NEWSD: A Realtime News Classification Engine for Web Streaming Data

Urooj Mohiuddin*, Hameeza Ahmed‡, Muhammad Ali Ismail†

Department of Computer & Information Systems Engineering,
NED University of Engineering & Technology University Road, Karachi-75270, Pakistan
Email: *urooj_cs50@yahoo.com, ‡hameeza@neduet.edu.pk, †maismail@neduet.edu.pk.

Abstract—News Explorer for Web Streaming Data (NEWSD) is a GUI based text mining tool developed for the classification of streaming web data. It provides a platform to perform text mining on news updates extracted from various selected online newspapers. Initially, the text based news data is fetched from social networking pages of the selected newspapers. The real time data gathering is immediately followed by the pre processing, extraction and classification. Classifiers namely NaïveBayes and J48 are employed to categorize the news updates according to their nature and semantics. The tool will lead towards a more aware society by constantly providing the relevant updates about the events.

Keywords—news explorer; classification; streaming web; text mining; real time tool

I. INTRODUCTION

Social web is one of the biggest revolutions of the present century. It has successfully achieved the task of establishing social relations amongst billions of its users across the globe [1]. It has been serving as the major contributor of growing amount of data on the World Wide Web due to the increased ratio of social networking. There exist a large number of social networking web sites namely Facebook, Twitter, LinkedIn, Google+, MySpace, Instagram, and many others. It is a nontrivial task to mine the real time data acquired from the social web as the data is streaming.

Text mining is an emerging area that attempts to extract meaningful information from natural language text. The mining process of unstructured textual data is done through the exploration and identification of interesting patterns. Text mining is strongly connected to machine learning, natural language processing, data mining, knowledge management and information retrieval [3], [4], [5], [6].

Recently, social networking sites are occupied with excessive real time data which is unstructured or semi-structured, heterogeneous, and mostly text based. Text mining has gained substantial importance as text data serves as the main data source on the web. It is a challenging task to acquire such continuous stream of data in order to extract knowledge from it, known as real time text analysis. This paper presents a tool developed, NEWSD, News Explorer for Web Streaming Data. It is used for the mining of real time news data from different online newspapers. The whole NEWSD is based on five phases namely data acquisition, preprocessing, feature extraction, classification and visualization. At first, real time data is acquired from selected online newspapers. The data gathering process is followed by its preprocessing and feature extraction. The extracted features serve as an input to the classification stage which is performed using Naïve Bayes and J48 classifiers. Classification is possibly the most popular predictive data mining technique and a discrete supervised machine learning method [8]. Currently, categorization is based on five different news activities including sports, politics, national, crime, and international. Finally, the classified news data is displayed through a user friendly GUI. Beside a tool, NEWSD is a classification engine that can be used to benefit a large portion of the society. Instead of visiting the online newspapers individually, user can access the related news from various newspapers by using NEWSD.

Rest of the paper is as follows: Section 2 discusses related work. A brief introduction of social web and text mining is provided in section 3. The complete methodology of NEWSD is discussed in section 4. Section 5 describes how to use the NEWSD, followed by section 6 highlighting the conclusions and future work.

II. RELATED WORK

In current internet world, where all types of data are being pumped in it, text analytics or intelligent text mining has started to get much attention. This scenario is being emerged with demand of new tools and technologies in domain of real time analytics. The majority of applications of real-time text analytics are addressing streaming data which is continuously generated on social web. Many companies are using text mining tools to get reviews about their brands, by government agencies in order to acquire the predictions of terrorist attacks, medical epidemics and other criminal activities. Also, the analytics is used by the companies in order to track blog posts and news feeds for financial reasons respectively [11]. Although the work published in [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, and 34] has performed the classification of news data, but none of them has come up as a complete, flexible and portable tool. The work presented in above has been done on part basis not as a whole. Some contributions worked on feature extraction method, while the other discussed the classification technique.
involved in text mining. They were unable to provide a single tool which would be capable enough to follow the complete cycle for the classification of web streaming news data. The work being presented in this paper is innovative in the sense that it deals with the development of a complete independent tool named as NEWSD. This flexible, portable and user-friendly tool is able to perform the mining of web streaming data collected from online newspapers. The complete standard text mining cycle is followed to achieve the desired objective and make the tool a standard. Also, a GUI is designed to facilitate the visualization of the tool.

