IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

A Comparative Study of Social Media Data Using Weka Tool

M. Saranya kala

Department of Computer Science, National College, Trichy, India

*Corresponding Author: saranyacs@nct.ac.in

Available online at: www.ijcseonline.org

***Abstract*—** Social media is a growing trend in the world today. It is being utilized by students, parents, businesses and religious organizations. Nowadays mostly every human being becomes addicted to social media, i.e. Facebook, Twitter and WhatsApp. Usages of social media are increasing in trends. They can build a personal network of friends that is connected to an open worldwide community. Information is now shared freely between the two. These parties can communicate either publicly or via the more discrete personal message. In this paper contains Facebook, Twitter and WhatsApp dataset like status and profile photo. The goal here is to analyze the time execution, Execution process and frequency by implementing weka tool. Here analogize the three algorithms, namely K-means, Bayesion algorithm and apriori algorithm. In this research process, the three algorithms used to find the time execution, Execution process and frequency which are predicting time consumes.

***Keywords*—** Facebook, Twitter, WhatsApp, Bayesion algorithm, K- Means algorithm, Apriori algorithm

I. INTRODUCTION

Social media sites such as twitter, Facebook provide great venues for students to share their experiences, vent emotion and stress, and seek social support. On various social media sites, students discuss and share their every day encounters in an informal and casual manner. The abundance of social media data provides opportunities to understand students' experiences, but also raises methodological difficulties in making sense of social media data for educational purposes. Just imagine the sheer data volumes, the diversity of internet slang, the unpredictability of location and timing of students posting on the web, as well as the complexities of students' experiences. Pure manual analysis cannot deal with the ever-growing scale of data, while pure automatic algorithms usually cannot capture the in-depth meaning within the data. Traditionally, educational researchers have been using methods such as surveys, interviews, focus groups, and classroom activities collect data related to students' learning experiences. These methods are usually very time consuming, thus cannot be duplicated or repeated with high frequency. The scale of such studies is also usually limited. In addition, when prompted about their experiences, students need to reflect on what they were thinking and doing sometime in the past, which may have become obscured over time. Casual posts and comments on social media (social network sites such as twitter) focuses into their educational experiences such as their problems, issues, suggestions about the students' study process. Some valuable information about student learning experiences can be inferred from the data gathered from such environments. It is very hard to analyze that information. Since the social media data keeps increasing in size it demands automation in data analysis. Then also the information inferred from those data needs human interpretation because it is the reflection of the students' crisis. A workflow that combines qualitative analysis and large-scale data mining techniques is generated. The project focused on students' posts to understand issues and problems in their educational experiences. Heavy workload, lack of awareness, social activities, and sleeplessness are some problems that students face as they go through the academic process.

II. RELATED WORK

Bogdan Batrinca, Philip C.&Treleaven [1] "Social media analytics: a survey of techniques, tools and platforms" Springer open access, AI &Soc (2015) 30:89–116. The easy availability of APIs provided by Twitter, Facebook and News services has led to an explosion of data services and software tools for scraping and sentiment analysis, and social media analytics platforms. This paper surveys some of the social media software tools, and for completeness introduced social media scraping, data cleaning and sentiment analysis. Perhaps, the biggest concern is that companies are increasingly restricting access to their data to monetize their content. It is important that researchers have access to computing environments and especially 'big' social media data for experimentation. Otherwise, computational social science could become the exclusive domain of major companies, government agencies and a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Arguably, the requirements are public-domain computational

environments and data facilities for quantitative social science, which can be accessed by researchers via a cloud-based facility.

G.ThirumaniAatthi, R.Aishwarya, R.Mallika, and Angel [2] "PREDICTION OF SOCIAL NETWORK SITES USING WEKA TOOL"International journal of advanced technology and science, volume 3, Isssue 1, Nov-2015. ISSN-2348.

People who are more interest to use the social Network for sharing or getting some information. The survey takes from students, workers, and other people. Overall the result from the survey is more number of students using social network sites. Because the student has use social network sites for information gathering, sharing and chatting purpose. The students are more interested to use the social networking sites than the working people and other people.

Andrzejewski, David, XiaojinZhu, Mark Craven, and Ben Recht [3]"Learning from Bullying Traces in Social Media" IJCAI, pages 1171–1177, 2011.

