International Journal of Advanced Science and Technology
Vol.103 (2017), pp.9-22
http://dx.doi.org/10.14257/ijast.2017.103.02

Trending News Analysis from Online Bangla Newspaper

Md. Mahfuzur Rahaman[1] and M. A. Alim Mukul[2]

[1]Lecturer, Dept. of CSE, SUST
[2]Dept. of CSE, SUST
[1]mahfuzsustbd@gmail.com, [2]mukul.sustcse@gmail.com

Abstract

People are always curios about the trend, everyone wants to find out "what's going on?" In the crowd of millions of websites and billions of texts in it, they often get stack. Nowadays newspaper is a media which people read it every day, not only the printed one but also the online version of it. Online newspaper, magazine and blogs are so much active that they provide and update themselves every hour. But what if, a busy person wants to take a look on the web for the recent hit news/Trending News? From this concept, we choose to research how it is possible to get the latest most trending news among all the Bangla online newspapers scattered data and build an automatic model so that it can update itself. Thus, anyone can know what's the most Trending Topic on the news now.

Keywords: *Trend analysis, data mining, n-gram, chi-square test, p-value, twitter*

1. Introduction

Newspaper plays a very vital role in today's world. People at the old age didn't know how to talk or communicate. Then they invented their way of communication. But what's going on at the other side of the world was always unknown to them. Slowly after decades' people invented Mail, Telegram, Telephone etc. In the 20th Century, everything and everybody is connected. If somebody wants to know about anything what is going on with the other side of the world, they can know it within a minute. Television, Newspaper, Social Media, Blog are so much active 24/7. People nowadays don't have to wait for the tomorrows newspaper. The Online newspaper has open a new era. Bangladesh isn't out of the role. Everything is now on the web. Printed newspaper is such a waste of paper, it also creates carbon which slowly enhancing the heat of the environment. The number of online newspaper readers is increasing day by day. There are smartphones in most of the people's hand nowadays. Social media like Facebook, Twitter etc. users are also sharing hit or trending news every day.

But there are lots of online newspaper on the web, everybody is busy. They cannot read all of them. Always people want to know – "What's Trending now?" In Politics, Sports, Fashion, Business, Stock exchange, Music, Literature, Culture, International Politics etc. One of the most popular social media site – 'Twitter' has done a great job. They are always ready to show the most Trending hot tweets and Hashtags among millions of tweets every day. Anybody can know what's even the Trending tweets of their country/region. But in online newspaper what people read every day, there is no website/service which is giving or showing the exact Trending news topic by which anybody can look out without not spending hours on the web.

The widespread and increasing availability of text documents in electronic form increases the importance of using automatic methods to analyze the content of text documents [1]. Our research topic is to find out those Trending Topic over millions of texts around all the online newspaper of Bangladesh. We tried to build a model which will find out the most recent Trending news topic.

ISSN: 2005-4238 IJAST

1.1. Motivation

People always run after or want to know about the trend, it's one of their behavior by born. What's going on outside the world – has always been our major concern after we have born and able to understand. There are many ways to know about what's going on around us, such as – Television, Book, Letter, Conversation, Video etc. In this Technological revolutionary time, most of us have got internet access easily. The access of everything is way much easier now. People use e-mail instead of mail or telegram nowadays. Newspaper has always been a friend to all of us. We keep updated about the world by reading it regularly. But it's also now on the web in way much smarter format.

Newspapers have always been the primary medium of journalists since 1700, with magazines added in the 18th century (which is also the 1700s) radio and television in the 20th century, and the Internet in the 21st century. An online newspaper is the online version of a newspaper, either as a stand-alone publication or as the online version of a printed periodical [2].

There are 49 online daily newspapers, 95 national news agencies and 37 international news agencies in Bangladesh. Trillions of data has been produced and spread every day through it. But is there any significance of it? Off course. Data is more valuable than oil in modern world. Those data give us rich source of information, because newspapers are an actor, arena and archive of public discourses at the same time.

But only a few of newspaper websites have this feature called most Trending, recent and visited news. On the other hand, first Bangla search engine – "Pipilika" has also a great feature that combines all of the newspapers most hit & recent news from all of the online newspaper.

