

Bachelor of Science in Computer Science and Engineering

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



Thesis

An Approach for News Trend Analysis Based on Online Bangladeshi Newspapers

Prepared By

Ponir Ahmed Khan (Std No: 084402)

Md. Touhidur Rahman (Std No: 084418)

Supervisor

Tareque Mohmud Chowdhury

Assistant Professor, Department of CSE

Islamic University of Technology

Department of Computer Science and Engineering

Islamic University of Technology (IUT)

Organisation of Islamic Cooperation (OIC)

Gazipur, Bangladesh

October, 2012

An Approach for News Trend Analysis Based on Online Bangladeshi Newspapers

Prepared By

Ponir Ahmed Khan (Std No: 084402)
Md. Touhidur Rahman (Std No: 084418)

Supervisor

Tareque Mohmud Chowdhury
Assistant Professor, Department of CSE
Islamic University of Technology

**This thesis was undertaken as a partial fulfillment of requirements for the
Bachelor of Science degree in Computer Science and Engineering
(B.Sc in CSE)**

Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Organisation of Islamic Cooperation (OIC)
Gazipur, Bangladesh
October, 2012

CERTIFICATE OF RESEARCH

This is to certify that the work presented in this paper is the outcome of the investigation carried out by the candidates under the supervision of Mr. Tareque Mohmud Chowdhury, in the Department of Computer Science and Engineering, Islamic University of Technology, Gazipur. It is also declared that neither of this thesis nor any part of theory has been submitted anywhere else for the award of any degree or diploma.

Authors:

Ponir Ahmed Khan
Student No: 084402 Signature

Md. Touhidur Rahman
Student No: 084418 Signature

Supervisor:

Tareque Mohmud Chowdhury
Assistant Professor, CSE Department Signature
Islamic University of Technology

Head of the Department:

Prof. Dr. M. A. Mottalib
Head, CSE Department Signature
Islamic University of Technology

Department of Computer Science and Engineering, IUT

October, 2012

Acknowledgement

By the grace and blessings of Almighty Allah we have finally completed our Thesis successfully. We express our heartiest gratefulness to Almighty Allah for giving us that capability.

First and foremost, we feel grateful and acknowledge our profound indebtedness to Professor Dr. M. A. Mottalib, Head, Department of Computer Science and Engineering, IUT and our Supervisor Mr. Tareque Mohmud Chowdhury, Assistant Professor, Computer Science and Engineering, IUT. Their endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice at all stages has made it possible to complete this project.

We would like to take this opportunity to thank our teachers for guiding us throughout this thesis and for giving us valuable advice when required. We would also like to thank our friends who helped us in this thesis.

Finally, we like to offer thanks to all my classmates, juniors and many others who contributed us in many ways during the thesis work.

Abstract

In the present world, there are many people who are talking in bangla language. We the Bangladeshi people use this language as a mother language. In our system we work with this bangla language.

There are many online bangla newspapers. We collect the articles from these online newspapers using web crawler program. In our case we use the prothom-alo online newspaper. Category wise we collect those articles. The main task of our thesis is to find the trend of the newspaper. That is, what type of things the daily newspaper having every day? What's the number of positive article and negative article? For this purpose we have gone through the newspaper's articles to find some method or rule that can give us the view about the article. We have used step by step process. We try to collect the feedback for each of the bangla words from the general people. Here feedback is positive, negative and neutral value. Then using these feedbacks and proposed rules we try to find the weight of the sentences and articles.

For weighting each word we have collected the feedback that I mentioned earlier. But for finding the weight of the sentences and articles we use some set of rules. In this thesis approach we have faced many problems with weighting the words, sentences and articles. This weighting technique is the core technique here. We have analyzed many newspaper articles to finally come with a technique called "addition rule" to weight the sentences. We have mentioned many problems and ambiguities here. Through our proposed algorithm we have tried to solve those problems and show the trend of the newspaper.

List of Figures

<u>Figure name</u>	<u>Page</u>
Figure 3.1: Step by step stages of our proposed system	05
Figure 3.2: Screen shot of our online interface for collecting the feedback	06
Figure 3.3: Statistics of the feedback of collected words	07
Figure 3.4: Step by step method for collecting the newspaper articles	08
Figure 3.5: Flowchart for finding the weight of the article	11
Figure 4.1: Step by step process for experiment and evaluation	17
Figure 4.2: Result of Manual approach	18
Figure 4.3: Result of the System	18
Figure 4.4: Screen shot of the online interface for finding the weight of the	19

List of Tables

<u>Table name</u>	<u>Page</u>
Table 3.1: List of Bangla বিভক্তি	10
Table 3.2: List of positive and negative words from the above articles	16
Table 4.1: Result table of the system	20
Table 4.2: Result table for the manual approach	20

Table of Contents

	Acknowledgement	i
	Abstract	ii
	List of Figures	iii
	List of Tables	iv
	Table of Contents	v
Chapter 01	Introduction	01
	1.1 Background	01
	1.2 Motivation	02
	1.3 Related Work	02
	1.4 Problem Description	03
Chapter 02	Objective	04
	2.1 Methodology	04
Chapter 03	Our Proposal	05
	3.1 Data Collection	05
	3.1.1 Feedback Collection	06
	3.1.2 Collection of Newspaper Article	07
	3.2 Data Processing	09
	3.2.1 Getting Individual Word	10
	3.2.2 Weighting Technique	11
Chapter 04	Experiment	17
	4.1 Manual Result	17
	4.2 System Result	18
	4.3 Comparison between the Two Results	20
Chapter 05	Evaluation	21
	5.1 Some other Problems	21
Chapter 06	Discussion and Conclusion	24
	References	25
	Appendix	26

Chapter 1

Introduction

Trend analysis is a form of comparative analysis that is often employed to identify current and future scenario of particular thing. In a word we can say that, Trend Analysis refers the practice of collecting information and attempting to spot a pattern, or trend, in the information. An aspect of technical analysis that tries to view the current scenario of information based on past data or current data. Trend analysis is based on the idea that what has happened in the past or present gives viewers an idea of what will happen in the future. So trend analysis is highly used to view the future or present scenario depending on the past or present data.

In our work, we will gather information about the bangla words, sentences, and show their trend based on their meaning on the sentence and article respectively. This will ultimately help us to find the trend of the newspaper.

1.1 Background

As Bengali is the 4th language among all of the language in the world, we are working with this language. We are using Bengali newspaper for applying our trend analysis. In our country, it's a long time we are used to read newspaper. Newspaper is a part of our life & almost every people read the newspaper. But we haven't any thing that shows what we are having almost every time in the news. That means we are having positive news or negative news. So we have decided to make an analysis graph or chart with last six months news database, that what types of news actually the newspaper contains. We classified news into two category i.e. positive and negative news. Making the Trend Analysis is our ultimate goal of our research. In our research, we will analyze the Bengali newspaper i.e. every article and summery of the newspaper to find the positive and negative news so that we can find how many positive and negative news are there. This analysis result will predict the current and future situation of the country with a graph and histogram. In our research we give more emphasize in the heading of the newspaper as well as

full article. By giving the positive and negative weight on each word and with proper calculation we will get the result. We will use some set of rules to get the almost accurate result.

1.2 Motivation

Still now in our country no work has been done regarding area like Bengali newspaper trend analysis. So this topic seems to us very challenging. We are doing our analysis using Bengali newspaper in every categories of the paper. As we are used to read daily newspaper in our day to day life, so we do our research Bengali news trend analysis so that we can build such a system which will analyze the article and heading of newspaper and give result about it. Our main goal is to figure out the types of news automatically i.e. positive news or negative news. Bengali news trend analysis totally Bangla language based research. Giving weight on each and every word is necessary according to their positivity and negativity. We determine the types of article by calculating the given weight on word. By proper calculation we are able to find the total weight of every sentence in article that gives us result and scenario of that article what actually it means. By analyzing the articles of newspaper we get the result whether the article is positive or negative. So we can get the present scenario of our country depending on the positive or negative news. By this we are able to get the idea about our country. We analyze the every category of newspaper such that commerce and economics, country news, first page etc. after analyzing the news, positive news and negative news are shown in a graph or a chart so that viewers can get the scenario easily.

