SIMILARITY OF TRENDING NEWS: A CASE STUDY OF BANGLADESH

Sabbir Ahmed, Md. Zakib Uddin Khan, Mir Ummay Touhida, Shahinuzzaman Shawon
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh.
sabbir4402@diu.edu.bd, zakib287@diu.edu.bd, touhida15-1910@diu.edu.bd, shawon3573@diu.edu.bd

Abstract: News production and its spreading have rapidly been changed by the social networking sites for instance Facebook. All the Bangla news exists on social media is in textual format which is unstructured as well. Different techniques of Text mining play a vital role in order to convert those Bangla unstructured news into formative knowledge. As there are lack of analysis regarding Bangla news of Facebook posts have been introduced, present study looks for drawing a pattern that refers a constructive knowledge from huge amount of data. To accomplish that, three newspapers have been chosen, namely ProthomAlo, Juganthor and Daily NayaDiganta. Facepager tool has been used to extract data from the Facebook pages of aforementioned newspapers and later data was processed through Spyder, an environment to run Python program. Consequence stated that 'Bangladesh National Election' along with the 'Political Issues' received maximum coverage followed by the recent 'Football World Cup'. Besides, the most frequent newspaper that shares posts on Facebook is Juganthor followed by Daily Naya Diganta and ProthomAlo, respectively. It is also to be said that there is a significant resemblance between Juganthor and Daily Naya Diganta in posting identical posts on Facebook.

Keywords: News similarity, trending news, word cloud, text mining, Bangladeshi newspaper, social media news analysis.

1. Introduction

Newspapers plays a vital role in our everyday life and undoubtedly a great source of day to day information. People around the world had to wait for printed newspaers every morning in pre internet era. But with the development of world wide web it has become easier to access newspapers at present as there are various news portals available in online to provide news updates as fast as possible. At the same time availability of hand held devices made accessing these online newspaser more easy around the globe [1]. Emergence of social media has added a new dimension on the availability of these newspaers as most of the newspaper are using there own approaches in social media to reach more and more readers for their portals where social media is defined by [15] as an online based innovation which let a user to share his or her thoughts, information, interests and various types of expression through this virtual network. Among the popular social media sites Facebook, Twitter and Linkedin are used in Bangladesh frequently for social communications and networking which includes updating status, video, audio or photographs sharing [2] and amount of activitivies proves that users spend a lot of time in this platform. This changes in social behavior of the people of this country made our traditional journalism to change their approaches to share their news contents or breaking news. As a result to reach maximum number of readers popular newspaper of the country are sharing their news in social platform regularly.

From the context of data, the news provided by these newspapers into the social media, mostly are text data which is considered as unstructed data from the Text Mining point of view. Text mining is particularly used as the way of bringing out indefinite and practical models or information from a summation of enormous and unstructured data or corpus [3-4]. On the other hand Information Retrieval (IR), Computational linguistics and data mining are also some popular research fields those are often incorporated by Text Mining which is one of the reknown branches of data mining [5]. Text Mining techniques are used every now and then in keyword extraction, topic detection, topic modeling, document clustering, sentiment analysis and text summarization. Similarly, Natural Language Processing is another research field highly correlated with Text Mining which often deals with enormous amount of unstructured textual data [6]. News generation and consumption of User Generated Content has made the recent media discussion, a possible field of research.

In the present Bangla is one of the most popular mother toungues and is currently the number sixth most widely spoken language all over the world. Therefore, it is obvious that a huge umber of bangla contents are shared everyday by social media users and among

those newspaper contents are very common where the source of these contents are undoubtedly the newspaper itself. They post a good number of news everyday in social media and again they are also a huge source of information from text mining point of view. This research aims to explore a text mining technique using this huge opportunity as there have not been enough literature on the analysis of Bangla newspaper text and in order to achieve our research goals the following research questions have been adopted:

I) What are the most frequent connected words that newspapers Facebook pages post on Facebook?

