



University of
KURDISTAN
Hewlêr

Department of Computer Science and Engineering

Big Data Processing SE505

Semester II, Academic year 2021/22

Kurdish Profanity Processing

Team members:

Mohammed Sardar Noori

Mahmood Yashar

Shaimaa Salih

Yasameen Sami

lecturer:

Dr. Hossein Hassani



Table of Contents

1.Introduction	4
3. Data Collection	7
Meta Data.....	12
4.Data Exploration	14
4.1 Identifying Data Characteristics.....	14
4.1.1 Volume	14
4.1.2 Velocity	15
4.1.3 Variety	15
4.1.4 Veracity.....	15
4.1.5 Value	15
4.2 Data Acquisition and Filtering.....	16
4.3 Data Validation and Cleansing	18
4.Data Analysis	20
6. Visualization	28
7. Conclusion and Future Works.....	32
8. Acknowledgment.....	32
9. Lesson Learned.....	33
9.1 Mahmood Yashar (Data analyzer)	33
9.2 Mohammed Sardar Noori (Team Leader).....	33
9.3 Shaimaa Salih (Data Collector & Documenter)	35
9.4 Yasameen Sami (Data Visualizer)	35
10. References.....	36

List of Tables

Table 1, shows detail of each data that have been collected during the project.	13
Table 2, Details of meetings and its achievements.....	38

List of Figueres

Figure 1 Facepager Software.....	7
Figure 2 Facepager API.....	8
Figure 3 Extracted types of data.	8
Figure 4 Adding Nodes	9
Figure 5 ID of pages	9
Figure 6 Data Parameters	10
Figure 7 Fetching Data	10
Figure 8 CSV file of datasets.....	11
Figure 9 Running python code on Google Colab	11
Figure 10 Extracted data from Facepager.....	16
Figure 11 Filtered data.....	17
Figure 12 only movie lines extracted.	18
Figure 13 As you can see this analysis result shows (Çek) in two different k.	19
Figure 14 Screenshot from Kurdînûs, it shows converting process to Kurdish Unicode.....	19
Figure 15 MapReduce processing.	20
Figure 16 Java code for counting words in Hadoop.....	21
Figure 17 Changing directory to Hadoop	22
Figure 18 Starting namenode and datanode	22
Figure 19 Starting nodemanager and resource manager	22
Figure 20 Putting data file to the HDF.	23
Figure 21 printing all file inside HDFS folder	23
Figure 22 analyzing the data with mapreduce.	24
Figure 23 An error occurred in reading Kurdish language.....	25
Figure 24 putting the output of the Hadoop to the c directory	25
Figure 25 output of the processed data with notepad ++.....	26
Figure 26 Code of comparing the bad word with processed output.....	27
Figure 27 Number of repeated bad word in each content.....	27
Figure 28 percentage of inappropriate words used by Rudaw.	28
Figure 29 The number of inappropriate words used in Vejin, a Kurdish literature website.	29
Figure 30 The number of inappropriate words used in NRT, a political website.	29
Figure 31 The number of inappropriate words used in Twitter replies/comments.	30
Figure 32 Most used bad words in a Book text corpus.	30
Figure 33 Shows that football and sport websites have least number of using bad words.	31
Figure 34 The number of inappropriate words used in NRT2, an entertainment website.	31
Figure 35 Most used bad words in movies corpus text.	32

1.Introduction

Without the internet, the world would not be what it is today. It affects almost every part of our lives, including how we live, work, socialize, shop, and play. However, internet connectivity is a relatively new phenomenon that has dramatically altered the globe in a relatively short period of time. The internet has evolved from a new tool for the US military to stay in touch with the human race's always-connected heartbeat in just a few decades.[1] With each passing year, an increasing number of people have gotten internet access. Therefore, as the number of user-generated web content expands, so does the amount of inappropriate and/or undesirable content. Several academic communities are working on detecting and managing such content: computer vision research is focusing on detecting unsuitable images, while natural language processing technology has progressed to recognize insults.

Additionally cyberbullying and negative content has increased dramatically as the number of Internet users has grown. Cyberbullying is criminal conduct that involves the persistent use of hostile expressions such as text, images, and sounds online through IT equipment such as computers to cause mental and physical harm to others. Because of the rise of smartphones, the age of cyberbullying perpetrators is gradually dropping. Furthermore, victims of cyberbullying have committed suicide, making it a severe social issue. On many websites, community administrators are primarily in charge of eliminating offensive content. The flow of user-generated content on many sites, on the other hand, quickly overwhelms community managers' abilities to control it properly. Therefore, to reduce negative content on the internet and especially cyberbullying, profanity filtering is a decent starting step in content management especially in user-generated content (UGC).[2]

A profanity filter is a sort of software that examines UGC in online forums, social platforms, markets, and other locations to remove profanity. Swear words, words connected with hate speech, harassment, and other words are censored by moderators. Kurdish language, like any other language, is subjected to cyberbullying, despite the fact that several profanity projects have been done before, we were unable to find any for the Kurdish language. That is why we wanted to be

the first to attempt anything like this for our community since there has never been any previous profanity filtering or detection for the Kurdish language done before.

However, profanity detection software remains ineffective. They are simple to get around and quickly grow stale, as they cannot adjust to misspellings, abbreviations, or the rapid expansion of profane language. On the other hand, they have their own set of acceptable behaviors; what is acceptable in one group may not be in another. In addition, detecting profanity in Kurdish is more difficult than in English since the standardized Central Kurdish (Sorani) keyboard/Unicode has not been used by every operating system and individual users. Furthermore, Sorani still doesn't have an orthography standpoint.[3]

In this paper, we're proposing a way for detecting and counting profanity on Kurdish websites and social media, with the goal of analyzing the profanity percentage on each website to determine whether it is appropriate for children to visit. We have to apologize that this paper contains some offensive words in Kurdish in the profanity\black list.

2. Problem Definition

Fastlink is an internet service provider in the Kurdistan Region of Iraq. The company rapidly grew in size and popularity after its founding. As a result, the company began bringing new services to participants to compete with other companies. One of the services was introduced in 2015 called family SIM cards, which assured parents that their children would be prevented from seeing and accessing inappropriate and violent websites while using the Internet including "Drugs, Alcohol and Tobacco, Gambling, Pornography, Suicide and Self-Harm, Weapons and Violence".[4] To activate the service, participants must buy a new SIM card for 15000 IQD and pay an extra 1000 IQD from their monthly subscription plan via Fastpay.

Customers of the service initially gathered a good number. However, according to the company's investigation, service customers were declining month by month, in a way that the company thought of retiring the service. Before that, the company conducted a survey among participants

to find out why the service participants were not continuing. As a result, they found out that parents knew their children were still accessing some inappropriate content that the service should have prevented them from accessing. According to the survey, 68% of the parents noted that their children had seen some inappropriate contents in Kurdish, so they would no longer subscribe in the service.

As a solution, we as 4 juniors in Fastlink decided to implement in-house profanity detection project for the company and conduct research on Kurdish-language contents over the Internet and find those websites that can be suspended through the services, they will give it another chance to the offer and will resume its marketing plans.

In order to start the project, we've created KPI for making Fastlink and its customer satisfaction.

KPI

- Increasing sales 8% yearly.

3. Data Collection

Regardless of the subject of study, data collecting is the first and most significant phase in the research process therefore we've used different ways for collecting data. Some of the data are collected from other researchers who are working on Kurdish language and they provided those data under public domain or other free licenses.[5][6] On the other hand, we collected data from social media with Facepager software which allows us to collect the data from Twitter, Facebook including comments posts and also tweets as shown in the figure 1.[7], [8] Furthermore, we gained some kurdish movie subtitles/scripts from Kurdcinama.com, also some of it from organizations such as the UN Migration Agency (IOM).

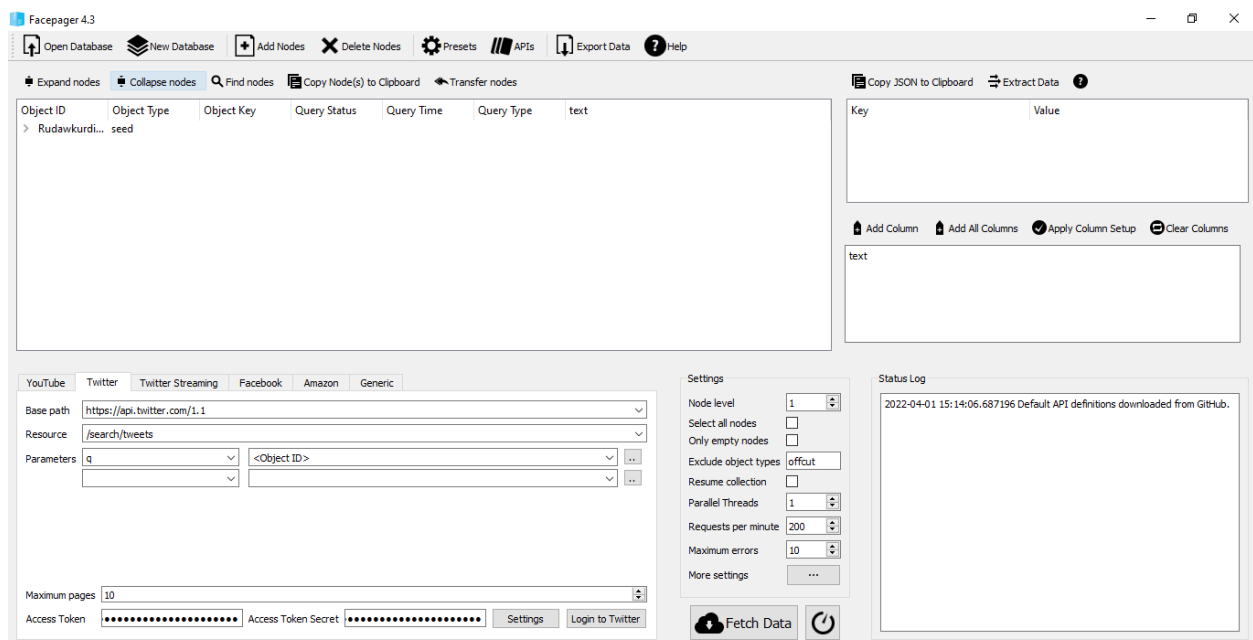


Figure 1 Facepager Software

Facepager allows us to collect types of data which provides the variety of the data that can be tweets, comments, retweets, posts and others through API of Facebook and Twitter as shown in the figure 2.

