M.Sc. Engg. (CSE) Thesis

Developing a Concept-Level Polarity Detection Model through Generation of a Rule Based Semantic Parser for Bengali Sentences

Submitted by

Md Fazle Rabbi

1015052096

Supervised by

Dr. Muhammad Masroor Ali

Submitted to

Department of Computer Science and Engineering

Bangladesh University of Engineering and Technology

Dhaka, Bangladesh

in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

March 2019

Dedicated to my parents

Candidate's Declaration

I, do, hereby, certify that the work presented in this thesis, titled, "Developing a Concept-Level Polarity Detection Model through Generation of a Rule Based Semantic Parser for Bengali Sentences", is the outcome of the investigation and research carried out by me under the supervision of Dr. Muhammad Masroor Ali, Professor, Department of CSE, BUET.

I also declare that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Md Fazle Rabbi
1015052096

The thesis titled "**Developing a Concept-Level Polarity Detection Model through Generation of a Rule Based Semantic Parser for Bengali Sentences**", submitted by Md Fazle Rabbi, Student ID 1015052096, Session October 2015, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering and approved as to its style and contents on March 11, 2019.

Board of Examiners

1. ______________________ 11.03.2019

Dr. Muhammad Masroor Ali — Chairman (Supervisor)
Professor
Department of CSE, BUET, Dhaka

2. ______________________

Dr. Md. Mostofa Akbar — Member (Ex-Officio)
Professor and Head
Department of CSE, BUET, Dhaka

3. ______________________

Dr. M. Kaykobad — Member
Professor
Department of CSE, BUET, Dhaka

4. ______________________

Dr. Md. Rifat Shahriyar — Member
Associate Professor
Department of CSE, BUET, Dhaka

5. ______________________

Dr. Md. Mahbubur Rahman — Member (External)
Professor
Department of CSE
Military Institute of Science and Technology, Dhaka

Acknowledgement

I would first convey my heartfelt thanks to my thesis supervisor Professor Dr. Muhammad Masroor Ali for being so generous in guiding me whenever I ran into a trouble spot or had a question about my research or writing. He steered me in the right direction whenever I needed it. His patience guidance helped me in all the time of writing of this thesis. I could not have imagined having a better supervisor and mentor for my MSc study and onward.

Besides my supervisor, I would like to thank the rest of my thesis committee: Professor Dr. Md. Mostofa Akbar, Professor Dr. M. Kaykobad, Associate Professor Dr. Md. Rifat Shahriyar, and Professor Dr. Md. Mahbubur Rahman for their insightful comments and encouragement, but also for the hard question which incented me to widen my thesis from various perspectives.

I would be always grateful to the Bangladesh Army for giving me the opportunity to attend my MSc study. I would also be thankful to MIST authority for sparing me as and when required. I also acknowledged the cooperation of my colleagues, who were always at my side.

Finally, I must express my profound gratitude to my spouse for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. Last but not the least, I would like to thank my parents and family for inspiring me spiritually throughout my life in general.

Dhaka
March 11, 2019

Md Fazle Rabbi
1015052096

Contents

List of Figures

List of Tables

Abstract

Public opinion over the Internet is getting importance with the rapid growth of online content every day. The sentiment of public opinion is considered a valuable piece of information in every interaction of human life. Concept-based approaches are the recent evolution in sentiment analysis, which is intended to infer the semantic and affective information associated with natural language opinion. Sentiment analysis at the concept level introduces a new opportunity for information retrieval like polarity detection, especially for a less privileged language like Bengali. In this work, a rule-based semantic parser is developed to generate the parse tree for a Bengali sentence. Concepts are extracted from the parse tree exploring the dependency among the constituents of the sentence. A domain specific classification model is proposed to detect the polarity of the concepts which in turn are used to find the sentence polarity through the parse tree traversal. Here, the AffectiveSpace is used as a knowledge base. Training data on targeted domain is generated from online contents using term frequency and inverse document frequency (tf-idf) where the concepts are labeled as positive, negative and neutral. The model uses the Linear Discriminant Analysis (LDA) to classify the training data where 81.8 percent of original grouped concepts correctly classified. The performance of the polarity detection method is evaluated using the precision and recall method. The overall accuracy for concept-level polarity detection is 70.24 percent. Whereas the accuracy at the sentence level is 65.63 percent for the simple sentence, and 73.77 percent for the complex or compound sentence, which can be considered an acceptable range for a less privileged language like Bengali. One of the limitations of the work is its failure to achieve the desired level of abstraction in forming the concept due to the language complexity of Bengali. Therefore, it is fully dependent on the terms available within the sentence and translate those to English for mapping in the AffectiveSpace. However, An independent dependency parser for Bengali can be generated by integrating the language morphology along with the language syntax to extract the concept with a high level of abstraction. Moreover, the generation of a Bengali affect space can be of great use in the field of NLP.

Chapter 1

Introduction

With the rapid development of the World Wide Web, human activities over Internet is growing vast. People have an inherent curiosity to discover what others are thinking. Public opinion is also getting importance in every interaction of personal, professional, social and political sectors. Online newspaper, blogs, discussion groups, tweets and comments on social and electronic media are a great source of public opinion. The opportunities to retrieve the public opinion from these unstructured data have opened up the area of research on Natural Language Processing (NLP) especially, the Sentiment Analysis (SA). The sentiment of the public opinion is considered a valuable piece of information to the business organization, social workers, government bodies and even law enforcement agencies for decision making. Polarity detection at sentence or concept level further enhance the applications.

1.1 Background

Existing works on NLP [8] can be broadly divided into two main categories. Firstly, the rule based approaches are focused on the construction of NLP tools like Parts of Speech (POS) tagger, Named Entity Recognizer etc. Since these methods explore the relationship among the lexicons within the sentence or document, therefore, these are language and domain independent and greatly rely on some knowledge bases. On the other hand, machine learning approaches utilize the knowledge through training data and use some classification methods for categorization of the sentiment. However, most of the recent work on sentiment analysis integrates both the approaches to overcome the limitations of both, and thereby, improve the efficiency of the works. In addition, recent evolution in the field of sentiment analysis known as affective computing [2] is getting popularity in the NLP research community. The novelty of this paradigm is its capability to merge the linguistic techniques with common-sense computing. Thereby, it facilitates to properly deconstruct the text into concepts, hence, improve the accuracy of polarity detection.

Polarity is detected in document, sentence or concept level. Feature based classifications are efficient in sentence or document level, but perform poorly in concept level as they rely on its contextual meaning. As concepts are independent of languages, it facilitates using available resources in other languages such as SentiWordNet [9] and WordNet Affect [7] for sentiment analysis. A significant number of words can take different polarities when associated with other words to form a concept. Concept can be categorized within the affective space generated from WordNetAffect intended to infer the conceptual and affective information for a specific domain. Therefore, AffectiveSpace [3] is considered a good resource to improve the efficiency of sentiment analysis at concept level.

Bengali being the national language of Bangladesh as well as the sixth highest spoken language in the world [10] draws the attention of the NLP researchers. The contents in Bengali over the Internet are also increasing every day through social media, news portals, blogs and other online platforms. These offer a huge volume of unstructured text data for information retrieval. Not much effort has been taken up in Bengali languages on polarity detection at concept level. The scarcity of NLP tools and linguistic resource makes the NLP task in Bengali more difficult and inconsistent. Parsing the sentence is the precondition for any NLP task. Whereas, no recognizable parser is available for Bengali text that can decompose the Bengali sentence into terms and key phrases depending on semantic and syntactic relationship among them [11]. Moreover, the complexity of Bengali grammar and ambiguities in sentence structures raise the challenges for working in Bengali. However, recent interest of the researchers in the field of Bengali opens a wide variety of scopes to work.

Domain dependency is inherited to any language [12]. It is more significant in determining the sentiment at concept or sentence level. For instant, a single concept may contain multiple sentiments depending on the context of the text. As a consequence, domain specific methodologies offer better performance than generalized method in sentiment analysis. Knowledge bases used for NLP tasks still provide very limited coverage for domain specific words. On the other hand, training data mainly contributes to domain specification of any method. Prior to any NLP task parsing the text or sentences is very crucial for the subsequent stages of the work. This solicits an independent parser in Bengali for deconstructing text into phrases hence, extract the concepts.

The performance of NLP approaches can vary due to the nature and quality of the data set. Preparation of the training data is the most important process in a classification method as it highly influences the performance of the classifier. On the other hand, the generation of a good standard corpus is highly important to evaluate the method. Though a notable number of good standard corpus is available in rich languages like English, Chinese, French; there is hardly any Bengali corpus to be mentioned [13]. Therefore, construction of well annotated corpus is a prerequisite for an NLP task, especially, in Bengali.

1.2 Problem Definition

Concept-based sentiment analysis is intended to infer the semantic and affective information associated with natural language opinions. There is hardly any existing method for sentiment analysis at the concept level in Bengali. A concept can take different polarities based on the context domain. To address the fact, domain specific methods are preferable in sentiment analysis. Moreover, the concepts are not just the terms within a sentence rather, refers to a group of terms, key phrases or even clauses that contain an idea. Therefore, concept extraction from the sentence is subtle.

Prior to performing the polarity detection, generation of an independent semantic parser for Bengali sentences is a requirement. The influence of modifiers and the dependency among the concepts need to be addressed with proper attention to increase the accuracy of the polarity detection. Integration of rule based and machine learning method facilitates resolving the limitations of existing approaches, hence improve the performance.

Construction of the data set demands special care as it affects the performance of the NLP task in subsequent stages of the work. The challenges for Bengali are more for the scarcity of NLP tools needed for annotation of unstructured text data. Annotation of the text with the appropriate POS tagging to each token in a sentence presents an idle corpus for parse tree generation. In this context, the accuracy of the POS tagger is very crucial as it affects the subsequent stages of the work.

1.3 Research Aim and Objective

Based on the problem definition and the applications, objectives for the thesis are formulated with specific aims. The objectives of this thesis with specific aim are listed as followings:

1. To generate a rule based parse tree generator for the Bengali language to break the text into clauses and deconstruct the clauses into key phrases that will be used as input to concept extractor.

2. To develop a method to extract the concept from the parse tree that is likely to express sentiment for polarity detection.

3. To derive a classification model to determine the concept valence and hence, determine the sentiment based on the cognitive and affective information associated with natural language.

4. To develop a simple method for polarity detection at sentence level without using so many NLP resources to detect the sentiment of a sentence.

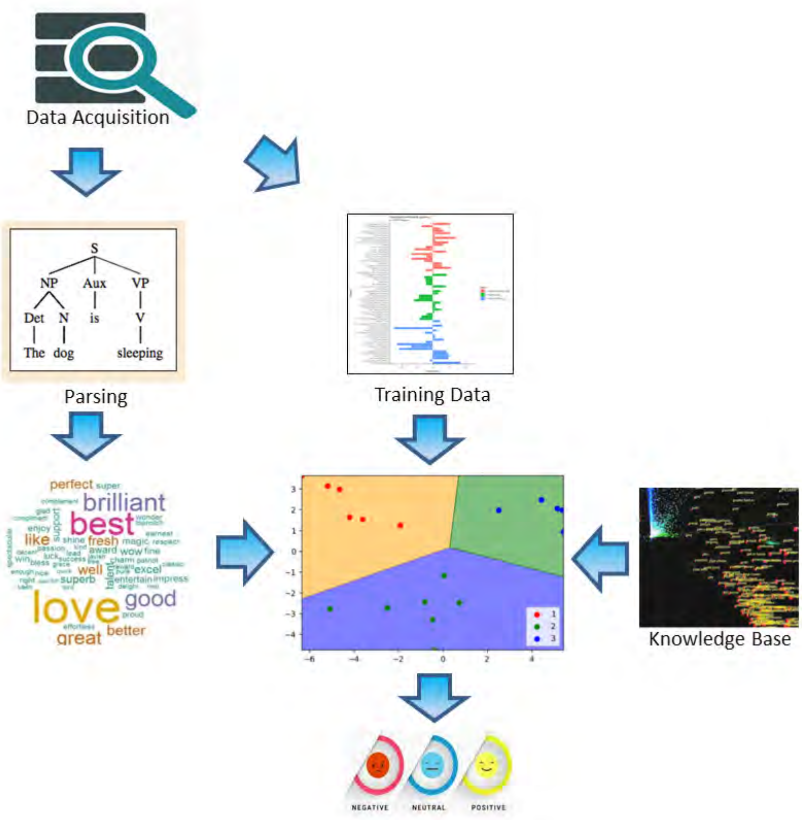

Figure 1.1: An overview of the proposed work.

1.4 Overview of the Work

In this work, a classification model is proposed to detect the polarity of the concept; hence, the sentence polarity is determined. Before proceeding to the classification of concepts, a methodology is developed for extracting the concepts from the annotated data set. Experimental data set will be generated from popular daily newspapers and online blogs on the field of a targeted domain. The data set is annotated with POS tags. Then, a rule based semantic parser decomposes the sentence into a parse tree representation. Concepts are extracted from the parse tree using some tree traversal rules developed for this purpose. The polarity of the concept is detected using the classification model. A pictorial overview of the proposed work is presented in Figure 1.1. A stepwise enumeration of the proposed methodology is summarized as follows:

1. At first, the data acquisition is carried out from available resources on a targeted domain. The annotated data set is generated through tokenization and POS tagging using standard NLP tools like; tokenizer, POS tagger, etc. Depending on the syntactic and semantic relations among the tokens, a rule based parse tree is generated to decompose a sentence into clauses, key phrases, modifiers and other constituents. The parse tree generated for each sentence also exhibits the dependency of the modifiers and the tokens within a sentence.

2. Next, the concepts containing sentiment are extracted from the parse tree. The method for extracting the concepts is developed relying on the syntactic and semantic association among the key phrases and terms within a sentence. Semantic relations are mostly dependent on linguistic morphology and language grammar. Whereas, the syntactic relations explore POS information and the stemming varieties. Identification of this relationship is very effective for detecting the dependency among the tokens and modifiers. The dependencies of the modifiers being a vital factor in determining the sentence polarity are also be extracted in this phase.

3. Concepts are the root constituents for the polarity detection model proposed in this work. AffectSpace is used as a knowledge base in this model to explore the affective reasoning among the concepts. Linear discriminant analysis is used as a classification model with three classes named as positive, negative and neutral. The model uses the training data on the targeted domain to initiate the classifier which in turn, accepts the concepts and determines the polarity class.

4. Finally, the polarity of the sentence is determined to analyze the relationship of the concepts and modifiers. A tree traversal in the reverse order is executed starting from the concepts to the root of the parse tree of the sentence. Here, the rules are highly sensitive to the dependencies that were detected in the earlier stage. The modifiers are very crucial to handle as these can alter the polarity of the sentence at any stage of the tree traversal in the sentence.

5. On completion of the classification, the evaluation of the experimental result is carried out using test data. Performance of each individual method along with the overall outcome of the model is explained through result analysis. We use the precision and recall metrics for evaluating the output of the model. In case of parse tree and concept extractor, we compare the output of our methods with the expected result. Evaluating metrics such as; precision, recall and F1 score values are calculated for each of the three classes. Moreover, the accuracy of every class and the overall accuracy is also determined for the test data to estimate the overall performance of the model. An endeavor is made to explain the causes of exceptions in the performance of the methods presented in this work. In addition, a comparison is made between binary classification and classification with tree classes for better analysis of the outcome.

The proposed work is based on Bengali language. However, some resources and knowledge bases from English are also been used. As the linguistic morphology and language grammar for each language are different from each other, therefore, the parse tree generation and concept extraction are language dependent. On the other hand, the classification method highly relies on the dictionary meaning of the concepts which can be considered language independent. Since the meaning of the concept may take different form based on the literature or sentence context, they are domain dependent.

1.5 Thesis Contribution and Final Outcome

The contribution of this thesis is two-fold. Firstly, we explore the language dependencies in the field of NLP and focus on a less privileged language like Bengali. In this context, our concentration focus on sentence parsing, which is language dependent and is a prerequisite to any NLP task. The parser is capable of generating parse tree through decomposition of a sentence that is very rare of this kind in Bengali. In addition, the dependencies among the constituents of the sentence and the modifiers are also detected in the process of parsing. Though the rule based parser is generated focusing the concept extraction for the proposed classification model, it can be used for other NLP task as well. Thereby, an NLP tool of this kind can facilitate wide varieties of opportunities in the field of Bengali language processing works.

Secondly, we introduce the concept as a constituent to be used in polarity detection. Concept level sentiment analysis is a novel integration to the field of Bengali of this kind. In performing the classification task, we utilize the domain dependencies of the concepts to improve the accuracy of the polarity detection model. The novelty of this model is the blending of the semantic and affective information associated with natural language. In addition, the concept-based approaches step away from the blind use of keywords and word co-occurrence counts but rely on the emotion categorization to detect polarity.

The contributions can be more visualized over the possible outcomes of the thesis. A set of tools and models with their specific capabilities are developed based on this work. The outcome of this thesis are listed as follows:

1. A semantic parser to generate the parse tree for Bengali sentence with the capabilities to identify dependencies the constituents of the sentence.

2. A concept extractor which can generate the concept containing sentiment from the parse tree.

3. A novel model for concept level polarity detection using affect space for the specific domain of context. The domain can be altered through the generation of training data only.

4. Finally, A polarity detection method for determining the sentence polarity which utilizes the interrelationship and dependencies among the concepts and modifiers.

Though the classification method is independent of the language, the domain dependencies improve the performance of the model to detect the polarity at the concept level. Use of

AffectiveSpace as knowledge base improve the accuracy of the model by exploring the affective associations among the concepts.

