iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** XII **Month of publication:** December 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

An Opinion Mining By Manhattan Clustering Using Decision Tree Feature Selection

P. Kavya Sri[1], K.C. Ravi Kumar[2], Dr. S. Anitha Reddy[3]
M.Tech, Associate Professor , Professor & HOD , Department of CSE, JNTUH, Hyderabad, AP, INDIA

***Abstract:** In present era Opinion mining plays a major character in text mining applications, and people are more depend on the web for many actions like purchasing, investment, business markings, etc. These applications led to a young generation of companies and products meant for online market awareness, online content monitoring and reputation management. Opinion mining is a procedure in which it deals with impressions, sentiments, and the subjectivity of text. The scalable distance based Clustering Algorithm enables the identification of topics within discussions in web social networks and their development. The predefined set of clustering is valuable in web opinion clustering. Features are extracted from the data for classifying the sentiment. Feature selection has gained importance due to its contribution to save classification cost with regard to time and computation load. This paper is an attempt to review and evaluate the various techniques used for opinion mining analysis and the main focus is on feature selection for Opinion mining using decision tree based feature selection. The suggested method is evaluated using IMDb data set, and is compared with Principal Component Analysis (PCA).*

***Keywords:** Opinion Mining, Hirarichal Clustering,Feature Selection ,Imdb, Inverse Document Frequency (IDF), opinion classification , Principal Component Analysis (PCA), Leaningr Vector Quantization(LVQ).*

I. INTRODUCTION

Opinion mining (or sentiment analysis) has attracted great interest in recent years, both in academia and industry due to its potential applications. A text understanding technology, Opinion mining assists people locate relevant opinions in a large review collection volume. One of the most promising applications is analysis of opinions in social networks.

An opinion mining technology based search engine shows potential to address this issue. Lots of people write their opinions in forums, microblogging or review websites. This data is very useful for business companies, governments, and individuals, who want to track automatically attitudes and feelings in those sites. An opinion-mining tool pores over product reviews for extraction of opinion units saving them in opinion databases. Opinion Mining is the process to extract the opinions expressed in user-generated content. Basically the opinions are categorized into two types:

1. Direct - Sentiment Expressions on some objects. Eg: products, topics, persons etc.
2. Comparisons to find the similarities and differences between more than one object. Eg: car x is cheaper than car y.

The content of the message is fewer focused. Messages are usually short in length ranging from a few words to couple paragraphs. The terms used in the messages are thin because different users may use different conditions to discuss the same topic. The volume of web opinion messages is vast and ever increasing in a massive rate. Opinion mining in a broad sense is defined as the computational study of opinions, sentiments and emotions expressed in texts. Opinions exist on the Web for any entity or object (person, product, service, etc.), and for the features or components of these objects, like, a cell phone battery, keyboard, touch screen display, etc.

Opinion-search engines arrange information based on opinion and not on the document. Hence, product review information is accessed quickly/easily. A very basic step of opinion mining and sentiment analysis is feature extraction. Figure 1 shows the process of opinion mining and sentiment analysis.

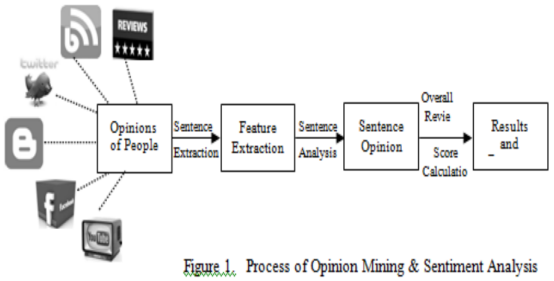


Figure 1. Process of Opinion Mining & Sentiment Analysis

The conventional document clustering techniques that work well in clustering regular documents usually do not work well in web opinion clustering. The conventional clustering characteristics like assigning all documents into clusters or having predefined set of clusters may not be applicable to web opinion content Analysis. Opinions and its related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining, with the sudden growth of social media (e.g., reviews, forum discussions, blogs, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision making . State-of-the-art opinion mining techniques are divided into 2 camps, i.e. attribute-driven methods and sentiment-driven methods.

Their basic idea is to use either attribute or sentiment keyword to locate opinion candidates through application of certain opinion patterns (involving attributes/sentiment keywords) for extraction of sentiment expressions filtering false opinion candidates. A drawback with this method is that they yield higher *precision* at the cost of large *recall* loss as generalization ability is not implied. The problem is mainly caused by out-of-vocabulary (OOV) attributes and OOV sentiment keywords being encountered in natural language review text. There are many challenges to Sentiment analysis. The first is an opinion word considered positive in one situation and negative in another. The second challenge is that people express opinions in various ways. Conventional text processing is based on the fact that limited differences can be identified between two text pieces which does not change meaning much.

