes (NB) classification model is used for sentiment analysis to classify tweets into positive and negative. Even if NB is a simple, probability-based classification model, many researchers like [2] [6] have used sentiment classification. In marketing and analysis, customer loyalty is a fundamental issue in terms of consumer success. As with hotel customer habits, they will pass on the mouth to mouth to other people when they are given excellent service. Text extraction or data retrieval is often done using analytical methods or manuals from the document collection store.

The research method will generate knowledge that can boost revenues and services from various text mining perspectives. Analyses of sentiments are used to find views of a given object from the author [5]. An opinion study on a commodity is a sentiment evaluation analysis. Sentiment analysis is based on the Natural Language Processing (NLP), the analysis of text and certain measured sections to delete or exclude unnecessary parts to interpret the pattern of the term negatively / positive [7] [10]. For sentiment analysis, the use of data mining algorithms has been extensive in the past. Now let's look at how powerful NB is and how widely it was employed as an essential classification data mining algorithm. NB classification is used for seismic and nuclear explosion detection.

A Researcher [7] said that an artificial immune system has also suggested self-adapting attribute weighting for the NB classification. In addition, NB grading techniques are often utilized when weighing features. These weights rarely deteriorate the output in experiments [8] compared to a simple algorithm for classification by NB. Classification strategies of NB also use the frequency approach to detect DDOS connections [6].

The classification NB is also used with T1 weighted MRI scans for the ischemic stroke classification. The passive indoor position classification is also achieved with NB classification, while the final results show that the algorithm is as accurate as 86 percent [5]. Negative class information is also performed in the text classification with the naive classification of Bayes and was executed very well in the results[4]. Displacement-then-confront attack and estimation of the offline server process of the total measurement overhead are used to maintain the privacy of NB classification techniques.

III. PROBLEM STATEMENT

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

3.1 The Problem

Nowadays people mostly believe what is written on the social media and take decision based on that so providing an accurate review to the reader is an important task. People who mostly like food do visit a couple of restaurants and taste them mostly people prefer those restaurants which has a good review rate on the social media, so the project is about collecting all the reviews from the customers and conducting an analysis basing on their reviews and providing them a better review is done. On average 600 million tweets are posted daily in that at least 4-5% of tweets are about restaurants.

If the restaurant business is tweeted about in twitter, and there are 100 tweets about the restaurant it becomes difficult for business owner to read all tweets and keep track on all tweets and analysis if the reviews are positive or negative.

3.2 The Proposed Solution

The proposed system will eliminate the need for the restaurant owner to go through all tweets and analyse them. A machine learning model will be trained and tested using multinomial naïve's bayes theorem. The trained model will be able to go through all the tweets and will be analysis the sentiment of the tweets, it will be able to tell if the tweets are positive or negative. This will eliminate the use for the restaurant owner to go through all the tweets to know about the sentiments of the tweets.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Impact Factor: 7.301
Volume 3, Issue 4, April 2023

IV. SYSTEM DESIGN

4.1 Proposed System Architecture

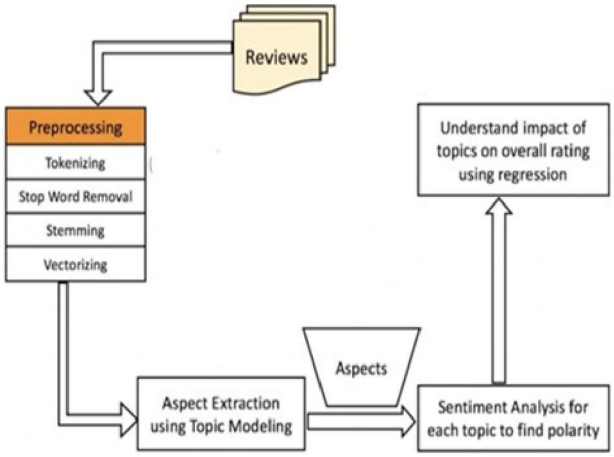

Fig. 1 System architecture

After importing the dataset, classification is done into labels and features followed by data pre-processing. The textual reviews are converted into numerical format. The model is trained and tested. Thereafter, topics are found out from reviews and overall sentiment score is evaluated for each