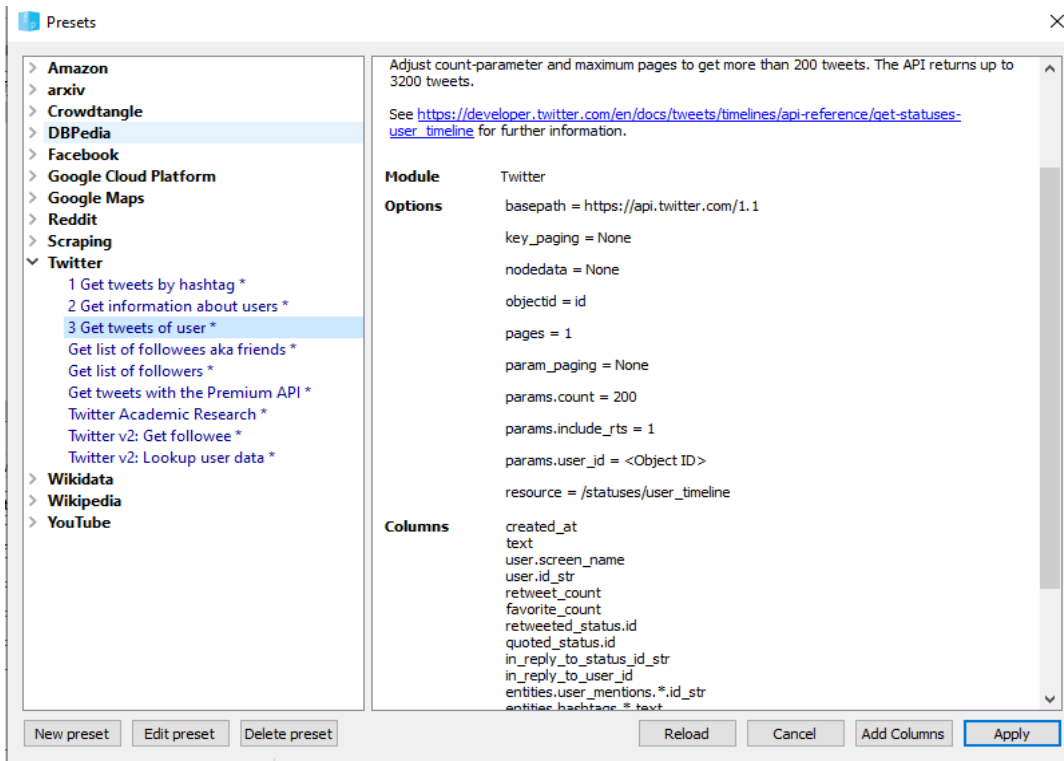


Figure 2 Facepager API

After setting the API of the data it will also show you what type of data will be collected to its database as such text, id, and others, type of columns can be modified by adding columns or clear the columns which is not necessary for the processing as shown in figure 3.

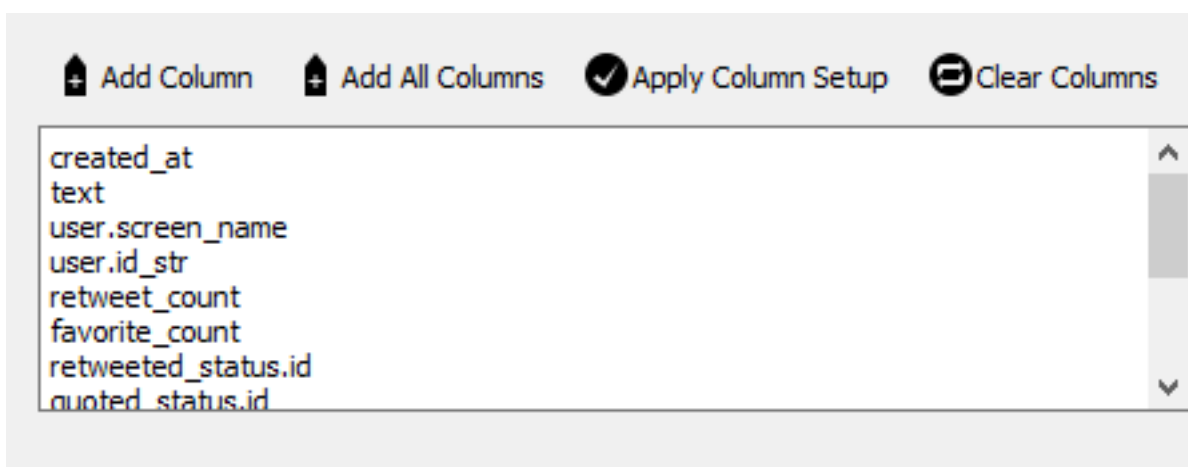


Figure 3 Extracted types of data.

The software has a feature in which allows you to select the type and amount of pages to be collected but firstly it needs to create a database and save it on your pc then to create a node as shown in figure 4 the node stores the id of the page, then the id of the page will be stored in the node for the data to be extracted as shown in the figure 5.

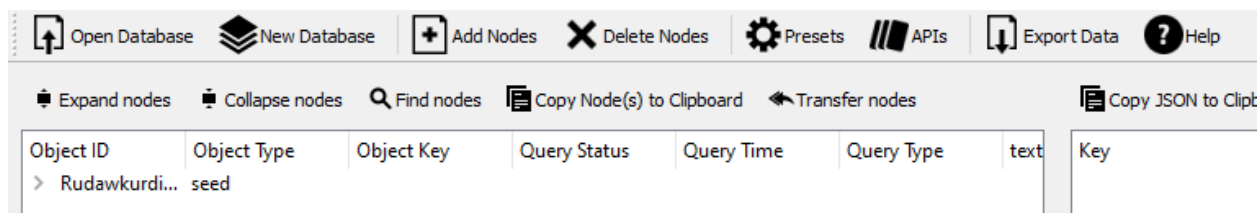


Figure 4 Adding Nodes

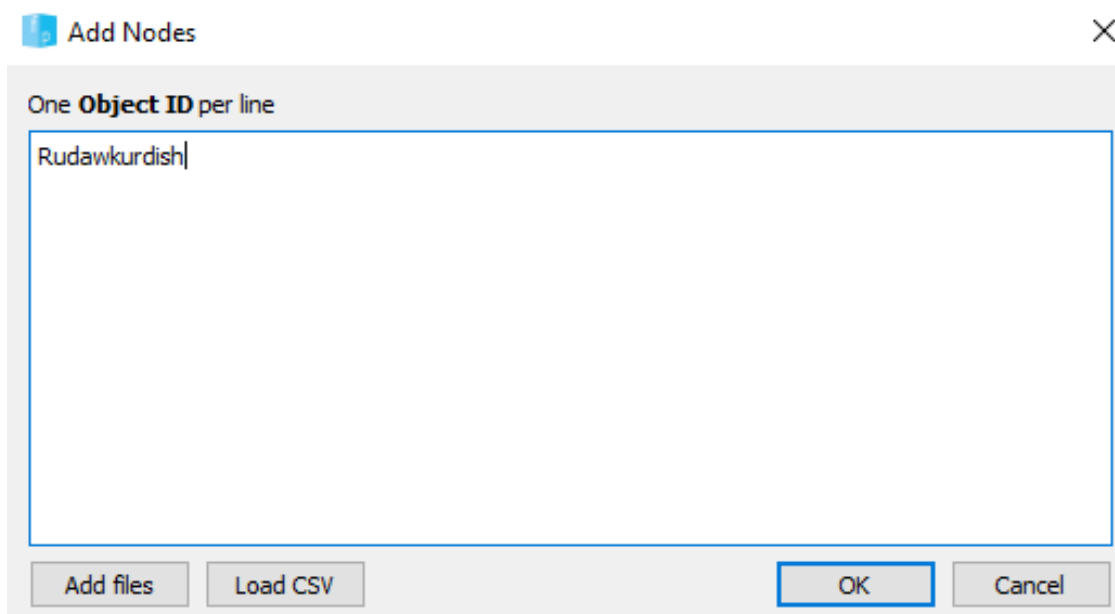


Figure 5 ID of pages

Then the parameter, maximum pages also can be decided for the amount of data to be collected as shown in the figure 6.

Figure 6 Data Parameters

After setting the parameters and types of the data the software will fetch the data depending on the parameters that are defined as shown in the figure 7.

