International Journal of Advanced Research in IT and Engineering

ISSN: 2278-6244
Impact Factor: 6.111

SENTIMENT ANALYSIS OF USER'S VIEWS USING MACHINE LEARNING

Rupinder Kaur, M.tech Research Scholar, Ambala College of Engineering and Applied Research, Devsthali, Ambala

Ashok, Assistant Prof., Department of Computer Science & Engineering, Ambala College of Engineering and Applied Research, Devsthali, Ambala

Abstract: *Sentiment Analysis or Opinion Mining is an important concept in today's world and due to the increased use of media it has become a huge source of database. Since everybody in the modern era is involved with some social media platform, the public mood is hugely reflected in the social media platform today. This study proposes to utilize this source of information and predict the all sentiments of public towards the food price in India expressed over twitter and twitter API is used for extracting live tweets. Oauth is used as handler and tweets are filtered for specific keywords and location using latitude and longitude data. The tweets are saved into a database. They first preprocessed for elimination of stop word, special characters, short words etc, after that stemming and tokenization steps are applied and TF-IDF score is calculated for all the keywords. A term document matrix (TDM) is created which is fed into the classifiers for classification. KNN and Naïve Baye's has been analyzed in this study and Hybrid algorithm using them was designed. The results of KNN and Naïve Baye's classifier in sentiment classification were found to be significant while the hybrid-KNN outperforms the Naïve Baye's Classifiers in terms of accuracy..*

Keywords: *Opinion Mining, Sentiment Analysis, KNN, Naïve Baye's Classifier, Food price.*

I. INTRODUCTION

Human life is filled with emotions and opinions. One cannot be imagined without emotions and opinions. They play a vital role in nearly all human actions and lead the human life by influencing the way they think, what they do and how they act. The recipients of the information do not only consume the available contents on web, but also can change this content and generate new data of information. In today's world of social media users can comment on already existing information, can book mark pages. They can also share their plans, news and knowledge with online communication. In this way, the entire community becomes a writer, in addition to being a reader. Internet users can port their data, give opinions and get feedback of other users through different medium like blogs, forums and

social networks etc. The increasing popularity of different personal publishing services is increasing day by day and in coming future this increase is expected to continue. Thus the opinionated information on web will become so large that it cannot be handled manually so need of automated Sentiment Classification of online opinions, reviews of information is desire of current and future scenario. Recently, most of researchers have focused on this area [1]. They fetch opinionated information to analyze and summarize the opinions expressed on web by different automatically with computers. Until now, all researchers have evolved several techniques to the solution of the problem. Current-day Information Retrieval (IR) and Natural Language Processing (NLP) is the crossroad for the Opinion Mining and Sentiment Analysis and share some characteristics with other disciplines such as text mining.

1.1 Use of Social Media

The popularity of social media can be accessed by the fact that almost 90% of internet users use it for some context or other. Some are active member of social networking; some use it for blogging and micro blogging. Other usage includes online video sharing, ecommerce etc [14]. Day by day, the number of internet users is also increasing. With increase of the user involvement on web their contribution to online data is also enhanced. One of contribution of this trend is providing online reviews in social networking sites. These reviews help users to take better decisions about the product for which reviews were placed. [12]. Hence to provide automation, we are studying sentiment analysis. Sentiment analysis is the modern method which helps to analyze huge amount of data to extract sentiments associated with the data.

1.2 Type of Social Media Applications

There are many social media platforms like Face book, Twitter, LinkedIn, You tube etc. In today's world they tend to be an integral part of almost every one's life. Figure 1.1 shows usage of different social media platform by the users of different age group.

1.3 Need for Analyzing Social Media Data

The use of social media is increasing day by day. Increasing growth of social media users over internet has also increased their participation in various discussions and activities simultaneously. In case of a product, users are tries to express the rating views about the product [12].These reviews are equally important for buyers and companies.

