A Novel Clustering Approach Based Sentiment Analysis of Social Media Data

Neha, Amit Garg
Department of Computer Science engineering
Indus Institute of Engineering and Technology, Jind

***Abstract* - Opinion Mining is an important concept in today's world and due to the advent of social media it has become a huge source of database. Since almost everybody in the modern era is involved with some social media platform, the public mood is hugely reflected in the social media today. This thesis proposes to utilize this source of information and predict the sentiments of public towards a particular topic. Food price crisis is being studied here in this thesis and public opinion is predicted for the topic. Twitter data is utilized for the same and live tweets of Indian origin are extracted using twitter API called 'tweepy'. Oauth is used as handler and tweets are filtered for specific keywords and location using latitude longitude data. The tweets are saved into a database. They first preprocessed for removal of spam, special characters, url, short words etc. The tweets are then stemmed and tokenized and TF-IDF score is calculated for all the keywords. Feature selection is applied on it using Chi-Square and information gain. A term document matrix (TDM) is created which is fed to the classifiers for classification. Two classifiers has been analysed in this thesis: KNN and Naïve Baye's. The results of both the classifier has ben found to be satisfactory while the KNN outperforms the Naïve Baye's Classifiers in terms of accuracy. Thus a novel method is designed for opinion mining of Indian tweets regarding food price crisis.**

***Keywords* - Opinion Mining, Sentiment Analysis, KNN, Naïve Baye's Classifier, Food price Crisis**

I. INTRODUCTION

Human life is filled with emotions and opinions. We cannot imagine the world without them. Emotions and opinions play a vital role in nearly all human actions. They lead the human life by influencing the way we think, what we do and how we act. Having an access to large quantities of data through internet and its transformation into a social web is no longer an issue, as there are terabytes of new information produced on the web everyday that are available to any individual[4]. Even more importantly, it has changed the way we share information. The receivers of the information do not only consume the available content on web, but in turn, actively annotate this content and generate new pieces of information. Today people not only comment on the existing information, bookmark pages and provide ratings but they also share their ideas, news and knowledge with the community at large. In this way, the entire community becomes a writer, in addition to being a reader [7]. The existing mediums like Blogs, Wikis, Forums and Social Networks where users can post information, give opinions and get feedback from other users on different topics, ranging from politics and health to product reviews and travelling. The increasing popularity of personal publishing services of different kinds suggests that opinionated information will become an important aspect of the textual data n the web. Recently, many researchers have focused on this area [1]. They are trying to fetch opinion information to analyze and summarize the op inions expressed automatically with computers. This new research domain is usually called Opinion Mining and Sentiment Analysis [6]. Until now, researchers have evolved sever al techniques to the solution of the problem. Current-day Opinion Mining and Sentiment Analysis is a field of study at the crossroad of Information Retrieval (IR) and Natural Language Processing (NLP) and share some characteristics with other disciplines such as text mining and Information Extraction[15].

Use of Social Media

Almost four out of five users of internet use social media for some or other context. Some of these include friendship networks, blogging and micro-blogging sites, content and video sharing sites, e-commerce sites etc [14]. The involvement and contribution of the users on the web is increasing day by day. One such contribution is reviews of users in social networking sites. The current trend of giving online reviews enables users to take better decisions who want to use a particular service or purchase a particular product. It helps them to check the popularity of the product. It also enables them to extract the positive or negative features of the products by reading reviews [12]. But manual analysis of such a huge amount of reviews can lead to biased decision. So to provide automation, we are studying sentiment analysis. Sentiment analysis is the modern methodology which analyze huge amount of data to extract sentiments associated with the data. The growth of internet has a special significance in online service.

