OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

ANALYSIS AND IDENTIFICATION OF COMPETITORS FROM UNSTRUCTURED DATASETS

Shruti .G. Regal[1], Prof. B. R. Solunke[2]

Student, Department of Computer and Engineering,N. B. Navale Sinhgad College of Engineering, Solapur, India Solapur University, Solapur-413255[1]

Professor of Department of Computer and Engineering, N. B. Navale Sinhgad College of Engineering, Solapur, India Solapur University, Solapur-413255[2]

Abstract*: The Internet is widely used for the propagation of propaganda as well as the sharing of thoughts and points of view. This research proposes the use of sentiment analysis techniques to decode web forum posts in a variety of languages. The article examines the usefulness of stylistic and syntactic features in classifying emotions in English and Arabic language. To account for Arabic linguistic features, unique feature extraction elements have been added. Additionally, the Entropy Weighted Genetic Algorithm (EWGA) is developed, which is a hybridized genetic algorithm that uses the information gain heuristic to select features. EWGA's objective is to improve continuity and to provide more accurate evaluation of critical functions. The suggested features and methods are tested using baseline movie review data from the United States and the Middle East, as well as web site postings from other countries. The experimental results obtained by integrating EWGA and SVM show that the device works well, with an accuracy of over 95% on the benchmark data collection and over 93% on both the US and Middle Eastern forums. Stylistic features significantly increased results in all test beds, and EWGA outperformed alternative feature selection approaches, demonstrating the efficacy of these features and techniques for document level sentiment classification.

.***Keywords***-*Competitors features, Hotel, E-Commerce, Seller Competitors*

----------∴ ∴ ∴----------

I INTRODUCTION

The competitive value of identifying and monitoring industry rivals is an inherent result of many market problems. Monitoring and distinguishing a firm's rivals has been researched previously. Data mining is the optimal method for mining rivals to manage such massive amounts of data. Online item ratings contain a variety of knowledge on consumers' opinions and desires, which can be used to obtain a general understanding of rivals. Numerous scholars previously published works in the literatures that intelligently and effectively processed such large amounts of consumer data. For instance, several research on online reviews have indicated that they would capture item opinion analysis from online reviews at various levels. Begin by developing a prioritized list of product specifications and advantages, accompanied by a table showing whether or not each of the rivals meets them. Indicate with a check mark which characteristics and advantages the rivals provide. Features are relatively straightforward; a product either has them or it does not. On the other hand, benefits are not as straightforward and can be reported solely based on customer reviews. Now, use this table to compare the product or service offered by your competitors. Consider the following: How does your product stack up against your nearest competitors? What qualities and advantages distinguish the product from others? How about theirs? The more distinct qualities and advantages your commodity possesses, the more secure your market share.

II. LITERATURE SURVEY

Z. Ma, G. Pant, and O. R. L. Sheng, “Mining competitor relationships from online news: A network-based approach,”[1]. As present Electronic Commerce Research and Applications, 2011. They suggested that business cites in

online news would be used to build an intercompany network whose structural properties could be used to infer competitor relationships between firms. They provoke three strong conclusions. First, the intercompany network is analyzed; second, the technical characteristics are analyzed; and third, the degree of output is quantified. These approaches take more time, work on only new data, and do not work with massive datasets.

R. Decker and M. Trusov, "Estimating aggregate consumer preferences from online product reviews,"[2]. As present *International Journal of Research in Marketing*, vol. 27, no. 4, pp. 293–307, 2010. They showed that customer reviews are accessible online for a wide range of product categories and provided an econometric method for translating the abundance of individual consumer opinions available by online product reviews into aggregate consumer preference results. Only ratings are used in this paper to locate rivals, which results in less precision.

C. W.-K. Leung, S. C.-F. Chan, F.-L. Chung, and G. Ngai, "A probabilistic rating inference framework for mining user preferences from reviews,"[3].As present *World Wide Web*, vol. 14, no. 2, pp. 187–215, 2011. . They suggest a novel probabilistic rating inference system, called PREF, for extracting consumer expectations from ratings and then translating them to numerical rating scales. PREF employs proven linguistic modeling methods to retrieve opinion terms and product characteristics from feedback, but it is time consuming and inefficient on broad datasets

E. Marrese-Taylor, J. D. Vel´asquez, F. Bravo-Marquez, and Y. Matsuo, "Identifying customer preferences about tourism products using an aspect-based opinion mining approach,"[4].As prasent *Procedia Computer Science*, vol. 22, pp. 182–191, 2013. They're based on Bing Liu's aspect-based opinion mining approach, which has now been applied to the tourism industry. This article's thesis necessitates review, which is time intensive due to the use of unstructured datasets.