A. Feature-based sentiment analysis

This discovers targets on which opinions were expressed in a sentence, and determines whether opinions are positive/negative or neutral. An object could be a service, product, organization, individual, topic, event etc.

B. The problem of sentiment analysis

A scientific problem has to be defined before it is solved to formalize it. Formulation introduces basic definitions, core concepts/issues, sub-problems and target objectives. It is also a framework to unite different research directions.

C. Feature extraction

Feature based opinion mining mainly concentrates on specific features related to that particular entity taken in to consideration, based up on those features the entire performance capabilities are dependable.

II. BACKGROUND WORK

The task of manually scanning huge amounts of reviews is computationally burdensome and not practically implemented regarding businesses/customer perspectives. Online customer reviews are a significant informative resource useful for both potential customers and product manufacturers. Reviews are written in natural language and are unstructured-free-texts scheme in web pages. Opinion mining extracts tasks from documents, opinions as expressed by sources on a target. A comparative study on methods used for mining opinions from the newspaper article quotations. Its difficulty in being motivated by various possible targets and variety. Faster and accessible internet ensures that people search/learn from fragmented knowledge. Generally, huge volumes of documents and homepages or learning objects are returned by search engines without any specific order.

A. Sentiment Analysis

Opinions are central to almost all human activities because they are key influencers of our behaviors. Whenever we need to make a decision, we want to know others' opinions. In the real world, businesses and organizations always want to find consumer or public opinions about their products and services.

Individual customers also want to be familiar with the opinions of existing users of a product before purchasing it, and others' opinions about political candidates before making a voting decision in a political election. In the past, when an individual needed opinions, he/she asked friends and family. Web, individuals and organizations are increasingly using the content in these media for decision making. When an organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groupsWith the explosive growth of social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Acquiring public and consumer opinions has long been a huge business itself for marketing, public relations, and political campaign companies. Feature selection has gained importance due to its contribution to save classification cost with regard to time/computation load. Searching for essential features, a feature search method is through decision trees.

A new matrix learning scheme extending the Relevance Learning Vector Quantization (RLVQ), to a general adaptive metric.

By introducing a full relevance factors matrix in distance measure, correlations between features and classification scheme importance are considered and automated, a general metric adaptation happens during training. Large margin generalization bounds are transferred to this, leading to input dimensionality independent bounds. The algorithm was tested and compared to alternative LVQ schemes using artificial data set, benchmark UCI repository multi-class, and an issue from bioinformatics, recognition of splice sites for C.Compared to weighted Euclidean metric used in RLVQ and its variations, a total matrix powerfully represents data's internal structure correctly.This includes local metrics attached to all prototypes corresponding to piecewise quadratic decision boundaries.

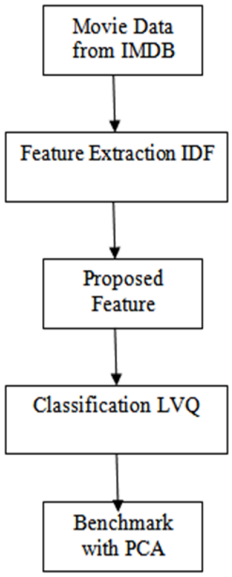


Figure 1: Flowchart Of Proposed Method

III. METHODOLOGY

The flowchart of the proposed methodology is shown in Figure 1 and the following sections details the steps in the proposed methodology.

A. IMDb Database

The IMDb is a large database with relevant and comprehensive information on movies- past, present and future. The latter was a collection of email messages between users of rec.arts.movies Usenet bulletin board.It began as a shell scripts set and data files. At some point, such data files became searchable with commands built by shell scripts. IMDb uses two methods to add information to a database: Web forms and e-mail forms. Information from submission procedures indicates that, it is simpler to use web forms rather than e-mail format, if only addition to information is an update.

B. Inverse Document Frequency (IDF)

Inverse document frequency (IDF) is a numerical statistic showing the importance of a word to a document, in a collection/corpus. It is used as a weighting factor in information retrieval/text mining. IDF value increases with the repeated appearance of a word in a document. IDF measures a word's ability to discriminate documents. Text Classification assigns a text document to a pre-defined class set automatically, using machine learning. Classification is based on significant words/key-features of text document. As classes are pre-defined, it is a supervised machine learning process.
Inverse Document Frequency (IDF) represents scaling. When a term occurs frequently in documents, its importance is scaled down due to lowered discriminative power. The is defined as follows:

$$\text{IDE(a)}= log \frac{1+|x|}{x_a}$$

$x_a$is documents set having term .

