**OPINION MINING ON YELP CUSTOMERS REVIEWS**

**Project Proposal**

**By:**

**Gargi Mrunal Kulkarni (CWID: 893210922)**

**Piyusha Kailash Zanjad (CWID: 893578021)**

**CPSC 531-01 (13298)**

**Spring, 2016**

**Professor: Dr. Chun-l Phillip Chen**

**Department of Computer Science**

**California State University, Fullerton**

**March 22, 2016**

**Table of Contents**

Abstract............................................................................................................................................5

Chapter 1: Introduction....................................................................................................................6

Chapter 2: Review of literature...........................................................................................................7

Overview of data mining.....................................................................................................7

Data Mining Techniques…………………………………….………..…..………….........8

Association…………………………………….………..………...….........8

Clustering …………………………………….………..………....….........8

Classification…………………………………….…………....…..….........8

Predication…………………………………….……….……..…..….........8

Decision Trees…………………………………….………..…..……........8

Opinion mining..............................................................................................................9

Opinion mining process and methods......…………………………………....10

Levels of Opinion mining……………………………………………………11

Supervised Machine Learning for opinion mining………………………….12

Support Vector Machine…………………….……………………………………...13

Separable data…………………………………………………………………………..14

Mathematical Formulation: Primal..……………………………………………..15

Mathematical Formulation: Dual...…...………………………………………….15

Non-Separable data………………………………………………………..……………..14

Mathematical Formulation: Dual...…...………………………………………….15

Nonlinear Transformation with Kernels……………………………………………………

Chapter 3: Methododogy...............................................................................................................16

Software and hardware requirements…….........................................................................16

System testing and validation procedures………………………………………………….

Unit Testing…….………………………………………………………………..

Functional Testing……………………………………………………………….

Performance Testing……………………………………………………………..

System installation………………………………………………………………………….

Chapter 4: Results and discussion………………………………………………………………

Chapter 5: Conclusion, Implications and recommendation……………………………………

References.....................................................................................................................................21

**List of Figures**

1. Figure 1: Knowledge discovery process….……………………….…………..……….……… 8

2. Figure 2: Data mining techniques……………………………………………………….…... ...9

3. Figure 3: Sentiment Analysis Process……………………………….….…..……….………...10

4. Figure 4: Sentiment Analysis methods ……………………………………………………….11

5. Figure 5: Levels of sentiment analysis………………………………………………………...11

6. Figure 6: Supervised classifier………………………………..……………...…………..……13

7. Figure 7: Naïve Bayes classifier…………………………..……………...…………………...14

8. Figure 8: Apache Mahout architecture…………………..……………...……………….……15

9. Figure 9: Algorithms present in Apache Mahout…………..……………...………….………16

10. Figure 10: Sentiment Analysis using Apache Mahout……..……………...……...…………17

**Chapter 1: Introduction**

Data Mining is an emerging technology which performs analysis of structured or unstructured data, to identify patterns and extract useful information from the same for decision making. Today, data to be dealt with is in large volume, Big Data. The traditional data analytical tools fail to extract the complete useful information from this Big Data. Thus, new Data Mining and Machine Learning algorithms and tools are invented for mining of big data to gain knowledge and apply them for decision making. There are several data mining algorithms invented to mine this big data to obtain valuable knowledge from it. Some of the data mining techniques are Association, Classification, Clustering, Prediction, Sequential patterns, decision trees, Combination, Long term processing etc. The data mining tools implement these algorithms for analysis and decision making. Different data mining tools available are Weka, Orange, Rapid Miner, R-Programming, and Apache Mahout Etc.

When data to be mined is in form of text, it is known as Text Mining. Sentiment Analysis is a form of text mining, natural language procession and prediction. It is also known as Opinion Mining, in which the person’s mood is predicted to be of ‘positive’, ‘negative’ or ‘neutral’ polarity. It is the study of person’s opinion, sentiments, emotions, attitude towards entities like organizations, services, individuals, issues/topics, events etc. Sentiment analysis plays an important role in every field from computer science to management and social science. Sentiment analysis has proved to be beneficial in enterprises/industries, business, politics, social media, etc.

