A Learning Dataset Aimed at Predicting the Feedbacks for Bengali Blogs

Shovra Das, Md. Hasibul Hasan, Md Shamsur Rahim and Mohammod Hafizur Rahman

***Abstract*— Since the innovation of blogging, social media began to explode its popularity. As huge amount of documents appearing in social media, there is an enormous opportunity for the analysis of such documents to find interesting relevance across this ocean of data. This work focuses on the analysis of documents appearing in blogs particularly in Bengali blogs. The key goal of this work is to prepare a machine learning dataset to analyse the data extracted from social media. Additionally the idea of a prediction model has been invoked which allows to forecast the number of feedbacks that a blog document is expected to receive in near future. For these experiments, blog documents were crawled from the Internet and processed to facilitate research.**

***Keywords*—text mining; web mining; social media; feedback prediction; machine learning.**

I. Introduction

The invention of blogging just exploded the popularity of social media. The previous few years there has been incredible growth of the significance of social media. While in the early days of social media they were more or less used as entertainment purpose mainly by some enthusiastic users (Buza 2014). Today social media are extensively used for promoting new products and services. While doing so, if negative opinions appear in social media about a company, it is very expected that the company might have to react quickly, in order to avoid projected losses.

With the popularity of the social networks, a large volume of documents are appearing in these networks. The analysis of these documents can facilitate enormous opportunity to find interesting relations and relevance across the ocean of data. However, for the analysis we need to take some application domain's specific special attributes into account. Especially the variety, uncontrolled and rapidly-updating contents of social media documents: e.g. when a blog post appears, users may post positive or negative feedback immediately.

We followed a specific set of actions to achieve the goal. The action involves the following major components: (i) the crawler, (ii) information extractors, (iii) data store and (iv) analytic components according to (Buza 2014). In this paper, we focus on the data preparation mechanism so that the analytic components can be developed. We constructed a machine learning dataset with the available data attributes and based on that we proposed a prediction model which is expected to predict the number of feedbacks that a document is expected to receive in the next H hours. For data preparation, we focused on the documents appearing in Bengali blogs and performed all sort of processing techniques including language specific pre-processing to support the idea. For these experiments we crawled documents from several popular Bengali blog sites to ensure the worth of our collected data.

II. Related works

Several mining techniques on social media have been studied by many researchers over the time, see e.g. (Reuter et al. 2011) and (Marinho et al. 2008). Most closely related to our work is the paper of (Buza 2014) targeted on varieties of topics and performed experiments with a sufficient variety of mining models.

Buza at [1] presented a proof-of-concept industrial application where they demonstrated that it is possible to predict the number of feedbacks that a blog document will receive in near future. For the prediction, they used various mining algorithms in their experiments with certain accuracy.

We are up to extend the concept developed by (Buza 2014) where he worked on the documents appeared on Hungarian blogs. But our focus is on Bengali blogs targeting the Bangladeshi business community. We have developed the learning data set according to (Buza 2014) and up to test the prediction model the author provided on the mentioned work. But in our case we are a little extensive about extracting the features to achieve a better accuracy. We have added some Post and Blogger specific statistical measures which might lead us to come up with an enriched model for feedback prediction.

Shovra Das is a Lecturer of Department of Computer Science at American International University-Bangladesh (AIUB), Banani, Dhaka-1213, Bangladesh. Email: shovra.das@aiub.edu
Md. Hasibul Hasan is a Lecturer of Department of Computer Science at American International University-Bangladesh (AIUB), Banani, Dhaka-1213, Bangladesh. Email: hasibul.hasan@aiub.edu
Md. Shamsur Rahim is a Lecturer of Department of Computer Science at American International University-Bangladesh (AIUB), Banani, Dhaka-1213, Bangladesh. Email: shamsur@aiub.edu
Mohammod Hafizur Rahman is an Assistant Professor of Department of Computer Science at American International University-Bangladesh (AIUB), Banani, Dhaka-1213, Bangladesh. Email: hafizur@aiub.edu

III. Background

A. The Problem:

As we know the all the relevant post and blog specific aspects of the documents appeared in the past along with the temporal information we aim to predict the number of feedbacks that a blog document is expected to receive in near future. The term future can be expressed in hours or days.

B. Domain-specific Concepts:

As the dataset was to be prepared aiming at proposing a Feedback prediction model we needed to define some domain-specific concepts and specify the problem boundary precisely

On our point of view a source produces documents. For example, on the site *somewhereinblog.net/blogger/*, new documents appear in a regular basis; therefore, we say that somewhereinblog.net/blogger/ is the source of these documents. The most relevant parts of a document are as follows:

(i) Post: Buza [1] defines a post as: "main text of the document: the text that is written by the author of the document, this text describes the topic of the document".