C. Feature Selection Based on Decision Trees

A decision tree is a k-array tree in which each internal node specifies a test on some attributes from input feature set representing data. Decision trees are popular methods for inductive inference. And every test results in branches, representing varied test outcomes. They are robust to noisy data and learn disjunctive expressions.Each branch from a node corresponds to possible feature values specified at that node. The algorithm begins with tuples in the training set, selecting best attribute yielding maximum information for classification. It generates a test node for this and then a top down decision trees induction divides current tuples set according to current test attribute values. At every node, the algorithm selects best partition data attribute to individual classes. The best attribute to partitioning is selected by attribute selection with Information gain . Decision tree induction is the learning of decision tree classifiers constructing tree structure where each internal node (no leaf node) denotes attribute test. Each branch represents test outcome and each external node (leaf node) denotes class prediction.

D. Learning Vector Quantization (LVQ)

Learning Vector Quantization (LVQ) is a local classification algorithm, where classification boundaries are locally approximated, the difference being that instead of using all training dataset points, LVQ uses only a prototype vectors set. This ensures efficient classification as vectors number needing storing or comparing is reduced greatly. Additionally, a carefully chosen prototype set also increase noise problems in the classification accuracy

E. Component Analysis

When input dimensions are large and components highly correlated, dimensions are reduced using PCA. PCA orthogonalises variables and resulting principal components with large variation and eliminates components with least variation from datasets. Artificial variables calculated are principal components used as predictor, criterion variable in the analysis When applied on a dataset PCA observes the following steps.

1) Covariance matrix is calculated.
2) Highest eigenvalues are principal components of dataset. Remove eigenvalues of less significance to form feature vector.
3) Mean subtracted from each data dimensions producing a data set with zero mean.
4) Eigenvectors and eigenvalues of the covariance matrix are calculated.
5) A new dataset is derived.

IV. CONCLUSION

Feature selection is needed for successful data mining applications, as they lower data dimensionality removing irrelevant features. Rapid advances in computer based high-throughput technique provided unparalleled chances for humans to expand production, services, communications, and research productions. In this paper, a feature selection for Opinion mining using decision tree is proposed. LVQ type learning models constitute popular learning algorithms due to their simple learning rule. The classification accuracy obtained by LVQ was 75%. However it was observed that the precision for positive opinions was quite low. This phenomenon was observed not only on LVQ but other classifiers including CART and Naïve Bayes.

V. ACKNOWLEDGMENT

We would like to thank everyone who has motivated an supported us for preparing this manuscript.

REFERENCES

[1] Balahur, A., Steinberger, R., Goot, E. V. D., Pouliquen, B., & Kabadjov, M. (2009, September). Opinion mining on newspaper quotations. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on* (Vol. 3, pp. 523-526). IET.
[2] Omar, N., Jusoh, F., Ibrahim, R., & Othman, M. S. (2013). Review of Feature Selection for Solving Classification Problems. *JISRI*, *3*.
[3] Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., & Liu, H. (2010). Advancing feature selection research. *ASU Feature Selection Repository*.
[4] El-Halees, A., & Gaza, P. (2011). Mining Feature-opinion in Educational Data for Course Improvement. *International Journal of New Computer Architectures and their Applications (IJNCAA)*, *1*(4), 1076-1085.
[5] Metzler, D. (2008, October). Generalized inverse document frequency. In*Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 399-408)
[6] Yacob, Y. M., Sakim, H. M., & Isa, N. M. (2012). Decision tree-based feature ranking using manhattan hierarchical cluster criterion. *International Journal of Engineering and Physical Sciences*, *6*.
[7] The Internet Movie Database Ltd. Internet movie database. http://www.imdb.com.
[8] Ratanamahatana, C. A., & Gunopulos, D. (2002). Scaling up the naive bayesian classifier: Using decision trees for feature selection.
[9] Grbovic, M., & Vucetic, S. (2009, June). Learning vector quantization with adaptive prototype addition and removal. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on* (pp. 994-1001). IEEE.
[10] Jeevanandam Jotheeswaran et al., (2012), Feature Reduction using Principal Component Analysis for Opinion Mining. IJCST, Volume 3, Issue 5, May

