CHAPTER 3

METHODOLOGY

The purpose of this chapter is to introduce the ways this research has been conducted, including the creation of the resources (knowledge base, concept polarity lexicon, and datasets) and evaluation of resources, algorithms (proposed MCSAlgo), and techniques (feature or concept extraction and preprocessing). Therefore, this chapter has presented methods of current research in its first section. The following section has described the ways of creating a Bengali knowledge base. The third section of this chapter has emphasized the creation process of the Bengali concept polarity lexicon. This chapter then presented the details of a proposed algorithm of multilingual sentiment analysis at the concept level. This chapter also presented a method of creating annotated datasets. Besides, this chapter has illustrated the selecting, testing, and evaluation process of feature or concept extraction techniques and their combinations. In addition, the searching process of preprocessing techniques and their combinations are presented in the next section. Finally, this chapter has presented the evaluation matrix and tools used in this research.

3.1 Methodology

This section describes the workflow of this thesis. The workflow is shown in Figure 3.1. As per the objectives, this workflow consists of two parts 1) resources such as knowledge base, lexicon, and datasets creation (for RO1 and RO3), part 2) resources, techniques, and algorithms evaluation part. In the first part, the resources are created.

Here, one Bengali knowledge base such as BanglaSenticNet (details in Section 3.2), one Bengali polarity lexicon (details in Section 3.3), one English, and one Bengali students feedback datasets (details in Section 3.5) are created.

In a later section, resources, techniques, and algorithms are evaluated. The evaluated resources are knowledge base (BanglaSenticNet, SenticNet 5), concepts polarity lexicons (Bengali and English), and datasets (created datasets (students feedback datasets (Bengali and English), one English (IMDB[12]), and one Bengali (cricket dataset[13]) baseline datasets (for RO1 and RO3). The baseline resources such as SenticNet 5, English polarity lexicon, IMDB, and cricket datasets were used for validation purposes. The evaluated algorithms are NB, SVM, LSTM, and proposed (MCSAlgo) (for RO2). Besides, evaluated techniques are feature and concept extraction techniques and their combinations (details in Section 3.6; for RO4), preprocessing techniques, and combinations (details in Section 3.7; for RO5).

The evaluation steps start with testing knowledge bases. It is tested on all the datasets, preprocessing techniques, and algorithms used in this thesis. In the beginning, the data were grabbed from each created dataset. The data of each dataset were sent to the preprocessor as sentences. After successful preprocessing, the data were sent to concept extractors. They were extracting the concepts using the rule defined in Section 3.2.1. The extracted concepts were used to create bag-of-concepts (BOC) and matched with the knowledge bases (BanglaSenticNet for Bengali concepts and SenticNet 5 for English concepts). If the matched found, the corresponding polarity was extracted from the knowledge bases and determine the accuracy using NB, SVM, LSTM, and MCSAlgo.

[12] https://www.imdb.com/interfaces/
[13] https://www.mdpi.com/2306-5729/3/2/15/htm

This process continues until the sentence ends. The polarities of individual sentences are aggregated to get the polarity of each dataset. The same process is applied in evaluating concept polarity lexicons. Moreover, the evaluation of created or baseline datasets is not separately discussed in this section, as every experiment involves datasets. Therefore, those experiment values are the evaluation status of the datasets.

Finding optimal feature or concept extraction techniques is achieved by extracting them individually and in combination. The individual features or concepts and their combination are evaluated using the left and right dotted rectangle processes. The evaluation was conducted on the created and baseline datasets. The evaluation results from both sides were merged to get the combination results. This research used BOC and BOF with different machine learning algorithms such as NB, SVM, and LSTM for sentiment classification. The algorithms were trained using extracted English and Bengali BOF in a feature-based approach. The polarity detection for each sentence matched a BOF with test data and continues until the last sentence. In the concept-based approach, two knowledge bases have been adopted, one contains English concepts (i.e., SenticNet 5), and the other contains Bengali concepts (i.e., BanglaSenticNet for sentiment classification. The algorithms were trained using both the knowledge bases. The extracted BOC from the previous section was then compared to find the polarity with the above algorithms. The polarity values found with different methods are then aggregated to find the sentence's absolute and comparative polarity and continue until the last sentence.

It is mentionable that, to meet the objective of finding optimal preprocessing techniques or their combinations has been achieved first selecting preprocessing techniques and determining the combinations. The preprocessing was applied on both the

created and baseline datasets. After applying each type of preprocessing technique and their combinations on the datasets, the preprocessed data were analyzed for sentiments using two distinct ways (such as feature and concept-based approaches) as indicated in the dotted rectangle. The left-hand side does the analysis again in two separate ways. In its first way, two-third of the pre-processed labeled data was used as training data, and one-third was used as test data. The three most popular machine-learning algorithms, such as NB, SVM, and LSTM, were trained and tested to find the polarity.

Secondly, features were extracted from each sentence of the pre-processed data and constitute bag-of-features (BOF) of Bengali and English sentences. This time, the same algorithms were trained using standard polarity lexicons (such as opinion-lexicon-English, and NRC-Emotion-Lexicon-v0.92-Bangla) and extracted BOF (details in Section 3.6). The polarity detection for each sentence matched a BOF and lexicons with test data and continues until the last sentence. The aggregated polarity value of the latter method is then compared with the previous one to find the final polarity. On the other hand, the right-hand side analyzes the sentiments using the concept-based approach discussed in the previous paragraph.

As per general practice, the best classifier was determined by using the following indexes such as accuracy (is the percent of the prediction were correct), precision (is the percent of the positive predictions were accurate), recall (is the percent of positive cases that were caught). F-score (is the weighted average of the recall and the precision).