(ii) Metadata: additional information about post such as category, number of times read, no of likes etc.

(iii) Feedbacks: opinions of social media users about the main text of the document often referred to as comments. Temporal features are most relevant to our work.

(iv) Blogger specific statistical information such as how long the blogger is writing post, how many feedbacks she/he has received, how many comments she/he has given etc. These attributes are taken to analyze the popularity of a blogger which can be significant for a specific post to get maximum feedbacks.

Fig. 1: Domain-specific concepts

C. Preparing the dataset for Machine Learning:

To feed the regression algorithms the instances are needed to be vectors. We have resolved two issues:

(i) We have transformed the instances (blog documents) into vectors,

(ii) We have generated train data.

Buza at [1] has demonstrated that if the data in such a manner is prepared then models like neural networks (multilayer perceptron's in particular), Radio Basic Function(RBF)-networks, regression trees (REP-tree, M5P-tree) and multivariate linear regression, nearest neighbour models and bagging out of the ensemble models might be used to predict the desired result.

D. Feature Extraction:

For turning the documents into vectors, we extract the following features from each document:

(i) Basic features: the number of feedbacks and how they increased or decreased with respect to a base time **B.**

(ii) Textual features: the most frequently used bag of words features.

(iii) Weekday features: On which day of the week the post was published and for which day the prediction need to be predicted.

(iv) Post Specific Features:

- No of times Read
- No of Likes
- Text Length
- Post external Links Count
- Post Image Count
- Post Video Count (YouTube video only)

(v) Blogger Specific Features:

- Blog Visit Count
- Blogger's total Post Count
- No of comments by Blogger
- No of comments Blogger received
- Blogger's Age (Blogging duration)

We would like to select some date and time in the past and simulate as if the current date and time would be the selected date and time. We refer the selected date and time as base time. As we actually know what happened after the base time, we know how many feedbacks the documents received in the next particular hours after that base time.

IV. Methodology: the dataset preparation

The data originates from Bengali blog posts. The raw HTML-documents of the blog posts were crawled and processed and therefore each instance corresponds to a blog post. The overall set of actions we performed is illustrated in the diagram right below:

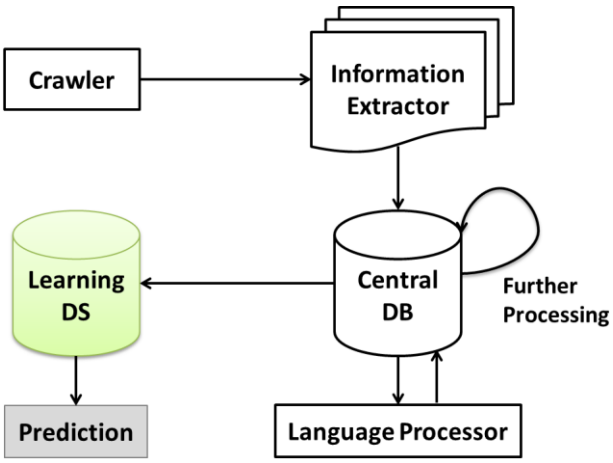

Fig. 2: System Diagram

(vi) Crawler: We used *irobotsoft* to crawl documents from the internet effectively.

(vii) Information Extractor:

– Hypertext Query Language (HTQL): We used HTQL to retrieve specific formatted data from the crawled document

– Regular Expression: RE was used to find formatted data in addition to HTQL

(viii) Language Processor

– Natural Language Toolkit (NLTK): We had to use Python NLTK to chunk the Bengali text into words and to tag the parts of speech so that we could take significant words. We have considered the most frequent words which are either verb or adjectives.

– Custom Translator: Some custom SQL query was built to translate the Bengali numbers into English to ease the further processing.

(ix) Storage: The central data set is in RDBMS (MySQL) which is easily exportable to CSV File to ease the analyzer software's.

(x) Further Processing: We had to develop some custom algorithms using PHP and Java to create the attributes with binary indicator (e.g.: Bag of Words, Weekday Features)

V. Results and Discussion

A. Data Statistics:

We have collected 11668 documents from different Bengali blogs among them 2589 documents (appeared within 1st December to 22nd December) are processed for demonstration purpose. These 2589 documents received a total 17024 feedbacks. We crawled documents within the period from October 2014 and December 2014.

