2020 23rd International Conference on Computer and Information Technology (ICCIT), 19-21 December, 2020

Bangla News Trend Observation using LDA Based Topic Modeling

Kazi Masudul Alam
DGTED Lab, Computer Science and Engineering
Khulna University
Khulna, Bangladesh
kazi@cseku.ac.bd

Md. Tanvir Hossain Hemel
DGTED Lab, Computer Science and Engineering
Khulna University
Khulna, Bangladesh
hemel1638@cseku.ac.bd

S.M. Muhaiminul Islam
DGTED Lab, Computer Science and Engineering
Khulna University
Khulna, Bangladesh
eshan1640@cseku.ac.bd

Aysha Akther*
DGTED Lab, Computer Science and Engineering
Khulna University
Khulna, Bangladesh
aysha@cseku.ac.bd

2020 23rd International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-2244-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCIT51783.2020.9392719

***Abstract*—Topic Modelling is an essential field of natural language processing (NLP) that can be considered as a type of statistical model for extracting the abstract topics that have occurred in a collection of documents. Bangla is among the most popular and used languages around the world and nowadays innumerable Bangla texts are generated through digital and social media. So the significance of extracting knowledge from these data is invaluable for various sectors. However, the number of works in this field is inadequate because of the lack of proper datasets, tools, and applications. Therefore, preparing a convenient dataset in Bangla can be a great help for topic modeling as well as for other NLP related research. In this paper, we have addressed some of those complications by creating a proper dataset. Also, we have demonstrated a method of observing the Bangla media trend by applying Latent Dirichlet Allocation (LDA) on newspaper articles. The result of our experiment suggests that the proposed method can be an admissible way of utilizing news media data to observe media trends overtime properly.**

***Index Terms*—Bangla News, Bangla Corpus, N-Grams, Topic Modelling, LDA**

I. Introduction

This is an era of an increasing number of real-time multimedia data generated in all types of digital and analog sources all around us. In order to retrieve the true essence of knowledge from this enormous data, we have to build important resources such as corpus to understand the words and their relationships. Then, we can use these tons of data to extract interesting patterns such as trending topics from media. Topic modeling tools and techniques can be useful to conduct other NLP operations. Topic modeling finds the collection of topics or recurring patterns that represents the main essence of a document from a pile of content [1]. While the English language has a rich set of tools and techniques for language processing and research in this field has touched diverse dimensions [2] [3], Bangla is way behind in terms of research and resources building.

Bangla is used every day by more than 250 million people in the world. It is the primary language in Bangladesh and secondary language in India [4]. So this language has created its own place on the news media or other knowledge-sharing platforms. It can be said that Bangla will join the contest of the action ground of the NLP in near future because of having a huge collection of Bangla newspaper, Bangla Wikipedia, Bangla literature, Bangla news portals, blogs, eBooks, web pages, search engines, etc. Among lots of discussions, various topics become trendy at a certain time. Every day a new topic can be popular as that topic is discussed by most of the people in news media. From tons of news around us, everybody wants to know what are the most important and frequently discussed topics around us at any specific time. But, the lack of available and standard corpus in Bangla restricts language researchers to develop important language processing applications.

The word "corpus" originated from Latin, which means "body/mass". In the case of NLP, a corpus is a large structured set of processed text that is electronically stored [5]. Text in the corpus may include written or spoken language and do not belong to a single subject field. Also, a specialized corpus can be designed to focus on specific research goals in mind. Examples of specialized corpus include Cambridge and Nottingham Corpus of Discourse in English (*CANCODE*[1]) and Michigan Corpus of Academic Spoken English (*MICASE*[2]). Except English, some of the European languages, as well as Japanese, have popular corpus building activities. Unfortunately Bangla has very few reusable works.

[1]https://www.nottingham.ac.uk/research/groups/cral/projects/cancode.asp
[2]https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple

978-1-6654-2244-4/20/$31.00 © 2020 IEEE

Working with Bangla language is critical because of its complex grammatical structure [6]. Besides the scarcity of datasets, and tools discourage researchers to walk on this ground. A text corpus is the basis of linguistic phenomena study such as spelling checking, morphological structure, word sense disambiguation, language evolve over time, etc. In this paper, we first present a specialized Bangla news corpus of 70,000 news articles which is a part of an ongoing data collection effort of 1 million news articles. Our developed corpus comprised of over 14.8 million words of different categories. We present several N-grams of our developed news corpus for further analysis. Secondly, we have applied the topic modeling algorithm, LDA, to find interesting patterns from the dataset. Topic modeling is one of the effective methods to find hidden structures in the collection of documents [7]. We apply it to find current trends happening in the news world. As topic modeling gives us the topics that are mostly repeated, we cannot easily understand which topic stands for which matter. To resolve this issue, we apply labeling to the LDA topics.