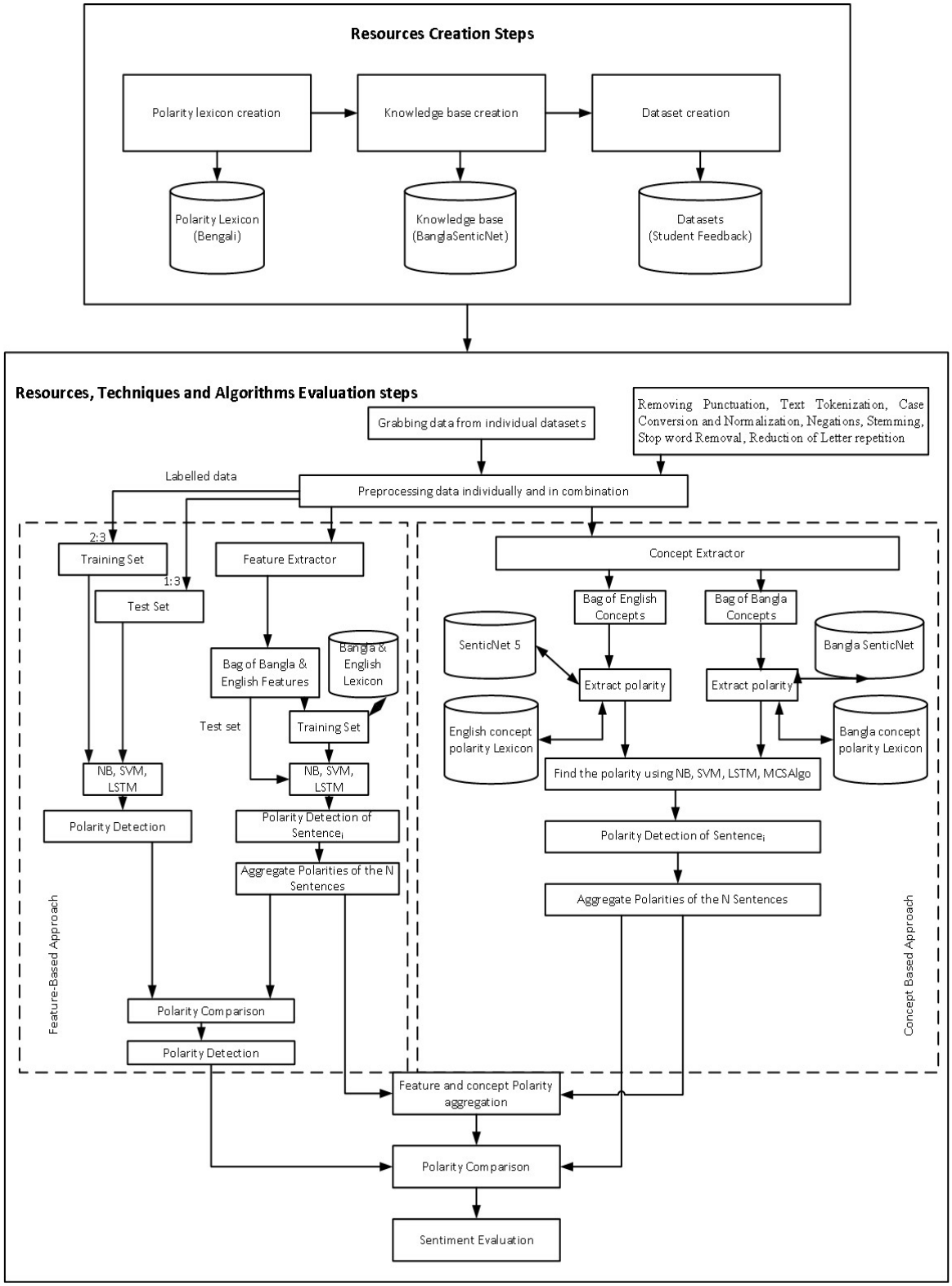

Figure 3.1: Flow Chart of this Research

3.2 Knowledge Base Creation

The main objective of this thesis is to create a Bengali knowledge base. This section focus on the creation process of a knowledge base. Knowledge Base is the most important source for finding the implicit meaning of the words and sentences. As per literature, a famous knowledge base, namely 'SenticNet' version 1 to 5, is widely used for dealing with English sentences. SenticNet is all about concept-level sentiment analysis, which is, performing work such as sentence polarity detection and sentiment or emotion recognition by forcing semantics and sentics instead of exclusively depending on the co-occurrence of the words. SenticNet consists of semantics (concepts), sentics, and polarity related to 100,000 natural languages (NL) concepts. Each input concept is related to five semantically related concepts. Sentics are emotion categories expressed in affective dimensions such as Pleasantness, Attention, Sensitivity, and Aptitude. Polarity value is considered from -1 (extreme negativity) to +1 (extreme positivity). SenticNet is also known as sentic computing (integrating machine learning, psychology, commonsense reasoning, linguistics), focusing on natural language concepts and sentence structure rather than only focusing on statistical approaches (Poria et al., 2018). Literature shows, this resource is also converted to 40 other languages except for Bengali.

This study tried to fill the gap of the inadequacy of the Bengali Knowledgebase by creating a knowledge base parallel to that of SenticNet5 but using a different method to deal with Bengali sentences and named it as 'BanglaSenticNet.'. The new method of Bengali knowledge base creation is shown in Figure 3.2 and the following algorithm. The

sample of a translated knowledge base is presented in APPENDIX 3. The process starts with the data collection for the KB building. The KB is concept-based and general; therefore, the data was collected (web contents and Reviews) from different relevant dictionaries, reports, newspapers, internet sources, social media, and generic knowledge sources such as SenticNet 5[5]. In order to meet the requirement, the concepts and their semantics in SenticNet 5 are translated to Bengali concept by using Google English to Bangla Translator or other English to Bangla Dictionaries. The data from all the sources are fed to the PoS tagger. The PoS tagger tags the words of the sentences. This research has used two PoS taggers, such as Stanford CoreNLP[14] and Stanford Log-linear PoS Tagger[15] , and do some manual tagging due to the inappropriateness found in some of the tagging. The words with PoS tagged are then sent to the concept extractor. The concept extractor used different dependency rules mentioned in the concept extraction section to extract the relevant concepts. This study has used two concept extractors such as Aylien[16] and Cttsai[17] , with some manual extracting stepped to extract the relevant concepts.

Each concept is then sending to the score-generating component of the process. This component is used to create the score of each concept using different existing dictionaries such as NRC emotion lexicon[20], Bangla Dictionary master[21], lexical DB Bangla master[22], and two estimation methods such as Probability scoring and Information-Theoretic scoring. The details of these two methods are described later in

[14]https://stanfordnlp.github.io/CoreNLP/
[15]https://nlp.stanford.edu/software/tagger.shtml
[16]https://aylien.com/text-api/concept-extraction/
[17]https://github.com/cttsai/concept-extractor

this section. The score calculator aggregates the score from all these sources and stores it in a sentiment matrix consisting of concepts, the semantics of the concepts, and their corresponding score. In the next stage, the concepts and semantics of the concepts with similar meanings are clustered using the K-means clustering algorithm. The maximum average score in the cluster is selected for each concept and assign that score to the concept. The process ends with assigning emotions to a concept from SenticNet 5 by matching the concept itself or the semantics of the concept.

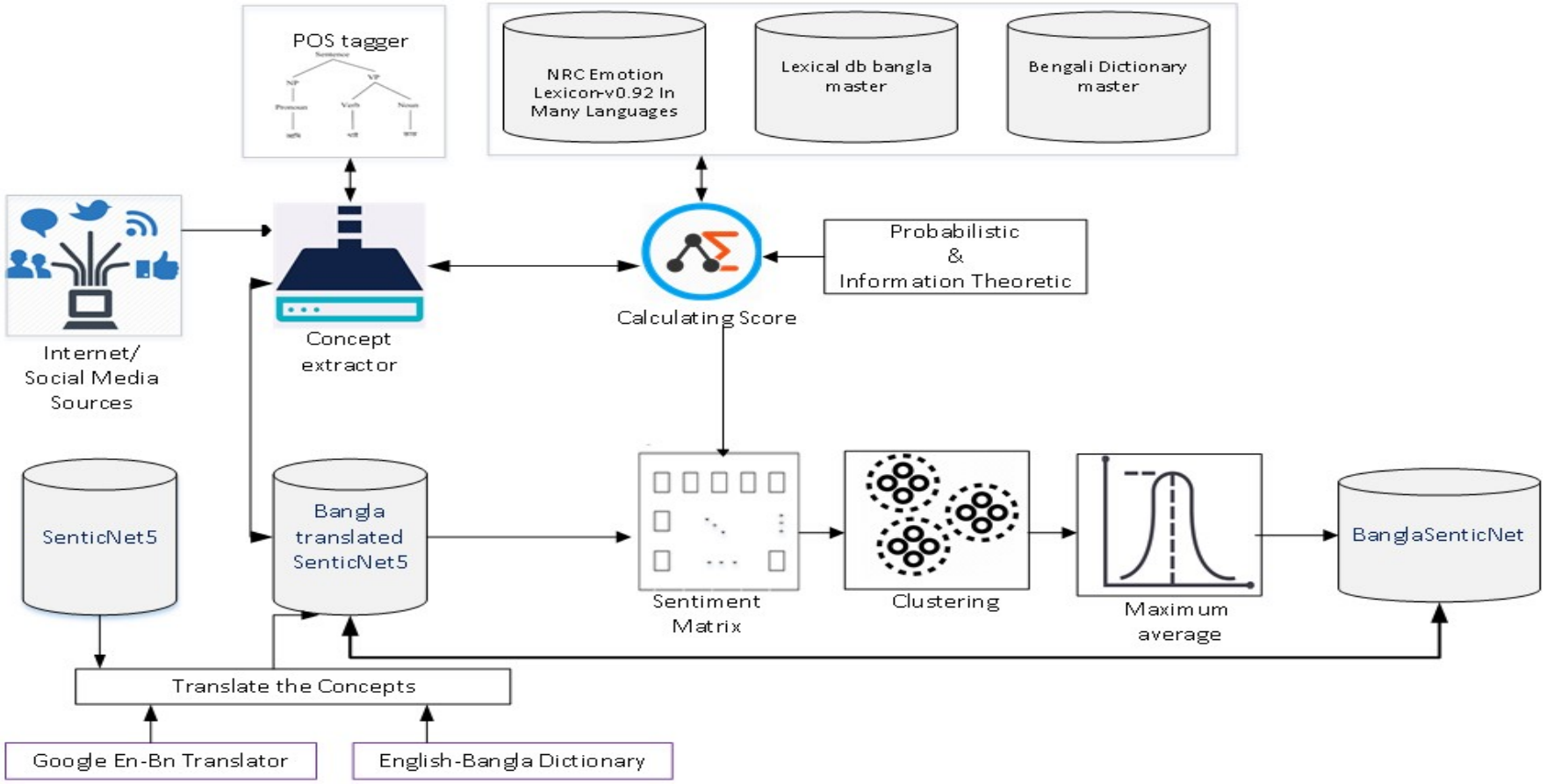

Figure 3.2: Knowledge Base Creation Process

Algorithm: Construction of BanglaSenticNet

Input: Internet and social media sources, SenticNet 5
Output: BanglaSenticNet
Notation: *C*→ Concepts in SenticNet, *SC*→ Semantics of Concepts in SenticNet, *S*→Review Sentences

1: **Begin:**
2: **for** each *C, SC,* and *S* **do**
3: Tag the PoS in each concept of the sentences using PoS tagger
4: Extract the concepts as per the defined concept extraction rules using concept extractor
5: Calculate the Probability score $Sc_{\text{prob}}(\text{c})$ & Information Theoretic score $Sc_{it}(c)$ for each concept using values from different existing dictionaries in the following equations and find the average score $Sc(c)$.

$$Sc_{\text{prob}}(\text{c}) = p(pve|c) - p(nve|c)$$

$$Sc_{it}(c) = \big(pve(c) - nve(c)\big) \times \left(\log \frac{N}{df_c}\right)$$

6: $$Sc(c) = \frac{Sc_{prob}(c) + Sc_{it}(c)}{2}$$

7: $SSc(c)$ = Extract the corresponding score of the concepts from the SenticNet 5
8: Form the sentiment matrix with concepts and its corresponding scores received from $Sc(c)$ and $SSc(c)$
9: Cluster the concepts and semantics of the concepts along with their scores using K-means clustering algorithms by

$$\text{C}_\text{i} = \left(\frac{1}{\text{d}_\text{i}}\right) \sum_{\text{j}=1}^{\text{d}_\text{i}} \text{x}_\text{i}$$

$$\text{K} = \sum_{\text{k}=1}^{\text{j}} \sum_{\text{i}=1}^{\text{m}} \left\| \text{x}_\text{i}^{\text{k}} - \text{c}_\text{k} \right\|^2$$

10: Final_score= Select the maximum average value of each concept obtained through $Sc(c)$ and $SSc(c)$ values for concepts and semantics of concepts in each cluster
11: Assign the final_score to the concept of BanglaSenticNet
Assign the semantics and emotions from SenticNet 5 to BanglaSenticNet for matched concepts
12: **end for**
13: **end**

3.2.1 Concept Extraction

In sentiment analysis, feature extraction or concept extraction techniques are considered exceptionally important. The textual data classification and sentiment detection solely depend on the successful feature and concept extraction techniques. As our primary concern is the knowledge base, and concepts are the keys to the knowledge base, this section has emphasized concept extraction. Concept extraction is the process of breaking the texts (Sentences) into concepts (or Clauses).