1.6 Thesis Outline

This thesis is further organized in four more chapters. Chapter 2 discusses the existing works related to our problem definition. Through a detail literature review, it highlights the scope of the works based on the limitation found in the state of the art. At the end of this chapter, the research questions for the thesis are enumerated. With a view to better understanding the methodology of the proposed work, we explain some of the NLP fundamentals along with the NLP techniques used for this work in Chapter 3. We also provide a precursory idea on the NLP resources and knowledge bases related to our model. The mathematical models used for the classification scheme are also explained in this chapter. In Chapter 4, we briefly represent the methods developed in the proposed work. The working procedure for each method is explained through examples and applications. Finally, the experimental analysis is performed in Chapter 5 to evaluate the performance of the model using the test data-set. Finally, we will conclude the thesis by summarizing the proposed work and highlighting some future scope of the work.

Chapter 2

Related Works

Early works on Natural Language Processing (NLP) [14] are focused on the construction of NLP tools like parts of speech (POS) tagger, named entity recognizer (NER) etc. With the advent of World Wide Web, the research on NLP has been diversified to multiple branches like text summarization, subjectivity detection, opinion mining and sentiment analysis. Opinion mining or sentiment analysis is an active area of study in the field of NLP, to find an automated way to determine the expression or emotion from text. Research on sentiment analysis has two main directions: 1. feature-based opinion mining which identifies the features within a review sentence to determine the orientation of the opinion, 2. sentiment classification to determine the polarity of opinionated sentence or document. This chapter will discuss the present state of the related works on sentiment analysis specially focused on polarity detection. Section 2.1 will discuss the different methods of sentiment analysis and Section 2.2 will discuss the related works on semantic parsing and concept extraction. Works on sentiment analysis in Bengali will be discussed in Section 2.3 and scope of the works found through related works will be highlighted in Section 2.4. Finally, an endeavor will be made to formulate the research questions in Section 2.5 to be address in this work.

2.1 Sentiment Analysis

Existing works in sentiment analysis [8] can be broadly grouped as rule based approaches and machine learning approaches. The former one offers high syntactic sensitivity, but performs poor in context or domain dependent classification whereas; the later is time-consuming and expensive to manage the large amounts of training data necessary for good performance. Concept-based approaches [15] are the recent evolution in sentiment analysis which are intended to infer the contextual and affective information associated with natural language opinions.

2.1.1 Rule Based Approaches

Most of the rule based approaches [8] are lexicon based, though some corpus based methods are also introduced. Lexicon based approaches explore the syntactic and semantic relationship among the lexicons within a corpus for sentiment classification at sentence or document level. Lexicon-based approaches mainly rely on sentiment lexicon, i.e. a collection of pre-generated sentiment terms, phrases and even idioms, developed for traditional genres of communication. These methods identify the sentiment lexicons applying some rules based on language syntax. POS information is most commonly used to exploit syntactic relation among the lexicons within a sentence. Polarity of these sentiment words are determined by exploring the semantic orientation of the word. Adjectives have been employed as features by a number of researchers [16, 17].

One of the earliest proposals for the data-driven prediction of the semantic orientation of words was developed for adjectives [18]. Here [19], a method is proposed for determining the semantic orientation of terms through gloss classification [19]. Sentence or document level polarity is determined using the semantic association among the lexicons. Therefore, they rely primarily on the underlying sentiment or opinion words. These sentiment words are generally labeled according to their semantic orientation as either positive or negative [20]. NLP resources like SentiWordNet [9], Affectnet [21] are very popular to acquire the semantic orientation at word or phrase level. Semantic orientation of sentiment lexicons is annotated as positive, negative or neutral using SentiWordNet for sentiment/ opinion classification in [22]. Sometimes, classification at sentence or document level is simply subjected to keywords and word co-occurrence counts within the corpus. In [23], a lexicon-based method is proposed to determine the orientation of opinion bearing words in a review sentence. The work basically counts the positive and negative words in the corpus and classifies the sentence as positive if the number of positive words is more than the number of negative words, else classifies as negative. A more sophisticated method based on relaxation labeling has been introduced in [24]. A lexicon-enhanced method for the sentiment analysis of user generated reviews has been introduced in [1] which incorporates many NLP tools and resources along with the rule-based classification scheme (Figure 2.1).

Rule based approaches have some major shortcomings. Firstly, it cannot handle the context and domain dependent words. The accuracy of sentiment classification can be influenced by the context of the items to which it is applied [25]. This is because, the same word can take different sentiment polarity in different domains or contexts. For instance, an observation was presented in [26] that "unpredictable" is a positive description for a movie plot, but a negative description for a car's steering abilities. Without prior knowledge of context, there is probably no way to determine the semantic orientation of a context dependent opinion word by looking at only the word itself. Obtaining such knowledge on the huge number of opinion words from a domain expert or user is also not scalable. A holistic lexicon-based approach is propose

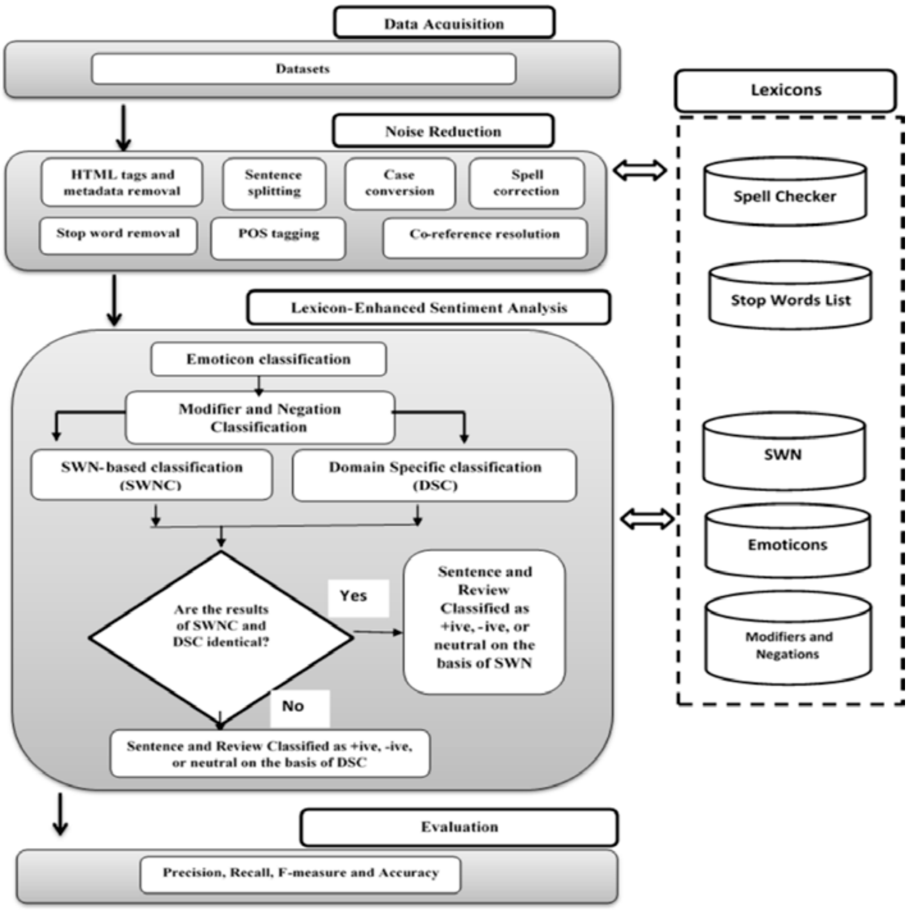

Figure 2.1: Rule based classification scheme integrated with different NLP resources (collected from [1]).

to deal with the limitations of lexicon based approaches in [27]. Contrary to locked in the current sentence, it exploits the external information and evidence in other sentences within the document. It also infers the linguistic attributes in natural language expressions to infer orientations of opinion words. Thereby, they also incorporate the semantic relations among the opinion words within the corpus. Secondly, occurrences of multiple conflicting words in a sentence raise dubiety in determining sentence polarity. For instance, the sentence, "he is honest but crabby", contains two conflicting words and makes the sentence polarity inexplicable. One of the pertinent solutions to address these problems is machine learning approaches.

2.1.2 Machine Learning Approaches

Several works on sentiment analysis and opinion mining have been done towards using machine learning approaches. These approaches are popular for wide range of coverage to the domain or context specific words. They also reduce the syntactic and semantic sensitivity of the lexicons within the corpus and improve the performance in sentence or document level. On the contrary, large amount of training data is required for good performance. Therefore, learning the model is time consuming and expensive for any NLP task. Some machine learning methods like Naive Bayes Classifier (NBC), Conditional Random Field (CRF), and Support Vector Machine (SVM) have been used for sentiment classification. NBC is a probabilistic classification method based on Bayes' theorem and used for both classifications as well as training purposes. SVM analyzes data and defines decision boundaries by having hyper-planes. In binary classification problem, the hyper-plane separates the document vector in one class from other class. CRF is a discriminating model, which can easily integrate various features.

The aspect of sentiment classification is categorization with positive and negative sentiments in [28]. Three different machine learning algorithms have been applied, such as: NBC, SVM, and Maximum Entropy (ME). In this framework, bag-of-words has been used as features to implement the machine learning algorithms. According to the analysis, SVM outperforms every other classifier in predicting the sentiment of a review. A comparison of results obtained by applying NBC and SVM classification algorithm is presented in [29] and almost the same result is observed.

The proposed work in [30], presents the classification of Chinese comments based on word-to-vector and SVM. It improves the performance of SVM by using word-to-vector tool to cluster similar features in order to acquire the semantic features in selected domain. In the later part, the lexicon based and POS based feature selection approaches are adopted to generate the training data in a specific domain. Though the method is very useful for domain specific sentiment analysis, it fails to handle the influence of modifiers which can change the valence of the features within the domain.

Existing works using CRF, demonstrate that it performs well for aspect based sentiment analysis. For example, with CRF, sentiment classification has been performed in sentence and document level in [31]; the work [32] identifies opinion expressions from news-wire documents. CRF is well known for sequence labeling tasks. To conform, the context information of the aspect terms are sequentially leveled in aspect based sentiment analysis [33]. Text representation and additional positional features of CRF have been improved by MFE-CRF [34] that introduces Multi-Feature Embedding (MFE) clustering based on the CRF to improve the effect of aspect term extraction.

The performance of these approaches can vary due to the nature of the data sets used, the quality of the training data set and the various technical approaches used internally. Since the performance of a machine learning methodology heavily depends on the choice of data representation, much effort is given for data analysis to build powerful feature extractor. Variant in mathematical model can also improve the performance of these methodologies. Recently, fuzzy logic and deep learning [35] approaches are emerging as powerful computational models that discover intricate semantic representation of text automatically from data without feature engineering.

2.1.3 Concept Based Approaches

Opinions and sentiments do not occur only at document-level, nor are they limited to a single valence or target. In recent works, text analysis granularity has been taken down to phrase level. Rule based methods are efficient for lexicon based sentiment analysis, whereas, at document-level, machine learning techniques have reached to a satisfactory level of efficiency. Unfortunately, these approaches are still far from being able to infer the cognitive and affective information associated with natural language. Approaches of these kinds are mostly dependent on knowledge bases that are still too limited to efficiently process text at document and sentence-level. These methods, however, are not semantically very strong because of their reliance on parts of text in which opinions are explicitly expressed, i.e. verbs, adjectives and adverbs. Moreover, such text analysis granularity might still not be enough as a single sentence or words may contain different opinions about different context or domain. In order to properly infer the meaning of the text within a specified context or domain, a NLP system must be developed to extract concepts rather than specific opinion words or phrases within the sentence.

A novel paradigm to sentiment analysis has been introduced in [2] that merges linguistics, common-sense computing, and machine learning for properly deconstructing natural language text into concepts and, hence, for improving the accuracy of polarity detection. ConceptNet [36], a freely available large-scale commonsense knowledge base is introduced as an integrated natural-language-processing tool-kit that supports many practical textual-reasoning tasks over real-world documents. In ConceptNet, the nodes of WordNet are extended from purely lexical items (words and simple phrases with atomic meaning) to higher-order compound concepts, which can devise an action verb with its relevant arguments (e.g. 'buy food', 'drive to store'). ConceptNet is a directed graph in which the concepts are represented as nodes and the edges are labeled with attributes that are sometime composed through common sense and interconnected the concepts (Figure 2.2). A linguistic resource named WordNet-Affect [7] has been presented for the lexical representation of affective knowledge from WordNet.

Efficiency of concept-level sentiment analysis depends on anticipating the affective valence of unknown multi-word expressions. In the significant work [3], a new data-set is built up by applying blending technique on ConceptNet and WordNet-Affect to represent as a single matrix. Truncated singular value decomposition (TSVD) is performed on the matrix for dimension reduction that yield a 50-dimensional vector space, which they named AffectiveSpace (Figure 2.3). AffectiveSpace is a major step out in concept based sentiment analysis in which common sense and affective knowledge coexist. In AffectiveSpace, different vectors represent affective valence, ways of making binary distinctions among concepts or emotions. Affective states of the concepts go from strongly positive to null to strongly negative within the space. Therefore, concepts with the same affective valence are likely to have similar features. Likewise, concepts expressing the same type of emotion tend to position closer to each other in affective space. However, similarity or affinity among the concepts in the affective space does not depend

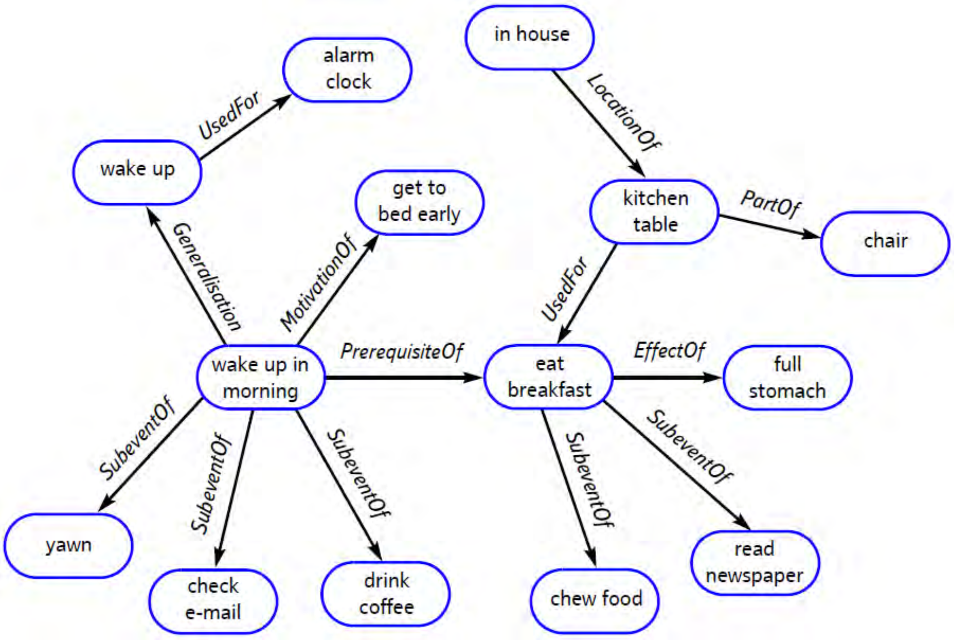

Figure 2.2: Fragment of a ConceptNet (collected from [2]).

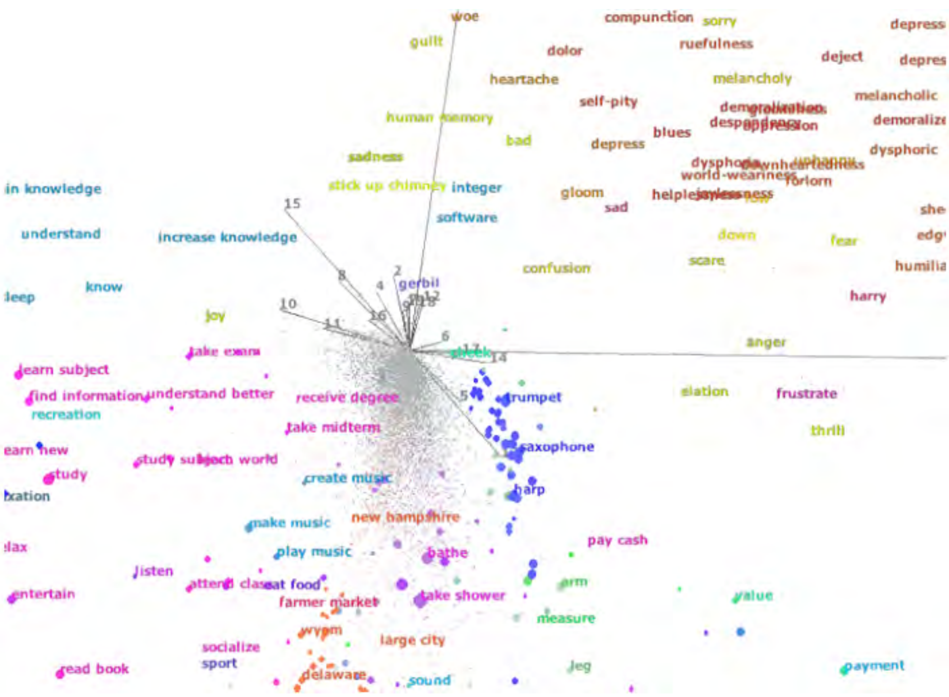

Figure 2.3: A sketch of AffectiveSpace [3]. Affectively positive concepts (in the bottom-left corner) and affectively negative concepts (in the up-right corner) are floating in the multi-dimensional vector space.

on absolute position of the concepts rather, on their relative position.

