Safe News for Kids

Devopriya Tirtho

160204033@aust.edu

Shafin Rahman

160204040@aust.edu

Introduction: The term "**news**" refers to facts regarding **current affairs**. This can be done by several means, including **word - of - mouth**, **publishing**, **mail services**, **television**, **electronic correspondence**, and spectators' witness. **Battle**, **democracy**, **politics**, **education**, **wellness**, **the climate**, **the economy**, **industry**, **culture**, and **entertainment** are all popular news coverage subjects, as are **sporting** activities and quirky or uncommon events. Since ancient times, government declarations surrounding **royal rituals**, **rules**, **taxation**, **human health**, and **terrorists** have been labelled news.

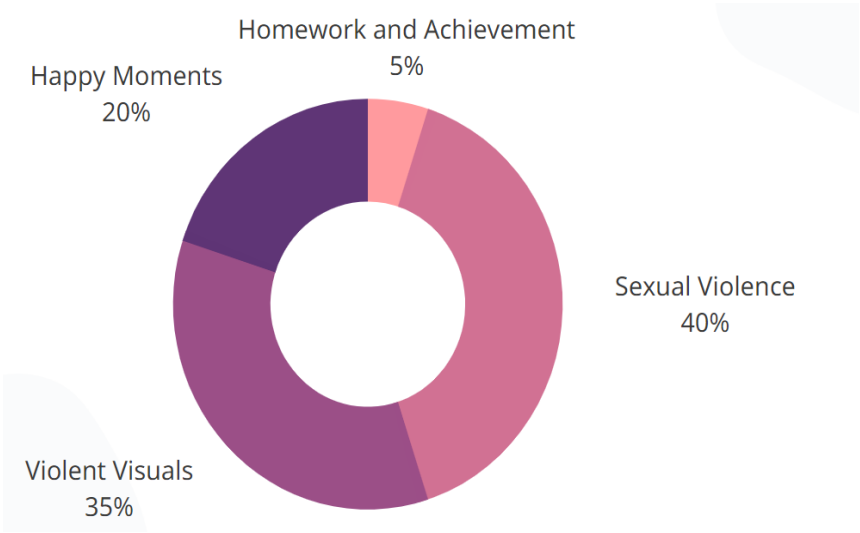
From the definition of a **newspaper, it can be stated that** a **newspaper** is a daily or monthly newsletter that provides **written** information on national affairs that is usually typed in black ink on a whitish backdrop. Newspapers can address a broad range of topics, including **politics**, **economy**, **sports**, and **entertainment**. They also reported significantly, including **opinion articles**, **weather reports**, **reviews of local programs**, **biographies**, **political cartoons**, and **discussion forums**.

Newspapers contain news that helps us learn about our surroundings and get an intuition to act based on the happenings. The impact of **information** and **newspapers** is immense in our day-to-day life.

From **children** to **old-aged** people, everyone read the newspaper. The number of **child-readers** is not less. This is very obvious that every parent wants their child to be proactive, to be

concerned about the world whereas, it is also true that the desirable types of news are not continuously published. Parents are worried about their children that their children should avoid the information. For printed newspapers, it is not always possible to censor news topics for their children. For the online newspaper, it is quite possible to impose a **filtered** news representation or **parental control** feature to get their children away from any **abusive**, **toxic** news that may detrimentally affect their minds.

As **computer science** is developing day-by-day, it is not fair to push our child into a dangerous situation. Children tend to think a lot; they try to visualise what they perceive. Here is a graph of the factors which impact most a child's brain.



Here, we can see that the most affecting factors are the terrible things that impact a child's mind awfully. While happy moments and achievements in school or off school have an impact, **sexual molestation** and **violent visuals** are **numerous**.

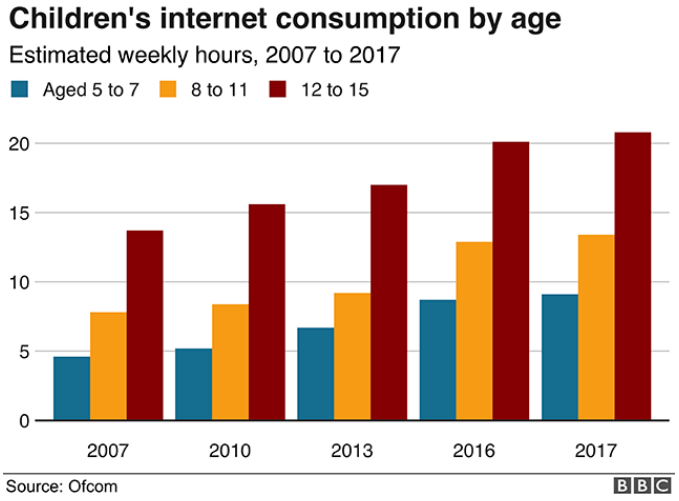
Our work is to make a safer world for children by imposing a **filtered** or **parental control-based online newspaper system**. Whenever a child searches for reading news, the newspaper itself wants to learn about the reader's age and make a **personalised** reading environment for everyone. The filtering will be imposed for **every age-group**, but, firstly, we work on **children** as they are the most sensitive. Our work is to **filter toxic and abusive news** from an online newspaper so that a child can read news freely, and their state of mind will not be affected. Most importantly, we start our work with **Bengali** newspapers to be imposed for every language in the future. Our main goal is to **automate** the filtering process. Whenever news is published, it will

be **categorised** automatically according to the **age-group**. Our work can be implemented in online newspapers, and for the children, we will make an abusive-content-free reading environment.

Motivation: For every work, we need the **enforcing** factor, which pushes our mind to do the task. For doing this project, we also feel the urge and dive into making this happen. The triggering factors which help us to think about this problem are listed below:

- Firstly, the present world is evolving so quickly. Everyone needs to be updated. From this list of people, we cannot throw out the participation of children. They are the future of a nation and the world. For making a better world, we need to fetch the potential out of a child. A child's mind is full of imagination, and what a child sees and hears, he/she tries to regenerate the whole scenario in his/her mind with his/her vision. To trigger the imagination, we need to present them in a **nonabusive** and **comfortable** environment. The first **motivating factor** to work for **children** is the thinking to make their world better and see a better world tomorrow.

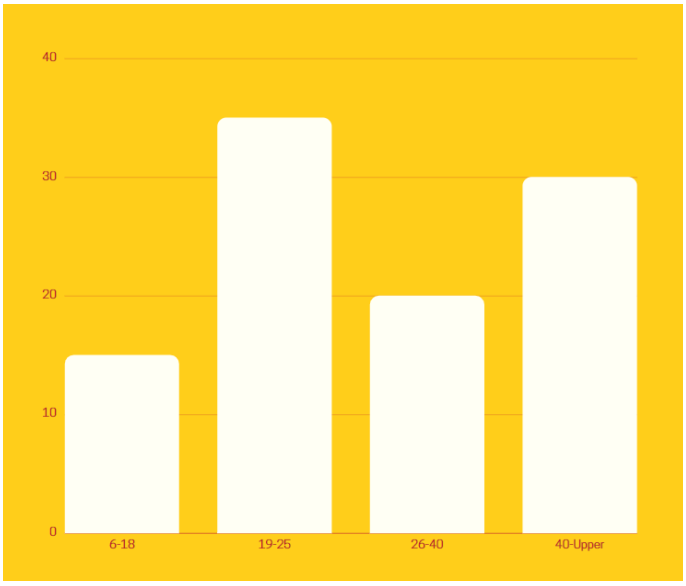
- Secondly, we focused on data and thought about the sector which should be picked. Then, we discovered a **surprising** factor that blew our minds.



This is the rise of **internet consumption** of children from the year **2007 to 2017**. At present, a child, on average, spends **more than 10 hours** on the internet weekly. When we look deep into the fact, children spend most of their time playing games and reading

blogs and newspaper posts. These statistics help us to choose a problem which involves their daily activities on the internet.

- Thirdly, there comes a question about which **sector** we should pick to work in. After studying the number of readers of different age groups, we have found that the number of child newspaper readers is more than 10%, and most of them read online newspapers. Here is an illustration of the readers of the newspaper.



- After studying the children's mental health, we have summarised some of the ideas. Our

primary focus is to work for kids of the **age group 6 to 18**. Various studies show that

children need **affirmation**. They are very naive about toxic behaviours and events. They tend to **flourish** when they encounter **positive** events **happening** around them. They need a platform from which they can find their **interests** and be updated about **current affairs**. As some social media do not let children open accounts, many tend to read **blogs** and online **newspaper posts**. That is the fact which influences us to choose the topic. As kids need a **healthy**, **positive** environment wherever they are and spend a lot of time on the **internet**, we can work on **filtering** news for them so that toxic and **abusive** contents may not **influence** their **gentle minds**.