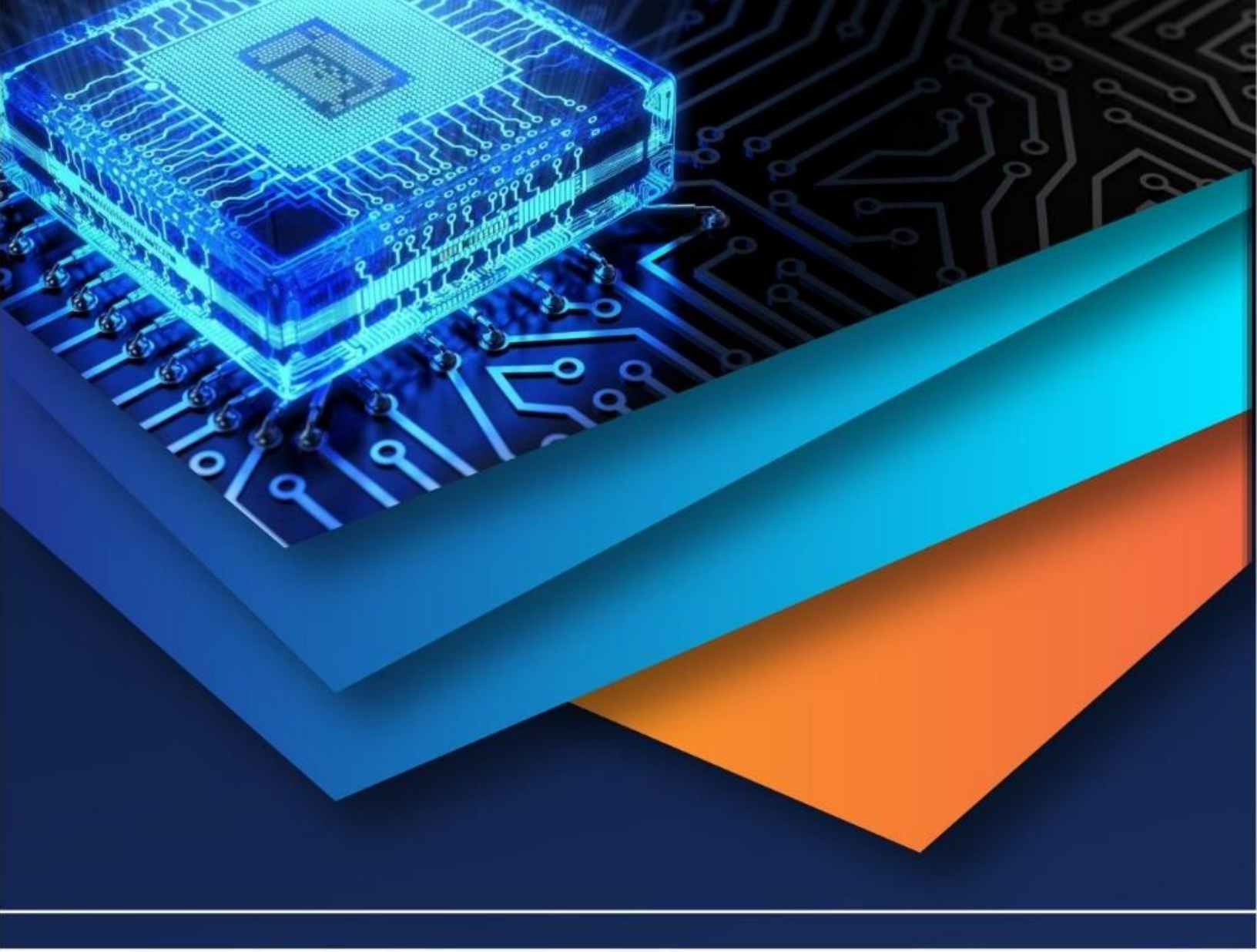
2012 (P 118 – 121).
[11] Gayatri, N., Nickolas, S., & Reddy, A. V. (2010). Feature selection using decision tree induction in class level metrics dataset for software defect predictions. In *Proceedings of the World Congress on Engineering and Computer Science* (Vol. 1, pp. 124-129).
[12] B. Liu. \Web Data Mining: Exploring hyperlinks, contents, and usage data," Opinion Mining.Springer, 2007.
[13] B. Pang & L. Lee, \Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales." Proceedings of the Association for Computational Linguistics (ACL), pp. 15124,2005.

AUTHOR

P KAVYA SRI[1], pursuing Post Graduate in Master of Technology with specialization in Computer Science and Engineering, JNTUH, Hyderabad, AP, India. My interested research area is Data Mining, Opinion Mining, Data Analysis.

K.C. RAVI KUMAR[2], Completed Post Graduate in Master of Technology with specialization in Computer Science and Engineering, JNTUH, Hyderabad, AP, India. His interested research area is Opinion Mining, Data Analysis, Data Security.

Dr. S.ANITHA REDDY[3] is the Professor of CSE Dept, in Sridevi Womens Engineering College,JNTUH, Hyderabad, AP, India. She received her M.Tech. from JNTU Hyderabad and Ph.D in the stream of Computer Science and Engineering in the specific area of Data mining. Her interested research topic is Data Mining, Opining Mining , Data Security and Analysis.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY