NEWSD: A Realtime News Classification Engine for Web Streaming Data

Urooj Mohiuddin*, Hameeza Ahmed‡, Muhammad Ali Ismail†
High Performance Computing Centre (HPCC)
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 preprocessing, feature 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, **N**ews **E**xplorer for **W**eb **S**treaming **D**ata. 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. The end users can easily get NEWSD, the tool / news classification engine freely available at [37].

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 done through the exploration and identification of required patterns. It has been playing a significant role in varying applications. Some of the applications exist in real time analytics, sentiment analysis, spam filtering, publishing media, content management & categorization, financial markets, healthcare, and many others. Real time text is capable to analyze text that is streaming continuously on social web [11].

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. 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. 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

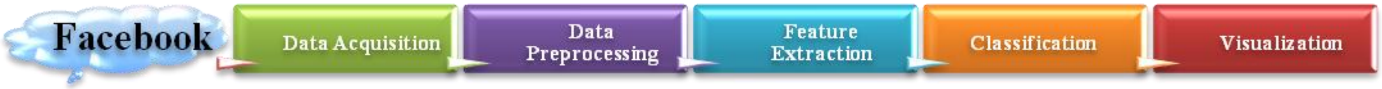


Fig. 1. Components of NEWSD

currently selected newspapers. Also, a training source is prepared by collecting and labeling few news data samples in order to perform the process of machine learning in a supervised manner.

B. Preprocessing and Feature Extraction

In the acquisition phase, the data obtained through Facebook is raw; containing redundant quotations, commas, question mark, full stops and other ambiguous symbols. The gathered data is thus inappropriate for further processing. Hence, these undesirable marks are removed in the initial phase of preprocessing. This initial cleaning is then followed by advanced preprocessing along with feature extraction which is performed using WEKA API. The filtering technique which is adopted for the entire process of preprocessing and feature extraction is known as StringToWordVector. It is mainly responsible to transform a string attribute into a vector of numeric attributes. Also, the filter is able to perform the process of tokenization and indexing by breaking text words into word stems [13]. Consequently, the input text is successfully transformed into vector by this stage of NEWSD.

C. Classification

The obtained vector is then passed through the classification stage. It is the most critical step representing the actual mining of text. The supervised learning is done in two phases namely training and testing. In the training phase, the system is trained using sample trained data. The testing phase predicts the values of new test data based on the training. In NEWSD, the training data is fixed comprising of news samples that are large enough to perform the learning process appropriately. The labels are manually assigned to the training file. The news categorization task is performed on test data which is real time and keeps on changing. But, the test data has no pre-assigned labels as it would be inappropriate to label each fetched instance manually. Thus, this entire task of classification and label assignment of the test data based on the training is handled by NEWSD. Presently, only two classifiers namely NaiveBayes and J48 are used by the classification engine.

D. Visualization

The role of a presentable and user friendly GUI is very significant in NEWSD. It enables the end users to visualize the entire news classification process in an appropriate manner. The GUI of NEWSD is designed using Java Swing Toolkit. Swing provides a large number of visualization facilities to the end user as it acts as the principal GUI toolkit for the Java language.

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 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 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

The project NEWSD has all the capabilities to overcome the existing solutions for news classification. The process of mining of text for the acquisition of knowledge is both complex and challenging. The major goal of NEWSD is to facilitate the end users who have become the regular consumers of social web. It would greatly reduce the cost of searching and exploring by automatically mining data from different news sources and displaying the combined results together. The significance of the data source is higher as the collected data is real time involving continuous variations. The classification of web streaming data is an important achievement of NEWSD. The industry and academia both can be benefitted by using the tool in order to perform the mining of individual sources. The portability, flexibility, and obustness of NEWSD would greatly increase its usability. The