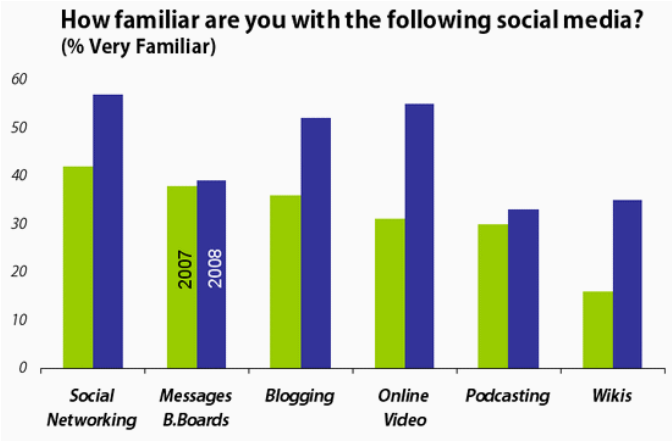


Figure 1.1: Depicting the Users of each type of social media platform

Type of Social Media Applications

There are many social media platforms which has become an integral part of people's lives these days. Some of the famous and most popular among them are:

- Facebook
- Twitter
- Linkedin
- Quora
- GooglePlus
- Youtube

The patter on of users in the specific platforms is shown in Fig. 1.2.

Social Networks Used by US Internet Users, by Age, July 2014
% of respondents in each group

	14-17	18-34	35-54	Total
Facebook	63.7%	83.2%	74.1%	76.8%
YouTube	81.9%	77.6%	54.2%	66.4%
Twitter	31.0%	38.7%	28.3%	32.8%
Instagram	56.4%	37.2%	16.0%	28.5%
Google+	24.6%	25.0%	20.4%	22.7%
LinkedIn	1.5%	15.9%	20.0%	16.6%
Snapchat	36.8%	21.1%	4.2%	14.2%
Tumblr	23.8%	15.6%	5.7%	11.5%
Vine	31.8%	15.5%	3.5%	11.1%
WhatsApp	8.0%	9.8%	4.0%	6.8%
reddit	8.0%	8.5%	3.9%	6.2%
Flickr	3.6%	3.9%	6.9%	5.4%
Pinterest	3.6%	2.0%	0.6%	1.5%

Note: n=1,093 ages 14-54; use at least once a week
Source: NuVoodoo as cited in press release, Aug 28, 2014
178953 www.eMarketer.com

Figure 1.2: User pattern of each social media

Advantages of Social Media

Social Media plays an important role in our life. Several merits of social media use include:

- Compelling and relevant content finds the attention of future customers and increase brand visibility
- Response facility to almost instantly to industry developments and become famous in your field
- It is very cheaper than traditional promotional and advertising activities
- Social content can indirectly encourage links to website content by appearing in general search results, improving search traffic and online sales
- Deliverance improved customer service and respond effectively to feedback
- Customers can find the seller easily through new channels, generating more leads
- Improved loyalty and advocacy from the contacted customers

Need for Analyzing Social Media Data

The use of social media is increasing day by day and this is represented by the no of monthly use as shown in Fig. 1.2. Increasing growth of social media users over internet has also increased their participation in various discussions and activities simultaneously. In case of a product, reviews of users will help to take many important decisions about the services of the product [12].

1. Sentiment Analysis

Sentiment analysis is a text classification problem which deals with extracting information present within the text. This extracted information can be then further classified according to its polarity as positive, negative or neutral.

2. *Text Mining*
   Text mining is branch of NLP (Natural Language Processing), i.e. used to extract automatically meaningful information from unstructured information which is usually textual data. This extracted information is transformed into numeric values and thereafter used by different data mining algorithms [8].
3. *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 develops algorithms and permit machine to evolve behaviors based on experimental data. In some special cases, it is difficult to represent an exact relationship from input to outputs. Machine learning is then expected to permit machines to adjust their algorithm in such a manner that it's expected future performance enhances [13].

KNN Algorithm

KNN is type of instance based learning or lazy learning. In this learning the function is approximately locally and all computation is deferred until classification. It is simplest of all machine learning algorithms. In KNN classification, the output is class membership. An object is classified by majority votes of its neighbors by the object being assigned to class most common among its k nearest neighbor (k is positive small integer). The nearest neighbor is determined using similarity measure usually distance functions are user. Following are the distance function used by KNN [15].

Feature selection methods

The mutual information measure provides a formal way to model the mutual information between the features and the classes. This measure was derived from the information theory. The point-wise mutual information (PMI) $M_{i(w)}$ between the word w and the class I is defined on the basis of the level of co-occurrence between the class I is defined on the basis of the level of co-occurrence between the class I and word w.