4.2 Use Case Diagram

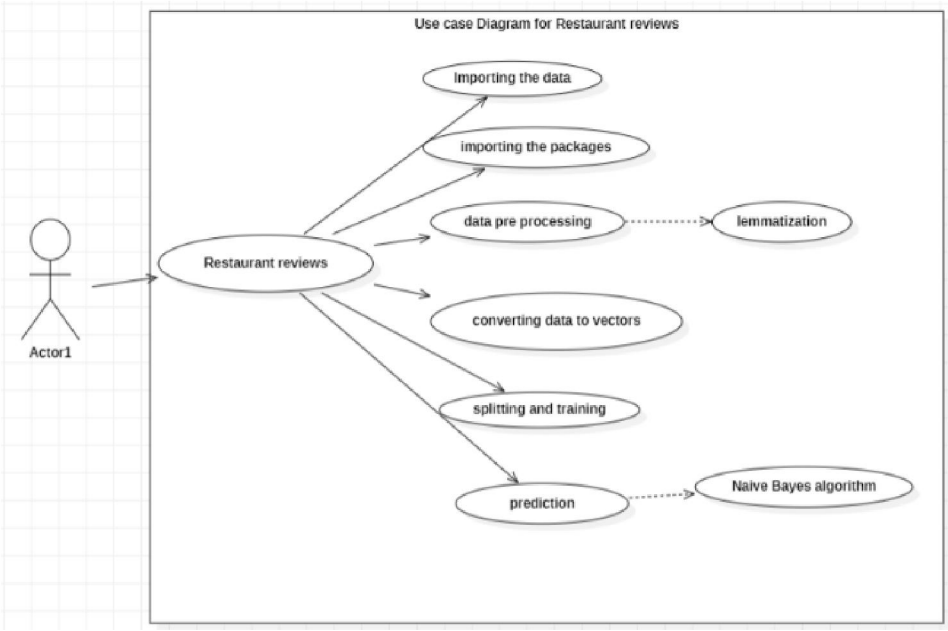

Fig. 2 Use case diagram

Following the dataset and libraries import, data is pre-processed using (lemmatization and POS tagging). Words are converted into vectors. The model is trained using multinomial naive Bayes algorithm to make predictions and its performance is evaluated using a confusion matrix. Then topics are identified from given reviews using LDA and polarity scores are calculated using WordNet.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

4.3 Collaboration Diagram

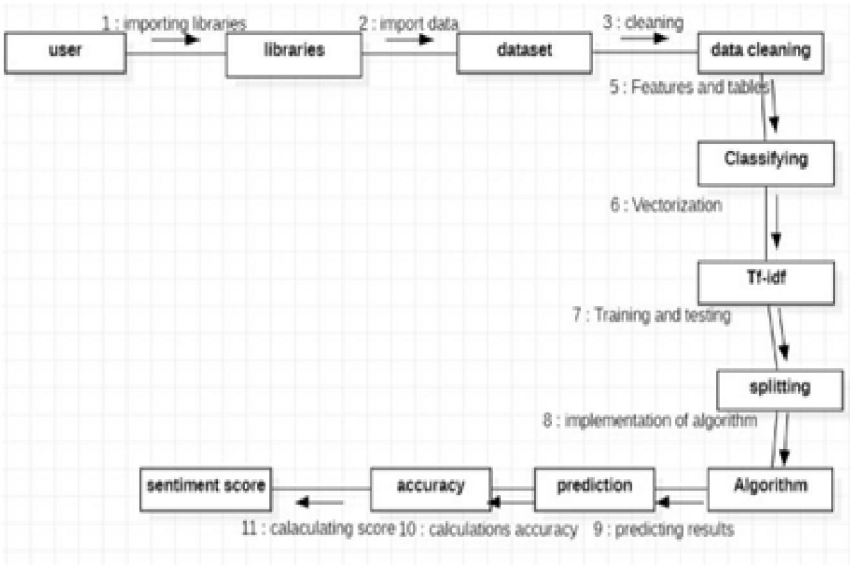

Fig. 3 Collaboration diagram

Data pre-processing techniques are applied that includes the removal of stop words, lemmatization, removal of punctuations and null values after importing the dataset followed by division of data into features and labels. Vectorization is done using Tf-idf vectorizer. In this stage depending on frequency of words each word is assigned a particular value. The model is trained and tested for making predictions on which accuracy is calculated, continued by topic modelling done based on relatedness among words. Finally, the sentiment score is calculated.