1.3 Related Work

Actually, still now there is no system related to our research i.e. Bangla news trend analysis. In Last year, research was done related to this area which had some ambiguities and problems.

In our research we try to figure out those problems with appropriate solution. Working in this area i.e. using Bengali language news trend analysis is little bit difficult because of difference in Bangla grammar with English grammar.

1.4 Problem Description

Working with Bangla language is pretty challenging. Because word categorization is complex in Bangla language than English language.

There have differences between English grammar and Bangla grammar. It is difficult to separate and give weight Bangla words on the basis of positive and negative. Conflicting words in the article would be handled very carefully.

Weight of Every sentence of the article gives the result about that article whether article shows the positive or negative news.

Some articles contain negative news with positive heading which may give incorrect result. Sometimes it may be seen that actual meaning of the article doesn't match with the result we find after our calculation. Literally or theoretically article may be positive or negative but mathematically not.

Chapter 2

Objective

Our main goal is to build an automated system which will show the trend of the newspaper for particular day, particular week, and particular time range. In a certain time this system starts and analyzes the newspaper automatically and stores them in a database. Here we Show the present scenario of our country in various fields like politics, finance, business, sports etc. after completion of analysis. Categorization of the words and give them proper weight. For showing the final goal, we have to find the proper weight of the each word and each sentences and the whole article. This is the most critical part of our thesis.

2.1 Methodology

To achieve our goal, we follow some steps and rules for categorizing the Bangla words depending on their meaning and use of purpose. Because, before giving weight we have to separate the word according to their meaning. We are adding the weight of words in a sentence to find the weight which leads us to find the meaning of that article is positive or negative. So our methodology includes -

- ⇒ Categorizations of Bengali words and give them proper weight.
- ⇒ Use this categorization; we have to find the weight of the sentence and then the whole article.
- ⇒ We will propose a method for finding those weights.
- ⇒ We will verify our weighting method by checking the success and error rate.
- ⇒ Finally, build an automated system which will show the trend of the newspaper for particular day, particular week, and particular time range based on that weighting method.

Chapter 3

Our Proposal

Our system consists of three stages. Those stages are work together in the system for showing the result and performing the other action. These stages are step by step processing stages for our thesis approach. Below figure shows that –

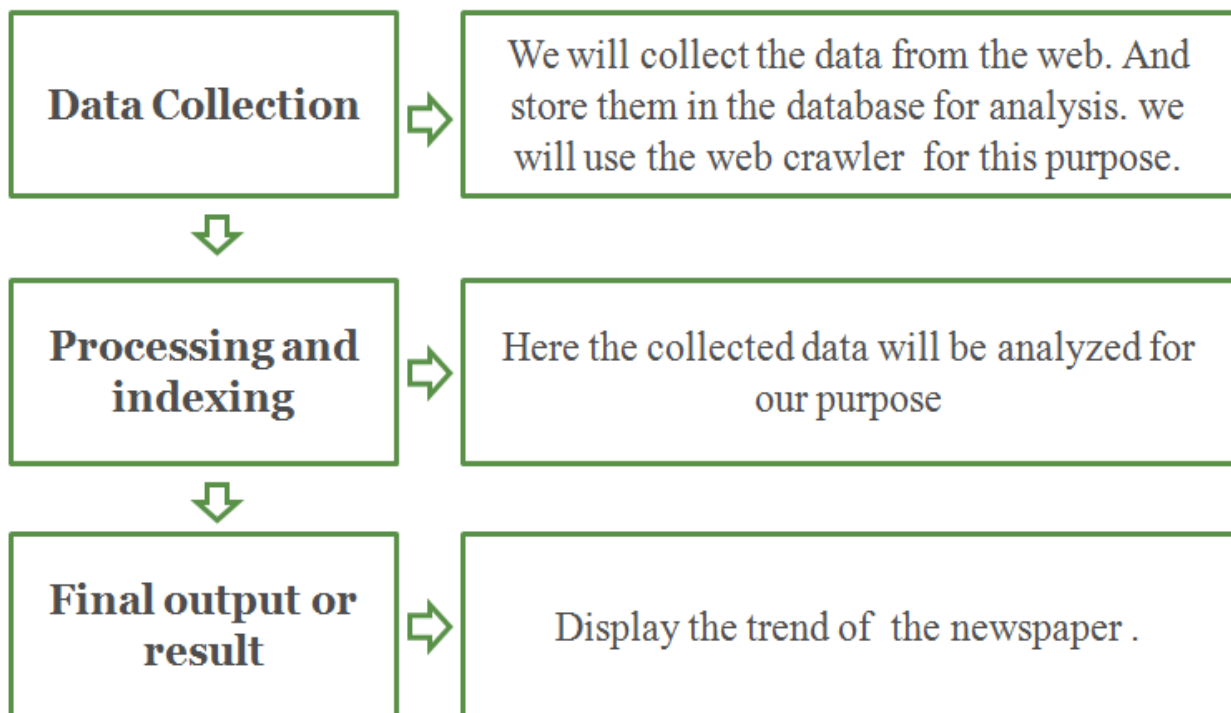


Figure 3.1: Step by step stages of our proposed system.

3.1 Data Collection

In this stage, we collect the bangla words and bangla newspaper articles. For collecting bangla words we use an online bangla word database and using this database, try to collect the positive






and negative feedback for each word from the general people. And in the next step we try to gather the articles from the online newspaper.

3.1.1 Feedback Collection

In this phase we collect a bangla word database from the internet. We use this database to create our own database for collecting the positive and negative feedback for each word. For getting the feedback from the general people, we built an online interface (www.shabda.net). Using this interface people can give their views about the bangla words. The meaning of the words and type of the words are given there. For better understanding of the meaning of the word, Google suggestion is also given there. People can assign positive, negative and neutral value to each word.

Categorization of Bangla Words

পৃথিবীর সব ভাষাতেই পজিটিভ এবং নেগেটিভ অর্থবোধক শব্দ রয়েছে। আমরা বাংলা ভাষার এই জাতীয় শব্দ গুলোর একটি তালিকা তৈরী করার উদ্যোগ নিয়েছি। যে তালিকার মাধ্যমে সহজেই আমরা বাংলা ভাষার পজিটিভ এবং নেগেটিভ শব্দ গুলোকে শ্রেণীবিভাগ করতে পারি। এই তালিকা তৈরীর জন্য আমরা আপনাদের একান্ত সহযোগিতা কামনা করছি। ধন্যবাদ।

Bangla Words	Type	Meaning	Trend	Remarks
ফাঁপানো	 বি.	বায়ুপূর্ণ করা	<input type="text"/>	<input type="text"/>
রেসলা	 বি.	অশ্বারোহী সৈন্য.	<input type="text"/>	<input type="text"/>
কলাপোড়া খাওয়া	 বি.	ব্যর্থ হয়ে পড়ে থাকা	<input type="text"/>	<input type="text"/>
বিদ্ধ	 বিণ.	পীড়িত, দুঃখাহত.	<input type="text"/>	<input type="text"/>
জগদ্বাসী	 বি.	পৃথিবীর অধিবাসী	<input type="text"/>	<input type="text"/>

[Log Out](#)

Figure 3.2: Screen shot of our online interface for collecting the feedback.

In the database, there are more than 62000 words. So to collect the feedback, we face a great challenge. Out of 62000 words we able to collect feedback for 10504 words. Where 1955 are positive words, 2265 are negative words and 6284 are neutral.