II) To what extend do the newspapers share posts on Facebook?

III) To what extend do the newspapers cover identical issue?

2. Literature Review

In [1,7,8] authors have claimed that a huge number of writers and readers are attracted consequitively by social media. Success of many social media and networks lies in the amount users they are able to contain successively and mass media is considered outdated due to the success of social media now a days [2] as social media have become the a source of contemporary news of newspapers and here the users have the option to choose which the news of his own interest.

An in depth study revealed different techniques of text mining by [8] where the researchers have focused on different models on social networking and different online applications where classification and clustering are observed as two most popular approaches for mining unstructured text data. In another work [9], researchers applied RapidMiner, a popular built in tools for data scientists, in order to analyze unstructured English text into a quantifiable data collected from different social media where the portals have posted their news.

“Arab Spring” was the core of the research by [10] when the authors tried to concentrate on sentiment analysis from the users’ posts in social media. Lexical analysis along with Support Vector Machine and Naïve Bayes were used used inorder to classify users sentiments. Another research was conducted by [11] where the researchers claimed that decision making could be easier for the business owners if they deploy text mining approaches to analyze customers reviews. They have used Facebook page of SAMSUNG mobile in order to collect data and for necessary analyses in support of their findings.

The structured approach has been suggested by the researchers to analyze social media data that include only the comments in English language. To extract quantifiable data from social media, researchers of [11] outlined a straightway approach to existing knowledge. The consequence of such quantification can be performed in studies, surveys and the plan of decision-making frameworks. Yet, the research failed to notice regularly changing example and progression of Facebook users.

With increasing popularity of social media it is also observed that engaenet of students with this platform has increased parallely and they have been found using this platform for clarifying diffierent contents which is again a huge source of research data for text ming. Researchers of [12] considered students casual discussions via web-based networking media concentrating on their emotions, opinions and worries about their learning knowledge. An example of 25.000 engineering students’ tweets related with their school life was examined by the researcher. The consequence of the investigation uncovered that various problems for example study load, sleep agony and lack of social engagement.

Moreover, an investigation was focused on extracting knowledge from university students’ information available on social media sites. Using K-means, a data mining technique to extract constructive information of educational sector, the author of [13] conducted a questionnaire for university students from different field of studies and analyzed the answers through data mining technique. Facebook, Orkut, and Twitter are most frequent sites used by the university students, study revealed.

Nevertheless, for the new researchers text mining provides huge scope for learning and at same time gaining experiences of Natural Language Processing as it does not help only to reduce a huge volume of manual task but also helps to find out different hidden pattern in those unstructured data. However, one of the vital aspects of study which is newspapers' social networking data analysis seems to be overlooked, though significant research on social media data mining has already been performed. To be more precise, no research has been recorded regarding unstructured Bangla news analysis yet. Therefore, current study seeks to have meaningful information after analyzing a large scale of data sets extracted from three popular newspapers' Facebook pages.

3. Research Methodology

Text mining is the way of bringing out constructive data from different text which is also referred as text analytics. Text analysis includes data recovery to consider frequency, data extraction, information mining procedures including connection and association, visualization of data and predictive analysis. The general objective is, basically, to transform unstructureddata into information for analysis through Natural Language Processing (NLP) and analytical strategies.

Due to the inconsistency of natural texts, it is often complicated to mine unstructured data by Natural Language Processing(NLP), statistical modeling and machine learning. It usually causes ambiguous consequence due to inconsistent syntax, semantics along with slang, double intenders and sarcasm. However, the method adopted for this research is shown in figure 1 and the steps are described in the consequently.