Query Type		text
Twitter:/search/...	{"created_at": "...	https://t.co/4ujYp5NZI6 سوێد: به دووری نازانین رووسیا هێرشمان بکاته سهر
Twitter:/search/...	{"created_at": "...	RT @Rudawkurdish: سهری سالتا کورد: بۆ گهلی کوردی به زمانی کوردی ئیسرائیل به
Twitter:/search/...	{"created_at": "...	@Rudawkurdish Hahaha
Twitter:/search/...	{"created_at": "...	@Rudawkurdish 😂😂. خودی خوه به فربهیت دی چن.
Twitter:/search/...	{"created_at": "...	مام قازی دانیشتوو به کی گوندی ئاخجه مه شته له خورماتوو، ده لیت، نه و که سه ی فاکسینی کو
Twitter:/search/...	{"created_at": "...	گهنجیک له ههولێر له بهر ساردیی که شوهه و به به تانییه وه چوو به سهریان بۆ سهر چای قهر
Twitter:/search/...	{"created_at": "...	@ShkoKurdi @RudawEnglish @Rudawkurdish Isn't @Rudawkurdish basicly run by...
Twitter:/search/...	{"created_at": "...	RT @Rudawkurdish: ئامه د نه ورۆزی ئامه د: https://t.co/IQyFmeuuY1
Twitter:/search/...	{"created_at": "...	Turkish Kurds!! Turkish Kurds? What?? Seriously what is a Turkish Kurd? Who is run...
Twitter:/search/...	{"created_at": "...	ده مباح: شاهۆ نه مین هه وائی نوئی پتیه https://t.co/IQyFmeuuY1
Twitter:/search/...	{"created_at": "...	مه ممه د حه لبوسی @AlHaLboosii سهرۆکی په ره له مانی عێراق به بۆنه ی نه ورۆزه وه له په یا
Twitter:/search/...	{"created_at": "...	@Rudawkurdish ههههم
Twitter:/search/...	{"search_metad...	
Twitter:/search/...	{"created_at": "...	@Rudawkurdish نینینینین
Twitter:/search/...	{"created_at": "...	malicijski alibiz, videacija, nikakvih, lažnih, šaržiranih, https://t.co/Hi

Figure 7 Fetching Data

Then the data can be Exported to a CSV file then we filtered the data in the excel sheet as shown in the figure 8.

[illegible]

To scrap Kurdish news websites, Beautiful Soup library is used in the python programming language.[9] We run the code through Google Colab[10] as shown in figure 9.

Figure 9 Running python code on Google Colab

Meta Data

File Name	Type	Date creation	Last Update	Source/Owner	Size	Category	Extra information
Kurdish Bad Words	.txt	25th March 2022	21st May 2022	<ul style="list-style-type: none"> Central Kurdish Wikipedia youswear.com 	2.94 KB	-	229 words
Awena	.txt	28th April 2022	28th April 2022	awene.com	18.5 MB	Politics	Article range id: (2344-12450)
Rachleken	.txt	7th April 2022	7th April 2022	rachlaken.com	23.5 MB	All	Article range id: (20837-32837)
Radio Nawa Website	.txt	28th April 2022	28th April 2022	radionawa.com	10.7 MB	Politics	Article range id: (20837-32837)
Wshe	.txt	28th April 2022	28th April 2022	wishe.net	6.9 MB	All	-
Payam	.txt	5th April 2022	5th April 2022	peyam.net	16.2 MB	Politics	-
NRT	.txt	5th April 2022	5th April 2022	nrttv.com	23.1 MB	Politics	Article range id: (100-9600)
NRT2	.txt	5th April 2022	5th April 2022	nrttv.com/nrt2	11.2 MB	Entertainment	-

Yariga	.txt	28th April 2022	28th April 2022	Yariga.net	263 KB	Sports	Article range id: (8932-10470)
Vejin	.txt	28th April 2022	28th April 2022	books.vejin.net/ck	32.2 MB	Literature	-
Xelk	.txt	7th April 2022	7th April 2022	xelk.org	5.47 MB	All	-
IOM Social Media-2015-19	.xlsx	19th May 2022	19th May 2022	IOM	718 KB	Migration	-
IOM Social Media-2020	.xlsx	19th May 2022	19th May 2022	IOM	431 KB	Migration	-
IOM Social Media 2021	.xlsx	19th May 2022	19th May 2022	IOM	1.03 MB	Migration	-
IOM Stories	.txt	19th May 2022	19th May 2022	IOM	1.3 MB	Migration	-
Movies	.srt	1st May 2022	1st May 2022	Kurdcinama.com	282.8 MB	Movies	Each movie generally contains 700-2000 lines.
Twitter-data	.txt	–	8th June 2020	Fatih Kurt	3.98 MB	Social Media	-

Table 1, shows detail of each data that have been collected during the project.

4.Data Exploration

4.1 Identifying Data Characteristics

The Data Identification stage is responsible for defining the datasets and sources that will be used in the analysis project. As previously stated, our data was gathered from a variety of sources including social media, websites, and books. A variety of datasets have been identified from the data collected:

- **Structured Data:** Excel sheet data from the migration agency are considered to be structured data. The data is organized by columns of each social media source, links and content and topics.
- **Unstructured Data:** This category of the dataset includes posts and comments from various social media accounts that are all text files, as well as the majority of our collected data. In addition, we collected Kurdish website content texts from various Books.
- **Semi-Structured Data:** The subtitles we got from the movies are considered to be Semi-Structured data, as Semi-structured data does not correspond to relational databases like Excel or SQL, but it still has some organization because of semantic components like tags.

As we're aware that data to be considered Big Data, a dataset must have some features that must be accommodated in the analytic environment's solution design and architecture. The following are the data characteristics of our dataset:

4.1.1 Volume

Big Data solutions are expected to process a significant amount of data, which is expected to continue to expand. We've collected approximately 2 GB of Kurdish texts from social media posts and comments, news, books, poems and movies subtitles.

4.1.2 Velocity

Velocity (in-flow of data) may be more important than volume because it might provide us with a greater competitive advantage. Sometimes it's preferable to have a small amount of data in real-time rather than a large amount at a slow speed. Since our data is collected from websites, movies and social media mostly, because of its widespread availability and popularity as a global conversation driver, social media has become associated with "big data." It is commonly mentioned as a textbook example of what comprises "big data" in today's data-drenched world because of its vast size, fast update speed, and wide range of content types.

4.1.3 Variety

There must be a multiplicity of data types and formats to be deemed Big Data. Poems, novels, Facebook posts, and tweets are all unstructured and incredibly varied. Movies' subtitles and social media comments are semi-structured.

4.1.4 Veracity

Data veracity refers to quality of the data where the data has been collected, how it was collected and how it will be analyzed. Since the data are collected from the Social media, books, movies, web and comments they are accurate data and the aim is to get the data from the places where Fastlink company did not filter them before.

4.1.5 Value

The value of data helps us in processing data and obtaining accurate results. Even if there is a lot of diversity and a lot of data, if it has no value, it is useless and will produce inaccurate results. Hence our data is collected from different resources including social media, poetry, websites, movies, and novels. Since most of these resources are among the most popular and widely read Kurdish social media, they will be useful in making decisions throughout the data processing.

4.2 Data Acquisition and Filtering

Following the collection stage The data is next submitted to automatic filtering to remove any corrupt data or data that has been determined to be irrelevant to the analysis goals.

For the data we collected in multiple pages of social media including (pages, websites, books, movie subtitle) some of the data they were not ready to be processes they needed to be cleansed for example as shown in the figure below when we were using facepager software tool to extract data from the social media the data was mixed between comments, id of the user and other different characters including path; "id"; "parent_id";"level"; "object_id" ;"object_type" ;"object_key" ;"query_status" ;"query_time"; "query_type" ;"message"; "created_time"; "updated_time";"error.message" as shown in figure 10.

[illegible]

Figure 10 Extracted data from Facepager

An online tool has been used to filter text only. As we got Kurdish movie subtitles from a third party source (Kurdicnama.com). Those subtitle files are stored in .srt files which include the number of lines, timestamps and movie dialogues as shown in the figure below. We removed time and line counters, then we save extracted remain texts into “.txt” files, as indicated in figure 12.[11]



Figure 12 only movie lines extracted.

4.3 Data Validation and Cleansing

The Data Validation and Cleansing stage is responsible for developing frequently complicated validation rules as well as deleting any known invalid data that cannot be used for analysis. In our case, offline ETL operation used to validate and cleanse data, since we're processing Kurdish language and due to the fact that operating systems don't provide a standard inbuilt Central Kurdish Keyboard from the beginning, several keyboard layouts with different code developed by individuals and those keyboards are still used by people social media and organizations.

Some keyboards use (ك U+0643) while others use (ک U+06A9),[12] look at Figure 13. Furthermore, some new keyboards has a special letter called (ه U+06BE) which different from (ە), this letter uses for writing some special words such as (گوناھ).[13] This issue brought a big challenge for language processing.

