CHAPTER 4

DATA ANALYSIS AND INTERPRETATION

This chapter described different phases of data analysis concerning knowledge base, lexicon, and dataset creation. The phases described related to data analysis for lexicons and knowledge base creation is the research question, collecting data for knowledge base and polarity lexicon (data acquisition, born-digital data), data preprocessing, quality data, compiling data, labels and metadata, data analysis, and interpretation for the knowledge base and polarity lexicon. The steps related to data analysis for lexicons and knowledge base creation are research question, collecting data, organizing data(data consolidation, data separation and conversion, language identification and data separation, convert to single language data), pre-processing(cleaning, transformation, and reduction), annotation, method selection, data analysis, and interpretation.

4.1 Knowledge Bases

4.1.1 Research Question

What are the impacts of creating or not creating knowledge bases or polarity lexicons for the multilingual sentiment analysis?

4.1.2 Collecting Data for Knowledge Base and Polarity Lexicon

This study tried to create the Bengali Knowledgebase to deal with Bengali sentences and named it 'BanglaSenticNet.' The new method is shown in Figure 3.2. The process starts with collecting data in the form: 1) Data acquisition 2) Born-digital Data. The KB is concept-based and general; therefore, the data were collected from different relevant dictionaries (hard copies), reports, newspapers, internet sources, social media, and generic knowledge sources such as SenticNet 5[5], NRC emotion lexicon[20], Bangla Dictionary master[21], lexical DB Bangla master[22].

4.1.2.1 Data Acquisition

This section is so named to mean that the data collected in this category need conversion before placing them as raw data. Generally, these data are available in hard copies (dictionaries, reports, newspapers) or not directly readable soft copies (scan documents). The data from all these sources are collected and converted to digital form first.

4.1.2.2 Born-digital Data

This section is so named to mean that the data collected in this category need no conversion before placing them as raw data. These data are created initially in a digital form, which in short is named "Born-digital Data." Generally, these are collected from different internet sources (Web pages and blogs), social media (Facebook), and generic knowledge sources such as SenticNet 5[5], NRC emotion lexicon[20], Bangla Dictionary

master[21], and lexical DB Bangla master[22] and did not convert because these data are already in digital form.

4.1.3 Data Preprocessing

Dealing with data need preprocessing in all the research practices. The knowledge base and polarity lexicon also require high-level preprocessing to bring our most important concepts. Therefore, this research has applied different preprocessing techniques such as text tokenization, removing punctuation, removing numbers, removing emoticons, case conversion and normalization, stemming, lemmatization, stop word deduction, reduction of letter repetition, and negations on the collected data. Consider the following example (this text is taken from the raw data) before and after preprocessing applied.

Before preprocessing:

Bengali-> R8 $$ দূর্লব কিছু ছবি LOLIIIIIIIII

English-> R8 $$ Some Rare pictures LOLIIIIIIIII

After preprocessing:

Bengali-> দূর্লব ছবি

English-> Rare pictures

4.1.4 Quality Data

Quality data has a significant role in knowledge base and polarity lexicon creation. It also has equal importance in the evaluation process. Therefore, this research

has paid a great emphasis on data selection. The data are only selected if it relates to the domain of study (student feedback, cricket, and movie review). Considering the following two examples:

Example 1:

Bengali-> আমার ভালবাসার আভাসভূমি প্রিয় দেশ

English-> My country is my place of love

This sentence has no relation with the domain mentioned above; therefore, these data are not included in the quality data list.

Example 2:

Bengali-> ইন্টারনেট সুবিধা বাড়াতে হবে

English-> The internet facilities should be increased

In this example, the data are related to student feedback. However, all these keywords are not that important. In this particular example, “internet” (ইন্টারনেট), “facility” (সুবিধা) and “increase” (বৃদ্ধি) are essential keywords. Therefore, only those data are kept for knowledge base and polarity lexicon building.