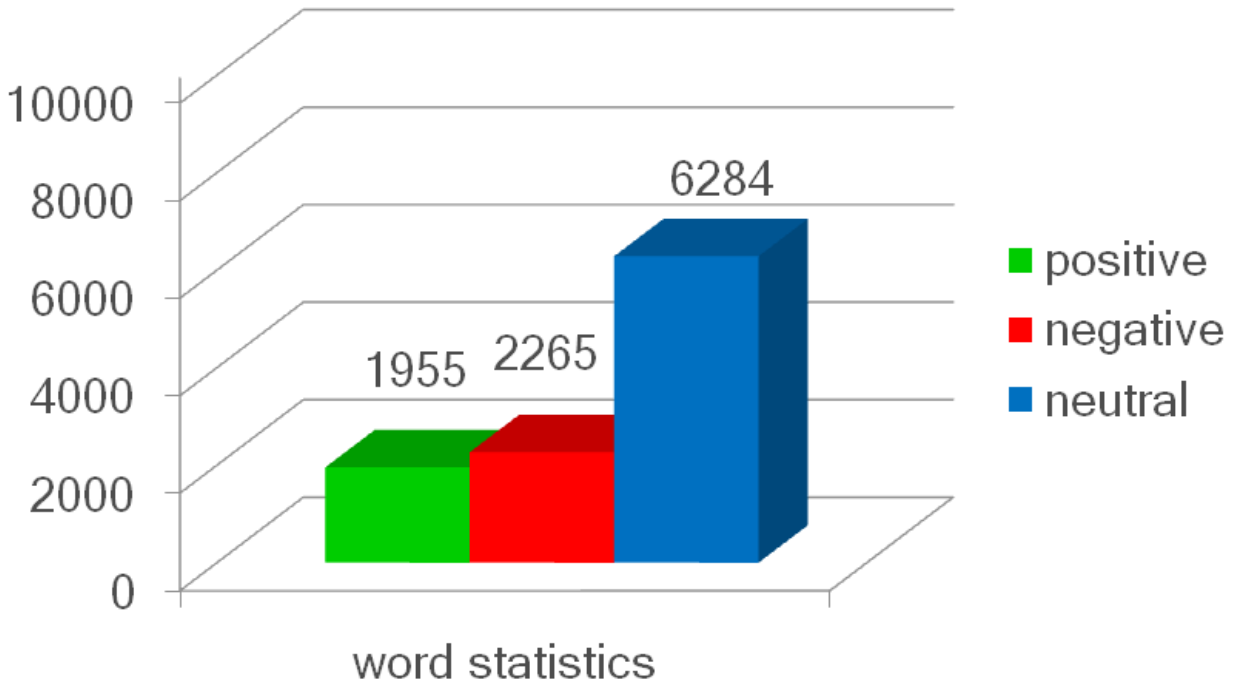


Figure 3.3: Statistics of the feedback of collected words.

3.1.2 Collection of Newspaper Article

For this purpose, in our thesis we consider the online prothom alo newspaper i.e. www.prothom-alo.com. There are many categories in here. From these categories, we collect article from categories that includes - *প্রথম পাতা*, *সারাদেশ*, *বিশাল বাংলা*, *রাজধানী*, and *অর্থ-বাণিজ্য*. For collecting articles we are using web crawler program.

A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion.

It is one type of bot, or software agent. In general, it starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies.

How we are approaching in this phase are given below –

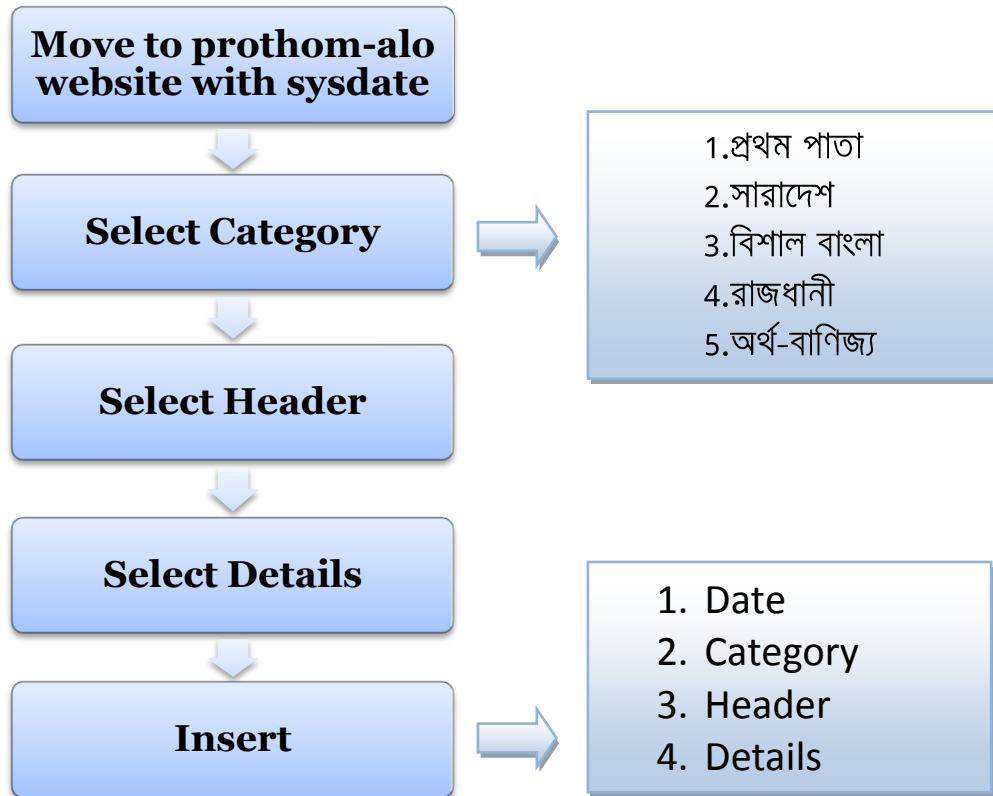


Figure 3.4: Step by step method for collecting the newspaper articles

Here, we are inserting the date, category, article header, and article details into our database for the processing purpose.

3.2 Data Processing

This is the main phase of our system. Here we use our previously mentioned database of Bangla words with feedback. Before processing the articles of the newspaper based on our system, we have gone through the articles manually for achieving or setting rules for finding the weight of the article i.e. is it positive or negative?

When we have analyzed newspaper we face many problems and ambiguities about how to get the trend of the newspaper. Human can easily identify what the paper says, is it giving us any positive idea or negative one. But when we think about doing that with the system, we face many problems. Those are –

⇒ How to get the individual word from the sentence. As for example -

“আকুতে কমছে লেনদেন” how to get the word “কম” from “কমছে”

⇒ How and what word will be weighted in the sentence? There are more than sixty two thousands (62,000) words in our Bangla word database, from that may be 10-12 thousands word will be weighted with negative and positive value.

When we try to weight the sentences based on the weight of the words we face many other problems. Like

⇒ How to deal with the sentences enrich with positive words but actual meaning is negative and vice - versa.

⇒ How to deal with “classification type” sentences such that,

“শিক্ষা খাতে বাঁধা গুলো হল প্রান্তিক জনগোষ্ঠীর নিকট শিক্ষা সেবা পৌঁছানো, প্রাথমিক শিক্ষার মান নিশ্চিত করণ,”

⇒ How to deal if there is a conflict between the positive word with the name of any place or person.

⇒ How to deal with proverbs. Like –

“নাচতে না জানলে উঠান বাঁকা”

⇒ How to deal with “double meaning” words. i.e. There are words which give both the positive and negative meaning in different sentences such words are –

“কমা, বাড়া, হ্রাস, বৃদ্ধি, সাহায্য, সহায়তা, সহযোগিতা, না, নামা, পিছিয়ে, বাধাগ্রস্ত, উর্ধ্বগতি, নিম্নগতি” etc.