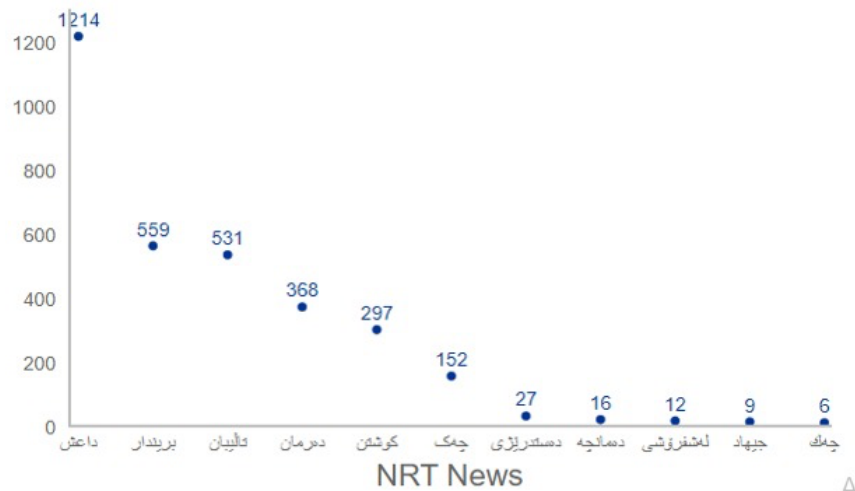


Figure 13 As you can see this analysis result shows (Çek) in two different k.

For this project, we followed a standard layout which was approved by KRG in 2014.[14][15] To convert our data, we use a ready made open source tool called (Kurdînûs) to convert our data into that standard unicode format as shown in figure 14.[16], [17] We divide text files into small files in order to be processed by Kurdînûs and then merge all files by this MS windows command (Copy *.txt output.txt).

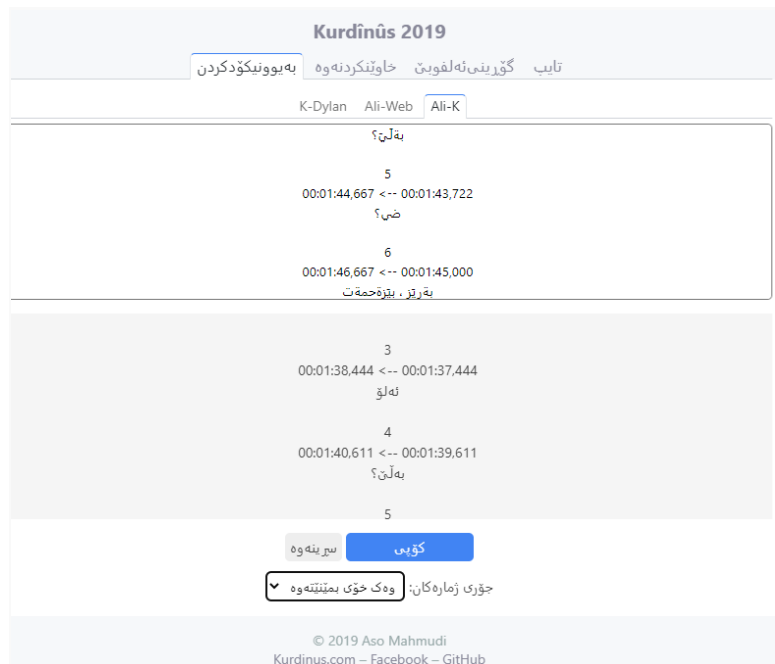


Figure 14 Screenshot from Kurdînûs, it shows converting process to Kurdish Unicode.

4.Data Analysis

Data analysis is a process used by the data analyzer to clean, reduce, transform the data based on the business. It helps to reduce large amounts of different data into smaller fragments, which will help the data analyzer to understand the data. The main goal is to discover useful information from the data. In this step we will analyze the data which is filtered, extracted, and cleansed in the previous stages. We use the data analysis to understand the relation between the result and according to the KPI which was identified previously, to find the most frequent profanities within the kurdish websites and social media.

The framework which is used to analyze the Batch processing (offline processing) data in Apache Hadoop which is an open-source software platform for analyzing the data in reliable, Scalable, distributed computing, which will allow for the distributed processing of the large amount of datasets across different clusters of computers. Apache hadoop is designed to scale up from a single server to multiple machines. Apache Hadoop implements MapReduce processing framework which used to map the way to process data by splitting and mapping the data and reducing tasks shuffle and reduce large amounts of data in smaller as shown in the figure 15.

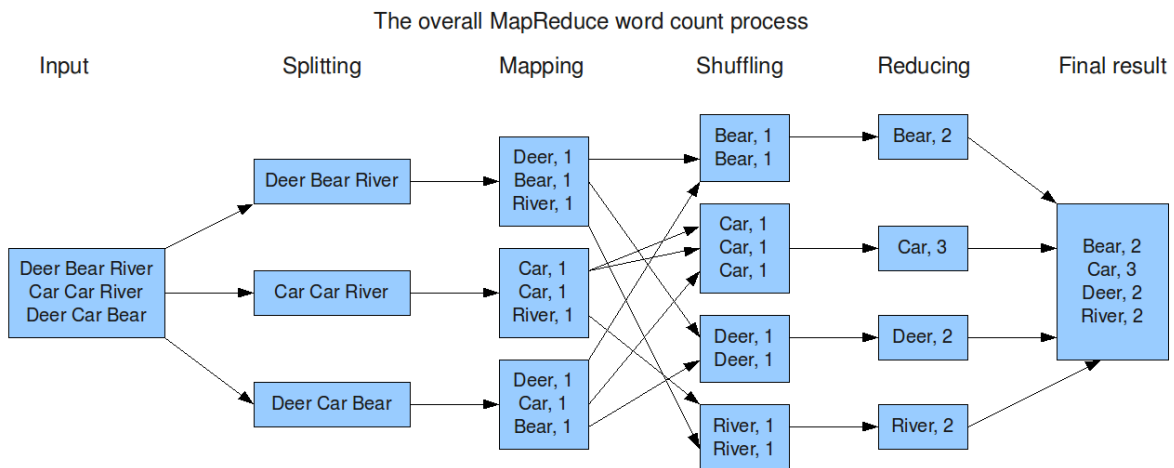


Figure 15 MapReduce processing.

To start using Apache hadoop first it must be installed and configured since this framework works with java language it requires a correct version of the JDK(java development environment) which in our case was (1.8.0_321). In our case we installed hadoop in the microsoft windows 11 operating system, after installing the hadoop there were some issues with the package we ensured to solve all issues before starting the analysis process. After doing all the configuration we started to begin processing. To do the word count we wrote a code for doing the word count as shown in the figure below.[18]

```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {

    public static class TokenizerMapper
        extends Mapper<Object, Text, Text, IntWritable>{

        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(Object key, Text value, Context context
            ) throws IOException, InterruptedException {
            StringTokenizer itr = new StringTokenizer(value.toString());
            while (itr.hasMoreTokens()) {
                word.set(itr.nextToken());
                context.write(word, one);
            }
        }
    }

    public static class IntSumReducer
        extends Reducer<Text, IntWritable, Text, IntWritable> {
        private IntWritable result = new IntWritable();

        public void reduce(Text key, Iterable<IntWritable> values,
            Context context
            ) throws IOException, InterruptedException {

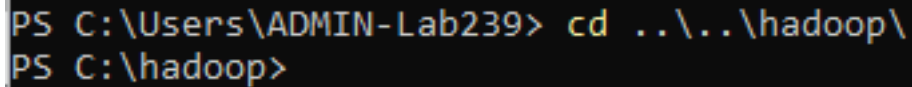
            int sum = 0;
            for (IntWritable val : values) {
                sum += val.get();
            }
            result.set(sum);
            context.write(key, result);
        }
    }

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "word count");
        job.setJarByClass(WordCount.class);
        job.setMapperClass(TokenizerMapper.class);
        job.setCombinerClass(IntSumReducer.class);
        job.setReducerClass(IntSumReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

Figure 16 Java code for counting words in Hadoop.

After filtering the data then we started to use hadoop:

1. First step is to use the hadoop we need to go to the hadoop folder by executing this command `cd ../../hadoop\`

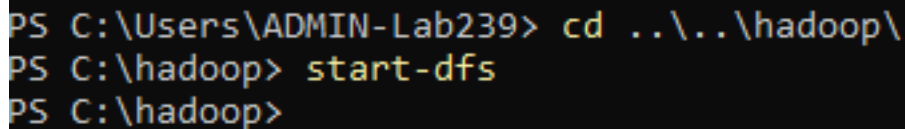


```
PS C:\Users\ADMIN-Lab239> cd ../../hadoop\  
PS C:\hadoop>
```

A terminal window with a black background and yellow text. The first line shows the command 'cd ../../hadoop\' being executed from the path 'C:\Users\ADMIN-Lab239'. The second line shows the prompt 'PS C:\hadoop>' indicating the directory change was successful.

Figure 17 Changing directory to Hadoop

2. Second, starting with the namenode and datanode. DataNode is responsible for storing the actual data in HDFS, to start it we need to execute this command line `PS C:\hadoop> start-dfs`



```
PS C:\Users\ADMIN-Lab239> cd ../../hadoop\  
PS C:\hadoop> start-dfs  
PS C:\hadoop>
```

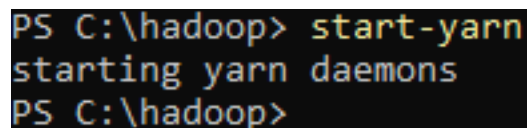
A terminal window with a black background and yellow text. The first line shows the directory change command. The second line shows 'start-dfs' being executed. The third line shows the prompt 'PS C:\hadoop>'.

Figure 18 Starting namenode and datanode

3. The following command need to be used to start nodemanager and resource manager: `PS C:\hadoop> start-yarn`

NodeManager (NM) is YARN's per-node agent, and takes care of the individual compute nodes in a Hadoop cluster

ResourceManager (RM) is the master that arbitrates all the available cluster resources and thus helps manage the distributed applications running on the YARN system.