Topic modeling divides the whole corpus into segments. Algorithms such as LSI, pLSI, lda2vec, and LDA are applied for topic modeling. In case of LSA, the words that have similar meaning appears in same text [7]. In case of pLSI, it models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of topics [8]. Again LDA is a generalization of the pLSI model, which is equivalent to LDA under a uniform Dirichlet prior distribution [9]. LDA is a generative statistical model. LDA assumes to have a sparse Dirichlet prior which has the intuition that the document cover only a few topics and that topic have only a few words which can be called keywords of that topic.

The rest of the paper is organized as Section II discusses the background and state of the art on Bangla corpus building and LDA applications. Section III gives details about the Bangla corpus built for this research, Section IV discusses the media trend analysis procedure, Section V further describes the experimental setup and observations. Finally, Section VI concludes the paper with possible future works.

II. Related Work

Though few notable works have been conducted on the Bangla corpus most of the studies are limited to analyzing language phenomena. Studies on other NLP applications such as knowledge engineering on Bangla corpus are still very limited. In [10] authors studied text categorization and classification based on keyword extraction. Authors of [11] proposed a Bangla question-answering system, where the authors used Cosine Similarity, Jaccard Similarity, and Naïve Bayes algorithms to obtain the relationship between the QA. In [12] authors applied a modified Bangla VADER for identification of Bangla text sentiment polarity. Authors of [6] attempt to develop a balanced corpus in Bangla which under development. Authors of [13] focuses on font and language detection and Unicode conversion of Bangla text for automatic Bangla corpus creation. In [14] authors developed a corpus from a particular newspaper's articles of a particular year.

Authors of [15] use LDA on a dataset of 400 news articles to identify labeled topics and sub-topics. In [16], topic modeling was applied to Bangla, where topics are extracted using Doc2Vec approach and compared with the LDA method. Authors of [17] show an approach to summarize information of news portal, blogs, books, etc. by using the extractive method. Their summarization extracts the basic ideas of the topic by word and sentence frequency and showed whether the topic is relevant or not. Authors of [18] use LDA to find topics and sentiments on news corpus collected from www.prothom-alo.com newspaper. Also, they have classified the news using similarity. Their dataset also contains Bangla comments from Facebook.

In this article, our objective is to extract the ongoing media trends in a time period by using topic modeling. In the field of politics, advertising, entertainment world, media trend analysis is of immense interest. We first describe a specialized Bangla news corpus of 70,000 news articles. At present, our developed corpus comprised of over 14.8 million words of different categories. We present the unigrams, bi-grams, and tri-grams from our developed news corpus for further analysis. Secondly, we have applied the topic modeling algorithm-LDA to find interesting patterns from the dataset. Since LDA doesn't label the extracted topics, we developed a labeling method to easily understand the topic of interest.

III. Proposed Bangla News Corpus

A properly balanced corpus in Bangla is yet to be developed. It is one of the key difficulties that any researchers face while working with Bangla language-related mining problems. In order to find the media trend ongoing in news articles, we first built a Bangla corpus with 70K articles. Our goal is to develop a rich corpus[3] of a million articles. At first a dataset of 40K Bangla news articles has been collected from *kaggle*[4]. As this dataset was not balanced in all the categories, additional 30K Bangla news articles were crawled from different Bangla newspapers to improve the impartiality of the dataset. We only collected news articles with UTF-8 encoding as they don't need any encoding conversion.

A. Data Preprocessing

Bangla is grammatically very complex language. So it is a big challenge to preprocess the documents without deprecating the hidden meaningful topics. Any document is consist of a large number of words from where we just

[3]Bangla News Corpus, https://github.com/dgted/BanglaNewsCorpus
[4]https://www.kaggle.com/zshujon/40k-bangla-newspaper-article

need smaller set of words that are the key words of that document.