In today’s world, customer service plays a vital role in the success of any business. A leading bank, Barclays, recently launched a customer service portal for customers above 18 years of age. The service got criticized online, specifically by youth under the age of 18. A series of negative comments on the internet gave a clear review about the service to Barclays. The service ended up being modified by making it accessible to customers of all age. Thus is the value of customer service in any business and thus opinion mining demands high levels of significance.

There are various methodologies [with different algorithms and data mining tools] in which sentiment analysis can be carried out. In this Project we are mining the Yelp Dataset to classify the reviews as positive or negative depending on the review stars. Yelp is an American multinational corporation that publishes customers’ reviews for local business, SeatMe and Eat24. The algorithm used in this project is SVM (Support Vector Machine) to classify reviews as positive or negative. The SVM classifier is implemented in Python and the results are compared with its Weka Tool counterpart. The study is beneficial for various businesses that post their products or services on Yelp as it would estimate a general public response regarding advertisements and business postings.

Thus, the project would work as a great tool for opinion and feedback analysis i.e. opinion mining which would lead to creation of better business solutions for any organization.

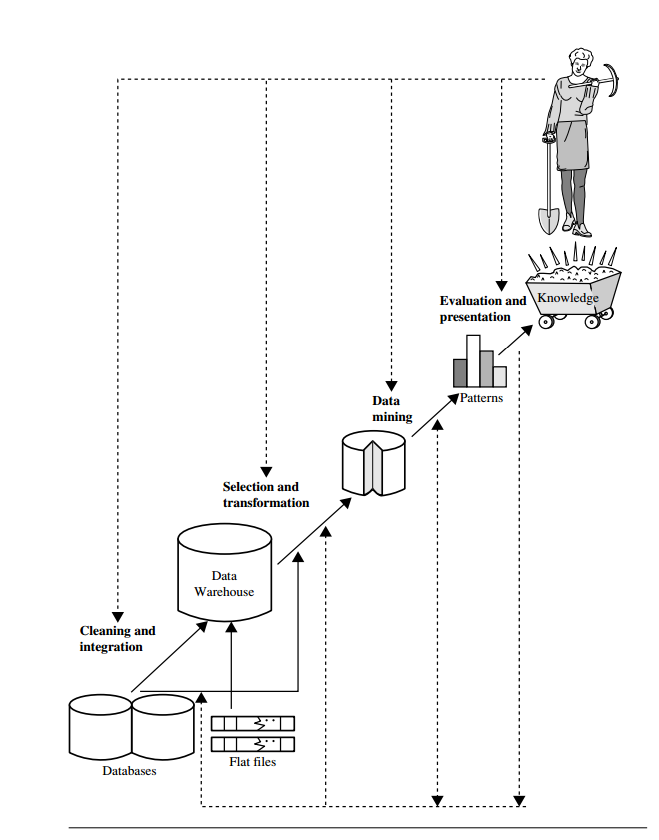
**Chapter 2: Review of Literature Review**

**Overview of Data Mining**

Data mining, also known as data or knowledge discovery, is referred to as the process of analyzing data from different point of views and to summarize that data into some useful information. It can also be defined as the process of extracting knowledge from large data sets. The technical definition of data mining would be the process of finding the patterns among large number of fields in huge relational databases. Data mining software is one of the various analytical tools used for analyzing the data which can be used to predict the future trends permitting businesses to take knowledge-driven decisions. The data mining process includes a various number of steps for discovering the knowledge, which are described below (Han, J., Kamber, M., & Pei, J., 2012):

1. Data cleaning and integration: In this step, the unwanted inconsistent data is removed and data from various sources is combined.
2. Data selection and transformation: In this step, the data applicable for analysis is collected and is transformed to required format for data mining
3. Data mining: In this step, the learning methods are applied to data to extract learn patterns in the data. This is an important step in data mining.
4. Evaluation and presentation: In this step, the learned patterns are evaluated to gain knowledge and this gained information is presented to the user.