Social media as a large-scale, near real-time, dynamic data source for the study of bullying. Social media offers a broad range of bullying traces that include but go beyond cyberbullying. In the present paper, we have identified several key problems in using social media to study bullying and formulated them as familiar NLP tasks. Our baseline performance with standard off-the-shelf approaches shows that it is feasible to learn from bullying traces. Much work remains in this new research direction. In the short term, we need to develop specialized NLP tools for processing bullying traces in social media, similar to (Ritter et al., 2011; Liu et al., 2010), to achieve better performance than models trained on formal English. In the long term, we need to tackle the problem of piecing together the underlying bullying episodes from fragmental bullying traces.

III. METHODOLOGY

3.1 Data Mining:

Data mining is the process of discovering interesting knowledge, such as associations, patterns, changes, significant structures and anomalies, from large amounts of data stored in databases or data warehouses or other information repositories.

3.2 Social media:

Social media are computer-mediated technologies that allow to creating and sharing of information, ideas, career interests and other forms of expression via virtual communities and networks. The variety of stand-alone and built-in social media services currently available introduces the challenges of defining. However, there are some common features.

The term social media are usually used to describe social networking sites such as:

Facebook – an online social networking site that allows users to create their personal profiles, share photos and videos, and communicate with other users

Twitter – an internet service that allows users to post "tweets" for their followers to see updates in real-time.

WhatsApp – WhatsApp is a cross-platform instant messaging service for Smartphone. It uses the internet to make voice and video calls, text messages, images, GIF, videos, documents, user location, audio files, phone contacts and voice notes.

LinkedIn – a networking website for the business community that allows users to create professional profiles, post resumes, and communicate with other professionals and job-seekers.

Telegram - Telegram is a free cloud-based instant messaging service. Telegram clients exist for both mobile and desktop systems. Users can send messages and exchange photos, videos, stickers, audio, and files of any type. Telegram also provides optional end-to-end-encrypted messaging.

Snapchat – an app for mobile devices that allows users to send and share photos of them doing their daily activities.

3.3 WEKA Tool:

WEKA Tool is an open source data mining tool that provides data mining and machine learning procedures including data loading and transformation, data preprocessing and visualization, modelling, evaluation, and deployment.

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tools. The new machine learning schemes can also be developed with this package. WEKA is open source software issued under the GNU General Public License.

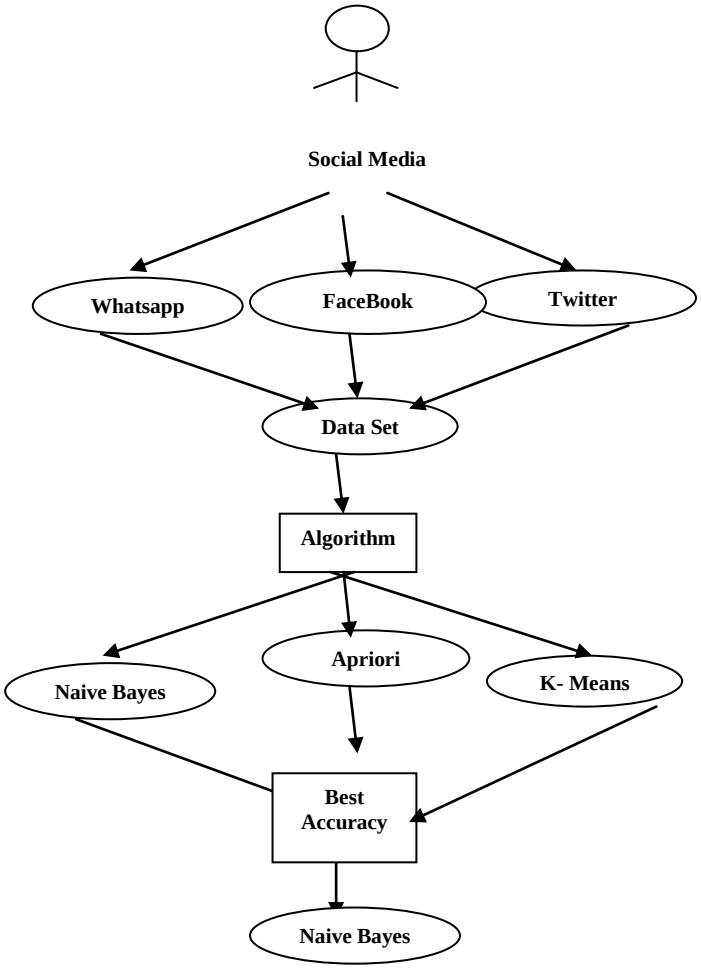
The goal of this Tutorial is to help you to learn WEKA Explorer. The tutorial will guide you step by step through the analysis of a simple problem using WEKA Explorer preprocessing, classification, clustering, association, attribute selection, and visualization tools. At the end of each problem there is a representation of the results with explanations side by side. Each part is concluded with the exercise for individual practice. By the time you reach the end of this tutorial, you will be able to analyze your data with WEKA Explorer using various learning schemes and interpret received results.