1) Tokenization and Punctuation removal: In the tokenization phase, the texts are split into sentences, as well as the sentences split into words. As in Bangla language, words are normally separated by white spaces or by punctuation marks so tokenization is comparatively easy in Bangla than in some other languages where white space does not resemble word boundary. Punctuation marks (e.g. ,!.?"':; etc.) are removed from the corpus. After tokenization and punctuation removal we got 1,48,67,270 word tokens and 4,60,231 unique word tokens. We didn't remove numbers because some numbers have great significance in the sense of topics such as ১৯৭১, ৬ দফা etc.

2) Stop words removal: The words those are less significant in a document but occur frequently (e.g. অবশ্য, অথবা, অনেকে, এবং, অথচ, অন্য, আছে, ও, অন্তত, আবারো etc.) can be considered as stop words. We have created a dictionary having 415 such stop words and removed those words from the corpus in one version. After stop word removal the number of unique tokens became 4,59,845 in one version of the corpus. Here is an example of tokens from a sentence before and after removing stop words. Before it was ব্যবসায়ী, ও, আড়তদারেরা, বলছেন, গুজবের, কারনে, হঠাৎ, পেয়াজের, দাম, আবারও, বেড়েছে . After the removal ব্যবসায়ী, আড়তদারেরা, গুজবের, পেয়াজের, দাম, বেড়েছে .

3) Stemming: Stemming is the process of reducing individual words to their root form. We have used an open-source Bangla Stemmer [5]. which splits suffix out from Bangla word token. Here is an example of tokens of a sentence before 'ব্যবসায়ী', 'আড়তদারেরা', 'গুজবের', 'পেয়াজের', 'দাম', 'বেড়েছে' and after stemming 'ব্যবসায়ী', 'আড়তদার', 'গুজব', 'পেয়াজ', 'দাম', 'বেড়' .

B. Developing N-gram profiles

In this step, unigrams, bigrams, and trigrams have been extracted from both the stemmed and non-stemmed words. An n-gram is a contiguous sequence of n items from a given sample of text. We extracted those n-grams from the preprocessed data. We have created three dictionaries of unigrams, bigrams, and trigrams. In this study of media trend analysis of news corpus, the dictionary of unigrams is our Bag-Of-Words for topic modeling. Before stemming we got 460231 unigrams, 11657 bigrams, and 961 trigrams. After stemming we got 321182 unigrams, 10349 bigrams, and 916 trigrams. A list of unigrams, bigrams, and trigrams from our corpus along with their total count in the training corpus is given in Table I.

[5]https://github.com/banglakit/bengali-stemmer/tree/dev/bengali$_stemmer$

Table I: Unigram, Bigram and Trigram list

Unigram, Bigram, Trigram	Count
প্রেসিডেন্ট, আওয়ামী লীগ, হাজার কোটি টাকা	75839, 11527, 1667
রাজনীতি, নির্বাচন কমিশন, প্রধানমন্ত্রী শেখ হাসিনা	37224, 4301, 1593
ডিএমপি, বাংলাদেশ ব্যাংক, আইনশৃঙ্খলা রক্ষাকারী বাহিনী	36415, 3876, 689
সীমান্ত, প্রধানমন্ত্রী শেখ, প্রধান নির্বাহী কর্মকর্তা	34323, 2606, 470
অনুমোদিত, ওয়েস্ট ইন্ডিজ, উপজেলা স্বাস্থ্য কমপ্লেক্স	31921, 2370, 456

C. Developing TF-IDF profile

TF-IDF (Term frequency-inverse document frequency), is a weighting scheme often used in information retrieval and text mining. It is a numerical statistic that is intended to reflect how important a word is to a document in a corpus. The TF–IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word. The TF-IDF value of all the words has been calculated to have an insight into the total corpus.

D. Developing Cluster of the Corpus

Clustering is the grouping of particular sets of data based on their characteristics, according to their similarities. K-means clustering is one of the most popular clustering algorithms in machine learning. We used K-means to find the tokens of the clusters in the corpus presented in Table II.

Table II: Cluster excerpt of the Bangla corpus

Cluster 0	অভিনয়শিল্পী, নাটক, আত্মীয়, করছি, দাফন, কবর
Cluster 1	তিনি, অভিনেতা, বিখ্যাত, যোগ, কদিন, আরেকজন
Cluster 2	বেসাতি, একবচন, বহুবচন, অভিনয়, মিথ্যাশ্রয়ী, কল্পনাশ্রয়ী
Cluster 3	মিথ্যা, কথা, শুন, খারাপ, যাক, বোঝা, আসল, সত্যি, ব্যাপার

IV. Proposed Media Trend Observation Model

Our first step toward the topic modeling was to build the Bangla news article corpus. Detailed steps are described in Figure 1. After extracting the topics, proper labels have been given to the topics by training them with some true topic (some news articles with known category). Then labeling method has been implemented to label the topics extracted from the LDA model to see the change in the recent media trend. Table III represents the notations and meaning of the notations we have used throughout the rest of the article to describe the media trend extraction system.