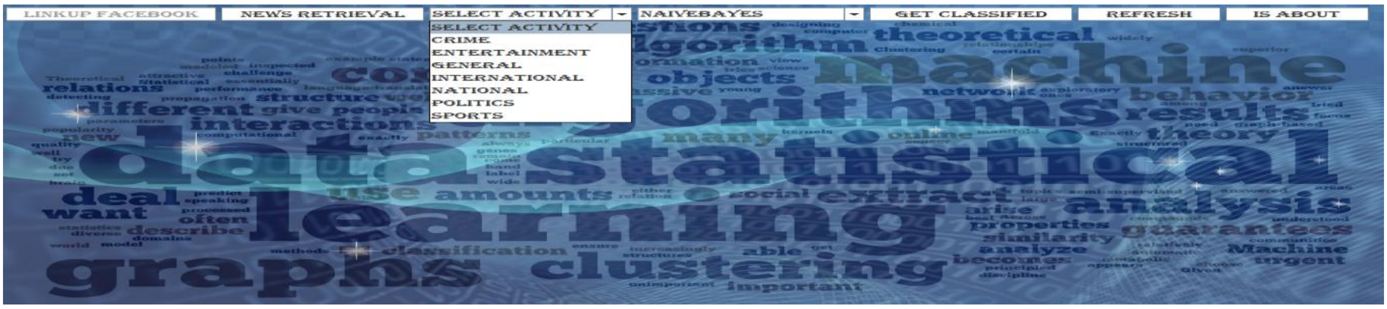


Fig. 2. NEWSD Interface

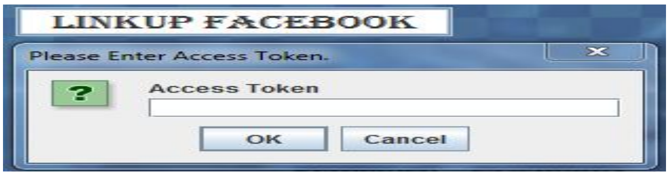
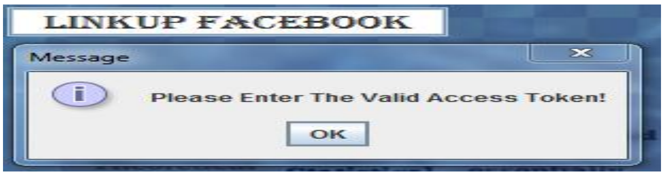
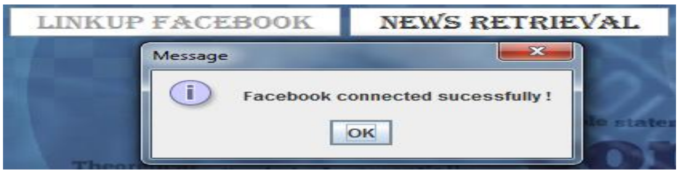


Fig. 3. Access Token Request



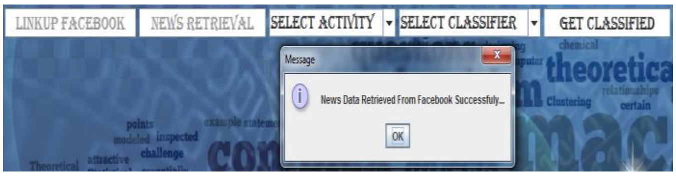
Please Enter The Valid Access Token!

Fig. 4. Access Token Error Message



Facebook connected sucessfully !

Fig. 5. Access Token Success Message



News Data Retrieved From Facebook Successfuly...