- For nurturing a kid's health, we need to make a **proper** environment everywhere. It is the duty of every parent. As a **computer engineer**, we feel **responsible** for creating an **appropriate** environment for children on the internet by making relevant applications and software. As many works are going on the visuals, we try to work on texts, **especially** on **Bengali** text.

- Our project focuses on people **aged 6 to 18**. The people of this group are **compassionate** and always need **supervision** for doing any task. As the world is becoming more and more active, every parent does not have enough time to make proper supervision. From the perspective of **social significance**, our project helps those kids who spend time reading online posts. Our project will create a **suitable** environment so that any embarrassing news does not come up to them. This may help them be **free of toxicity** and find more **interest in reading newspapers**.

- Our project can be implemented as an **application** or **filtering software** used by **newspaper agencies** to publish their news online. According to our project, the information, which includes **toxic and abusive events** and **words**, will be **filtered out**. When a person enters a website, he/she will be asked to give input on their age, and according to their age, the news feed will be classified so that no one will encounter any embarrassing news which may affect their mental health. Another **application** of our project can be imposed **externally** as **parental control software**. When someone turns on the software, it automatically filtered out toxic news from the website and only

showed the reader's filtered information. So our project can be acquired by newspaper agencies and can be implemented as external software.

Challenges: From defining the task to completing it, we faced some **difficulties**. Here we are

illustrating the challenges we faced.

- Firstly, our project is based on the **Bengali language**, we have searched on the web for **Bengali newspaper data**. The dataset we were looking for, which should be **clean** and **labelled**, could not find. In this link Prothom Alo [2013 - 2019], we have found a dataset of the newspaper of **'The Prothom Alo'**.

- After finding the dataset, we have faced the issue of making it **proper** for our **project**. That was an **extensive** dataset of about **two lakhs ninety thousand** rows where many rows contain **'null'** values. We needed to process the dataset for making it **suitable** for our project.

- Then, the main task emerged as the dataset was not classified according to our problem. As it is a **deep learning problem**, **gold data** is required. For making the dataset **gold**, we, along with one of my cousins, make **proper labelling** of about **seven thousand data**.
- Then an issue occurred as we worked with the **Bengali language**. When we tried to **tokenize** the newspaper contents without the pre-defined **bnltk tokenizers**, we observed that the **Bengali punctuation** marks are omitted. The actual Bengali word could not be retrieved. So, we need to handle the problem by using a **lambda** function. Then, another problem occurred. The line ending symbol in Bengali, which is **‘ ।’ (**দাড়ি**)** could not be split. After handling these issues, we made the **vocabulary appropriately**.

- After finalizing everything, we were unsure about which method to choose. We choose to work with ‘**Bag-of-Words**’ with **count-based** representation. We have also worked with **‘TF-IDF’**, which resulted in **very bad results**. We will talk about the models in the later part of the report.

Related Works: Before starting working on our project, we have studied some of the **similar** projects which have been done before. Here are some related project works which have been done:

- **Sentimental Analysis(Bengali)- (Roy 2017)**
- **Bangla (Bengali) sentiment analysis classification benchmark dataset corpus- (Sazzed 2021)**
- **Sentiment analysis in Bengali via transfer learning using multi-lingual BERT- (Islam 2020)**
- **Development of a Bangla news classification system- (Sirajus 2019)**
- **Child-friendly news- (NewsWise 2021)**

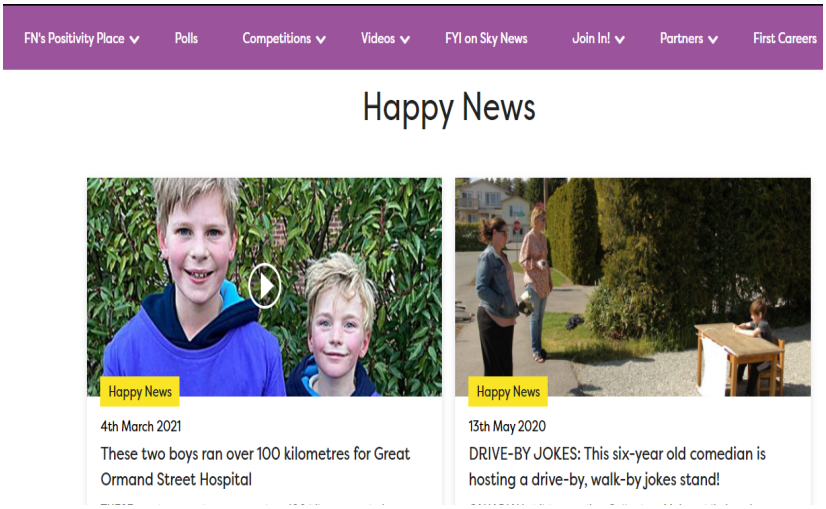
About the Works: The previous works we have mentioned have been done before, which helped us to think about our project more and find some **distinctive** features out of works. Here, we are trying to **represent** the mentioned in short and present their **short-comings**.

- After studying the paper **'Sentimental Analysis(Bengali)**' (Roy 2017), we have got a firm idea about **Bengali sentiment analysis** through **natural language processing**. They worked on **Bengali tweets** and made a simple project of analyzing the **sentiments** of the tweets as **positive** and **negative**. They used **HMM** for **POS tagging** and an **SVM classifier** for classifying the sentiments. In their work, they did not impose **deep learning** for classifying the **sentiments** and the sentiments are classified into only positives and negatives.
- The second work is **not a project**. This is a **dataset** about **Bengali corpus** based on the **youtube comments** on Bengali drama, which can be used further for **hate comment classification** and omission.
- The next work named '**Sentiment analysis in Bengali via transfer learning using multi-lingual BERT**' uses **transfer learning** for **sentiment analysis**. They have worked on **3 classes** of sentiments. Their model gives **71% accuracy** for **two** classes while the model shows **60% accuracy** for **three** classes' classification. Their work is done by using the dataset which is **manually labeled**. It may **differ** from **others' perspectives**.
- This work has the most similarity with our proposed work. The article named '**Development of a Bangla news classification system**' was written according to the work of classifying **Bengali news**. They have proposed three terms, such as:
    - **Filtering According to Collaboration**
    - **Content-based Filtering**
    - **Subscription-based Personalization**

  Among these, they have chosen **'Content-Based Filtering'** to work on. But in our opinion Their work **lacked** proper **implementation** as they have tried to make a **web application** where users can search and read **category-based** news. They worked with **n-gram** based text with **various classifiers** to achieve **greater accuracy**. As they proposed to make a **'Content-based** Filtering' on a news article, they end up making a web application where the **categorical** news is presented.
- The next article is based on the listing of newspapers that are **child-friendly**. The article is named **'Child-friendly news'**. In this article of **'The Guardian'**, the writer enlightens us with a list of papers which are child-friendly. The list of some of the papers are here:
    - **BBC Newsround**

- This newspaper publishes news about **international**, **national** and **political** aspects and also publishes **inspiring** stories about children.

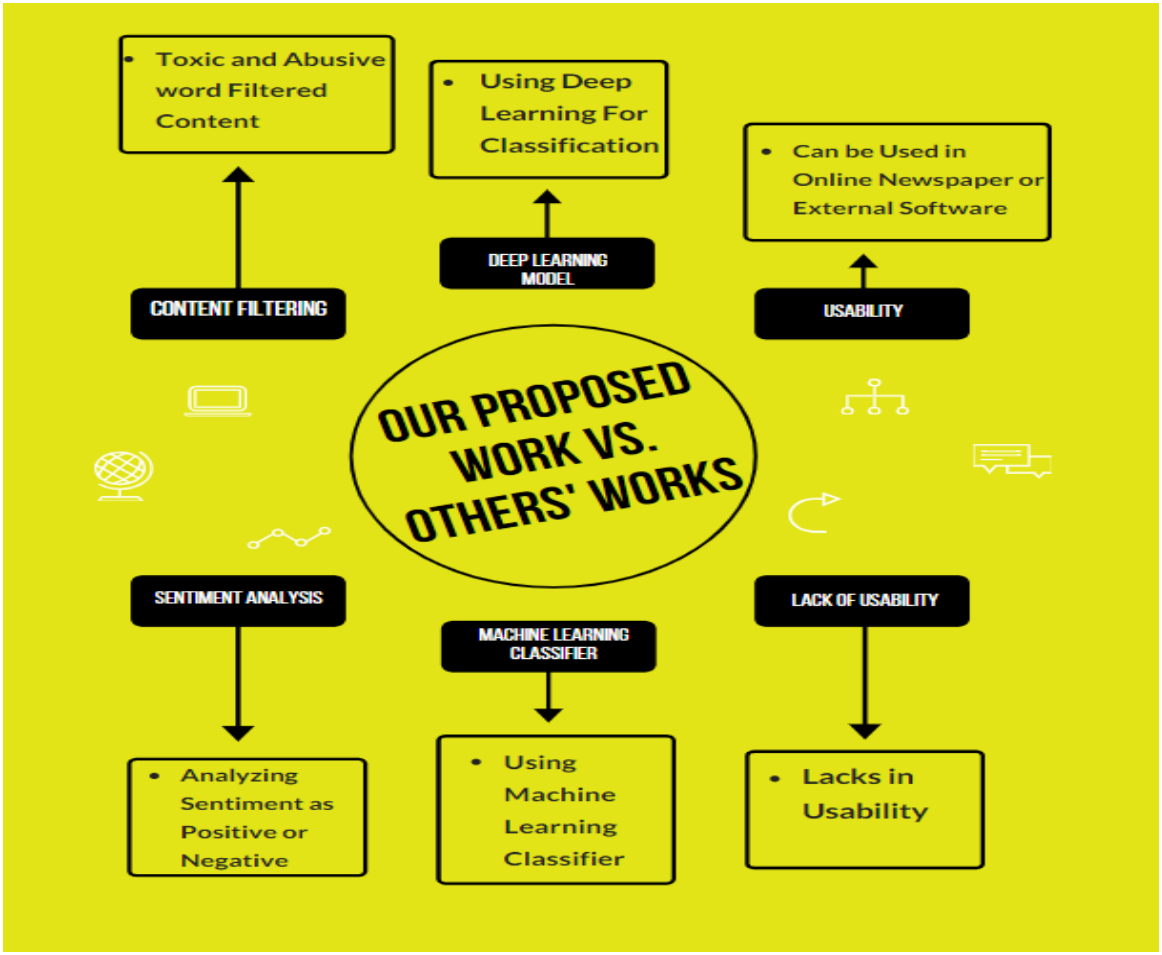
- **The Day**
  - This online newspaper is a **subscription-based** paper where they post writings about current affairs.
- **Dogo News**
  - The news is written here from an **international** perspective.
- **First News Live**
  - This is a great source of **positive** news sites only for children.