A new paradigm known as Sentic Computing is presented in [4]. In the Sentic Computing, four dimensions are taken as basis to classify the affective states: Sensitivity, Attention, Pleasantness and Aptitude. The transition between different emotional states within the same affective

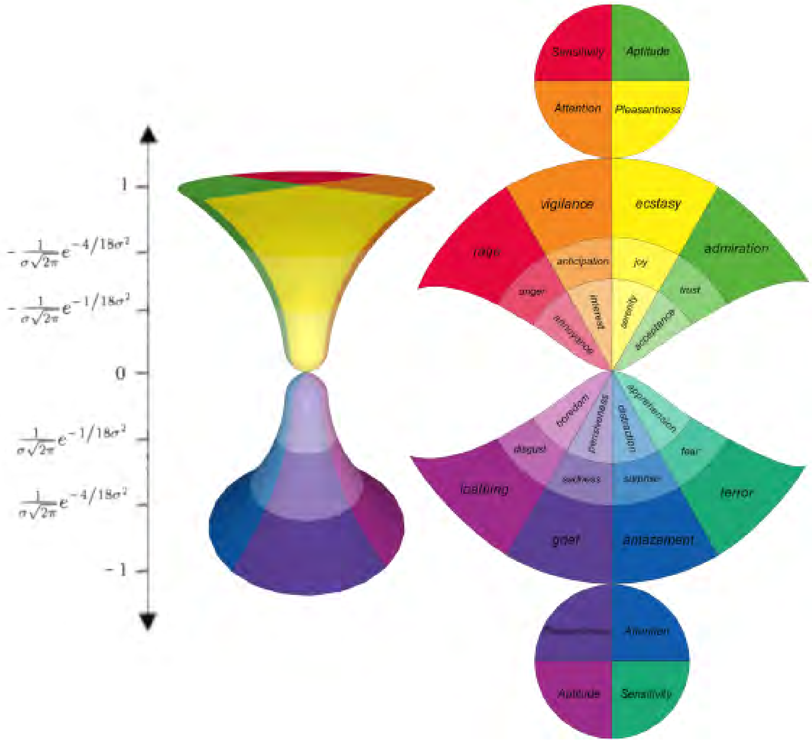

Figure 2.4: The 3D model and the net of the Hourglass of Emotions [4]. Since affective states go from strongly positive to null to strongly negative, the model assumes an hourglass shape.

dimension, when depicted, generates a symmetric inverted bell curve shape. Mapping of this curves of possible emotions leads to an hourglass shape (Figure 2.4), hence, the name the model is hourglass of emotion [37]. In this framework, the vertical dimension represents the intensity of the different affective dimensions, while the radial dimension represents emotion categorization. Though the Hourglass model is designed for emotion detection, it also used for polarity detection tasks. Since polarity is strongly connected to attitudes and feelings, it is defined in terms of the four affective dimensions in [4], according to the formula:

$$p = \sum_{i=1}^{N} \frac{\text{Pleasentness}(c_i) + |\text{Attention}(c_i)| - |\text{Sensitivity}(c_i)| + \text{Aptitude}(c_i)}{3N} \tag{2.1}$$

Where, c_i is an input concept, N the total number of concepts. The major observation in determining the polarity using hourglass emotions through the equation-1 are: firstly, measuring the emotion is very complex and costly, secondly, emotions like attention and sensitivity always contribute positively and negatively respectively, which may not be true in all cases.

The concept based approaches step away from blind use of keywords and word co-occurrence counts and dependency on dictionary based NLP resources. They allow the aggregation of conceptual and affective information, but rely on the emotion categorization, for example, hour glass of emotion to detect polarity. Some emotions do not provide any valence, and sometimes, can be universally positive or negative. This kind of emotion increases the complexity of the polarity detection method using hourglass emotions. On the contrary, affective space provide a uniform association among the concepts within the space. Therefore, determining the concept valence from affective information rather than emotion categorization will provide a simpler and efficient polarity detection model. Moreover, an attempt to find a domain specific feature subspace of affect space will further enhance class separability.

2.2 Sentence Parsing and Concept Extraction

In the scientific paper [38], a high performance syntactic and semantic dependency parser is presented. This system takes a sentence as input and performs syntactic and semantic annotation using the CoNLL 2009 format [39]. The dependency parser is based on Carrerass algorithm [40] and second order spanning trees. Overall performance of the semantic analyzer consists of several sub-tasks, such as: tokenization, lemmatizing, tagging, sentence parsing and dependency detection. Application of context free grammar rule is very crucial to explore the dependency among the tokens.

A rule based technique to parse Bengali sentence using context free grammar rules is presented in [11]. This work analyzes Bengali sentences in three phases. Firstly, lexical analysis phase performs sequential scanning of the characters' stream and groups them into tokens or lexicons. Secondly, Syntax analysis phase groups the lexicons having a collective meaning. Finally, semantic analyzer ensures that the discrete input components fit together meaningfully. This phase is highly application-dependent and is regulated by the norms and rules of the concerned natural language. One of the achievements of their work is the capability of handling the complex and compound Bengali sentences. However, they still failed to determine the dependency among the sentences as well as modifiers.

One of the notable works in Sentic Computing [2] presents a semantic parser with a view to facilitating concept extraction from the text. The role of the semantic parser is to break text into clauses and hence deconstruct such clauses into concepts, to be fed later to affect space. Knowledge on the semantics associated with text and some affective information associated with such semantics are often sufficient to perform tasks such as emotion recognition and polarity detection. In this connection, the parser uses the POS information to break the text into clauses, then, in the next step, verb and noun chunks are separated following the Stanford Chunker (Figure 2.5) [41]. Finally, a POS based bi-gram algorithm is applied to extract the concept implementing some association rules.

The major disadvantage of this parser is the language dependency. Against this backdrop, efficiency of NLP tools like POS tagger plays an important role in the overall performance of the parser, hence, extraction of concepts. Therefore, it is necessary to generate a separate independent semantic parser, particularly for a less privileged language like Bengali.

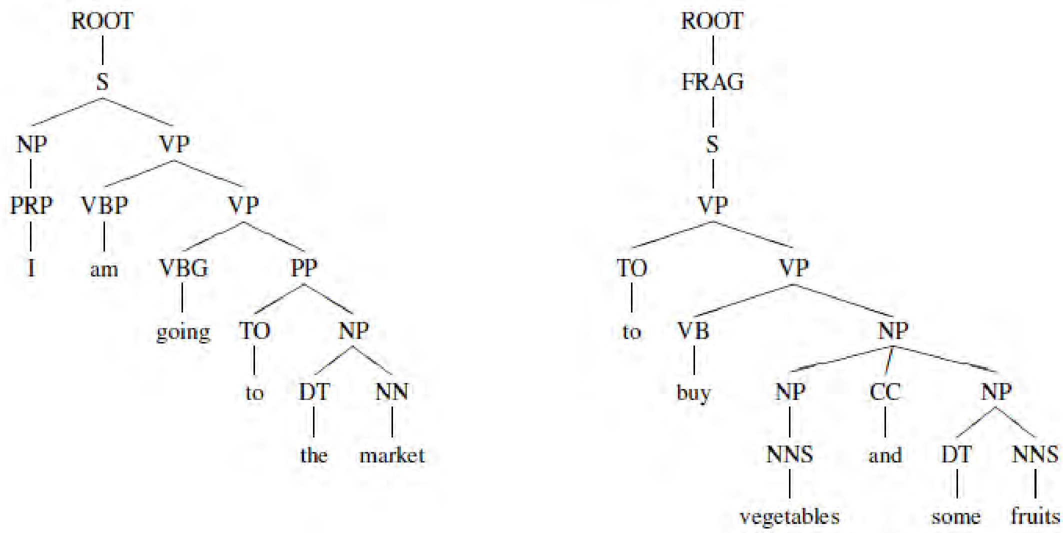

Figure 2.5: Parse Tree generated by Stanford Parser [5].

2.3 Sentiment Analysis in Bengali

Bengali is the national language of Bangladesh and the second most spoken language in India as well as the primary language of West Bengal. It is also the sixth most spoken language in the world [10]. But, there are very few achievements in the field of NLP for Bengali. With the advent of World Wide Web, online contents are increasing a lot over the Bengali blogs, newspapers and social media. Researchers have also shown their interest in Bengali NLP task over the years. Most of the works focus on generation of NLP tools like Bengali POS tagger, NER, sentence parser and to some extent of text categorization. A few number of NLP resources are also available, for example: Bengali SentiWordNet, WordNetaffect, stop word list, etc.

A computational technique for developing Bengali SentiWordNet is proposed using English-Bengali bilingual dictionary and English sentiment lexicons in [42]. In that endeavour, transformation of the WordNetAffect in Bengali was carried out in [43]. The lists are updated with the synsets retrieved from the English SentiWordNet to make adequate number of emotion word entries. A few number of works on development of Bengali parser is also available. As mentioned earlier, a Bengali sentence parser was presented applying the context free grammar rules in [11]. Some works on subjectivity detection and sentiment classification are also available, which are mostly learning base computational methods. An effort is made for subjectivity detection and opinion polarity identification from Bengali news text using CRF [43] and SVM [44] respectively.

Though the interest of NLP researchers in Bengali is increasing, there is an acute scarcity of NLP tools and resources. Complexity of Bengali grammatical structures is a substantial hurdle for the parsing of Bengali sentences, such as, the verb can be placed in any position in the sentence. Sometimes, more than one or two verbs can form a single verb. For example, the sentence "অভিযোগ করা যেতে পারে " ([obɦidʒog kra dʒete pare], Can be complained), contains three auxiliary verbs. In this breath, the adjective and adverb can also be positioned as 'pre' or 'post' modifier to the noun or verb respectively. Heterogeneity among the literature, communicative writing and spoken form is another ambiguity for Bangle language. For instance, the literature form of the sentence "মোহাম্মদ একজন ভাল ছেলে " ([moɦammɔd ɛkdʒn bɦalɔ tʃʰele], Mohammad is a good boy) can take the form "মোহাম্মাদ ছেলেটি ভাল " ([moɦammɔd tʃʰeleti bɦalɔ], Mohammad is a good boy) in communicative writing. Whereas, the spoken form can be much crisp as "মোহাম্মাদ ভাল " ([moɦammɔd bɦalɔ], Mohammad is a good boy), where the verb is concealed within the sentence. These types of heterogeneity is very complex to handle in NLP task. Unlike other languages domain dependency is also a challenge in sentiment classification or polarity detection in Bengali. For example, consider the following sentence:

"কেউ ইভ টিজিংয়ের শিকার হলে কাছের থানায় গিয়ে লিখিত অভিযোগ করা উচিত " ([keu ibɦtidʒiŋer ʃikar ɦole katʃʰer tɦanae gie obɦidʒog kra utʃit], If someone is victim of eve teasing, she should complain to nearby police station).

In this context, the phrase "অভিযোগ করা " ([obɦidʒog kra], complain against) carries a positive sentiment though the phrase is generally used as negative sense. However, the integration of semantic association along with the syntactic parsing can handle those more efficiently.

2.4 Scope of the Work

Rule based approaches are popular for their accessibility and simplicity in lexicon based methods. On the other hand, feature based classifications are efficient in sentence or document level and provide limited coverage to domain dependency. There are some major weaknesses of these methods as discussed in early sections of this chapter. Firstly, the poor recognition of opinion words when modifiers are involved in influencing the sentiment valence of the sentence. A modifier can turn the polarity of an opinion word or phrase from neutral to negative or positive. Secondly, reliance on surface features ignoring the affective association among the words within the sentence. Concept based approach can be an effective solution to sentiment analysis resolving the above-mentioned limitations. Concept level sentiment analysis is intended to infer the semantic and affective information associated with natural language opinions. Concept can be categorized within the affective space generated from WordNetAffect, and allows the aggregation of conceptual and affective information. Moreover, an attempt to find a domain specific feature subspace of AffectiveSpace will further enhance class separability. It also offer the scope of building a classification model to determine the concept valence and hence, determine the sentiment associated with natural language.

Prior to any NLP task, parsing the sentence is very crucial for the performance in subsequent steps. Considering the complexity of the Bengali language structure and diversity in semantic composition, an independent semantic parser is necessary for deconstructing Bengali text into key phrases. It will also enable to determine the dependency among the key phrases as well as modifiers for Bengali language, and hence, generate a parse tree. Extraction of concepts from the parse tree before performing the classification will make the model more economic as well. To keep it simple and efficient, a rule based method can be developed to extract the concepts that are likely to express sentiment for polarity detection from the parse tree.

With the rapid expansion of the Internet, human interaction with the web is rising every moment. Users from different corners of the world are generating a huge volume of data over the World Wide Web. Along with English, varying amounts of data are growing in many other languages, particularly, over the social media sites like Facebook, Twitter and various web portals. Online blogs and newspapers in native languages are rapidly gaining popularity as they allow people to share and express their views about topics in public platform. In this regard, Bengali as the sixth most spoken language in the world, is also moving forward. Therefore, Bengali newspapers or blogs can be used to generate Bengali corpus for various NLP tasks.

2.5 Research Questions

Bringing an end to the discussion on related works, we can formulate the following issues as our research questions to be addressed in this thesis.

[RQ1] *What is the significance of developing an independent semantic parser for different languages in NLP?*

Efficiency of any NLP task is highly influenced by the parser to extract the desired constituent of the text. The syntactic parser, no doubt, is language dependent as the syntax for every language is unique to each other. On the contrary, the uniqueness of semantics of the languages is inconclusive to the NLP researchers. Related works reveal that: (a) a unified semantic parser can be developed based on the idea of universal language [45] to reduce the complexity of language dependency. (b) an independent semantic parser can be developed to increase the task efficiency . Some of the researchers come up with the idea to generate independent task specific semantic parser within the language.

[RQ2] *What genetic factors should be considered in determining the constituents of the language for sentiment analysis and, how can those be adopted to extract from text?*

Researchers have introduced different levels of text constituents like lexicons, keywords, concepts, key phrases, even clauses and sentences. With their merits and demerits, selection of appropriate constituent of the text has immense importance. Particularly, in sentiment analysis where the sentiment or opinion can vary a lot with the level of language granularity.

[RQ3] *How can the rule based and machine learning methods be integrated to develop an efficient classification model, eliminating the shortcomings of both?*

From the study of related works it was found that, most of the works are focused on either rule based or machine learning methods in carrying out sentiment analysis tasks. It will be very interesting as well as motivating to improve the efficiency of the sentiment classifier by integrating both methods with their positive contributions. In the same breath, determining a befitting mathematical model is very essential for its performance.

We try to overcome the limitations in the state of the art and explore the scope of the work in developing the thesis. The subsequent stages of this thesis will try to focus on these research questions in developing the methodology, generating the data set, and evaluating the model. Thereby, it enables to answer these questions at the end of this thesis.

Chapter 3

Preliminaries

This chapter will discuss the preliminary knowledge on NLP which will be helpful to develop the methodology of our proposed work. Section 3.1 will focus on the fundamentals of NLP research with the emphasis on defining the basic terminologies and primitives to sentiment analysis. Section 3.2 and Section 3.3 will provide a brief idea on NLP techniques and resources respectively, used throughout this thesis. Finally, some mathematical model used in the subsequent chapters to this thesis will be explained in Section 3.4.

3.1 NLP Fundamentals

This section will define various terminologies that are widely used in the field of NLP. Knowledge on these topics is primitive for understanding and developing our methodology. In the process of discussion, our aim is to disclose the abstractions prevailed in various terms like lexicon, concept, corpus, opinion, sentiment and semantic dependency.

3.1.1 Lexicon

A lexicon is the vocabulary or a coordinate term of the vocabulary of a language or subject. It can be referred to as a dictionary that includes or focuses on lexemes. Items in the lexicon are called lexemes, or lexical items. In the context of NLP, it can be defined as Definition 1.

Definition 1. *(Lexicon) A lexicon is the unit semantic constituent in the linguistic analysis which must contain meaning or expression in a language.*

It differs from the token in the sense that the lexicon can be one or a set of tokens. Whereas, the tokens are the unit granularity in the textual representation which may not express any meaning. An ordinary dictionary is an example of a lexicon. As the outline or format of a dictionary is presented for human use, it is inappropriate for computational use like NLP. A notable difficulty is the explications of the meaning of each word are themselves enumerated in natural language. Lexicon resources are illustrated in a machine readable format that can be interpreted by the computational application, such as, WordNet. Instead of a definition, WordNet uses the synonymy to represent the relationships among the words or lexicons, as between shut and close or car and automobile. Synonyms are grouped into unordered sets called synsets.

3.1.2 Concept

According to the Oxford dictionary [46], “Concept is an idea or a principle that is connected with something abstract”. It can be referred to mental representations that are used to discriminate between objects, events, relationships, or other states of affairs. In NLP, a concept can be viewed as a single or group of words that incorporate the common sense knowledge with the dictionary meaning, applying commonsense reasoning to natural language processing and understanding, thereby, disclose the abstraction expressed within the text. In a work of cognitive science [47], Concept has been stated as the constituent of thought, and in principle, the thought is unbounded. It can be formally defined as Definition 2.

Definition 2. *(Concept) Concept is the linguistic constituent extracted from the sentence or documents that blend the common sense knowledge with dictionary meaning applying commonsense reasoning to infer the semantic and affective information associated with natural language.*

Concepts are learned inductively from the sparse and noisy data applying some common sense knowledge bases which are usually developed by exploring the semantic synset and their application in the different context of the language. For instant [36], from the sentence “I got fired today”, the computer reader would not know what to think. Someone can not infer any substantial meaning from the lexicons “got fired”. However, the common sense knowledge base can reason out something things about “getting fired”; someone gets fired for lack of skill. As a consequence of getting fired, someone will not be able to meet his daily expenses. Likewise, in the sentence “I had a long day”, the concept, “long day” does refer to the length of the day, rather; will infer the knowledge as a hectic day, which causes tiredness.