Table 1: User pattern of each Social Media

Social Network Used by U.S Internet Users, by Age, July 2016 % of respondents in each group				
	14-17	18-34	35-54	TOTAL
Face book	63.8%	83.4%	74.2%	76.8%
YouTube	81.9%	77.2%	54.2%	66.4%
Twitter	31.0%	38.7%	28.3%	32.8%
LinkedIn	1.5%	15.9%	20.0%	16.6%
WhatsApp	8.0%	9.8%	4.0%	6.8%

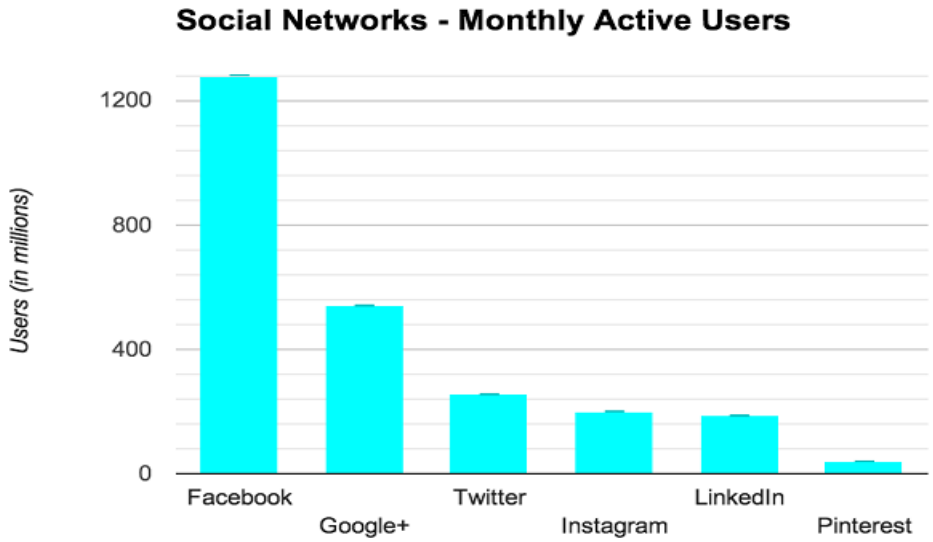

Figure 1: Depicting the Users of each type of social media platform

1.4 Sentiment Analysis

Sentiment analysis is a text classification problem which deals with extracting useful information present within the web documents. This extracted data can be then further classified according to its polarity as positive, negative or neutral. Sentiment analysis is widely used in business and government intelligence, Named Entity Recognition.

Sentiment analysis is also about finding subjectivity or objectivity of the opinion. Subjectivity is about someone's personal review whereas objectivity is the opinion given by an expert. For example: doctor's opinion about the patient on the basis of observed symptoms comes under the objectivity.

1.5 Definition of Sentiment Analysis

It can be defined as a computational task of extracting sentiments from the opinion. Some opinions represent sentiments and some opinions do not represent any sentiment.

Sentiment analysis is a natural language processing and information extraction task. This aims to extract writer's feelings expressed in comments or reviews expressed over web.

1.6 Levels of Sentiment Analysis

Sentiment analysis is defined as a classification process. There are 3 main classification levels in sentiment analysis:

- **Document Level**

  Identify if the document have (product reviews, forum posts) expressed opinions and whether opinions are positive negative or neutral.

- **Sentence Level**

  The task at this level goes to the sentences and found whether each sentence expressed positive, negative or neutral opinions.

- **Attribute Level**

  Extract object attribute (e.g. image quality, zoom size.) that are subject of opinion and opinion orientations (positive, negative or neutral).

1.7 Machine learning

Machine learning is technique by which a device modifies its own behavior due to the result of its past experience. This is systematic way which design algorithms and permit machine to evolve behaviors based on experimental data. Machine learning approaches can be divided into two categories:

- Supervised Learning
- Unsupervised Learning

II. LITERATURE SURVEY

Alexander Trilla [15] Used text classification scheme based on Multinomial Naïve Bayes to deals with Twitter messages. The effectiveness of this technique was evaluated using TASS-SEPLN twitter data sets and it achieved maximum macro averaged F1 measure rate of 36.28%. The accurate results provided by TASS-SEPLN organizers indicate that the proposal based on MNB was rather effective.