4.1.4.1 Compiling Data

In this sub-section, the dependency rules from Section 3.2.1 (concept extraction) were applied to the data from the quality data sub-section. These rules helped derived the concepts for both knowledge base and polarity lexicon. For instance, the concept derived from above example 2 using dependency rules is ”increase_facility” (সুবিধা_বৃদ্ধি).

4.1.5 Labels and Meta Data

This section deals with one of the significant and challenging issues of knowledge base and polarity lexicon creation. The issues are labeling and meta data tagging to the concepts. Polarity lexicons need the only assignment of concepts polarity value. However, knowledge base demands assigning polarity labels and semantics of the concepts too with the concepts. Besides, Pleasantness, Attention, Sensitivity, Aptitude values are also added. Therefore, in the lexicon, the assigned polarity value for ”increase_facility” (সুবিধা_বৃদ্ধি) is 0.75 (methods in Section 3.2 and 3.3). However, the values for the same concept in the knowledge base are Pleasantness (0.827), Attention (0.732), Sensitivity (0), Aptitude (0.769), polarity level (positive), and semantics ('ইন্টারনেট_অ্যাক্সেস', 'অনন্য_সুযোগ', 'দক্ষতা_বৃদ্ধি', 'দ্রুত_ভ্রমণ', 'সক্রিয়'). It is clear from the methodology (Section 3.2) that different values for the concept could also be derived using the semantics of that concepts. For instance, if the knowledge base does not have the concept ”increase_facility” (সুবিধা_বৃদ্ধি), then the semantics such as 'ইন্টারনেট_অ্যাক্সেস' could be used to derived the value of ”increase_facility” (সুবিধা_বৃদ্ধি).

4.1.6 Data Analysis and Interpretation for the Knowledge Base and Polarity Lexicon

A sample of the polarity lexicon is presented in APPENDIX 2. Table 4.1 and Figure 4.1 show the overall statistics of the lexicon used in this research. Different lexicons such as SenticNet 5 English polarity lexicon[18], opinion-lexicon-English[19], NRC-

[18] https://sentic.net/downloads/

Emotion-Lexicon-v0.92-Bangla[20], Bengali Dictionary master[21], and lexical_db_bangla-master[22] were used in this research to evaluate the performance of the Bengali concept polarity lexicon (created in this research). The table and figure also reveal that existing lexicons have very few numbers of words except SenticNet 5. Besides, the lexicons are reported to have different numbers of words or concepts. SenticNet 5 has the highest, and lexical_db_bangla-master has the lowest number of words or concepts. The lexicons are created by assigning different types of polarity such as 'positive,' 'negative,' and some cases 'neutral.' Revising available resources seems no available Bengali concept polarity lexicons; only word-level polarity lexicons are available. Therefore, the Bengali concept polarity lexicon could be considered the rich resource for CLSA with 72433 concepts, where 36309 concepts are positive, and 27085 concepts are negative. However, no neutral concepts are found. There is a difference in the number of positive and negative concepts, where more positive concepts are available than negative concepts. However, these problem is dealt with allowing the same number of concepts in the analysis. Moreover, to include all the positive concepts in the analysis, the concepts are selected on a k-fold basis. The experiment continues till the last fold of concepts included in the analysis.

[19] https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/tree/master/data/opinion-lexicon-English
[20] https://github.com/sebastianruder/emotion_proposition_store/tree/master/NRC-Emotion-Lexicon-v0.92
[21] https://github.com/MinhasKamal/BengaliDictionary
[22] https://github.com/abhishekgupta92/lexical_db_bangla

Table 4.1: Statistics of Lexicons

Polarity Lexicons	Words/Concepts			
	Total	Positive	Negative	Neutral
SenticNet 5 English lexicon[18]	100000	54937	45063	0
Created Bengali lexicon	72433	36309	27085	0
opinion-lexicon-English[19]	6790	2007	4783	0
NRC-Emotion-Lexicon v0.92 Bangla[20]	14182	2312	3324	8546
Bengali Dictionary-master[21]	17297	5325	6812	5160
lexical_db_bangla-master[22]	3396	1200	1800	396