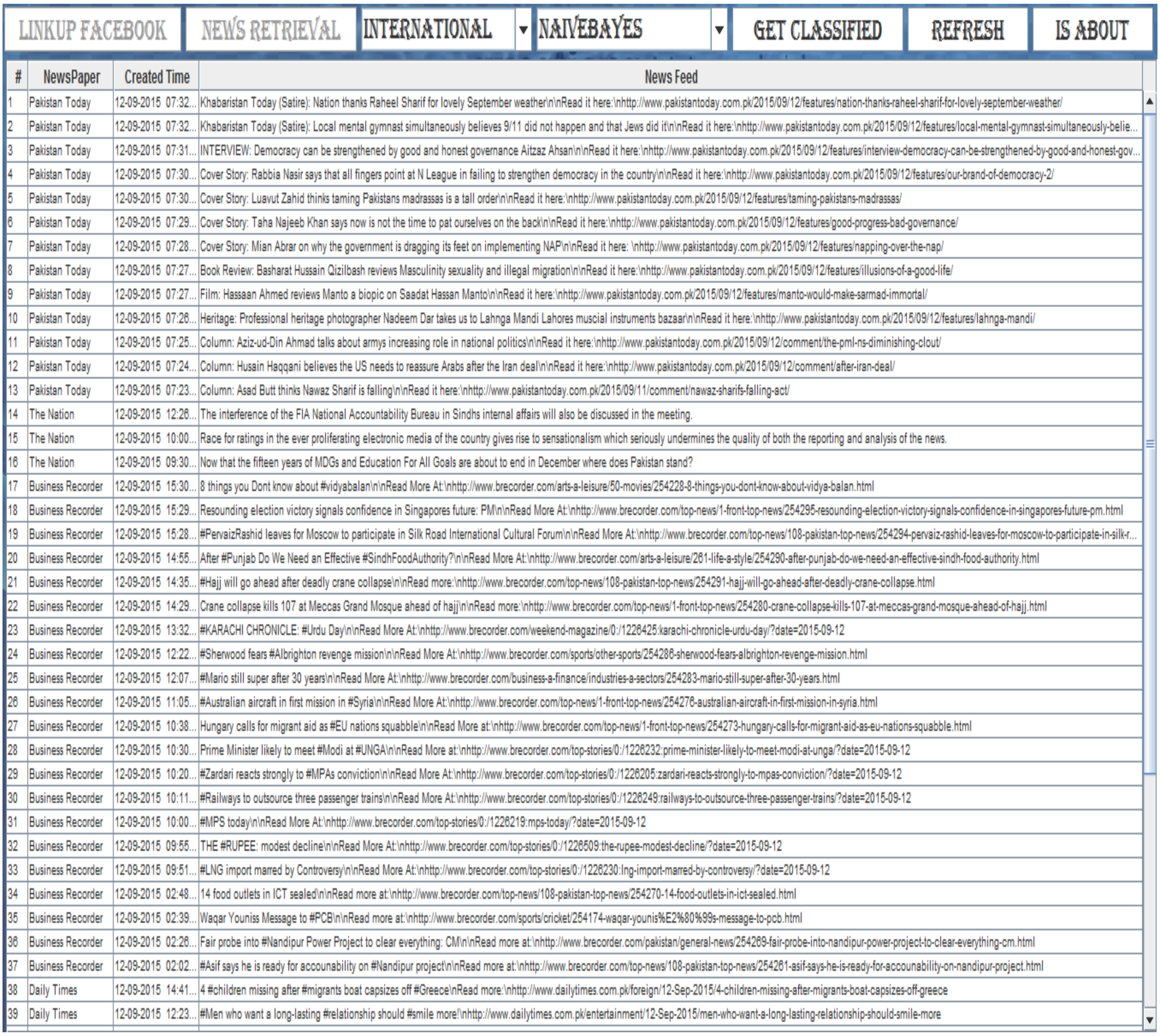
Fig. 6. Data Acquisition Completion

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 NaïveBayes, accuracy is 54.821%; while the accuracy is reduced to 52.6003% for classifier J48 as shown in Table 1. It is inappropriate every time to determine the classification accuracy of the unlabeled test data, as the results are required to be verified manually with different classifiers. Therefore, a rough estimate of the classification accuracy of the news results is calculated manually for individual activities of both the classifiers 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 Naïve 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

Activities	**Naïve Bayes Classification Accuracy (%)**	**J48 Classification Accuracy (%)**
Crime	80	72
International	75	68
National	50	44
Sports	40	20
Entertainment	20	16
Politics	86	75
General	78	70
Total (Test Samples)	61.2857	52.1428
Total (Cross Validation in Training Samples)	54.8291	52.6003