3.4 Naive Bayes Algorithm:

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.



3.5 Apriori algorithm

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

Apriori is designed to operate on databases containing transactions. Other algorithms are designed for finding association rules in data having no transactions, or having no timestamps. Each transaction is seen as a set of items. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time, and groups of candidates are tested against the data.

3.6 k-means clustering

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the k-means algorithm. The algorithm is often presented as assigning objects to the nearest cluster by distance. Using a different distance function other than (squared) Euclidean distance may stop the algorithm from converging.[citation needed] Various modifications of k-means such as spherical k-means and k-medoids have been proposed to allow using other distance measures.

If you are using *Word*, use either the Microsoft Equation Editor or the *MathType* add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation *or* MathType Equation). "Float over text" should *not* be selected.

IV. RESULT AND DISCUSSION

Weka Tool is a broad area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence, and database systems, for the analysis of large volumes of data. Social network analysis applications have experienced tremendous advances within the last few years due in part to increasing trends towards users interacting with each other on the internet. There have been a large number of Algorithms rooted in these fields to perform different data analysis tasks. In this paper, the comparison on the performance of Classification Algorithms was executed on the dataset. To start with the entire dataset is

categorized into 3 subsets. The entire attribute set includes 102 attributes which is very vast and hence feature reduction is performed to identify the highly relevant attribute for the target variable. The selected attributes were given as input to various Classification Algorithm and the error rates were analyzed and compared. From the results it is clear that in all the subsets considered for the research Apriori Algorithm produced less error rates when compared to all other Algorithms while executing in WEKA tool.

TABLE I: EXECUTION TIME IN SECONDS

Execution Time in seconds			
DATASET	Navie Bayes	K-Means	Apriori
Face book	0	0.02	0.08
Twitter	0	0.05	0.10
Whatsapp	0	0.2	0

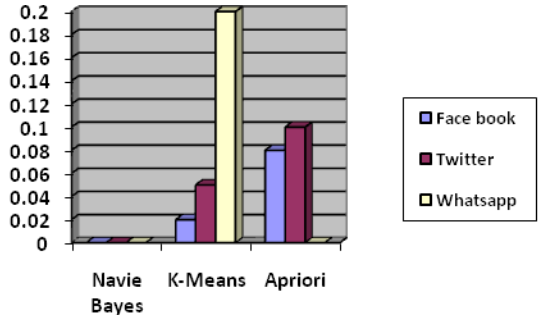


Figure 1. Execution time comparison.

The mean value theorem for integrals:

If $f(x)$is a continuous function on the closed interval l[a, b],then there exists a number in the closed interval such that

$$\int_a^b f(x)\,dx = f(c)\cdot(b-a)$$

The theorem basically just guarantees the existence of the mean value rectangle.

So, if you divide its area, $\int_a^b f(x)\,dx$, by its base, $(b-a)$, you get its height, $f(c)$.

Now Consider the Mean Value Calculation for Facebook Dataset with Naive Bayes Algorithmas an example from above table

Assigning a=0.958, b=1

Determine the calculation for Mean with above data set

1-0.958=0.042(It shows the Result for Facebook Results using Naive Bayes Algorithm)

TABLE 2: MEAN VALUE

Mean Value			
DATASET	Navie Bayes	K-Means	Apriori
Face book	0.042	0.475	0
Twitter	0.043	0.476	0
Whatsapp	0.478	0.07	0

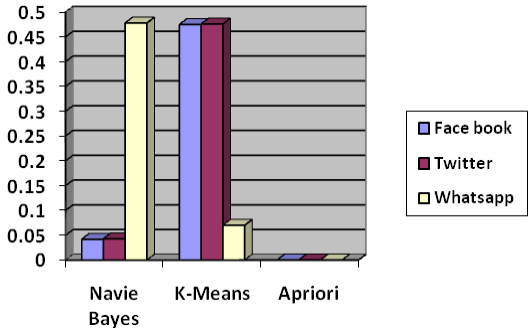


Figure. 2. Mean value comparison.