III. PROBLEM DEFINATION

Web mining is a form of data mining that uses the World Wide Web (www) as the main data base. It can provide everything from web content to server logs and anything in between. In order to stay competitive, market research is necessary. We suggest a competition mining algorithm for defining and analyzing various available rivals based on the goods they produce, which can be helpful when introducing a new product or comparing a company product to other products in the market segment.

IV. PROPOSED SYSTEM

A formal definition of two items' competition based on their appeal to separate consumer groups in their market. An method overcomes previous work's dependency on scarce comparative data extracted from text. A systematic methodology for identifying the various categories of customers in a given market and estimating the number of customers that belong to each category. A highly efficient method for locating a given item's top-k competitors in very large datasets.

Module 1: Consumer Module: In this module, an administrator may insert information about things such as cameras, hotels, restaurants, and recipes. Following that, the administrator will review all submitted item information, customer requests, and desires. Finally, the top-k competitors from a given item are classified using CMiner++. The Customer-based features are developed in the second module. In this module, customers may submit queries for any object, such as cameras, hotels, restaurants, and recipes. Initially, the data set for cameras, hotels, restaurants, and recipes was developed. Gather consumer requirements from the customer list.

.**Module 2: Algorithm CMiner++ The module:** Following that, we present CMiner++, a precise algorithm for locating the top-k competitors of a given object. The skyline pyramid is used by our algorithm to minimize the number of elements that must be considered. Given that we are only interested in the top-k competitors, we will calculate the score of each candidate incrementally and stop when it is certain that the top-k has appeared.

*Module 3: The Skyline Operator module:****In this module is perform The skyline is a well-studied term that denotes the subset of points in a population that are not dominated by another point. Skyline refers to the skyline of a group of things i. (I). The skyline definition contributes to the following lemma:***

Module 4:Lemma1: Given the skyline Sky(I) of a collection of items I and an item I, let Y include the k Sky(I) items that are most competitive with i. Then, if j Y or if j is occupied by one of the items in Y, an object j I can only be among the top-k competitors of i. The list of elements I, the set of attributes F, the object of interest I the number k of top competitors to retrieve, the set Q of questions and their odds, and the skyline pyramid DI are all included in the input. The algorithm retrieves the objects that control I first, using masters I (line1). These products clash with I to the full degree possible. If there are at least k such objects, we record them and draw a conclusion (lines 2-4). Otherwise, we bind them

to Top and deduct k from our budget (line 5). The vector LB keeps the lowest lower bound from the present top set and is used to prune candidates (line 6). In line 7, we define the set of candidates X as the union of objects in the first layer of the pyramid and those already in the Top. This is accomplished by invoking GETSLAVES (TopK,). CMiner feeds the collection of candidates X to the UPDATE TOPK() routine, which prunes objects dependent on the LB threshold, in each iteration of lines 8-17. It then uses the MERGE() function to change the Top set, identifying the products with the highest competition from Top [X. Since X and Top are both sorted, this can be done in linear time. Line 13 sets the pruning threshold LB to the worst (lowest) score among the current Top. Finally, GETSLAVES() is used to extend the candidate range by adding objects that are occupied by those in X.