III. SOCIAL WEB AND TEXT MINING

A. Social Web and Web Streaming Data

Social web is mainly concerned with creating links amongst its users and World Wide Web by a set of social relations. These online interactions serve as the foundation of the online activities namely education, gaming, online shopping, and social networking websites [1]. With the advent of social web, World Wide Web has become a source of overloaded information due to the increased ratio of social networking. Social networking sites can be defined as web-based services that allow groups of individuals to share mutual experiences [2]. There exist a large number of social networking web sites including Facebook, Twitter, LinkedIn, Google+, MySpace, Instagram, YouTube, Pinterest, Flickr and many others. These sites encourage fast update of any information and sharing of knowledge on a vast scale. Data stream derived from an ordered set of instances made available over time on the web it is known as Web Streaming data. The streaming process is indicated by the continuous transmission of information in real time at a steady speed. There are number of sources generating web streaming data including webpages, sensors, media, weather stations, and many others. But, among all those social networking sites have the major share.

B. Text Mining

Text mining has become a promising and challenging research area due to the growing use of social web. It is strongly connected with data mining, natural language processing, machine learning, knowledge management and information retrieval. Text mining can be defined as the process of discovering useful information from semi-structured or unstructured text [5]. The information retrieval is process of discovering useful information from semi-structured data. These news activities include sports, political, entertainment, national and international affairs. Finally, the classification results are depicted through a user friendly GUI. The complete NEWSD has been developed in Java using number of APIs namely Restfb, Facebook Graph and WEKA. The basic components of NEWSD are illustrated by Fig. 1.

1) Preprocessing and Feature Extraction:

The term preprocessing and feature extraction are closely related to each other in text mining. These steps play a crucial part to determine the quality of classification phase. The selection of significant keywords and discarding the less important words is done at this stage. The preprocessing technique greatly facilitate the mining process by transforming the raw unstructured data into structured format. The text preprocessing is also known as text normalization or tokenization. It is mainly comprised of 5 transformations namely lexical analysis of the text, stemming, elimination of stopwords, index terms selection, and thesauri. The identification of relations and facts in text remain the main goal of feature extraction. It basically deals with the transformation of text data into numerical features. The specific counting tokenization, and normalization technique is also known as Bag of Words approach [7], [12], [35], [36].

2) Classification:

It is used to categorize an unknown observation into a set of categories, by getting trained from a set of training data comprising of ample observations with known category membership. It is supervised machine learning technique. A number of text based classifier techniques exist; two of them used in the tool, are discussed below.

Naive Bayes: A Naive Bayes classifier is a fast, simple, and easy to implement probabilistic classifier, which is based on Bayes’ theorem. Naive Bayes classifiers can be efficiently trained, in a supervised learning scenario. They perform quite well in many complex real world situations, in spite of their simplified design. They require small amount of training data for parameters estimation [8].

J48: It is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database which results in classification for that tuple. J48 uses divide-and-conquer algorithm to split a root node into a subset of two partitions till leaf node (target node) occur in tree [9], [10].

IV. METHODOLOGY

NEWSD is a text mining tool developed specifically for the classification of news data. Initially, the real time news data is acquired by multiple Facebook pages. The data gathering process is followed by its preprocessing and feature extraction. The separation of appropriate words from the whole text is known as extraction process. The obtained word vector will serve as input for the classification stage which is responsible to categorize the news activities as per the text data. These news activities include sports, political, entertainment, national and international affairs. The classification results are depicted through a user friendly GUI. The complete NEWSD has been developed in Java using number of APIs namely Restfb, Facebook Graph and WEKA. The basic components of NEWSD are illustrated by Fig. 1. The detailed description of NEWSD is as follows:

A. Data Acquisition

The task of real time data gathering is done by using a Java API known as Restfb. It is a light weight Java implementation of Facebook Graph API [12]. Initially, the connection is established between NEWSD and Facebook by the access token. The Facebook connectivity will enable NEWSD to retrieve the real time news updates via Facebook pages of 12.
The GUI of NEWSD is designed using Java Swing Toolkit. ENTIRE NEWS CLASSIFICATION PROCESS IN AN APPROPRIATE MANNER. SIGNIFICANT IN NEWSD. IT ENABLES THE END USERS TO VISUALIZE THE D. CLASSIFIERS NAMELY NAIVEBAYES AND J48 ARE USED BY THE TRAINING IS HANDLED BY NEWSD. PRESENTLY, ONLY TWO CLASSIFICATION AND LABEL ASSIGNMENT OF THE TEST DATA BASED ON THE LABEL EACH Fetched INSTANCE MANUALLY. Thus, THIS ENTIRE TASK OF DATA HAS NO PRE-ASSIGNED LABELS AS IT WOULD BE INAPPROPRIATE TO DATA WHICH IS REAL TIME AND KEEPS ON CHANGING. BUT, THE TEST TRAINING FILE. THE NEWS CATEGORIZATION TASK IS PERFORMED ON TEST APPROPRIATELY. THE LABELS ARE MANUALLY ASSIGNED TO THE SAMPLES THAT ARE LARGE ENOUGH TO PERFORM THE LEARNING PROCESS NEWSD, THE TRAINING DATA IS FIXED COMPRISING OF NEWS COLLECTED DATA IS REAL TIME INVOLVING CONTINUOUS VARIATIONS. THE SIGNIFICANCE OF THE DATA SOURCE IS HIGHER AS THE DIFFERENT NEWS SOURCES AND DISPLAYING THE COMBINED RESULTS FACILITATE THE END USERS WHO HAVE BECOME THE REGULAR NEWSD INITIALLY COLLECTS THE NEWS DATA FROM 12 SELECTED NEWSPAPER USING THEIR FACEBOOK PAGES. THEN CLASSIFICATION IS PERFORMED ACCORDING TO THE CHOSEN NEWS CATEGORY. ALSO, FEW RELATED DETAILS NAMELY NEWSPAPER NAME AND TIME OF CREATION ARE SHOWN ALONG WITH THE NEWS FEED. IT IS CLEAR THROUGH THE SNAPSHOTS HOW USER IS ABLE TO VIEW THE NEWS OF HIS CHOICE BY USING NEWSD. THE NEWS OF ONE’S INTEREST IS DISPLAYED IN 2 SNAPSHTOS HOW USER IS ABLE TO VIEW THE NEWS OF HIS CHOICE BY USING NEWSD.