- **Inside Science**
  - The newspaper posts about **scientific** events
- **News for Kids**
  - It publishes regular **crackdowns** of **public affairs** written especially for youngsters, with meanings of **new vocabulary** The bulk of the stories come from the United States, but others are from around the world. Access is free with advertising, but an **ad-free membership** is available for a charge.
- **Space Scoop**
  - **Space Scoop** is a **science** newspaper website for children between the ages of 8 and up that features frequent articles from a range of external scientific institutions. Podcasts are indeed available.
- **Twinkl NewsRoom**

- Every day, **Twinkl NewsRoom** publishes a **primary-age-appropriate** news report and activity.

Proposed Work: We proposed a project which helps the children by **filtering out** the **toxic** and **abusive** contents out of the newspaper so that their minds do not be affected badly. Here is an illustration of our proposed work and the missing terms in others' works:

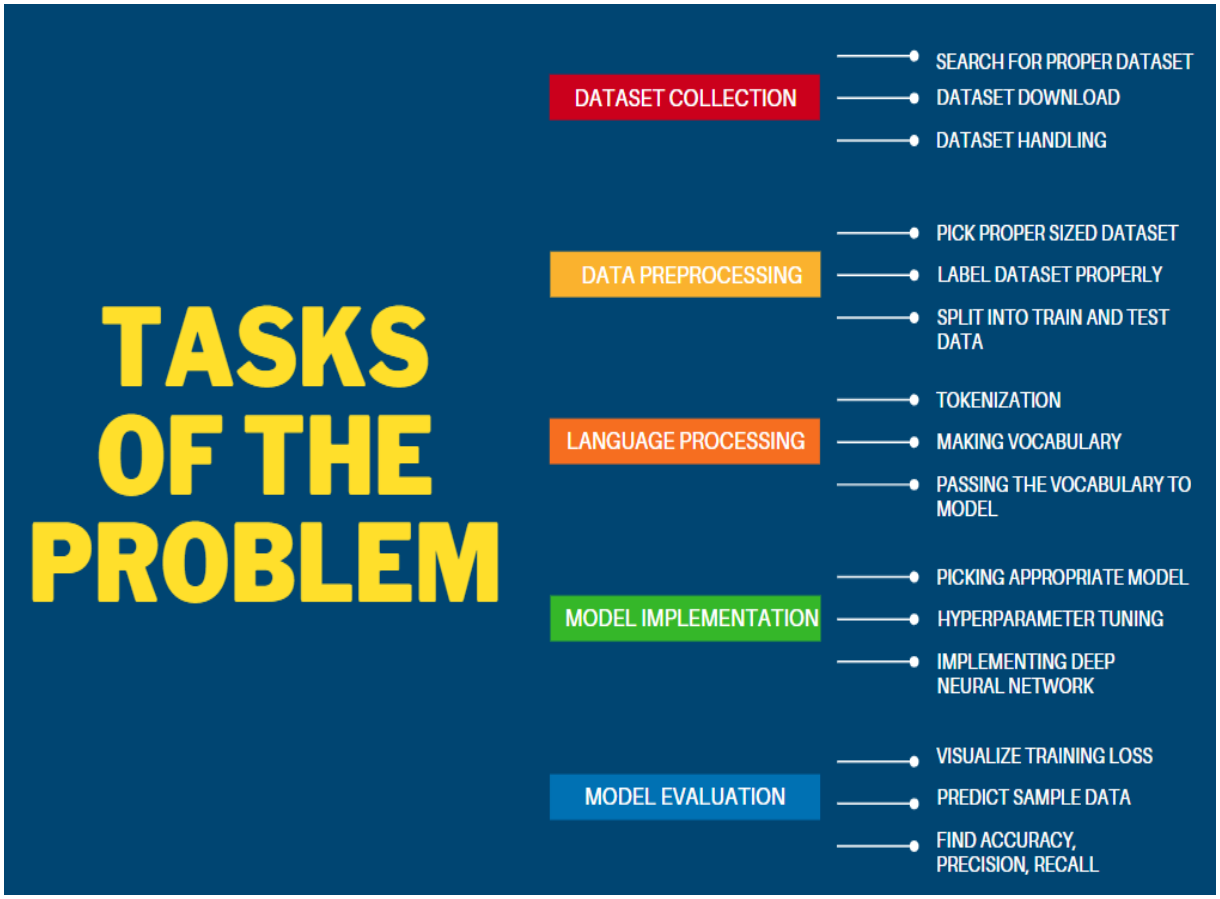


The things we will introduce in our project are:

- Firstly, we work on **Bengali language**. The number of works with **Bengali language** is so **little**
- We will work on a **mode** of **sentiment analysis** which implements the filtering of **toxic** and **abusive content-free** news while others work with comments or social media post to classify the sentiments

- Our work is done through **deep learning algorithm** while others work with machine learning algorithms
- The usability of our work is for **online newspapers** and also for **software systems** while others' work has no greater usability for children

Project Objective: The project we are going to **implement** is divided into some **tasks** and the tasks are also divided into some **subtasks**. Here is the illustration:



We have divided our project into five tasks. These are:

- **Data Collection:** In this process, the searching of a **proper** dataset and collecting the data from an **authorized** source is included.
- **Data Preprocessing:** We tried to pick the **proper** size of the dataset and handle issues of the dataset in this phase. The **splitting** into **train** and **test** portions was also done here.
- **Language Processing:** This process included the **tokenization** of **Bengali corpus** and making the proper **vocabulary** with proper **stemming** and **tokenization**.

- **Model Implementation:** The next phase consisted with picking the **appropriate model**, tuning the **hyperparameters** and implementation of the **deep neural networks**.
- **Model Evaluation:** The last task was about evaluating the model with proper **metrics**.