3.1.3 Corpus

In linguistics and NLP, corpus (literally Latin for body) refers to a large and structured set of texts that are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. A formal definition can be as Definition 3.

Definition 3. *(Corpus) A usually large collection of documents that can be used to infer and validate linguistic rules, as well as to do statistical analysis and hypothesis testing.*

Construction of a gold standard annotated corpus is a colossal task which may be referred to data acquisition phase of any NLP task. The text in a corpus may be annotated with various syntactic and semantic information for specific language which assists in performing the various NLP task. The performance of the NLP task like sentiment analysis greatly depends on the precision of the annotation. Several corpora are available for the most widely used languages. To be mentioned, Brown Corpus [1] contains a wide collection of texts of different genres including newspapers, fiction, scientific text, legal text, and others. Lancaster Oslo Bergen (LOB) corpus is the British interpretation of Brown corpus. As an example, a fragment of an English Brown Corpus annotated with POS tag is highlighted below:

```
Daniel/np personally/rb led/vbd the/at fight/nn for/in the/at measure/nn
,/, which/wdt he/pps had/hvd watered/vbn down/rp considerably/rb since/in
its/pp$ rejection/nn by/in two/cd previous/jj Legislatures/nns-tl ,/,
in/in a/at public/jj hearing/nn before/in the/at House/nn-tl Committee/nn
on/in-tl Revenue/nn-tl and/cc-tl Taxation/nn-tl ./.
```

3.1.4 Semantic Dependency

In the book [48], linguistics dependency is categorized into three major types: syntactic, semantic and morphological dependency. Syntactic dependency understandably depends on sentence structure and language grammar. Few languages like English, French, Chinese, etc. have already achieved some satisfactory level of efficiency in developing syntactic dependency grammar. While working in sentiment analysis, we are mostly interested in semantic dependency. Semantic dependency can be effectively inferred through the conception of 'predicate' and 'argument' of the sentence. Reasonably, the argument is semantically dependent on the predicate where predicate can be recognized through the language syntax. In some instances, the semantic dependency can overlap the syntactic dependency. However, it can be sometimes opposite or even entirely independent of each other. One of the simplest definition of semantic dependency, found in the literature [48] is stated as Definition 4:

Definition 4. *(Semantic Dependency) The word form w_2 is said to be 'Semantically Dependent' on the word form w_1 in the given utterance if and only if the meaning of w_1 is or (includes) a predicate and the meaning of w_2 is an argument to this predicate in this utterance (defined in the book [48]).*

For example, in the sentence "I saw[w_1] allen's[w_2] wife[w_3] going[w_4] to market[w_5]", w_1 does not carry semantic dependency to w_2, though, it has syntactical dependency shown as $w_1 \rightarrow w_3$. However, the dependency $w_3 \rightarrow w_2$ is overlapping between syntactic and semantic. In the

[1] A tagged corpus of about a million words put together at Brown University during the 1960s and 1970s.

context of sentiment analysis, concept can be extracted exploring these dependencies, such as, the dependency $w_3 \rightarrow w_2$ can form a concept 'allen-wife'. Likewise, $w_4 \rightarrow w_5$ can form a concept 'go-market'[c_2]. In the same context, dependencies among the concept can also be generated as 'go-market'[c_2] semantically depends on 'allen-wife'[c_1] not on w_1 and represented as $c_2 \rightarrow c_1$.

One of the popular semantic dependency Extraction tool for English is Stanford dependency developed by "Stanford Natural Language Processing Group" [2]. Stanford dependency finds the dependency from the sentence discussed in the previous example as followings:

```
[nsubj](saw-2, I-1)
[nmod:poss](wife-5, allen-3)
[dobj](saw-2, wife-5)
[acl](wife-5, going-6)
[nmod](going-6, market-8)
```

Here, dependencies are extracted through textual relation depending on the grammatical rules.

3.1.5 Opinion

Opinion is one's views or judgment on a specific aspect of a thing or object. It is mostly used against review of product or things. In NLP, opinion has structured representation with its several components extracted from unstructured text. Definition 5 presents the formal definition of opinion as stated in [8].

Definition 5. *(Opinion) An opinion (or regular opinion) is a quintuple, (e_i; a_ij; oo_ijkl; h_k; t_l), where e_i is the name of an entity, a_ij is an aspect of e_i, oo_ijkl is the orientation of the opinion about aspect a_ij of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k. The opinion orientation oo_ijkl can be positive, negative or neutral, or be expressed with different strength/intensity levels. When an opinion is on the entity itself as a whole, we use the special aspect GENERAL to denote it[Stated at [8]].*

In NLP, an opinion generally consists of five components mentioned in Definition 5. However, it can vary depending on the application. For example, sometimes, it is not necessary to know the opinion holder if the task is to summarize opinion from different people. On the contrary, sometimes, we want to know more information like the sex of the opinion holder. One of the most effective contributions of the definition is its capability to present a structured text from the unstructured one. Thereby, the NLP task like sentiment analysis, opinion summarization become more simple and efficient.

[2]The Natural Language Processing Group at Stanford University is a team of faculty, postdocs, programmers and students who work together on algorithms that allow computers to process and understand human languages.

3.1.6 Sentiment

Not surprisingly, there is some confusion among researchers and professionals in differentiating between opinion and sentiment. In the sense of dictionary meaning, these two are almost complementary to each other and sometimes used as a synonym. However, distinct differences can be found in its applications in NLP, where, Sentiment referred to attitude, thoughts or mood of a text and opinion is defined as view or judgment on a specific aspect of a matter.

In the context of NLP, opinion mining extracts and analyze people's opinion about an entity through converting the unstructured text into structured one while sentiment analysis search for the sentiment words/expression in a text and then analyze it. Sentiment analysis most precisely depends on the identification of sentiment or opinion bearing words and their affective relation to finding the mood of the text. Then, various classification methods are used to reach a decision about the valence of the mood of the text. Sentiment analysis can be carried out in document, sentence or even phrase level.

3.2 NLP Techniques

NLP task requires a series of subtasks to perform, most of which are common to all. Various tools have been developed using different techniques for performing NLP task in different stages of language evaluation. This section will discuss different integrals tools for NLP and their techniques with a view to finding the suitability to use in our proposed methodology for sentiment analysis.

3.2.1 Tokenization

The first step to any NLP task is the tokenization of the sentences. Tokenization is a fundamental technique to split a sentence or document into tokens. Tokens are the smallest constituent of the text of a language, which, can be used later on to form the semantic constituents of the language such as, keywords, phrases, and concepts. It can be formally defined as definition 6.

Definition 6. *(Tokenization) Tokenization is the process of splitting up the text into units called tokens. The tokens may be words or number or punctuation mark. Tokenization does this task by locating word boundaries. The terminating point of a word and the beginning of the next word is called word boundaries. Tokenization is also known as word segmentation (collected from [49]).*

The process of tokenization relies on language structure and conventions. For example, tokens of language such as English, Bengali are separated by space delimiter. Tokens are not necessarily always to be meaningful words, as the punctuation marks, articles, and prepositions, such as 'the', 'a', 'and' etc. These tokens are sometimes, called stop words and removed from the corpus, as they do not generally contribute greatly to the semantic evaluation of the content. As a fundamental technique, many tokenization tools are available, such as Stanford Tokenizer, OpenNLP, Natural Language Toolkit (NLTK) tokenizer etc.

3.2.2 POS Tagging

POS tagging is the technique that analyzes the syntactic and semantic information depending on language grammar. According to Wikipedia, "In corpus linguistics, POS tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph". POS tagging is also a sequential labeling problem. It plays an important role to distinguish opinion or sentiment words within a sentence as opinion words are usually adjectives, adverbs, and sometimes nouns or combination of nouns.

Though the POS tagging technique mostly relies on the semantic association among the words within the sentence. It can be important for semantic categorization of the tokens as well. For an instant, the word 'can' may take several meanings based on its relative use in the sentence. It may be used as an auxiliary verb to form a question, another form, may be a container for holding food or liquid. Another variation is being a verb denoting the ability to do something. Distinguishing such a specific meaning facilitates to explore the sentiment of the word in the context of the text more efficiently. Most of the POS taggers follow the rule based methods and mostly depends on a dictionary to determine the tag to be used. Besides, statistical methods like hidden Markov model, CRF model are also contributing a lot to improve the performance of the taggers. Several POS taggers are available, especially, for language like English, French, Chinese. NLTK provides a very handful POS tagger. For example, an online demonstration of NLTK POS tagger output the following version of a plain text after performing POS tagging:

```
Important/JJ gains/NNS have/VBP been/VBN made/VBN in/IN the/DT past/JJ
two/CD decades/NNS in/IN the/DT participation/NN of/IN women/NNS in/IN
science/NN ,/, engineering/NN ,/, and/CC biomedical/JJ disciplines/NNS
at/IN the/DT undergraduate/NN and/CC graduate/VB levels/NNS in/IN
the/DT United/NNP States/NNPS
```

Very few works found in POS tagging for Bengali languages. Though some POS taggers are available, their performance is also not very encouraging [50].

3.2.3 Parsing

While POS tagging provides lexical information, parsing obtains syntactic information. Parsing represents the grammatical structure of a given sentence with the corresponding relationship of

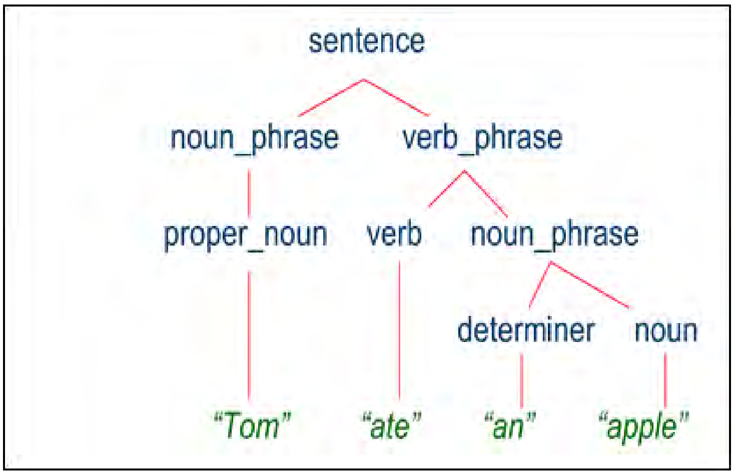

Figure 3.1: Generation of a Parse Tree using POS information.

different constituents. Comparing to POS tagging, parsing provides richer structural information [51]. The outcome of the parsing would be a parse tree. As a simple example, the sentence "Tom ate an apple" is the root, intermediate nodes such as noun phrase and verb phrase are the child nodes of the root, hence, called non-terminals. Finally, the leaves of the tree 'Tom', 'ate', 'an', 'apple' are called terminals. The tree is shown in (Figure 3.1):

Apart from its capabilities of syntactic and grammatical annotation, the parser also provides dependency among the terms of the sentence. In the context of sentiment analysis, the parser plays a very influential role to form the concepts from the corpus, hence, extract the dependencies among them as discussed in Section 3.1.4. One of the enhanced version of it is semantic dependency parser. To be mentioned, the Stanford dependency parser is an example for one of this kind which produces a dependency tree from the annotated corpus. For instance, the Stanford dependency [6] of the sentence, "Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas" are as following (Figure 3.2):

3.3 NLP Resources

3.3.1 Concept Net

ConceptNet was first introduced in [36], motivated with the significance of large-scale commonsense knowledge bases to textual information management. It is a freely available commonsense knowledge base and natural-language-processing tool-kit which supports many practical textual-reasoning tasks over real-world documents including topic-gisting, analogy-making, and other context oriented inferences. The knowledge base is optimized for making practical context-based inferences over real-world text. It is a graphical representation of a semantic network, where the nodes represent the concepts with their attributes as edges associated with each other. Presently, it consists of over 1.6 million edges connecting more than

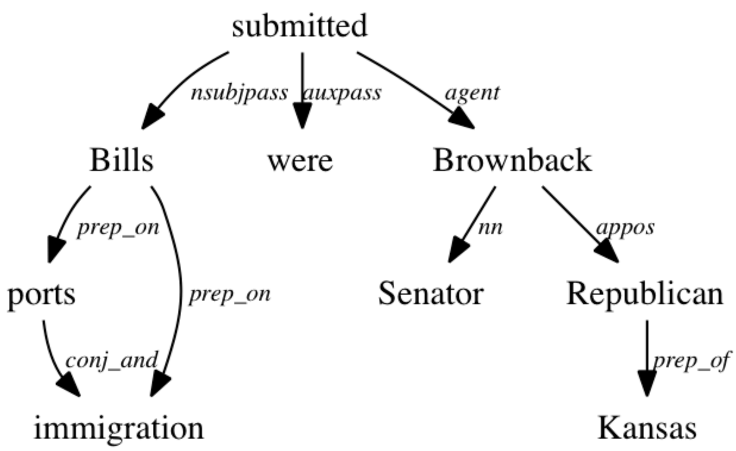

Figure 3.2: Stanford Dependency Tree [6].

300 000 nodes. A partial snapshot of actual knowledge in ConceptNet is given in Figure 2.2 of Chapter 2.

In the field of NLP, the conceptNet has opened up new diversity in sentiment analysis. Instead of depending on the key phrases and the dictionary biased knowledge bases, ConceptNet utilizes the commonsense knowledge base to extract the concept which carries related information on the context of the text to be analyzed. Thereby, it excels the contextual commonsense reasoning over real-world texts. The concept level sentiment analysis also offers more efficiency and accuracy in the classification of sentiment.

3.3.2 WordNet-Affect

WordNet-Affect, first proposed in [7], is a linguistic resource for the lexical representation of affective knowledge, developed starting from WordNet [52]. The knowledge base is built by assigning a number of WordNet synsets to one or more affective labels (a-labels). In particular, the affective concepts representing emotional states, are identified by synsets marked with the a-label 'emotion', but there are also other a-labels for concepts representing moods, situations eliciting emotions or emotional responses. Figure 3.3 shows an example of a list of a-level emotion and their corresponding synset.

The purpose of the WordNet-Affect was to incorporate the lexical information with the affective information. Lexical information includes the correlation between English and Italian terms, POS, definitions, synonyms, and antonyms. Affective information is a reference to one or more of the three main kinds of theories on emotion representation: discrete theories (based on the

A-Labels	Examples
EMOTION	noun anger#1, verb fear#1
MOOD	noun animosisy#1, adjective amiable#1
TRAIT	noun aggressiveness#1, adjective competitive#1
COGNITIVE STATE	noun confusion#2, adjective dazed#2
PHYSICAL STATE	noun illness#1, adjective all_in#1
EDONIC SIGNAL	noun hurt#3, noun suffering#4
EMOTION-ELICITING SITUATION	noun awkwardness#3, adjective out_of_danger#1
EMOTIONAL RESPONSE	noun cold_sweat#1, verb tremble#2
BEHAVIOUR	noun offense#1, adjective inhibited#1
ATTITUDE	noun intolerance#1, noun defensive#1
SENSATION	noun coldness#1, verb feel#3

Figure 3.3: A-Labels and corresponding example synsets (collected from [7])

concept of cognitive evaluation), basic emotion theories and dimensional theories.

To build the WordNet-Affect, an initial set of affective words directly or indirectly referring to mental (e.g. emotional) states was build. By mapping the senses of affective words to their respective synsets, the "affective core" was identified. Then, a subset of WORDNET containing all synsets in which there are at least one word of the affective word list, and rejected those synsets that are not recognized as affective concepts were prepared. Finally, an automatic check for the coherence of the affective information inside the synsets was performed. The results have shown that the synsets are a good model for the representation of affective concepts.

3.3.3 AffectiveSpace

The latest form of the knowledge base in the field of emotion reasoning is the AffectiveSpace. In [3], a blending technique was applied on ConceptNet and WordNet-Affect to build a suitable knowledge base for emotive reasoning. Blending performs inference over both sources simultaneously to combine two sparse matrices linearly into a single matrix, so that, it can all be represented in the same matrix. Singular value decomposition (SVD) is applied on the blend of ConceptNet and WordNet-Affect to form the affective space (illustrated in Figure 3.4), in which common sense and affective knowledge are in fact combined, not just concomitant.

Though affective space is mainly generated for sentic computing [4], it is a very effective resource for opinion mining and sentiment analysis. Information sharing properties of truncated SVD, suggest that concepts with the same affective valence are likely to have similar features, i.e. concepts concerning the same emotion tend to fall near each other in AffectiveSpace. For instances, 'love' and 'affection' tend to be nearer to each other, in contrast, 'honesty' and 'corruption' will fall far from each other. Therefore, affective space will enhance the performance of sentiment classification. An enhanced version is proposed as AffectiveSpace-2 [53]. It represents a novel vector space model for concept-level sentiment analysis that allows reasoning by analogy on natural language concepts, even when these are represented by highly dimensional semantic features.