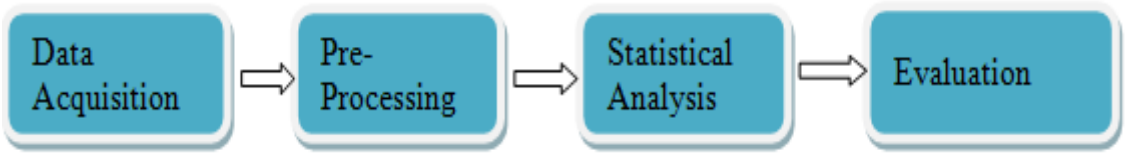

Figure 1: Overall methodology

3.1 Data Collection Procedure

In the current research, data has been extracted from the Facebook page of three different Bangladeshi newspapers, namely, ProthomAlo, Juganthor and Daily NayaDigantha. A Facebook graph API named ''Facepager'' is used to extract from the Facebook page. The extracted data is stored in a local database which is later exported into a CSV form. The current study seeksto analyze the Facebook pages' posts of aforementioned Bangladeshi Newspapers. The total number of extracted Facebook posts of three newspapers is 15,093**.**

3.2 Data Preprocessing

Data preprocessing is considered as one of the most vital part of present study. The collected raw data from different the Facebook pages of newspapers are not quantifiable. In order to get the expected outcome through the analysis process, all the raw data sets need to be pre-processed. Data preprocessing includes the following aspects-

Data Filtering

A good number of missing datais found after the extraction of data from Facebook by 'Facepager'. All the missing values has been filtered through the elimination of missing data.

Excluding Irrelevant variables

A number of irrelevant variables are identified. All the irrelevant variables need to be excluded in order to enhance the efficiency of collected data sets.

Removal of Special Characters

The collected raw data sets contain some special characters that include all the Bangla and English punctuations along with all the numbers different digits.

Tokenization

Tokenization splits a pre-defined document of data into pieces. All the extracted Bangla texts are tokenized that refers the representation of each word.

Removal of Stop-Words

Stop words are defined as the most commonly used words in any languages such as articles, prepositions, pronouns, verbs and adverb. There are number stop words belong to Bangla Language. All the stop words are removed in order to enhance the data quality before the commencement of analysis phase.

After having the above mentioned procedure of Pre-processing accomplished, a clean data set gets prepared for text analysis to turns out the expected outcome. The aforementioned procedures are as follows-

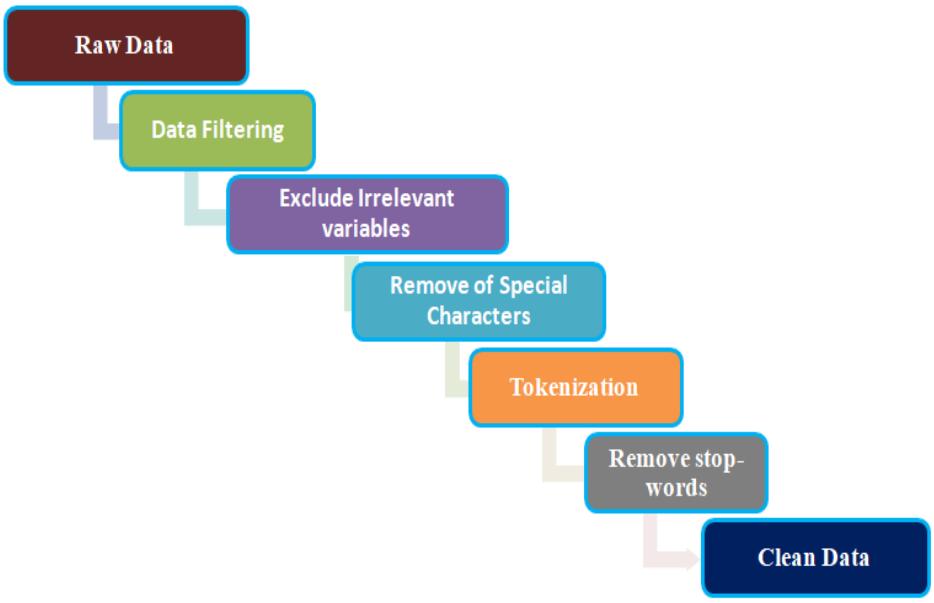

Figure 2: Document Processing steps

Overall statistics of the data is presented in the table 1 which presents that total 15093 news were collected from facebook from 3 different newspapers namely Prothom Alo, Jugantor, Daily Naya Diganto. Among these newspapers maximum number of news were collected from the Prothom Alo.