The method starts with the breaking of text such as "I was going to the university" and "আমি বিশ্ববিদ্যালয়ে যাচ্ছিলাম" into concepts or clauses as shown in figure 3.3 figure 3.4 respectively. As the grammatical structure of Bengali and English sentences is not the same, English correspondence of Bengali terms is used in Figure 3.4 and throughout the description to understand Bengali text better.

For concept extraction, the following formations of concepts were used in this study (Poria et al., 2014; Agarwal et al., 2015; Naadan et al., 2018):

1. **Subject Noun Rule** – Given the sentence, "The lecture was excellent," if the word 'Lecture' is a subject noun and if it has a relation with the word 'Excellent', then the concept (Excellent-lecture) was extracted.

   Bangla: Given the sentence, "বক্তৃতা চমৎকার ছিল," if a word 'বক্তৃতা' is a subject noun and if it has a relation with the word 'চমৎকার' then the concept (চমৎকার-বক্তৃতা) was extracted.

2. **Joint Subject Noun and Adjective Complement Rule** - Given the sentence, "The university provides many services, and services are good", if a word 'university' is a subject noun and if it has a relation with a word 'services', whereas 'services' has a relationship with adjective complement 'good' then the concept (good- university) was extracted.

   Bangla: Given the sentence, "বিশ্ববিদ্যালয় অনেক সেবা প্রদান করে, এবং সেবা ভাল", if a word 'বিশ্ববিদ্যালয়' is a subject noun and if it has a relation with the word 'সেবা', whereas 'সেবা' has a relation with adjective complement 'ভাল' then the concept (ভাল-বিশ্ববিদ্যালয়) was extracted.

3. **Adjective and Clausal Complements Rules** – The concept (reputation, good) was extracted from the sentence like 'The university reputation is good',.

   Bangla: From the sentence like "বিশ্ববিদ্যালয় খ্যাতি ভাল", the concept (ভাল,খ্যাতি) was extracted.

4. **Negation** – negation plays an essential role in text concept extraction as it is used as a complement. So, the sentence like 'I did not like the university' the concept (not, like) was extracted.

   Bangla: negation plays an essential role in text concept extraction as it is used as a complement. So, from the sentence like আমি বিশ্ববিদ্যালয় পছন্দ করিনি' the concept (পছন্দ , না) was extracted.

5. **Single Word Concepts** – the study has extracted the following words from the text, such as POS (verb, adjective, noun, adverb) and single word concepts that lie

in the multi-word concepts because of redundant information. For example, the concept ‘Hall,’ which already exists in the concept ‘Lecture Hall’, was extracted.

Bangla: Concept ‘কক্ষ’, which already exists in the concept ‘বক্তৃতা কক্ষ’ was extracted.

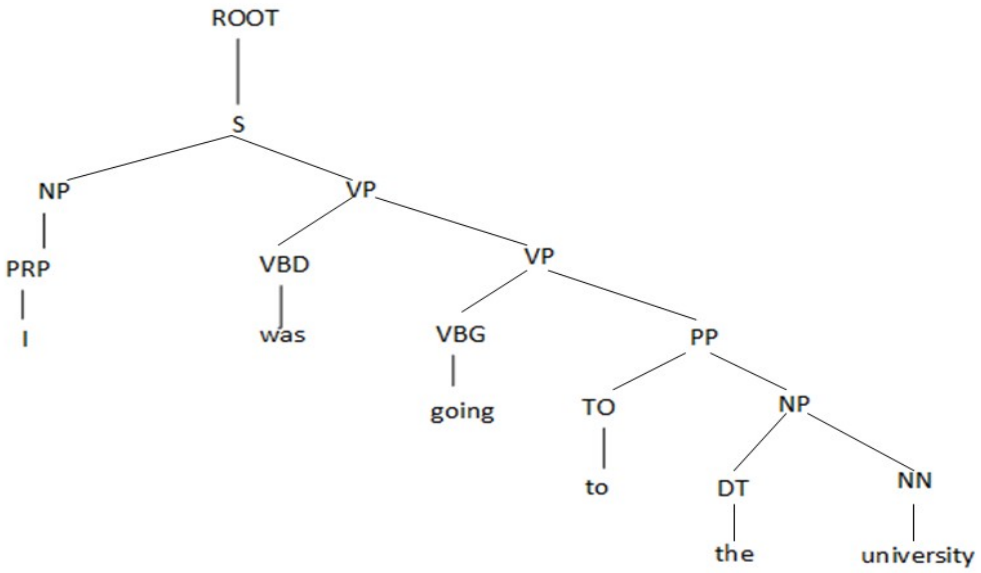

Figure 3.3: Parse Tree of the Sentence “I was going to the university”

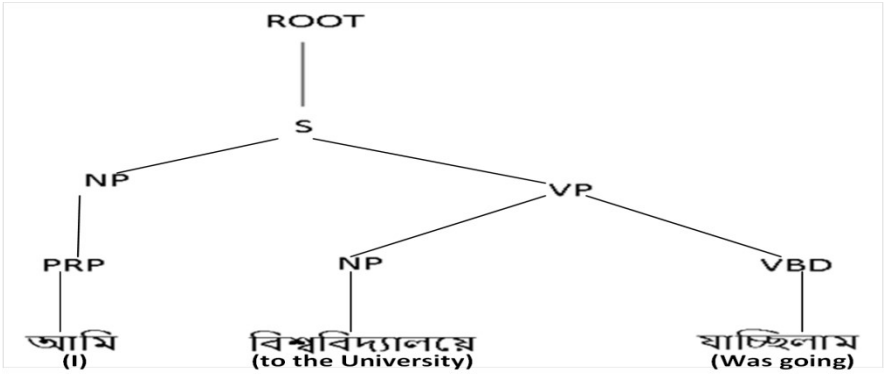

Figure 3.4: Parse Tree of Sentence “আমি বিশ্ববিদ্যালয়ে যাচ্ছিলাম”

3.2.2 Estimating Concepts Score

The sentiment score of each concept is estimated by averaging the probability score Scprob(c) and the score of the information-theoretic approach Scit (c). The score $SSc(c)$ is obtained from SenticNet 5 for fine-tuning results. Finally, determine the score using the maximum averaging method on the above two results.

The probability score $Sc_{\text{prob}}(\text{c})$ of a concept c is obtained by computing the difference between p (pve|c), i.e., the probability of concept c being positive, and p(nve|c), i.e., the probability of concept c being negative. The computed value is also known as posterior probabilities and mathematically expressed as follows:

$$\text{Sc}_{\text{prob}}(\text{c}) = \text{p(pve|c)} - \text{p(nve|c)} \tag{3.1}$$

where:

$$p(pve|c) = {p(pve) \times p(c|p)}\Big/{p(c)} \tag{3.2}$$

$$p(nve|c) = {p(nve) \times p(c|nve)}\Big/{p(c)} \tag{3.3}$$

and

$$\text{p(c|pve)} = \frac{{\gamma n_{c5^{\#}} + n_{c4^{\#}}}\Big/{\sum_{c'} (\gamma n_{c'5^{\#}} + n_{c'4^{\#}})} + 1}{D_{dic} + 1} \tag{3.4}$$

$$p(c|nve) = \frac{{\gamma n_{c1^{\#}} + n_{c2^{\#}}}\Big/{\sum_{c\prime} (\gamma n_{c\prime 1^{\#}} + n_{c\prime 2^{\#}})} + 1}{D_{dic} + 1} \tag{3.5}$$

p(pve) is the percentage of concepts that belong to the positive class; $p(nve)$ is the percentage of concepts belong to the negative class. Moreover, $p(c)$ is the total frequency

of concept *c*.*p*(c|*pve*) denotes the probability of detecting a concept c, given a positive class; and *p*(c|*nve*) denotes the probability of detecting a concept c given a negative class. $\gamma n_{c5^{\#}} + n_{c4^{\#}}$is the amount of *c* occurs in the positive class for review or comment *s* in corpus *S*; $\gamma n_{c1^{\#}} + n_{c2^{\#}}$ is the amount of *c* that occurs in the negative class for review or comment *s* in corpus *S*.$\sum_{c'}(\gamma n_{c'5^{\#}} + n_{c'4^{\#}})$ is the number of appearances of each concept in a positive class; and $\sum_{c'}(\gamma n_{c'1^{\#}} + n_{c'2^{\#}})$ is the number of appearances of each concept in a negative class. The size of the dictionary is denoted byD_{dic}. γ is described as the weight factor of the concept (1= entirely positive, -1= entirely negative, 0= neutral). *c* = $5^{\#}$ is the number of times c appears in 5-numbered reviews (Labille et al., 2016).