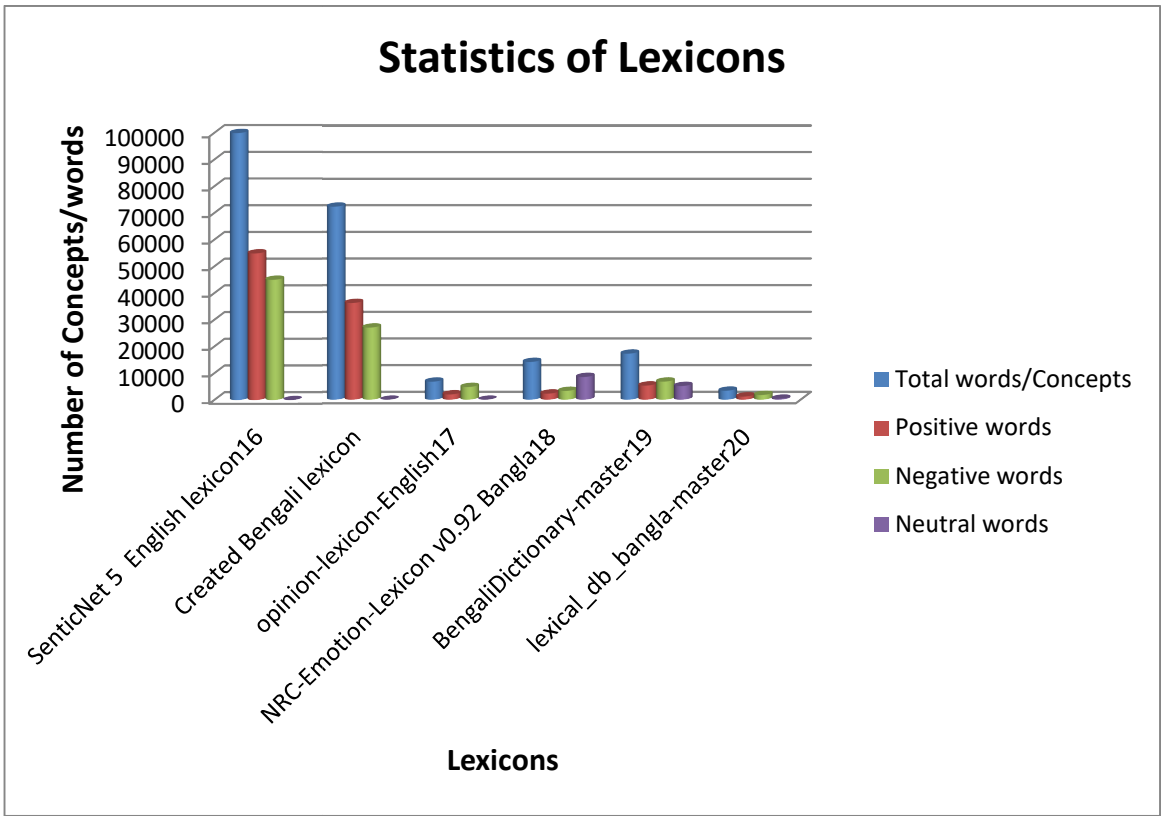

Figure 4.1: Statistics of Bengali and English Lexicons

Some sample examples of BanglaSenticNet entries are represented in Table 4.2 and appendix 3. We have ignored all the states in Table 4.2 other than positive and negative states for easy representation, though the basic version of BanglaSenticNet contains all the emotional states, as shown in Table 4.3. However, due to the importance, the semantics of concepts are shown. These semantics can be considered helpful metadata that could help analyze data even if main concepts are not available in the dataset. These semantics also help in establishing semantic relatedness between concepts and derive chain relationships.