B. Attribute Information:

Total 218+ Attributes were created to support the idea. These are as follows:

- Number of feedbacks before the base time (Extensive)

- Number of feedbacks after the document appears (Extensive)

- Binary indicator for Bag of words (200)

- Binary indicator for weekdays (7)

- Post specific attributes (6)

- Blogger specific attributes (5)

C. Sample Dataset snapshots:

AUTO_PID	BBURL BlogBaseURL	BS Blog Source	BVC Blog Visit Count	BPC Blogger Post Count	BCC Blogger Comment Count
2	somewhereinblog.net	mreedul1986/30000479	15566	74	1345
3	somewhereinblog.net	habiburjewel/30000895	17753	189	196
4	somewhereinblog.net	meshkatmahmud/30000894	2210	33	108
5	somewhereinblog.net	pothik01/30000893	1089	11	19
6	somewhereinblog.net	wellcomejuwel/30000892	16999	126	84
7	somewhereinblog.net	ishaqomar71/30000889	2931	37	61
8	somewhereinblog.net	jasim7324/30000888	170	11	23
9	somewhereinblog.net	shapno2003/30000887	5004	48	198

Fig. 3: Dataset of Blog Posts (Part-1)

BRCC Blogger Received Comment Count	PDT Post Date Time	PRC Post Read Count	PLC Post Like Count	PFC Post Favourite Count	PCL Post Content Length	PCLC Post Content Links Count	PCIC Post Content Image Count	PCYTVC Post Content Youtube Video Count
1596	2014-12-20 01:30:00	934	6	15	10741	11	73	0
532	2014-12-22 01:17:00	30	0	0	1606	0	0	0
88	2014-12-22 01:15:00	19	0	0	997	0	0	0
37	2014-12-22 01:09:00	73	0	2	3975	1	2	0
256	2014-12-22 00:46:00	16	0	0	1965	0	0	0
69	2014-12-22 00:39:00	36	0	0	3551	0	0	0
38	2014-12-22 00:29:00	4	0	0	458	1	0	0
199	2014-12-22 00:05:00	29	0	0	474	0	0	0

Fig. 4: Dataset of Blog Posts (Part-2)

AUTO_CID	CDT Comment Date Time	CommentText	PID
1	2014-12-14 10:51:00	. তিনলু বলেছেন: অনেক ভালো . ২	928
2	2014-12-14 11:28:00	. সুমন কর বলেছেন: ভাল লাগলো।	928
3	2014-12-14 10:50:00	. তিনলু বলেছেন: লিখে যান . ১৫ , ২০১৪ ৭:৫৮. আহমেদ জী...	929
4	2014-12-14 11:24:00	. এহসান সাবির বলেছেন: ভালো পোস্ট। ১৫ , ২০১৪ ৮:০০....	929
5	2014-12-14 11:30:00	. সুমন কর বলেছেন: সময় নিয়ে পড়ার মতো পোস্ট। কেমন আ...	929
6	2014-12-15 09:32:00	. তুষার কাব্য বলেছেন: বেশ গুরুগম্ভীর পোস্ট... জ্ঞা...	929
7	2014-12-15 03:04:00	. খাটাস বলেছেন: বুঝলাম আমার বোধ শক্তি তত টা তীব্র ...	929
8	2014-12-15 04:53:00	. সোহানী বলেছেন: বিশ্লেষণ এ ভালো লাগা ....ও পোস্টে...	929
9	2014-12-15 08:02:00	. মাঈনউদ্দিন মইনুল বলেছেন: বিশাল বিষয়.... সুন্দর আ...	929
10	2014-12-15 08:41:00	. নেক্সাস বলেছেন: সুন্দর পোষ্ট। . ১৬ , ২০১৪ ১:৪৭. ...	929

Fig. 5: Dataset of Posts Comments

CommentsCount	PDT Post Date Time	CDT Comment Date Time
78	2014-12-14 18:20:00	2014-12-15 00:11:00
54	2014-12-14 11:16:00	2014-12-15 00:02:00
39	2014-12-14 14:04:00	2014-12-15 00:07:00
39	2014-12-14 14:04:00	2014-12-15 00:07:00
28	2014-12-15 11:34:00	2014-12-15 00:10:00
26	2014-12-15 08:24:00	2014-12-15 08:30:00
26	2014-12-15 11:08:00	2014-12-15 00:31:00
24	2014-12-14 01:20:00	2014-12-15 00:12:00
21	2014-12-14 10:41:00	2014-12-15 03:04:00
21	2014-12-14 19:27:00	2014-12-15 02:06:00

Fig. 6: Dataset of Comments Count within H hours after the blog post

D. Discussion:

The prepared dataset is in such an extensive format so that it can facilitate any sort of data (see Fig. 7) needed to feed the machine learning algorithms. Then the generated set can be exploded into test set and training set according to the need of the designated algorithm. Therefore supervised learning can be applied to any extent of analysis (in our case feedback prediction).