Aisopos and Fotis [4] have studied some serious challenges associated with respect to Micro blog content. Some of these are the applicability of sentiment analysis over past and different classification methods caused by their inherent characteristics of content. To

resolve them, author introduced a method that relies on two orthogonal and complementary sources of evidence: context-based method captured by polarity ratio and content-based features acquired by n-gram graphs. Both the methods are language-neutral and tolerant to noise; guarantee high robustness and effectiveness in the manner author are considering. To ensure this approach can be applied into practical applications with large amount of data, aim should be enhancing its time efficiency. Thus author propose alternative sets of features having low extraction cost, explore dimensionality reduction techniques and discretization techniques and also experiment with multiple different classification

Ortigosa et. al [1] proposed a novel method for sentiment analysis in social site giant Face book that, starting from the messages of its users, supports: (i) to extract useful information about the Face book users' sentiment polarity, which reflected from the user's messages; and (ii) to model the users' normal sentiment polarity and to analyze significant emotional changes in user. Author has implemented this method in Sent Buk which is a Face book application also presented in the paper. In general, in order to take decisions based on information and emotions of the users, it is necessary for a system to get and store this information. One of the most reliable procedures to fetch information about user's emotion consists of asking them directly to fill in questionnaires which helps to detect their option. However, for a particular user this task can be too time-consuming and tedious.

Agarwal and Apoorv [3] worked with micro blog data named as Twitter and manufacture models to classification of the "tweets" into positive and negative sentiment or they can be neutral. Author build novel models for two classification: first one is a binary task of classifying sentiment of users into positive and negative classes and secondly is a 3-way task of sentiment classification of users into positive, negative and neutral. Author experimented with two types of models: (1) unigram model which is a feature based model (2) a tree 30 kernel based model. They build a new tweet representation, for the tree kernel based model. They take a unigram model, which work well for sentiment analysis for Twitter data in the past. Result indicates that a unigram model is really a hard baseline. Feature based model that used 100 features gives similar accuracy as compared to the unigram model that

used about 10,000 features. Tree kernel based model gives improvement outperforming both these models by a significant margin.

Balahur and Alexandr [5], identified that the major difference between subjective texts type (like movie or product reviews) is that their target is unique and clearly stated across the text. Following various efforts of annotation and the analysis of the issues encountered, it was realized that news opinion mining is different from that of other text types. They identified 3 subtasks that need to be addressed: defining the target; separating the bad and good news content from the sentiment expressed which is good and bad; and finally analysis of clearly mentioned opinion that is expressed not ambiguously, not needing understanding or the utilization of world knowledge. Furthermore, they distinguish 3 not similar views on newspaper articles, which have to be handling differently while analyzing sentiment.

Horakova and Marketa [8] presented a model which collects tweets from social networking sites and thus provide a view of business intelligence. In the framework, there are two layers in the sentiment analysis tool, the layer of data processing and the layer of sentiment analysis. Data processing layer deals with data collection and data mining, while sentiment analysis layer use a application to present the result of data mining.

III. PROPOSED METHODOLOGY

This section describes the various techniques were applied for the fulfillment of following objectives.

- To study existing algorithms for Sentiment Analysis.
- Analysis and Classification of User's views about products expressed over the web using machine learning techniques

The various text mining algorithm and streaming of twitter API are discussed in this section. The process starts with the extraction of tweets followed by preprocessing of the extracted tweets. Then Classifier algorithm has to be applied on it to identify the polarity.

Data extraction: The twitter API was used for tweet extraction. The major steps involved in *development* of the framework for live streaming of tweets begin with setting up an account on twitter.

- Set up your account on twitter
- Go to site of dev.twitter.com
- Create a new app and register for it
- Change access level to Read, write and access messages
- Generate security id and secret number
- Generate access token id and secret token number
- Save them to be utilized for streaming

Auth handler was used for streaming the tweets. Filters are applied on it using the track filter. The tweets were filtered by two ways.