As for example:

আকুতে **কমে** গেছে বাংলাদেশের লেনদেন

রমজান মাসে চালের মূল্য **কমার** সম্ভাবনা রয়েছে

আবারও **বাড়লো** বিদ্যুতের দাম

দেশের জনগনের নাগরিক সুবিধা **বাড়াতে** সরকারী উদ্যোগ

Here in the first two sentences, the word “কম” gives negative and positive meaning respectively. And in the next two sentences, the word “বাড়া” gives negative and positive meaning respectively. In that case the weight of this kind of words depends on the words Those affect their meaning.

These are some of the problems and ambiguities that we have encountered. In this stage of our system, we try to set some rules to solve these problems. We have tried to gather as many information as possible to set the rule. So that it can give us more accurate precise result.

3.2.1 Getting Individual Word

For getting the individual words from the sentences is not easy like the English language. In bangla language there are several grammatical concepts which make it tougher. There is a grammatical concept called “বিভক্তি” which makes this work tougher. There is several number of this. These are

Table 3.1: List of Bangla বিভক্তি

বিভক্তিরনাম	বিভক্তি
প্রথমা বা শূন্য বিভক্তি	০, অ
দ্বিতীয়া বিভক্তি	কে, রে
তৃতীয়া বিভক্তি	দ্বারা, দিয়া (দিয়ে), কর্তৃক
চতুর্থী বিভক্তি	কে, রে
পঞ্চমী বিভক্তি	হইতে (হতে), থেকে, চেয়ে
ষষ্ঠী বিভক্তি	র, এর
সপ্তমী বিভক্তি	এ, য়, তে

After analyzing these in the sentences we come to the conclusion that some of the characters and words need to be discarded from the words in the sentence to get the individual words. These characters and words are –

‘ের’, ‘ে’, ‘ও’, ‘গুলো’, ‘গুলো’, ‘গুলি’, ‘গণ’, ‘কে’, ‘র’, ‘রও’, ‘ভাবে’, ‘ভাবেও’, ‘ই’, ‘তে’, ‘য়’, ‘ত’

As for example –

দেশগুলো = দ ে শ গুলো

বাংলাদেশের = বা ং ল া দ ে শ ের

পণ্যের = প ন্য ের

দেশকে = দ ে শ কে

Here, we discard those selected words and characters from the words to get the individual words দেশ, বাংলাদেশ, পণ্য, দেশ respectively.

3.2.2 Weighting Technique

For this purpose we will use the following step by step process. This is –

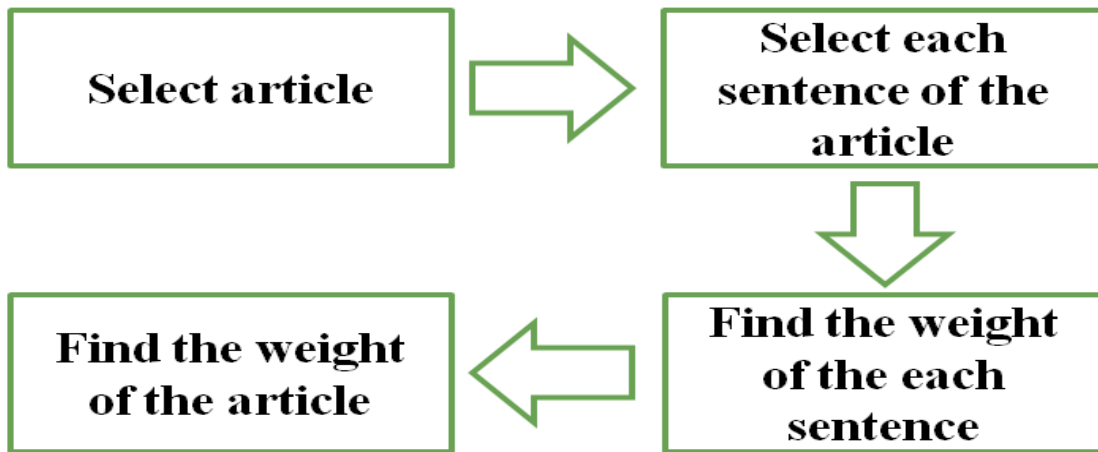


Figure 3.5: Flowchart for finding the weight of the article.

For determining the weight of the sentence of an article we introduce rule in our research which can be considered as primary rule. This is an addition rule. That is –

“Add the weight of the all weighted words present in the sentence”

Sum of all weights of all positive and negative words in a sentence = Sentence weight

If we find the result positive integer then it will be positive and if negative integer then will be negative.

As for examples:-

Positive and negative words are denoted by ‘p’ and ‘n’ respectively.

“সহস্রাব্দের উন্নয়ন(p=1) লক্ষ্য(p=1) অর্জনের(p=1) সময়কালের দুই-তৃতীয়াংশ পথ পাড়ি দিয়েছে বাংলাদেশ” [p=3](positive sentence)

“তবে উদ্বেগজনক(n=-1) বিষয় হল নবজাতক মৃত্যু(n=-1) হার হ্রাস(n=-1) পাচ্ছে খুব ধীরগতিতে(n=-1)” [n=-4](negative sentence)

“বন্ধনারী(n=-1) শিকারী(n=-1) বাংলাদেশি পোশাকশ্রমিকেরা” [n=-2](negative sentence)

“গুলশানে সৌদি দূতাবাসের কর্মকর্তা গুলিতে(n=-1) নিহত(n=-1)” [n=-2](negative sentence)

For finding the weight of the article we have proposed the rules in below

Number of positive sentences – Number of negative sentences = Article weight

If there is a miss match with the heading weight and the article weight we will consider the article weight as the final weight of that article.

Value ‘0’ will be considered as positive weight. It’s applicable for both sentences or heading and articles.

Here we are giving some of the articles that have been analyzed using these rules.

বঞ্চনার(n=-1) শিকার(n=-1) বাংলাদেশি পোশাকশ্রমিকেরা =[-2/

লন্ডন অলিম্পিকে পোশাক সরবরাহকারী প্রতিষ্ঠান পুমা, নাইকি ও অ্যাডিডাসের পক্ষে পোশাক তৈরির কাজে নিয়োজিত আছেন অনেক বাংলাদেশি শ্রমিক। তাঁদের তৈরি পোশাক পরেই হয়তো এবারের অলিম্পিকে অংশ নেবেন অনেক দেশের অ্যাথলেট। কিন্তু বিশ্বের সবচেয়ে বড়(p=1) এই ক্রীড়ানুষ্ঠানের মাধ্যমে পুমা, নাইকি বা অ্যাডিডাস তাদের পণ্যের প্রসার(n) বাড়ালেও(p=1) বাংলাদেশে বসে এই পোশাকগুলো যাঁরা তৈরি করছেন, সেই শ্রমিকেরা হচ্ছেন চরম(n=-1) বঞ্চনার(n=-1) শিকার(n=-1)।=[-6/ ন্যূনতম (n=-1)মজুরিরও(p=1) কম(n=-1) টাকায় অনেক(p=1) দীর্ঘ(n=-1) সময় ধরে কাজ করতে হচ্ছে এই শ্রমিকদের ।=[-1/ শুধু তাই নয়, শারীরিক ও মানসিকভাবেও তাঁদের নির্যাতনের(n=-1) শিকার(n=-1) হতে হচ্ছে বলে অভিযোগ(n=-1) করেছে ব্রিটিশ দাতব্য প্রতিষ্ঠান ওয়ার অন ওয়ান্ট 1/[-3/