A. Determine optimum number of topics

Topic modeling methods give us as many topics as we want. But to get the most accurate topics from the datasets, an optimum number of topics are extracted. To get the optimum number of topics, topic coherence is a

Table III: Table of Notations

Notations	Meanings
T	Set of Topics
L	Set of Labels
M	Matrix indicating relevance of each document with Topics from T
N	Matrix indicating relevance of each Topic with Labels from L
X	Normalized Matrix represents Per Topic Label distribution in %
W	Set of collected news in weekly basis
Y	Matrix indicating relevance of news from each week with Topics
Z	Matrix indicating relevance of news from each week with Labels

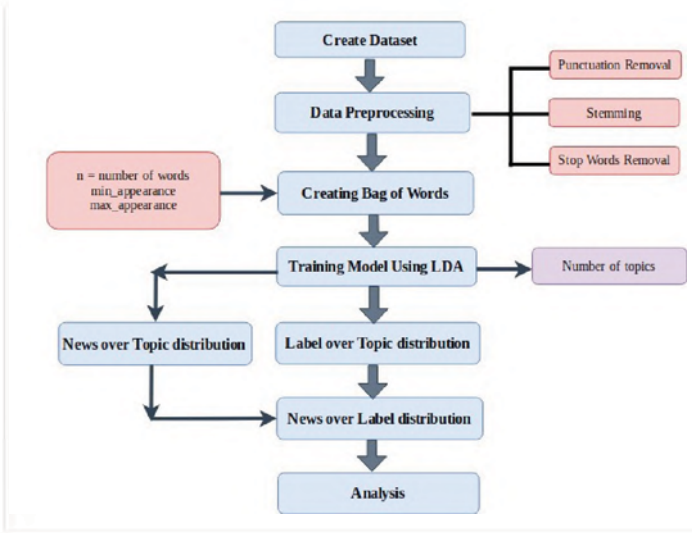

Figure 1: Media trend analysis working procedure

popular method. For our dataset, the highest coherence value occurred for 6 topics. So here we have extracted six topics from using LDA.

B. Training model using LDA

We have set the number of topics 6 which has appeared to be the optimum number for this dataset during coherence measurements. Also we have trained our topic model M using LDA on *Training Dataset* for 6 topics as $T = \{t_0, t_1, t_2, t_3, t_4, t_5\}$

C. Assigning Label to Each Topic

For each $l_x \in L$ we had a set of documents $D_x = \{d_{x1}, d_{x2}...d_{xn}\}$ which are the representative of l_x. We've computed relevance of each document in D_x with model as a matrix M_x. In our dataset, we have a label set L with 6 labels as $L = \{$*National*, *Sports*, *International*, *Technology*, *Economics* and *Others*$\}$

$$M_x = p(t_i|d_{xj}) \tag{1}$$

Any cell [t,d] of M_x represents correspondence of document d_x with topic t ,where i and j are index of elements of T and D. Now for each $t \in T$ and each $l_x \in L$ we calculated matrix N_{T*L} as

$$P(t|l_x) = \frac{\sum_{n=0}^{n=j} p(t|d_{xn})}{j} \tag{2}$$

Where j is the number of documents in D_x, T is topic set and L is label set. Here each cell from N represents relevance of a particular label from L with a particular topic from T.

We have per label topic distribution as N_{T*L} which is shown in Table IV. We have calculated another Matrix X_{T*L} from Matrix N_{T*L} where value of each cell is normalized and each row represents per topic label distribution in percentage. For any r, r_{t_r*L} from N_{T*L} represents any *row*. We calculated coefficient to be multiplied with elements of row for each row k_r as Equation 3

$$k_r = \frac{100}{\sum_{n=0}^{size(T)} N_{T_r*L_n}} \tag{3}$$

For Example, From Table IV, k_1 can be calculated as Equation 3.

$$k_1 = \frac{100}{0.19 + 88.88 + 4.07 + 1.02 + 0.10 + 9.08} = 0.97 \tag{4}$$

By multiplying k_r with value of each cell from the row_k normalized value of corresponding cell achieved. Thus to get Matrix X_{T*L}, each value of row_1 is multiplied by 0.97. Matrix X_{T*L} is represented in Table V where each topic can be viewed as a formation of different types of categories. For example, in Topic 3 Label-Topic Correspondence value for National category is very high but other categories also keep their participation on that topic.