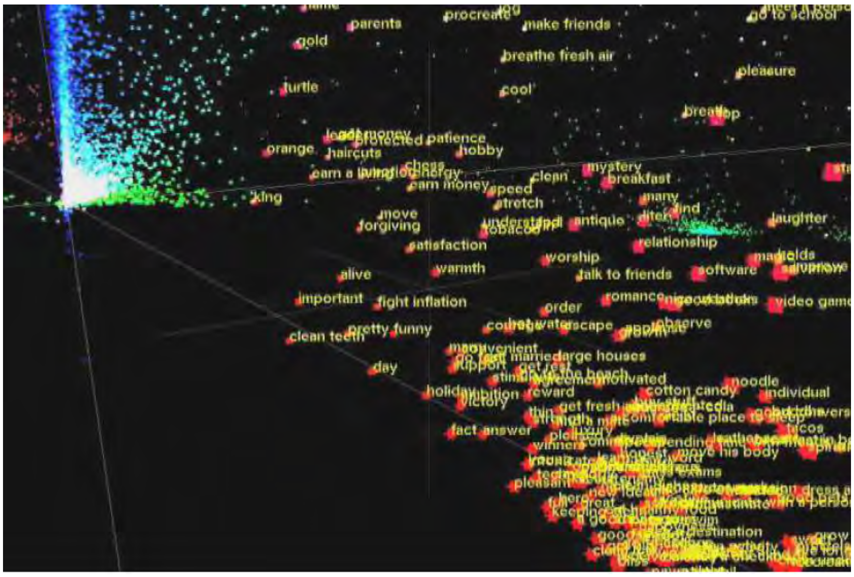

Figure 3.4: A Sketch of the AffectiveSpace (collected from [3]).

3.4 Mathematical Models

There are several mathematical models used in different machine learning or even rule based methods for NLP tasks like sentiment classification, subjectivity prediction, sentic computing etc. This section will highlight the basics of the model singular value decomposition (SVD) and linear discriminant analysis as our proposed model will use these methods for the classification of concepts.

3.4.1 Singular Value Decomposition (SVD)

Singular value decomposition takes a rectangular matrix of gene expression data (defined as A, where A is a $n \times p$ matrix) in which the n rows represent the genes, and the p columns represent the experimental conditions. The SVD theorem states as Equation (3.1):

$$A_{n\times p} = U_{n\times n} S_{n\times p} V^T_{p\times p} \tag{3.1}$$

Where:

- U and V are orthogonal
- the columns of U are the left singular vectors and $U^T U = I_{n\times n}$,
- V^T has rows that are the right singular vectors and $V^T V = I_{p\times p}$.
- S (the same dimensions as A) has singular values and is diagonal and

The SVD represents an expansion of the original data in a coordinate system where the covariance matrix is diagonal and vice versa. The eigenvalues and eigenvectors of AA^T and A^TA make up the columns of V and U respectively. These values need to be Calculated to perform the SVD. Also, the singular values in S are square roots of eigenvalues from AA^T or A^TA and always real numbers. Equation (3.1) can be expressed using summation notation as Equation (3.2).

$$a_{ij} = \sum_{k=1}^{1} u_{ik} s_k v_{jk} \tag{3.2}$$

One of the popular application of SVD is dimension reduction. The singular values are the diagonal entries of the S matrix and are arranged in descending order, where the larger values correspond to the vectors in U and V that are more significant components of the initial A matrix. A tranculated SVD can be created with a reduced number of singular value p following Equation (3.3):

$$a_{ij} \approx \sum_{k=1}^{1} u_{ik} s_k v_{jk} \tag{3.3}$$

where $p < n$ is the number of singular values that will remain.

Equation (3.3) can be used for data compression by storing the truncated forms of U, S, and V in place of A and for variable reduction by replacing A with U. The principal components of A form a low-rank approximation of the original data by often discarding all but the first k components. This factorization allows the row space of A and the column space of A to be projected into a common space by the transformations U and V. We can think of these spaces as containing two types of objects, which we can represent as row and column vectors of A, which are related to each other by the values where they meet.

3.4.2 Linear Discriminant Analysis (LDA)

LDA is a probabilistic classification method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The particularity of LDA is its capability to model the distribution of predictors separately in each of the response classes, and then, use Bayes' theorem to estimate the probability.

For instant[3], an observation needs to be classified into one of K classes, where K $\geq$ 2 and Π_k be the overall probability that an observation is associated with the k^{th} class. Then, let $f_k(x)$ denote the density function of X for an observation that comes from the k^{th} class. That is, $f_k(x)$ is large

[3] Mathematical equations and their explanation used in this section are mostly followed to the platform "Towards Data Science" sharing concepts, ideas and code (https://towardsdatascience.com/).

if the probability that an observation from the k^{th} class has $X = x$. Then, Bayes' theorem states the Equation (3.4) to predict the class ($Y = k$) for the observation ($X = x$).

$$Pr(Y = k \mid X = x) = \frac{\Pi_k f_k x}{\sum_{l=1}^{k} \Pi_l f_l(x)} \tag{3.4}$$

The challenge here is to estimate the density function $f_k(x)$. It consists of statistical properties of data, calculated for each class. For a single predictor, the density function can be assumed as a normal distribution (Equation (3.5)):

$$f_k(x) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp^{(-\frac{1}{2\sigma_k^2}(x-\mu)^2)} \tag{3.5}$$

Now, to maximize $P_k(x)$ for the observation $X = x$, it will plugin the density function in $P_k(x)$ and take the log to generate Equation (3.6):

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\Pi_k) \tag{3.6}$$

The Equation (3.6) above is called the discriminant function for class k given input x. As it is linear, hence, the name linear discriminant analysis. The boundary equations of the two classes with equal distributions are represented in Figure 3.5. In reality, LDA cannot exactly calculate the boundary line rather, makes use of the following approximation:

- For the average of all training observations, called as mean, μ_k

$$\mu_k = \frac{1}{n_k} \sum_{i:y=k} x_i \tag{3.7}$$

- For the weighted average of sample variances for each class, called as σ^2

$$\sigma^2 = \frac{1}{n-k} \sum_{k=1}^{k} \sum_{i:y=k} (x - \mu_k)^2 \tag{3.8}$$

For multiple predictors, the same properties calculated over the multivariate Gaussian distribution with a class-specific mean vector, and a common covariance matrix. Figure 3.6 shows an example of a correlated and uncorrelated Gaussian distribution.

To accommodate multiple predictors, the discriminant equation remains the same but it is expressed using vector notation as Equation (3.9):

$$\delta_k(x) = x^T \sum^{-1} \mu_k - \frac{1}{2}\mu_k^T \sum^{-1} \mu_k + \log(\Pi_k) \tag{3.9}$$

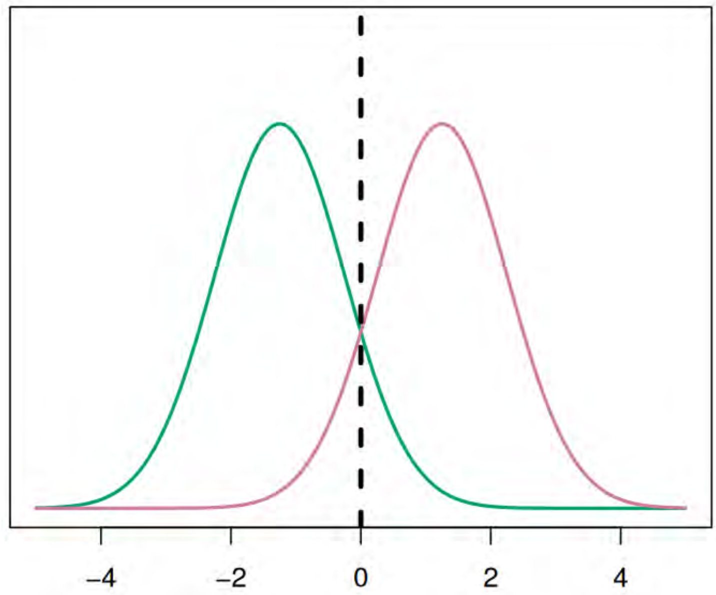

Figure 3.5: Boundary line to separate 2 classes using LDA.

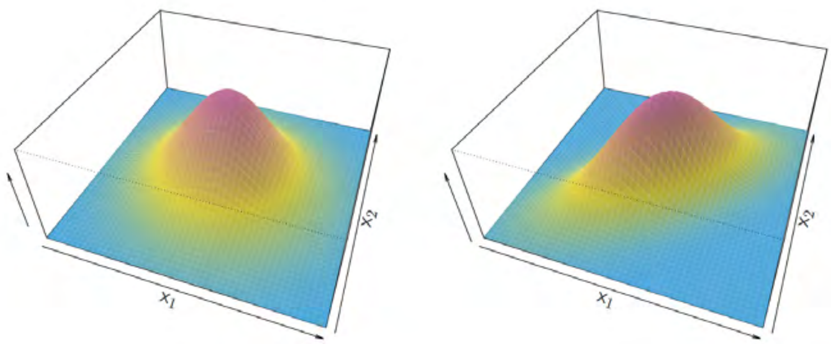

Figure 3.6: Uncorrelated (left) and correlated (right) normal distribution.

In sentiment classification, LDA can be utilized as an effective classifier, especially, in feature based text classification. The features of the concept can be taken as independent variables to categorize them in different classes like positive, negative or neutral, which can be considered as dependent variables. LDA is also a very effective tool for dimension reduction. In the process of generating classification function coefficients, it automatically discards the features which have no effective contribution to determine the classes of the dependent variables. Thereby, LDA offers a composed classification function with fewer parameters.

Chapter 4

Proposed Methodology

In this chapter, we present our proposed methodology. At the very outset, we will present an overview of the methodology in Section 4.1. Then, the construction of our methodology will be discussed in four steps in four subsequent sections. To step forward to our methodology, firstly, we will discuss the process of generating annotated data-set in Section 4.2. Thereafter, a rule based semantic parser will be presented in Section 4.3. In Section 4.4, we will discuss the technique of concept extraction from the parse tree and, determine the dependencies of the modifiers. Finally, we will present the concept level polarity detection model and explain the process of determining the polarity of the sentence in Section 4.5.

4.1 Overview of the Methodology

Proposed methodology presents a polarity detection model to detect polarity at the concept level and eventually, determines the polarity of the sentence in a specific domain or context in Bengali. This model will use the AffectiveSpace (discussed in Chapter 3) as a knowledge base. On the other hand, concepts on the desired domain are selected from AffectiveSpace and grouped as positive, negative and neutral to be used as training data for the model. Corpus is prepared with annotated text which is used as input to the model. The output of the model are the concepts with their polarity. Workflow diagram of the proposed methodology is presented in Figure 4.1 which highlights the overview of methodology.

To start with, a rule based semantic parser takes the annotated data as input which produces a parse tree for each sentence. The parse tree also contains the dependencies among the child nodes of a single root node. The outcome of the semantic parser is fed to the concept extractor which exploits the semantic dependencies among the tokens based on POS and extracts the concepts from the parse tree. Thereby, we get all possible concepts from out corpus which are constituents containing some opinion words for the sentiment of the sentence. On the other end, the training data along with the AffectiveSpace are fed to the model, in which the concepts in training

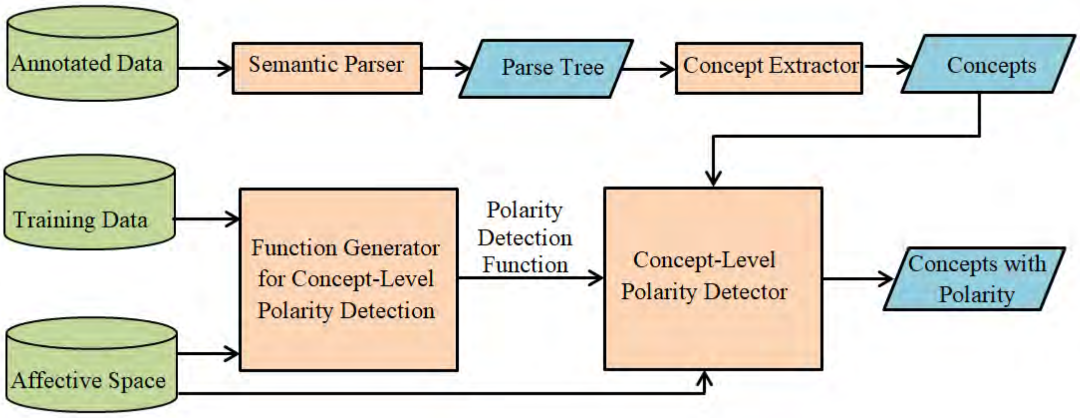

Figure 4.1: Workflow diagram of the proposed methodology.

data are mapped in AffectiveSpace to know a set of features known as ‘eigenmood’. LDA, the mathematical classifier used by the model generates the classification function coefficient for each class corresponding to positive, negative and neutral. This step is termed as function generator for concept level polarity detection which generates the polarity detection function using the classification function coefficient values produced by LDA for each class. Finally, concepts extracted from the corpus are mapped to affective space to get the corresponding egenmood values for the polarity detection function. These values in turn, fed to the function for each class to produce the classification value. Comparing the classification values of the classes of a concept, the polarity for the concept is determined. This step is termed as concept level polarity detector which produces the concepts with their polarity. In the end, the sentence polarity is determined through a back propagation technique on the parse tree. To gain better insight, the proposed methodology is described broadly in four steps in four subsequent sections.

4.2 Data Acquisition

Unlike any other research work, our foremost initiative was to generate a data set. As it was mentioned in Chapter 2 about the scarcity of annotated data set, especially, for the less privileged language like Bengali, we develop an annotated data set as a part of our research work. Firstly, we build a collection of sentences in natural Bengali language crawled from news editorials and blog. Then, we apply NLTK POS tagger for Bengali on each sentence. The POS tagging process involves assigning the right POS marker to all the words in a sentence (or corpus). According to Bengali grammar, it has five kinds of POS with several sub-kinds of each. Though there are some deviations in the POS category of Bengali from the English, most of the tagger available for Bengali uses the English categorization as a universal POS tags. To achieve better syntactic representation, particularly for sentiment analysis, we have also used the universal POS tag set generally used for English. For instant, the sentence “ইভ টিজিং বা উত্ত্যক্ততার শিকার হলে কারও সংকোচ করা উচিত নয় ” ([ibʱtidʒiŋer ba uttktɔtar ʃikar ɦɔle karɔ ʃŋkotʃ kra utʃit nɔi], If someone is victim of eve teasing or harassing, she should not be ashamed of that) annotated with POS using the Bengali NLTK POS tagger is as follows:

[ইভ/JJ টিজিং/NN বা/CC উত্ত্যক্ততার/NN শিকার/NN হলে/VM কারও/CC সংকোচ/NN করা/VM উচিত/JJ নয়/VM]

To achieve gold standard annotated corpus, improvement of accuracy of the tagger is essential, particularly when the performance of the tagger is not very encouraging. To be mentioned here, "কারও /CC" ([karo], someone) is tagged as conjunction based on detecting the 'ও ' ([o], and) as conjunction. Whereas, it is clearly a pronoun. Likewise, The NLTK POS tagger has no 'VAUX' (auxiliary verb) tag for Bengali, rather it considers the words 'হলে [ɦɔle]/VM' and 'করা [kra]/VM' as main verb (VM), which are actually contemporary to auxiliary verb used in English. Although Bengali has several forms of the verb, we categorize them in two types, such as main verb (VM) and auxiliary verb (VAUX), to facilitate the parsing. Therefore, we re-tag the above mentioned tokens as 'হলে [ɦɔle]/VAUX' and করা [kra]/VAUX. In the same context, the POS tag used for 'শিকার /NN' ([ʃikar], hunt) and 'সংকোচ /NN' ([ʃŋkotʃ], ashamed) are reformed as 'শিকার /VM' [ʃikar] and 'সংকোচ /VM' [ʃŋkotʃ]. On the other hand, the word tag 'নয় /VM' ([nɔi], not) is wrongly tagged as 'VM' which should be a negation(NEG) annotation. Thereby, we get a more comprehensive annotation of the aforementioned sentence as under:

[ইভ/JJ টিজিং/NN বা/CC উত্ত্যক্ততার/NN শিকার/VM হলে/VAUX কারও/PRP সংকোচ/VM করা/VAUX উচিত/RB নয়/NEG]

To limit the scope of our work, we improve the accuracy of the POS tagging with manual annotation for the false tagging as some of them mentioned above. Doing so, we have introduced the tag set for our work as listed in Table 4.1. For the simplicity; quantifiers, intensifiers, and interjections are categorized as modifiers and further subcategorized as pre-modifier and post-modifier depending on their position related to modified words as they play an important role in the performance of dependency parser. As negation has a very effective contribution in determining the sentiment polarity, therefore, we use the tag 'NEG' for all negation words in the corpus. In addition, we recognize the named entity (NE) early on and tag them as NE in the corpus, with the intention to make our corpus application specific to our methodology. As the NE has no significant contribution in determining the sentiment polarity, therefore, tagging them early on will reduce the computational complexity.

Table 4.1: Tag list used in annotated data with examples.

Category	Tag	Example
Noun	NN	প্রতিক্রিয়া ([prətikriɛ], reaction), উন্নয়ন ([unnɔn], development)
Pronoun	PRP	আমাদের ([amader], us), তাদের ([tader], their)
Main Verb	VM	বলা ([bɔla], say), শোনা ([ʃona], listen)
Auxiliary Verb	AUX	হয় ([ɦɔi̯], is), হবে ([ɦɔbe], will)
Adjective	JJ	ভালো ([bʱalɔ], good), সুন্দর ([ʃundɔr], beautiful)
Adverb	RB	দ্রুত ([drutɔ], quickly), সুচারুভাবে ([ʃutʃarubʱabe], skillfully)
Conjunction	CC	এবং ([ebɔŋ], and), অথবা ([ɔtʰɔba], or)
Preposition	PSP	জন্য ([dʒɔnnɔ], for), সাথে ([ʃatʰe], with)
Pre Modifier	PREMOD	অতি ([ɔti], too), খুব ([kʰub], very)
Post Modifier	POSMOD	বটেই ([btei̯], indeed), তর ([tɔrɔ], more)
Named Entity	NE	বেগম রোকেয়া [begɔm rokea], বাংলাদেশ ([baŋladeʃ]
Negation	NEG	না ([na], no), নাই ([nai̯], not)

Besides generating the corpus to be used in the proposed model, we use the 'AffectiveSpace' as a resource to incorporate the semantic affectivity among the concepts. To generate the training data, we extract a set of high frequency concepts on the targeted domain from the news portals, blogs, and literature available and label as positive, negative and neutral. While preparing training data, our focus was on those words of the targeted domain which may change the sentiment polarity in other domain.