4.4 Sequence Diagram

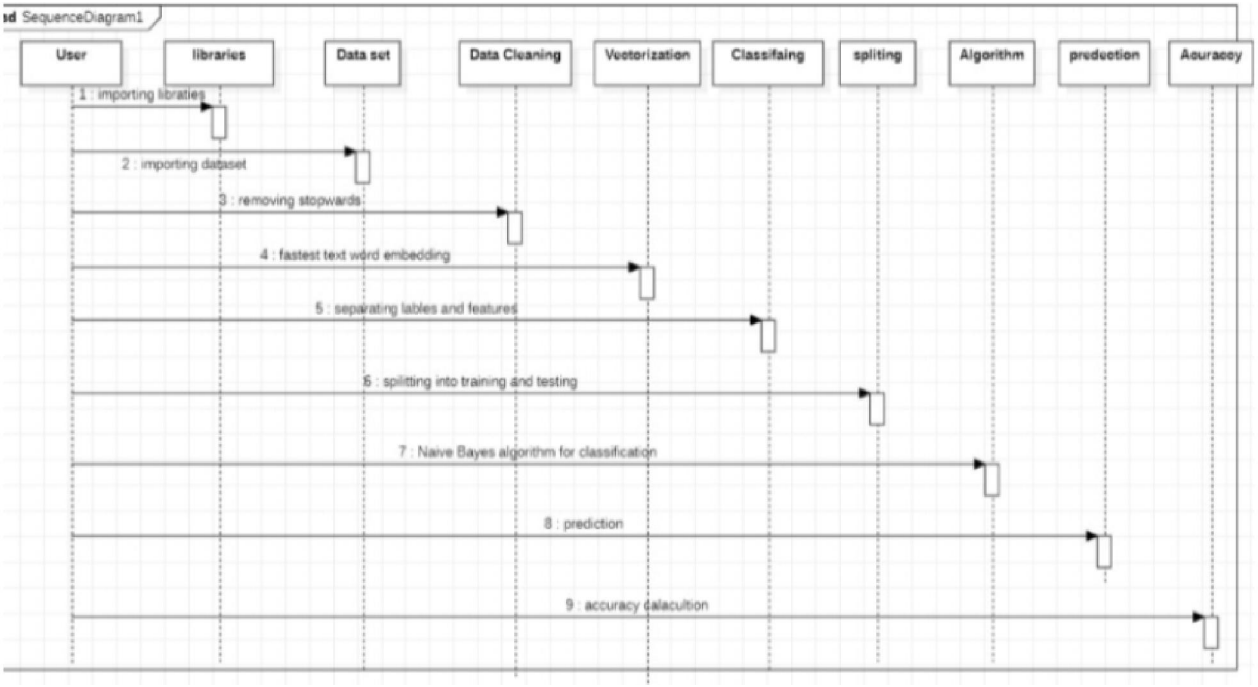

Fig. 4 sequence diagram

This diagram depicts a process in which the end user can perform all functions of importing, data cleaning, classifying and splitting.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

It includes importing dataset and splitting into training and testing sets Nltk- for pre-processing Pandas-for data analysis, Naïve bayes algorithm for prediction, LDA-for topic modelling, Word net-to evaluate sentiment score and interface to perform tasks.