Dummy Evaluation: Here, we can **visualize** the actual implementation of our **project**. Firstly, we think of a **Bengali newspaper** where various news articles are written. The illustration is given below:

জুন

সংবাদপত্র

নারায়ণগঞ্জে সংঘর্ষ: বিজিবির আহত সদস্যের মৃত্যু

10 June 2020

নারায়ণগঞ্জে ঢাকা-চট্টগ্রাম মহাসড়কের বিভিন্ন অংশে গতকাল সোমবারের সহিংসতার ঘটনায় বিজিবির আহত ওই সদস্যের মৃত্যু হয়েছে। জওয়ান লাভলু আজ মঙ্গলবার ঢাকা মেডিকেল কলেজ হাসপাতালের নিবিড় পরিচর্যাকেন্দ্রে চিকিৎসাধীন অবস্থায় মারা যান। এ নিয়ে ওই সংঘর্ষে বিজিবির দুই ও পুলিশের দুই সদস্যসহ ২১ জন নিহত হলেন। ঢাকা মেডিকেল কলেজ হাসপাতালের পুলিশ ক্যাম্পের দায়িত্বে থাকা উপপরিদর্শক (এসআই) গাজী মোজাম্মেল হক প্রথম আলো ডটকমকে লাভলুর মৃত্যুর ঘটনাটি নিশ্চিত করেছেন। রাজধানীতে ধ্বংসযজ্ঞ চালানোর পরদিন গতকাল ঢাকা-চট্টগ্রাম মহাসড়কের নারায়ণগঞ্জ অংশের বিভিন্ন এলাকায় তাণ্ডব চালায় হেফাজতে ইসলাম। তারা গাছ ফেলে সড়ক অবরোধ করে এবং হাইওয়ে পুলিশের ফাঁড়িসহ বিভিন্ন স্থাপনা ও সড়কে আগুন দেয়। গতকাল ভোর থেকে দুপুর পর্যন্ত নারায়ণগঞ্জ সদর উপজেলার সিদ্ধিরগঞ্জ, সাইনবোর্ড থেকে সানারপাড়-শিমরাইল হয়ে সোনারগাঁ উপজেলার কাঁচপুর পর্যন্ত পাঁচ কিলোমিটার এলাকায় সংঘর্ষের ঘটনা ঘটে। ফলে বেলা তিনটা পর্যন্ত ঢাকা-চট্টগ্রাম ও ঢাকা-সিলেট মহাসড়কে যান চলাচল বন্ধ থাকে।

ট্রাকের চাপায় স্কুলছাত্রসহ দুজন নিহত

10 June 2020

পাবনা সদর উপজেলায় গতকাল মঙ্গলবার ট্রাকের চাপায় আল আমিন (২৫) নামের এক ব্যক্তি নিহত হয়েছেন। এদিকে গত সোমবার রাজশাহীর বাগমারা উপজেলায় হাফিজুর রহমান (১৩) নামের এক স্কুলছাত্র একইভাবে নিহত হয়.....

মাহমুদউল্লাহ-নাসিরের নৈপুণ্যে ২৪৭

সময় অধিনায়ক মুশফিকুরের সংগ্রহ ছিল ৩২ রান। টেস্টাই স্যাতরার শিকারে পরিণত হওয়া উদ্বোধনী ব্যাটসম্যান তামিমের সংগ্রহও ছিল একই। সাকিব আল হাসান আউট হন ১৮ রান করে।

আপনার রাশিফল

৯ মে. ২০১৩ আজকের এই তারিখে জন্মগ্রহণ করায় পাশ্চাত্যমতে আপনি বৃষ রাশির জাতক বা জাতিকা।

সাবিনাদের ড্র

কলকাতায় আইএফএ একাদশের সঙ্গে কাল গোলশূন্য ড্র করেছে বাংলাদেশ।

Here, the illustration is a **dummy** illustration. From the picture, we see that there are various types of news regarding **politics**, **sports**, **horoscope**, **national** etc. For a child we want to filter

the news page and the newspaper will be then **abusive** content free. So, after classifying the news articles the news can be presented like this:

Title	Content	Threat
ধ্বংসস্তূপে আর কত লাশ?	সাভারে ধসে পড়া রানা প্লাজার ধ্বংসস্তূপে আর কত লাশ আছে? গতকাল বুধবার একের পর এক লাশ বেরিয়ে আসতে থাকে এমন কথা শোনা যায় সেনাবাহিনীর একজন উদ্ধারকর্মী মুখে। তিনি বলেন. ধ্বংসস্তূপ টানলেই বেরিয়ে আসছে বিকৃত হয়ে যাওয়া লাশ। লাশগুলো পরিণত হয়েছে প্রায় কঙ্কালে। ঢাকা জেলা প্রশাসনের নিয়ন্ত্রণ কক্ষের তথ্য অনুযায়ী. গত ২৪ এপ্রিল সকালে ভবনটি ধসে পড়ার পর থেকে গতকাল রাত সোয়া ১০টা পর্যন্ত ধ্বংসস্তূপ থেকে উদ্ধার করা হয়েছে ৮০৭টি লাশ। এর মধ্যে গত মঙ্গলবার রাত সোয়া ১০টা থেকে গতকাল রাত সোয়া ১০টা পর্যন্ত উদ্ধার হয় ১০১টি লাশ। হাসপাতালে চিকিৎসাধীন অবস্থায় ১১ জনের মৃত্যু সহ ভবনধসে সব মিলিয়ে এ পর্যন্ত মৃতের সংখ্যা দাঁড়িয়েছে ৮১৮ জন। তথ্য অনুযায়ী. এ পর্যন্ত স্বজনদের কাছে হস্তান্তর করা হয়েছে ৬২৪টি লাশ। হস্তান্তরের অপেক্ষায় সাভার অধরচুদর উচ্চ বিদ্যালয় মাঠে ৩০টি এবং ঢাকা মেডিকেল কলেজ ও স্যার সলিমুল্লাহ (মিটফোর্ড) মেডিকেল কলেজের মর্গে আছে ১০০ টি লাশ।	Yes
নতুন টাচ স্ক্রিন ল্যাপটপ	আসুসের নতুন একটি টাচ স্ক্রিন ল্যাপটপ কম্পিউটার বাজারে এসেছে। ভিভোবুক এস ৩০০ এ মডেলের এই ল্যাপটপে আছে ১৩ দশমিক ৩ ইঞ্চির পর্দা। এর আবরণ তৈরি হয়েছে অ্যালুমিনিয়ামে। এতে রয়েছে ১.৭ গিগাহার্টজ গতির ইন্টেল কোর আই-৫ প্রসেসর. ৪ গিগাবাইট র‍্যাম. ৫০০ গিগাবাইট হার্ডডিস্ক ইত্যাদি। উইন্ডোজ ৮ অপারেটিং সিস্টেমে এটি। গ্লোবাল ব্র্যান্ড (প্রা.) লিমিটেড আনা এ কম্পিউটারের দাম ৬৫ হাজার টাকা। —বিজ্ঞপ্তি	No
আবার শীর্ষস্থান হারালেন সাকিব	আইসিসি সেরা ওয়ানডে অলরাউন্ডার লড়াই এখন এমন. যেন কে কত খারাপ করতে পারে! জানুয়ারিতে সাকিব আল হাসানের শীর্ষস্থান কেড়ে নিয়েছে মোহাম্মদ হাফিজ। কিন্তু দক্ষিণ আফ্রিকা সিরিজে পাকিস্তান অলরাউন্ডারের বাজে পারফরম্যান্সে মার্চে আবার এক নম্বরে উঠে যান সাকিব। এবার জিম্বাবুয়েতে সাকিবের বাজে পারফরমন্যাস শীর্ষে তুলে দিল হাফিজকে। টেস্ট-ওয়ানডে দুটোই এখন দুই নম্বর অলরাউন্ডার সাকিব। ৪২২ রেটিং পয়েন্ট নিয়ে সিরিজ শুরু করেছিলেন সাকিব। প্রথম ওয়ানডেতে ১ রানে আউট আর কোনো উইকেট না পাওয়ায় পয়েন্ট কমে দাঁড়ায় ৩৯৯। পরের ম্যাচে ৩৪ রান করার পর ১ উইকেট নিলে পয়েন্ট কমে	No

	হয় ৩৯৫। শেষ ম্যাচের ১৮ রান ও বিনা উইকেট পয়েন্ট নামিয়ে এনেছে ৩৮৭ তে। হাফিজের পয়েন্ট ৪০০। তিন ম্যাচে ৫৩ রানে ১ উইকেট। তিন ম্যাচের সিরিজ এর চেয়ে বাজে কেটেছে সাকিবের বলতে গেলে একবারই। ২০০৭ সালে শ্রীলঙ্কায় ১ উইকেট নেওয়ার পাশাপাশি রান করতে পেরেছে মাত্র ১৭।	

Based on the classification of the news, the **filtered** online newspaper will look like this:

ছায়ানটে রবীন্দ্র উৎসব

10 June 2020

কবিগুরু রবীন্দ্রনাথ ঠাকুরের ১৫২তম জয়ন্তী ছিল গত বুধবার। এ উপলক্ষে গতকাল বৃহস্পতিবার ছায়ানট সংস্কৃতি-ভবন মিলনায়তনে শুরু হয়েছে দুই দিনের রবীন্দ্র উৎসব। ছায়ানটের শিল্পীদের অংশগ্রহণে নৃত্যগীতি পারবি না কি যোগ দিতে এই ছন্দে যে' পরিবেশনের মধ্য দিয়ে অনুষ্ঠান শুরু হয়। এরপর স্বাগত কথনে অংশ নেন ছায়ানটের সাধারণ সম্পাদক খায়রুল আনাম শাকিল। রবীন্দ্রনাথের জীবন ও কর্ম নিয়ে আলোচনা করেন ছায়ানটের সভাপতি সন্জীদা খাতুন। এরপর শুরু হয় গানের পালা। তে।