1.2. Concept of Trending Data

A pattern of gradual change in a condition, output, or process, or an average or general tendency of a series of data points to move in a certain direction over time, represented by a line or curve on a graph [3]. Trends in online media are commonly called trending topics. A topic refers to a subject matter of discussion or conversation that is agreed on by a group of people. Such a topic becomes trending when it experiences a sudden spike in user interest or engagement. Therefore, the ability to measure user interest and engagement is necessary.

Among the vast information available on the web, social media streams capture what people currently pay attention to and how they feel about certain topics. Awareness of such trending topics plays a crucial role in many application domains such as economics, health monitoring, journalism, finance, marketing, and social multimedia systems [4].

Trends in factors such as rates of disease and death, as well as behaviors such as smoking are often used by public health professionals to assist in healthcare needs assessments, service planning, and policy development. Examining data over time also makes it possible to predict future frequencies and rates of occurrence.

Studies of time trends may focus on any of the following:

- Patterns of change in an indicator over time
- Comparing one time period to another time period
- Comparing one geographical area or population to another
- Making future projections [5]

We can sense about many trending topic around us, but when it needs a automation system – what would do most of the multinational business organization? We will take a look of a few of them.

Twitter, nowadays are most hit social microblogging website. They analyze every tweet every day and publish them on a corner site of the website. Like this:

Figure 1. Trends of U.S.A. on Twitter

Recently Google has done some great job about finding the trend. U.S.A. election is now is a huge factor for all over the world. Donald Trump and Hilary Clinton are the main two candidates of this election. Google releases some of the major questions which people want to know by searching Google.

Figure 2. Top Trending Question on Donald Trump

Figure 3. Top Trending Questions on Hillary Clinton

After the presidential debate, we have a graph which has shown the situation about Donald Trump in Twitter.

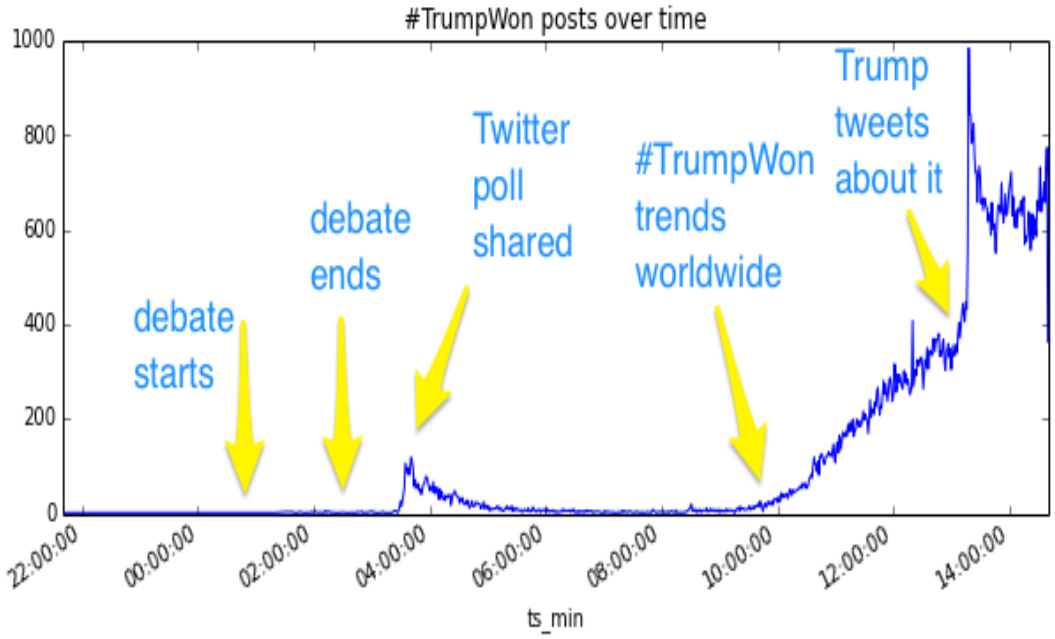


Figure 4. Trending Hashtag #TrumpWon over time on Twitter

It cannot happen overnight, there are lots of analyze behind it. On the other hand, for trend detection when we look on the internet, we go on two kind of websites. They are: Newspaper and Online social media. The most popular social media website in our country is Facebook, but they don't have any trend analysis feature yet. And the people of our country don't use twitter most.