দ্য অবজারভারের এক প্রতিবেদনে বলা হয়েছে, পুমার পোশাক সরবরাহকারী একটা প্রতিষ্ঠানে, দুই-তৃতীয়াংশ শ্রমিকের সঙ্গে কথা বলে জানা গেছে, তাঁদের মারধর(n=-1) করা হয়, ধাক্কা(n=-1) দেওয়া হয় বা চুল ধরে টানা হয়।’ অ্যাডিডাসের জন্য পোশাক সরবরাহকারী আরেকটি প্রতিষ্ঠানের নারীশ্রমিকেরা অভিযোগ(n=-1) করেছেন, তাঁদের ওড়না খুলে ফেলতে বাধ্য(n=-1) করা হয়।=[-4/

পুমার পোশাক সরবরাহকারী এক পোশাক কারখানার শ্রমিক হাজেরা খানম (৩২) বলেছেন, তাঁরা প্রায়ই আমাকে খাপ্পড়(n=-1) মারে, লাথি(n=-1) মারে, ধাক্কা(n=-1) দেয়।=[-3/ প্রায়ই আমাদের অকথ্য(n=-1) ভাষায় গালাগাল(n=-1) করে।=[-2/ মানসিকভাবে আমাদের চরমভাবে(n=-1) আহত(n=-1) করা হয়।=[-2/’ একই প্রতিষ্ঠানের আরেক শ্রমিক পপি আক্তার নিজের দুর্দশার(n=-1) কথা জানিয়ে বলেছেন, আমাকে খুবই(p=1) বাজে(n=-1) বাজে(n=-1) নামে ডাকা হয়।=[-2/ প্রায়ই চড়(n=-1) খাপ্পড়(n=-1) খেতে হয়।=[-2/ টেবিলের উপরে দাঁড় করিয়ে রাখা হয়। এমনকি চাকরি থেকে বরখাস্ত(n=-1) ও জেলে পাঠানোর হুমকিও(n=-1) দেওয়া হয়।’=[-2/

অলিম্পিকের এবারের আসরে প্রায় ১০ কোটি ইউরোর ক্রীড়া পোশাকসামগ্রী বিক্রির আশা(p=1) করছে অ্যাডিডাস। =/1/নাইকির দখলে(n=-1) রয়েছে যুক্তরাজ্যের ৪০০ কোটি ডলারের ক্রীড়াসামগ্রীর ১৮ শতাংশ ।=[-1/ এবারের অলিম্পিকে তারা স্পনসর করবে যুক্তরাষ্ট্র, চীন, জার্মানিসহ প্রায় ২৫টি দেশের অ্যাথলেটদের। অন্যদিকে বিশ্বের দ্রুততম(p=1) মানব উসাইন বোল্টসহ বেশ কয়েকটি দেশের অ্যাথলেটকে স্পনসর করবে পুমা ।=[1/ কোম্পানিগুলো তাদের পণ্যের প্রসার(p=1) ও বিপুল(p=1) পরিমাণ মুনাফার সুযোগ(p=1) পেলেও তাদের পণ্যসামগ্রী উৎপাদনের(p=1) সঙ্গে জড়িত(n=-1) শ্রমিকদের বঞ্চনার(n=-1) খোঁজ(p=1) রাখে না(-1)

বলেই প্রতীয়মান হয়। =/3/ ব্রিটিশ দাতব্য প্রতিষ্ঠান 'ওয়ার অন ওয়াক্টের' ক্যাম্পেইন ও পলিসি ডিরেক্টর গ্রেগ মুট্টিট বলেছেন, অ্যাডিডাস, নাইকি, পুমার মতো কোম্পানিগুলো শ্রমিক বঞ্চনার (n=-1) সুযোগ (p=1) নিয়ে বিপুল (p=1) পরিমাণ মুনাফা করে। =/1/ কিন্তু এই কোম্পানিগুলো যদি অলিম্পিকে অংশ নেওয়া দল ও অ্যাথলেটদের স্পনসর করে মুনাফা করতে চায়, তাহলে তাদের অবশ্যই শ্রমিকদের সম্মানের (p=1) সঙ্গে বিবেচনা (p=1) করতে হবে। =/2/ তবে প্রতিবেদন প্রকাশের (p=1) পরপরই সামগ্রিক বিষয় খতিয়ে দেখার আশ্বাস (p=1) দিয়েছে কোম্পানিগুলো। =/2/ নাইকির একজন মুখপাত্র এ প্রসঙ্গে বলেছেন, চুক্তিবদ্ধ (p=1) পোশাক কারখানার কাজের পরিবেশ কেমন, তা নাইকি খুব গুরুত্ব (p=1) দিয়ে বিবেচনা (p=1) করে। =/3/ নাইকির সব সরবরাহকারী প্রতিষ্ঠানকে অবশ্যই আমাদের নীতিমালা (p=1) মেনে চলতে হয়। =/1/ যে অভিযোগগুলো (p=1) তোলা হয়েছে, তা আমরা তদন্ত (p=1) করে দেখব এবং খুব শিগগিরই (p=1) এ ব্যাপারে ব্যবস্থা নেব। =/3/ অ্যাডিডাসের পক্ষ থেকে বলা হয়েছে, বাংলাদেশে নিযুক্ত আমাদের সব সরবরাহকারী পোশাক কারখানা নিয়মিত (p=1) নিরীক্ষণ (p=1) করা হয়। =/2/ সেখানকার শ্রমিকদের সঙ্গে নিয়মিত (p=1) কথা বলে কাজের পরিবেশ সম্পর্কেও খোঁজখবর (p=1) নেওয়া হয় একটা এনজিওর মাধ্যমে। =/2/ বাংলাদেশ থেকে পোশাক সরবরাহকারী একটা প্রতিষ্ঠান বেআইনিভাবে (n=-1) শ্রমিকদের দিয়ে দীর্ঘ (n=-1) সময় কাজ করায় -এর প্রমাণ (p=1) পাওয়া গেছে বলে স্বীকার (p=1) করেছে পুমা। =/0/ তবে খুব দ্রুতই (p=1) এই সমস্যা (n=-1) সমাধানের (p=1) আশ্বাস (p=1) দেওয়া হয়েছে বলে জানিয়েছে পুমার একজন মুখপাত্র। =/2/

The number of positive sentences = 10

The number of negative sentences = 12

Total weight of the article = (The number of positive sentences - The number of negative Sentences)

$$= 12 - 10 = -2 \text{ (Negative article)}$$

গুলশানে সৌদি দূতাবাসের কর্মকর্তা গুলিতে($n=-1$) নিহত($n=-1$)= $[-2]$

(রাজধানীর গুলশানে দুর্বৃত্তদের ($n=-1$) গুলিতে ($n=-1$) খালাফ আল আলী নামে সৌদি দূতাবাসের একজন কর্মকর্তা নিহত ($n=-1$) হয়েছেন।) $[-3]$ (গতকাল সোমবার রাতে গুলিবিদ্ধ(n)= -1 অবস্থায় তাঁকে রাজধানীর ইউনাইটেড হাসপাতালে ভর্তি করা হলে আজ মঙ্গলবার ভোর পাঁচটার দিকে তাঁর মৃত্যু(n)= -1 হয়।) $[-2]$

(নিহত(n)= -1 খালাফ সৌদি দূতাবাসের দ্বিতীয় সচিব (হেড অব সৌদি সিটিজেনস অ্যাফেয়ার্স) ছিলেন।) $[-1]$ (তিনি সৌদি নাগরিকদের সুযোগ($p=1$)-সুবিধার($p=1$) বিষয়টি দেখতেন।) $[2]$ (গুলশান খানার পরিদর্শক (তদন্ত) নূরে আজম সাংবাদিকদের বলেন,) (গতকাল রাত দেড়টার দিকে গুলশানের ১১৭ নম্বর সড়কের কনসুলেট অব পর্তুগাল ভবনের নিরাপত্তা কর্মী জুলফিকার আলী খালাফের গুলিবিদ্ধ(n)= -1 হওয়ার ঘটনা খানায় ফোন করে জানান।) $[-1]$ (এরপর পুলিশ আহত (n)= -1 অবস্থায় তাঁকে উদ্ধার (p)= 1 করে ইউনাইটেড হাসপাতালে নিয়ে যায় ।) $[0]$ (সেখানে চিকিৎসাধীন অবস্থায় ভোরে তাঁর মৃত্যু(n)= -1 হয়।) $[-1]$ (পরে সৌদি দূতাবাসের কর্মকর্তারা আজ সকালে হাসপাতালে গিয়ে তাঁকে সনাক্ত (p)= 1 করে।) $[1]$