'কন্যার কানে দুল আর খোঁপায় সাদা ফুল' নামের এই প্রদর্শনীতে শিল্পকর্ম রয়েছে ৩৬টি। নারীদের নিয়েই কাজ করেছেন শিল্পী। নারীরা এখন সমাজের নানা স্তরে কাজ করছেন। এর মধ্যেও তাঁরা সংসার সামলাচ্ছেন এবং নিজেদের সুন্দর করে উপস্থাপনের জন্য প্রসাধন. রূপচর্চার মতো বিষয়গুলোও অব্যাহত রেখেছেন। শিল্পী তাঁর কাজে নারীর জীবনের এসব দিকই তুলে ধরেছেন। প্রদর্শনী ১৫ মে বেলা ১১টা থেকে সন্ধ্যা সাতটা পর্যন্ত খোলা থাকবে।

ছায়ানটে রবীন্দ্র উৎসব

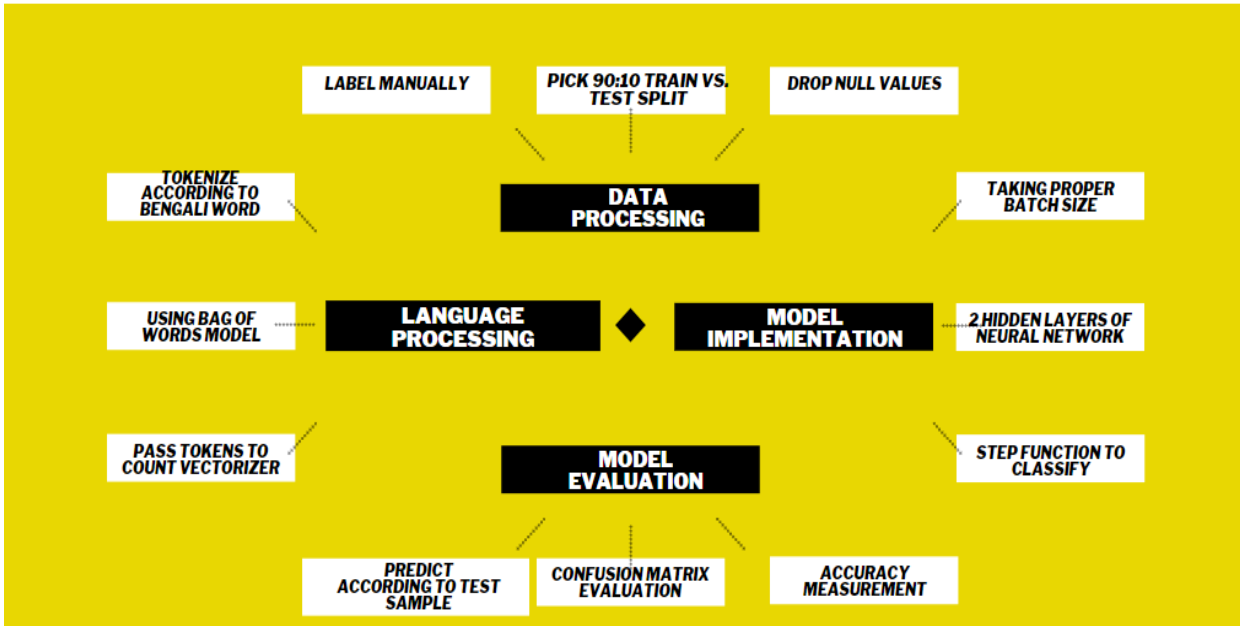
সমঝোতা না হওয়ায় হতাশ যুক্তরাষ্ট্র

10 June 2020

একটি অবাধ, সুষ্ঠু ও বিশ্বাসযোগ্য নির্বাচনের লক্ষ্যে বাংলাদেশের রাজনৈতিক দলগুলোর মধ্যে এখনো কোনো মতৈক্য না হওয়ায় হতাশা জানিয়েছে যুক্তরাষ্ট্র। গত সোমবার ওয়াশিংটনে মার্কিন পররাষ্ট্র দপ্তরের উপমুখপাত্র মেরি হার্ফ নিয়মিত ব্রিফিংয়ে তাঁর দেশের এই হতাশার কথা জানান।ঢাকায় গতকাল মঙ্গলবার দুপুরে বিএনপির চেয়ারপারসন খালেদা জিয়ার সঙ্গে দেখা করার পর মার্কিন রাষ্ট্রদূত ডানল অট্রিট মন্তব্য করেছেন. বিরোধী দলকে স্বাধীনভাবে মতামত প্রকাশের সুযোগ করে দেওয়া সরকারের দায়িত্ব। আর বিরোধী দলের দায়িত্ব হচ্ছে শান্তিপূর্ণভাবে সেই সুযোগ কাজে লাগানো।একই সময়ে যুক্তরাজ্যের হাইকমিশনার রবার্ট শিবসন আওয়ামী লীগের সাধারণ সম্পাদক এবং স্থানীয় সরকার, পল্লী উন্নয়ন ও সমবায়মন্ত্রী সৈয়দ আশরাফুল ইসলামের সঙ্গে দেখা করে রাজনৈতিক সংকট নিরসনে সংলাপের তাগিদ দেন।

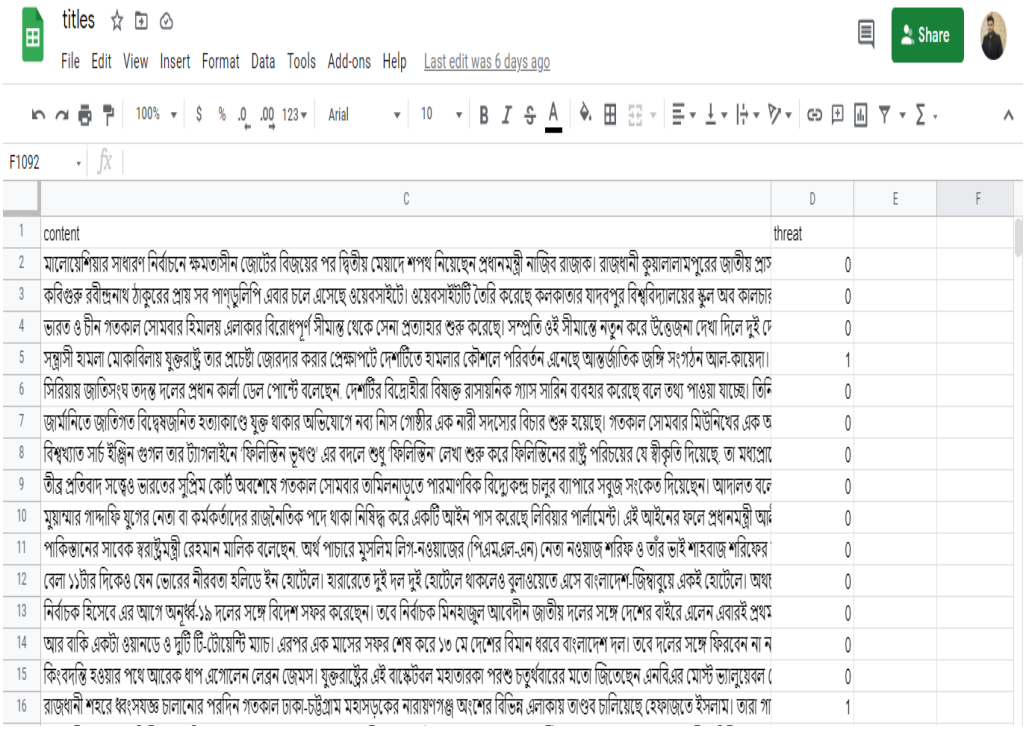
Methodologies: We have divided our problems in **small** groups and tried to solve them one by one. Here is an illustration that how we solved our problem:

METHODOLOGIES



We have divided our whole work into **four subtasks** and then tried to solve these accordingly. The **methodologies** of our project are given below:

- **Data Processing**:
    - In the main dataset, the data were sorted by year. We have taken **one thousand** rows from **each year** and make a **seven thousand** row's dataset.
    - After making a short dataset, we **labeled** each news content according to the **context of threat**. If the news seemed to be **threatful** or **harmful** for kids we assigned **1** otherwise we assigned **0**.