1.3. Top and Recent News Finding from Bangladeshi Online Newspapers

Pipilika is the first search engine of Bangladesh. They have launched a great feature about recent, top and most hit news. Everybody can get update about the recent, hit and top most trending news from different newspaper in this website's new feature. Every news is clustered here, so if anybody want to take a look at a specific news, they will get all of the newspaper link at once. But there are limitations too, top trending news they are showing are collected from the newspaper's website who have got the top trending news feature, they couldn't affiliate all of the newspaper at once. But overall, it's awesome. We remember when we were eagerly waiting for the news of hanging Kader Molla, we were looking at this website's most recent news section for hours. They gather all of the newspaper's news in one place at a time & it's really very useful. Let's take a look at this:

Figure 5. Pipilika Recent News(Beta) Feature Screen Shot

2. Background Study

2.1. Twitter Trending Data Analysis

Twitter regularly updates their top trending hashtag topics based on most people's tweet. There are mainly 3 algorithms behind it, we will describe them briefly below.

(1) **Simple counting:** There are 400 million tweets published every day and 4600 tweets per second. After gathering all the tweets on twitter, we want to observe in a time period, we have to follow some steps. Find the n-gram of every tweet and store it in a dataset. Then tokenize them and count every n-gram's frequency. Now

arrange them by descending order, but the stop words always dominate. So, we have to remove them as well. We have to establish a baseline of expected frequencies based on history and compare their current frequencies to baseline. Then calculate the ratio. It works great, but low past frequency terms could get artificially inflated. Sandboxing, thresholds and smoothing will drawback latency. We also need a better statistic than simple ration to capture relative growth.

(2) **Pearson chi-squared test**: A common test for such a goodness of fit experiment is the Pearson chi-squared test. Chi-square value uses to determine for trend detection.

$$\chi^2 = \sum \frac{(observed\ count\ -\ expected\ count)^2}{expected\ count}$$

Calculate it from the n-gram & counting dataset. We will get our chi-square value. If the p-value > 0.05 then it isn't statistically significant, but if p-value < 0.05 then the null hypothesis is not rejected.

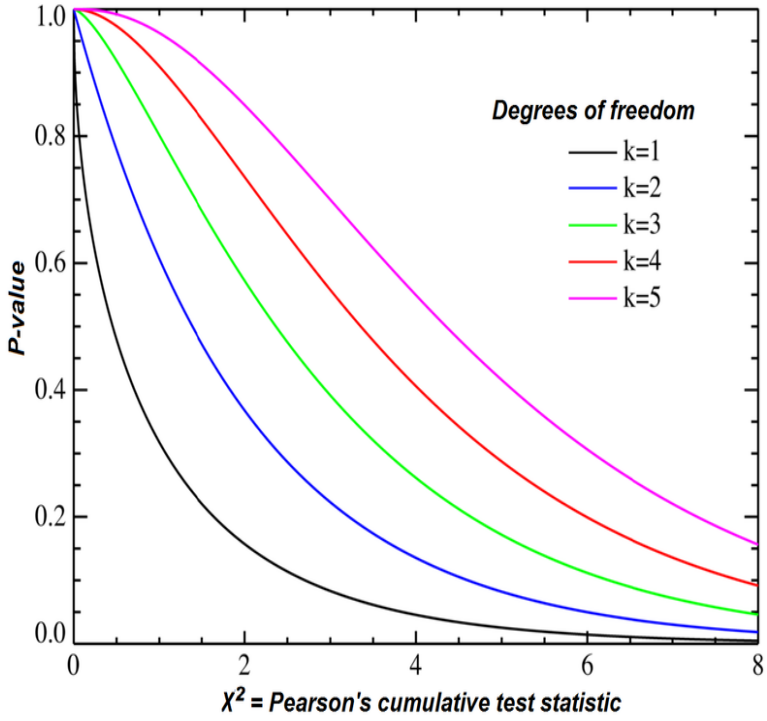


Figure 6. Chi-squared Distribution, showing X2 on the x-axis and P-value on the y-axis [15]

(3) **Per-topic models:** Autoregressive (AR) model had been used to predict trend detection:

$$\log(St) \sim \log(St-1) + \log(St-12) + xt + e$$

Linear regression estimation:

$$\log(St) \sim w1\log(St-1) + w2\log(St-12) + w3xt + w4et$$

But this will directly translate to trend detection problem:

$$\log(ft) \sim \log(ft-24hours) + \log(ft-1week) + et$$

We will need richer feature set to find trending topic in that way, it's also harder to compute and update. We can maintain more statistics than just expected value such as: periodicity, standard deviation and model them [6].