- Filter by content
- Filter by location

Due to the policies of twitter the filtering is not absolutely correct and there might be a similar tweet which doesn't lie in the filtered bandwidth. The content filtering is done using the following keywords:

- Mehngai
- Food cost
- Inflation
- Food Security
- Prices of vegetables
- Price Rise

The location filtering was performed by using a 'location' filter available with tweepy. The location filter works on the basis of latitude and longitude of the place. A bounding box has to be formed in which the location filter works. Any tweets sent from that bounding box is streamed.

This has utilized the following settings:

South West Longitude=73 degrees

South West Latitude=15 degrees

North East longitude=85 degrees

North East Latitude=27 degrees

Using these settings the tweets are extracted and saved in a database.

For further processing Text mining was applied on the filtered tweets

Pre-processing: Tokenization, stop word removal and stemming are the main steps of preprocessing. Tokenization divides textual description into tokens by removing punctuation marks. Then stop words are performed that remove unnecessary information (conjunctions, interjections and articles) from datasets. Stemming on reduced datasets is performed to reduced terms into their root terms. Porter's stemming algorithm is used to perform stemming.

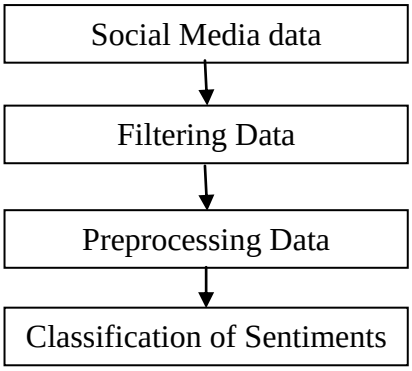

Figure 2: Sentiment Analysis process

Step 1: Preprocessing: To distill unstructured data to structured format this step is used. There are different preprocessing steps used in Text mining such as tokenization, stop word removal and stemming. These algorithms are discussed below.

i. *Tokenization*: There are two types of tokenization i.e. partial and full tokenization. For removing commas, full stop, hyphen and brackets we use this step. It divides the whole text into separate tokens to explore the words in document.

ii. *Stop word removal*: The purpose of this process is used to reduce conjunction, prepositions, articles and other frequent words such as adverbs, verbs and adjectives from data. Thus it reduces textual data and system performance is improved.

iii. *Stemming*: For reduction of words into their root word e.g. words like "Processing", "processed" has it root word "process" stemming process is used. The purpose of stemming is to represent the words to only terms in their document. There are various tools to perform stemming such as Lovins Stemmer, Porters Stemmer, Paice/Husk Stemmer, Dawson Stemmer, HMM Stemmer.

iv. *Weighting Factor*: - Features are extracted from overloaded large datasets.TF-IDF (Term frequency- Inverse document frequency) [7] score is generally is used to give

weight to each term. TF-IDF is multiply of term frequency and inverse document frequency.

$$\text{TF-IDE} = n_w^d \log_2\left(\frac{N}{N_w}\right)$$

Where n_w^d = frequency of word w in document d.

N= total document and N_w = document containing word w.

v. *Term-document matrix* – After all steps of the preprocessing Term- document matrix is created from the text available in documents. Rows in matrix represents document in which word appears and columns represent the words that are extracted from documents. The TF-IDF score filled in the cell of matrix.

Machine Learning Approaches

There are different methods to design machine learning algorithms. The purpose of ML algorithms is to use observations as an input and this can be a data, information and past experience. To improve the performance of instances we used MI algorithms, which can be done by any classifier by trying to classify the input pattern into set of categories or to cluster unknown instances. There are two categories supervised and unsupervised approach [9].

KNN Algorithm

The K means algorithm takes the centroid of a cluster as mean value of the points within cluster. It randomly selects k, number of objects in dataset, each of which initially represents a cluster mean. For each of remaining objects, an object is assigned to cluster to which it is most similar, based on Euclidean distance. The algorithm iteratively improves within cluster variations. The iterations continue until assignment become stable.