```
PS C:\hadoop> start-yarn  
starting yarn daemons  
PS C:\hadoop>
```

A terminal window with a black background and yellow text. The first line shows 'start-yarn' being executed. The second line shows the output 'starting yarn daemons'. The third line shows the prompt 'PS C:\hadoop>'.

Figure 19 Starting nodemanager and resource manager

- Then to process the data we need to copy the folder to the HDFS file in the hadoop. Otherwise the data cannot be processed so we must copy all the data to the HDFS file in the hadoop by this command **PS C:\hadoop> bin/hdfs dfs -put c:/kurword/* /kurword**

```
PS C:\hadoop> bin/hdfs dfs -put c:/kurword/* /kurword
```

Figure 20 Putting data file to the HDF.

- The command means there is a folder named kurword inside the c copy all by * then put inside the hdfs folder
- Then now we can check if the HDFS folder consist of kurword by this command **PS C:\hadoop> bin/hdfs dfs -ls /kurword**

```
PS C:\hadoop> bin/hdfs dfs -ls /kurword
Found 21 items
-rw-r--r-- 1 ADMIN-Lab239 supergroup 466201 2022-05-15 13:15 /kurword/Allmovies.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 19440325 2022-05-15 09:43 /kurword/Awana.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 24214912 2022-05-15 09:43 /kurword/NRT NEWS.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 24652174 2022-05-15 09:43 /kurword/Rachleken.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 11232879 2022-05-15 09:43 /kurword/Radio Nawa Website.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 7233422 2022-05-15 09:43 /kurword/Wshe.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 269421 2022-05-15 09:43 /kurword/Yariga.net.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 259746629 2022-05-15 13:48 /kurword/books.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 2345241 2022-05-16 10:17 /kurword/facebookkurdistan24.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 3489120 2022-05-16 10:17 /kurword/facebookrudawposts.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 16703692 2022-05-15 09:43 /kurword/milletpress.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 11707514 2022-05-15 09:42 /kurword/nrt2.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 16191716 2022-05-15 09:43 /kurword/peyam (1).txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 4170331 2022-05-15 13:44 /kurword/twitter-data.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 14313 2022-05-16 10:17 /kurword/twitterkurdistantv.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 33825 2022-05-16 10:17 /kurword/twitterkurditan24.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 191490 2022-05-16 10:17 /kurword/twiterrudawposts.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 17116097 2022-05-15 09:43 /kurword/vejin part1.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 16816864 2022-05-15 09:43 /kurword/vejin part2.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 2219377 2022-05-16 10:17 /kurword/xandanfacebook.txt
-rw-r--r-- 1 ADMIN-Lab239 supergroup 5735114 2022-05-15 09:43 /kurword/xelk.org.txt
PS C:\hadoop>
```

Figure 21 printing all file inside HDFS folder

- Data processing starts the data in the hdfs folder by this command
PS C:\hadoop> bin/yarn jar C:\hadoop\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.2.2.jar wordcount /kurword/Awana.txt / Awanaprocess
Which means do the MapReduce process as a word count inside the hdfs folder named kurword and inside it there is a txt file named Awana.txt

```

PS C:\hadoop> bin\yarn jar C:\hadoop\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.2.2.jar wordcount /kurword/Awena.txt /Awenaprocess
2022-05-15 10:48:21,005 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2022-05-15 10:48:21,586 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ADMIN-Lab239/.staging/job_1652596461883_0001
2022-05-15 10:48:21,731 INFO input.FileInputFormat: Total input files to process : 1
2022-05-15 10:48:21,802 INFO mapreduce.JobSubmitter: number of splits:1
2022-05-15 10:48:21,930 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652596461883_0001
2022-05-15 10:48:21,931 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-05-15 10:48:22,071 INFO conf.Configuration: resource-types.xml not found
2022-05-15 10:48:22,072 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-05-15 10:48:22,274 INFO impl.YarnClientImpl: Submitted application application_1652596461883_0001
2022-05-15 10:48:22,380 INFO mapreduce.Job: The url to track the job: http://DESKTOP-018HQSV:8088/proxy/application_1652596461883_0001/
2022-05-15 10:48:22,597 INFO mapreduce.Job: Running job: job_1652596461883_0001
2022-05-15 10:48:29,439 INFO mapreduce.Job: Job job_1652596461883_0001 running in uber mode : false
2022-05-15 10:48:29,440 INFO mapreduce.Job: map 0% reduce 0%
2022-05-15 10:48:34,497 INFO mapreduce.Job: map 100% reduce 0%
2022-05-15 10:48:39,555 INFO mapreduce.Job: map 100% reduce 100%
2022-05-15 10:48:39,571 INFO mapreduce.Job: Job job_1652596461883_0001 completed successfully
2022-05-15 10:48:39,637 INFO mapreduce.Job: Counters: 54
  File System Counters
    FILE: Number of bytes read=5528735
    FILE: Number of bytes written=11529259
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=19446429
    HDFS: Number of bytes written=4673203
    HDFS: Number of read operations=8
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
    HDFS: Number of bytes read erasure-coded=0
  Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=3769
    Total time spent by all reduces in occupied slots (ms)=2694
    Total time spent by all map tasks (ms)=1769
    Total time spent by all reduce tasks (ms)=2694
    Total vcore-milliseconds taken by all map tasks=3769
    Total vcore-milliseconds taken by all reduce tasks=2694
    Total megabyte-milliseconds taken by all map tasks=3859456
    Total megabyte-milliseconds taken by all reduce tasks=2758656
  Map-Reduce Framework
    Map input records=136072
    Map output records=1392807
    Map output bytes=24870594
    Map output materialized bytes=5528735
    Input split bytes=104
    Combine input records=1392807
    Combine output records=218030
    Reduce input groups=218030
    Reduce shuffle bytes=5528735
    Reduce input records=218030
    Reduce output records=218030
    Spilled Records=436000
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=105
    CPU time spent (ms)=4186
    Physical memory (bytes) spilled=613052416

```

Figure 22 analyzing the data with mapreduce.

8. After the processing finished we wanted to show the result which is shown in the figure 23 the kurdish word has some problem. They are unicode hadoop cant show the same output so by this command **bin/hdfs dfs -cat /Awenaprocess/*** it shows us the output.

10. A Folded of notepad++ created to shows the correct output as shown in the figure below.

```

3283 1 1
3284 1 1
3285 1 1
3286 1 1
3287 12 12
3288 1 1
3289 1 1
3290 1 1
3291 1 1
3292 1 1
3293 2 2
3294 8 8
3295 1 1
3296 1 1
3297 14 14
3298 4 4
3299 14 14
3300 2 2
3301 1 1
3302 6 6
3303 1 1
3304 1 1
3305 1 1
3306 1 1
3307 1 1
3308 1 1
3309 1 1
3310 1 1
3311 3 3
3312 1 1
3313 1 1
3314 1 1
3315 1 1
3316 11 11
3317 1 1
3318 3 3
3319 7 7
3320 4 4
3321 6 6
3322 8 8
3323 2 2
3324 27 27
3325 1 1
3326 1 1
3327 1 1
3328 1 1
3329 1 1
3330 1 1
3331 20 20
3332 1 1
3333 5 5
3334 2 2
3335 1 1
3336 34 34
3337 4 4
3338 1 1

```

Figure 25 output of the processed data with notepad ++

The code found the word count for the data and we found a code that does those word counting processes as shown in the figure below.

11. Still the process isn't finished. We need to compare them with a list of bad words and sort them. Which bad word is used the most, we created a code to do this comparison and sort them as shown in the figure below.

PS C:\hadoop> & C:/ProgramData/Anaconda3/python.exe c:/Users/ADMIN-Lab239/Desktop/Bive-main/Bive-main/main.py

```

import os # operating system for file
from collections import defaultdict # collection package for extra datastructure

with open('Bive-main/badword.txt', encoding='utf-8') as infile: # opening file
    badwords = infile.read().split('\n')

#COUNTING PART
os.chdir("Bive-main/kurdish news data")
output = ''
for file in os.listdir():
    wordcount=defaultdict(int)
    with open(file, encoding='utf-8') as infile:
        words = infile.read()
        words.replace('\n', ' ')
        words = words.split(' ')
        for word in words:
            wordcount[word] += 1
# OUTPUT PART
output += file + '\n' # FILE ONLY HAVE THE NAME OF THE FILE THE NAME OF THE txt THEN GO DOWN FOR THE BAD WORDS DATA
badwords_count = [] #NEW LIST CONERT DICTIONARY OF THE KEY VALUE PAIR ( WORD:5) CANNOT BE SORTED [(WORD,5 )] THIS CAN BE SORTED ONLY THE
for badword in badwords:
    badwords_count.append((badword, wordcount[badword])) # [(5:خۆکوشتن)]

badwords_count.sort(key=lambda x: -x[1])# SORTING BY THE VALUE THE NUMBERS THE HIGHS FIRST
for word, count in badwords_count:
    output += f'{word}: {count}\n' # GET THE BAD WORD WITH ITS VALUE
output += '\n'

with open('./output.txt', 'w', encoding='utf-8') as o:
    o.write(output) # PRINT THE OUTPUT IN THE OUTPUT.TXT FILE

```

Figure 26 Code of comparing the bad word with processed output

12. Then the output was correct as shown in the figure below.