খালাফ গুলশানের ১২০ নম্বর সড়কের ২২/এ নম্বর বাড়ির চার তলায় একাই থাকতেন। ওই ভবনের নিরাপত্তা(কর্মী তাপস রেমা জানান, গতকাল রাত সাড়ে ১১টার দিকে খালাফ বাসা থেকে বের হন। এরপর তিনি আর ফিরে আসেননি। তিনি বলেন, খালাফ প্রায়সময়ই নৈশভ্রমণে বের হতেন। আবার রাতেই ফিরে আসতেন।

পুলিশ জানিয়েছে, মৃতদেহ ইউনাইটেড হাসপাতালের হিমঘরে রাখা হয়েছে। পুলিশের গুলশান অঞ্চলের উপকমিশনার লুৎফুল কবিরও বিষয়টি নিশ্চিত করেছেন ।

The number of positive sentence =3

The number of negative sentence = 5

Total weight of the article = (The number of positive sentences -The number of negative

Sentences)

= 3 – 5 = -2 (Negative article)

Table 3.2: List of positive and negative words from the above articles

Positive(ইতিবাচক)(1)	Positive(ইতিবাচক)(1)	Negative(নেতিবাচক)(-1)	Negative(নেতিবাচক)(-1)
<p>প্রসার, মজুরি, পরিমাণ, মুনাফা, সুযোগ, উত্পাদন, খোঁজ, সম্মান, বিবেচনা, প্রকাশ, বিষয়, আশ্বাস, চুক্তিবন্ধ, পরিবেশ, গুরুত্ব, নীতিমালা, তদন্ত, সামগ্রিক</p>	<p>ব্যবস্থা, নিযুক্ত, নিরীক্ষণ, খোঁজখবর, প্রমাণ, স্বীকার, সমাধান, বড়, অনেক, দীর্ঘ, দ্রুততম, বিপুল, প্রতীয়মান, অবশ্যই, শিগগির, নিয়মিত, দ্রুত, সুযোগ, সুবিধা, উদ্ধার</p>	<p>বঞ্চনা, শিকার, নির্যাতন, অভিযোগ, মারধর, ধাক্কা, বাজে, জড়িত, বেআইনি, বাধ্য, থাপ্পড়, মার, লাথি, দুর্বৃত্ত, গুলি, নিহত</p>	<p>গালি, দুর্দশা, চড়, বরখাস্ত, হুমকি, দখল, সমস্যা, চরম, ন্যূনতম, কম, অকথ্য, গুলিবিদ্ধ, মৃত্যু, আহত</p>

Chapter 4

Experiment

We have proposed some rules for finding the weight of the article. Now we will show our experimenting result. The above mentioned rules are applied here to find the trend. For our experiment purpose, we take the articles of the 09th September, 2012 from the prothom alo website. Here we take the articles of the selected category that we have mentioned earlier.

We follow a step by step process for our experiment. This includes –

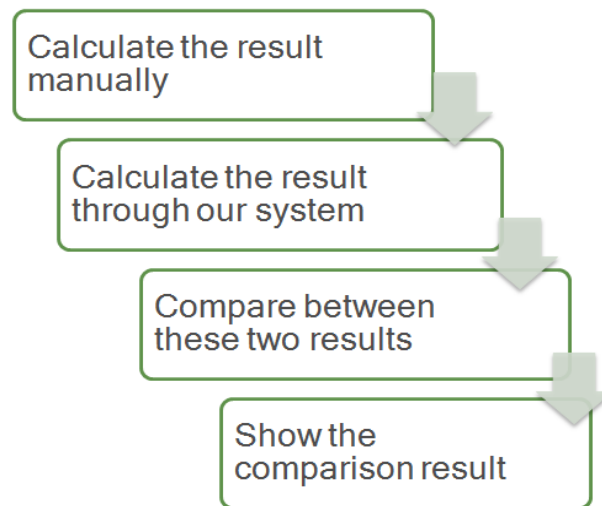


Figure 4.1: Step by step process for experiment and evaluation

We have collected a total of 112 articles from 5 categories of the newspaper. Where 5 articles are from “প্রথমপাতা”, 27 articles are from “সারাদেশ”, 62 from “বিশাল বাংলা”, 11 from “রাজধানী” and 7 from “অর্থ-বাণিজ্য”.

4.1 Manual Result

We gather the information about the collected articles manually. We go through the articles like we read the paper and assign the trend i.e. positive or negative to each article. These trends of the articles are the actual trend that a general people can find while reading the newspaper.

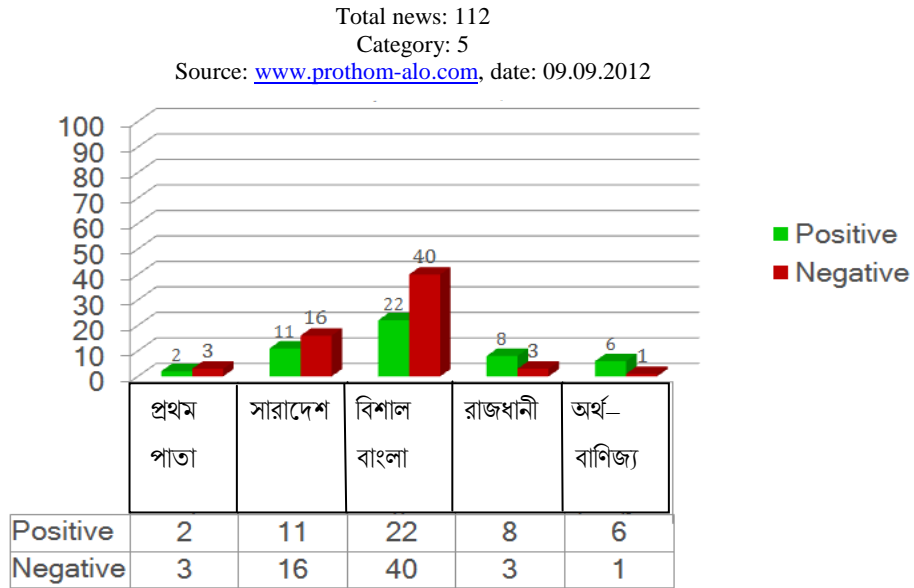


Figure 4.2: Result of Manual approach

4.2 System Result

Here we will analyze the collected data based on our rules and through the system. Earlier we have mentioned some rules. The result of using those rules on the collected data is given below.