Table 1: Data Corpus

Extracted Data from	Number of Data
ProthomAlo	5882
Juganthor	4686
Daily NayaDiganta	4525
Total Data	15093

4. Experimental analysis

To answer the very first research question, word frequency technique has been applied on the extracted and pre-processed clean data set of different three newspapers. It is clearly noticed that the top fifteen most frequently used words across the newspapers are "বাংলাদেশ", "সরকার", "প্রধানমন্ত্রী", "সংসদ", "নির্বাচন", "আওয়ামী", "লীগ", "ফুটবল", "বিএনপি", "বিশ্বকাপ", "শেখ", "হাসিনা", "খালেদা", "জামায়াত", "জিয়া". All those frequently used words across the newspapers clearly indicate the upcoming election of Bangladesh along with the political issues and the recent Football World Cup. The correlation among the newspapers expresses that newspapers' Facebook posts are well concerned about the talked topic issues. The mentioned keywords refer the correlation between the words related to election and name of the politicians along with their political party. And the other keywords clearly indicate the most recent prestigious event of Football and Bangladeshi newspapers are well aware about such global event. Besides, figure 3 demonstrates that the newspaper named Juganthor's Facebook page covers most frequent news regarding the aforementioned issues followed by Daily NayaDiganta and ProthomAlo, respectively.