2.2. Google News Lab

Google announced the launch of a new site called News Lab, a destination that aims to connect journalists with programs, data and other resources to aid in their reporting. The site will feature a number of tools for newsrooms, including tutorials and best practices on how to use Google products in reporting, as well as provide access to the recently updated Google Trends service and more. It will also showcase Google's numerous efforts surrounding new media partnerships and citizen reporting [7].

2.3. Google Trends

Google Trends is a public web facility of Google Inc., based on Google Search, that shows how often a particular search-term is entered relative to the total search-volume across various regions of the world, and in various languages. The horizontal axis of the main graph represents time (starting from 2004), and the vertical is how often a term is searched for relative to the total number of searches, globally. Below the main graph, popularity is broken down by countries, regions, cities and language. Note that what Google calls "language", however, does not display the relative results of searches in different languages for the same term(s). It only displays the relative combined search volumes from all countries that share a particular language. It is possible to refine the main graph by region and time period. On August 5, 2008, Google launched Google Insights for Search, a more sophisticated and advanced service displaying search trends data. On September 27, 2012, Google merged Google Insights for Search into Google Trends [8].

2.4. Pipilika Recent News

Recent News search works based on various online newspapers who have got this feature of their own. Pipilika just crawls the data from the Top and Recent news section. They don't analyze any of the content of the news. But pack up all of the news and bind them together is a great feature. They have done some clustering and put them together in the website. It's certainly a unique work.

3. Methodologies

3.1. Data Mining from Online Bangla Newspaper

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [9].

Data mining is an important part of our research. Here data will play the role of fuel in our work. There are a lot of daily online newspaper out there. Every single day millions of data is published and produced. They published news about both Bangladesh and the world. Those data can be very useful if it can be used properly. Top trending news search; predict about fashion, technology, business, art & culture, international politics and sports can be measured with those data. But they are not arranged properly, they are scattered. A few newspapers have the feature of getting their data easily. So, in that case, we need a good Crawler by which we can access and store all the data of our desired newspapers. We will need both date, news headline, place and the contents within it.

Pipilika's crawler is very efficient. So, we choose it to crawl Prothom Alo and Bdnews24's one year's data with it.

The online portal of *Prothom Alo* (www.prothom-alo.com) is the number 1 Bangladeshi website in the world. This portal is accessed by 1.6 million visitors from 200 different countries and territories across the globe with 60 million pageview per month. The e-paper site of *Prothom Alo* (www.eprothomalo.com) is also the Number 1 e-paper Web site of Bangladesh. From 160 countries, 465 thousand visitors access this website with more than 26 million pageview per month. On an average, each of the visitors stays for 20 minutes in this Web site. Based on Facebook fan following, *Prothom Alo* is one of the leading corporate houses of Bangladesh. Till November 2015, 6.75 Million people are following this newspaper through Facebook. This is the biggest FB Fan page for any organization in Bangladesh [10].

bdnews24.com is also one of Bangladesh's leading newspaper/media groups. With an Alexa ranking placing it consistently among the world's top websites. It has also persisted as the country's leading English language as well as bilingual English and Bengali website. bdnews24.com ranks among the top 10 of all websites consulted by Bangladeshis [11].

So, we choose those 2 most popular and reliable newspaper for mining the data.

3.2. Finding n-gram Dataset

An n-gram is a sub-sequence of n-items in any given sequence. In computational linguistics n-gram models are used most commonly in predicting words (in word level n-gram) or predicting characters (in character level n-gram) for the purpose of various applications [1].

An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram", size 3 is a "trigram". Larger sizes are sometimes referred to by the value of n, *e.g.*, "four-gram", "five-gram", and so on. An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a $(n-1)$–order Markov model. n-gram models are now widely used in probability, communication theory, computational linguistics (for instance, statistical natural language processing), computational biology (for instance, biological sequence analysis), and data compression. Two benefits of n-gram models (and algorithms that use them) are simplicity and scalability – with larger n, a model can store more context with a well-understood space–time tradeoff, enabling small experiments to scale up efficiently [12].