Input DateTime:
2014-12-05 10:51:00
Input Post Hour:
2
Submit

Blog Url	Blog Source	Blog Visit Count	Blogger Post Count	Post ID	Post Date Time	Post Week Day	Post Comments Count
somewhereinblog.net	tawhidhappy/29997357	4459	154	2114	2014-12-05 09:26:00	Friday	2
somewhereinblog.net	plnk/29997363	7713	22	2111	2014-12-05 10:38:00	Friday	2
somewhereinblog.net	Milton787/29997461	18071	169	2039	2014-12-05 10:14:00	Friday	2
somewhereinblog.net	rksiddique/29997361	28606	184	2112	2014-12-05 10:17:00	Friday	1
somewhereinblog.net	Nisorgohappy/29997355	29934	1148	2115	2014-12-05 09:16:00	Friday	1
somewhereinblog.net	kasperAumee/29997446	42	4	2053	2014-12-05 08:59:00	Friday	1
somewhereinblog.net	faniloy/29997465	6865	54	2036	2014-12-05 10:23:00	Friday	1

Fig. 7: Generated dataset based on specific query

E. The Proposed Model:

We strained to define the problem above through machine learning, especially through regression models. We aim to predict the number of responses which a blog-entry might receive during the next H hours (e.g. H=24 hours).

Based on the collected data, additionally, we propose the following model to support our idea:

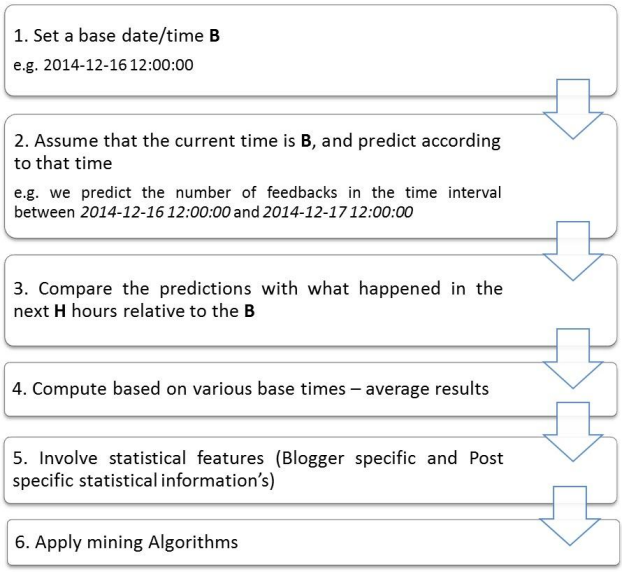

Fig. 8: The prediction model

We select a base date/time B from the past assuming that the selected one is the present date/time. The advantage is, we actually know the number of feedbacks received by the blog entries during the next H hours. For these particular cases we know the value of the target. We considered only the blog posts which were published within one, two, or three days earlier relative to B, as the older entries usually do not get a significant number of feedbacks.

Similarly, we select a time interval in the past and select different B, hence compute the target value and feed the result to the regressor which tends us preparing the training set. Again we select another time interval and calculate the true value for different B where the prediction model is unaware of the actual values. Then, we compare the true and predicted value which allows us to evaluate the prediction model quantitatively.

Additionally, we enhance the prediction model by feeding Blogger and Post specific Meta information's as mentioned in Section III-D. We train the regresor with the blog documents appeared in the past in several time intervals which will lead us to predict the number of feedback received by the blog documents of the present.

VI. Conclusion

The use of social media is growing at an astronomical rate. Over the last decade, the impact of social media grew unbelievably. In this work we have extracted the basic attributes required to apply mining algorithms. In particular, we aimed to predict the number of feedbacks that blog documents receive. Supervised learning can be performed where applicable algorithms are: Regression trees, Neural networks, K-NN, Linear Regression etc.