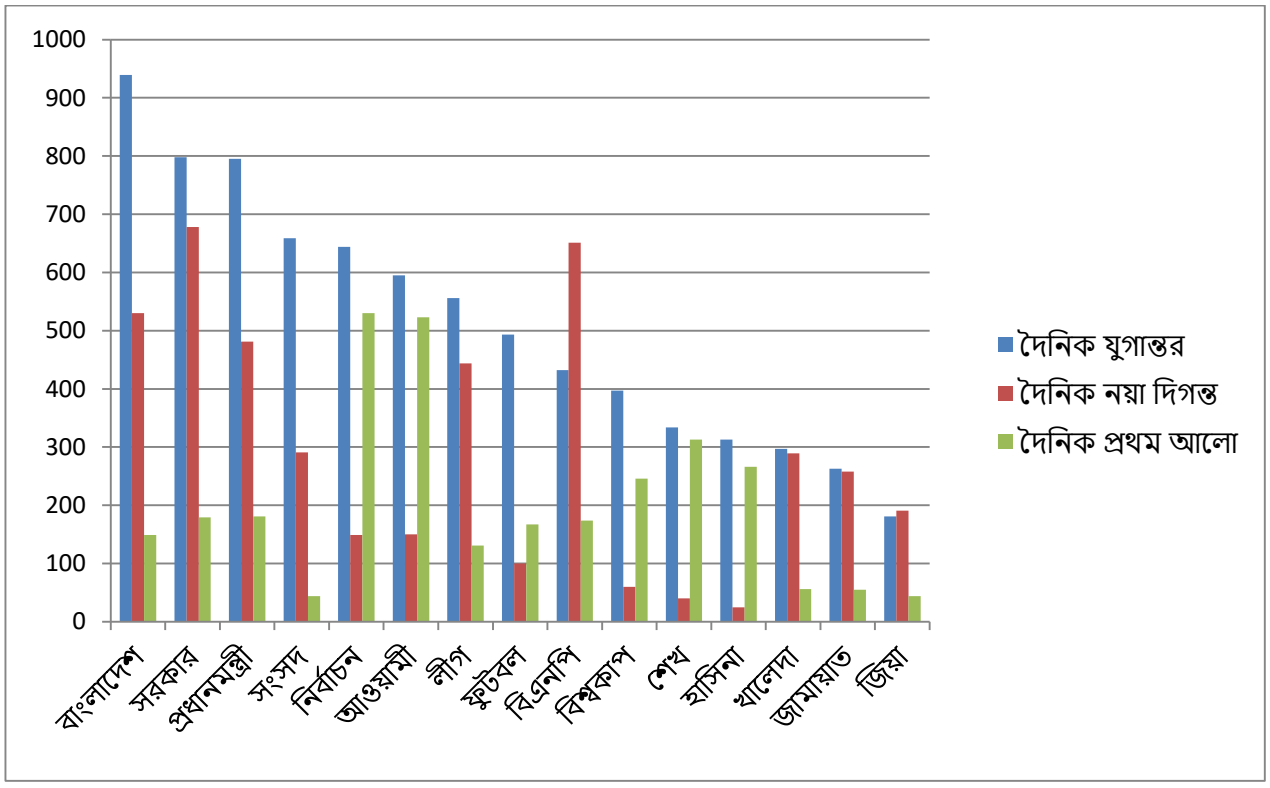

Figure 3: Allocation of word frequency across the newspapers

To answer the second research question, the extracted data has been analyzed through the word frequency technique to indicate the newspaper which shares most of the posts on its Facebook page. In figure 4 illustrates the most repeatedly shared posts regarding the

Bangladesh election along with political issues and the most recent football world cup were by the newspaper “Juganthor’’ followed by “Daily NayaDiganta” and “ProthomAlo”, respectively. It’s clearly said that the newspaper named “Juganthor” was most biased in sharing the recent trends on their Facebook page compared to other newspapers.

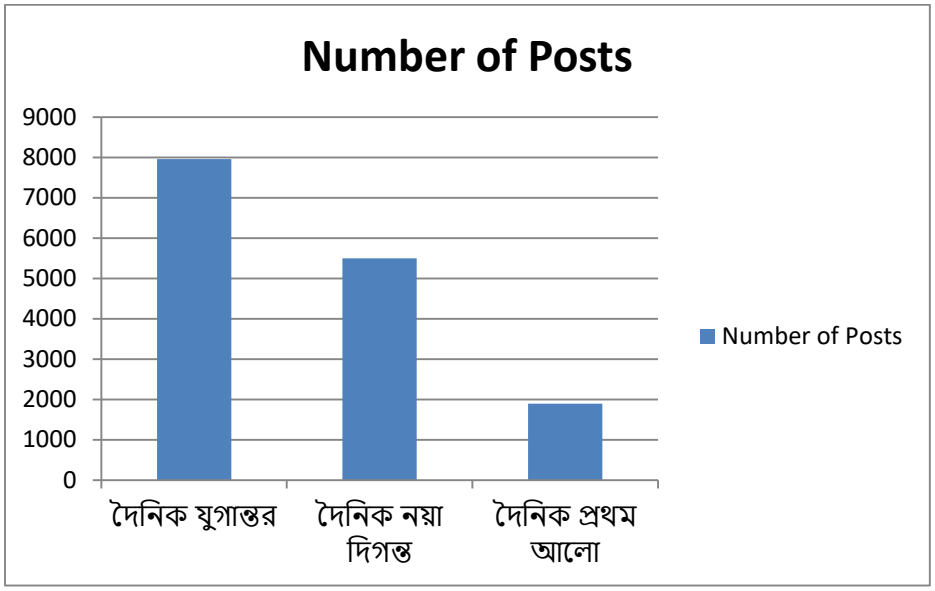

Figure 4: Number of posts shared by the newspapers on Facebook

To answer the last research question, in accordance with [15], similarity operators have applied to the extracted and processed clean data set to determine the topics which are identical to each other. As per Figure 5, the newspaper “Juganthor” refers the value (1.0), “Daily NayaDigantho” denoted by the value(2.0) and value(3.0) stand for “ProthomAlo”. It has been observed that there is a correlation between “Juganthor” and “Daily NayaDigantho” in posting Facebook posts that are relatively identical to each other. The percentage of posting identical posts on Facebook between “Juganthor” and “Daily NayaDiganta” is 52% which considered as the highest. And the percentage relationship between “NayaDiganta” and “ProthomAlo”, “Juganthor” and “ProthomAlo”are 30% and 18%, respectively.