Both Table 4.3 and Figure 4.2 show the Statistics of Bengali and English SenticNet (common sense knowledge base) used in this research. For English, a benchmark knowledge base such as senticnet5 was used, which consists of 100000 concepts and 500000 semantics of those concepts (details in Section 3.2) and consists of some emotional states such as positive concepts, negative concepts, sadness, disgust, joy, surprise, anger, fear, admiration, and interest. For example, SenticNet 5 has 55191 positive concepts and 45193 negative concepts. However, after carefully consolidating concepts from different sources (as described above), BanglaSenticNet has 30000 Bengali concepts (may be extended in further research) and 150000 semantics of those concepts where, for example, 15126 concepts are positive, and 14875 concepts are negative. Table 4.4 and Figure 4.2 also reveal that the concepts that correspond to emotional states are not the same. The SenticNet 5 and BanglaSenticNet both have maximum concepts (36789 and 10306) on emotion state ‘admiration’. At the same time, emotion states having minimum concepts are ‘fear’ in SenticNet 5 with 10731 concepts and ‘surprise’ in

BanglaSenticNet with 4570 concepts. The problem with number of concepts in knowledge bases is handled by allowing the same number of concepts while analyzing.

Table 4.2: Some example of BanglaSenticNet Entries

Concept name	Polarity level	Polarity Value	Semantics1	Semantics2	Semantics3	Semantics4	Semantics5
একটু	Negative	-0.79	অন্তত	ছোট	ছোট_পরিমাণ	অভাব	ক্ষতিকারক
একটু_ক্ষুধার্ত	Positive	0.65	পূর্ণ_হও	ক্ষুধা_দূর_হয়ে_যাওয়া	পূর্ণ_অনুভব	ক্ষুধা	পূর্ণ
একটু_নির্দিষ্ট'	Positive	0.065	ক্ল্যান	একসাথে_সুখী	অনেক_মানুষ	মাসিমা_চাচা	মানব_গ্রুপ
অনেক	Positive	0.335	অনেক	প্রচুর	বড়_পরিমাণ	প্রচুর	ভাল_সংখ্যা
অনেক_বই	Positive	0.081	লাইব্রেরি	পুরানো_নিয়মাবলী	পাকপ্রণালীর_বই	পত্রিকা	গ্রন্থাগারিক

Table 4.3: Statistics of Bengali and English SenticNet

				Emotion States							
SenticNet	**Number of concepts**	**Positive concepts**	**Negative concepts**	**sadness**	**disgust**	**Joy**	**surprise**	**Anger**	**fear**	**admiration**	**interest**
English	100000	55191	45193	27394	31351	31608	16145	22036	10731	36789	26523
Bengali	30000	15126	14875	7862	8933	8926	4570	6139	3132	10306	7208