References

[1] Buza, K. (2014). Feedback prediction for blogs. In Data Analysis, Machine Learning and Knowledge Discovery (pp. 145-152). Springer International Publishing.
[2] Marinho, L. B., Buza, K., & Schmidt-Thieme, L. (2008). Folksonomy-based collabulary learning (pp. 261-276). Springer Berlin Heidelberg.
[3] Mishne, G. (2007, March). Using Blog Properties to Improve Retrieval. In ICWSM.
[4] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.
[5] Pinto JPGS (2008) Detection Methods for Blog Trends. Report of Dissertation Master in Informatics and Computing Engineering, Faculdade de Engenharia da Universidade do Porto
[6] Reuter, T., Cimiano, P., Drumond, L., Buza, K., & Schmidt-Thieme, L. (2011, May). Scalable Event-Based Clustering of Social Media Via Record Linkage Techniques. In ICWSM.
[7] Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining (Vol. 1). Boston: Pearson Addison Wesley.
[8] Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
[9] Yano, T., & Smith, N. A. (2010, May). What's Worthy of Comment? Content and Comment Volume in Political Blogs. In ICWSM.
[10] Complete History of Social Media: Then And Now. http://smallbiztrends.com. Retrieved January 12, 2015, from http://smallbiztrends.com/2013/05/the-complete-history-of-social-media-infographic.html
[11] Social media. Wikipedia. Retrieved January 22, 2015, from http://en.wikipedia.org/wiki/Social_media#cite_note-1
[12] Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. Business horizons, 54(3), 241-251.
[13] Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008, February). Finding high-quality content in social media. In Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 183-194). ACM.
[14] "State of the media: The social media report 2012". Featured Insights, Global, Media + Entertainment. Nielsen. Retrieved 15 November 2014.
[15] Tang, Q., Gu, B., & Whinston, A. B. (2012). Content contribution for revenue sharing and reputation in social media: A dynamic structural model. Journal of Management Information Systems, 29(2), 41-76.
[16] "The U.S. Digital Consumer Report". Featured Insights, Global, Media + Entertainment. Nielsen. Retrieved 25 November 2014.
[17] Social media mining. Wikipedia. Retrieved January 22, 2015, from http://en.wikipedia.org/wiki/Social_media_mining
[18] Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social media mining: an introduction. Cambridge University Press.
[19] Bastos, M. T. (2014). Shares, pins, and tweets: News readership from daily papers to social media. Journalism Studies, (ahead-of-print), 1-21.
[20] Runge, K. K., Yeo, S. K., Cacciatore, M., Scheufele, D. A., Brossard, D., Xenos, M., ... & Su, L. Y. F. (2013). Tweeting nano: How public discourses about nanotechnology develop in social media environments. Journal of nanoparticle research, 15(1), 1-11.
[21] Gerhards, J., & Schäfer, M. S. (2010). Is the internet a better public sphere? Comparing old and new media in the US and Germany. New media & society.
[22] "The High-Level Business Impact of Social Media". smallbusiness.chron.com. Retrieved January 22, 2015, from http://smallbusiness.chron.com/highlevel-business-impact-social-media-38816.html
[23] "The impact of Social Networks on developing countries". smallbusiness.chron.com. Retrieved January 22, 2015, from http://www.diplointernetgovernance.org/profiles/blogs/the-impact-of-social-networks-1

Shovra Das completed his BSc in Computer Science & Engineering and MSc in Information and Database Management from Amarican International University-Bangladesh. His research interest is mainly focused on Data Mining, Web Mining, Image processing, Software Engineering and Data & knowledge management. Currently he is working as a Lecturer in American International University-Bangladesh (AIUB) in the department of Computer Science.

Md. Hasibul Hasan received his B.Sc. degree in Computer Science and Software Engineering from American International University Bangladesh (AIUB), Dhaka, Bangladesh, in 2013, and the MS degree in Computer Science from American International University Bangladesh (AIUB), Dhaka, Bangladesh, in 2015. In 2013, he joined the Department of Computer Science and Engineering at Bangladesh University, as a Lecturer, and from 2015 he has been with the Department of Computer Science at American International University Bangladesh (AIUB) as a lecturer. His current research interests include Data mining, Data Security, Natural language Processing, Sensor networking.

Md. Shamsur Rahim obtained his B.Sc in Computer Science & Software Engineering and MSc. in Computer Science from American International University-Bangladesh, Dhaka, Bangladesh. His research interest includes Cyber Physical Systems, Big Data, Human Computer Interaction and Software Engineering. Currently, He is working as a Lecturer at the department of Computer Science in American International University-Bangladesh.

Mohammod Hafizur Rahman is working as Assistant Professor in American International University-Bangladesh (AIUB) under the department of Computer Science. He completed his BSc and MSc also from Amarican International University-Bangladesh. His research areas are Internet of things (IoT), Wireless Sensor Network, Data Warehouse & Mining, Text Mining, Embedded Technologies, and Software Engineering etc.