Table IV: Topic-Label Correspondence

	National (জাতীয়) %	Sports (খেলা) %	International (আন্তর্জাতিক) %	Technology (তথ্যপ্রযুক্তি) %	Economy (অর্থনীতি) %	Others (অন্যান্য) %
Topic 0	5.17	1.39	9.15	5.59	2.34	24.79
Topic 1	0.19	88.88	4.07	1.02	0.10	9.08
Topic 2	3.00	2.04	73.65	2.22	0.34	18.45
Topic 3	72.43	3.76	5.24	3.50	12.98	13.51
Topic 4	2.60	1.51	7.15	17.05	81.02	14.34
Topic 5	16.61	2.42	0.74	70.62	3.22	19.93

Table V: Row normalized topic label correspondence

	National (জাতীয়)	Sports (খেলা)	International (আন্তর্জাতিক)	Technology (তথ্যপ্রযুক্তি)	Economy (অর্থনীতি)	Others (অন্যান্য)
Topic 0 %	10.68	2.87	18.89	11.54	4.83	51.19
Topic 1 %	0.18	86.04	3.94	0.99	0.10	8.79
Topic 2 %	3.01	2.05	73.87	2.23	0.34	18.51
Topic 3 %	65.01	3.37	4.70	3.14	11.65	12.13
Topic 4 %	2.10	1.22	5.78	13.79	65.51	11.60
Topic 5 %	14.63	2.13	0.65	62.20	2.84	17.55

V. Experimental Result and Discussions

Media trend testing dataset is collected from different weekly news as a separate collection. For every collection, percentage of classified news in different topic is calculated. For the test set $W = \{w_0, w_1, w_2...w_n\}$ a Matrix $Y_{T\times W}$ of $T \times W$ dimension using Equation 1 and Equation 2 is calculated. This represents ratio of news in each week that are classified in different topics. Table VI represents weekly news distribution over topics in percentage where

each week contains forty news articles. Each $w \in W$ is written as $w_n = \{d_{n0}, d_{n1}, d_{n2}...d_{n39}\}$

Table VI: Week news distribution over topic in (%)

	Week 1	Week 2	Week 3
Topic 0	12.05	8.15	16.13
Topic 1	55.97	4.38	12.32
Topic 2	7.95	5.10	18.25
Topic 3	6.90	7.24	24.18
Topic 4	14.02	71.59	20.93
Topic 5	3.11	3.54	8.19

Using Equation 1 and Equation 2 we calculate six topics on already trained model. In Table VI the percentage of participation of different topics in each week can be seen e.g. in week 1, 55.97% news has discussed the matter(s) which are in Topic 1. In week 2 that number dropped down to 4.38%. But still those topics are mixture of two or more categories. To get weekly data distribution over labels rather than topic uniquely a new Matrix Z is constructed from weekly news distribution over topic Matrix(Y) and Label-Topic distribution Matrix(X). Matrix Z representing the distribution of weekly news over different label.

Each values $z \in Z$ are calculated as Equation 5

$$z = Z_{t,l} = \sum_{t=0}^{size(T)} Y_{tw}^{T} * X_{tl}\% \tag{5}$$

where $Y_{T\times W}^{T}$, is transpose of Matrix $Y_{T\times W}$.

From Table V we got label distribution over the topics. Here each of the labels were distributed among all the topics more or less. From this table we can see that *Topic 3* is a combination of 65.01% *National*, 3.37% of *Sports*, 4.70% of *International*, 3.14% of *Technology*, 11.65% of *Economy* and 12.13% of *Others*. For this context of application more then 50% participation of any single category has been considered as the label of that particular topic. Thus the labelling result is Topic 0 Others (অন্যান্য) , Topic 1 Sports (খেলা) , Topic 2 International (আন্তর্জাতিক) , Topic 3 National (জাতীয়) , Topic 4 Economy (অর্থনীতি) , Topic 5 Technology (তথ্যপ্রযুক্তি) .