II. LITERATURE SURVEY

In this paper Ortigosa and Alvaro proposed a novel method for sentiment analysis in social site giant Facebook that, starting from the messages written by its users, supports: (i) to extract useful information about the Facebook users' sentiment polarity (whether it is positive, neutral or negative), which reflected from the messages written by users; and (ii) to model the users' normal sentiment polarity and to analyze significant emotional changes in user. Author has implemented this method in SentBuk which is a Facebook application also presented in the paper [1]. SentBuk fetch messages written by Facebook users and classify them accordingly to their polarity, giving the results to the users via an interactive interface. This app also supports emotional swing detection, friend's emotion detection and user's classification according to the messages retrieved, and statistics, among others. The classification method used in SentBuk is based on a hybrid approach: it combines two approaches: lexical-based and standard machine-learning techniques. The results acquired through this approach indicate that it is possible to perform sentiment analysis in Facebook easily with high accuracy (83.27%). In present scenario, with the advancement of e-learning, having information about the users' sentiments is very useful [10].

According to Pak and Alexander [2], Micro blogging is becoming a very popular communication and knowledge sharing tool among Internet users globally. Billions of users exchange their knowledge on different aspects of life each day. In this paper, Agarwal and Apoorv [3] explained one such popular micro blog named as Twitter and build models to classifying 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 second is a 3-way task of classifying sentiment of users into positive, negative and neutral. Author design a new representation for tweets, for the tree kernel based model. Author use a unigram model, which work well for sentiment analysis for Twitter data in the past. Result from paper indicates that a unigram model is really a hard baseline [4]. The major difference these subjective texts have with published news articles is that their target is unique and clearly stated across the text. Following different annotation efforts and the analysis of the issues encountered, we realized that news opinion mining is different from that of other text types [5].

According to Jebaseeli and A. Nisha, Opinion mining or Sentiment Analysis refers to identification and classification of the viewpoint or opinion expressed in the text span; using information retrieval and computational linguistics [6]. Twitter, one of the biggest and most popular social web site which contains unstructured data. In order to analysis such a data we need effective methodology which can process huge volume of data. Therefore, in this paper, Mahalakshmi R and Suseela S propose a method of sentiment analysis on twitter b y using Hadoop and its ecosystems that will process the large volume of data on a Hadoop and the MapReduce function will perform the sentiment analysis [7]. Social network analysis has been define in [8] as an assumption of the importance of relationships among interacting units, and the relations defined by linkages among units are a fundamental component of network theories. Social network analysis has emerged as a key technique in modern sociology. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, information science, organizational studies, social psychology, and sociolinguistics 1. In 1954, Barnes [8] started to use the term systematically to denote patterns of ties, encompassing concepts traditionally. Afterwards, there are many scholars expanded the use of systematic social network analysis. Due to the growth of online social networking site, online social networking analysis becomes a hot research topic recently.

Data processing layer deals with data collection and data mining, while sentiment analysis layer use a application to present the result of data mining [9].

In fact, companies manufacturing such products have started to poll these micro blogs to get a sense of general sentiment for their product. Many times these companies study user reactions and re ply to users on micro blogs. Social media continues to gain increased presence and importance in society. Public and private opinions about a wide variety of subjects are expressed and spread continually via numerous social media, with twitter being among the timeliest. Social media has become one of the biggest forums to express ones opinion. Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state (the emotional state of the author when writing), or the intended emotional communication (the emotional effect the author wishes to have on the reader [10].

III. PROPOSED METHODOLOGY

This chapter describes the various techniques we have applied for the fulfillment of our objectives as described in previous chapter. The various text mining algorithm and streaming of twitter api are given in this chapter. The first step starts with the extraction of tweets followed by preprocessing of the extracted tweets. Then Classifier algorithm has to be applied on it.

Data extraction: The twitter API named as 'tweepy' has been used in this thesis for the extraction step. The major steps involved in development of the framework for live streaming of tweets begin with setting up an account on twitter.

- Set up an account on twitter
- Go to 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 is used for streaming the tweets. Filters are applied on it using the track filter. The tweets are 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:

- Food price Crisis
- Food price
- Food Security
- Inflation
- Vegetable Prices
- Tomato Prices
- Onion prices
- Price Rise
- Onion Price

The location is done 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 thesis 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.
Text mining is applied on the filtered tweets for further processing.