The information-theoretic (IT) approach is proposed by Burnham and Anderson in 1998 and is a technique that uses the Akaike information criterion. IT score sc_{it} (c) is calculated based on Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF helps identify the importance of the concept in a document (Crnic, 2011). IT score is also calculated in the same way as probability score, i.e., the distinction between the frequency of positive concept's score $pve(c)$ and negative concept's score $nve(c)$ of the document and its inverse document frequency, and mathematically defines as:

$$Sc_{it}(c) = \big(pve(c) - nve(c)\big) \times (\log \frac{N}{df_c}) \tag{3.6}$$

where:

$$\text{pve(c)}=\gamma \text{brtf}\,(\text{c}_{5^{\#}}) + \text{brtf}_{\text{c}}(\text{c}_{4^{\#}}) \tag{3.7}$$

$$nve(c) = \gamma brtf(\text{c}_{1^{\#}}) + brtf_c(c_{2^{\#}}) \tag{3.8}$$

brtf is used to denote balanced relative term frequency of concept c and is proposed by Labille et al.(2016); *Npos* and *Nneg* denote the total number of positive and

negative reviews, respectively; *N* stands for a total number of comments or reviews. The terms c and γ are used for the same purpose as in the probability approach.

In order to get the benefits from both the approaches, the literature has tested by averaging the score of both the approaches and found a more accurate score than simply using them separately (Labille et al., 2016). Therefore, we have averaged the scores $Score_{prob}(c)$ and $Score_{it}(c)$ by using the following generic equation:

$$Sc(c) = \frac{Sc_{prob}(c) + Sc_{it}(c)}{2} \tag{3.9}$$

3.2.3 Sentiment Matrix Formation and Clustering

The sentiment matrix is formed with the concepts, semantics of the concepts, and their corresponding polarity score. The information of concepts, semantics of the concepts, and their polarity score are taken from both the SenticNet 5 and upper block of Figure 3.2, (i.e.1) the concepts that are extracted from social media or other internet sources and 2) polarity scores that are determined with probability theory and information-theoretic approach). The matrix value is then sent to the clustering algorithm.

For clustering, we have used the K-means clustering algorithm. K-means is an unsupervised learning algorithm proposed by J. MacQueen in 1967 (Likas et al., 2003), where data points $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ such as concepts and semantics of the concepts with their corresponding polarity score are divided into $C = \{c_1, c_2, \ldots\ldots, c_d\}$ number of clusters considering the nearest mean values of the cluster centers. The reason for using this algorithm is the scalability in terms of data sizes, ensures intermingling, easily

adjustable to new examples, simplicity of usage, fast execution efficiency, and high performance.

The algorithm runs n number of times and ends once all the concepts and semantics of the concepts with polarity scores are classified successfully into appropriate clusters. In each loop, it calculated a new cluster center and reorganized the values to the nearest clusters using the following equation:

$$C_i = \left(\frac{1}{d_i}\right) \sum_{j=1}^{d_i} x_i \tag{3.10}$$

Where d_i stands for the number of data points at the i^{th} cluster. The objective function was minimized using the following equation:

$$K = \sum_{k=1}^{j} \sum_{i=1}^{m} \left\| x_i^k - c_k \right\|^2 \tag{3.11}$$

Where m denotes the number of data points; j is the number of clusters; $\left\| x_i^k - c_k \right\|^2$ is used as a distance measure between x_i^k (data point) and c_k (cluster center).

3.2.4 Final Concept Score Determination

The final score in each cluster is obtained by selecting the maximum average value of each concept through $\boldsymbol{Sc(c)}$ and $\boldsymbol{SSc(c)}$. The score is then assigned to the corresponding concept of BanglaSenticNet. Finally, the semantics and emotions of that concept are derived from SenticNet 5 and assigned to the matched concepts of BanglaSenticNet.

3.2.5 Knowledge Base Evaluation

The evaluation of knowledge bases (SenticNet 5, BanglaSenticNet) is done for datasets, algorithms (NB, SVM, LSTM, and MCSAlgo), concept extraction techniques, and their combinations (stated in Section 3.2.1), preprocessing techniques and their combinations (stated in Section 3.7). For each type of resource or technique, the evaluation of knowledge bases is done separately and finally averaged to find the outcome. The outcome is then compared with the outcome of state-of-art research. The experiment results and their related evaluation are shown in Section 5.1.

3.3 Lexicon Creation

An objective of this thesis is to create the Bengali concept polarity lexicon. This section focus on the creation process of the polarity lexicon. The Lexicon is considered the essential resource from among the resources used for emotion detection. There are many English lexicons available (as discussed in Section 2.5); however, only a few Bengali lexica exist. Therefore, to improve the performance of the SA, creating an enriched Bengali lexicon is essential. The study used the following algorithm for Bengali lexicon creation. The creation process starts with gathering existing Bengali and English benchmark lexicons. The process then matched the Bengali and English words from different lexicons with the benchmark Bengali lexicons. If the match is found in the lexicon, then the benchmark lexicon remains the same; otherwise, words from other lexicons are added (translated if needed) to the benchmark lexicons. For new words or concepts, the polarity is found as per the process followed in Section 3.2. Final polarity is determined by averaging values from P_{bl} and P_{ol} (if these two values differ); otherwise,

P_{bl} value is the final polarity. A sample of the polarity lexicon is presented in APPENDIX 2.

Algorithm: Construction of Bangla sentiment polarity lexicon

Input: Most of the existing English and Bangla polarity lexicon
Output: Bangla sentiment polarity lexicon
Notation: W_{bl}→ words in benchmark Bengali lexicon, W_b→ Bangla words, W_e→English words, P_{bl}→ Polarity of final Bengali lexicon, P_{ol}→ Polarity of other lexicons

1: **Begin:**
2: **for** each word in the lexicon (W_b or W_e)**do**
3: **if** W_{bl}=W_b or W_e (same meaning) **do**
W_{bl} with polarity remains the same
4: **else if** $W_{bl} \neq W_e$**do**
Translate to Bengali
Calculate the polarity as per Section 3.2
add to W_{bl}
5: **else**
Add the Bengali word already not available in W_{bl} and found in other Bengali lexicon
Calculate the polarity as per Section 3.2
add to W_{bl}
6: **end if**

7: **for** each word of the final lexicon, **do**
8: **while***i*=1.....n **do**
9: **if** $P_{bi} \neq (P_{ol})_l$ **do**
Average them and Assign new polarity to W_{bl}
10: **else**
Polarity remain the same
11: **end if**
12: **end while**
13: **end for**
14: **end**

The evaluation of polarity lexicons (English, Bengali) is done for datasets, algorithms (NB, SVM, LSTM, and MCSAlgo), concept extraction techniques, and their combinations (stated in Section 3.2.1), preprocessing techniques and their combinations (stated in Section 3.7). For each type of resource or technique, the evaluation of polarity lexicons are separately and finally averaged to find the outcome. The outcome is then

compared with the outcome of state-of-art research. The experiment results and their related evaluation are shown in Section 5.1.

3.4 Proposed Algorithm for Multilingual Sentiment Analysis at Concept Level

This section describes the proposed algorithm for finding user sentiment at the concept level (using SenticNet5 and BanglaSenticNet) from the multilingual text to meet one of its objectives. The algorithm is named *"MCSAlgo"* and stands for "multilingual concept-level sentiment analysis algorithm." The pre-processed data are sent through MCSAlgo, and this algorithm extracts the concepts for each sentence by tokenizing them to n-gram and placing '_' between them. The algorithm then matches each concept to knowledge bases and lexicons such as SenticNet 5, BanglaSenticNet, and English and Bengali concept polarity lexicon (details in Section 3.2 and Section 3.3). The matched concepts return the intensity value from those knowledge bases and lexicons; otherwise, it returns null. Besides, the algorithm matches each concept to SenticNet 5 and BanglaSenticNet for finding the pleasantness value, attention value, sensitivity value, aptitude value, primary mood, secondary mood, polarity label, polarity value of each concept. If the concept does not match the concepts in the knowledge bases mentioned above, then the algorithm matches the concepts' semantics. The algorithms find the pleasantness value, attention value, sensitivity value, aptitude value, primary mood, secondary mood, polarity label, and polarity value for each match. The algorithm then compares the polarity value with and without semantics for each found concept and selects the highest polarity. This algorithm is written in such a way so that it could be

used in MLSA and discover different states of the sentiment; for instance, this thesis has worked with the polarity value only.