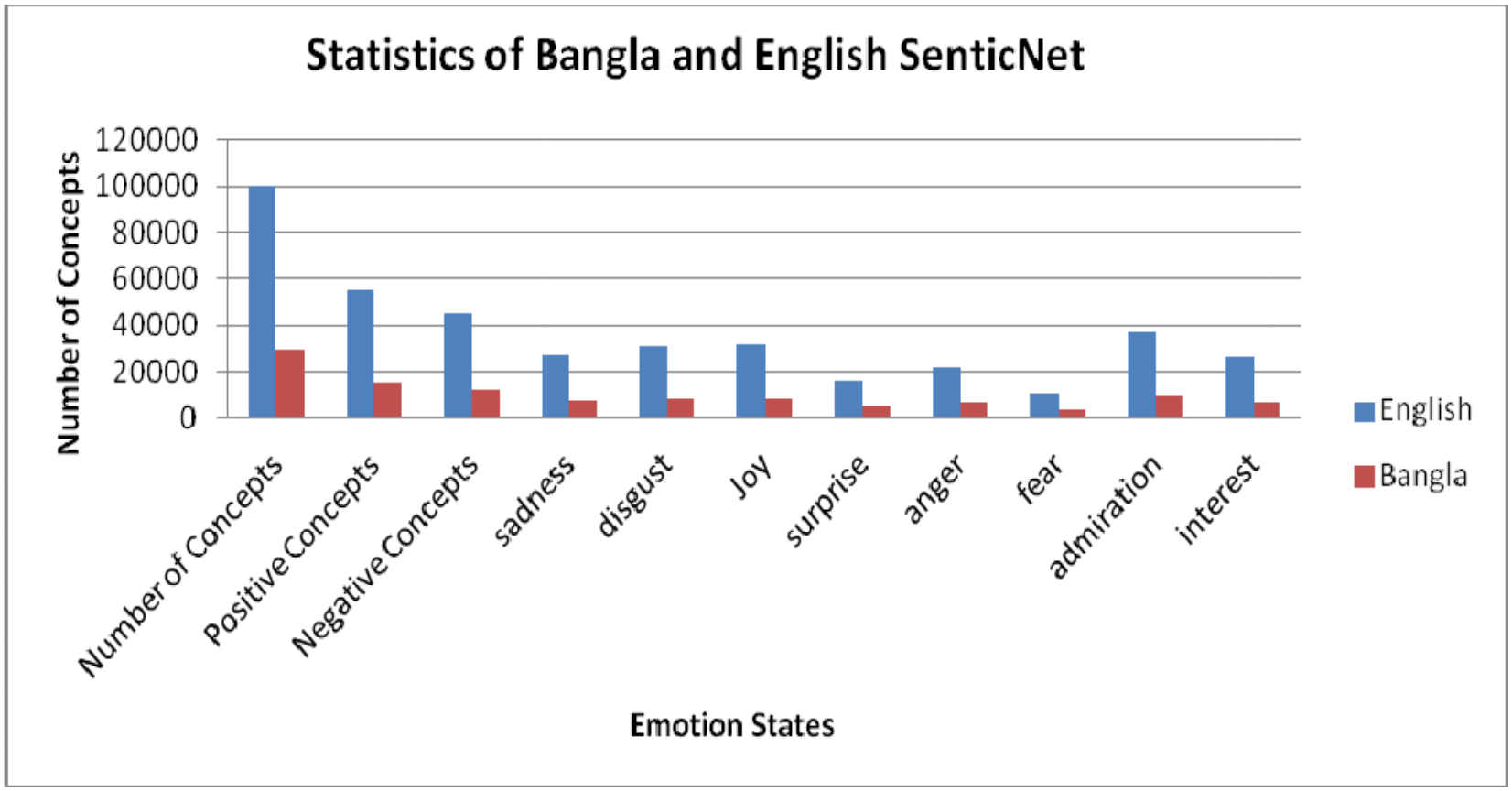

Figure 4.2: Statistics of Bengali and English SenticNet

4.2 Data Sets

Social media, especially Facebook, has observed a massive growth of regular posts and their related comments in recent years. The users are free to post and comment in any language, and there is a lack of explicit mechanisms to reconcile the information expressed in different languages into the useful data set. In most cases, the contents of Facebook expressed in different languages remain useless. Therefore, there is a need for data sets (is the unit to compute the information accessible in the public open data repository) to be built with a data collection system to get helpful information. The data collection system is a computer-based system that assists data gathering, allowing certain, well-formed information to be collected systematically, finally, facilitating data analysis to be done on the information[23].

4.2.1 Research Question

What are the impacts of creating or not creating Bengali and English Student feedback datasets?

4.2.2 Collecting Data for Dataset

One of the essential tasks of any research is the collection of data. Similarly, significant is to identify sources of the data. This research aims to mine the data of the student feedback from among many possible domains. Hence, social networking sites such as Facebook, Twitter, and weblogs are considered data sources for the said domain. However, this research used only Facebook as the primary source of data as

[23] https://en.wikipedia.org/wiki/Data_collection_system

Facebook messages are not limited in size like tweets and become more eligible for study in the said domain.