Pre-processing: Pre-processing steps on textual description of bug reports are performed. It includes tokenization, stop word removal and stemming. 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 are performed to reduced terms into their root terms. Porter's stemming algorithm are used to perform stemming.

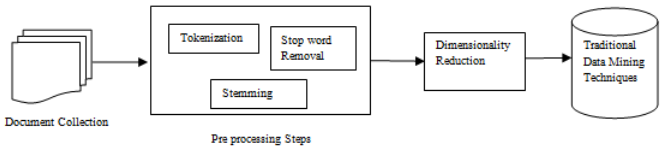


Figure 4.1: Text Mining

Steps in text mining: The different steps performed in text mining are as follows:

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

i. *Tokenization*: The purpose of tokenization is to remove all the punctuation marks like commas, full stop, hyphen and brackets. 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 eliminate conjunction, prepositions, articles and other frequent words such as adverbs, verbs and adjectives from textual data. Thus it reduces textual data and system performance is improved.

iii. *Stemming*: Stemming is used to reduce the words to their root words e.g. words like ”computing” ,”computed” and “computerize” has it root word “compute”. The purpose of stemming is to represent the words to only terms in their document. There are different algorithms to perform stemming such as Lovins Stemmer, Porters Stemmer, Paice/Husk Stemmer ,Dawson Stemmer, N-Gram Stemmer, YASS Stemmer and 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 multiplication of term frequency and inverse document frequency.

$$\mathbf{TF-IDE} = n_w^d \, log_2\left(\frac{N}{N_w}\right) \qquad (1)$$

Where n_w^d = frequency of word w in document d.
N= total document and N_w = document congaing word w.
v. *Term-document matrix* – After initial steps of preprocessing text in documents is converted into term- document matrix. Rows in matrix represents document in which word appears and columns represent the words that are extracted from documents. The cell of matrix is filled with TF-IDF score.
Mining the reduced data with traditional data mining techniques
Classification, clustering and predictive methods are applied to the reduced datasets using data mining techniques to analyze the pattern and trends within data.

Term -Document matrix- The Term- Document Matrix (TDM) is created. Each column in matrix represents the terms occurring in documents and row represents id of each bug report. The cells of matrix are filled with TF-IDF score. If term is not present in the particular bug reports then cell is filled with zero.
Dimensionality Reduction – After preprocessing steps, dimensionality reduction is performed. Here original TDM (term document matrix) is replaced with smaller matrix by using a SVD (singular value decomposition technique). This technique discards unimportant word and relevant and important word are filtered out. The new matrix is generated of terms and documents.
Feature Selection- Feature selection methods are used to retrieve the most informative terms from corpus of datasets. In our research, we have used two feature selection methods info gain and CHI square methods. These methods are applied on TDM matrix to reduce matrix.

Creating dictionary of terms- The terms obtained after applying feature selection are sorted in descending order according to their weights. The top m- terms are used for creating dictionary. The dictionary contains the terms that help in specifying the severity levels of each bug report.

Machine Learning Approaches
There are various approaches to design machine learning algorithms. The purpose of ML algorithms is to use observations as input and this observation can be a data, pattern and past experience. Thus Ml algorithms use to improve the performance of instances, which can be done by any classifier by trying to classify the input pattern into set of categories or to cluster unknown instances. As the nature of ML algorithms it enhances its performance from past experience or by receiving feedback. It can be divided into two categories supervised and unsupervised approach [9].
Supervised: In supervised learning, the instances are labeled with known or target classes labels. Here before classification the dataset knows the target class. Thus it is very helpful for the problems which have known inputs.
Unsupervised: In unsupervised learning, the algorithm groups the instances by their similarities in values of features and makes different clusters. In it no prior class or clusters are given, the algorithm itself defines their clusters automatically and statistically.

KNN Algorithm

KNN is type of instance based learning or lazy learning. In this learning the function is approximately locally and all computation is deferred until classification. It is simplest of all machine learning algorithms. In KNN classification, the output is class membership. An object is classified by majority votes of its neighbors by the object being assigned to class most common among its k nearest neighbor (k is positive small integer). The nearest neighbor is determined using similarity measure usually distance functions are user. Following are the 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 $\{(a_1, b_1), (a_2,b_2), (a_3,b_3) \ldots\ldots (a_N, b_N)\}$ is training datasets.