Algorithm: Proposed algorithm for multilingual concept-level sentiment analysis (MCSAlgo)

Input: Pre-processed Data
Output: Polarity (Positive, Negative, Neutral)
Notation: *SN*→ SenticNet, *C*→Concepts, *P(Si, I_i)*→Polarity of Sentence without semantic, *P(POV$_i$, I_i)*→ Polarity of Sentence with semantic, *BOC* →Bag of Concepts, *BSN*→ BanglaSenticNet, *EBCP*→ English and Bangla Concept Polarity lexicon, *S*→Sentence, *I*→Intensity, *SC*→ Semantics of Concepts, *PV*→ Pleasantness Value, *AV*→ Attention Value, *SV*→ Sensitivity Value, *APV*→ Aptitude Value, *PM*→ Primary Mood, *SM*→Secondary Mood, *PL*→Polarity Label, *POV*→ Polarity Value, *WOS*→ Concepts without semantics, *WS*→ Concepts with Semantics
Begin:

1: **for** each sentence $S_1, S_2, \ldots\ldots, S_n \in S$**do**
2: **while***i*=1.....n **do**
3: *BOC*← Extract concept for $S_i$by tokenizing to n-gram and place '_' between them
4: **for** each concept $C_1, C_2, \ldots\ldots, C_n \in BOC$**do**
5: **while** i=1.....n **do**
6: *M*←Match C_i to *EBCP* or *SN* or *BSN* and return 1 for 'yes' and 0 for 'no'
7: **if** *M*==1 **do**
8: I_i+← Extract the Intensity of the Concept$_i$
9: **else**
10: Return null;
11: **end if**
12: *M*←Match $C_i$to *SN* or *BSN* and return 1 for 'yes' and 0 for 'no'
13: **if** *M*==1 **do**
14: PV_i= Extract *PV* of concept C_i;
15: AV_i = Extract *AV* of concept C_i;
16: SV_i= Extract *SV* of concept C_i;
17: APV_i = Extract *APV* of concept C_i;
18: PM_i = Extract *PM* of concept C_i;
19: SM_i = Extract *SM* of concept C_i;
20: PL_i = Extract *PL* of concept C_i;
21: POV_i = Extract *POV* of concept C_i;
22: **else**
23: *M*←Match $C_i$to *SC* in *SN* or *BSN* and return 1 or 2 (match semantics of other concepts)
24: for 'yes' and 0 for 'no'
if *M*==1 or *M*==2 **do**
25: PV_i= Find the concept of semantics, Extract *PV* of concept C_i, and average them;
26: AV_i = Find the concept of semantics, Extract *AV* of concept C_i, and average them;
SV_i = Find the concept of semantics, Extract *SV* of concept C_i, and average them;
27: APV_i = Find the concept of the semantics, Extract *APV* of concept C_i, and average them;
PM_i = Find the concept of the semantics, Extract *PM* of concept C_i, and average them;
28: SM_i = Find the concept of semantics, Extract *SM* of concept C_i, and average them;
PL_i = Find the concept of semantics, Extract *PL* of concept C_i, and average them;
29: POV_i = Find the concept of semantics, Extract *POV* of concept C_i, and average them;

```
        else
30:       Return null;
        end if
31:     end if
      end while
32:  end for
     if P(Si,I_i)> P(POV_i,I_i) do
33:          WOS_i+←Concepts without semantics is better and corresponding value is
           extracted
34:     else
35:          WS_i+←Concepts with semantics is better and corresponding value is extracted
36:     end if
37:     The highest value among WOS and WS is selected as the final polarity
38:     end while
     end for
```

The evaluation of MCSAlgo is done concerning knowledge bases (SenticNet 5, BanglaSenticNet), polarity lexicons (English, Bengali), Datasets (English and Bengali, created and baseline), concept extraction techniques (stated in Section 3.2.1), preprocessing techniques and their combinations (stated in Section 3.7). For each type of resource (knowledge bases, polarity lexicons, and datasets) and techniques (concept extraction and preprocessing), the evaluation with MCSAlgo has been done separately and finally averaged to find the outcome. The outcome is then compared with state-of-art research and other standard algorithms (NB, SVM, and LSTM). The experiment results and their related evaluation are shown in Section 5.1.

3.5 Dataset Creation from Multilingual Data

One of the objectives of this research is to create datasets. In order to create a comprehensive multilingual dataset (which contains many variables and labels from different languages within the same dataset (Weesie, 2005)), there requires appropriate

principles and process, which in turn create similar challenges for the researcher in this field (Dang-Nguyen et al., 2017). So it becomes necessary to address those cases seriously to reach an appropriate solution. Figure 3.6 is such an attempt that elucidates a process for multilingual dataset creation. The process starts with collecting multilingual data (e.g., English, Bengali).

The collection of data is done following the method shown in Figure 3.5. The method starts with identifying domain (i.e., Student feedback) and sources (i.e., Facebook pages and groups) of data (i.e., comments and status), where a diverse set of data are available. The data were grabbed from those sources using application programming interfaces (API) and python crawler in the next step. Both structured and unstructured data were traced out by different tracing approaches, such as self-involved and topic-based, as the research aims to know the institution's overall sentiments. In the next step, the standard practice is the preprocessing of data. However, here (as per Figure 3.6), the data are separated into English and Bengali first and then sent to the pre-processor. The preprocessing techniques (discussed in Section 3.7) are applied to the data if applicable and required data.

After separation and pre-processing the comments and status, each comment and status is then annotated with the labels ‘positive’, ‘negative,’ and ‘neutral’ as applicable. The annotations are done using the comments of a human expert, existing datasets, and internet sources. The annotation or polarity assigning algorithm is given below in this section. Finally, the research got well-formed datasets. However, for easy data processing and due to the unavailability of tools, the Bengali and English data are separately

processed. Therefore, this requirement is fulfilled by separating one multilingual dataset and creating two English and one Bengali dataset.