The data on Facebook are semi-public and could only be collected from groups or pages. Facebook allows data to be easily accessed from public groups; however, public groups of educational institutions mostly do not contain essential data as the audiences of those groups are reluctant to disclose the institution's negativity outside the world. Moreover, the private groups contain factual information about the institution but are not publicly available and need additional permission for access. Therefore, data were collected from some public pages, groups, and some private groups (with necessary authorization from the admin of those groups), related to the target institution 'International Islamic University Chittagong (IIUC)'. This institution is selected for easy access to private group data.

The expressed views (such as comments, status, and updates of status) in the Facebook as mentioned above pages and groups were collected with the use of Facebook application programming interfaces (API), also known as "Graph API" and python crawler. The used script has successfully grabbed the required data and saved it in the Comma Separated Value (CSV) file. Generally, the expressed views on Facebook contain both structured data (such as user profile, number of likes, spatial data, thematic and temporal data) and unstructured data (such as content shared by users, comments, etc.).

Both structured and unstructured data were traced out by different tracing approaches such as *self-involved* - the approach is suitable, for example, when the individual (for instance, tutor or an Institution) is interested in knowing themselves looking at the students post about them on social media. This research has collected

all the posting done in predefined external sources (Facebook pages/groups) in favor of their name, as those found useful in many cases.

Topic-based approach - This approach is suitable, for example, when an individual tutor or institution is highly interested in knowing specific educational topics (e.g., keywords) from the expressed opinions in their predefined pages or groups. In this approach, exciting topics are carefully chosen in advance to achieve a high level of data comprehensiveness. The more the keywords to be analyzed, the more the keywords need to be considered in advance (Stieglitz et al., 2012). In this research, topic-based is also used, as the aim is to know the institution's overall sentiments.

4.2.3 Organizing Data

The data were organized for processing using the following simple steps:

4.2.3.1 Data Consolidation

Once successfully grabbed the data, the next task is selecting and consolidating data from different pages and groups, which is easier if the groups and pages are in a similar context or institution. In this research, the data of different groups and pages of the said institution is selected and consolidated.

4.2.3.2 Data Separation and Conversion

It is essential to separate and convert the structured and unstructured data to bring those in a unique format. This research separated and converted the data manually due to difficulties in implementing through algorithms and the presence of multi-lingual and high-volume unstructured data.

4.2.3.3 Language Identification and Data Separation

A challenging part of multilingual dataset creation is identifying diverse lingual data and carefully separating it from collected data for easy processing. The collected data were a mixture of English and Bengali lingual data and separated based on language similarity.

4.2.3.4 Convert to Single Language Data

It is recommended to work with the different lingual datasets separately (English and Bengali) for easy processing. However, this research has worked with both separate and multilingual datasets. The dataset is made multilingual by bringing the data from both lingual datasets created in the early stage.

4.2.4 Data Pre-Processing

4.2.4.1 Data Cleaning

Data cleaning is an essential step in making datasets perfect for further processing. This research step dealt with missing values, data inconsistencies, and relevant noise in the collected data. Here, unnecessary punctuations, stop words,

letters, irrelevant words, symbols, irrelevant comments and status, extra spaces, and numbers were removed.

4.2.4.2 Data Transformation

The collected data were normalized, which reduced data redundancy and different forms of anomalies. The normalization was done by creating new attributes, aggregating attributes, and finally, discretization of data. Here, a new attribute name, 'polarity,' was created. The status and comment attributes were aggregated. Finally, labeled the data into finite value 0 or 1.

4.2.4.3 Data Reduction

This step is the final step of the data preparation process. In this step, unnecessary variables (except comments and polarity variables) and cases (comments and status) were reduced to bring the dataset into the desired format.