V. USING NEWSD

NEWSD IS A REAL TIME TOOL ENCOMPASSING ALL THE MAJOR FUNCTIONALITIES THAT ARE NECESSARY FOR A TEXT MINER. IT INCLUDES ACQUISITION OF FACEBOOK STREAMING DATA, ITS PREPROCESSING & FEATURE EXTRACTION, CLASSIFICATION AND VISUALIZATION RESPECTIVELY. THE MAIN INTERFACE OF NEWSD IS DISPLAYED IN FIG. 2. IT CAN BE CLEARLY SEEN HOW THE USER IS PROVIDED WITH VARIETY OF OPTIONS NAMELY LINKUP FACEBOOK, NEWS RETRIEVAL, ACTIVITY SELECTION, AND TEXT CLASSIFICATION. IN ORDER TO ALLOW PERIODIC NEWS FEED UPDATES A REFRESH OPTION IS PROVIDED. ALSO, THE TOOL IS USER FRIENDLY WITH ATTRACTIVE AND GRACEFUL GUI. THE USER IS REQUIRED TO ACCESS THE FUNCTIONALITIES IN A SEQUENTIAL MANNER. INITIALLY, AN ACCESS TOKEN IS ENTERED BY THE USER AS SHOWN IN FIG. 3. IN ORDER TO ESTABLISH THE LINK BETWEEN NEWSD AND FACEBOOK. THE INCORRECT TOKEN WOULD RESULT IN AN ERROR WHILE THE CORRECT ONE WOULD INDICATE SUCCESS AS SHOWN IN FIG. 4. AND 5 RESPECTIVELY. AS, THE USER GETS CONNECTED TO FACEBOOK, THE COLLECTION OF DATA IS STARTED FOR THE SPECIFIED NEWSPAPER PAGES. THE COMPLETION OF ACQUISITION PROCESS IS DEPICTED IN FIG. 6. THE COLLECTION PROCESS IS FOLLOWED BY CLASSIFICATION. THE CLASSIFICATION OPTION IS RESPONSIBLE TO PERFORM ALL THE REQUIRED PHASES NAMELY PREPROCESSING, FEATURE EXTRACTION AND CLASSIFICATION RESPECTIVELY. THE NEWS CATEGORY CAN BE SPECIFICALLY Chosen FROM A LIST OF ACTIVITIES NAMELY SPORTS, POLITICS, CRIME, NATIONAL, INTERNATIONAL, GENERAL, AND ENTERTAINMENT RESPECTIVELY. ALSO, THE CLASSIFIER CAN BE SELECTED FROM THE LIST. FINALLY, THE CLASSIFIED NEWS DATA AS PER SELECTED CATEGORY IS SHOWN IN FIG. 7. THE CHOSEN ACTIVITY IS INTERNATIONAL WHICH MEANS THE MINED INTERNATIONAL NEWS IS DISPLAYED WITH SOME POSSIBLE CLASSIFICATION ERROR AS WELL. THE NEWSD INITIALLY COLLECTS THE NEWS DATA FROM 12 SELECTED NEWSPAPER USING THEIR FACEBOOK PAGES. THEN CLASSIFICATION IS PERFORMED ACCORDING TO THE CHOSEN NEWS CATEGORY. ALSO, FEW RELATED DETAILS NAMELY NEWSPAPER NAME AND TIME OF CREATION ARE SHOWN ALONG WITH THE NEWS FEED. IT IS CLEAR THROUGH THE SNAPSHTOS HOW USER IS ABLE TO VIEW THE NEWS OF HIS CHOICE BY USING NEWSD. THE NEWS OF ONE’S INTEREST IS DISPLAYED IN 2 STEPS WHICH ARE LINKING AND CLASSIFICATION. ALSO, USER IS ABLE TO SEE THE RECENT NEWS UPDATES BY REFRESH OPTION IN ORDER TO AVOID THE REESTABLISHMENT OF CONNECTION.