V SYSTEM ARCHITECTURE

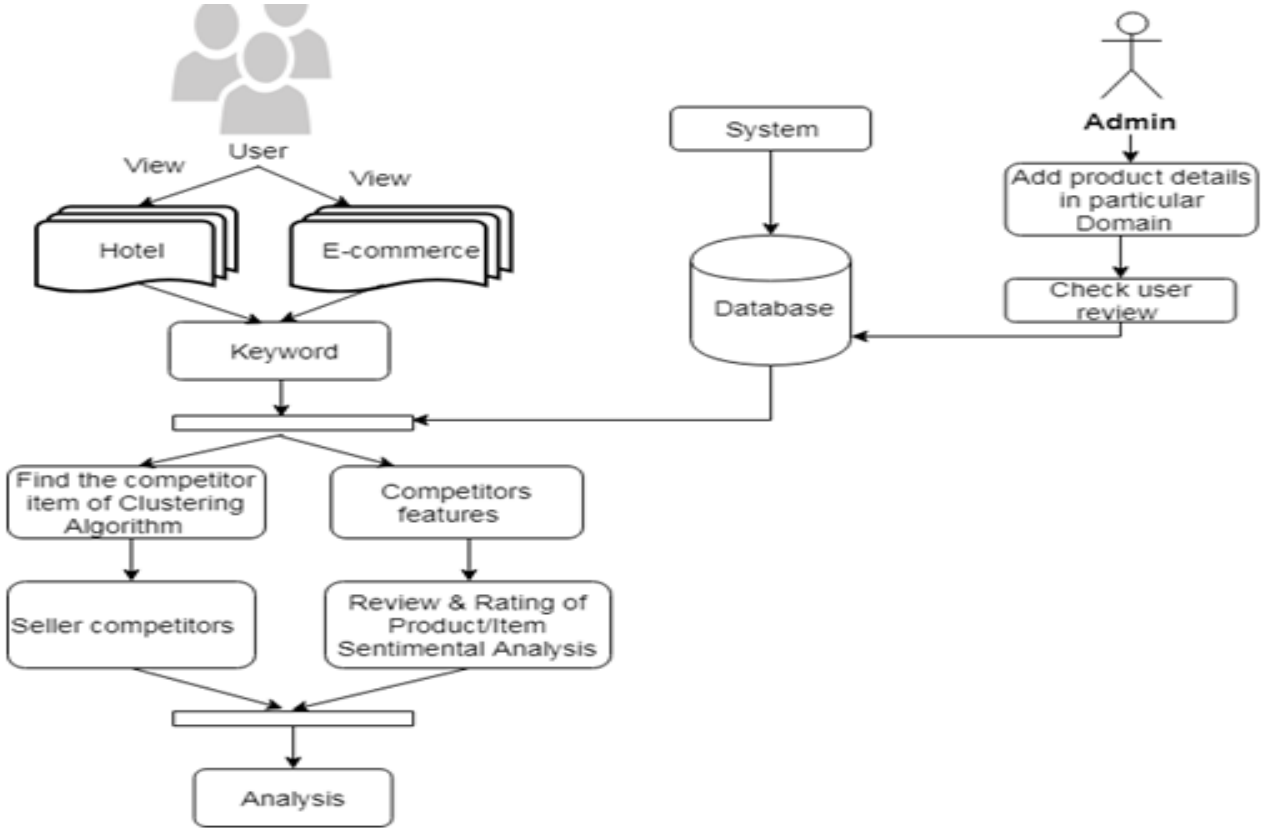


Fig: System Architecture

VI ALGORITHM INTERFERANCE

Clustering is an unsupervised problem in which natural groups are verified in the function space of input data. There are a variety of clustering algorithms to choose from, and no one algorithm is appropriate for all datasets. How to implement, fit, and use top clustering algorithms in Python using the scikit-learn machine learning library. It involves automatically classifying data using the regular classification system. Clustering algorithms, in contrast to supervised learning (e.g., mathematical modeling), see the input data simply and classify natural groups or clusters in feature space. A cluster is a dense area in the feature space where domain instances (observations or data rows) are more likely to be found nearby than in other clusters. In addition to a core (the centroid), the cluster can have a boundary or extent. The core (the centroid) may be a sample or a point function space.

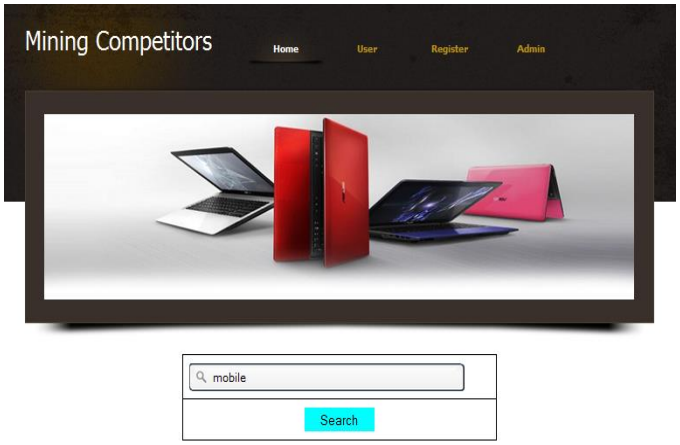
Centroid-based Clustering- Unlike the hierarchical clustering mentioned below, centroid-based clustering organizes data into non-hierarchical clusters. The most commonly used centroid-based clustering algorithm is k-means. The efficiency of centroid-based algorithms is limited by their sensitivity to initial conditions and outliers. Since k-means is an inexpensive, reliable, and simple clustering algorithm, it is the subject of this course.