The figure below represents the knowledge discovery process.



*Figure 1.* Knowledge discovery process (Source: Han, J., Kamber, M., & Pei, J., 2012)

Though Data Mining is the important step in knowledge discovery process, in every field the term ‘Data Mining’ is used to refer the entire process. Thus Data Mining can be defined as the process of learning patterns from a large amount of data in order to gain knowledge.

**Data mining techniques.** Data sources for mining can be traditional databases, data warehouses, web, other data storage systems etc. Data from all this sources is in a variety of forms. There are various techniques implemented to mine this large volume varied data. We will briefly examine each of the data mining techniques below. (Han, J., Kamber, M., & Pei, J., 2012)

***Association.*** It is one of the best data mining technique. In this, a pattern is discovered based on the relationship between items in same transaction. Retailers use association techniques to predict the customers buying habits. It includes different types such as multi-level, multidimensional and quantitative.

***Clustering.*** It is a technique that makes a meaningful cluster of objects which have similar characteristics. For example, by using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books about a particular topic, they would only have to go to that shelf instead of looking for the entire library. It includes different types such as partitioning methods, hierarchical agglomerative methods, density based methods, grid based methods, and model based methods.

***Classification.*** It is classic data mining technique based on machine learning. It is used to classify each item in a set of data into one of a predefined set of classes or groups. In this method, a model with training dataset is developed and the outcome of the datasets is predicted based on the trained model. For example, fraud detection. It includes Bayesian classification, Support Vector Machine (SVM), Neural Networks, classification based on associations.

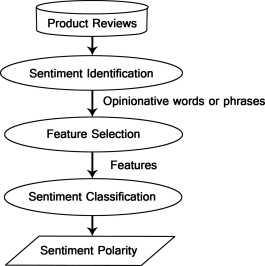
***Prediction.*** It finds the relationship between independent variables and relationship between dependent and independent variables. Results are predicated using previously known variables. For example, it can be used in the sales to predict profit. It includes linear regression, multi-variate linear regression, non-linear regression, multi-variate non-linear regression.

***Decision trees.*** In decision tree technique, the root of the decision tree is a simple question having multiple answers. Each answer then leads to a set of questions that help us determine the data so that we can make the final decision based on it. For example, decision tree can be used to determine whether or not to play tennis. It includes classification and regression tree (CART).

**Opinion mining (or Sentiment Analysis)**

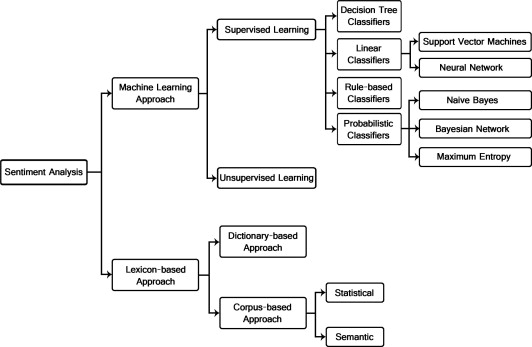
Opinion mining also known as sentiment analysis, is the study of the person’s opinion about any entity. It is a natural language processing in which the opinion of the person is classified to be of positive, negative or neutral polarity. The data source for sentiment analysis is varied, from web [blogs, social media, online customer reviews] to hand written documents. It uses different machine learning algorithms or lexicons based methods to extract words/phrases from the text and classify them as positive or negative or neutral opinion (Liu, B., 2012).

**Opinion mining and methods.** The following figure shows the opinion mining process for product reviews as the input text:



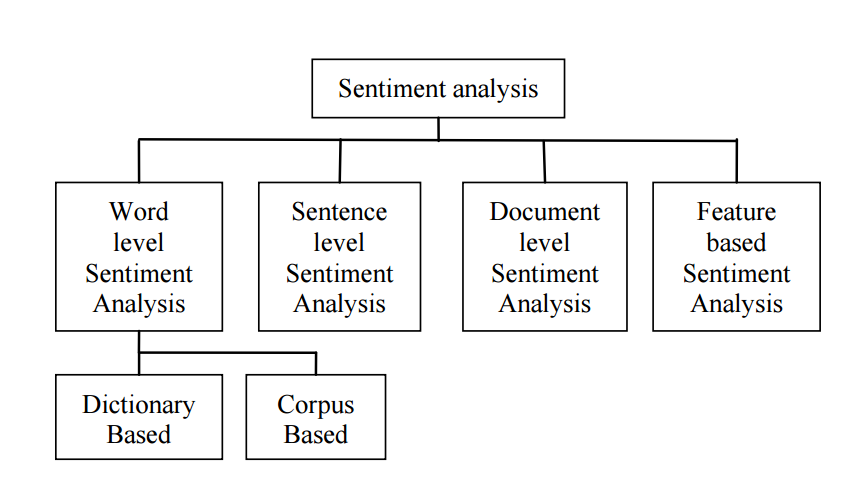
*Figure 2*. Opinion mining or sentiment analysis process (Source: Sentiment analysis algorithms and applications: A survey. (n.d.))

The product reviews, the words/phrases are extracted from the input in Sentiment Identification. Depending on the features [POS tags N-grams, frequency, negations etc.] of the words, they are selected in Feature Selection creating the feature vector. These features are then classified in Sentiment Analysis to detect the Sentiment Polarity (Liu, B., 2012). The sentiment classification methods are given in the figure below:



*Figure 3*. Sentiment Analysis methods (Source: Sentiment analysis algorithms and applications: A survey. (n.d.))

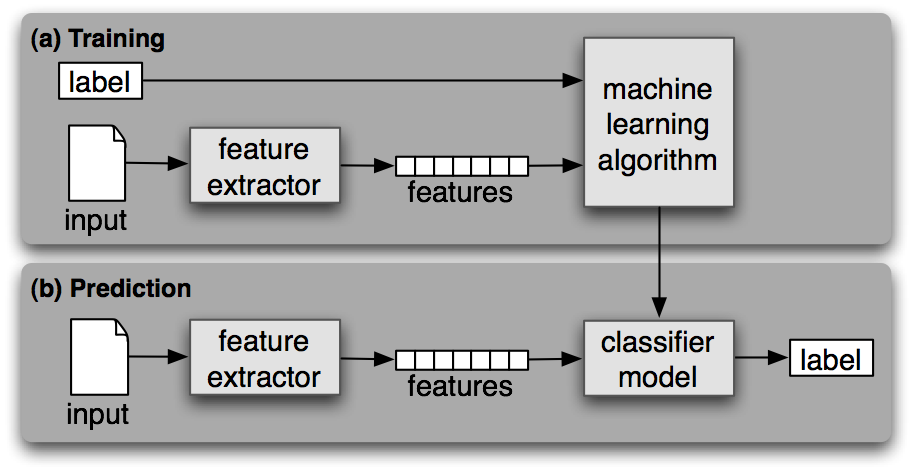
**Levels of Opinion Mining or Sentiment Analysis.** Opinion mining or Sentiment analysis can be carried out at different levels that is word level, sentence level, document level and aspect level. The figure below shows the different levels.



*Figure 4*. Levels of Opinion mining or Sentiment Analysis (Source: Sentiment analysis algorithms and applications: A survey. (n.d.))

Document level sentiment analysis calculates the opinion presented by the subjective and objective sentences in the document as a whole. At Sentence Level, the subjective sentences are analyses and opinion polarity is detected. The word level sentiment analysis forms the basis of document and sentence level sentiment analysis as the phrases in sentences are analyzed for positivity or negativity of the sentiment. There two approaches in which word level sentiment analysis is carried out: one, the Dictionary based approach, in which the small list of words is created with known polarity and the list in extended with synonyms and antonyms [assigning each to appropriate polarity] from online words source, an online dictionary and other is Corpus based approach in which the word polarity is determined from its co-occurrence with another known polarity word, relying on syntactic and statistic methods. In aspect level or feature based sentiment analysis the person’s opinion about the features of the entity is analyzed, which can’t be extracted from sentiment or document level analysis, as different person can have different opinion about the features of the entity (Liu, B., 2012).