KNN is lazy learning type of algorithm. In this learning the function is accurately local and all computation is deferred until classification. It is simplest of all machine learning algorithms. In KNN classification, the class membership is output. An object is classified by majority votes of its neighbors by the object being assigned to class which is most common among its k nearest neighbor (k is positive small integer). The nearest neighbor is calculated by using similarity measure usually distance functions are user. There is some distance function used by KNN [50].

Euclidean Distance Function $\sqrt{\sum_{i=1}^{N}(a_i - b_i)^2}$

Manhattan Distance Function $\sum_{i=}^{N}|a_i - b_i|$

Where {(a_1, b_1), (a_2,b_2) ,(a_3,b_3)…… (a_N, b_N)} are training datasets.

In KNN algorithm all the distance from testing data point to training data point are computed. Then all testing points are sorted according to the ascending order. Then class labels are added for each K nearest neighbors and sign of sum are used for checking prediction. Finding value of K in K-nearest neighbor is more challenging task.

As choosing smaller value of k. e.g. by choosing K=1 may take lead to risk of over fitting and choosing larger value of K e.g. K=N may take lead to under fitting problem. Therefore optimal value of K has been taken between the values 3-10, which gives better result.

Strengths of KNN

1. It is relatively efficient and scalable in processing on large number of dataset.
2. It is often terminate at local optimum.

Weakness

1. Applicable only when mean is defined.
2. Unable to handle noisy data.
3. No guarantee to converge to global optimum

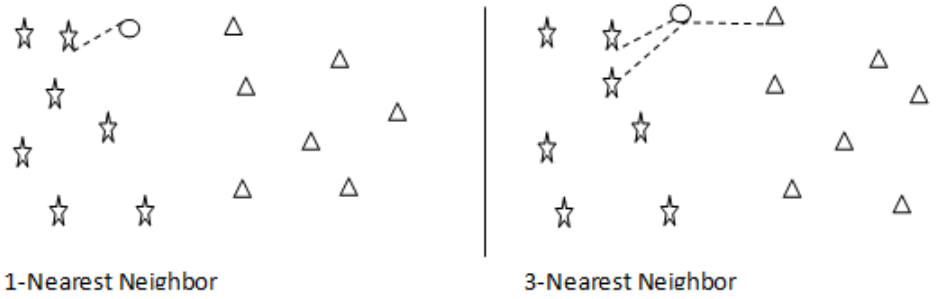

Figure 3: Working of KNN Algorithm

Naïve Bayes Algorithm

Bayesian classifier is statistical classifier. It can predict class membership probabilities such as probability that a given tuple belongs to a particular class. Bayesian classifier has a minimum error rate in comparisons to other classifiers.

The algorithm is named after popular statistician Thomas Bayes who proposed Bayesian theorem. The Naïve bayes algorithm is also based on Bayesian theorem. This theorem supposes that all the attributes are conditionally independent to each other. This assumption is also called class conditional independence. In this algorithm, conditional probability for every attribute with respect to certain class level is calculated. The new document is classified using sum of probabilities for each class [12] . The classifier is easy to build and useful when there is large datasets.The classification framework is briefly discussed as follows:

Suppose we have D set of tuples and each tuple has attribute vector X(x1, x2, x3 , ... xn) of n dimensions. Let there are k number of classes C1, C2, C3... Ck. The classifier predicts X belongs to Ci if

$$P(\frac{C_i}{X}) = P\frac{C_j}{X}$$

for 1<= j <= k, j < > i

Posterior probability is calculated as

$$P\left(\frac{C_i}{X}\right) = \frac{P\left(\frac{X}{Ci}\right) P(c_i)}{P(X)}$$

Problems with Existing Approach:

Problems with KNN

- A shortcoming of the *k*-NN approach is that it is sensitive to the local structure of the data.
- KNN doesn't know which attributes are more important.
- Doesn't handle missing data gracefully.

Problems with Naïve Bayes

- Most important disadvantage of Naive Bayes is that it has strong feature of independence assumptions.
- In classification tasks you need a big dataset. You can use Naïve Bayes classification algorithm with a small data set but precision and recall will keep very slow.