RMSE: Root Mean Square Error (RMSE) is the standard deviation of the residuals (predictionerrors). Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

TABLE 3: ROOT MEAN VALUE

Root Mean Value			
DATASET	Navie Bayes	K-Means	Apriori
Face book	0.458	0.460	0
Twitter	0.431	0.476	0
Whatsapp	0.480	0.480	0

NowConsidertheRootMeanValueCalculationforFacebookDatasetwithNaiveBayesAlgorithm as an example from above table The formula is:

$$RMSE = \sqrt{(f-o)^2}$$

Determine the calculation for Root Mean with above dataset

RMSE =1-0.542 =0.458(It shows the Result for Facebook Results using Naive BayesAlgorithm).

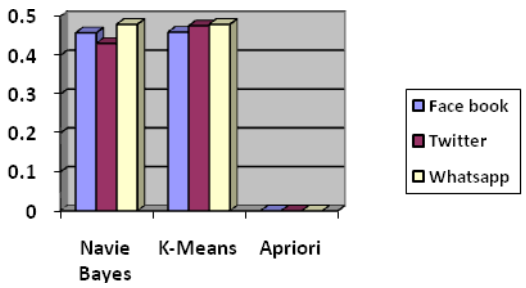


Figure. 3. Root Mean value

In this paper provides a workflow for analyzing social media data for analysis purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user generated textual content. In this paper social media data we are implement in weka tool to compare time execution, frequency and an execution process by using three algorithms namely K-means, Bayesion algorithm and aprior based on this three algorithm we find out which one is best. This paper gives the comprehensive and theoretical analysis of three algorithms. The study describes the technical specification, features, and area for each selected tool along with its applications. By employing the study of choice and selection of tools Naive Bayes is the best in execution.

In future, Fuzzy tech has to be implemented to provide the most comprehensive support for your target platform and application area. Due to differences in the capabilities of the supported hardware platforms, technical limits apply to the size of the fuzzy systems. Analyzing the collective posts can also add as a significant part.

REFERENCES

[1]. Bogdan Batrinca, Philip C.&Treleaven "Social media analytics: a survey of techniques, tools and platforms" Springer open access, AI &Soc (2015) 30:89–116.

[2]. G.ThirumaniAatthi, R.Aishwarya, R.Mallika, and Angel "PREDICTION OF SOCIAL NETWORK SITES USING WEKA TOOL"International journal of advanced technology and science, volume 3, Isssue 1, Nov-2015. ISSN-2348.

[3]. Andrzejewski, David, XiaojinZhu, Mark Craven, and Ben Recht [3]"Learning from Bullying Traces in Social Media" IJCAI, pages 1171–1177, 2011.

[4]. Boyd, D., & Ellison, N., (2007), 'Social network sites: Definition, history, and scholarship', Journal of Computer-Mediated Communication, 13(1),.Retrieved August 22, 2009 from http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.htm , pp.1-11

[5]. Christopher,C.(2008) Executive Briefing: Social network for business associations. Retrieved August 20, 2009 from http://haystack.cerado.com/html/haystack_directory.php. PP.18.

[6]. Dan, M. (2009). Social networking for dentists – made easy. Retrieved August 22, 2009 from http://www.dental-tribune.com/articles/content/id/315/scope/news/region/usa.

[7]. DiMicco. J., Millen. D., Geyer. W., Dugan, C., Brownholtz, B., and Muller, M. (2008) Motivations for Social Networking At Work. In Proceedings of The 2008 ACM Conference on Computer Supported Cooperative Work, San Diego, CA, USA, November 08 - 12, 2008, pp.711-720

[8]. David,R.(2007) YouTube for Your Business; Computerworld. Retrieved August 20, 2009 from http://www.pcworld.com/article/133278/youtube_for_your_business.html.

[9]. Emin , D. & Cüneyt , B. (2007) Web 2.0 - an Editor's Perspective: New Media for Knowledge Cocreation. International Conference on Web Based Communities (2007),pp 27-34. [7]. Facebook Adds Marketplace of Classified Ads (2007-05-12). Retrieved August 24, 2009 from www.physorg.com/news98196557.html .