**Supervised Machine Learning.** Supervised learning works on classification. In classification process, the class label is assigned to the input. The class labels are pre-defined. The supervised classifier is built based upon the training set of inputs, corpora, having specific label for each input. Input are considered to be independent from each other. The figure below shows the Supervised Classification:



*Figure 5*. Supervised classifier (Source: Bird, S., Klein, E., & Loper, E., 2009)

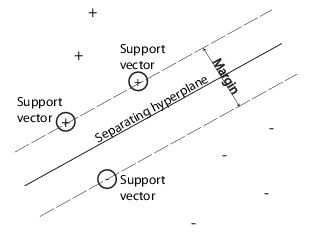
Supervised Classification is divide into two steps. In first Training step, in which the model, a pair of labels and features, is generated by machine learning algorithm. The features of the known input are extracted by feature extractor creating a feature vector, which is then fed to machine learning algorithm along with label to develop the classifier model. In second Prediction step, the unknown input is fed to same feature extractor which extract features and gives to classifier model to predict labels (Wu, X.et al, 2007; Bird, S., Klein, E., & Loper, E., 2009).

**Support vector machine**

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. The operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. Twice, this distance receives the important name of **margin** within SVM’s theory. Therefore, the optimal separating hyperplane maximizes the margin of the training data.

SVM contains two types of classifiers i.e., linear classifier and non- Linear classifier .Linear classifiers are simple and computationally efficient. However for nonlinearly separable features, they might lead to very inaccurate decisions. We may trade simplicity and efficiency for accuracy using a nonlinear classifier. (Support Vector Machines for Binary Classification. (n.d.))

**Separable Data.** Support vector machine (SVM) can be used when the data has exactly two classes. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points. The support vectors are the data points that are closest to the separating hyperplane; these points are on the boundary of the slab. The following figure illustrates these definitions, with + indicating data points of type 1, and – indicating data points of type –1.



*Figure 6*. Separating hyperplane in SVM (Source: Support Vector Machines for Binary Classification. (n.d.))

***Mathematical Formulation: Primal.*** The data for training is a set of points (vectors) *xj* along with their categories *yj*. For some dimension *d*, the *xj* ∊ *Rd*, and the *yj* = ±1. The equation of a hyperplane is

*f* (*x*)=*x*′*β*+*b*=0

where *β* ∊ *Rd* and *b* is a real number.

The following problem defines the *best* separating hyperplane (i.e., the decision boundary). Find *β* and *b* that minimize ||*β*|| such that for all data points (*xj*, *yj*),

*Yj f (Xj)*. ≥1.

The support vectors are the *xj* on the boundary, those for which *Yj f (Xj)*. ≥1.

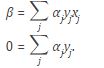
For mathematical convenience, the problem is usually given as the equivalent problem of minimizing ||*β*|| .This is a quadratic programming problem. The optimal solution (ˆ*β*,ˆ*b)*enables classification of a vector *z* as follows:



ˆ*f* is the *classification score* and represents the distance *z* is from the decision boundary.

***Mathematical Formulation: Dual.***It is computationally simpler to solve the dual quadratic programming problem. To obtain the dual, take positive Lagrange multipliers *αj* multiplied by each constraint, and subtract from the objective function:





Substituting into *LP*, you get the dual *LD*:



which you maximize over *αj* ≥ 0. In general, many *αj* are 0 at the maximum. The nonzero *αj* in the solution to the dual problem define the hyperplane, which gives *β* as the sum of *αjyjxj*. The data points *xj* corresponding to nonzero *αj* are the *support vectors*.

The derivative of *LD* with respect to a nonzero *αj* is 0 at an optimum. This gives



In particular, this gives the value of *b* at the solution, by taking any *j* with nonzero *αj*.