Density-based grouping Clustering—Density-based clustering groups together regions with a lot of examples. As long as dense areas can be connected, arbitrary-shaped distributions are possible. This algorithms struggle with data with a broad variety of densities and dimensions. Furthermore, these algorithms are not designed to allocate outliers to clusters.

VII IMPLIMENTATION DETAILS

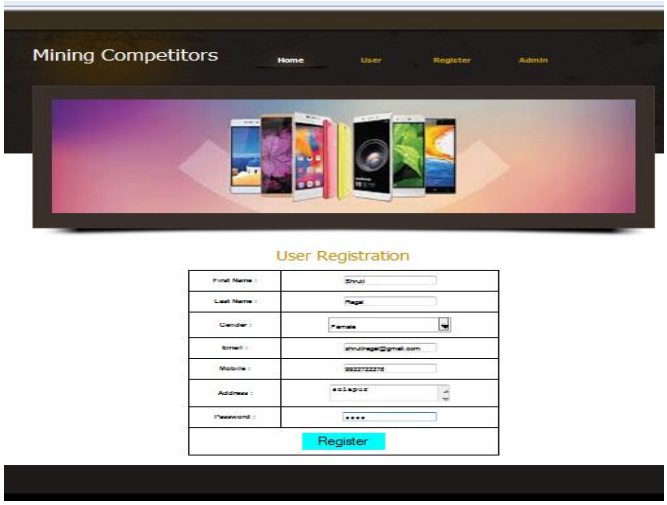
The method of finding rivals for products or materials is referred to as "mining rivals." When a new user logs in for the first time, he or she is taken to the home page. The user must first register with their email address and password before clicking the register button. Returning to their home page, the user must then press the login button, where they must enter their email address and password. Finally, depending on the administrator's provision of the home page, the administrator has the authority to determine whether or not the user is a registered user. The device is given two identities: Admin and User. Admins may input product information into the system, such as consistency and price. He or she will have access to all consumers and will be able to view and review all consumer feedback and ratings to decide which product is the perfect fit for the customer. The commodity is now inexpensive, and papers on it can be published.

Mining Competitors Home Page:



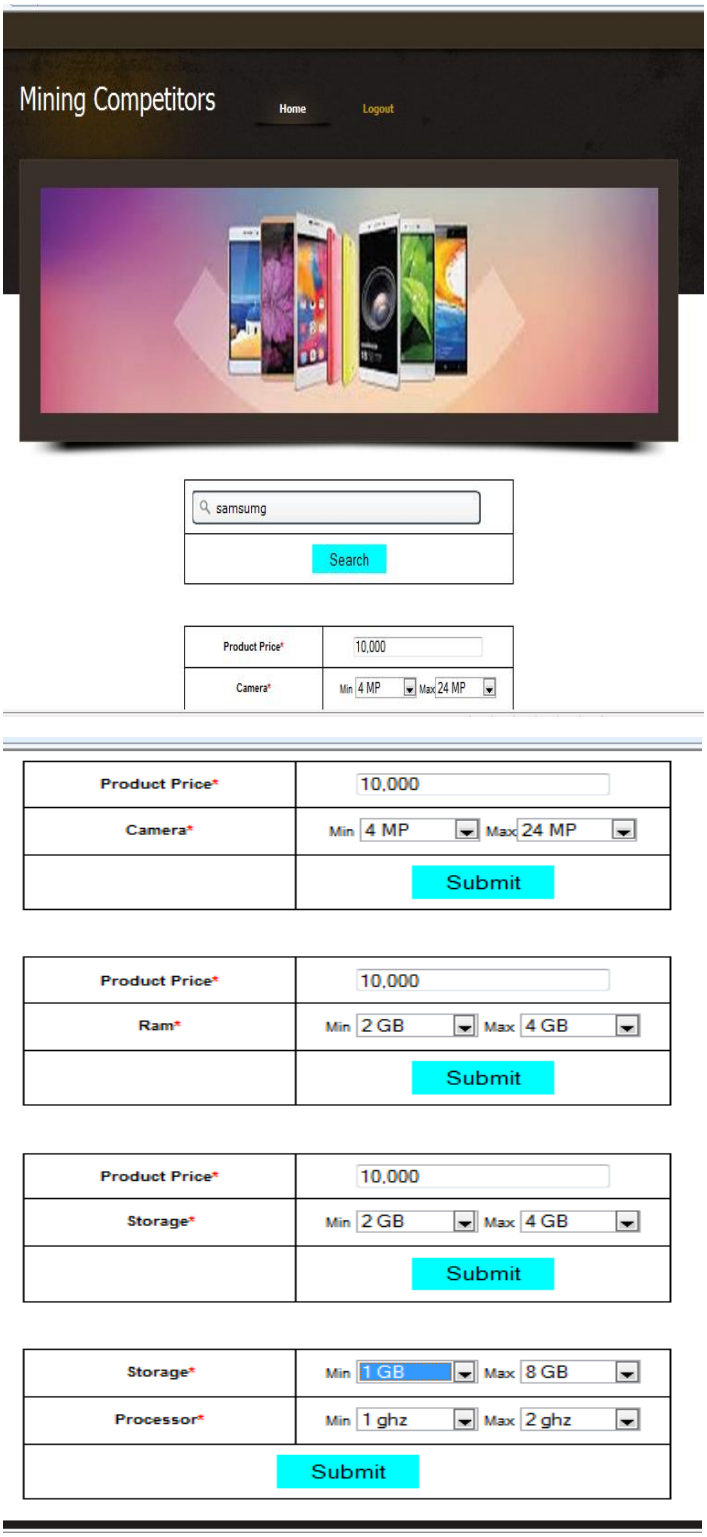
Welcome User:

In this the welcome User this message is displayed and the user is able to see the products available for e.g. Mobile, laptop.



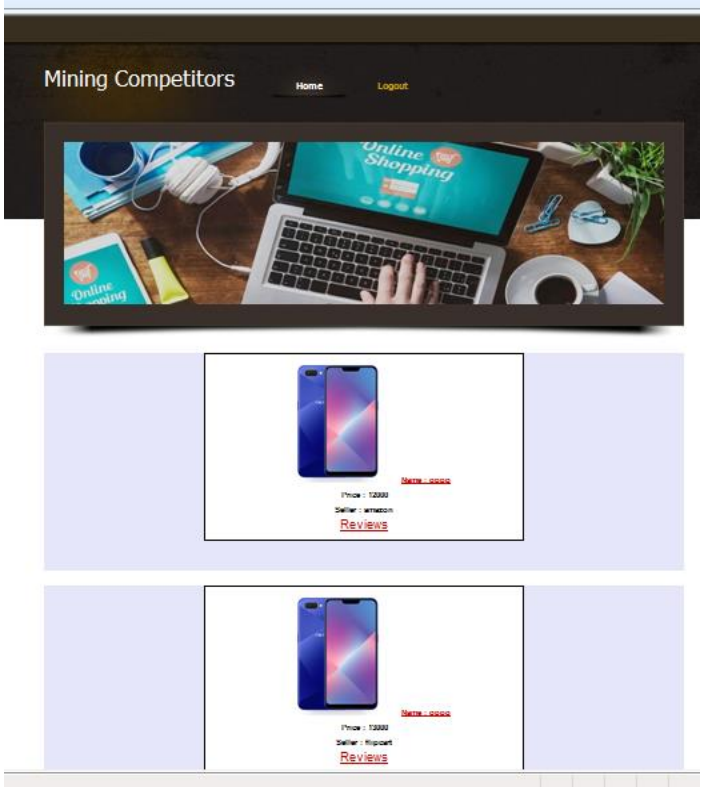
Search Page of a Product:

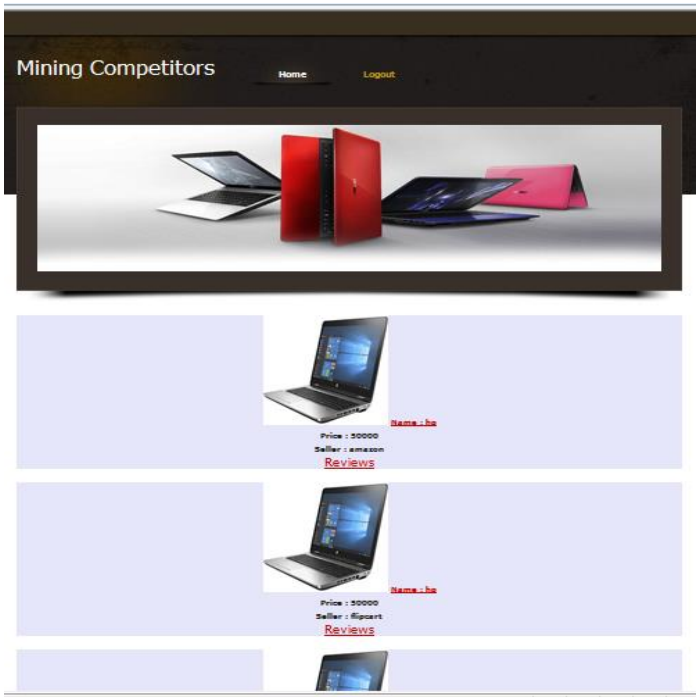
In this user is able to see all the available product or items of different models and company too. If user just click on the search button for e.g. mobile then the admin will display all the product with its quality, price ad reviews or user can search by particular product for e.g. redmi, oppo, redmi note, etc. with filling details below.



Product:

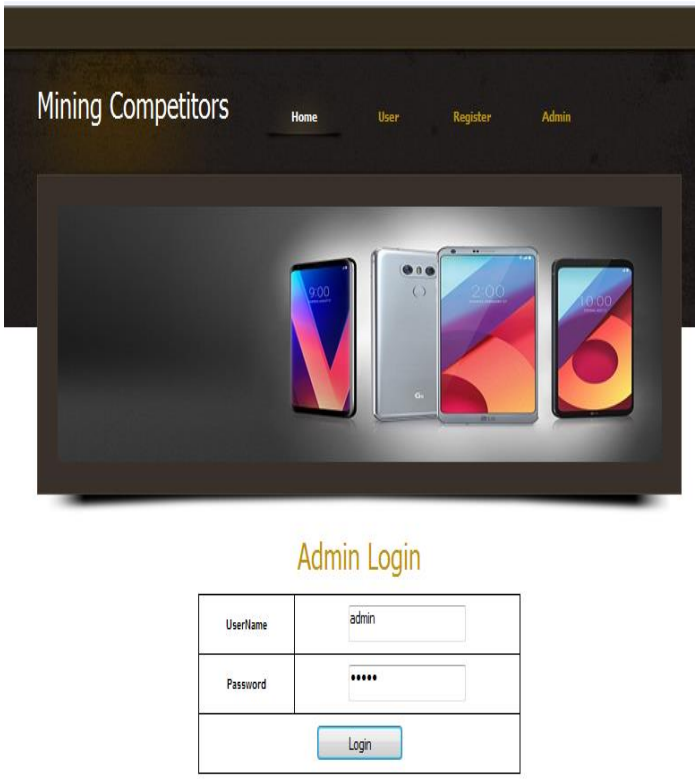
These are available product with its name, price, selling and review at the user side.