Proposed KNN Hybrid algorithm

A KNN Hybrid algorithm is composition of KNN algorithm and Naïve Bayes algorithm. Here Naïve Bayes algorithm is embedded in KNN algorithm to take advantages of both

algorithms. By this approach we will able to handle noisy database and missing values also. Hence the capabilities of KNN are enhanced by embedded it with Naïve Bayes.

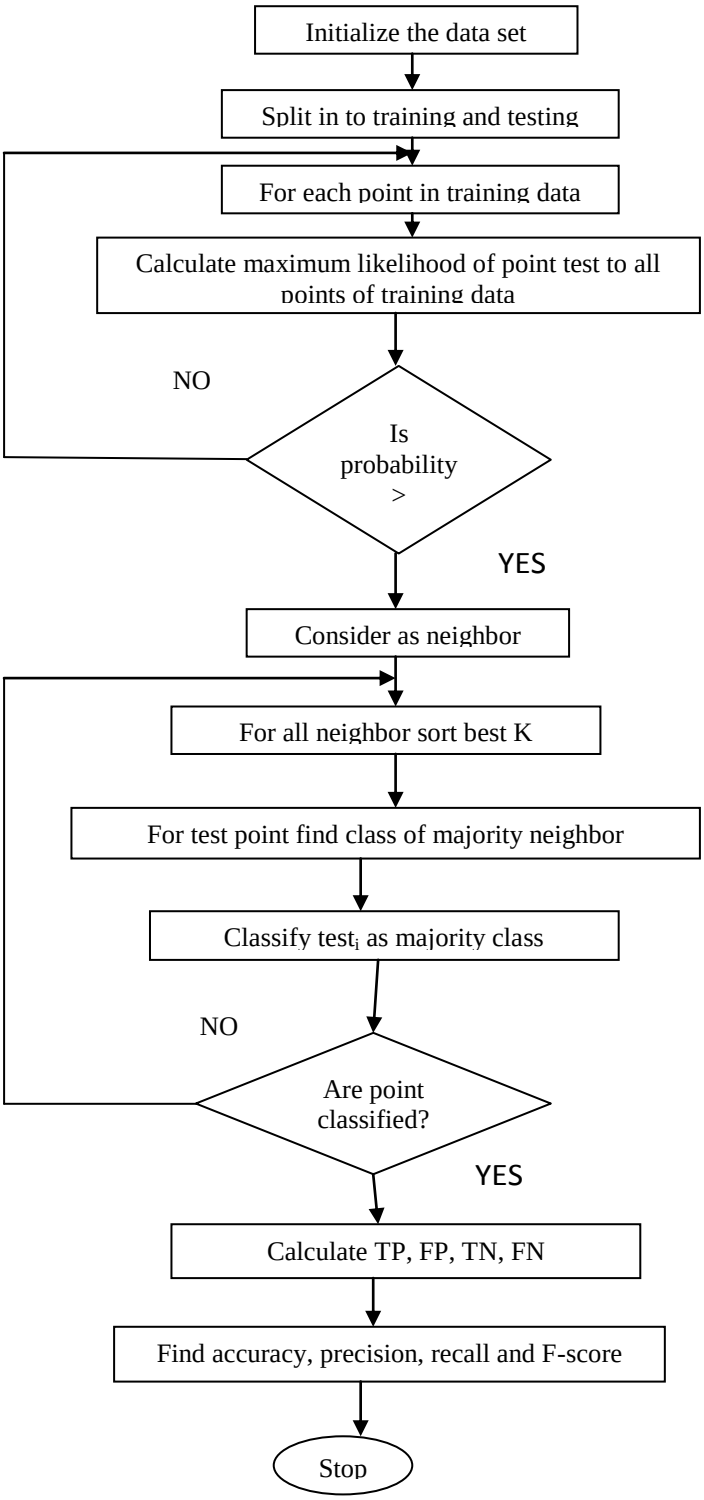

Figure 4: KNN Hybrid Algorithm

IV. RESULTS AND DISCUSSION

This Section presents the results obtained by various methodologies applied on the dataset discussed in the previous section. The results are verified by running the simulations for repeated number of times. The opinions are mined and analyzed for public response.