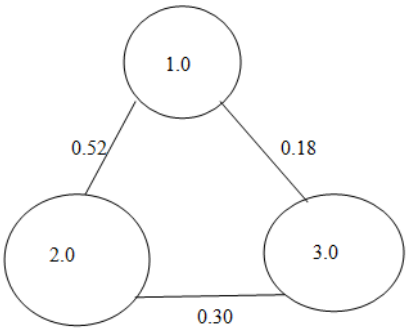

Figure 5: Resemblance across the newspapers

5. Conclusion

To transform unstructured Bangla accessible texts of social media into a constructive knowledge, three different Facebook pages of Bangladeshi newspapers were chosen. To accomplish task, the study went through the different steps till its accomplishment. Finally, constructive information was gained from a vast amount of unstructured data which was 15,093. To conclude, nowadays

media in the present globalized world has not been the same as it was but no precise analysis regarding Bangla text analysis despite having a vast amount of data on social media. Present study addressed some techniques to process quantifiable Bangla news and covert into qualitative knowledge. A pattern can be introduced for Bangla texts that might result a constructive knowledge from comprehensive data without going through the complicated procedures that include the various stairs of pre-processing and analysis phase. A pattern can be introduced for Bangla texts that might result a constructive knowledge from comprehensive data without going through the complicated procedures that include the various stairs of pre-processing and analysis phase.

References

1. Mhamdi, C. (2016). Transgressing media boundaries: News creation and dissemination in a globalized world. *Mediterranean Journal of Social Sciences*, *7*(5), 272-272.
2. Alejandro, J. (2010). Journalism in the age of social media. *Reuters Institute Fellowship Paper*, *5*, 1-47.
3. Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases* (Vol. 8, pp. 65-70). sn.
4. Feldman, R., & Dagan, I. (1995, August). Knowledge Discovery in Textual Databases (KDT). In *KDD* (Vol. 95, pp. 112-117).
5. Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J*, *2*(1), 127-133.
6. Salloum, S. A., Al-Emran, M., & Shaalan, K. (2016). A survey of lexical functional grammar in the Arabic context. *International Journal of Computing and Network Technology*, *4*(03).
7. Comscore Media Matrix. (2008). Huffington Post and Politico leadwave of explosive growth at independent political blogs and news sites this election season. Retrieved from http://www.comscore.com/press/release.asp?press=2525
8. Sifry, D.: State of the blogosphere. Retrieved from Technorati (2008). http://technorati.com/ blogging/state-of-the-blogosphere
9. Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., ... & Li, H. (2015). A survey on text mining in social networks. *The Knowledge Engineering Review*, *30*(2), 157-170.
10. Shaalan, K., Hassanien, A. E., & Tolba, F. (Eds.). (2017). *Intelligent natural language processing: trends and applications* (Vol. 740). Springer.
11. Hamouda, S. B., & Akaichi, J. (2013). Social networks' text mining for sentiment classification: The case of Facebook'statuses updates in the 'Arabic Spring'era. *International Journal Application or Innovation in Engineering and Management*, *2*(5), 470-478.
12. Chan, H. K., Lacka, E., Yee, R. W., & Lim, M. K. (2014, December). A case study on mining social media data. In *2014 IEEE International Conference on Industrial Engineering and Engineering Management* (pp. 593-596). IEEE.
13. Chen, X., Vorvoreanu, M., & Madhavan, K. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on learning technologies*, *7*(3), 246-259.
14. Singh, A. (2017). Mining of Social Media data of University students. *Education and Information Technologies*, *22*(4), 1515-1526.
15. Huang, A. (2008, April). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (Vol. 4, pp. 9-56).