```

207 Awena.txt
208 1038 ده‌مێش
209 454 بڕیندار
210 184 تیرۆر
211 78 چه‌ك
212 72 كوشتن
213 68 چه‌ك
214 65 ده‌ستدرئێژی
215 52 ده‌رمان
216 50 كوشتن
217 20 ده‌مانچه
218 12 ئاگیرین
219 12 له‌شفرۆشی
220 9 جیه‌اد
221 9 جاش
222 8 خۆكوشتن
223 7 لاقه‌كردن
224 7 قومان
225 4 پۆزن
226 4 سێكس
227 4 سێكس
228 3 تالێیان
229 3 ده‌شیش
230 3 له‌شفرۆش
231 2 سمع
232 2 كێز
233 2 گێل
234 2 كه‌ر
235 2 لاقه‌
236 1 ئێكوژ
237 1 جیه‌اد
238 1 گه‌وج
239 1 گون
240 1 قوون

```

Figure 27 Number of repeated bad word in each content

6. Visualization

Visualization is displayed graphically using graphs and charts for a better comprehension of data analysis results, making it more approachable to most people, as the data studied is useless if it cannot be read by business owners. When properly aligned, data visualization may give a quicker route to assist decision making and become a tool to convey information vital in any data analysis. Various strategies and technologies are used to appropriately represent data in order to demonstrate patterns and trends. To illustrate the analytical results for our research, tableau and Microsoft Excel tools were used. In this project, Pie chart and Scatter used to represent semi-structured data from difference social media, websites and movie subtitles, as shown in figures below.

We used Microsoft Office Excel to demonstrate the percentage of inappropriate terms used by social media users (see figures 28 and 31), as well as movie subtitles (see figures 35). We compared each source separately for this purpose because the inappropriate words were different. Figure** shows the number of inappropriate words used by the Vejin website, which was also estimated using Microsoft Office Excel.



Figure 28 percentage of inappropriate words used by Rudaw.

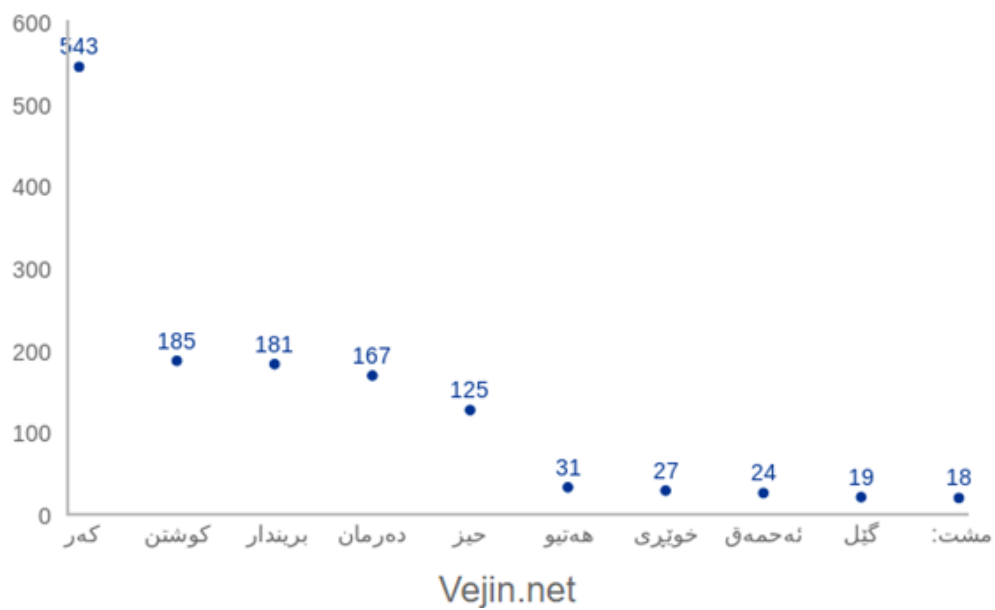


Figure 29 The number of inappropriate words used in Vejin, a Kurdish literature website.

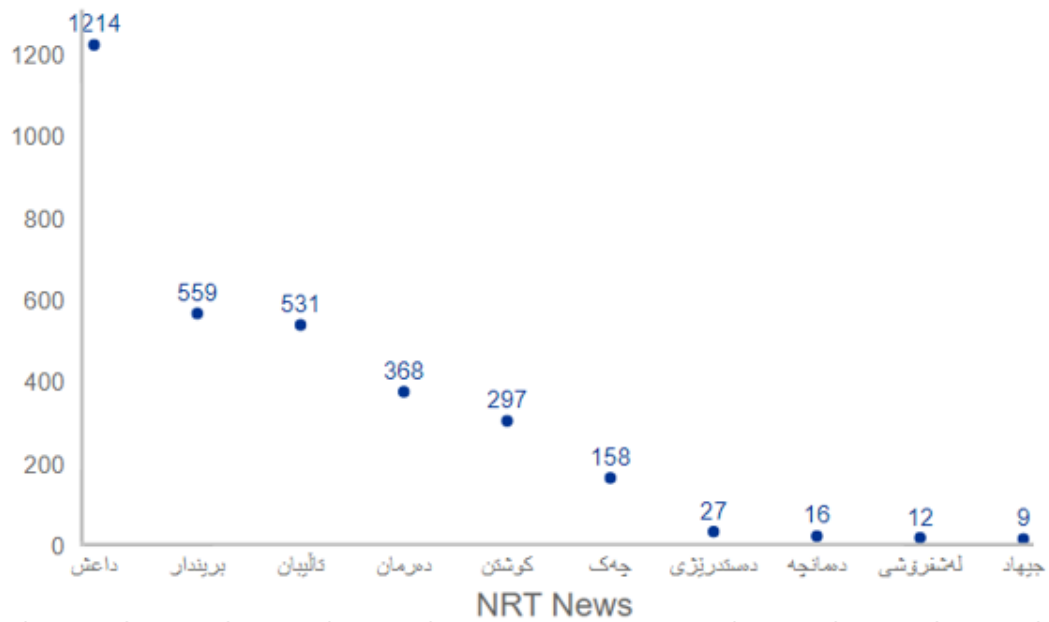
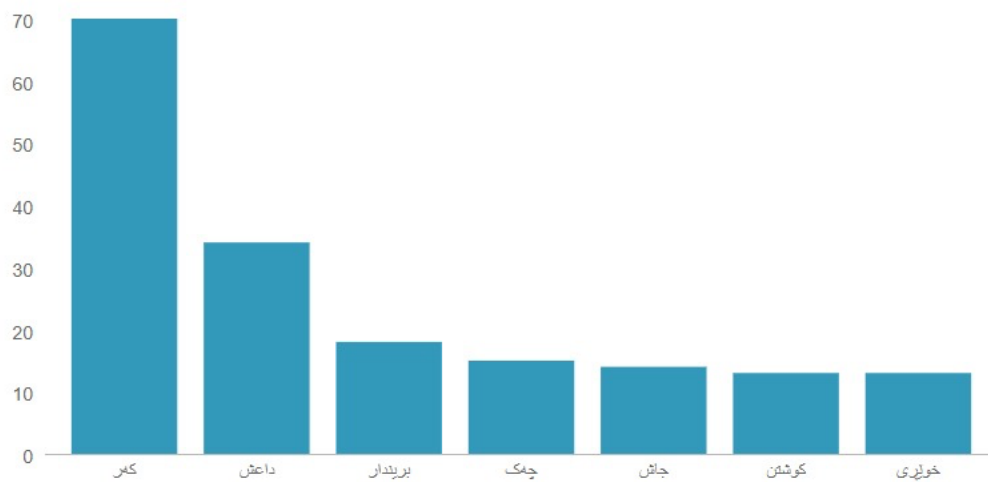


Figure 30 The number of inappropriate words used in NRT, a political website.



Twitter Posts

Figure 31 The number of inappropriate words used in Twitter replies/comments.

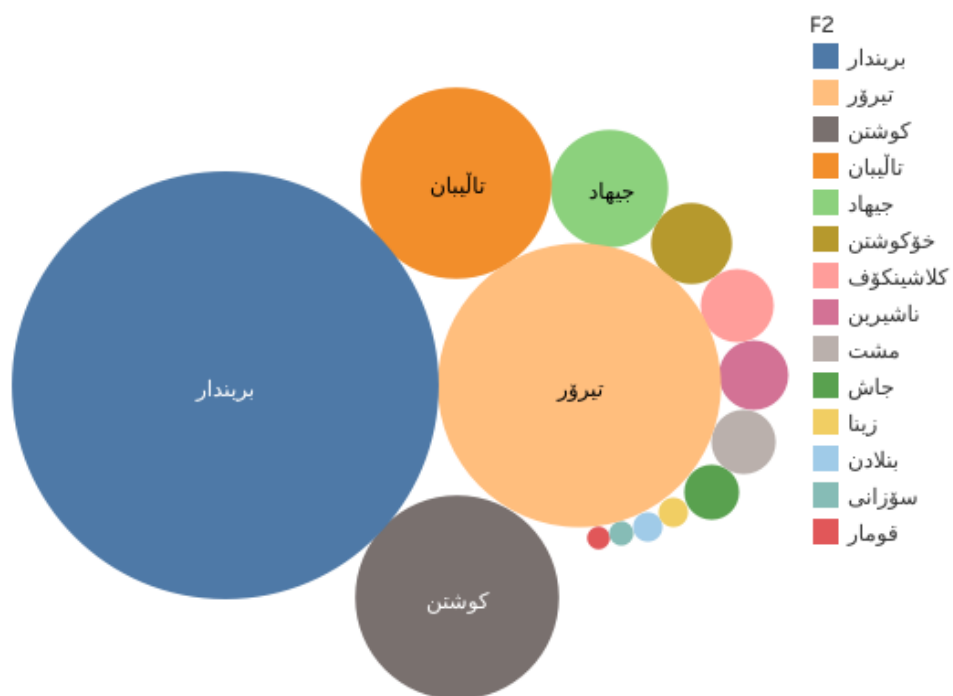


Figure 32 Most used bad words in a Book text corpus.

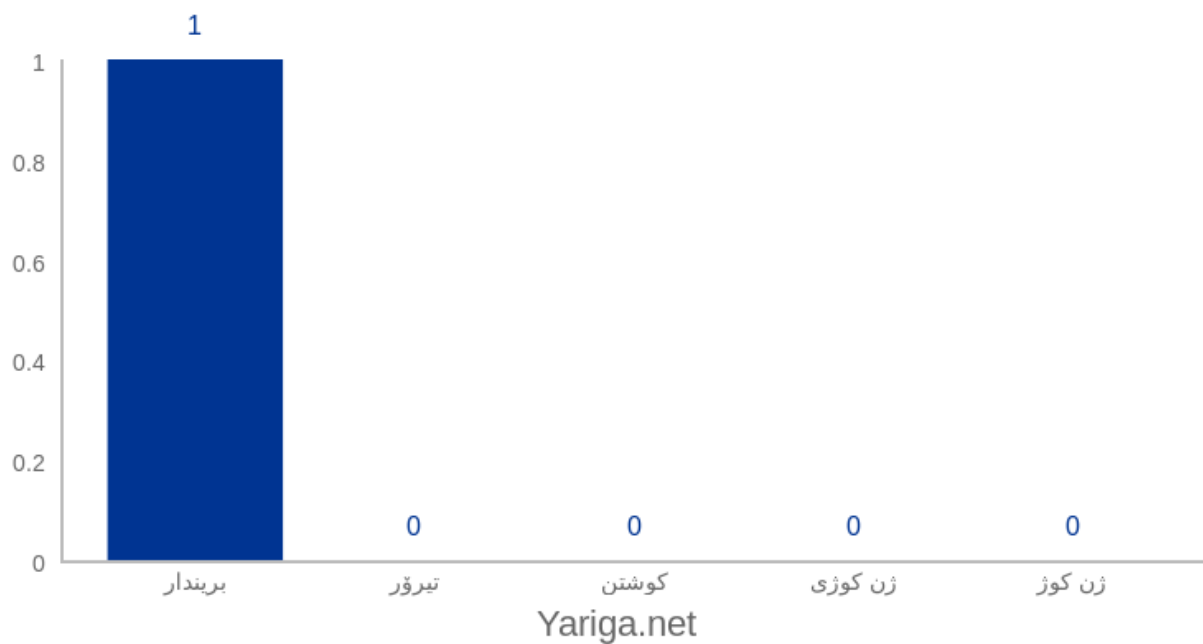


Figure 33 Shows that football and sport websites have least number of using bad words.

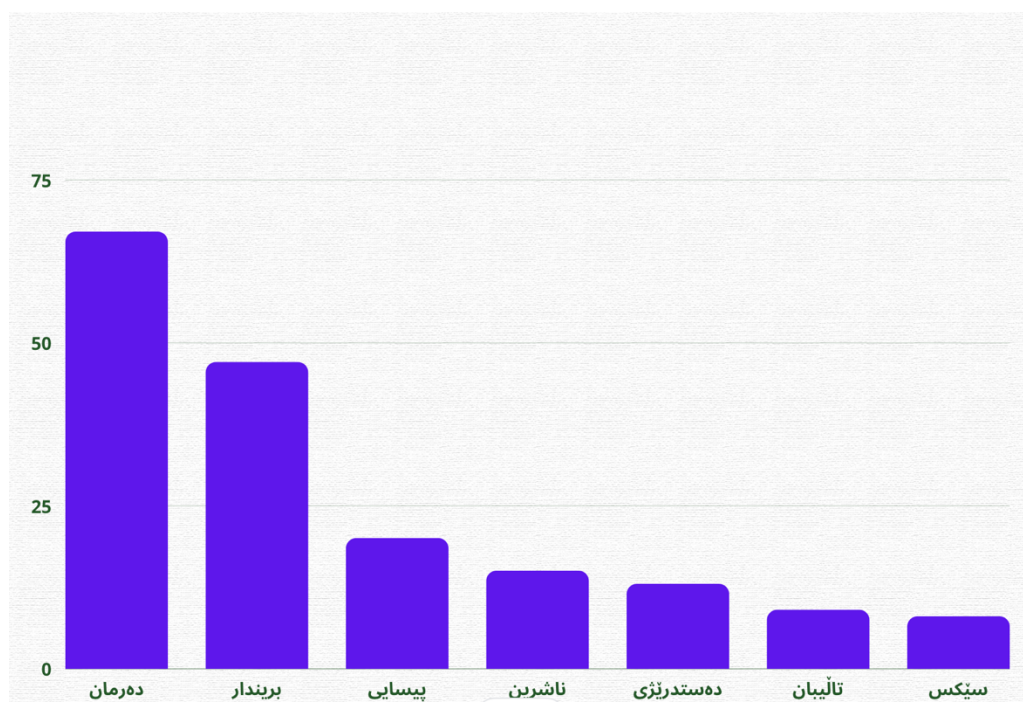
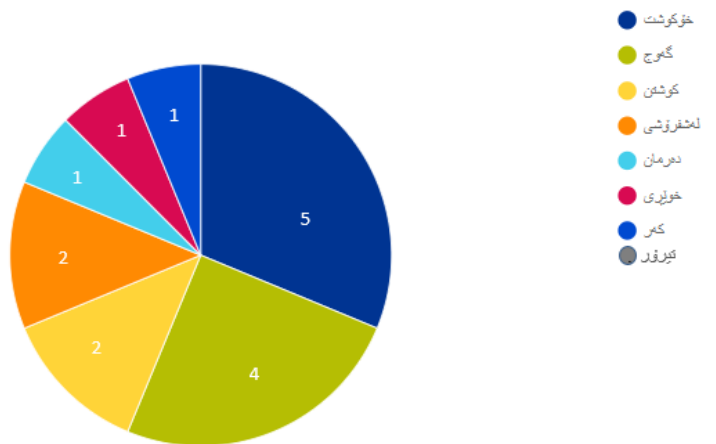


Figure 34 The number of inappropriate words used in NRT2, an entertainment website.

Movies



chartblocks

Figure 35 Most used bad words in movies corpus text.

7. Conclusion and Future Works

In this project, we analyzed different Central Kurdish (Sorani) resources. Yet, the project needs to collecting data and processing it continuously in order to be more effective for decision making. The limitation of our work is not supporting Kurdish Kurmanji (Northern Kurdish), Zazaki and other dialects. However, with providing data, we can follow same procedure for them as well.

In future, having a proper Kurdish speech-to-text generator will provide more data, so we can process profanity on video and audio files as well.

8. Acknowledgment

To complete this report we benefited from various open source projects and codes including Hadoop for analyzing, Facepager and published text corpuses for data collection stages. Sometimes only saying “Thank you” is inadequate. Thus, we decided to publish our report under “Creative Commons Attribution-NonCommercial 4.0 International License” and published our codes & data contribution on Github in return for what we gain from the open source community.[19]

9. Lesson Learned

Member names sorted in alphabetical order:

9.1 Mahmood Yashar (Data analyzer)

I am Mahmood, I was assigned to do the data analysis of this project, and since my role was to do the data analysis, I did some research about the Apache Hadoop how to install software and how the Apache Hadoop is working and how to process those data's which were collected by my colleges. For both unstructured and semi-structured I used Apache Hadoop to process the data's, so I started to learn how to install Hadoop and how to use the Hadoop, it took me some time to fix the errors before starting the Hadoop, after fixing the errors then I started to learn how to process those data's, I applied the word count class which I found multiple tutorial about the word count which simply counts the repeated word inside the data's. but I found that the data's cannot be processed just by writing command to the Hadoop to process the data's, inside the Hadoop there is HDFS (Hadoop Distributed File System) which the data's must be stored there to be processed otherwise the data cannot be processed, after processing the data I faced another issue the Hadoop doesn't support Kurdish language it was counting the words but it was messing the Kurdish word then I tried to open the processed data with notepad++ to don't mess the Kurdish word because till now Hadoop doesn't support Kurdish language. After getting the output still the process wasn't complete because we have to compare the processed words with the bad words that we have which is around 200+ bad words the focus was not on all data's we wanted to know how many times those bad word where repeated for each resource, then I created a simple algorithm just to compare the processed data with the bad word and sort them form highest amount of bad word to least amount.