V. SYSTEM OVERVIEW

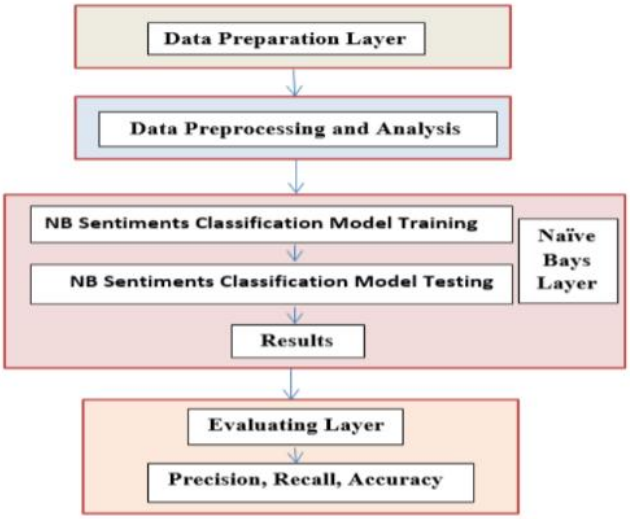

Fig. 5 System working flowchart

5.1 Data Collection Layer

The data set utilised in this paper was downloaded from Kaggle. The dataset consists of tweets that represent actual datasets from conversations concerning customers' evaluations of eateries. 1000 tweets from users from various restaurants make up the dataset. The user who participated in tweet construction, the text, and the sentiment score of the tweet. During the dataset preparation procedure, the data are cleaned by performing data selection and stop-word removal. The tweet's data set contains a huge amount of stop words and punctuation. Consequently, this procedure helps with better evaluation for the tweet data. The primary dataset, which contains all the information and data for this work, determines the dataset choice.

5.2 Data Pre-processing and Analysis Layer

The technique should be followed in a number of steps. The lower-case Trim method changes all of the letters in the uniforms so that "NoiSy" becomes "noisy," all of the letters being in little, unmixed bricks. Stopword deletion is a technique for getting rid of words like "the,""in,""an," and "a," which are frequently found in languages but serve no useful purpose. The technique, which occasionally happens and typically has no meaning, is the elimination of punctuation, such as "-, /,:,? After the text has been pre-processed, the next step is to categorize the sentiments using the NB approach.

5.3 Naive Bayes Layer

The third layer uses the NB Sentiment Classification model to divide the tweets into positive and negative categories. For ease of use, the properties of a unigram used for text analysis are applied to sentiment categorization. The fundamental tokens supplied in each tweet serve as the foundation for the document's word matrix. For each of the matrices, the term's frequency is calculated.

NB's classification model predicts the sentiment in the training phase for the testing dataset based on the fundamental of frequency. In the given example, 70% of the data set is used for training and 30% for testing. The NB classification model bases its calculations on the likelihoods and probabilities that a specific circumstance will arise or a particular item belongs to a class. The algorithm is used to generate the NB model, which categorises the feelings into positive and negative categories.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

5.4 Evaluating Layer

To evaluate the effectiveness of a classification algorithm, various assessment metrics are employed. A contingency table, often referred to as a confusion matrix, is frequently used to assess the performance of classification algorithms. It is an easy and clear method of displaying the classification results according to classification accuracy. It is determined by estimating the number of correctly identified class examples (true positives), the number of correctly identified examples that do not belong to the class (true negatives), and the number of examples that were either not correctly identified as class examples (false negatives) or incorrectly identified as class examples (false positives). This matrix of confusion is made up of these four counts.

Let's assume that TP refers to the emotions that the system accurately identifies as positive. Then, TN stands for the feelings that the system correctly classifies as negative, FP for the emotion that the system mistakenly classifies as positive but is actually negative, and FN for the sentiment that the system mistakenly classifies as negative but is actually positive. You can retrieve the accuracy (ACC), error rate (ERR), precision, and recall as

$$ACC = \frac{\sum TP + \sum TN}{\sum \text{Total Population}}$$

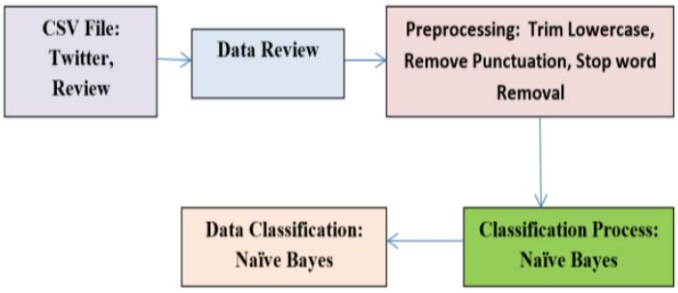

Fig. 7 System workflow

VI. RESULT AND ANALYSIS

A popular online forum for posting opinions on various themes in status updates is Twitter result discussion. In this study, tweets were collected and trained on using a Python script. The tweets were collected via Coma Separate Files (CSV), which are used to assess client happiness and displeasure with the server provider. Following the creation of the training set, data analysis is performed by uploading it into the NB classifier. The information was gathered and then loaded as a data frame with 1000 comments into the Python library. The likes or dislikes were displayed, which indicates that there are two labels or targets: the positive class means one and the negative class means 0.