4.3 Parse Tree Generation

In the proposed work, a semantic parse tree is developed depending on the syntactic affinity among the POS of the token. The parser takes the annotated corpus as input and generates a parse tree for each sentence as its output. The aim of the parse tree generator is to deconstruct the text into tokens exploring the syntactic dependencies. The parse tree will be later on fed to the concept extractor to construct the concepts. Doing so, we have formulated the following rules depending on the syntactic association through the POS tagged with each token:

[RULE- 1] *Parse the complex or compound sentence into simple sentences:* If more than one verb clause parse is available then each verb clause with its preceding words forms a simple sentence.

[RULE- 2] *Parse the simple sentence into subject and predicate:*

(a) Add token to subject until first NN/PRP occurs.

(b) Check the next token: If NN/NPN/ (MOD/NEG)/CC add to subject and repeat the process, else Parse as subject.

(c) Parse the rest of the sentence except subject as object.

[RULE- 3] *Parse the Subject into tokens:*

(a) If the subject is a single word parse into a single token.

(b) If any MOD/NEG is present parse associated token with MOD/NEG, else parse into a single token.

(c) For each token with MOD/NEG parse into toke and MOD/NEG marked as token MOD/NEG :

[RULE- 4] *Parse the Object into tokens:*

(a) Parse object into the verb clause and others along with associated MOD/NEG

(b) For parsing the verb clause and others follow the parsing rule for subject

The first step to parse the sentence is to decompose the complex and compound sentence into simple sentences. Bengali simple sentence usually contains a single verb clause and ends with a finite closing auxiliary verb. There is hardly any verb represented with a single token. Almost, in all the cases, the verb clause ends with a finite closing auxiliary verb, such as, থাকা [tʰaka], করা [kra], হওয়া [ɦɔa], পারা [para] which can be used within or at the end of a sentence. For example, in the sentence "মেয়েটি ইভটিজিং এর শিকার হওয়াই থানায় অভিযোগ করে " ([meeti ibʰtidʒiŋ er ʃikar ɦɔai tʰane ɔbʰidʒog kre], The girl complain to police for being victim of eve-teasing), the auxiliary verb 'হওয়াই ' [ɦɔai] is used inside the sentence, whereas, 'করে ' [kre] is used at end of the complex sentence. Table 4.2 shows some examples of verb phrase with a finite closing auxiliary verb with their different forms based on the position in the sentence. The parser detects the verb clause in the complex and compound sentence and applies the rule-1 to deconstruct simple sentences as a child of the complex or compound sentence.

Table 4.2: Various forms of auxiliary verb depending on the position within the sentence.

Verb Phrase	Auxiliary Verb Inside the Sentence	Auxiliary Verb at the end of the Sentence
দাঁড়িয়ে থাকা ([dar̃ie tʱaka], stand)	থাকালে, থাকায় [tʱakle, tʱakaî̯]	থাকি, থাক, থাকে [tʱaki, tʱak, tʱake]
অভিযোগ করা ([obʱidʒog kra], complain)	করলে, করায় [kɔrle, kraî̯]	করি, কর, করে [kɔri, ʁorɔ, kɔre]
শিকার হওয়া ([ʃikar ɦoa], victim)	হলে, হওয়াই [ɦole, ɦɔai]	হই, হও, হয় [ɦɔî̯, ɦɔô, ɦoî̯]
বলতে পারা (bɔlte para], say)	পারলে, পারাই [parle, paraî̯]	পারি, পার, পারে [pari, parɔ, pare]

The next step is to deconstruct the simple sentence into subject and predicate. As subject must contain one or more noun or pronoun, therefore, it detects the first noun or pronoun in the sentence to form the subject and concatenate the succeeding noun, pronoun, conjunction and if any modifier or negation token is present. As these succeeding elements form the noun clause, we consider those to be part of the subject. In the same context, the preseeding tokens of the subject are actually dependent on the subject and included in the subject as per the rule-2. After detecting the subject rest of the sentence are considered to be the object. Rule-3 is generated to deconstruct the subject into dependent tokens. Here, all the tokens will be independent except any negation or modifier is present. In case of negations and modifiers, the token will be parsed along with its modifiers. It is important to explore the dependency of the modifiers or negation as it will reverse the polarity of the sentiment of the associated concept. On the other hand, if the negation is associated with the verb clause then, it will negate the overall sentiment of the sentence. Otherwise, if it is associated with an adjective or any other opinion word, it will only reverse the polarity of the concern concept. Therefore, we generate the rule-4 which parse the object into verb clause and other first, then parse the subsequent phases.

Figure 4.2 shows a parse tree generated following the rule we discussed in this section for a complex sentence. However, the negations in the parse tree are associated with the concern tokens only. In the next section, while generating the concept we will see how negation or modifiers are associated with the concepts or sentences.

4.4 Concept Extraction and Dependency Detection

In this section, we proposed a technique for concept extraction and in the process, we will detect the dependency as well. A concept can be a single word or a chank of words depending on the semantic association of the terms in the sentence. As we discussed in Chapter 3,

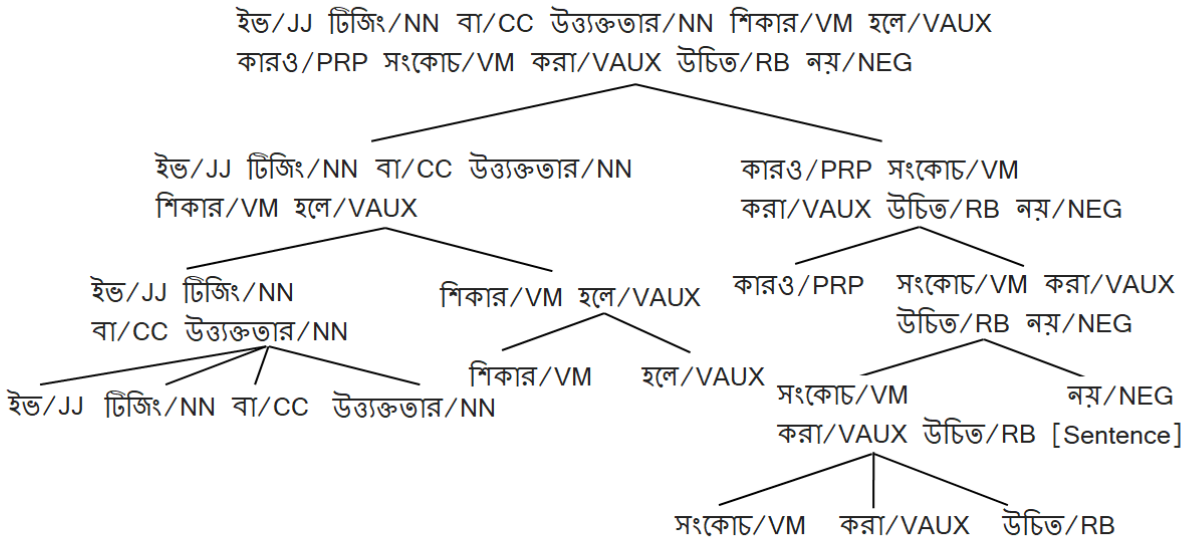

Figure 4.2: Rule based parse tree of a complex sentence generated from annotated data.

the concept can be formed combining two or more words which add some abstraction over the dictionary meaning of the constituents. Combining the words sometimes brings massive semantic metamorphosis, especially, for a language like Bengali with semantic diversity. For instant, "শিকার করা " ([ʃɪkar kra], hunt) may be considered as a neutral sentiment. On the other hand, if we use 'হওয়া ' [ɦɔa] instead of 'করা ' [kra], then "শিকার হওয়া " ([ʃɪkar ɦɔa], victim) is used to express sentiment like "someone is victim of something", which exhibits a negative sentiment. Therefore, the right choice of constituents of the text is very important in concept extraction particularly, in sentiment analysis.

POS of the tokens have a semantic association among them depending on their syntactic relation within the sentence. For an instant, when a noun is preceded by an adjective, the combination is generally a single entity. For Example, 'কর্মজীবী ' ([krmɔdʒibi], working) is an adjective and 'নারী ' ([nari], woman) is a noun. When the noun is preceded by the adjective, it became a single concept like "কর্মজীবী নারী " ([krmɔdʒibi nari], working woman). In this context, the concept does not simply indicates the meaning as "the woman who works" rather it represents the knowledge as "the woman who earns wages through regular employment outside the home" and generally exhibits a positive sentiment. On the contrary, "noun + adjective" combination will not be considered as a concept as there will be hardly any dependency. In the same context, consecutive or dependent nouns can form a single concept. Such as, 'নারী ' ([nari], woman) and 'নিরাপত্তা ' ([nirapɔtta], safety) form a concept "নারী নিরাপত্তা " [nari nirapɔtta] which means "woman safety" that indicates some specific measures for the safety of woman. If there is any stem in the preceding noun, like, "নারীর নিরাপত্তা " [narir nirapɔtta] means the safety of a woman. Since the combination does neither enhance the original meaning nor indicate any other knowledge, therefore, it will not be considered as a concept. However, some of the tokens with POS like a noun, adjective, and VM remaining stand lone can also be considered as a concept.

The rules based on the POS combination that are taken into consideration to form a concept are listed below:

1. NOUN + NOUN: If two nouns are consecutive and the preceding one has no steaming then it indicates a single entity and added to concept list.
2. ADJECTIVE + NOUN: If the adjective is preceding the noun then the combination is added to concept list as the adjective is obviously dependent on the noun.
3. MAIN VERB + AUXILIARY VERB: In Bengali auxiliary verb can influence the semantic orientation of the main verb. Hence, they are considered a concept as a verb phrase.
4. ADVERB + VERB PHRASE: As the adverb will modify the action of verb therefore, the combination is added to the concept.
5. NOUN, ADJECTIVE: Besides the combinations above a stand-alone noun and adjective can have an impact on the sentiment of the sentence, hence, taken as concept.

After generating the parse tree, we traverse the parse tree from leaf nodes towards the root and extract concepts applying the rules listed above. While identifying the concept we try to extend the combination of tokens to the highest level. However, there is no use of going beyond two levels up from the child node in traversing the parse tree as we will not find any POS combination at the sentence level. Concept extracted from the parse tree in the Figure 4.2 are shown in Figure 4.3. The concepts are highlighted in bold and shown their association within the sentence.

One of the most important tasks in this stage is to identify the dependencies. For the simplicity, we are mostly concentrated on the negation. If any negation is available with any token then it will be considered to be associated with the concept that contains the token. However, if the negation is associated with a verb phrase then it will be associated with the concern simple sentence. Because, the verb contains the momentum of the sentence and the negation of the verb indicates the negation of the sentence polarity. The dependency identification rule can be represented as follows:

[RULE- 1] If the root of the negation is a verb parse, level the negation as “Sentence NEG”

[RULE- 2] If the root of the negation is a concept without a verb phrase , level the negation as “Concept NEG”

At this juncture, we discuss the process of the concept extraction and dependency detection from the parse tree in Figure 4.2. The concept extracted are as follows:

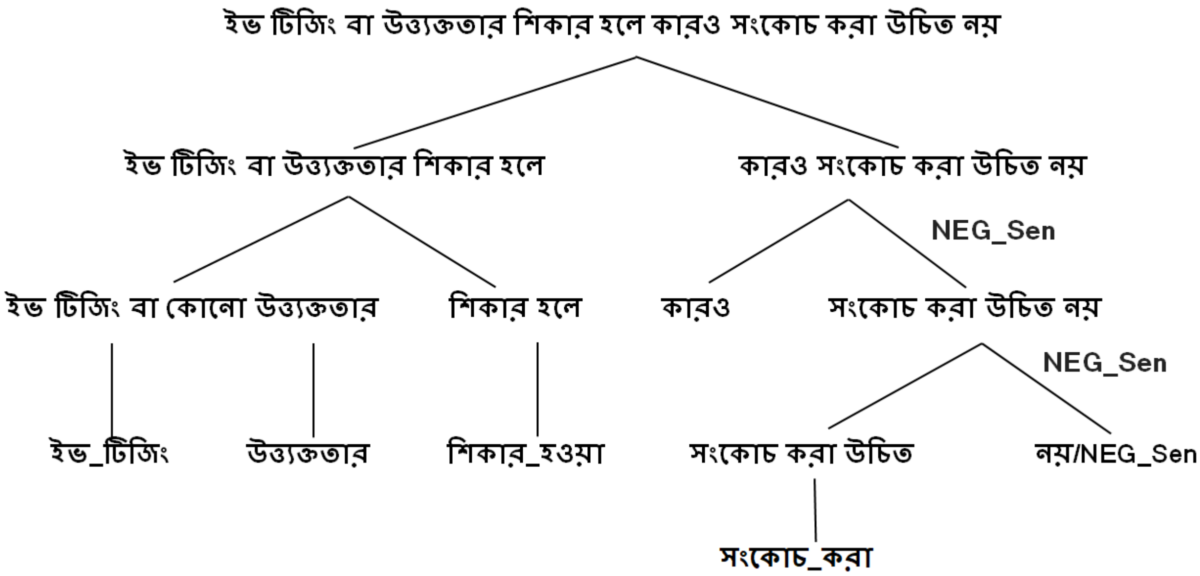

Figure 4.3: Concepts extracted from parse tree and their association within the sentence.

(“ইভ _টিজিং ” [ib^{ɦ}tidʒiŋ], eve-teasing) (“উত্যক্ততা ” [uttktɔta], harassment) (“শিকার _হওয়া ” [ʃikar hɔa], to be victim of) (“সংকোচ _করা ” [ʃŋkotʃ kra], to be ashamed of)

The first concept was taken following the rule “ADJECTIVE+NOUN” where two different tokens with independent meaning form a concept that represents a specific meaning as “harassing the girls or woman”. Likewise “শিকার _হওয়া ” [ʃikar hɔa] and “সংকোচ _করা ” [ʃŋkotʃ kora] have been formed following the rule “MAIN VERB + AUXILIARY VERB”. Both the concept represent some action like “to be victim of” and “to be ashamed of” respectively and carry negative sentiment, whereas, ‘শিকার ’ ([ʃikar], hunt) and ‘সংকোচ ’ ([ʃŋkotʃ], shame) stand-alone may contain neutral sentiment. On the other hand, ‘উত্যক্ততা ’ [uttktɔta] is a stand alone concept which carries the default meaning with its negative sentiment. To identify the dependency of the negation, we can see that the root of the negation in the Figure 4.2 verb phrase. Therefore, we consider it to be negation for the simple sentence “কারও সংকোচ করা উচিত নয় ” ([karɔ ʃŋkotʃ kra utʃit nɔi], someone should not be ashamed of) and alter the valence of the sentence while determining the sentence polarity.

4.5 Polarity Detection

In this section, we will present our proposed “concept level polarity detection model”. In the process, we will discuss the construction method and integration of affective space followed by the working procedure of the model in detecting the polarity of the concept. At later part, we will discuss the procedure to detect the sentence polarity using the parse tree in the example from the previous section.

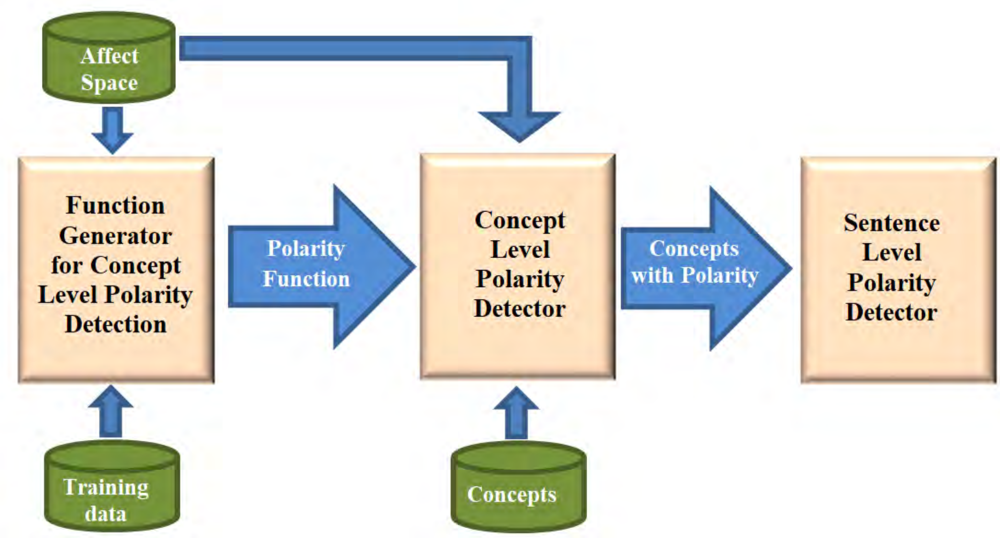

Figure 4.4: Overview of the concept based polarity detection model.

4.5.1 Construction of the Polarity Detection Model

The proposed model for detection of the polarity of a concept is inspired through the affective categorization model widely known as Hourglass of Emotion [37]. AffectiveSpace is used as a knowledge base for the proposed model. The affective relation among the concepts in AffectiveSpace is the inspiration for the proposed methodology. We use the LDA as a mathematical model to train the model with training data. On the other hand, the concepts extracted from the corpus are considered to be input to the model and output the concept with its sentiment polarity. Figure 4.4 represent the overview of the proposed model which can be described in two phases.