In KNN algorithm all the distance from testing point to training point are computed. Then these all testing points are sorted ascending order. Then class labels are added for each K nearest neighbors and sign of sum are used for prediction. The value of k in k-nearest neighbor is challenging task. As choosing smaller value of k. e.g. by choosing k=1 may lead to risk of over fitting and choosing larger value of k e.g. k=N may lead to under fitting. Therefore optimal value of k has been chosen between the values 3-10, which gives better result.

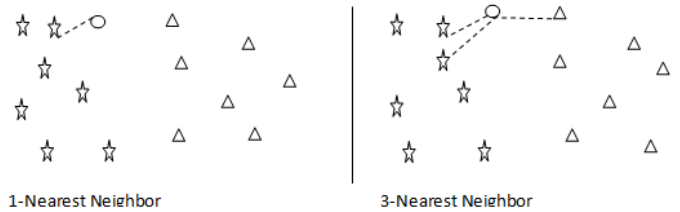


Figure 4.1 Working of KNN Algorithm

Algorithm: KNN (D, k, $\hat{x}$)

D is training dataset, N training examples are paired as $(x_1, y_1), (x_2, y_2) \ldots (x_N, y_N)$.

[] ⟶ an empty list and ⊕ ⟶ used to append in list.

Prediction on $\hat{x}$ (testing data point) is called $\hat{y}$

1. ⟵ S []
2. for n=1 to N do
   ⟵ $\hat{x}$
3. S S ⊕ <d (x_n,), n> // store distance to training example n
4. end for
5. ⟵ S SORT(S) // put lowest-distance objects first
6. ⟵ $\hat{y}$ 0
7. For K=1 to K do
8. ⟵ <dist, n> S_K // n this is the kth closest data point
9. ⟵ $\hat{y}$ $\hat{y}$ +y_n // vote according to the label for the nth training point
10. end for
11. return SIGN($\hat{y}$) // return +1 if $\hat{y}$ > 0 and 1 if $\hat{y}$ < 0

Application of K nearest neighbor

1. Nearest Neighbor based Content Retrieval- It is one of the important applications of K-Nearest neighbor e.g. if the content is video and it is used for retrieving videos that is closest to given video [11].
2. Protein-Protein interaction and 3D structure prediction- KNN is used to predict the structure of Gene and graph based KNN is used to predict the interaction of protein.

Naïve Bayes Algorithm

The algorithm is named after famous statistician Thomas Bayes who proposed Bayesian theorem. The Naïve bayes algorithm is also based on Bayesian theorem. This theorem assumes that all the attributes are conditionally independent to each other. In this algorithm, conditional probability for each 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

$$P\left(\frac{C_i}{X}\right) = P\left(\frac{C_j}{X}\right) \quad \text{for } 1<=j<=k,\ j <> i \qquad (1)$$

Posterior probability is calculated as

$$P\left(\frac{C_i}{X}\right) = \frac{P\left(\frac{X}{Ci}\right) P(c_i)}{P(X)} \qquad (2)$$

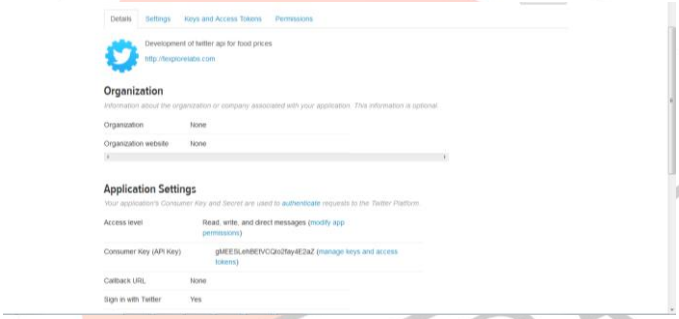
Application of Naïve Bayes

1. Text Classification- The classifier is well known for its most efficient learning capability for classification of text document [13].
2. Spam filtering –Spam filtering makes use of the classifier to identify spam mails and filter out them from legitimate mail. E-mail filter such as SpamBayes, SpamAssassin and Bogofilte are example of filter that uses Naïve bayes classifier.
3. Hybrid Recommender System- It is proposed a unique switching hybrid recommendation approach by combining a Naïve Bayes classification approach with the collaborative filtering.