Finally, when is started this course I didn't know that organization use big data to find information about those data especially this project made me to learn how the data's are useful and how to process those data and get the results from the data's.

9.2 Mohammed Sardar Noori (Team Leader)

First, Thanks to my team members for choosing me as team leader and gave me this opportunity. I've been team leader before on my graduation project for BSc degree. However, this experience was completely different. beginning from choosing members, unlike the bachelor degree

experience this time we randomly selected team members and I got chance to work with people which I didn't know about their personality, academic, technical backgrounds & skills. Unfortunately, one of our members (Shaimaa) traveled for one month. Nevertheless, we were still lucky because we knew that from day one itself not in the middle of works.

All of mentioned challenges forced me to think about building communication channel and chemistry between members. During the first week we finalized the way to communicate through online tools/platforms, yet we feel lacks of chemistry and co-working. We solved this issue by spendings some times and talking with each other in order to know about each to start the project and dividing tasks. From my previous experience, I thought online meeting is not effective way to process and perform well in projects. Well, this project proved me wrong that we can do that with online as well. All the thing that you need is vision and having a bigger picture of what are we doing.

After that, I tried to use all human resources to engage the project. Despite the fact that we distributed the roles and responsibilities based on skills, profession and interest. I let each member to engage and work with each other in almost stages to boost process, group thinking, better communication and learning about each big data processing. At the end, our group working seems as peer to peer rather than in a hierarchy system.

Overall, I really enjoyed working on this project; starting from leading team to practicing each stage of big data processing, finding solutions for technical challenges and even writing a single line command for Linux/MS Windows. Doing those technical stuffs teaches me that even for manager and higher position it will be really helpful to have technical background, that is not only for communication but for sort of coworking as well. Furthermore, I understood better in open-source community as it was really help us to complete this project.

9.3 Shaimaa Salih (Data Collector & Documenter)

My responsibilities included data collection (with the help of Mahmud and Yasmin) and meeting documentation and reporting part of the project with Mohammed.

Another thing I learned was that keeping up with my teammates while I was away for work-related trainings was challenging, especially with our Software engineering classmates because we didn't know each other before, despite having established good communication channels, so it was challenging to keep up via online meetings.

9.4 Yasameen Sami (Data Visualizer)

As a visualizer my role was to collect the processed data from the data analyzer Mahmood which proceeded those data by Hadoop the data where not easily understandable so for this purpose we use visualization. As the result using charts or graphs to portray large amounts of complex data is easier on the human brain than trawling through spreadsheets or reports, visualizing data is particularly important when it comes to expressing data to customers.

Through undertake this project, I was in charge of final step of project which is effectively the visualizing data by employing numerous tools. When it came to visualizing semi-structured data, I used Microsoft Office Excel. Moreover, I utilized a platform called Tableau to make the outcome easier to understand of unstructured data, which allowed me to show the data in a variety of ways utilizing various sorts of visualization.

This project taught me how to organize my time and collaborate with other team members in order to conduct online and face-to-face meetings. This task also prompted me to learn more about data analytics and become acquainted with other visualization tools, such as Tableau, which allowed me to visualize the analytical results in a variety of ways.

10. References

- [1] J. Naughton, “The evolution of the Internet: from military experiment to General Purpose Technology,” *Journal of Cyber Policy*, vol. 1, no. 1, pp. 5–28, Jan. 2016, doi: 10.1080/23738871.2016.1157619.
- [2] S. O. Sood, J. Antin, and E. F. Churchill, “Profanity use in online communities,” *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1481–1490, 2012, doi: 10.1145/2207676.2208610.
- [3] S. Ahmadi, “A Rule-Based Kurdish Text Transliteration System,” *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 18, no. 2, Jan. 2019, doi: 10.1145/3278623.
- [4] Fast-link.com, “Family Offer,” 2022. https://www.fast-link.com/Offer_Details_New/38 (accessed May 19, 2022).