The dual is a standard quadratic programming problem. For example, the Optimization Toolbox™ [quadprog](http://www.mathworks.com/help/optim/ug/quadprog.html) solver solves this type of problem.

**Nonseparable Data.** The data might not allow for a separating hyperplane. In that case, SVM can use a *soft margin*, meaning a hyperplane that separates many, but not all data points. There are two standard formulations of soft margins. Both involve adding slack variables *ξj* and a penalty parameter *C*.

The *L* 1-norm problem is:



Such that,



The *L*1-norm refers to using *ξj* as slack variables instead of their squares. The three solver options SMO, ISDA, and L1QP of [fitcsvm](http://www.mathworks.com/help/stats/fitcsvm.html) minimize the *L*1-norm problem.

The *L*2-norm problem is:



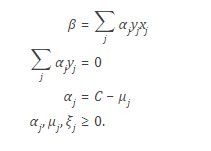
subject to the same constraints.

In these formulations, you can see that increasing *C* places more weight on the slack variables *ξj*, meaning the optimization attempts to make a stricter separation between classes. Equivalently, reducing *C* towards 0 makes misclassification less important.

***Mathematical Formulation: Dual.***For easier calculations, consider the *L*1 dual problem to this soft-margin formulation. Using Lagrange multipliers *μj*, the function to minimize for the *L*1-norm problem is:



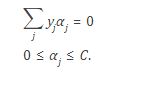
where you look for a stationary point of *LP* over *β*, *b*, and positive *ξj*. Setting the gradient of *LP* to 0, you get



These equations lead directly to the dual formulation:



subject to the constraints



The final set of inequalities, 0 ≤ *αj* ≤ *C*, shows why *C* is sometimes called a *box constraint*. *C* keeps the allowable values of the Lagrange multipliers *αj* in a "box", a bounded region.The gradient equation for *b* gives the solution *b* in terms of the set of nonzero *αj*, which correspond to the support vectors.

**Nonlinear Transformation with Kernels.** Some binary classification problems do not have a simple hyperplane as a useful separating criterion. For those problems, there is a variant of the mathematical approach that retains nearly all the simplicity of an SVM separating hyperplane.

This approach uses these results from the theory of reproducing kernels:

There is a class of functions *G*(*x*,*y*) with the following property. There is a linear space *S* and a function *φ* mapping *x* to *S* such that

*G*(*x*,*y*) = <*φ*(*x*),*φ*(*y*)>.

The dot product takes place in the space *S*.

This class of functions includes:

Polynomials: For some positive integer *d*,

*G*(*x*,*y*) = (1 + *x*′*y*)*d*.

Radial basis function (Gaussian): For some positive number *σ*,

*G*(*x*,*y*) = exp (–(*x*–*y*)′(*x* – *y*)/(2*σ*2)).

Multilayer perceptron (neural network): For a positive number *p*1 and a negative number *p*2,

*G*(*x*,*y*) = tanh (*p*1*x*′*y* + *p*2).

The mathematical approach using kernels relies on the computational method of hyperplanes. All the calculations for hyperplane classification use nothing more than dot products. Therefore, nonlinear kernels can use identical calculations and solution algorithms, and obtain classifiers that are nonlinear. The resulting classifiers are hypersurfaces in some space *S*, but the space *S* does not have to be identified or examined.

**Chapter 3: Methodology**

The basic idea of the project is to classify the reviews into two categories i.e. positive or negative and based on this predicting the nature of future reviews. For this, we are using SVM (Support Vector Machine) algorithm for the classification. We are implementing the SVM classifier using python’s scikit-learn library. Also, we will be using NLTK (Natural Language Toolkit) library in python which is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python programming language. Along with the implementation in python, we are also using a tool known as “WEKA” and the results from both will be compared. Weka is a collection of machine learning algorithms for data mining tasks. The data which we are using for data mining is collected from the website of “Yelp”. The data collected is of customer reviews in JSON format.