The following step involves trimming lower case characters or standardising the letters to small letters in order to prepare analysis data for training and test data. The following pre-process outlines how terms and punctuation that occasionally appear but have no use in the document are removed. In web mining, a confusion matrix is required. The confusion matrix was created after developing and implementing the classification model for classifying comments as positive or negative

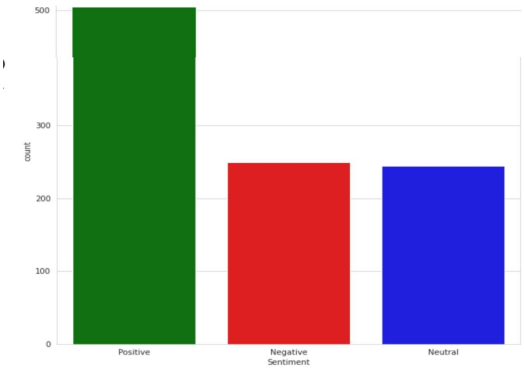

Fig. 6 Graph analysis

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

VII. CONCLUSION

In this project we were able to tell how the customer's opinion was on the restaurant from the tweets on the twitter platform. The model which was trained using the multinomial navies' bayes was able to analyze the sentiments of tweets about the restaurant. The trained model was able to tell if the tweets were positive or negative. This trained model can not only be used for the analysis of restaurant reviews on twitter, but can also be used for the analysis of any business or any product. With use of this model, we were able to achieve the accuracy of 75.2%, precision of 0.72 and 0.77 of recall. The accuracy of the model can be increased by training the model with additional datasets and testing

REFERENCES

[1]. Journal of Transport & Health, 16, 2020, 100842; Barakat Albadnai, Ranghuashai,"Commuters' Satisfaction with Public Transportation."

[2]. A unique text mining approach for scholar information extraction from web material Chinese, X. Xie, Y. Fu, H. Jin, Y. Zhao, Peng ceh, Kexin zhang and W. Cao, Future Generation Computer Systems, 111, 2020, 859–872.

[3]. B. Liu and L. Zhang, "A Survey of Opinion Mining and Sentiment Analysis in Mining Text Data," Springer, Boston, MA, 2012, pp. 415–463. Applications of modelling and simulation, 5, 2021, 166–172, M. M. HAMAD et al.

[4]. "Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis," P. Tiwari, S. Kumar, V. Kumar, and B. K. Mishra, Cognitive Analytics: Concepts, Methodologies, Tools, and Applications, IGI Global, 2020, 689–701

[5]. S. N. Alves-Souza, L. Vilela Leite Filgueiras, F. G. Contratres, and L. Trends and Advances in Information Systems and Technologies, 2018, 122-132; S. DeSouza, Sentiment analysis of social network data for cold-start relief in recommender systems

[6]. H. U. Khan, Classifying web forum postings' mixed emotions using lexical and nonlexical elements, Journal of Web Engineering, 16(1- 2), 2017, 161–176.

[7]. U. Ishfaq, K., Dipak R, Dr Kavita and H. U. Khan Journal of Web Engineering, 16(5-6), 2017, 505-523. Iqbal, "Identifying the Influential Bloggers: A Modular Approach Based on Sentiment Analysis."

[8]. H. N. Aljohani, T. Amjad, U. Khan, A. Daud, R. A. Abbasi, and J. Modeling to find important bloggers in the blogosphere: A survey, S. Alowibdi, Computers in Human Behavior, 68, 2017, 64–82.

[9]. MIIB: A metric to identify most influential bloggers in a community, H. U. Khan, A. Daud, and T. A. Malik, PLOS ONE One, 10(9), 2015, e0138359.

[10]. Mahmood, H. U. Khan, Zahoor-ur-Rehman, and W. Khan, "Query based information retrieval and knowledge extraction utilising Hadith datasets," 13th International Conference on Emerging Technologies (ICET), Islamabad, Pakistan, 1-6, 2017.