An app named 'rupinder kaur' was created. This was utilized for streaming.

A consumer Key is generated. Next figure shows the Keys generated which will be used for streaming.

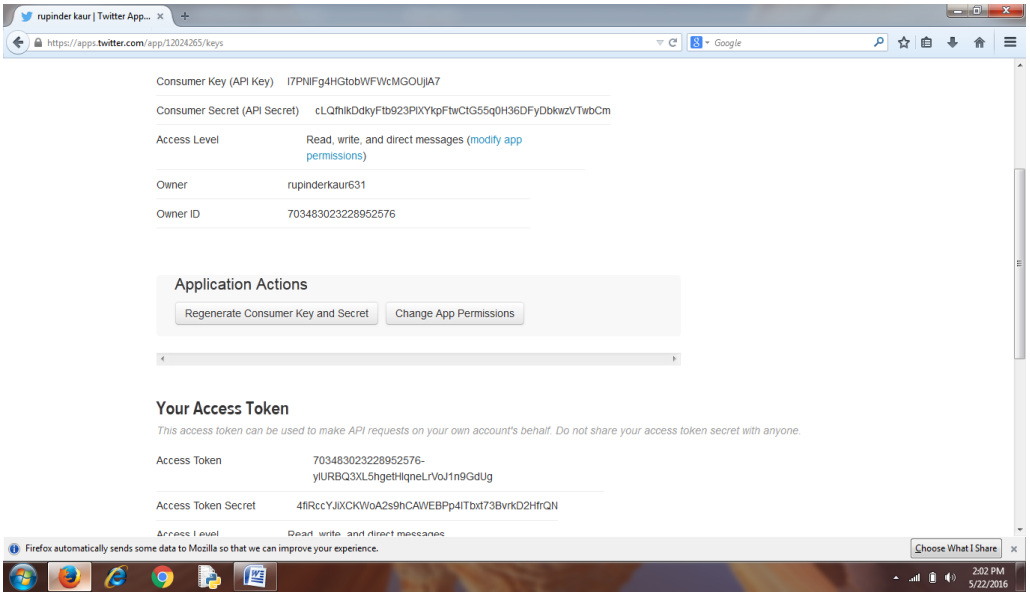

Figure 5: Key Used for Streaming

An array of the tweets is created and term document matrix is created using TFIDF score as shown below.

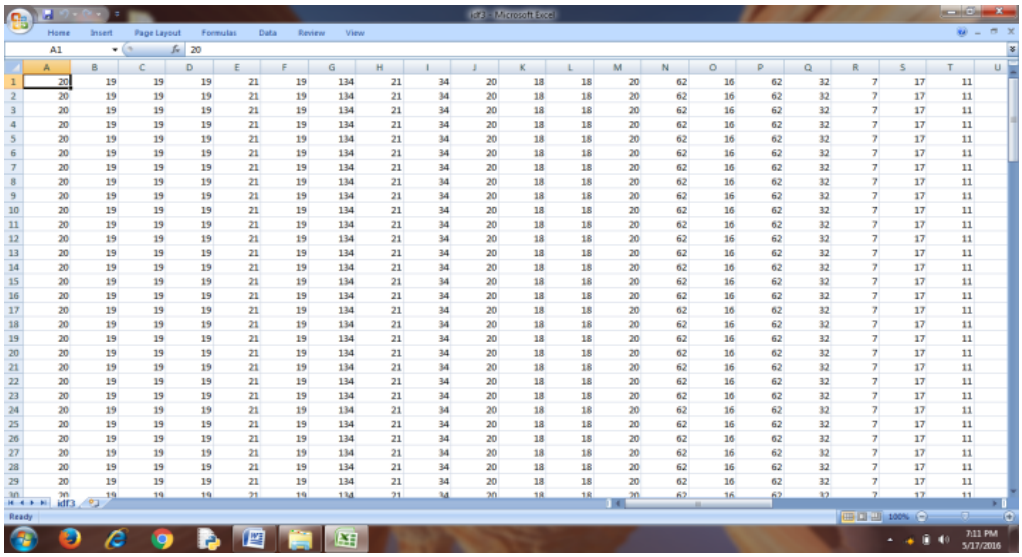

Figure 6: Creation of Term Document Matrix using TFIDF

• **Naïve Bayes**

When Naïve Bayes algorithm was applied on data of different sizes max accuracy achieved was 82%. The result of Naïve Bayes classifiers are shown in table 2.