**Hardware and Software requirements**

Operating System: Windows 7

Processor: 1 gigahertz (GHz) or faster 32-bit or 64-bit processor

RAM: 1 gigabyte (GB) RAM (32-bit) or 2 GB RAM (64-bit)

Hard disk space: 16 GB available hard disk space (32-bit) or 20 GB (64-bit)

Tools: Weka (Version 3.6)

Languages: Python (Version 3.5.1)

Libraries used in python: NLTK(Version 3.0) , scikit-learn (Version 0.17.1 ).

**System testing and validation procedures**

After the successful implementation of the application, we will test and validate it using the following testing methodologies.

**Unit testing.** Testing of individual software components or modules. Typically done by the programmer and not by testers, as it requires detailed knowledge of the internal program design and code may require developing test driver modules or test harnesses.

**Functional testing.** This type of testing ignores the internal parts and focus on the output is as per requirement or not. Black-box type testing geared to functional requirements of an application.

**Performance testing.** Term often used interchangeably with ‘stress’ and ‘load’ testing. To check whether system meets performance requirements. Used different performance and load tools to do this.

**System installation**

1. Download and install python using following link:

<https://www.python.org/>

1. Python libraries (scikit-learn and nltk) can be installed using following links:

<https://pypi.python.org/pypi/nltk>

<http://scikit-learn.org/stable/install.html>

1. Download and install WEKA tool using the following link:

<http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

**Chapter 4: Results and Discussion**

**Implementation using python**

For the opinion mining on yelp dataset, we have used SVM (Support Vector Machine) algorithm for the classification of customer reviews. Thus, the reviews can be classified as either positive or negative and nature of the next reviews are predicted. For this, we have implemented SVM in python using scikit-learn and NLTK library. The data was collected from the official website of yelp. The approach followed while doing the classification is given as follows:

1. The dataset was loaded and the reviews were separated into the classes of positive and negative based on the stars of rating.
2. The rating is out of 5 stars. So, if the stars are greater than “3”, then they are categorized as “1” else categorized as “0”.
3. The next step is to create a feature vector. A feature vector is an n-dimensional vector of numerical features that represent some object.
4. Then, we will build the classifier.
5. The next step is to do 5 fold cross-validation. Cross-validation (CV) is a standard technique for adjusting hyper parameters of predictive models. In K-fold CV, the available data S is partitioned into K subsets S1, …, SK.

Also, five class classification was implemented using SVM classifier. The output after running the python script is shown as follows:

**Implementation using weka tool**

Using the weka tool, we have performed the support vector machine algorithm on the yelp dataset. The data was available in JSON format. Hence, for processing it in WEKA, it was convert into CSV (Comma Separated Values) format. The analysis can be performed using one of the two libraries such as LibSVM and SMO.

They are both implementations of optimizers to find linear and non-linear support vector machines. SMO implements Platt's Sequential Minimal Optimization algorithm whereas LibSVM implements an SMO-like method but works faster than SMO.

Weka results output:

TP = true positives: number of examples predicted positive that are actually positive

FP = false positives: number of examples predicted positive that are actually negative

TN = true negatives: number of examples predicted negative that are actually negative

FN = false negatives: number of examples predicted negative that are actually positive

Weka Confusion Matrix if a is taken to be the positive class:

a b <-- classified as

Actual a=0 TP FN

Actual b=1 FP TN

Weka Confusion Matrix if a is taken to be the negative class:

a b <-- classified as

Actual a=0 TN FP

Actual b=1 FN TP

Recall is the TP rate also known as sensitivity.

Precision is TP / predicted Positive, also referred to as Positive predictive value.

**Chapter 5: Conclusion, Implications and recommendations**

**Conclusion**

On performing data mining on the yelp dataset using support vector machine algorithm, we successfully classified the stars as positive or negative. By using python, SVM got implemented with ease using NLTK libraries giving optimal performance and it classified the data with high precision. Also, by using the weka tool, it was not only easy to implement the algorithm without actually writing the algorithm but also it was easy to obtain optimum results. We studied different algorithms on the same dataset, but support vector machine was the best out of all the algorithms in terms of classification.