- [5] “ftkurt/kurdish-twitter-data: Kurdish twitter data repository for Kurmanji and Sorani dialects.” <https://github.com/ftkurt/kurdish-twitter-data> (accessed May 21, 2022).
- [6] “allekok/kurdish-data-sources: سەرچاوەکانی دەیتای زمانی کوردی.” <https://github.com/allekok/kurdish-data-sources> (accessed May 21, 2022).
- [7] “For researchers, accessing data is one thing. Assessing its quality another. - Algorithm Watch.” <https://algorithmwatch.org/en/research-data-quality/> (accessed May 22, 2022).
- [8] “strohne/Facepager: Facepager was made for fetching public available data from YouTube, Twitter and other websites on the basis of APIs and webscraping.” <https://github.com/strohne/Facepager> (accessed May 22, 2022).
- [9] “dolanskurd/kurdish_news: Kurdish News text corpus.” https://github.com/dolanskurd/kurdish_news (accessed May 21, 2022).
- [10] “Google Colaboratory – How to Run Python Code in Your Google Drive.” <https://www.freecodecamp.org/news/google-colaboratory-python-code-in-your-google-drive/> (accessed May 21, 2022).
- [11] “Extract Text from subtitle, remove timestamps.” <https://anatolt.ru/t/del-timestamp-srt.html> (accessed May 21, 2022).
- [12] Diyako Hashemi, “فێرگهی زمانی کوردی – ڕێنووس,” *Yagey Ziman*, Jun. 09, 2013. <http://diyako.yageyziman.com/%DA%95%DB%8E%D9%86%D9%88%D9%88%D8%B3/#10> (accessed May 19, 2022).
- [13] “ویکیپیدیا، نێنسیایکۆڵیپیدیاى نازاد - هـ.” <https://ckb.wikipedia.org/wiki/%DA%BE> (accessed May 19, 2022).
- [14] “Kurdish Unicode Keyboard.” http://unicode.ekrg.org/ku_unicodes.html (accessed May 19, 2022).
- [15] “رێمی کوردستان ستانداردی یونیکۆدی زمانی کوردی له سەرجههه زیرانی ههروێك و مهنچیرفان بارزانی سه- دامهزراوهكانی حكومهت به فهرمى كرد” *KRG*, Dec. 15, 2014. <http://previous.cabinet.gov.krd/a/d.aspx?s=040000&l=13&a=52717> (accessed May 19, 2022).
- [16] “Kurdînûs 2019.” <https://kurdinus.com/2019/Normalize.html> (accessed May 21, 2022).
- [17] “Kurdinus/kurdinusLibrary: JavaScript tools for normalization and transliteration of Kurdish texts.” <https://github.com/Kurdinus/kurdinusLibrary> (accessed May 21, 2022).

- [18] “Run Hadoop Wordcount MapReduce Example on Windows - SrcCodes.”
<https://www.srcodes.com/run-hadoop-wordcount-mapreduce-example-windows/>
 (accessed May 21, 2022).
- [19] “MohammedSardar/Bive: Bive is a Kurdish profanity language processing project.”
<https://github.com/MohammedSardar/Bive> (accessed May 22, 2022).

Meeting Documentation

Date / 2022	Communication Method	Outcomes/Decision
7th March	Face to face	<ol style="list-style-type: none"> 1. Group formed 2. Members introduced themselves and their technical skills. 3. Communication tools decided (Whatsapp, Google Docs and Google Meets)
9th March	Stand-up meeting	<ol style="list-style-type: none"> 1. Brainstorm session started and we decided to have one free week to choose either topic or find unstructured data. 2. Hadoop introduction by Mahmood
11th March	Whatsapp/Youtube	<ol style="list-style-type: none"> 1. Mohammed shared his idea via Youtube video. 2. We decided to continue one free week to bring new ideas.
15th March	Online: Google Meets	<ol style="list-style-type: none"> 1. All members agreed with Mohammed’s idea and shared their inputs. 2. We discussed how to narrate scenarios and how to collect data on that area. We choose Fastlink case study as our scenario.
Newroz Week	<ol style="list-style-type: none"> 1. Mahmood and Yasemin collected data from social media (Facebook & Twitter). 2. Shaimaa started to document meetings on Google Docs. 3. Mohammed opened the primary report on Google Docs & he traveled. 	
30th March	Online: Google Meets	<ol style="list-style-type: none"> 1. Primary report and presentation ready (All members involved) 2. Mahmood and Yasemin showed their data 3. Mohammed showed his data which collected from various Kurdish resources (News, Kurdish literature, films and other web contents)

28th March	Shaimaa traveled.	
1st April	Coordinating with some people in Fastlink	
2nd April	Primary report submitted	
4th April	Primary report presented in the class.	
17-24 April	Midterm exam break	
28th April	Online: Google Meets	<ol style="list-style-type: none"> 1. Scenario & KPI reviewed 2. Moving to open source idea approved 3. Data will be shared on GitHub 4. Searching to find analyzing process started 5. Data cleaning and exploration started
11th May	Shaimaa is back from her business trip.	
14th May	Online: Google Meets	<ol style="list-style-type: none"> 1. Resume working on reports 2. Data cleaning and exploration done
16th May	Online: Google Meets	<ol style="list-style-type: none"> 1. Analysis process done. 2. How to use Github explained. 3. Planning for finalizing the visualization, introduction and business case study section started.
17th May	Fastlink Meeting scheduled on Thursday	
18th May	In-person meeting	<ol style="list-style-type: none"> 1. Review all required parts 2. Exchanging ideas about visualization processes discussed.
19th May	In-Person Meeting	1- We met Fastlink Admin Manager
21st May	Online: Google Meets	<ol style="list-style-type: none"> 1- Yasemin showed visualization works. 2- Finalizing report discussed
22nd May	<ol style="list-style-type: none"> 1- Final Report finished. 2- Report submitted via Moodle. 	

Table 2, Details of meetings and its achievements.