We choose to analyze it up to tri-gram. The uni-gram, bi-gram and trigram. From the news crawling xml file, we have to run a program which will analyze the file and store all the n-grams into the dataset.

If we choose to discover the sentence – "আমি ভাত খাই" to n-gram. Then it will look like:

Table 1. বাংলা n-gram সেট

	বাংলা n-gram **সেট**
Unigram	আমি, ভাত, খাই
Bi-gram	আমি ভাত, ভাত খাই, খাই আমি
Tri-gram	আমি ভাত খাই

Our data is stored in XML format. So, we have to separate them by using JAVA SAX Parser. After that, we have to store every n-gram into database with id, date and the actual

word after trim (a library of java to remove the space before and after String). We will discuss about the whole storing process and the database below in 3.5 section.

3.3. Stem uni-gram Words

Stemming of uni-gram is also important. If we don't stem them apart, we cannot get the actual word. Because most of the time Article is attached with the words. An article is a word used to modify a noun, which is a person, place, object, or idea. Technically, an article is an adjective, which is any word that modifies a noun. Usually adjectives modify nouns through description, but articles are used instead to point out or refer to nouns [13]. Article is called "নির্দেশক বিভক্তি" in Bangla language. The most used article words usually are "খানা", "খানি", "গুলো", "গুলি", "য়োন", "টা", "টি", "য়েদেরকে", "েদেরকে", "দেরকে", "য়েদের", "েদের", "য়েরা", "দের", "েরা", "য়ের", "ের", "তে", "রা", "কে", "িনি", "নি". Removing the articles from unigram is one of the part of our work. We cannot remove the articles from bi-gram and tri-gram. Because then it will be a hazard.

3.4 Storing the n-grams into a Database

After finding n-gram and stem the unigram – it's time to save all the words into a database. We store the data with the date. Date is a significant measure in our work. Finding trending words is tricky, from a timeline we will calculate all the data by date. Also there will be a Primary ID for each of the words to keep actual track of every word. The database is look like this:

Table 2. Database for Storing n-grams

Column	Type
ID (primary key)	Integer
Date	Date
Word	Varchar
Count	Integer

If we find any match for both date and word, we will increase the value of count by 1. This is how this system requires for making it work. By doing this, we will get the actual frequency of every word set for every single date.

3.5. Removing the Stop Words

The stop words always dominate. So, we have to remove it from the main Dataset so that we can get the actual significant Trending words. There are 116 stop words in Bangla. Such as: অবশ্য, অনেক, অনেকে, অনেকেই, অন্তত, অথবা, অথচ, অন্য, আজ, আছে, আপনার, আপনি, আবার, আমরা, আমাকে, আমাদের, আমার, আমি, *etc*.

3.6. Pearson Chi-Square Test

The Chi-Squared Distribution is probably the most widely used distribution in Statistics today. We choose to implement it because it's way more efficient in discovering twitter's trending topic and researchers also suggest to implement it. It is most well-known for its use in Pearson's Chi-Squared Test which is used to measure goodness of fit. Chi-square value uses to determine for trend detection.

$$\chi^2 = \sum \frac{(observed\ count\ -\ expected\ count)^2}{expected\ count}$$

If we choose to find 1 months observed value, then we will have to run a query which will add all the frequency in 1 month. Then we have to divide it by 30 to measure the average frequency of everyday. And for the expected count measurement, we have to choose a day's dataset's frequency. Then we will get our chi square result.

But what if we get expected count = 0? We will add 1 for smoothing.

$$\frac{((O+1)-(E+1))^2}{E+1} = \frac{(O-E)^2}{E+1}$$

If the calculated P-value is less than 0.05, then there is a statistically significant relationship between the two classifications.