**Implications**

Opinions are so important that whenever one needs to make a decision, one wants to hear others’ opinions. This is true for both individuals and organizations. The technology of opinion mining thus has a tremendous scope for practical applications.

**Individual consumers.**  If an individual wants to purchase a product, it is useful to see a summary of opinions of existing users so that he/she can make an informed decision. This is better than reading a large number of reviews to form a mental picture of the strengths and weaknesses of the product. He/she can also compare the summaries of opinions of competing products, which is even more useful.

**Organizations and businesses.** Opinion mining is equally, if not even more, important to businesses and organizations. For example, it is critical for a product manufacturer to know how consumers perceive its products and those of its competitors. This information is not only useful for marketing and product benchmarking but also useful for product design and product developments.

**Recommendations**

The opinion mining can be used in the following major areas (Rahmat, H. P., 2014):

1) Purchasing Product or Service: While purchasing a product or service, taking right decision is no longer a difficult task. By this technique, people can easily evaluate other’s opinion and experience about any product or service and also he can easily compare the competing brands. Now people don’t want to rely on external consultant. The Opinion mining and sentiment analysis extract people opinion form the huge collection of unstructured content, the internet, and analyze it and then present to them in highly structured and understandable manner.

2) Quality Improvement in Product or service: By Opinion mining and sentiment analysis the manufactures can collect the critic’s opinion as well as the favorable opinion about their product or service and thereby they can improve the quality of their product or service. They can make use of online product reviews from websites such as Amazon, RottenTomatoes.com and IMDb.

3) Marketing research: The result of sentiment analysis techniques can be utilized in marketing research. By sentiment analysis techniques, the recent trend of consumers about some product or services can be analyzed. Similarly the recent attitude of general public towards some new government policy can also be easily analyzed. These all result can be contributed to collective intelligent research.

4) Recommendation Systems: By classifying the people’s opinion into positive and negative, the system can say which one should get recommended and which one should not get recommended.

5) Opinion spam detection: Since internet is available to all, anyone can put anything on internet, this increased the possibility of spam content on the web. People may write spam content to mislead the people. Opinion mining and sentiment analysis can classify the internet content into’ spam’ content and ‘not spam’ content.

6) Decision Making: People’s opinion and experience are very useful element in decision making process. Opinion mining and Sentiment analysis gives analyzed people’s opinion that can be effectively used for decision making.

**References**

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (n.d.). *An Overview of Sentiment Analysis in Social Media and its Applications in Disaster Relief* (Tech.).

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly.

Fan, W., & Bifet, A. (2013). *Mining Big Data: Current Status, and Forecast to the Future* (Tech.). SIGKDD Explorations.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques*. Morgan Kaufmann.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.

Owen, S., Anil, R., Dunning, T., & Friedman, E. (2012). *Mahout in Action*. Manning Publications.

Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis* (Tech.).

Qu, Y., Wu, P. F., & Wang, X. (2009). *Online Community Response to Major Disaster: A Study of Tianya Forum in the 2008 Sichuan Earthquake*. HICSS.

Natural Language Toolkit. (n.d.). Retrieved from http://www.nltk.org/

Scikit-learn Machine Learning in Python. (n.d.). Retrieved from http://scikit-learn.org/stable/

Support Vector Machines for Binary Classification. (n.d.). Retrieved from

<http://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html#bsr5oqx>

Sentiment analysis algorithms and applications: A survey. (n.d.). Retrieved from http://www.sciencedirect.com/

Weka 3: Data Mining Software in Java. (n.d.). Retrieved from

http://www.cs.waikato.ac.nz/ml/weka/

Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H. Steinberg, D. (2007). *Top 10 algorithms in data mining*. Springer.

Rahmat, H. P. (2014). Opinion Mining and Sentiment Analysis - Challenges and Applications. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 3(3), may 2014.