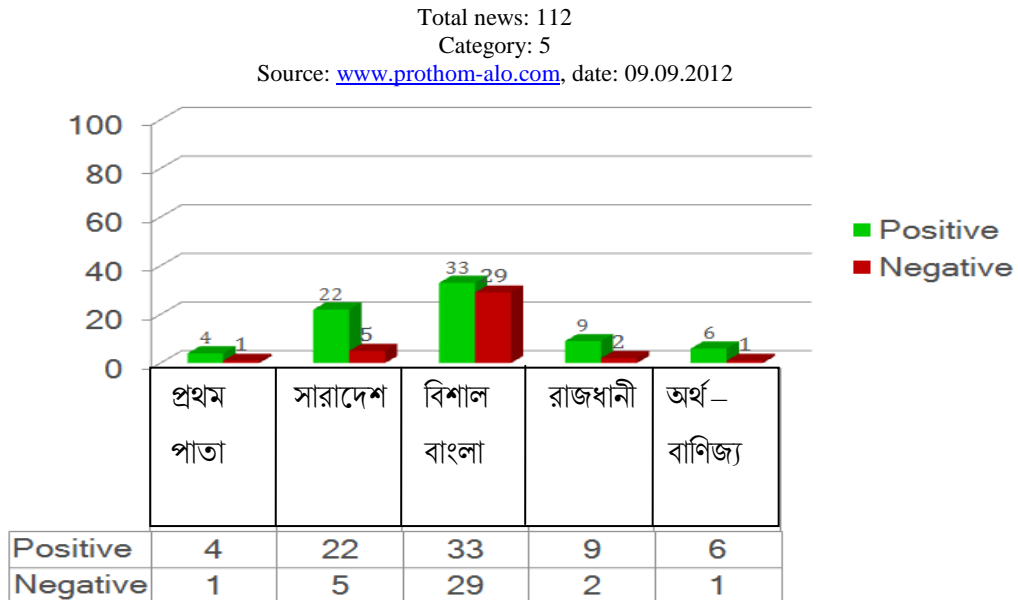


Figure 4.3: Result of the System

For finding the result through our system, we have made an online interface. Where the users have to paste the header and details of the article and the system will generate the result for the article. The screen shot of the online interface is given below.

INSERT ARTICLE HEADER

গাজীপুর কারাগারে পানিসংকট, দুর্ভোগ

INSERT ARTICLE DETAILS

থেকে ট্যাংকারে করে পানি এনে কারাগারের কর্মকর্তা-কর্মচারী ও বন্দীদের পানি সরবরাহ করা হচ্ছে। সূত্র জানায়, কারা তহবিলের টাকা দিয়ে তিনটি অকেজো অগভীর নলকূপ মেরামত করে সাময়িকভাবে সংকট সামাল দেওয়া হচ্ছে। কিন্তু পর্যাপ্ত পরিমাণে পানি না পাওয়ায় কর্মকর্তা-কর্মচারী ও বন্দীরা দুর্ভোগ পোহাচ্ছেন। সূত্র জানায়, কারাগারে ৫০ জনেরও বেশি কর্মকর্তা-কর্মচারী এবং ৮৮০ জন বন্দী রয়েছেন। প্রতিদিন কারাগারে দেড় থেকে দুই লাখ লিটার পানির প্রয়োজন। গাজীপুর গণপূর্ত বিভাগের উপসহকারী প্রকৌশলী মো. মিজানুর রহমান বলেন, কারাগারের আশপাশে বেশ কয়েকটি গভীর নলকূপ স্থাপিত হওয়ায় পানির স্তর নেমে গেছে। ফলে কারাগারের নলকূপ থেকে পানি উঠছে না। তিনি জানান, গভীর নলকূপটি ১৪০ ফুট গভীরে স্থাপন করা হয়েছিল। এটা প্রায় ৩৫০ ফুট গভীরে স্থাপন করা দরকার। গণপূর্ত বিভাগ সূত্র জানায়, নলকূপ মেরামতের জন্য কারা অধিদপ্তর পাঁচ লাখ টাকা বরাদ্দ দিয়েছে। অর্ধ পাওয়া সাপেক্ষে রোববারের (আজ) মধ্যে নলকূপটি মেরামত করে পানিসংকটের স্থায়ী সমাধান করা সম্ভব হবে বলে আশা করা হচ্ছে। কারাধ্যক্ষ আমজাদ হোসেন বলেন, পানি না থাকায় কারা কর্মকর্তা-কর্মচারী ও বন্দীদের দুর্ভোগ পোহাতে হচ্ছে। অন্য স্থান থেকে গাড়ি দিয়ে পানি এনে কিছুটা কাজ চালানো হচ্ছে।

The article details is Negative
The article header is Negative
Weights of the Article and Header matches and The article is Negative

Number of Positive Word = 14
Number of Negative Word = 29
Number of Neutral Word = 75

Figure 4.4: Screen shot of the online interface for finding the weight of the article

4.3 Comparison between the Two Results

Now we will compare between the two results. The comparison between them is given below.

Table 4.1: Result table of the system

Category	Positive news	Negative news	Incorrect answer
প্রথম পাতা	4	1	2
সারাদেশ	22	5	8
রাজধানী	9	2	2
বিশাল বাংলা	33	29	18
অর্থ-বাণিজ্য	5	2	2

Table 4.2: Result table for the manual approach

Category	Positive news	Negative news
প্রথম পাতা	3	2
সারাদেশ	11	16
রাজধানী	8	3
বিশাল বাংলা	22	40
অর্থ-বাণিজ্য	6	1

Here we see that, our system result give some wrong output compared to the manual approach. In the next section we will evaluate these results and show the calculation of the success and error rate.

Chapter 5

Evaluation

From the above experiment, we have seen that our proposed rules for finding the weight of the sentence and article generate some errors. We will show the error and success rate of our methods.

- Total articles in five categories = 112
- Correct result = 80
- Incorrect result = 32
- Success rate = 72%
- Error rate = 28%

For 112 articles, our system generates 72% success rate and 28% error rate. For larger amount of data it will generate more success rate. But in case of small amount of data it may fail to generate more accurate result. Why we are using the addition rule here? We have proposed this idea for generating weight or trend of the article because after analyzing of large amount of data we have seen that, in maximum case the positive articles are enriched with positive words and sentences. And in case of negative articles, the number of negative words and sentences are also bulky. Then why we have some errors in our system? There some other problems that we have also faced. These are discussed in the next subsection.

5.1 Some other Problems

There are some other problems we have confronted. Some human centric problems are there. Newspaper article is always based on the current event. There is always a possibility that present positive event or information or any other things may become negative in nature in future. Some news may give one trend generally but based on the current situation that news may give another trend. These situation justifications are not possible in our system.

There is much news in the newspaper which may present different meaning to the different peoples. Like Government news.

These problems are current event related and situation based. So there will always be a problem about finding the trend of these articles. It may happen that some are accepting our system result for particular article but some are not.

Human can think but machine cannot think. There is already a debate about the last sentence among AI experts. To get the fully accurate results, our system needs to be intelligent enough. This is not possible in the current situation. May be in the future it may become possible to make the machine think. Then it will be possible to generate an accurate result.

So those errors are there because of these problems and also for the earlier mentioned problems. Some problems may be solved through one process. We have literally gone through this process but in full measure. That is multiplication rule. This is –

We will add the weight of word excluding weight of one word which may have positive or negative weight.

Now we will multiply the sum of weights with the weight of those words which are excluded.

If there are more excluding words in a sentence then we add among them and then we multiply.

As for Example:-

“বর্তমানে দারিদ্র্য(-1) নেমে(-1) এসেছে সারে একত্রিশ শতাংশে” $[(-1)*(-1)]= 1$, positive

“এমডিজির সব লক্ষ্য(1) অর্জন(1) নাও(-1) হতে পারে” $[(1+1)*(-1)]= -2$, negative

From this multiplication rule we get the idea that, categorization of the words is the main problem here.

Some words which should be used to multiply are given here -

“কমা, বাড়া, হ্রাস, বৃদ্ধি, সাহায্য, সহায়তা, সহযোগিতা, না, নামা, পিছিয়ে, বাধাগ্রস্ত, উর্ধ্বগতি, নিম্নগতি” etc.

This is somewhat a theoretical idea. In future, its use can be extended through the proper categorization of the bangla words that should be used in multiplication.

Chapter 6

Discussion and Conclusion

We can say that finding the trend of the particular sentence is not too easy. We try to gather the ambiguities through our analysis and try to find a way to solve those ambiguities and problems. If it is possible to categorize the words completely and find some ways to solve those above mentioned problems and ambiguities, then we will be able to generate more accurate result.