4.2.5 Data Annotation

Data annotation plays a vital role in SA. However, the annotation is needed to be accomplished with accuracy and comprehensiveness. The machine learning algorithms work better with accurate training data, and these data come from a great deal of annotation. This research used both human experts and algorithms for data annotation. After separating and pre-processing the comments and status, those are then annotated with the labels 'positive', 'negative,' and 'neutral.' The accuracy is tested by both human experts manually and by machine learning algorithms. The

accuracy result achieved by the machine learning algorithm shows, the annotation is 99.32% accurate.

4.2.6 Method Selection

This research used the qualitative data analysis method to derive data via words, symbols, images, and observations and does not use statistics. There are many qualitative data analysis methods, such as content analysis, narrative analysis, and grounded theory. This research has used grounded theory for data analysis as this method has achieved a standard for social research. The collected data (as mentioned above) were first coded (labeled the data), then the data have been conceptualized (groups the code with similar ideas). The research then categorized the data (brought the concepts of similar ideas to entities. Finally, these entities provide us the theme of the dataset.

4.2.7 Data Analysis and Interpretation

Three datasets have been created with the process discussed and adopted above. Two of them are English and one Bengali dataset. Besides, one English and one Bengali dataset were used for validation purposes. Table 4.4 and Figure 4.3 illustrate the statistics of different datasets used in this research. From Table 4.4 and Figure 4.3, it is evident that this research considered three datasets for English with 3884, 4161, and 1000 comments in Dataset1, Dataset2, and Dataset3 respectively, where total comments are 9045, average words in comments is 54, In total, positive, negative, and neutral comments are 3699, 2591, and 2754 respectively. These statistics indicate that created datasets have a higher number of comments. These

types of datasets are chosen to understand the effect of data sample size in analysis. Thus reviewed the sample size till a sufficient sample size is achieved. The statistics also reveal that there are fewer negative comments, especially in dataset 2. This problem is dealt with using the same number of positive, negative, and neutral comments in the training.

Moreover, two datasets (Dataset1 and Dataset2) were considered for Bengali language text with 14766 and 2979 comments. On average, the comments contain approximately six words. In total, positive, negative, and neutral comments are 9118, 6272, and 2355 respectively. The same interpretation as English datasets could be drawn for this dataset. Also, resampling and restructuring of these datasets were continued until satisfactory evaluation results were achieved.

Table 4.4: Statistics of Datasets

Datasets	Number of comments	Total words	Average words in Comments	Positive comments	Negative comments	Neutral comments
English Dataset1	3884	193631	50	1213	1180	1491
English Dataset2	4161	280303	67	1992	906	1263
English Dataset3 (IMDB)[12]	1000	6855	7	495	505	0
Bengali dataset1	14766	99998	7	8552	4120	2094
Bengali dataset2 (ABSA-Cricket)[13]	2979	5016	2	566	2152	261