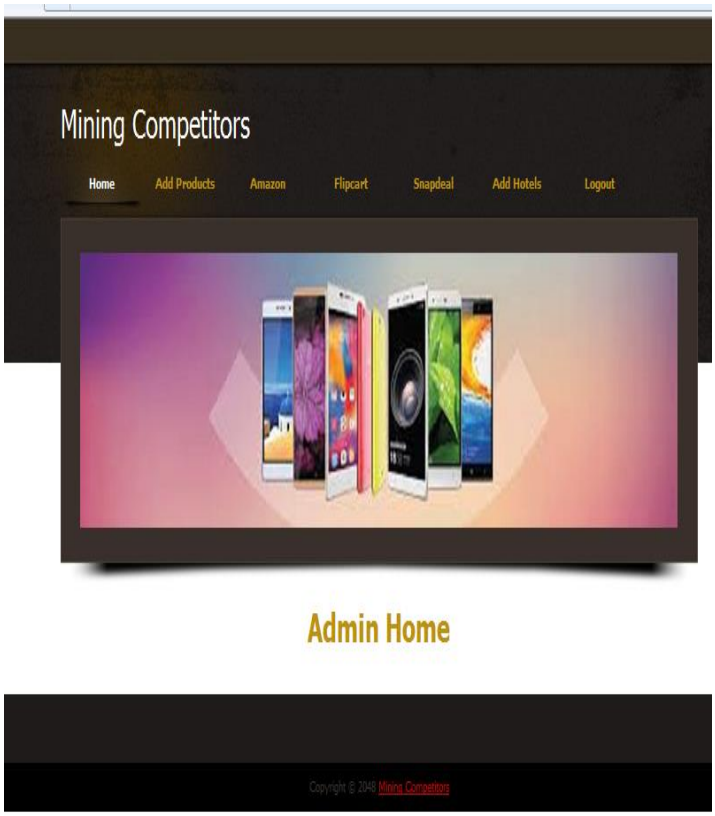
Admin login Page:

In this page admin has to login with user name as (admin) and password as (admin) and then click on login button.



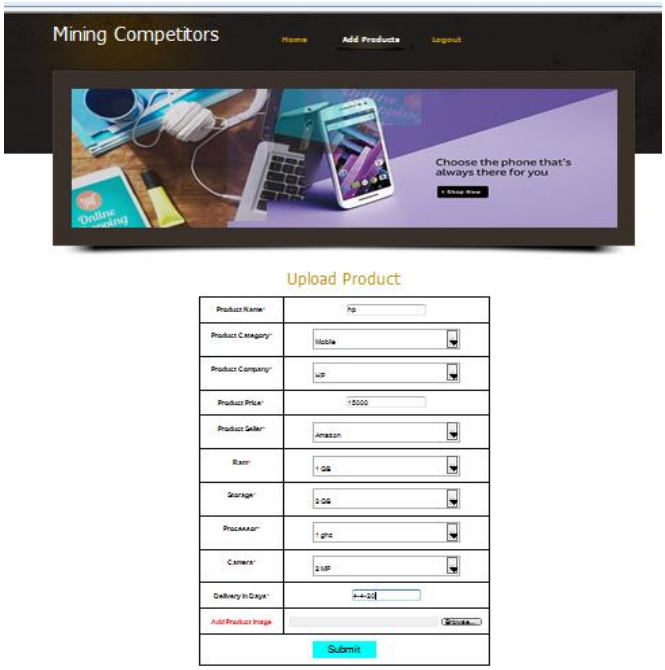
Admin home page:

In this page the admin can see all the product or items from different sites (amazon, flipcart and snapdeal) and admin can add the products too.



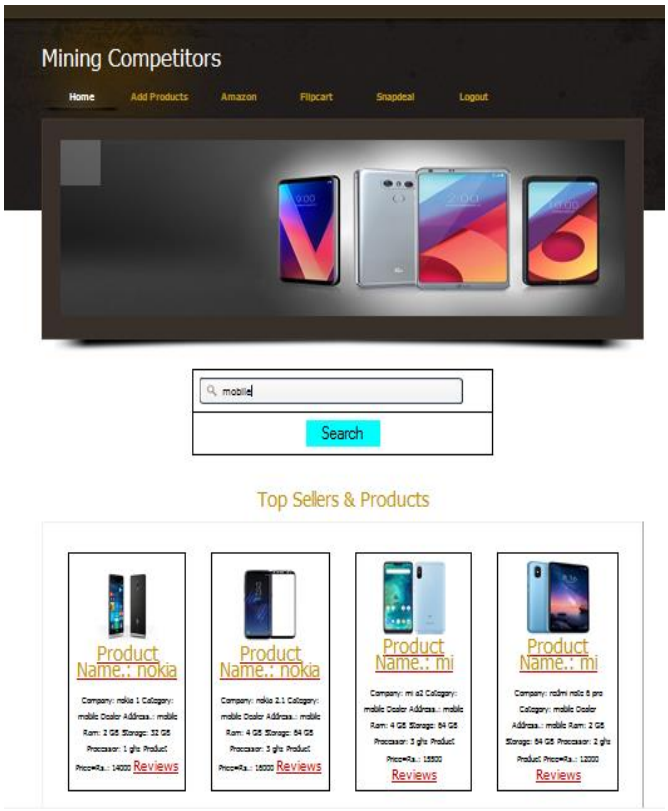
Add product :