titles

File Edit View Insert Format Data Tools Add-ons Help Last edit was 6 days ago

Share

F1092

	C	D	E	F
1	content	threat		
2	মালেয়েশিয়ার সাধারণ নির্বাচনে ক্ষমতাসীন জোটের বিজয়ের পর দ্বিতীয় মেয়াদে শপথ নিয়েছেন প্রধানমন্ত্রী নাজিব রাজাক। রাজধানী কুয়ালালামপুরের জাতীয় প্রাস	0		
3	কবিগুরু রবীন্দ্রনাথ ঠাকুরের প্রায় সব পাণ্ডুলিপি এবার চলে এসেছে ওয়েবসাইটে। ওয়েবসাইটটি তৈরি করেছে কলকাতার যাদবপুর বিশ্ববিদ্যালয়ের স্কুল অব কালচার	0		
4	ভারত ও চীন গতকাল সোমবার হিমালয় এলাকার বিরোধপূর্ণ সীমান্ত থেকে সেনা প্রত্যাহার শুরু করেছে। সম্প্রতি ওই সীমান্তে নতুন করে উত্তেজনা দেখা দিলে দুই দে	0		
5	সন্ত্রাসী হামলা মোকাবিলায় যুক্তরাষ্ট্র তার প্রচেষ্টা জোরদার করার প্রেক্ষাপটে দেশটিতে হামলার কৌশলে পরিবর্তন এনেছে আন্তর্জাতিক জঙ্গি সংগঠন আল-কায়েদা।	1		
6	সিরিয়ায় জাতিসংঘ তদন্ত দলের প্রধান কার্লা ডেল পোন্টে বলেছেন, দেশটির বিদ্রোহীরা বিষাক্ত রাসায়নিক গ্যাস সারিন ব্যবহার করেছে বলে তথ্য পাওয়া যাচ্ছে। তিনি	0		
7	জার্মানিতে জাতিগত বিদ্বেষজনিত হত্যাকাণ্ডে যুক্ত থাকার অভিযোগে নব্য নাৎসি গোষ্ঠীর এক নারী সদস্যের বিচার শুরু হয়েছে। গতকাল সোমবার মিউনিখের এক ত	0		
8	বিশ্বখ্যাত সার্চ ইঞ্জিন গুগল তার ট্যাগলাইনে 'ফিলিস্তিন ভূখণ্ড' এর বদলে শুধু 'ফিলিস্তিন' লেখা শুরু করে ফিলিস্তিনের রাষ্ট্র পরিচয়ের যে স্বীকৃতি দিয়েছে, তা মধ্যপ্রা	0		
9	তীব্র প্রতিবাদ সত্ত্বেও ভারতের সুপ্রিম কোর্ট অবশেষে গতকাল সোমবার তামিলনাড়ুতে পারমাণবিক বিদ্যুৎকেন্দ্র চালুর ব্যাপারে সবুজ সংকেত দিয়েছেন। আদালত বলে	0		
10	মুয়ান্মার গান্দাফি যুগের নেতা বা কর্মকর্তাদের রাজনৈতিক পদে থাকা নিষিদ্ধ করে একটি আইন পাস করেছে নিবিয়ার পার্লামেন্ট। এই আইনের ফলে প্রধানমন্ত্রী আ	0		
11	পাকিস্তানের সাবেক স্বরাষ্ট্রমন্ত্রী রেহমান মালিক বলেছেন, অর্থ পাচারে মুসলিম লিগ-নওয়াজের (পিএমএল-এন) নেতা নওয়াজ শরিফ ও তাঁর ভাই শাহবাজ শরিফের	0		
12	বেলা ১১টার দিকেও যেন ভোরের নীরবতা হলিডে ইন হোটেলে। হাররোতে দুই দল দুই হোটেলে থাকলেও বুলাওয়েতে এসে বাংলাদেশ-জিম্বাবুয়ে একই হোটেলে। অথ	0		
13	নির্বাচক হিসেবে এর আগে অনূর্ধ্ব-১৯ দলের সঙ্গে বিদেশ সফর করেছেন। তবে নির্বাচক মিনহাজুল আবেদীন জাতীয় দলের সঙ্গে দেশের বাইরে এলেন এবারই প্রথম	0		
14	আর বাকি একটা ওয়ানডে ও দুটি টি-টোয়েন্টি ম্যাচ। এরপর এক মাসের সফর শেষ করে ১৩ মে দেশের বিমান ধরবে বাংলাদেশ দল। তবে দলের সঙ্গে ফিরবেন না ন	0		
15	কিংবদন্তি হওয়ার পথে আরেক ধাপ এগোলেন লেব্রন জেমস। যুক্তরাষ্ট্রের এই বাস্কেটবল মহাতারকা পরশু চতুর্থবারের মতো জিতেছেন এনবিএর মোস্ট ভ্যালুয়েবল (	0		
16	রাজধানী শহরে ধ্বংসযজ্ঞ চালানোর পরদিন গতকাল ঢাকা-চট্টগ্রাম মহাসড়কের নারায়ণগঞ্জ অংশের বিভিন্ন এলাকায় তাণ্ডব চালিয়েছে হেফাজতে ইসলাম। তারা গা	1		

- Then we make the dataset to be split into 90:10 to train:test.

- **Language Processing:**
  - When we try to tokenize the Bengali news contents, the characters with bengali **symbols(** ◌ , ◌া, ◌ি, ◌ী, ◌ু, ◌ূ **)** were omitted. So, we had to handle the issue by calling the **tokenizer** method with a self defined lambda function as:
    ```
    CountVectorizer(tokenizer=lambda x: x.split())
    ```
  - Then, we have faced another problem while **tokenizing**. It was that, the Bengali words ending with **' ◌ঁ'(দাড়ি)** could not be split without the **stopping** symbol. So, we need to replace the stopping symbol with space to solve the problem.
  - After **tokenizing** the words, we need to **stack** them in a **vector** as the computer does not understand words. It only understands **numeric** values. For this representation, we chose the **Bag-of-Words**(**BoW**) model. This word embedding technique is described below:
    - When modeling text with machine learning algorithms, the **Bag-of-Words** (BoW) model is used to represent **text** data. The **Bag-of-Words** (BoW) paradigm is widely used in language modeling and text classification since

it is easy to understand. A bag-of-words is a text representation that represents the appearance of words in a document. It entails two steps:

- A **vocabulary** of **known** words.
- A measure of the **presence** of **known** words.

For our project we have used **count-based** vectors for representing our corpus.

- **Model Implementation**:
  - For the implementation of the model, we create a class named **'BagofWordsClassifier'** which is a **deep neural network** based classifier. We have used **two hidden layers** with activation functions **ReLU** and **LeakyReLU** respectively. We have picked the learning rate of **0.0001** and a batch size of **1000**.
  - After passing the tokenized count vector to the model, our model **trained** according to the training data.
  - Lastly, as we have two classes to classify, we have implemented a **binary step function** to **predict** our test samples.

- **Model Evaluation**:
  - For **evaluating** the model, we have passed the test dataset to our model to predict the class of each news content.
  - After classifying all the news contents, we have evaluated our model according to the **confusion matrix** and measured **accuracy**, **precision**, **recall** and **f1-score**.

Experiments:

Dataset: We have taken a dataset of **'Prothom Alo'** paper's contents from 2013 to 2019. The whole dataset has **5,84,710 rows** and **12 columns**. Here is the **year-wise** distribution of number of rows:

2013	2014	2015	2016	2017	2018	2019
59945	97335	102637	113552	82175	97335	31731

Here, we can see that the dataset is extensively huge and there are some **'null'** values in some rows.

5950	এ বছরও আমরা সমন্বিত ভর্তি পরীক্ষা পদ্ধতি চালু ...	0.0
5951	সন্তানদের ক্লাস বা কোচিং শেষ না হওয়া পর্যন্ত স...	1.0
5952	গত চার বছরে সরকার পরিচালনার বিভিন্ন দিক নিয়ে ধ...	1.0
5953	প্রশস্ত মেঘনার কিনারা ঘেঁষে একটি ভূখণ্ড। আদিতে...	0.0
5954	স্বামী-স্ত্রী ফেসবুকে চ্যাট করছিল। একপর্যায়ে ...	0.0
5955	NaN	NaN

So, we take one thousand rows from each year's data and make a modified dataset for our problem. We have also **omitted** extra columns and keep only the columns of **'title'** and **'content'** for our work. Further, we labelled our data according to the **threat**/**toxicity** presence. The distribution of rows of yearly data is presented here:

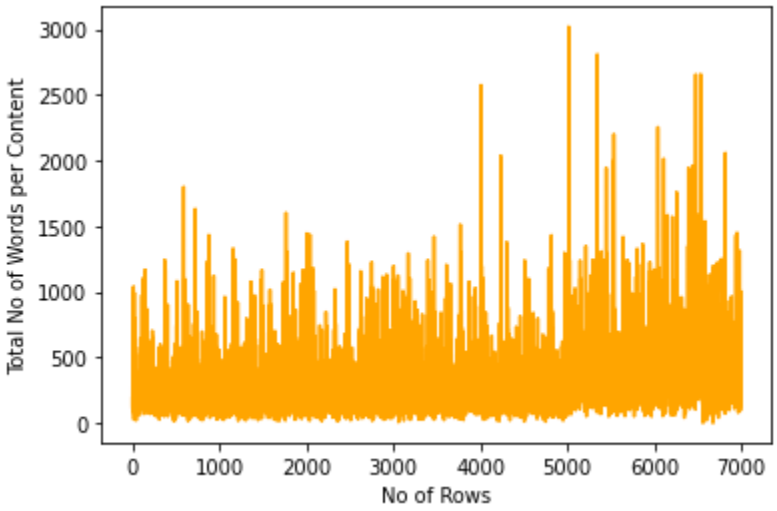
2013	2014	2015	2016	2017	2018	2019
1000	1000	1000	1000	999	1000	998