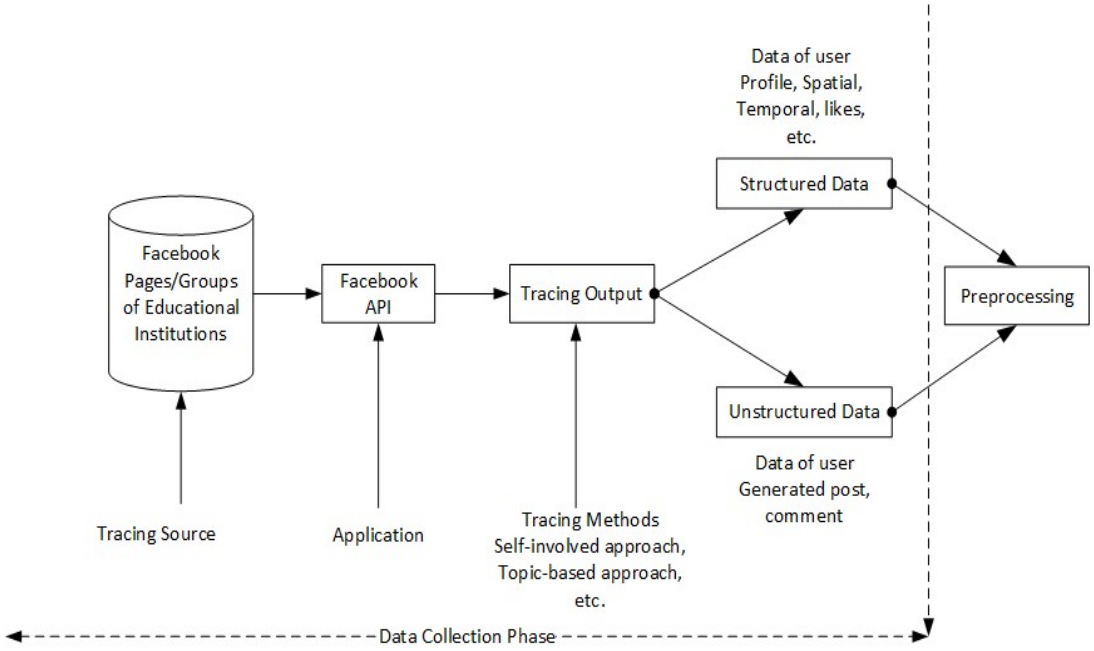

Figure 3.5: Data Collection Method

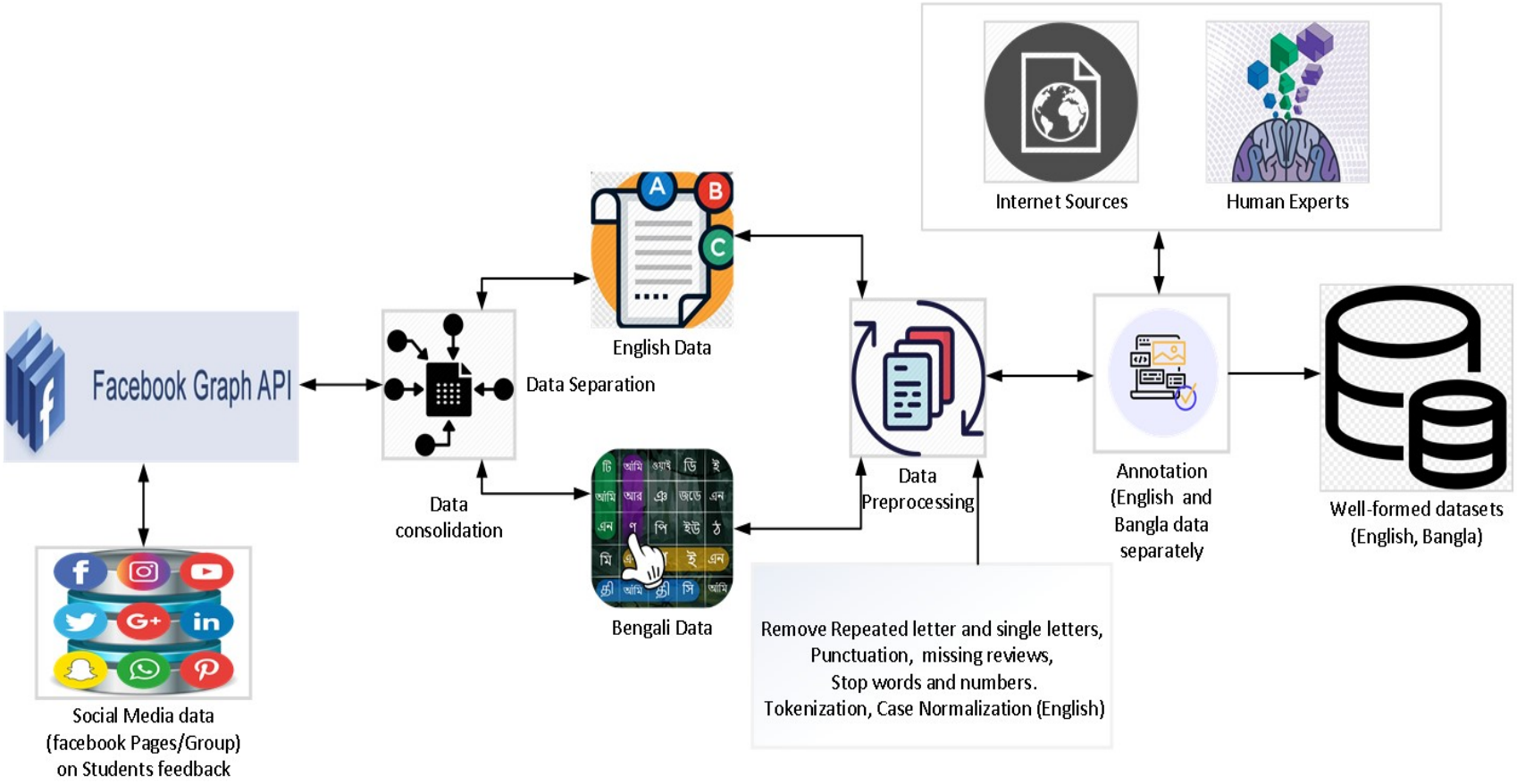

Figure 3.6: Annotated Multilingual Dataset Creation Process

The algorithm for assigning polarity to the comments is described in this section. The Facebook comments are filed separately after pre-processing in a CSV (comma-separated value) file. For each comment in the file, the problems of sentence boundary disambiguation are resolved by separating the comments by period, and the next word starts with a capital letter. Again, for each sentence, the study found the polarity by subtracting the term frequency of positive and negative words from the sentence using different internet sources and human experts. As per rules, if the polarity value (*P*) is found to between 0 and 1, then the sentence is designated as 'positive', and if the polarity value is less than 0, then polarity is 'negative'; otherwise, the polarity was termed as 'neutral'.

Algorithm: Assigning polarity to comments

Input: Facebook page after pre-processing comments
Output: Polarity (Positive, Negative, Neutral)
Notation:*S*→Sentence, *C*→Comment, *P*→Polarity

Begin:
for each Comment, **do**
  Resolve the Sentence boundary disambiguation problem by separating sentences by period, and the next word starts with a capital letter
  for each sentence $S_1, S_2, \ldots\ldots, S_n \in C$ **do**

$$p = \frac{1}{R}\sum_{r=1}^{R} p(l_r)$$

    Where *R* is the number of sources searched, and *p(lr)* is the polarity obtained from the sources *r*.
    Assign final polarity (Positive, Negative) to the words by calculating
    if *p*> 0 and p<=1 **do**
      Assign "Positive" polarity
    else if p< 0 **do**
      Assign "Negative" polarity
    else
      Assign "Neutral" polarity
    end if
  end for
end for
end

The evaluation of each dataset has been done concerning knowledge bases (SenticNet 5, BanglaSenticNet), polarity lexicons (English, Bengali), algorithms (NB, SVM, LSTM, and MCSAlgo), feature or concept extraction techniques and their combinations (stated in Section 3.6), preprocessing techniques and their combinations (stated in Section 3.7). For each type of resource or technique, the evaluation with datasets has been done separately and finally averaged to find the outcome. The outcome is then compared with the outcome of state-of-art research. The experiment results and their related evaluation are shown in Section 5.2.

3.6 Feature or Concept Extraction

One of the objectives of this research is finding the optimal features or concept extraction techniques for MLSA. The classification and sentiment detection from the textual data solely depends on the successful feature and concept extraction technique. The study uses the multilingual (English, Bengali) corpus, features, and concepts from both languages that needed to be extracted with equal importance.

Therefore, this research gives equal weight to feature and concept extraction techniques and their combinations for evaluating optimal one. This section only describes the feature extraction techniques that have been used in this study. However, the description of concept extraction rules is given in Section 3.2.1.

Generally, the Feature extraction process is the task of finding the appropriate characteristics of sentiment analysis of the pre-defined domain. The feature extraction phase consists of three steps (1) *Feature selection*- refers to selecting suitable features for

classification (2) *Feature weighting mechanism*- refers to assigning a weight to each feature for better extraction of solution (3) *Feature reduction mechanism*- refers to the reduction of selected features for improved classification. From different features, this research has extracted the following features to improve the performance of the analysis.

1. **Term Co-occurrence** - features such as unigram, bigram, trigram, and N-gram that lies together in the document.
2. **Term Frequency –** the number of occurrences of a particular term in a given review document.
3. **Part of Speech (POS) information** – POS has been used to remove uncertainty, i.e., to detect some useful sentiment indicators such as adverbs and adjectives.

The research has used the same feature or concept extraction techniques for both English and Bengali text. In requirement to extract the above text features or concepts (Section 3.2.1), the text is estimated based on semantic relatedness of words, which could learn from a large volume of unlabeled data (Liu, 2015). This research evaluates features or concepts in two different ways 1) sole and 2) combinations. Sole features or concepts are TF-IDF, unigram, bigram, trigram, POS, concepts (simple). The combinations of features and concepts are unigram+bigram, unigram+trigram, unigram+bigram+trigram, concepts+unigram, concepts+bigram, concepts+trigram.

The method of evaluating techniques such as features, concepts, and their combinations starts with extracting them. The extracted features, concepts, and combinations (stated in Section 3.6) were used to train and test classifiers such as NB, SVM, LSTM, and MCSAlgo. Every technique is evaluated separately and in combinations with other techniques on each dataset and algorithms. Finally, the

evaluations are compared with each other to find the optimal one. The experiments and their results are shown in Section 5.3. The following algorithm is proposed for finding optimal feature or concept extraction techniques and their combinations.

Algorithm: Finding the best pre-processing, feature, and concept extraction technique(s)

Input: Bangla English and Mix dataset (raw)
Output: Preprocessed data and overall polarity
Notation: D→Datasets, C→Bag of Concepts, U→ Unigram Dictionary, B→ Bigram Dictionary, P→Polarity, P_p→Positive Polarity, P_n→ Negative Polarity, RP→ Removing Punctuation, T→ Text Tokenization, CC→ Case Conversion, N→ Negations, SWD→ Stop word Deduction, S→ Stemming, RLR→ Reduction of Letter repetition
/*Finding optimal preprocessing techniques and its combinations*/

Begin:
for each dataset $D_1, D_2, \ldots\ldots, D_n \in D$ **do**
  while j=1…..n **do**
    while i=1.....n **do**
      Apply all (*RP,T, CC, N, SWD, S, RLR*) or a combination of two or more (e.g., *RP, T*) preprocessing techniques to D_j
      Determine Polarity (P_p, P_n) for each type combination using U to D_j
      if P_p-P_n> 0**do**
        $(P_p)_i$← Assign positive Polarity value
      else
        $(P_n)_i$← Assign negative Polarity value
      end if
        TP_j← Select the preprocessing type combination with the highest positive polarity
        TN_j←Select the preprocessing type combination with the highest negative polarity
        Compare among all TP_j and TN_jvalues
        BUP← Best preprocessing type combination using U
        BUV← value of best preprocessing type combination using U
    while end
  while end
end for
for each dataset $D_1, D_2, \ldots\ldots, D_n \in D$ **do**
  while j=1…..n **do**
    while i=1.....n **do**
      Apply all (*RP,T, CC, N, SWD, S, RLR*) or a combination of two or more (e.g., *RP, T*) preprocessing techniques to D_j
      Determine Polarity (P_p, P_n) for each type combination using B to D_j
      if P_p-P_n> 0 **do**
        $(P_p)_i$← Assign positive Polarity value
      else
        $(P_n)_i$← Assign negative Polarity value
      end if
      TP_j← Select the preprocessing type combination with the highest positive

polarity
TN_j←Select the preprocessing type combination with the highest negative polarity
Compare among all TP_j and TN_jvalues
BBP← Best preprocessing type combination using B
BBV← value of best preprocessing type combination using B
while end
while end
end for
for each dataset $D_1, D_2, \ldots \ldots, D_n \in D$**do**
while j=1…..n **do**
while i=1.....n **do**
Apply all (*RP,T, CC, N, SWD, S, RLR*) or a combination of two or more (e.g., *RP*, *T*) preprocessing techniques to D_j
Determine Polarity (P_p, P_n) for each type combination using C to D_j
if P_p-P_n> 0 **do**
$(P_p)_i$← Assign positive Polarity value
else
$(P_n)_i$← Assign negative Polarity value
end if
TP_j← Select the preprocessing type combination with the highest positive polarity
TN_j←Select the preprocessing type combination with the highest negative polarity
Compare among all TP_j and TN_jvalues
BCP← Best preprocessing type combination using C
BCV← value of best preprocessing type combination using C
while end
while end
end for
/* Finding optimal feature or concept extraction techniques and its combinations*/
if $BUV>BBP \&\& BUV>BCV$ **do**
BUP is the best preprocessing technique, and unigram is the best feature
else if $BBV>BUP \&\& BBV>BCV$ **do**
BBP is the best preprocessing technique, and Bigram is the best feature
else
BCP is the best preprocessing technique, and *C* is the best feature
end if
end