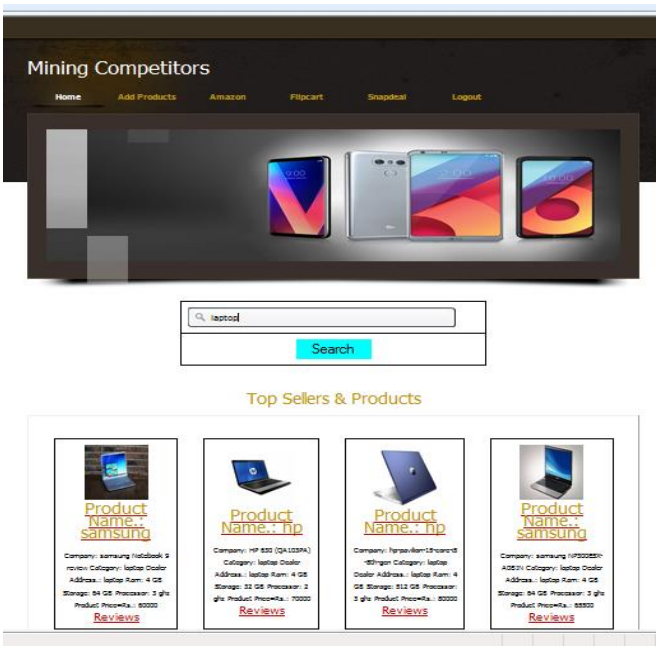
In this page admin add the products or items details which they wanted to upload their product .



Products:

These are the top sellers and products at the admin sides.





VIII.CONCLUSION AND FUTURE SCOPE

In various market domains, data mining is essential for discovering trends, forecasting, information discovery, and so on. Algorithms for machine learning are commonly used in a number of applications. Data mining techniques are used in any business application. Web mining techniques are needed to boost certain businesses or to have suitable competitors for the company to the customer. One method for analyzing competitors for the chosen products is competitor mining. In this article, we provided a detailed study of the competing mining algorithms, including their benefits and disadvantages. Finally, as compared to others, CMiner++ generated the shortest computation time. All baseline algorithms do not take into account the most important features and procedures. on the customer's business segment. This can be strengthened with further testing. The proposed method utilizes a competition mining algorithm to determine its consistency, and it distinguishes rival goods between the two objects depending.

REFERENCE

[1] Ding, X., Liu, B., Yu, P.S., 2008.A holistic lexicon-based approach to opinion mining. In: Proceedings of the WSDM'08.

[2] Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. 26 (3), 12:1–12:34

[3] Chen, L., Qi, L., Wang, F., 2012. Comparison of feature-level learning methods for mining online consumer reviews. Expert Syst. Appl. 39 (10), 9588–9601.

[4] Zhan, J., Loh, H.T., Liu, Y., 2009. Gather customer concerns from online product reviews – a text summarization approach. Expert Syst. Appl. 36 (2 Part 1), 2107–2115

[5] Jin, Jian, Ping Ji, and RuiGu. "Identifying comparative customer requirements from product online reviews for competitor analysis." Engineering Applications of Artificial Intelligence 49 (2016): 61-73.

[6] Saxena, Prateek, David Molnar, and Benjamin Livshits. "SCRIPTGARD: automatic context-sensitive sanitization for large-scale legacy web applications." Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011.

[7] R. Li, S. Bao, J. Wang, Y. Liu, and Y. Yu, "Web scale competitor discovery using mutual information," in ADMA, 2006.

[8] S. Bao, R.Li, Y.Yu, and Y.Cao, "Competitor mining with the web," IEEE Trans. Knowl. Data Eng., 2008

[9] G. Pant and O. R. L. Sheng, "Avoiding the blind spots: Competitor identification using web text and linkage structure," in ICIS, 2009.

[10] D. Zelenko and O. Semin, "Automatic competitor identification from public information sources," International Journal of Computational Intelligence and Applications, 2002