Every segment represents a specific topic which is a mixture of multiple categories because every topic is a collection of different categories. For example we have 55.97% news at *week 1* from *Topic 1*. Again *Topic 1* is composed of 86.01% of sports. Also *Sports* has its involvement at different weight in other topics too. So to get exactly how much *Sports* has been discussed over the whole week, exact portion of each category from different topic has been calculated in Figure 2. It can be seen as the exact ratio of the trending media topic in a specific week. This is the original portion of discussion which had been done on that particular time period according to our dataset. The reason behind the highest score of the sports event during this period we found *ICC World Cup 2019* was happening at that time. So most of the newspapers were covering the sports news on that time.

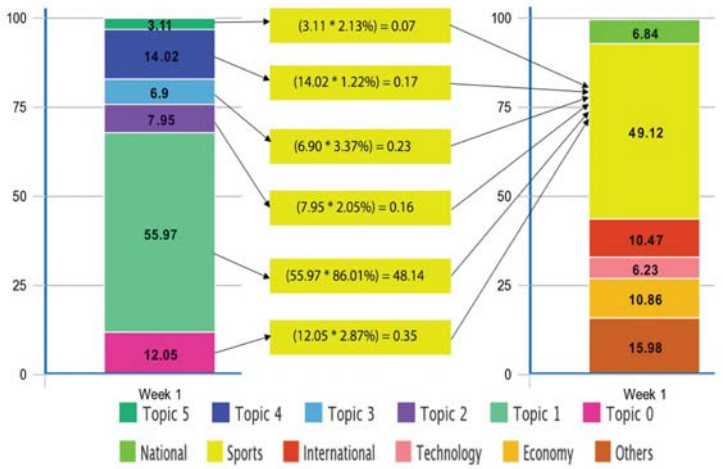

Figure 2: Label topic correspondence value distribution for a week

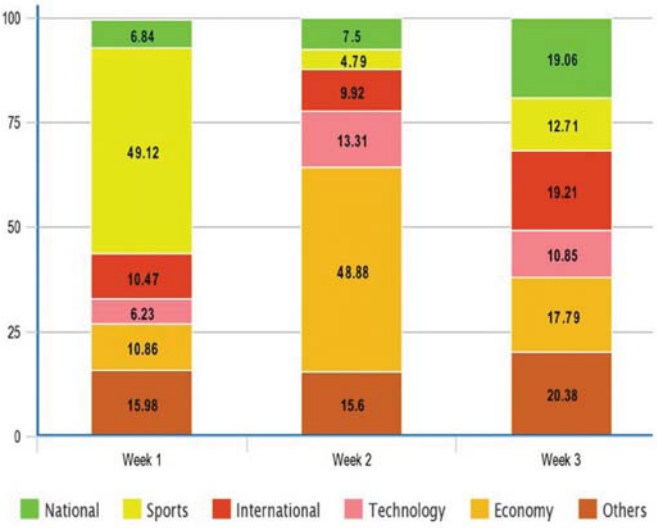

Figure 3: Label topic correspondence value distribution for all the three weeks

In the same way, exact ratio of trending media topic in all weeks are calculated as shown in Figure 3. From Figure 2, we can see that the portion of sports news are very significant (49.12%) during that period. Other topics such as national, international, technology, economy, etc. were also discussed in this period. At *week 2* we see strong presence of economy labeled news more than the *week 1*. That might because of price hike in the market of *Onion* during that time period was prevalent all newspapers. Again *week 3* does not tend to show any particular label. It can be said that not very significant event was occurring during this period.

At this point, we have conducted another experiment to examine the impact of the size of the training dataset over extracted trending media topic. For this experiment we have trained our system on six different training dataset size. We decreased training dataset size gradually by keeping the test data same. The resultant score for *Sports* of $week_1$ of changing training dataset size is shown in Figure 4. As expected, in FIgure 4 score for the topic is decreasing with decrease in training data size. When only 1000 news articles were used to train the model, score dropped to 19.14 from score 49.12 when the model was trained with 70,000 news articles. This result proves that though test data was same for both case, the system failed to extract substantial amount of information due to lack of training.