To start with, it takes the training data that includes the list of positive, negative and neutral concepts on the targeted domain. The concepts are then mapped on AffectiveSpace to find the appropriate matching within the affective space. In affective space, each concept contains 100 'eigenmood' values leveled as $(e_0, e_2, e_3, ..., e_{99})$. After the mapping, we get a training data set as a matrix in which, each row represents the concept with the eigenvalues as a column. The values in the last column are class level. Table 4.3 shows the data representation used by the model. The last column is the value for the class in the training data where positive, negative and neutral class are denoted by 1, 2 and 3 respectively.

Table 4.3: Concepts with their eigenvalues and class labels to train the model.

Concept	e_0	e_1	.....	e_{99}	Class
Social_development	0.9351	-0.0514	.....	0.0143	1
Sexual_victim	0.7651	-0.0904	.....	0.0214	2
Working_girl	0.8331	-0.0674	.....	0.0035	3
.....	.....	.....	.....	.....	.....

As we discussed in Chapter 3, the AffectiveSpace is constructed through the blending of ConceptNet and WordNetAffect to combine the commonsense and affective knowledge. Since the blending process explores the information sharing properties of truncated SVD to build the affective space, the 'eigenmood' values of the concept nearer to each other within the space tend to be similar. To categorize the sentiment, our aim is to discriminate among the positive, negative and neutral concepts based on their eigenvalues. In this context, LDA is an appropriate choice for its property of maximization of class separability. As we mentioned in preliminaries, LDA has the capability of modeling the distribution of predictors separately in each of the response classes, thereby, formulate a class specific mean vector and covariance matrix which is used to determine the significance of the features of the independent variable. In the proposed method, three sentiment classes are considered as dependent variables and the 'eigenmood' are taken as the independent variable for LDA. The model explores the dimension reduction properties to discard the less significant 'eigenmood' that contribute very less in discriminant function. At the same time, as the eigenvalues are sorted from strong to weak, it increases the efficiency of the dimension reduction. LDA generates two discriminant function which divides the space into three groups according to the three classes. Finally, the classification function coefficient for each class is generated by LDA as shown in Table 4.4.

Table 4.4: Classification function coefficient matrix.

Eigenmoods (E)	Classification Function Coefficient		
	Class-1	Class-2	Class-3
e_i	a_{01}	a_{02}	a_{03}
$--$	$--$	$--$	$--$
e_i	a_{i1}	a_{i2}	a_{i3}
Constant	$Constant_1$	$Constant_2$	$Constant_3$

From the classification function coefficient generated by the classifier, we formulate the classification function for each of positive, negative and neutral as Equation (4.1), Equation (4.2), Equation (4.3). These functions are fed to the polarity detector to determine the concept polarity (Figure 4.4).

$$\text{Score}_{positive}(\text{C}) = \sum_{i \in E} e_i a_{i1} + Constant_1 \tag{4.1}$$

$$\text{Score}_{Negative}(\text{C}) = \sum_{i \in E} e_i a_{i2} + Constant_2 \tag{4.2}$$

$$\text{Score}_{Neutral}(\text{C}) = \sum_{i \in E} e_i a_{i3} + Constant_3 \tag{4.3}$$

Where:

- C is the concept with its eigenvalue
- e_i is the i^{th} eigenvalue of the concept C
- a_{i1}, a_{i2}, a_{i3} are the i^{th} coefficient values for class-1 (positive), class-2 (negative) and class-3 (neutral) respectively
- $Constant_1, Constant_1, Constant_1$ are the constant value for class-1 (positive), class-2 (negative) and class-3 (neutral) respectively
- E is the set of significant eigenmood in for classification function generated

The next step is to determine the polarity of the concept that is extracted from the corpus. The polarity detector receives a concept as input. The concept is mapped in the AffectiveSpace to find the best match. The rows of the AffectiveSpace contain the eigenvalues for each concept. Eigenvalues of the concept are used as parameters to Equation (4.1), Equation (4.2), Equation (4.3). The output of the equations are positive, negative and neutral score for the concept. Finally, the polarity of the concept is determined as the polarity class with the maximum score.

4.5.2 Determination of the Polarity of the Sentence

The polarity of the sentence is strongly depends on the association of the concepts within the sentence. However, the dependency of the negation and modifiers can highly influence the sentence polarity. At the same time, the syntactic relation and POS information are also plays an important role in determining the sentiment polarity of the sentence. Therefore, we use the parse tree with concepts in Figure 4.3.

The technique for determining the polarity of the sentence can be described in the Figure 4.5. First of all, we label all the concepts with their polarity determined by the concept level polarity detector. Then, we traverse the tree from the concept node to the root following the rules described below.

1. If any node contains one concept the node will be labeled the polarity of the concept.
2. If any node contains two concepts then the polarity of the node will be labeled as follows:
   (a) If both are positive then the node is Positive.
   (b) If they are of opposite polarity then the node is negative.
   (c) If both are negative and the concepts are associated with conjunction then the node will be labeled as negative, else it will be positive.

3. If any node contains more the two concepts then the polarity label of the node will be determined pairwise following the rule 2.

4. If any sentence negation is available it will reverse the polarity of the simple sentence at the child node of the root. The token negation will reverse the polarity label of the immediate concept that contains the token.

5. From child node to parent node the label is determined as follows:

   (a) POSITIVE + POSITIVE = POSITIVE

   (b) POSITIVE + NEGATIVE = NEGATIVE

   (c) NEGATIVE + NEGATIVE = NEGATIVE

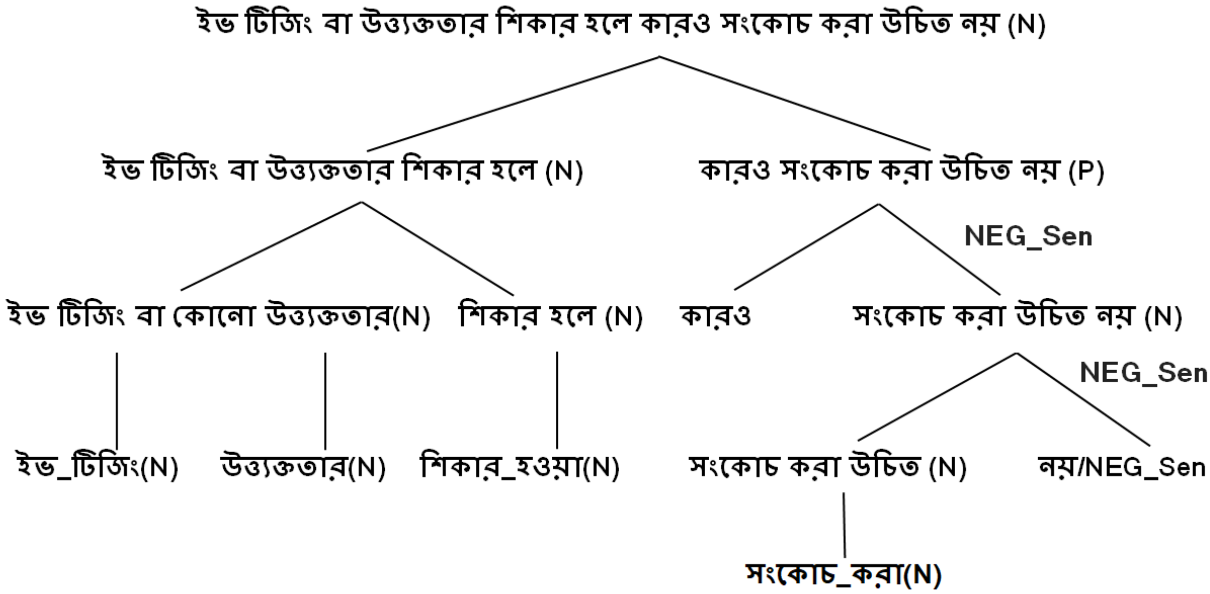

Figure 4.5: Sentence level polarity detection through tree traversal.

In Figure 4.5, both the concepts “ইভ _টিজিং ” ([ib^{ɦ}tidʒiŋ], eve-teasing) and “উত্যক্ততা ” ([uttktɔta], harassment) are labeled as negative (N). AS they are associated through conjunction their parent node will be labeled as (N) as well. On the other, stand-alone concept will label their parent node same as their label. As the negation is a sentence negation, therefore it has reversed the polarity of the simple sentence “কারও সংকোচ করা উচিত নয় ” ([karɔ ʃŋkotʃ kra utʃit nɔi], someone should not be ashamed) from negative to positive. Finally, the polarity of the sentence is labeled as (N) as the child node are opposite to each other.

In summarizing the methodology, the data set was built by semi-automated annotation by POS tagging followed by using the NE recognizer. The annotated data is used as input to the parse tree generator to build the parse tree. The concepts are extracted from the parse tree by the concept extractor. Then, the concepts are mapped to the AffectiveSpace to be used as input to the classification method named as concept level polarity detection model. Mapping of the concept is carried out through translation and matching of words that form the concept. LDA is used as classification and dimension reduction method to the model. The model generates a polarity detection function which takes the ‘eigenmood’ values of each concept to output the score of each polarity class. Finally, the sentence polarity is determined following a rule based tree traversal.

Chapter 5

Experimental Analysis

In this chapter, we will discuss the experimental analysis process and evaluate the performance of our proposed model through result analysis using our data-set. Initially, we will discuss the experimental data set in Section 5.2 followed by presenting an outline of the experimental setup in Section 5.1. Section 5.3 will give an overview on evaluation method and performance metrics used for evaluating the result and for analyzing the data-set. In Section 5.4, we will evaluate the performance of the model and represent some statistical data on experimental analysis. Finally, Section 5.5 will recapitulate the experimental analysis through an inquisitive discussion on performance evaluation of the model.

5.1 Experimental Setup

In this work, we implement our experiment using Python as a programming platform and carry out the statistical computation in SPSS. Doing so, we use the Python-2.3.15 version in windows set up. To perform the NLP tasks, we use the NLTK[1] plugin for Python, a leading platform for building Python programs to work with human language data. We utilize the natural language text processing API such as tagging and chunk extraction, phrase extraction and NE recognition. To our knowledge, NLTK is the only platform which offers the POS tagger and some other limited NLP tools for Bengali. However, the performance of this requires improvements to be used in an organized framework.

For implementing the classification model, we use the IBM SPSS software[2] platform to take advantage of the readily executable mathematical model. This platform offers all of the widely used classifiers with their traditional variation as well. We also incorporate the Python Integration Package for IBM SPSS Statistics with the SPSS. It allows us to build python programs that control the flow of command syntax jobs, read and write data, and procreate

[1] https://www.nltk.org/

[2] https://www.ibm.com/analytics/spss-statistics-software

Table 5.1: Statistics of the training data set.

Type of Source	Number of document	Number of Positive Concept	Number of Negative Concept	Number of Neutral Concept	Total Number of Concept
Blog	5	35	44	39	118
News Portal	3	24	25	20	69
Total	8	59	69	59	187

custom procedures. This integration facilitates the handling of large data set and carrying out experimental analysis more efficiently.

5.2 Experimental Data Set

We build a training data-set containing three sets of the concept named as positive, negative and neutral. The domain dependency of the proposed model relies on the domain of the training data. In this work, our targeted domain is "woman harassment". To generate the data set, we crawl through the various Bengali blogs, online news portals especially the editorials on woman harassment, as these sources mostly provides some opinions.

Our aim is to determine the most frequently used and the most influencing concepts from the available resources in the targeted domain. We use an improved term frequency-inverse document frequency (tf-idf) scheme proposed in [54] as a weighting factor to the terms in the corpus which can form a concept. In the process, the most weighting terms are selected and mapped to the affective space to determine the concept within the space. Then, the concepts are grouped according to the valence manually. Table 5.1 shows the overall statistics of the training data used in this work.

In the statistics of the training data, we observe that the number of the negative concepts is more than the positive. In reality, we also find that the number of negative terms is more than that of positive in various lexicon resources like SentiWordNet. Moreover, the NLP researchers also reached in consensus that the number of negative sentiment words is more than the positive in any language. As mentioned in methodology (Chapter 4), we develop a new domain specific corpus to evaluate the proposed model. Three paragraphs are considered, two forms the most popular online Bengali newspaper "Prothom Alo" and another one from a Bengali news portal on law "LawyersClub Bangladesh.com". The statistics of the corpus is highlighted in Table 5.2.

As the method finally finds the polarity at sentence level using the polarity of the concepts determined through the concept level polarity detection model, therefore, variation in the paragraph does not affect the performance of the model. However, they influence the contextual sense in the sentence and concept level.

Table 5.2: Statistics of the corpus to evaluate the model.

Paragraph	Number of Sentence	Number of Simple Sentence	Number of concept
Paragraph 1	10	17	49
Paragraph 2	9	23	44
Paragraph 3	13	21	73
Total	32	61	166

Table 5.3: Confusion Matrix for precision and recall.

Predicted \ Actual	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

5.3 Evaluation Method

Since the proposed model classifies the concepts as positive, negative and neutral, therefore we want to know the answer to the following question for evaluating the method:

1. What proportion of concepts of specific polarity in the corpus are classified correctly.
2. What portion of the classified concepts of specific polarity are correctly classified.
3. Finally, what percentage of the total number of concepts in the corpus are classified to their actual classes.

Precision and recall [55] are the perfect metrics to get the answers to the above question for evaluating the proposed model. For binary classification, precision means the percentage of the result that is relevant. On the other hand, recall means the percentage of total relevant result that are predicted correctly in comparison to their actual class. These metrics are realized more specifically with the mathematical notation using the confusion matrix given at Table 5.3. The true positive or true negative denotes the instances where the positive or negative case is predicted correctly to their actual classes. Whereas, the false positive means that an instance other than the positive is predicted as positive and the false negative means that an instance other then the negative is predicted as negative comparing to their actual class.

From the Table 5.3, we can find the precision and recall value with the Equation (5.1) and Equation (5.2) respectively.

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Result}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{5.1}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Result}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{5.2}$$

Accuracy is another metric for evaluating a classification model. It provides an idea on the overall performance of the model by detecting the percentage of the correctly identified instances with respect to the total number of instances in the corpus. It can be expressed in terms of the parameters of the confusion matrix as Equation (5.3) for a binary classification scheme. However, we also find the overall accuracy of the model by finding the total number of correctly classified concept comparing to the total number of concept in the test corpus. The overall accuracy of the model can be calculated by applying Equation (5.4).

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \tag{5.3}$$

$$\text{Accuracy of the model} = \frac{\text{Number of correctly classified concept}}{\text{Total Number of concept in the corpus}} \tag{5.4}$$

There is a trade-off between recall and precision based on a threshold value in maximization between the two. If we recall everything, we need to keep generating results which are not accurate, hence lowering the precision. A simpler metric known as F1 score, takes both precision and recall into account, and represents the balance between recall and precision. F1 Score is also viewed as the weighted average of precision and recall which is measured by the Equation (5.5). When one of the precision and recall is given emphasize over the other, the F1 score decreases. On the other hand, a higher value of F1 score expresses the desired harmonic balance between precision and recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{recall}}{\text{Precision} + \text{recall}} \tag{5.5}$$

Although the above metrics are suitable for binary classification, it is also used for multiclass classification. In that case, metric values for each class are determined considering a binary classification as one being the class itself and another is its complement. For instances, while building the confusion matrix for the neutral class, we consider neutral to be one of the classification states. Whereas, classification other the neutral is another state which can be compared to positive and negative for a binary classification scheme.

5.4 Result Analysis

In this section, we will evaluate the proposed method through result analysis based on the values of the evaluating metrics. To start with, we want to evaluate the performance of the parse tree and the concept extractor. As there is no traditional metric to evaluate the parse tree, we focused on evaluating the concept extractor. Then, we evaluate the performance of the classification model used for classifying the training data. Finally, we will analyze the accuracy of the polarity detection at concept level as well as sentence level.

Table 5.4: Performance of the concept extractor.

Number of Sentence	Number of Concept	Perfectly Extracted	Partially Extracted	Not Extracted
31	166	87	68	11

5.4.1 Performance Analysis on Concept Extraction

Parse tree generation is the precursory step for concept extraction. The performance of concept extractor depends on the proficiency of the parse tree, thereby, concept extractor exhibits the performance of the parse tree as well. the concept being a newly introduced entity in the field of NLP, very few works are available in the state of art. We hardly find any works at concept level in the Bengali language. Therefore, we evaluate the performance of the concept extractor by comparing the extracted concepts with the expected concepts within a sentence. In this work, about 53 percent of the concepts are extracted which match to our expected concepts and 41 percent of the concept was extracted partially.

Partial extraction is elucidated with an example such as, in the sentence “মেয়েটি ইভটিজিং বা উত্ত্যক্ততার শিকার হওয়াই থানায় অভিযোগ করে ” ([meeti ibʱtidʒiŋ ba uttktɔtar ʃikar ɦɔai tʱane ɔbʱidʒog kre], The girl complain to police for being victim of eve-teasing), both “ইভটিজিং এর শিকার ” ([ibʱtidʒiŋ er ʃikar], eve-teasing victim) and “উত্ত্যক্ততারর শিকার ” ([uttktɔtar ʃikar], harassment) are expected to be extracted as concept. Whereas, the first one was partially extracted as “ইভটিজিং ” ([ibʱtidʒiŋ], eve-teasing) even though the second one was accurately extracted as “উত্ত্যক্ততারর শিকার ” ([uttktɔtar ʃikar], harassment). However, 6 percent of the terms are not extracted, though they may form a potential concept. These terms need to form a concept along with some other terms which are already used in another concept. Therefore, the terms are ignored as the proposed concept extractor does not allow multiple uses of a term in a sentence. Table 5.4 represents the performance of the concept extractor.