Table 2:Result for Naïve Bayes Algorithm for different dataset

Total Test Record	Total Correct Records	Accuracy	Precision	Recall
42	28	82%	82%	74%
160	136	79%	80%	70%
312	159	81%	84%	76%

• **KNN**

When K- Nearest Neighbors algorithm was applied on data of different sizes then maximum accuracy of 83% was achieved. The result of KNN classifiers are shown in table 3.

Table 3: Result for KNN Algorithm for different dataset

Total Test Records	Total Correct Records	Accuracy	Precision	Recall
38	29	83%	78%	84%
436	324	81%	75%	84%
931	702	81%	75%	84%

• **KNN Hybrid**

Two classifiers have been analyzed in this: KNN and Naïve Baye's and a hybrid have been made using them. Hybrid scheme is the combination of two or more types, so we can design it in such a way that strengths of both types are maximized

When KNN Hybrid algorithm was applied on data of different sizes then it was observed that accuracy was enhanced as to Naïve Bayes and KNN. Precision of hybrid algorithm was found better than other two algorithms. The result of KNN classifiers are shown in table 4.

Table4: Result for Hybrid KNN Algo for

Total Test Record	Total Correct Records	Accuracy	Precision	Recall
42	33	84%	85%	76%
470	370	82%	80%	70%
1037	803	82%	84%	76%

V. CONCLUSION

A methodology for the classification of sentiments was developed in this study for food price data. Twitter API was used for streaming of tweets. The streamed tweets was filtered

for relevant content and stored in a database. The several steps of pre-processing were applied on it and the tweets were removed from special characters, stop word, tokenized, etc. Stemming was done to all words in order to extract the root words.

TF-IDF score based approach was utilized and the score was calculated for each tweets. The extracted features form a term document matrix which is utilized in the classification algorithm.

The results are found to be satisfactory and when comparative analysis was done between them it is found that KNN Hybrid outperforms Naïve Baye's and KNN Algorithm in terms of accuracy and precision. Thus an automated system is designed for opinion mining related to food price data.

REFERENCES

[1] Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. "Sentiment analysis in Facebook and its application to e-learning." Computers in Human Behavior 31 (2014): 527-541.

[2] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.

[3] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, 2011.

[4] Aisopos, Fotis, et al. "Content vs. context for sentiment analysis: a comparative analysis over microblogs." Proceedings of the 23rd ACM conference on Hypertext and social media. ACM, 2012.

[5] Balahur, Alexandra, et al. "Sentiment analysis in the news." arXiv preprint arXiv:1309.6202 (2013).

[6] Jebaseeli, A. Nisha, and E. Kirubakaran. "A Survey on Sentiment Analysis of (Product) Reviews." International Journal of Computer Applications 47.11 (2012).

[7] Scholar, P. G. "Big-SoSA: Social Sentiment Analysis and Data Visualization on Big Data."

[8] Horakova, Marketa. "Sentiment Analysis Tool using Machine Learning." Global Journal on Technology (2015).

[9] Gupta, Aditi, et al. "Sentiment analysis for social media." International Journal of Advanced Research in Computer Science and Software Engineering 3.7 (2013): 216-221.

[10] B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1 -2):1{135, 2008.

[11] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79{86, 2002.

[12] Twitter Sentiment Classification using Distant Supervision by Alec Go, Richa Bhayani, and Lei Huang.

[13] Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Association for Computational Linguistics, 2005.

[14] K.Nigam, J. Lafferty, and A. Mccallum. Using maximum (2016) entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61{67, 1999.

[15] Alexandre Trilla. ”Sentiment Analysis of Twitter messages based on Multinomial Naïve Bayes” (2012).