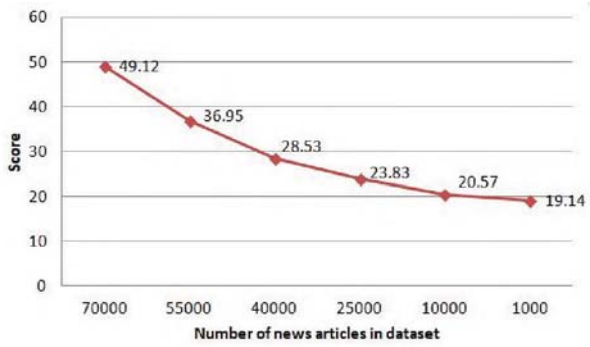

Figure 4: Change of accuracy depending on corpus data for Sports

VI. Conclusion and Future Works

Bangla language is lacking in enough NLP research and resources compared to other contemporary languages. In this article, we observe ongoing media trends in Bangla news articles in a period of time. In this regard, we built a balanced Bangla news corpus with 70,000 articles composed of 4,60,231 unique tokens and extracted different data features from the corpus. Later we have discussed a method to evaluate the model generated by topic modeling algorithm-LDA for taking out the ultimate knowledge. Also, we have observed other unseen texts with a more human-readable form by assigning user-defined Labels. Our experiments prove that corpus and labeled LDA is a good model for Bangla news topic modeling. Possible future works can be the use of lemmatization during prepossessing to find better root words. Lexical priors can be added as seed words to train the model and it will help to control the labeling more specifically and accurately using guided LDA. Integrated chatbot for Bangla can be a viable application on the model. Human to machine interaction using Bangla can be enriched. Also, adding more content to the corpus is an ongoing future work.

References

[1] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, vol. 6, no. 1, 2015.

[2] K. M. Alam, S. Hardy, A. Akther, and A. El Saddik, "Sms text based affective haptic application," in *Proceedings of Virtual Reality International Conference (VIRC 2011), April, Laval, France*, 2011.

[3] K. M. Alam, M. A. Rahman, A. El Saddik, and W. Gueaieb, "Adding emotional tag to augment context-awareness in social network services," in *2011 IEEE International Instrumentation and Measurement Technology Conference*, pp. 1–6, IEEE, 2011.

[4] K. A. Hasan, A. Mondal, and A. Saha, "A context free grammar and its predictive parser for bangla grammar recognition," in *2010 13th International Conference on Computer and Information Technology (ICCIT)*, pp. 87–91, IEEE, 2010.

[5] T. McEnery, *Corpus linguistics*, vol. 978019. Oxford University Press Inc, 2012.

[6] K. M. A. Salam, M. Rahman, and M. M. S. Khan, "Developing the bangladeshi national corpus-a balanced and representative bangla corpus," in *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1–6, IEEE, 2019.

[7] A. Akther, H.-N. Kim, M. Rawashdeh, and A. El Saddik, "Applying latent semantic analysis to tag-based community recommendations," in *Canadian Conference on Artificial Intelligence*, pp. 1–12, Springer, 2012.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[9] M. Girolami and A. Kabán, "On an equivalence between plsi and lda," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 433–434, 2003.

[10] M. Gope and M. Hashem, "Knowledge extraction from bangla documents using nlp: A case study," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–5, IEEE, 2019.

[11] M. Kowsher, M. M. Rahman, S. S. Ahmed, and N. J. Prottasha, "Bangla intelligence question answering system based on mathematics and statistics," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–6, IEEE, 2019.

[12] A. Amin, I. Hossain, A. Akther, and K. M. Alam, "Bengali vader: A sentiment analysis approach using modified vader," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–6, IEEE, 2019.

[13] A. I. Sarkar, D. S. H. Pavel, and M. Khan, "Automatic bangla corpus creation," tech. rep., BRAC University, 2007.

[14] K. M. Majumder and Y. Arafat, "Analysis of and observations from a bangla news corpus," 2006.

[15] S. C and V. S., "Statistical topic modeling for news articles," in *International Journal of Engineering Trends and Technology 31*, pp. 232–239, 2016.

[16] M. Mouhoub and M. Al Helal, "Topic modelling in bangla language: An lda approach to optimize topics and news classification.," *Computer and Information Science*, vol. 11, no. 4, pp. 77–83, 2018.

[17] S. Abujar, M. Hasan, M. Shahin, and S. A. Hossain, "A heuristic approach of text summarization for bengali documentation," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–8, IEEE, 2017.

[18] M. A. Helal, *Topic Modelling and Sentiment Analysis with the Bangla Language: A Deep Learning Approach Combined with the Latent Dirichlet Allocation*. PhD thesis, Faculty of Graduate Studies and Research, University of Regina, 2018.