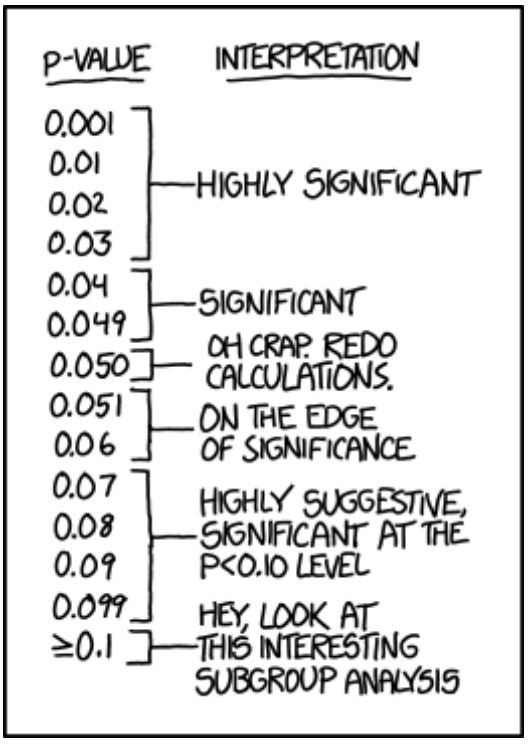


Figure 7. p Value Interpretation of chi-square Test [14]

The traditional method:

FORMULA FOR Z SCORE
USING PROPORTIONS

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

p-value method:

FORMULA FOR Z SCORE
USING MEANS

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Calculate it from the n-gram & counting dataset. We will get our chi-square value. If the p-value > 0.05 then it isn't statistically significant, but if p-value < 0.05 then the null hypothesis is not rejected.

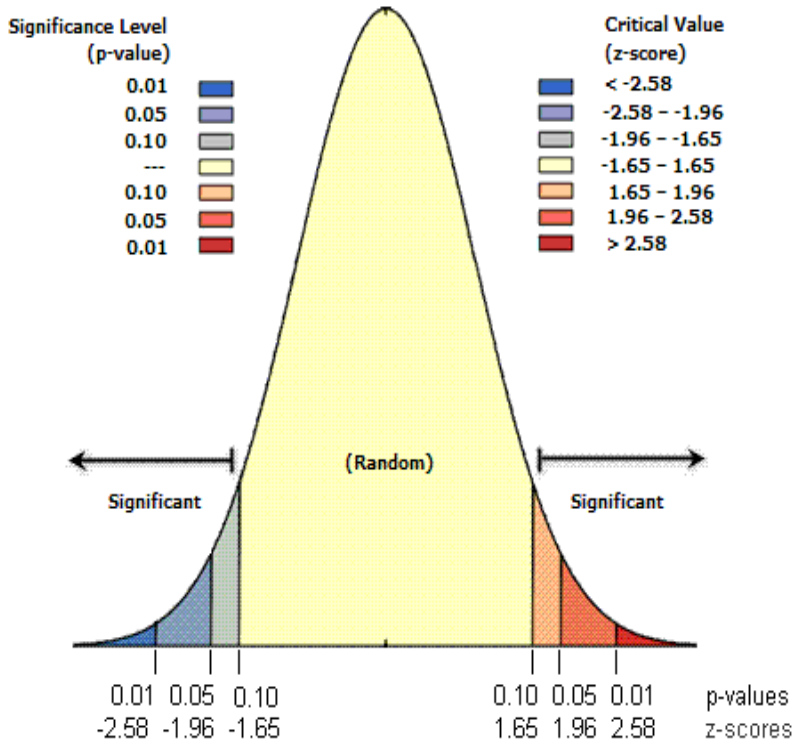


Figure 8. Visual interpretation of Distribution of Significance Level (p-values) and z-score in ArcGIS (ESRI, 2013c)

We can arrange the dataset to find the most trending topic in two ways. One is to store chi-square value and the p-value. The more the p-value is lesser than the significant value like 0.05 the more it's significant.

We analyze the whole dataset of n-gram counting and run the chi-square test & p-value method and store it in another dataset. Then we rearrange it in descending order. The greater value of chi-square we can measure it by a most trending topic.