#	NewsPaper	Created Time	News Feed
1	Pakistan Today	12-09-2015 07:32...	Khabaristan Today (Satire): Nation thanks Raheel Sharif for lovely September weather\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/features/nation-thanks-raheel-sharif-for-lovely-september-weather/
2	Pakistan Today	12-09-2015 07:32...	Khabaristan Today (Satire): Local mental gymnast simultaneously believes 9/11 did not happen and that Jews did it\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/features/local-mental-gymnast-simultaneously-belie...
3	Pakistan Today	12-09-2015 07:31...	INTERVIEW: Democracy can be strengthened by good and honest governance Aitzaz Ahsan\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/features/interview-democracy-can-be-strengthened-by-good-and-honest-gov...
4	Pakistan Today	12-09-2015 07:30...	Cover Story: Rabbia Nasir says that all fingers point at N League in failing to strengthen democracy in the country\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/features/our-brand-of-democracy-2/
5	Pakistan Today	12-09-2015 07:30...	Cover Story: Luavut Zahid thinks taming Pakistans madrassas is a tall order\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/features/taming-pakistans-madrassas/
6	Pakistan Today	12-09-2015 07:29...	Cover Story: Taha Najeeb Khan says now is not the time to pat ourselves on the back\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/features/good-progress-bad-governance/
7	Pakistan Today	12-09-2015 07:28...	Cover Story: Mian Abrar on why the government is dragging its feet on implementing NAP\n\nRead it here: \nhttp://www.pakistantoday.com.pk/2015/09/12/features/napping-over-the-nap/
8	Pakistan Today	12-09-2015 07:27...	Book Review: Basharat Hussain Qizilbash reviews Masculinity sexuality and illegal migration\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/features/illusions-of-a-good-life/
9	Pakistan Today	12-09-2015 07:27...	Film: Hassaan Ahmed reviews Manto a biopic on Saadat Hassan Manto\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/features/manto-would-make-sarmad-immortal/
10	Pakistan Today	12-09-2015 07:26...	Heritage: Professional heritage photographer Nadeem Dar takes us to Lahnga Mandi Lahores muscial instruments bazaar\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/features/lahnga-mandi/
11	Pakistan Today	12-09-2015 07:25...	Column: Aziz-ud-Din Ahmad talks about armys increasing role in national politics\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/comment/the-pml-ns-diminishing-clout/
12	Pakistan Today	12-09-2015 07:24...	Column: Husain Haqqani believes the US needs to reassure Arabs after the Iran deal\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/12/comment/after-iran-deal/
13	Pakistan Today	12-09-2015 07:23...	Column: Asad Butt thinks Nawaz Sharif is falling\n\nRead it here:\nhttp://www.pakistantoday.com.pk/2015/09/11/comment/nawaz-sharifs-falling-act/
14	The Nation	12-09-2015 12:26...	The interference of the FIA National Accountability Bureau in Sindhs internal affairs will also be discussed in the meeting.
15	The Nation	12-09-2015 10:00...	Race for ratings in the ever proliferating electronic media of the country gives rise to sensationalism which seriously undermines the quality of both the reporting and analysis of the news.
16	The Nation	12-09-2015 09:30...	Now that the fifteen years of MDGs and Education For All Goals are about to end in December where does Pakistan stand?
17	Business Recorder	12-09-2015 15:30...	8 things you Dont know about #vidyabalan\n\nRead More At:\nhttp://www.brecorder.com/arts-a-leisure/50-movies/254228-8-things-you-dont-know-about-vidya-balan.html
18	Business Recorder	12-09-2015 15:29...	Resounding election victory signals confidence in Singapores future: PM\n\nRead More At:\nhttp://www.brecorder.com/top-news/1-front-top-news/254295-resounding-election-victory-signals-confidence-in-singapores-future-pm.html
19	Business Recorder	12-09-2015 15:28...	#PervaizRashid leaves for Moscow to participate in Silk Road International Cultural Forum\n\nRead More At:\nhttp://www.brecorder.com/top-news/108-pakistan-top-news/254294-pervaiz-rashid-leaves-for-moscow-to-participate-in-silk-r...