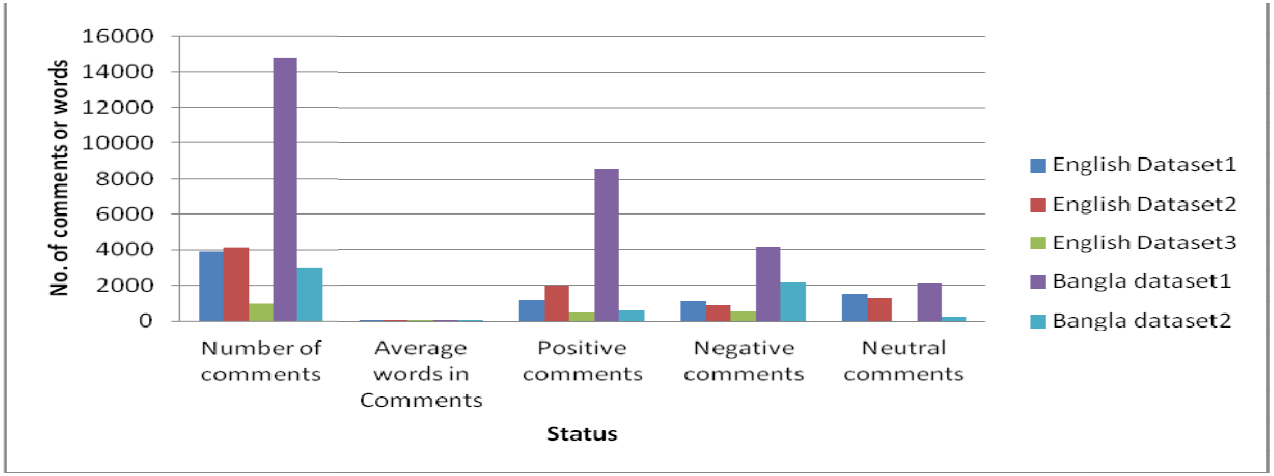

Figure 4.3: Statistics of Datasets

4.3 Summary of Data Analysis and Interpretation

Data analysis of two objectives such as knowledge base, polarity lexicon creation, and dataset creation are very important. This chapter has emphasized those two objectives and performed a deep analysis of data to find the appropriate data for SA. The process started with analyzing data for the knowledge base, which includes determining domain and types of data from the research question. This step helps complete the collecting data for the knowledge base and polarity lexicon. In this step, data were collected from both internal, external, and legacy sources, whichever felt relevant for Knowledge Base and Polarity Lexicon creation. Two forms of data were collected, one that needs conversion (data acquisition) and born-digital data. Once the data were successfully collected, the data were then preprocessed to bring the data in good shape. The research then selected the Quality data, which is the data having relation with the domain of study (student feedback, cricket, and movie review). The quality data are compiled with the dependency rules to derive the concepts for both knowledge base and polarity lexicon.

This research then does a challenging annotation job, where concept polarity lexicons were only assigned polarity values. However, knowledge base demands assigning polarity labels and semantics of the concepts too with the concepts. Besides, Pleasantness, Attention, Sensitivity, Aptitude values are also added. It is recommended through data analysis that there are no available Bengali concept polarity lexicons; only word-level polarity lexicons are available. Therefore, the Bengali concept polarity lexicon could be considered the rich resource for CLSA with 72433 concepts, where 36309 concepts are positive, and 27085 concepts are negative. Moreover, after carefully consolidating concepts from different sources, the Bengali knowledge base BanglaSenticNet is created with 30000 Bengali concepts and 150000

semantics of those concepts where, for example, 15126 concepts are positive and 14875 concepts are negative.

The dataset creation process starts with identifying data and domain according to the research question. The targetted domain was student feedback, and the data are qualitative. This research collected structured and unstructured data from predefined Facebook pages and groups using 'Graph API' and then organized the data using different steps such as consolidating, separating, and converting. The data are then separated (English, Bengali) based on language similarity in the next step. This research again converted the data to a single lingual dataset. It is recommended to work with the different lingual datasets separately (English and Bengali) and at once for easy processing.

Data pre-processing starts at this stage, where different techniques (unnecessary punctuations, stop words, letters, irrelevant words, symbols, irrelevant comments and status, extra spaces, and numbers removal) were applied to the organized data. This research then dealt with missing values, data inconsistencies, and relevant noise in the collected data. Once data cleaning is done, the research then starts transforming data to reduce data redundancy and different forms of anomalies. This research reduced unnecessary variables (except comments and polarity variables) and cases (comments and status) to bring the dataset into the desired format in its next step. Finally, data annotation is accomplished that labeled data as 'positive', 'negative,' and 'neutral.' The accuracy is tested by both human experts manually and by machine learning algorithms.

The statistics of the dataset show that the created datasets have a higher number of comments. These types of datasets are chosen to understand the effect of data sample size in analysis. Thus reviewed the sample size till a sufficient sample

size is achieved. The statistics also reveal that there is a less number of negative comments. This difference is dealt with using the same number of positive, negative, and neutral comments in the training.