3.7 Work Flow

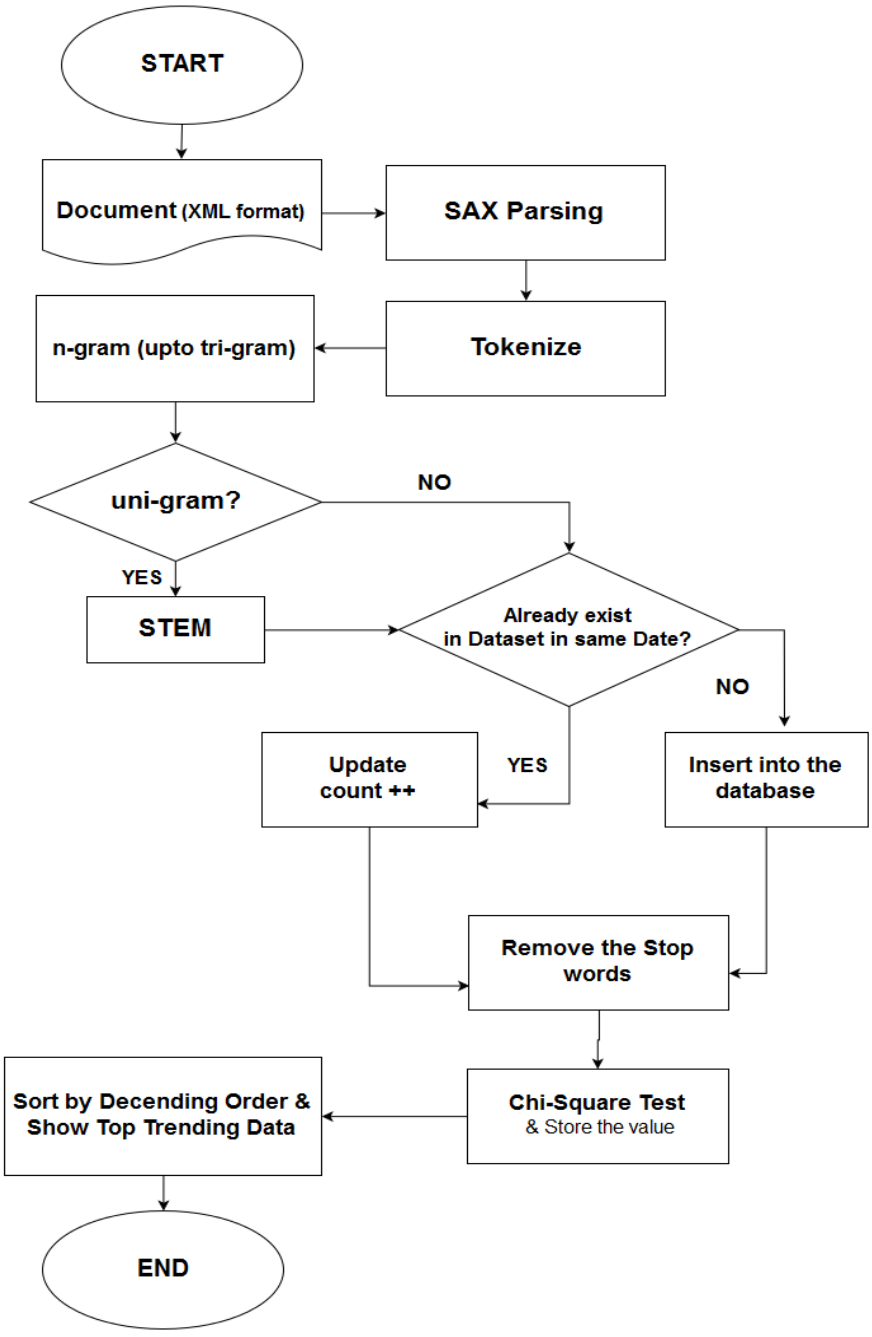


Figure 9. Flow Diagram for analyzing Trending Data

4. Result Analysis

After analyzing over millions of Data and from 40576 news of 3 months from Prothom-alo and bdnews24. we have got some of the Trending topic using the Algorithm. Some of the Trending Topics are below:

Table 3. Some Trending Topics

বাংলাদেশ	আওয়ামী লীগ	পুলিশ	নির্বাচন
বিএনপি	উপজেলা	উপজেলা নির্বাচন	আগুন
টাকা	রাজনৈতিক	জাতীয়	ভোট
মাশরাফি	বিশ্ববিদ্যালয়	ঢাকা	ক্রিকেট
ছাত্রলীগ	ভারত	ভোট	শেখ হাসিনা

We have also experimented about "হরতাল" when it used to happen frequently between 30 November 2010 and 11 February 2014. After analyzing on the data, we have got some results.