From the labelled data, the number of contents containing threat news and the number of contents which present non-harmful news are:

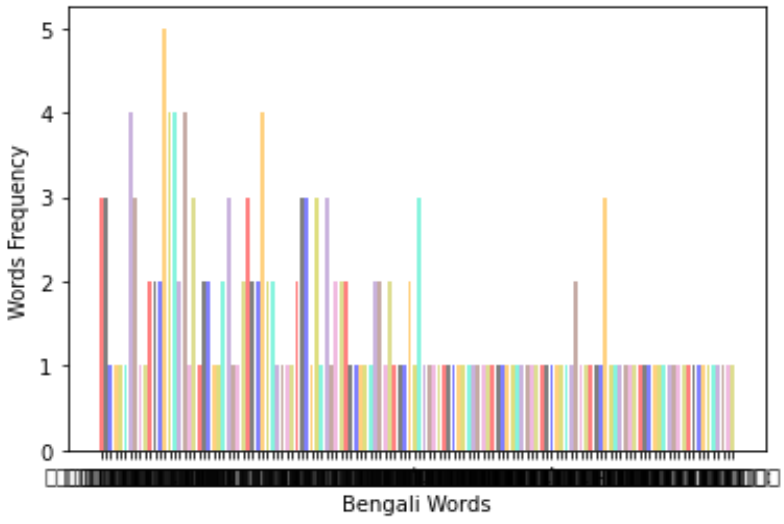
- Number of Samples of **Threatful** News: **3397**
- Number of Samples of **Non-threatful** News: **3600**

Here are some other statistics of the dataset:

- Average length of Contents: **304 Words/Content**
- **Frequency** of words in each content:



- **Frequency** of each word in a sample content:



- **Column Names**:
    - **Content**: **object**
    - **Threat**: **float64**
- **Frequency** of Some Toxic Words:

- হামলা: **597986**
- সৎরাস: **403036**
- নিহত: **361061**
- হৎযা: **645578**
- সহিংসতা: **367400**

Sample Presentation: The sample presentation of our dataset is given below:

Index	Content	Class
1	রাজধানী শহরে ধ্বংসযজ্ঞ চালানোর পরদিন গতকাল ঢাকা-চট্টগ্রাম মহাসড়কের নারায়ণগঞ্জ অংশের বিভিন্ন এলাকায় তা৷ডব চালিয়েছে হেফাজতে ইসলাম। তারা গাছ ফেলে সড়ক অবরোধ করে এবং হাইওয়ে পুলিশের ফাঁড়িসহ বিভিন্ন স্থাপনা ও সড়কে আগুন দেয়। এ সময় পুলিশ-র‍্যাব-বিজিবির সঙ্গে সংঘর্ষে ২০ জন নিহত হন। আহত হন দেড় শতাধিক। নিহত ব্যক্তিদের মধ্যে দুই পুলিশ ও বিজিবির এক সদস্য রয়েছেন।এ ছাড়া গতকাল হেফাজতে ইসলামের সড়ক অবরোধকালে চট্টগ্রামের হাটহাজারীতে পুলিশের গুলিতে সংগঠনটির দুই কর্মী এবং সেনাবাহিনীর এক সদস্যসহ ছয়জন নিহত হয়েছেন। হেফাজতের কর্মীরা সেখানেও তা৷ডব চালান। আর বাগেরহাটে নিহত হয়েছেন হেফাজতের এক কর্মী। এ নিয়ে গতকাল মোট ২৭ জন নিহত হলেন। আগের দিন রোববার রাজধানীতে ব্যাপক সহিংসতায় মারা যান ২২ জন।গতকাল সোমবার ভোর থেকে দুপুর পর্যন্ত নারায়ণগঞ্জ সদর উপজেলার সিদ্ধিরগঞ্জ. সাইনবোর্ড থেকে সানারপাড়-শিমরাইল হয়ে সোনারগাঁ উপজেলার কাঁচপুর পর্যন্ত পাঁচ কিলোমিটার এলাকাজুড়ে সংঘর্ষের ঘটনা ঘটে। এর ফলে বেলা তিনটা পর্যন্ত ঢাকা-চট্টগ্রাম ও ঢাকা-সিলেট মহাসড়কে যান চলাচল বন্ধ থাকে। পুলিশ বলছে. হেফাজতের অবরোধ সরাতে গেলে কর্মীরা হামলা চালান। এতে সংঘর্ষের সূত্রপাত হয়।	1
2	প্রাইম ব্যাংক জেলা ফুটবল লিগে কাল ময়মনসিংহে পুলিশ দল ৩-১ গোলে হারিয়েছে প্রতিশ্রুতি জনকল্যাণকে। কিশোরগঞ্জে কোর্টরোড ৮-১ গোলে হারিয়েছে মার্শাল আর্টকে।	0
3	জিম্বাবুয়ে থেকে দেশে ফিরেই আবার বিদেশে চলে যাবেন সাকিব আল হাসান। বিদেশ মানে যুক্তরাষ্ট্র। স্ত্রী উম্মে আহমেদ শিশির আছেন সেখানে। এরপর কাউন্টি খেলতে যেতে পারেন ইংল্যান্ডেও। সাকিবের এই গতিময় জীবনে এমনকি দুই বছর আগের ঘটনারও খুব বেশি স্থান নেই। নইলে আরেকবার জিম্বাবুয়েতে এসে কেন ভুলে থাকতে চাইবে দুঃসহ সেই স্মৃতির কথা!২০১১ সালের জিম্বাবুয়ে সফর সাকিবকে নামিয়ে এনেছিল আকাশ থেকে মাটিতে। টেস্ট. ওয়ানডে দুই সিরিজই হারার পর অধিনায়কত্ব হারালেন. বিদ্ধ হলেন সমালোচনার তিরে। দুই বছর আগের ক্ষত শুকাতে জয় এবার শুধু বাংলাদেশ দলের জন্যই নয়. সাকিবের জন্যও প্রয়োজন ছিল খুব। তবে টেস্ট সিরিজ ড্র হওয়ায় এখন কাঙ্ক্ষিত সেই জয় আসতে পারে কেবল আজ শেষ হতে যাওয়া ওয়ানডে সিরিজে। সিরিজ জয়ের জন্য তো বটেই. আজকের অলিখিত ফাইনাল জেতার কি সেই অর্থেও জরুরি নয়? সাকিব মানলেন না. 'না...এখন আর এসব নিয়ে এত বেশি টেনশন নেই। জীবন অনেক বদলে গেছে।	0
4	হবিগঞ্জে সাংবাদিক কামরুল হাসান ওরফে আলীম হত্যা মামলায় পাঁচজনকে যাবজ্জীবন কারাদ৷ড. পাঁচ হাজার টাকা করে জরিমানা. অনাদায়ে আরও ছয় মাসের কারাদ৷ড দেওয়া হয়েছে। অতিরিক্ত জেলা ও দায়রা জজ এ এইচ এম নিজামুল হক গতকাল মঙ্গলবার এ রায় দেন। দ৷ডপ্রাপ্ত ব্যক্তিরা হলেন নিহত ব্যক্তির চাচা আবদুল হান্নান চৌধুরী (৫৫). চাচি পিয়ারা বেগম (৪৫). চাচাতো ভাই শাহ আলম চৌধুরী	1

	(১৮). মকদ্দছ চৌধুরী (৩০) ও আজাদ চৌধুরী (২০)। মামলার সংক্ষিপ্ত বিবরণে জানা যায়. দৈনিক দিনকাল-এর নবীগঞ্জ উপজেলা প্রতিনিধি কামরুল হাসানের সঙ্গে আবদুল হান্নানের জমিজমা সংক্রান্ত বিষয় নিয়ে দীর্ঘদিন ধরে বিরোধ চলে আসছিল। ২০০৬ সালের ৭ নভেম্বর সকালে কামরুল বাড়ির পাশে নলকূপে হাত-মুখ ধুতে গেলে আসামিরা দেশীয় অস্ত্রশস্ত্র নিয়ে তাঁর ওপর হামলা চালান।	

Train / Test Split: We use **90:10** ratio for training and testing data.

- Data for **Training**: **6300**
- Data for **Testing**: **697**

Evaluation Metric: After the implementation of the model, our next task is to evaluate the model. For model evaluation, there are various matrices available. We choose to use the **confusion matrix** and **Cohen's Kappa** metric to evaluate our model. For evaluating the model according to the confusion matrix, we need to know about the terms of the confusion matrix. These terms are stated as:

- **True Positive**: It refers to the number of predictions where the classifier correctly predicts the positive class as positive.
- **True Negative**: It refers to the number of predictions where the classifier correctly predicts the negative class as negative.
- **False Positive**: It refers to the number of predictions where the classifier incorrectly predicts the negative class as positive.
- **False Negative**: It refers to the number of predictions where the classifier incorrectly predicts the positive class as negative.