20	Business Recorder	12-09-2015 14:55...	After #Punjab Do We Need an Effective #SindhFoodAuthority?\n\nRead More At:\nhttp://www.brecorder.com/arts-a-leisure/261-life-a-style/254290-after-punjab-do-we-need-an-effective-sindh-food-authority.html
21	Business Recorder	12-09-2015 14:35...	#Hajj will go ahead after deadly crane collapse\n\nRead more:\nhttp://www.brecorder.com/top-news/108-pakistan-top-news/254291-hajj-will-go-ahead-after-deadly-crane-collapse.html
22	Business Recorder	12-09-2015 14:29...	Crane collapse kills 107 at Meccas Grand Mosque ahead of hajj\n\nRead more:\nhttp://www.brecorder.com/top-news/1-front-top-news/254280-crane-collapse-kills-107-at-meccas-grand-mosque-ahead-of-hajj.html
23	Business Recorder	12-09-2015 13:32...	#KARACHI CHRONICLE: #Urdu Day\n\nRead More At:\nhttp://www.brecorder.com/weekend-magazine/0:/1226425:karachi-chronicle-urdu-day/?date=2015-09-12
24	Business Recorder	12-09-2015 12:22...	#Sherwood fears #Albrighton revenge mission\n\nRead More At:\nhttp://www.brecorder.com/sports/other-sports/254286-sherwood-fears-albrighton-revenge-mission.html
25	Business Recorder	12-09-2015 12:07...	#Mario still super after 30 years\n\nRead More At:\nhttp://www.brecorder.com/business-a-finance/industries-a-sectors/254283-mario-still-super-after-30-years.html
26	Business Recorder	12-09-2015 11:05...	#Australian aircraft in first mission in #Syria\n\nRead More At:\nhttp://www.brecorder.com/top-news/1-front-top-news/254276-australian-aircraft-in-first-mission-in-syria.html
27	Business Recorder	12-09-2015 10:38...	Hungary calls for migrant aid as #EU nations squabble\n\nRead More at:\nhttp://www.brecorder.com/top-news/1-front-top-news/254273-hungary-calls-for-migrant-aid-as-eu-nations-squabble.html
28	Business Recorder	12-09-2015 10:30...	Prime Minister likely to meet #Modi at #UNGA\n\nRead More at:\nhttp://www.brecorder.com/top-stories/0:/1226232:prime-minister-likely-to-meet-modi-at-unga/?date=2015-09-12
29	Business Recorder	12-09-2015 10:20...	#Zardari reacts strongly to #MPAs conviction\n\nRead More At:\nhttp://www.brecorder.com/top-stories/0:/1226205:zardari-reacts-strongly-to-mpas-conviction/?date=2015-09-12
30	Business Recorder	12-09-2015 10:11...	#Railways to outsource three passenger trains\n\nRead More At:\nhttp://www.brecorder.com/top-stories/0:/1226249:railways-to-outsource-three-passenger-trains/?date=2015-09-12
31	Business Recorder	12-09-2015 10:00...	#MPS today\n\nRead More At:\nhttp://www.brecorder.com/top-stories/0:/1226219:mps-today/?date=2015-09-12
32	Business Recorder	12-09-2015 09:55...	THE #RUPEE: modest decline\n\nRead More At:\nhttp://www.brecorder.com/top-stories/0:/1226509:the-rupee-modest-decline/?date=2015-09-12
33	Business Recorder	12-09-2015 09:51...	#LNG import marred by Controversy\n\nRead More At:\nhttp://www.brecorder.com/top-stories/0:/1226230:lng-import-marred-by-controversy/?date=2015-09-12
34	Business Recorder	12-09-2015 02:48...	14 food outlets in ICT sealed\n\nRead more at:\nhttp://www.brecorder.com/top-news/108-pakistan-top-news/254270-14-food-outlets-in-ict-sealed.html
35	Business Recorder	12-09-2015 02:39...	Waqar Youniss Message to #PCB\n\nRead more at:\nhttp://www.brecorder.com/sports/cricket/254174-waqar-younis%E2%80%99s-message-to-pcb.html
36	Business Recorder	12-09-2015 02:26...	Fair probe into #Nandipur Power Project to clear everything: CM\n\nRead more at:\nhttp://www.brecorder.com/pakistan/general-news/254269-fair-probe-into-nandipur-power-project-to-clear-everything-cm.html
37	Business Recorder	12-09-2015 02:02...	#Asif says he is ready for accounability on #Nandipur project\n\nRead more at:\nhttp://www.brecorder.com/top-news/108-pakistan-top-news/254261-asif-says-he-is-ready-for-accounability-on-nandipur-project.html
38	Daily Times	12-09-2015 14:41...	4 #children missing after #migrants boat capsizes off #Greece\nRead more:\nhttp://www.dailytimes.com.pk/foreign/12-Sep-2015/4-children-missing-after-migrants-boat-capsizes-off-greece
39	Daily Times	12-09-2015 12:23...	#Men who want a long-lasting #relationship should #smile more!\nhttp://www.dailytimes.com.pk/entertainment/12-Sep-2015/men-who-want-a-long-lasting-relationship-should-smile-more