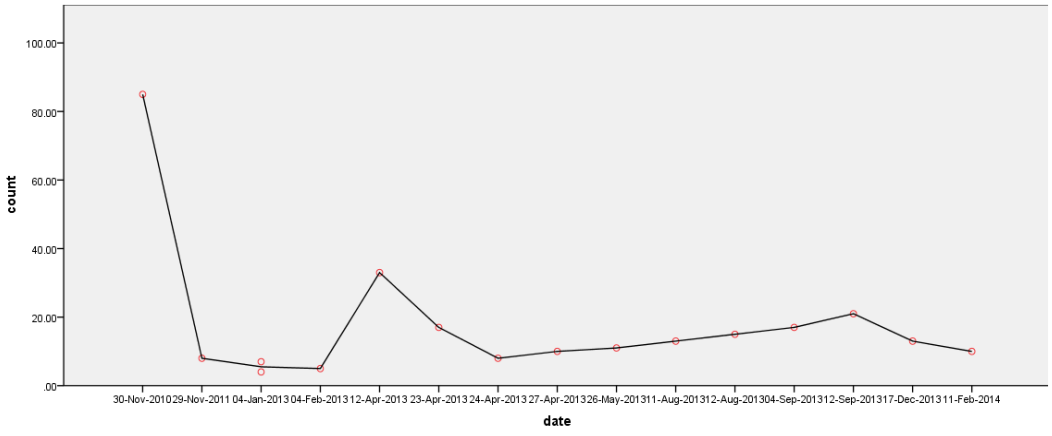
Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	3.162	.354
	Cramer's V	.953	.354
N of Valid Cases		16	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	160.000[a]	154	.354
Likelihood Ratio	74.860	154	1.000
Linear-by-Linear Association	6.442	1	.011
N of Valid Cases	16		

a. 180 cells (100.0%) have expected count less than 5. The minimum expected count is .06.



5. Limitations

- Data Mining and analyzing the whole thing is really a lengthy process, it takes so much time to extract from the document, tokenize, find n-gram, cross match with date & word and insert n-gram into the database. We have tried to reduce the complexity of this program as much as possible. But it's still a slow process.

- The stemmer we have used give us more than 80% accuracy, but it's still have so much limitation. We cannot find any proper stemmer which can give me over 95% accuracy. If we use a good stemmer, then the result of finding trending topic can enhance more.

- We have crawled and used 2 of the most popular online newspapers. But if we can get all of the newspaper's past 5 years' data, then we can find more significant trending data.

6. Conclusion

Finding and analyzing of Trending topic research is totally new in Bangla Linguistic. We haven't found any past research work on Trending topic/news. On the other hand, Trending topic research about Newspaper, Blog, Business, Social media microblogging, fashion, sports *etc.*, in English are very enrich, having good performance too. We have found Pearson's chi-square test very useful and significant statistically, so we choose to use it. But there can be various ways to enrich this research.

References

[1] M. Munirul, U. Naushad and K. Mumit, "Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus", International Conference on Computer and Information Technology, ICCIT 2006 23 December, 2006 Dhaka, Bangladesh.
[2] https://en.wikipedia.org/wiki/History_of_journalism
[3] http://www.businessdictionary.com/definition/trend.html
[4] T. Althoff, D. Borth, J. Hees and A. Dengel, "Analysis and forecasting of trending topics in online media streams", In Proceedings of the 21st ACM international conference on Multimedia (MM '13), ACM, New York, NY, USA, **(2013)**, pp. 907-916.
[5] https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/time-trend-analysis
[6] K. Tsioutsiouliklis, "Trend and Event Detection In Social Streams", **(2012)**.
[7] https://techcrunch.com/2015/06/22/google-launches-a-new-home-for-journalists-with-news-lab/
[8] https://en.wikipedia.org/wiki/Google_Trends
[9] https://en.wikipedia.org/wiki/Data_mining
[10] https://en.wikipedia.org/wiki/Prothom_Alo
[11] https://en.wikipedia.org/wiki/Bdnews24.com
[12] https://en.wikipedia.org/wiki/N-gram
[13] http://study.com/academy/lesson/what-are-articles-in-english-grammar-definition-use-examples.html
[14] https://xkcd.com/1478/
[15] https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