3.7 Preprocessing

Pre-processing is one of the critical phases in sentiment analysis because the views of the people are unstructured and contain many grammatical errors, which demands the task of data cleaning, punctuation symbols removal, etc. to be done to

minimize the noise from the dataset by removing such data that are semantically irrelevant (Polymerou et al., 2014). Besides, by selecting suitable pre-processing techniques, classification effectiveness and accuracy may be enhanced, and performance would be improved. One of the popular tools for unnecessary data remover is the NLTK3 library. Moreover, MLSA requires separate pre-processing of different lingual (e.g., English and Bengali) data. Therefore, as per the objectives, this research has conducted the following types of pre-processing techniques (individually and in combination) on the given text or sentences to improve the performance of multilingual sentiment classification.

1. **Text Tokenization:** Divided sentences into individual tokens or words.
2. **Case Conversion and Normalization:** Normalized the comments and status by converting in lowercase for easy searching from the dictionaries. This technique is applied to English text as bangle has no cases.
3. **Stemming:** Changed the words into their corresponding root or base to ease the morphological variety.
4. **Stop Word Deduction:** Removed the irrelevant words such as a, an, he, she, etc.
5. **Reduction of Letter Repetition**: Reduced the word and vocabulary size by minimizing the repetition of adjacent letters such as “Cuteeeee” to “Cutee”.
6. **Negations**: negated words such as *isn’t*, wasn’t, hasn’t, and haven’t replaced with word NEGATION to treat them equally.
7. **Removing Punctuation:** Removed those marks to ease the comparison of words from the vocabulary.

The algorithm presented in the previous section is proposed to find the best pre-processing, and Figure 3.7 shows the pre-processing topology. This topology illustrates the process of finding the best pre-processing technique or combination of techniques. Each pre-processing technique mentioned above was applied individually on the datasets (raw text). The polarity was then calculated after the application of each technique on the datasets. After using all the pre-processing techniques separately, the pre-processing technique combinations were then applied to the datasets and calculate the polarity. This process continued until having well-formed datasets. The combinations are named 'two preprocessing', 'three preprocessing' etc. Here, 'two preprocessing' means combining the results from two techniques: (stop word deduction + stemming), (stemming + negation), etc. This research achieves seven preprocessing technique combinations for English data and six for Bengali data. Each technique or technique combination is applied on both created and baseline datasets and measured the performance using NB, SVM, LSTM, and MCSAlgo. The comparative evaluation is done among different parameters of the confusion matrix to determine the optimal preprocessing technique. The experiments and their results are shown in Section 5.4.

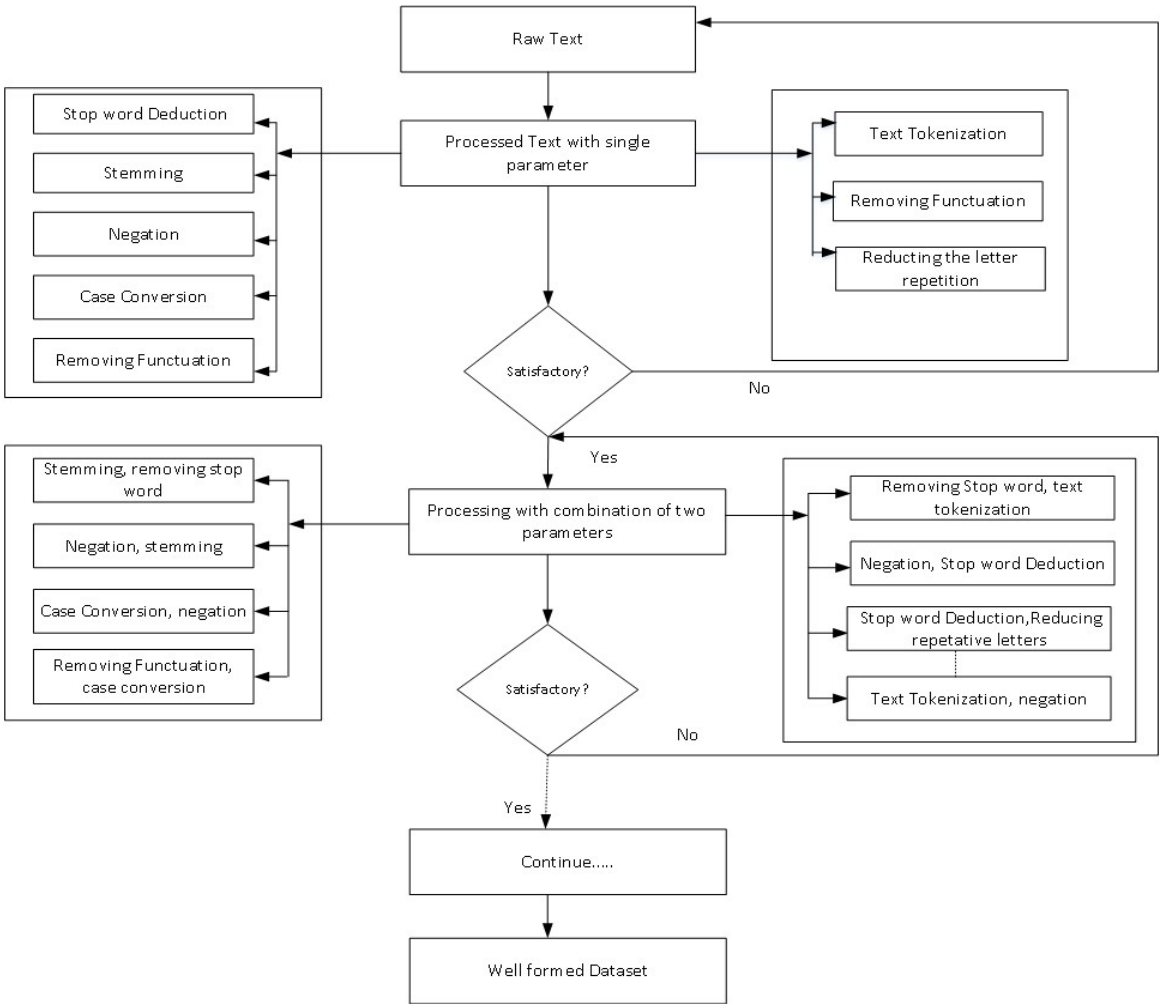

Figure 3.7: Pre-processing Topology

3.8 Performance Evaluation

This section describes different parameters of the sentiment classifier's performance evaluation, such as confusion matrix and its constituent parameters, for instance, accuracy, precision, recall, and F-score.

3.8.1 Model Evaluation Metrics

In this thesis, the performance of the applied classifier such as NB, SVM, LSTM, and MCSAlgo was evaluated using different measures of the confusion matrix such as accuracy, precision, recall, and F-score(with some exception, where only accuracy measure was used due to the massive requirements of calculation). Detailed confusion matrix (Davis et al., 2006; chai et al., 2014) are given below:

3.8.1.1 Confusion Matrix

The confusion matrix is a simple table that could exemplify the performance of the classifiers on any test set if the classifiers are already trained. Table 3.1 shows the sample of the confusion matrix. It consists of four sets of values such as actual NO, actual YES, predicted NO, predicted YES. For example, the test case with the same actual NO and predicted NO value is known as true negative (TN).