Fig. 7. International News Classification

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

[1] Russell, Matthew A. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. " O'Reilly Media, Inc.", 2013.

[2] Social Networking Sites. https://www.nsa.gov/ia/_files/factsheets/i73-021r-2009.pdf.

[3] Radovanović, Miloš, and Mirjana Ivanović. "Text mining: Approaches and applications." Novi Sad J. Math 38.3 (2008): 227-234.

[4] Gupta, Vishal, and Gurpreet S. Lehal. "A survey of text mining techniques and applications." Journal of emerging technologies in web intelligence 1.1 (2009): 60-76.

[5] Nahm, Un Yong, and Raymond J. Mooney. "Text mining with information extraction." AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases. Vol. 1. 2002.

[6] Witten, Ian H. "Text mining." Practical handbook of Internet computing (2005): 14-1.

[7] Feature Extraction. http://scikit-learn.org/stable/modules/feature_extraction.html.

[8] Ahmed, Hameeza; Ismail, Muhammad Ali, "Comparative Analysis of Activity Recognition Classifiers using Big Dataset," Presented at 30th IEEEP Multi-Topic International Symposium, March 25-26, 2015, Karachi, Pakistan.

[9] Tina R. Patil, Mrs. S. S. Sherekar "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications Vol. 6, No.2, Apr 2013.

[10] Gupta, D. L., A. K. Malviya, and Satyendra Singh. "Performance analysis of classification tree learning algorithms." IJCA) International Journal of Computer Applications 55, no. 6 (2012).

[11] Chakraborty, Goutam, Murali Pagolu, and Satish Garla. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. SAS Institute, 2014.

[12] Ramasubramanian, C., and R. Ramya. "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm." International Journal of Advanced Research in Computer and Communication Engineering2.12 (2013): 2278-1021.

[13] Nihil Obstat, " Text Mining in WEKA", http://jmgomezhidalgo.blogspot.com/2013/01/text-mining-in-weka-chaining-filters.html

[14] Ariki, Y.; Sugiyama, Y., "Classification of TV sports news by DCT features using multiple subspace method," Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on , vol.2, no., pp.1488,1491 vol.2, 16-20 Aug 1998.

[15] Chung-Hsien Wu; Chia-Hsin Hsieh, "Story Segmentation and Topic Classification of Broadcast News via a Topic-Based Segmental Model and a Genetic Algorithm," Audio, Speech, and Language Processing, IEEE Transactions on , vol.17, no.8, pp.1612,1623, Nov. 2009.

[16] Peng Wang; Rui Cai; Shi-Qiang Yang, "A hybrid approach to news video classification multimodal features," Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on , vol.2, no., pp.787,791 vol.2, 15-18 Dec. 2003.

[17] Ariki, Y.; Matsuura, K., "Automatic classification of TV news articles based on telop character recognition," Multimedia Computing and Systems, 1999. IEEE International Conference on , vol.2, no., pp.148,152 vol.2, Jul 1999.