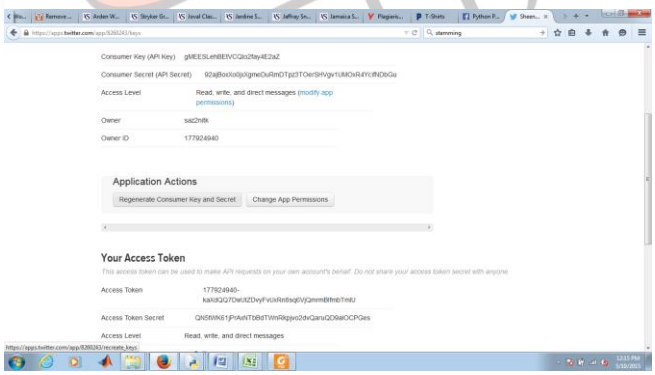
The results of the proposed method will be shown and discussed in the next chapter. The methodology has been designed for development of a framework to automatically provide public feedback for decision making regarding food related issues in India. The rise in food prices has been dealt with the public opinion regarding this will be shown in next Chapter.

RESULTS AND DISCUSSION

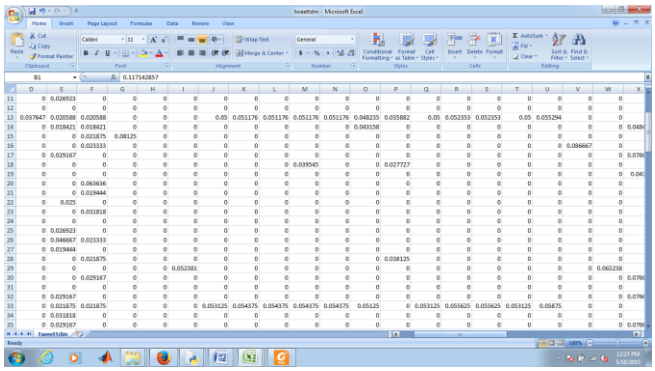
This Chapter shows the results obtained by various methodologies applied on the dataset discussed in the previous chapter. The results are verified by running the simulations for repeated number of times. The opinions are mined and analyzed fir public response.



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



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



• Naïve Bayes

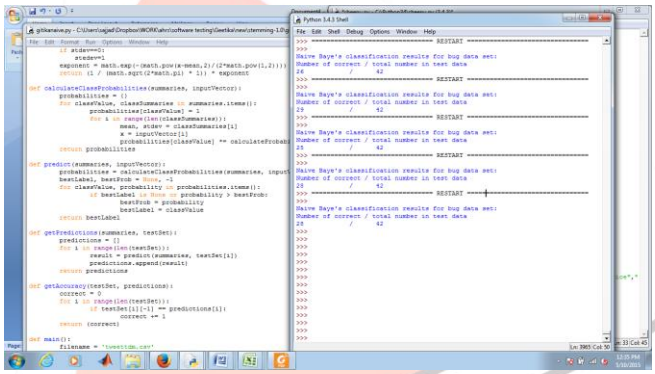
The result of Naïve Baye's Classifier is found to be 28 correct classified to that of total 42 tweets.
The Accuracy is calculated as :

Accuracy=28/42*100=66.66 %

• KNN

Value of K is taken as three and the result of KNN is found to be 29 correct tweets as compared to 39 total tweets.
The accuracy is calculated as:

Accuracy=29/39*100=76.31%



IV. CONCLUSION

A methodology for the classification of sentiments was developed in this thesis for food price crisis in Indian market. 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. Feature Selection was applied on it using Chi Square method and information gain. The extracted features forms a term document matrix which is utilized in the classification algorithm. Two classification algorithms are compared as shown in previous chapter.

The results are found to be satisfactory and when comparative analysis is done between them it is found that KNN outperforms Naïve Baye's Algorithm. Thus an automated system is designed for opinion mining related to food price crisis using Indian tweets.

V. 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 entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61{67, 1999.
[15] Mikheev, 1999 Andrei Mikheev. Feature lattics and maximum entropy models. Machine Learning, 1999.