Table 3.1: Confusion Matrix

Scale	Predicted NO	Predicted YES
Actual NO	True Negative (TN)	False Positive (FP)
Actual YES	False Negative (FN)	True Positive (TP)

3.8.1.2 Accuracy

It is the measure of "how often the classification models are correct in predicting?" It is the ratio between the correctly classified classes (YES and NO) and the total number of classes. It is mathematically calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.12}$$

3.8.1.3 Recall or Sensitivity

The measure of "how often the classifier predicted yes for actually yes cases." Recall or sensitivity is the ratio between true positive (TP) and actual yes and is calculated mathematically as:

$$Recall = \frac{TP}{TP+FN} \tag{3.13}$$

3.8.1.4 Precision

It is the ratio of true positive (TP) and predicted yes. It finds, "if the classifier predicts yes, how often the classifier is correct?" The precision of the classifiers is calculated as follows:

$$Precision = \frac{TP}{TP+FP} \tag{3.14}$$

3.8.1.5 F-score

It is one of the widely used measures of classification accuracy. It is the weighted average of recall and precision. It is sometimes called a dice similarity coefficient and is calculated as:

$$F - Score = \frac{2 \times Recall \times Precision}{Recall + Precisio} \tag{3.15}$$

The performance of both English and Bengali datasets, lexicons, knowledge bases, proposed and state-of-art machine learning and deep learning algorithms, different

preprocessing techniques and their combinations, different feature and concept extraction techniques and their combinations, are evaluated in this thesis to find the optimal resources, techniques, and algorithms. The total process is shown in Figure 3.8; an abstract view of Figure 3.1. The evaluation is done using a standard method known as a confusion matrix. We evaluated using four parameters of confusion matrices (accuracy, precision, recall, and F-measure).

The proposed algorithm (MCSAlgo) is intended for multilingual concept-level sentiment analysis; therefore, it is used to test only the knowledge bases such as SenticNet 5 and BanglaSenticNet and polarity lexicons (both English and Bengali). Each preprocessing technique and its combinations, feature, and concept extraction techniques and their combinations, both English and Bengali datasets, lexicons, and knowledge bases are tested using different machine learning (NB, SVM, and LSTM) algorithms separately and evaluated across each other. The evaluation results are presented in different sections of chapter 5.

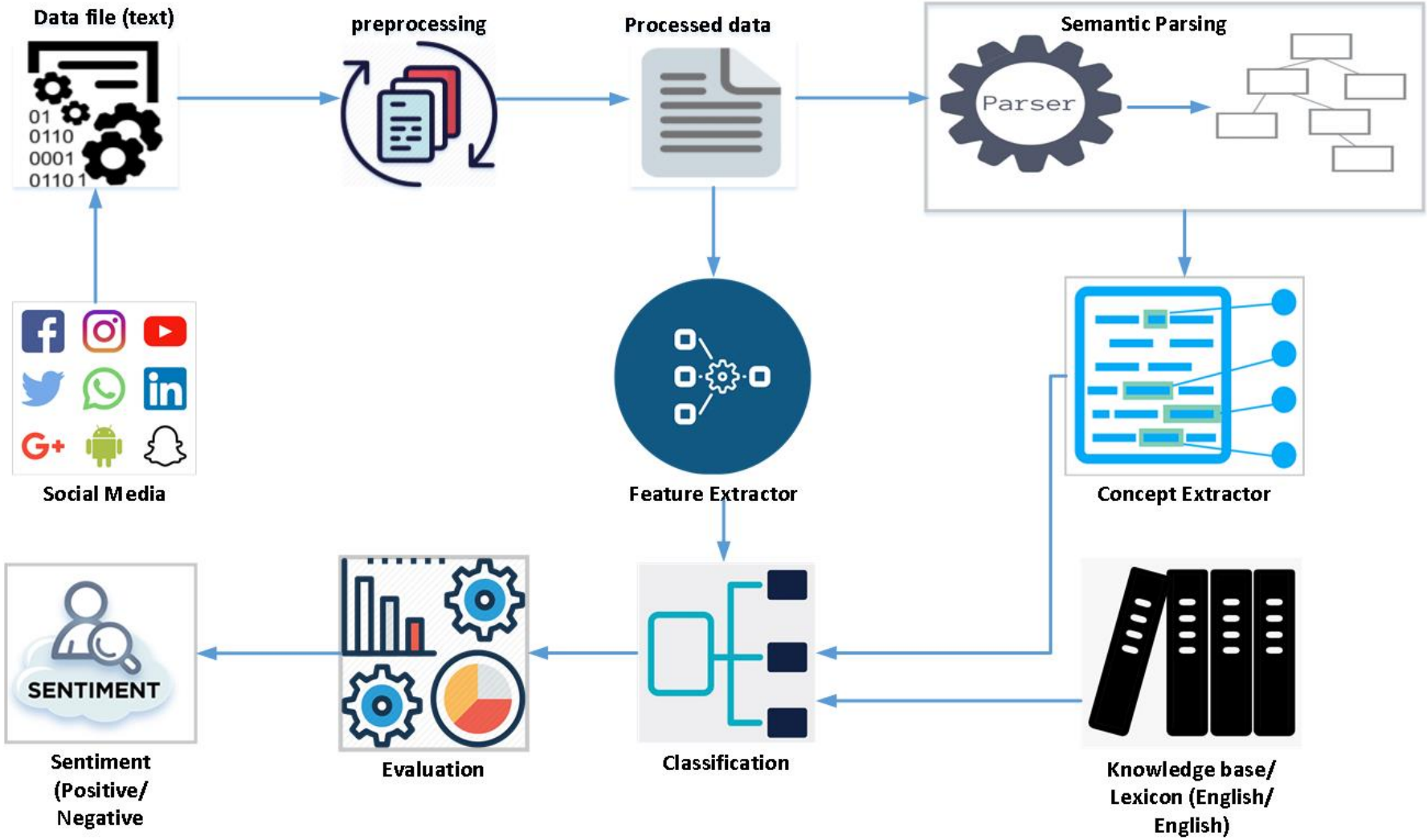

Figure 3.8: Abstract View of Figure 3.1 Showing Resources and Techniques Evaluation Process

3.9 Tools

The tools and algorithms used in different stages of this research are shown in Table 3.2. For instance, the data was collected from different data sources such as Facebook pages and groups. The data were collected using Facebook API (Graph API) and Python crawler tools. The preprocessing was done using Python, NLTK library, and R. The extraction is done based on features and concepts. Several machine learning algorithms (NB, SVM, LSTM, and MCSAlgo) and resources (knowledge base (SenticNet 5, BanglaSenticNet) and concept polarity lexicons) were used for sentiment analysis. Finally, two famous representation tools, Excel/SPSS, were used for data and result presentation.

Table 3.2: Overview of Tools/Algorithm

Steps	Tools/Algorithm
Data sources	Facebook Pages or Groups such as International Islamic University (IIUC) pages data; ABSA Cricket; IMDB
Data collection	Facebook API (Graph API), Python crawler
Data preprocessing	Python, NLTK library, R
Feature Extraction/Selection	Feature/concept based
Sentiment analysis	NB, SVM, LSTM, MCSAlgo, Concept-based (SenticNet, BanglaSenticNet)
Data/Results presentation	Excel/SPSS (Tabular and Graphical form)

3.10 Summary

This chapter has introduced the whole process of this research. Knowledge base is the most important source for finding the implicit meaning of the words and sentences. As per literature, a famous knowledge base, namely 'SenticNet' version 1 to 5, is widely

used for dealing with English sentences. Literature shows, this resource is also converted to 40 other languages except for Bengali. This study tried to fill the gap of the inadequacy of the Bengali Knowledgebase by creating a knowledge base parallel to that of SenticNet 5 but using a different method to deal with Bengali sentences. The knowledge base creation process and algorithm are described in detail in this chapter. This research explicitly demands concept extraction rules to be defined; thereby, many concept extraction rules are defined in this chapter.

The Lexicon is considered the essential resource for emotion detection. There are many English lexicons available; however, only a few Bengali lexica exist. This study proposed an algorithm for Bengali lexicon creation. The complete lexicon creation process and algorithm are presented in this chapter. The lexicon is named 'SenticNet 5 Bangla lexicon' and contains 72433 concepts.

This chapter also presented a proposed algorithm (MCSAlgo) for concept-level multilingual sentiment analysis. The algorithm is so proposed that it could be used to deal with any lingual data; for instance, only bilingual data such as English and Bengali data were tested.

Moreover, creating a comprehensive multilingual dataset containing many variables and labels from different languages requires appropriate principles and processes. This chapter has elucidated such a process for multilingual dataset creation and proposed an algorithm and framework. Besides, a framework for annotated data set creation is proposed.

One of this research aims to find the optimal features and concept extraction techniques for MLSA. In sentiment analysis, feature extraction or concept extraction

techniques are considered extremely important. The study uses the multilingual (English, Bengali) corpus, features, and concepts from both languages that needed to be extracted with equal importance. Therefore, this research proposes an algorithm for finding optimal feature or concept techniques. The feature extraction techniques considered in this thesis are TD - IDF (Simple), unigram, bigram, trigram, parts of speech, and concepts. The entire extraction process is described in this chapter.

With suitable pre-processing techniques, classification effectiveness and accuracy may be enhanced, and performance would be improved. In multilingual sentiment analysis, pre-processing needs to be done on the data of different languages (e.g., English and Bengali) separately. Therefore, as per the objectives, this research conducted the different types of pre-processing such as remove punctuation, negation, reduction of letter reputation, stop word deduction, stemming, tokenization, and case conversion (Individually and in combination) on the text or sentences supplied to improve the performance of multilingual sentiment classification. This study also proposed an algorithm to find the optimal preprocessing technique or their combinations. Total pre-processing steps are shown in the topology.

The evaluation procedure of knowledge bases, concept polarity lexicons, datasets, feature and concept extractions techniques, preprocessing techniques is also described in this chapter. Finally, the list of tools used in this research is presented. The next chapter gives the details of the data analysis and their interpretation.