[18] Agarwal, S.; Singhal, A.; Bedi, P., "Classification of RSS feed news items using ontology," Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on , vol., no., pp.491,496, 27-29 Nov. 2012.

[19] Setty, S.; Jadi, R.; Shaikh, S.; Mattikalli, C.; Mudenagudi, U., "Classification of facebook news feeds and sentiment analysis," Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on , vol., no., pp.18,23, 24-27 Sept. 2014.

[20] Xin Liu; Gao Rujia; Song Liufu, "Internet news headlines classification method based on the N-Gram language model," Computer Science and Information Processing (CSIP), 2012 International Conference on, vol., no., pp.826,828, 24-26 Aug. 2012.

[21] Dilrukshi, I.; de Zoysa, K.; Caldera, A., "Twitter news classification using SVM," Computer Science & Education (ICCSE), 2013 8th International Conference on , vol., no., pp.287,291, 26-28 April 2013.

[22] Chy, A.N.; Seddiqui, M.H.; Das, S., "Bangla news classification using naive Bayes classifier," Computer and Information Technology (ICCIT), 2013 16th International Conference on , vol., no., pp.366,371, 8-10 March 2014.

[23] Lauren, S.; Harlili, S., "Stock trend prediction using simple moving average supported by news classification," Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 International Conference of , vol., no., pp.135,139, 20-21 Aug. 2014.

[24] Kilic, E.; Tavus, M.R.; Karhan, Z., "Classification of breaking news taken from the online news sites," Signal Processing and Communications Applications Conference (SIU), 2015 23th , vol., no., pp.363,366, 16-19 May 2015.

[25] Duarte, E.M.; Braga, A.P.; Braga, J.L., "Internet economic news gathering and classification: a neural network software agent based approach," Neural Networks, 2002. SBRN 2002. Proceedings. VII Brazilian Symposium on , vol., no., pp.112,, 2002.

[26] Selamat, A.; Yanagimoto, H.; Omatu, S., "Web news classification using neural networks based on PCA," SICE 2002. Proceedings of the 41st SICE Annual Conference , vol.4, no., pp.2389,2394 vol.4, 5-7 Aug. 2002.

[27] Deng-Yiv Chiu; Chi-Chung Lee; Ya-Chen Pan, "A Classification Approach of News Web Pages from Multi-Media Sources at Chinese Entry Website-Taiwan Yahoo! as an Example," Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on , vol., no., pp.1156,1159, 7-9 Dec. 2009.

[28] Kroha, P.; Baeza-Yates, R., "A Case Study: News Classification Based on Term Frequency," Database and Expert Systems Applications, 2005. Proceedings. Sixteenth International Workshop on , vol., no., pp.428,432, 26-26 Aug. 2005.

[29] Limeng Cui; Fan Meng; Yong Shi; Minqiang 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.

[30] Wei Hu; Dong-Mo Zhang; Huan-Ye Sheng, "Vague events-based Chinese Web news classification," Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on , vol.3, no., pp.1545,1549 vol.3, 26-29 Aug. 2004.

[31] Rachmania, A.; Jaafar, J.; Zamin, N., "Likelihood calculation classification for Indonesian language news documents," Information Technology and Electrical Engineering (ICITEE), 2013 International Conference on , vol., no., pp.149,154, 7-8 Oct. 2013

[32] Dilrukshi, I.; de Zoysa, K., "Twitter news classification: Theoretical and practical comparison of SVM against Naive Bayes algorithms," Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on , vol., no., pp.278,278, 11-15 Dec. 2013.

[33] Gomes, H.; de Castro Neto, M.; Henriques, R., "Text Mining: Sentiment analysis on news classification," Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on , vol., no., pp.1,6, 19-22 June 2013.

[34] Billsus, Daniel, and Michael J. Pazzani. A hybrid user model for news story classification. Springer Vienna, 1999.

[35] Katariya, Ms Nikita P., et al. "TEXT PREPROCESSING FOR TEXT MINING USING SIDE INFORMATION." (2015).

[36] Text Mining. https://semantria.com/text-mining.

[37] NEWSD. http://www.neduet.edu.pk/hpcc/download.html