Bangla language is a very old language. To work with it is pretty challenging. We have tried our best to give some rules to successfully find the trend of the sentences and articles. Though it is not 100% correct, in the future we will try to gather more features that can help us to find more correct result.

Another approach can be taken in this purpose. This is machine learning and reinforcement algorithm.

References

- *An Approach for Building Bangla News Corpus and News Trend Analysis by Ali Md. Mahmudul Hasan Sohan, Omar Faruque, Md.Rezowan Sarker .*
- “বাংলা ভাষা ও সাহিত্য– ড. সৌমিত্র শেখর”
- “বাংলা ব্যাকরণ ও রচনা- দেলওয়ার মফিজ”
- বাংলা অভিধান- বাংলা একাডেমী
- Sailendra Biswas- সংসদ বাংলা অভিধান <http://dsal.uchicago.edu/dictionaries/biswas-bangala/>

Appendix

Database Connection

```
<?php
$db = mysql_connect('localhost','<database name>','<password>');
if(!$db)
{
die("not connected : " . mysql_error());
}
$db_selected = mysql_select_db("<select database>", $db);
if(!$db_selected)
{
die("can't connect:" . mysql_error());
}
mysql_query('SET CHARACTER SET utf8');
mysql_query("SET SESSION collation_connection = 'utf8_general_ci'");
?>
```

Data Collection

Here we are using “simple_html_dom.php” a web crawler program to collect the articles from the online newspaper. We are including this file.

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>Untitled Document</title>
</head>
<?php
include_once('simple_html_dom.php');
include("connect.php");
$ar = array(0=>2,1=>90,2=>51,3=>56,4=>50);
//echo $ar[0]." ".$ar[1];
$k = 0;
$dt = date("Y-m-d");
echo $dt;
$html = new simple_html_dom();
$html = new simple_html_dom();
```

```

        set_time_limit(0);
while($k < 5){
    $cat = $ar[$k];
$target_url = "http://www.prothom-alo.com/section/date/" . $dt . "/category/" . $ar[$k];
// $html = new simple_html_dom();
$html->load_file($target_url);

foreach($html->find('a[class=blackAnchor]') as $link)
{
    // $t = $link->href;
    // $d = strchr($t, "/detail/date/" . $dt . "/news/");
    $d = $link->href;
    if($d)
    {
        $d = "http://www.prothom-alo.com" . $d;
        $htm->load_file($d);

        foreach($htm->find('h2[class=title mb10]') as $title)
        {
            $t = $title->plaintext;
            // echo $t . "<br />";
        }

        foreach($htm->find('p[id=content]') as $link)
        {
            $det = $link->plaintext;
        }
        $qrun = mysql_query("INSERT INTO `data_collection`(`date`,
`category`, `heading`, `details`) VALUES ('$dt','$cat','$t','$det')");
        if($qrun) echo "Data Collected" . "<br />";
    }
}
}
}
?>
<body>
</body>
</html>

```

Trend Analysis

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>Untitled Document</title>
</head>

<body>
<table width="70%" border="1" align="center" style="margin-top:10px; margin-
bottom:10px" cellspacing="0" cellpadding="2">
<?php
include("connect.php");
function str_replace($search, $replace, $subject)
{
    $pos = strrpos($subject, $search);

    if($pos !== false)
    {
        $subject = substr_replace($subject, $replace, $pos, strlen($search));
    }

    return $subject;
}

$ar = array(0=>2,1=>90,2=>51,3=>56,4=>50);
$arr =
array(0=>'ের',1=>'ে',2=>'ও',3=>'গুলো',4=>'গুলি',5=>'গণ',7=>'কে',8=>'র',9=>'রও',10
=>'ভাবে',11=>'ভাবেও',12=>'ই',13=>'তে',14=>'য়',15=>'তর);
$a = $ar[0];
$dt = "2012-09-09";
$query = mysql_query("SELECT * FROM `data_collection` WHERE category = '51' AND date
= '$dt'");
while($res = mysql_fetch_array($query))
{
    $header = $res['heading'];
    if($header == "সংক্ষেপে সারা দেশ" || $header == "শোক" || $header == "মৃত্যুবার্ষিকী" ||
$header == "আপনার রাশিফল" || $header == "চারা বিতরণ" || $header == "নতুন কমিটি" ||
$header == "একনজরে"){break;}
    echo " <tr>
```

```

<td align="center"><font size="+2">".$res['heading']."</font></td>";
//$header = $res['heading'];
$header = explode(" ", $header);
$res1 = count($header)-1;
$j = 0;
$weight = 0;
while($res1 >= 0)
{
    //-----
    $wo = $header[$j];
    set_time_limit(0);
    $q = mysql_query("SELECT * FROM bwords WHERE trend_1 != '' and
bword = '$wo' GROUP BY bword");
    $numrow = mysql_num_rows($q);
    if($numrow == 0)
    {
        $flag = 0;
        while($flag < 16)
        {
            $wd = str_replace($arr[$flag], " ", $wo);
            if($wd != $wo)
            {
                $q = mysql_query("SELECT * FROM bwords WHERE
trend_1 != '' and bword = '$wd' GROUP BY bword");
                $numrow = mysql_num_rows($q);
                if($numrow != 0)
                {
                    while($r = mysql_fetch_array($q))
                    {
                        if($r['trend_1'] == 'positive')$weight++;
                        else if($r['trend_1'] == 'negative')$weight--;
                    }
                }
            }
            $flag++;
        }
    }
}
else
{
    while($r = mysql_fetch_array($q))
    {
        //echo $r['bword']."<br>";
        if($r['trend_1'] == 'positive')$weight++;
        else if($r['trend_1'] == 'negative')$weight--;
    }
}
}

```

```

    }
    //-----
    $res1--;
    $j++;
}
$sen = explode(" /", $res['details']);
$result = count($sen)-1;
$i = 0;
$weight_article = 0;
while($result >= 0)
{
    $word = explode(" ", $sen[$i]);
    $result1 = count($word)-1;
    $j = 0;
    $weight_sentence = 0;
    while($result1 >= 0)
    {
        $wo = $word[$j];
        set_time_limit(0);
        $q = mysql_query("SELECT * FROM bwords WHERE trend_1 != '' and
bword = '$wo' GROUP BY bword");
        $numrow = mysql_num_rows($q);
        if($numrow == 0)
        {
            $flag = 0;
            while($flag < 16)
            {
                $wd = str_ireplace($ar[$flag], $wo);
                if($wd != $wo)
                {
                    $q = mysql_query("SELECT * FROM bwords WHERE
trend_1 = '' and bword = '$wd' GROUP BY bword");
                    $numrow = mysql_num_rows($q);
                    if($numrow != 0)
                    {
                        while($r = mysql_fetch_array($q))
                        {
                            if($r['trend_1'] ==
'positive')$weight_sentence++;
                            else if($r['trend_1'] ==
'negative')$weight_sentence--;
                        }
                    }
                }
            }
        }
    }
}

```

```

        $flag++;
    }
else
{
    while($r = mysql_fetch_array($q))
    {
        if($r['trend_1'] == 'positive')$weight_sentence++;
        else if($r['trend_1'] == 'negative')$weight_sentence--;
    }
}
//-----
$result1--;
$j++;
}
$weight_article = $weight_article + $weight_sentence;
//echo $weight_article."<br />";
$result--;
$i++;
}
if($weight_article >= 0 && $weight >= 0 ) echo "<td align=\"center\"><font
size=\"+2\">Positive</font></td></tr>";
else if($weight_article < 0 && $weight < 0)echo "<td align=\"center\"><font
size=\"+2\">Negative</font></td></tr>";
else if($weight_article >= 0)echo "<td align=\"center\"><font
size=\"+2\">Positive</font></td></tr>";
else if($weight_article < 0)echo "<td align=\"center\"><font
size=\"+2\">Negative</font></td></tr>";
}
?>
</table>
</body>
</html>

```