VI. IMPACT OF NEWSD

end user is able to get the updates about the latest events occurred in the society by just installing NEWSD. The two properties namely light weight and less computation & resource intensiveness make NEWSD a promising and easy to use engine. Also, the latest updates gathered from multiple sources would greatly increase the social awareness amongst the masses. The classification system being presented is able to provide the breaking news about bomb blasts, strikes, political riots, cricket, football, and many others. This news broadcasting is of significant importance due to the fact that the updates are collected from different real time sources and the NEWSD provides one single platform in order to perform the mining of collected data. The gathering, mining and visualization of news data are the biggest achievements of NEWSD. Also, the tool has strong appeal not only for industry and academia but for the common masses as well. The ability to provide benefits to the society increases the worth of NEWSD to a greater extent.

The tools has also been evaluated for its performance. In order to evaluate the performance of individual classifiers, there exists primarily two techniques namely hold-out and cross validation. In case of NEWSD, initially cross validation is performed in training dataset using NaiveBayes and J48 classifiers. In case of classifier NaiveBayes, accuracy is 54.821%; while the accuracy is reduced to 52.6003% for classifier J48 as shown in Table 1. It can be seen how certain activities including crime, international, general, and politics which are taking place more frequently possess better accuracy as compared to national, sports, and entertainment activities. Also, the number of training samples for certain scenarios is kept higher. The average classification accuracy in case of classifier NaiveBayes is roughly estimated to be 61.2857% and for J48 the accuracy is around 52.1428%. Hence, the evaluation results clearly depict how Naive Bayes is more accurate as compared to J48 for both the scenarios namely cross validation and hold out respectively.

**TABLE I. CLASSIFICATION RESULTS OF NEWSD**

<table>
<thead>
<tr>
<th>Activities</th>
<th>Naive Bayes Classification Accuracy (%)</th>
<th>J48 Classification Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crime</td>
<td>80</td>
<td>72</td>
</tr>
<tr>
<td>International</td>
<td>75</td>
<td>68</td>
</tr>
<tr>
<td>National</td>
<td>50</td>
<td>44</td>
</tr>
<tr>
<td>Sports</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>Entertainment</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>Politics</td>
<td>86</td>
<td>75</td>
</tr>
<tr>
<td>General</td>
<td>78</td>
<td>70</td>
</tr>
<tr>
<td>Total (Test Samples)</td>
<td>61.2857</td>
<td>52.1428</td>
</tr>
<tr>
<td>Total (Cross Validation in Training Samples)</td>
<td>54.8291</td>
<td>52.6003</td>
</tr>
</tbody>
</table>
VII. CONCLUSIONS AND FUTURE WORK

Streaming data is tightly bounded with social networking in particular. The rapid increase in social networking has drawn the attention of both the users and developers in the designing of related tools and technologies dealing with web streaming data. One such promising tool known as NEWSD (News Explorer for Web Streaming Data) is presented in this paper. The idea behind NEWSD is innovative in the sense that it makes the complex process of news classification for multiple activities possible. In the tool, one can classify the news in various categories including crime, sports, national, politics, international, entertainment, and general activities respectively. The final classification results are able to group all the same activities in one set. In order to minimize the possible errors, option of selecting one of two well know classifiers namely NaiveBayes and J48 is also given. NEWSD has got all the major capabilities to provide greater ease to the end user with respect to news searching and classification. It reduces human effort and time by automatically displaying the updated news of interest.

Presently, NEWSD is incorporating news data of multiple online newspapers. The existing project can be extended to increase the number of news activities and classification techniques. Also, the project can be promoted at the organization level to mine the news updates of a single or group of organizations together. Also, the adaptability of NEWSD can further be increased by providing a NEWSD App for the handheld devices. These devices would certainly allow the end user to attain the benefits of portability and flexibility features of NEWSD in a much better way.
References


[29] Limeng Cui; Fan Meng; Yong Shi; Minjiang Li; An Liu, "A Hierarchy Method Based on LDA and SVM for News Classification," Data Mining Workshop (ICDMW), 2014 IEEE International Conference on , vol., no., pp.60,64, 14-14 Dec. 2014.