From these terms, we need to calculate Accuracy, Precision and Recall.

For our selected model:

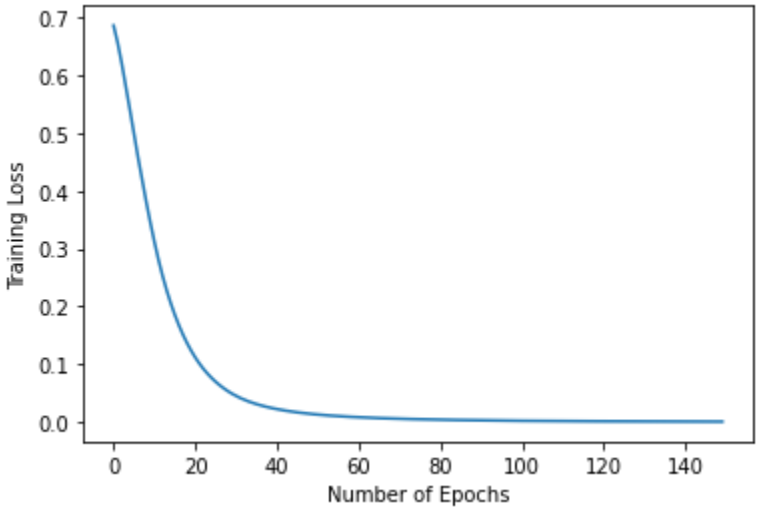
- **Accuracy= 78.00%** ; Refers to how close a measured value to a true value
- **Precision= 73.65%** ; Refers to how close the measured value to each other
- **Recall= 76.567%** ; Refers to the ratio of correctly predicted true predictions

As we are working on **Bengali Language** with a **preliminary** model of **Bag-of-Words**, our focus was to correctly predict the abusive contents from a newspaper. So, the **accuracy** and

recall values are more **important** to us. We choose to select a model with the moderate recall value along with a good accuracy value.

Another evaluation metric which we follow is **Cohen's Kappa** score. Cohen's Kappa metric is the metric which shows the **agreement** between two raters. As our problem is a **binary classification** problem, we decide to measure **Cohen's Kappa** value. In comparison to estimating average precision, Cohen's kappa considers imbalance in class distribution and can thus be more difficult to read.
The **Cohen's Kappa** value for our model: **55.25%**

Results: After the implementation of our model, we have plotted our **training losses** in a graph to visualize the distribution. The graph is shown below:



Here, we can see that the training loss starts from about **0.700**, and after running for **150** epochs, it becomes **0.001**. When the **training** occurs, it tends to reduce the loss **gradually**. After **20** epochs, the curve becomes much **straighter**.
When we evaluate our model according to the test dataset, we calculate the **accuracy** of our model from the **confusion matrix**. Here is the result of our model is shown below:

Word Embedding Model	Bag-of-Words

Technique	Count-based Model
Number of Hidden Layers	2
Nodes in 1st Hidden Layer	128
Nodes in 2nd Hidden Layer	64
Criterion	BCEWithLogitsLoss
Optimizer	Adam
Batch Size	1000
Epoch	150
Input Features	150978
Activation Function 1	ReLU Activation Function
Activation Function 2	LeakyReLU Activation Function
Activation Function of Output Layer	Linear Activation Function
Activation Function of Step Binary Function	Sigmoid
Training Loss after 150 Epochs	0.001
Prediction Criteria	0(Non Threatful Content): for <=0.4 else: 1(Threatful Content)
Accuracy	78.00%
Precision	73.65%
Recall	76.567%
Cohen's Kappa Score	55.25%

From the table, we can see that we use **two hidden layers** to design our **deep neural network**. For the binary step function, we use the value of **0.4** as the **deciding** value. The **activation functions** we use for our model is the **ReLU** and **LeakyReLU** activation function. After evaluating the model, we got an **accuracy** of about **78%.**

Trials: We have tried to get a **good result** from the dataset we have tried in various ways. Here is the **trial chart** of our experiment:

Setting No	**Batch Size**	**Epoch**	**Learning Rate**	**Word Embedding Model**	**Criterion**	**Optimizer**	**Result (Accuracy)**	**Precision**	**Recall**
1	800	100	0.0003	BoW	MSELoss	Adam	40.25%	40.25%	100%
2	900	150	0.0001	TF-IDF	BCEWithLogitsLoss	Adam	40.25%	40.25%	100%
3	1500	100	0.0002	BoW	BCEWithLogitsLoss	Adam	76.00%	68.78%	73.91%
4	1500	100	0.0005	BoW	BCEWithLogitsLoss	Adam	76.25%	68.96%	74.53%
5	6300	200	0.0001	BoW	BCEWithLogitsLoss	Adam	76.32%	72.54%	73.26%
6	6300	200	0.0001	BoW	BCEWithLogitsLoss	Adam	76.33%	72.55%	73.27%
7	900	150	0.01	BoW	BCEWithLogitsLoss	Adam	76.62%	70.59%	89.2%
8	1000	100	0.0001	BoW	BCEWithLogitsLoss	Adam	76.75%	69.31%	75.77%
9	900	150	0.0001	BoW	BCEWithLogitsLoss	Adam	77.00%	70.18%	74.54%
10	**1000**	**150**	**0.0001**	**BoW**	**BCEWithLogitsLoss**	**Adam**	**78.00%**	**73.65%**	**76.567%**
11	500	50	0.0005	BoW	BCEWithLogitsLoss	Adam	78.25%	73.41%	72.04%

From the chart, we get some intuitions for picking the right combination of hyperparameters.

- For this dataset, the **‘TF-IDF’** model does not work well. The training loss of the model was **very high** and did not give a **good result**.

- We have used the **Bi-gram** technique to see how the model works. While using the **Bi-gram** technique, the model gave an accuracy of **76%,** which was pretty similar to our normal model.
- **'MSELoss'** criterion function **does not** give **good results** for this dataset.
- When we choose to pick a **lesser amount** of **batch size**, our model works **very well**.
- **The Learning rate** between **0.0001** and **0.0005** works fine with the model.
- **'SGD'** optimizer needs a **higher number** of **epochs** for convergence while **'Adam'** optimizer batch size of **100** to **150** is quite **enough**.
- **A Greater** number of **batch sizes** needs a **greater number** of **epochs** to reduce the training loss.
- In our model, the **training loss** becomes nearly **0,** while the **validation loss** may be **greater** than the training loss. So, we can say that the model is somewhat **overfitting,** but the level of **overfitting may not be very bad** as the testing **accuracy** gives a **good result** of nearly **78%**.

Conclusion: Kids are the future of the world. Kids need to feel **affirmative** towards anything. We should not let them face any **embarrassing** situations. As we want to make a suitable online world for kids, our job is to make such software that helps kids **flourish** and practice their **imagination**. Our project tends to help kids by presenting **non-abusive content-based** newspapers by filtering out the **abusive** contents with deep neural networks and natural language processing. As our work is for the kids now, but our goal is to make a proper generalized newsfeed for every age group in the future. Our project is just a model now. In the end, we will use **RNN**, **FastText**, and many other **natural language methods** to increase our work's **accuracy** and try to implement the work as a **software system** so that anyone can use it easily.

Bibliography

Islam, Khondoker I. 2020. "Sentiment analysis in Bengali via transfer learning using multi-lingual BERT." ***https://arxiv.org*** **1, no. [Submitted on 3 Dec 2020] (12): 5.**

NewsWise. 2021. "Child-friendly news." *The Guardian*, 1 18, 2021. https://www.theguardian.com/newswise/2019/jan/18/child-friendly-news.

Pawlowsk, A. 2020. "Why racism can have long-term effects on children's health." *Today*, July 17, 2020. https://www.today.com/health/why-racism-can-have-long-term-effects-children-s-health-t186480.

Roy, Adrija. 2017. "Sentimental Analysis (Bengali)." https://github.com/abhie19/Sentiment-Analysis-Bangla-Language/blob/master/ANLP_REPORT_abhishek_Adrija.pdf. https://github.com/abhie19/Sentiment-Analysis-Bangla-Language/blob/master/ANLP_REPORT_abhishek_Adrija.pdf.

Sazzed, Salim. 2021. "Bangla (Bengali) sentiment analysis classification benchmark dataset corpus." https://data.mendeley.com. https://data.mendeley.com/datasets/p6zc7krs37/4.

Sirajus, MD S. 2019. "Development of a Bangla news classification system." http://lib.buet.ac.bd. http://lib.buet.ac.bd:8080/xmlui/handle/123456789/5371.