Proposed method failed to extract very few concepts of which most of them are adverb not positioned adjacent to the verb. Some of the auxiliary verbs sometime express the sentiment, hence, expected to be extracted as a concept. However, this performance can not be compared with other existing concept extractors as the proposed extractor is domain and application specific. Therefore, it may not perform well for other applications and languages to extract concepts from the sentence.

5.4.2 Evaluation on Classification of Training Data

The accuracy of the proposed method depends highly on the accuracy of the classifier while performing supervised classification of the training data. We will try to evaluate the performance of the classification of training data set with the integral metrics that are provided by the LDA in SPSS. SPSS provides some of the statistical as well as graphical analysis on the performance of

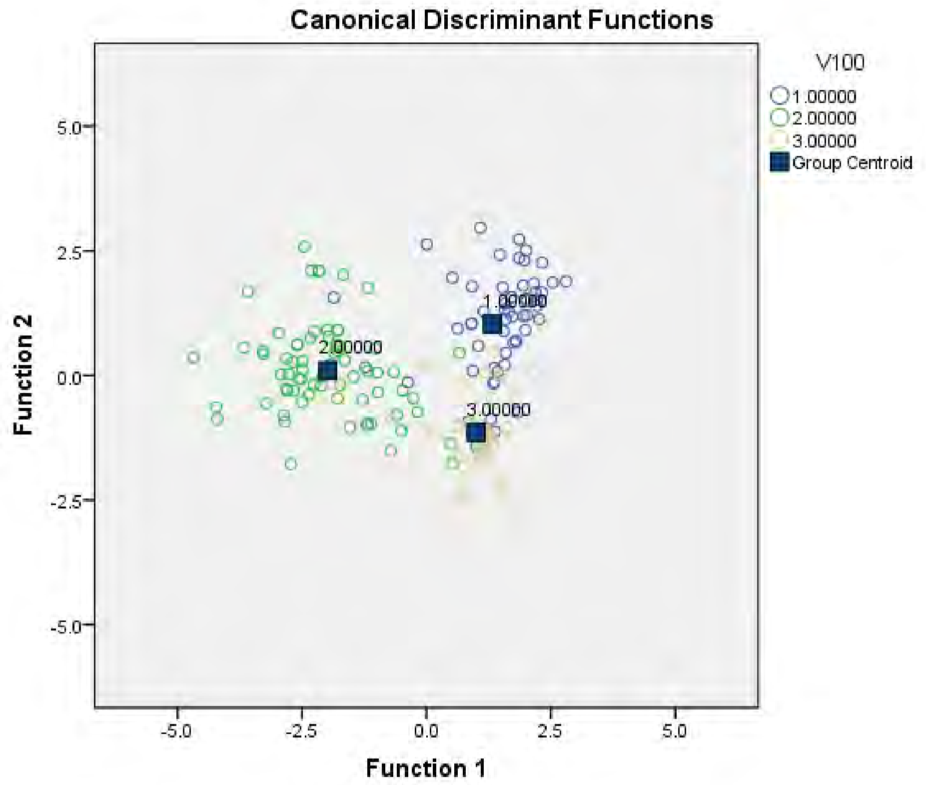

Figure 5.1: Positional overview of the data points based on the canonical discriminant function.

the classifier. Figure 5.1 shows the positional overview of the data point based on the canonical discriminant functions.

Based on the accuracy of the classification, SPSS provides a statistical summary of classification result in Figure 5.2. The summary result represents the performance of the classifier to classify the concepts correctly as per the classes assigned to the concepts. The overall accuracy of the classifier is 88.8 percent which means that only 11.2 percent of concepts have not been classified to their original classes. We can observe from the summary result that only one concept with negative class is classified as positive and five concepts of the positive class is classified as negative which are very negligible in the context of NLP.

A significant number of concepts from positive and negative class are classified as neutral and vice versa. In practical terms, some of the neutral concepts may shift their valence to positive or negative sentiment and contrariwise. As a matter of fact, there is a very low semantic gap between neutral and positive or neutral and negative. As a consequence, the positional distance between them in the affective space are also marginal and sometimes overlap each other. Therefore, most of the miss classification occurs between neutral and positive class or neutral and negative class in the affective space. Nevertheless, there is hardly any miss classification found between positive and negative. This is because the affective strength of the positive and negative concepts are usually very strong and position them opposite to each other in the

Classification Results[a]

		V100	Predicted Group Membership 1	2	3	Total
Original	Count	1	46	5	8	59
		2	1	61	7	69
		3	10	3	46	59
	%	1	78.0	8.5	13.6	100.0
		2	1.4	88.4	10.1	100.0
		3	16.9	5.1	78.0	100.0

a. 81.8% of original grouped cases correctly classified.

Figure 5.2: Summary result of classification of training data with three polarity classes.

Classification Results[a]

		V100	Predicted Group Membership 1	2	Total
Original	Count	1	54	5	59
		2	1	68	69
	%	1	91.5	8.5	100.0
		2	1.4	98.6	100.0

a. 95.3% of original grouped cases correctly classified.

Figure 5.3: Summary result of classification of training data with two polarity classes.

affective space.

This can be further analyzed through a classification scheme with a training data set with two polarity classes. Figure 5.2 shows the statistical summary of the classification result with two polarity classes. 95.3 percent of the concepts are correctly classified to their assigned class. The reason for the improvement in accuracy is already mentioned above. The neutral concepts are classified into positive and negative depending on their inclination towards positive and negative use in the proposed domain. Despite the better efficiency of binary classification, the exclusion of neutral class may affect the efficiency in determining the sentence valence as many neutral concepts will be considered as either positive or negative. Sometimes, it is cumbersome to deal with the concepts which are neither positive nor negative.

However, the binary classification is very effective to handle the concepts which shift their valence depending on the context of the sentence. Depending on the domain context, these concepts are classified into desired classes which also contribute to function coefficient matrix used to build the polarity detection function. Therefore, the binary classification scheme is sometimes preferable where the detection of sentence polarity is the primary task, even though, it is not effective for polarity detection at the concept level.

Classification Function Coefficients

	V100		
	1	2	3
V0	6.629	12.529	3.720
V1	9.068	8.187	4.700
V2	2.371	11.322	.865
V3	22.634	59.784	16.740
V4	.668	9.296	2.501
V5	-14.713	-16.101	-6.330
V6	7.852	-34.018	-13.047
V7	2.342	.441	-4.683
-	-	-	-
-	-	-	-
-	-	-	-
V70	-11.928	-10.158	-8.211
V81	-5.754	-2.293	-.924
V89	-1.671	-20.229	-11.460
(Constant)	-6.343	-9.098	-3.182

Figure 5.4: Fraction of classification function coefficient matrix.

5.4.3 Analysis of the Polarity Detection Model

At this stage, we focus on evaluating our proposed model for polarity detection at the concept level. As mentioned in methodology (Chapter 4), the performance of the classification function to determine the polarity of a concept depends on the classification function coefficient matrix generated by SPSS. Figure 5.4 shows a fraction of the classification function coefficient matrix generated based on our training data. It provides coefficient values for classification function for each of the three classes. It contains forty one eigenmoods along with values of the constant. As we discussed early that the strength of the eigenmood in the affective space decreases as the order increases. Therefore, the higher ordered eigenmoods have very minimal significance in determining the concept class. Therefore, LDA carries out tolerance test for each variable and ignores the eigenmoods failing the test to qualify for the coefficient matrix.

Performance of the proposed polarity detection model mainly relies on the efficiency of the polarity detection function. As the proposed model classifies the polarity in three classes as positive, negative and neutral, we will use the multiple class evaluation technique for the precision and recall method. Therefore, we generate the confusion matrix for each of the polarity classes of concepts to evaluate the performance of the polarity detector. Table 5.5, Table 5.6 and Table 5.7 show the confusion matrix for the positive, negative and neutral polarity classes of

Table 5.5: Confusion Matrix for the concept with positive polarity.

Predicted \ Actual	Positive	Not Positive
Positive	22	10
Not Positive	15	06

Table 5.6: Confusion Matrix for the concept with negative polarity.

Predicted \ Actual	Negative	Not Negative
Negative	35	15
Not Negative	10	19

Table 5.7: Confusion Matrix for the concept with neutral polarity.

Predicted \ Actual	Neutral	Not Neutral
Neutral	33	17
Not Neutral	15	60

concept respectively. Here, the polarity class of concepts, determine by the polarity detector is considered as predicted class, whereas, the actual class is determined by manual annotation. In the confusion matrix of each class, the class itself considered as positive, whereas, the class other than itself is considered as negative to conform with binary classification metric as shown in Table 5.3.

To evaluate the performance of the proposed model; precision, recall, F1 score and the accuracy for each of the polarity classes are calculated from the confusion matrix. We use the equations mentioned in Section 5.3 to calculate the values of the metrics for each polarity class as shown in Table 5.8. A system with high recall but low precision returns a high number of instances for desired class but many of the predictions are incorrect compared to their actual class. A system with high precision but low recall is just the opposite, returning very few results for the desired class, but most of which are predicted correctly as their actual class. An ideal system with high precision and high recall will maximize the instances of the desired class with a correct prediction to their actual class.

In observing the result of our polarity detection model, we find that the accuracy in classifying the positive class is lower than the other two classes. On the other hand, the negative class shows a very high recall value. This can be explained through semantic gaps among the concepts and the affective strength of the concepts in the affective space. As the affective strength of negative concepts usually very strong, therefore, they have very low semantic overlapping with other class and shows high recall value. On the contrary, the semantic gap between positive and negative class is very low and reduces the recall value for positive and neutral class. The spatial distribution of positive and neutral concepts in the affective space is also overlapping each other.

Table 5.8: Values of the metrics for the polarity classes of concept.

Polarity Class	Precision	Recall	F1 Score	Accuracy	Model Accuracy
Positive	68.75	59.46	61.97	52.83	70.24
Negative	70.00	77.77	73.68	68.35	
Neutral	66.00	68.75	67.35	74.40	

Table 5.9: Performance evaluation for polarity detection at sentence level.

Simple Sentence		Complex/ Compound Sentence	
Count	Accuracy	Count	Accuracy
61	73.77	32	65.63

This spatial overlapping influences to miss classify between positive and negative class which reduces the precision as well.

However, the precision values for all the classes are almost equal and are also well accepted in the context of polarity classification as an NLP task for a less privileged language like Bengali. The balance between recall and precision can be observed through the F1 score. The F1 score for the negative class is higher than the other classes. This is because of the high precision and recall and the maintenance of balance between the two measures. The F1 scores for other classes being greater than 0.50 are also in an acceptable range. However, the F1 score is not convenient for evaluating the proposed method, as it completely ignores the true negatives. The overall accuracy of the model shows that 70.40 percent of the total number of concepts are classified correctly. Acquirement this performance is considered promising, especially, for an ambiguous language like Bengali.

In order to evaluate the performance of polarity detection at the sentence level, we compare the polarity of a sentence determined by the proposed model with the actual polarity of the sentence. However, discovering of the actual polarity of a sentence is sometimes cumbersome, especially, for the complex and compound sentence where different parts or clauses of the sentence express different polarity. In case of the simple sentences, these ambiguities are very limited as they normally inherit a single context. Therefore, we evaluate the accuracy both for simple sentences as well as the original sentences as presented in the documents. Table 5.9 shows the performance of polarity detection at sentence level both for simple sentences and the complex or compound sentences.

We observe that the accuracy for polarity detection of the simple sentences is higher than the accuracy of complex or compound sentence. This is due to the complexity raised in the instances where different parts of the complex or compound sentence contain opposite polarity. In most cases, the model transmutes the polarity of a complex or compound sentence to negative if any part of the sentence contains a negative simple sentence. In the same context, we find that the number of false negative is much higher than the number of false positive out of the total miss classification.

5.5 Discussion

To our knowledge, there is hardly any work that finds the polarity at concept level in the Bengali language. Few works determine the polarity of the keywords or key phrases using various lexical resources which are mainly dictionary based approaches. In addition, the proposed model is domain dependent and language specific. Comparison of the model to the other methods may raise some miss leading result. Therefore, we evaluate our model by comparing the predicted result with the actual or desired result. The overall accuracy for polarity detection at concept level being above 70 percent is considered to be an acceptable range for a less privileged language like Bengali. However, some of the concept level polarity detection model in rich languages like English, French, Chinese shows accuracy up to 80 percent. This is possible due to the availability of the highly efficient NLP tools for those languages.

Though the accuracy for polarity detection of simple sentences is very high, the efficiency is reduced in case of the complex and compound sentences. Variations in polarity of the simple sentences as parts of the complex and compound sentence raise the ambiguities. An efficient dependency parser can solve the problem and improve the efficiency of the polarity detection at the sentence level. This requires a detail linguistic analysis incorporating the morphological and grammatical influences which are beyond the scope of this thesis. However, the performance of the parse tree in accordance with the concept extractor shows a satisfactory result especially in our application to determine the concept level polarity in the domain of woman harassment.

Chapter 6

Conclusion

This chapter will conclude the thesis with a brief summery of the works in Section 6.1 including the performance evaluation. Section 6.2 will highlight some challenges encountered in this thesis. Finally, scope of the future works will be enumerated in Section 6.3.

6.1 Contribution of the Work

With the increasing importance of autonomous information retrieval from online content, sentiment analysis becomes a popular area of research in the field of NLP. Bengali being one of the most significant languages in the world, has a scarcity of NLP tools and resource. Particularly, there is hardly any appropriate parser that explores the Bengali language structure comprehensively. This provides the motivation for developing an independent parser for Bengali sentences. The methodology follows a rule based approach which explores the language morphology and syntactical structure. This parser is capable of decomposing the Bengali sentence into clauses, key phrases and terms. The final outcome of the parser is a parse tree which also exhibits the dependencies among the constituents of the sentence.

In the latter part of this thesis, a concept level polarity detection model is introduced using a learning method which uses the AffectiveSpace as a knowledge base. The motivation behind using the AffectiveSpace is to infer the semantic and affective information associated with natural language opinions. Moreover, concept-based approaches are the recent evolution in sentiment analysis that step away from blind use of keywords and word co-occurrence counts to detect polarity. Since the concept usually shifts the valence depending on the contextual domain, therefore, the model is developed as domain specific to increase the classification accuracy while detecting the polarity.

The proposed model finds the polarity of the concepts that are extracted from the parse tree generated by the rule based semantic parser. Doing so, it determines a polarity detection function for each of the positive, negative and neutral classes through classification model. LDA is used as a mathematical model in which the training data are applied to generate the polarity detection function. Prior to the classification, the concepts are mapped to the AffectiveSpace to get the eigenmood values which are considered to be the classification features for each concept. The polarity detection function finds the positive, negative and neutral score for each concept applying the eigenvalues as the parameters to the functions. Comparing the scores of the polarity detection functions, the final polarity of a concept is determined. Finally, the polarity of the sentence is determined through the tree traversal in reverse order. In this process, the resultant polarity for a sentence highly depends on the mutual relationship among the concepts and dependencies of them on the modifiers.

Both the training data and the test data were developed using traditional NLP tools. However, the annotation of the data set requires some major improvement as it highly affects the subsequent stages of this work. The improvement was done through manual annotation. A comprehensive evaluation through result analysis based on precision and recall is carried out using the test data set. The evaluation shows a satisfactory performance of the work for polarity detection at concept level as well as sentence level.

6.2 Challenges

In the implementation phase of the thesis, one of the major challenges was to perform NLP task in Bengali. There is an acute scarcity of appropriate NLP tools with good performance. The complexity of Bengali language structure and grammatical directives is also required special attention. Since we use the AffectiveSpace that is built in English, therefore, translating the concepts from Bengali to English and mapping them in the space was sometime ambiguous. Some of the concepts may be translated to different English concepts which yield the confusion in selecting the appropriate one.

6.3 Scope of Future Work

From the challenges and the explanation of result analysis, we can highlight some of the scope of the future works to step forward from this thesis and improve the performance as well. The future works are listed as follows:

1. An independent dependency parser can be generated for Bengali, which is capable of identifying the dependency among the concepts as well as the modifiers. It requires an exhaustive study of Bengali grammar and linguistic structure to explore the relationship among the constituents of the sentence. The dependencies among the POS and their variation within the sentence based on stem categories need to address to increase the

efficiency of the parser. Integration of morphological knowledge can have a great impact on the performance of the dependency parser.

2. The concept extractor presented in this thesis extracts the concepts containing only the terms found within the sentence. To achieve a higher level abstraction while extracting the concepts from a sentence, the new word, terms or phrases may be of great utilization. Thereby, the performance of the concept extractor might be improved, hence, improves the accuracy of the polarity detection model.

3. Besides the difficulties in translating the concepts that are mentioned earlier, it sometimes drops out some linguistic information as well. This may have some significant impact in the context of sentiment analysis at the sentence level. Therefore, an affective space in Bengali can be generated for an exact representation of concepts extracted from sentences. This requires a huge effort in collecting the data set to build the knowledge base. However, a transformation method can be applied to mirror the English AffectiveSpace to Bengali where the approximation must be made with great care.

4. There is a lot of room for improvement in polarity detection at the sentence level. Once the polarity of the sentence elements is known, the sentence polarity can be determined more efficiently without using any NLP resources. Here, the approximation of resultant polarity of the clauses or phrases from the polarities of the child node is very crucial. Semantic affinity along with the syntactic dependencies can improve the performance of sentence level polarity detection.

The concept level sentiment analysis presented in this thesis is a new integration of this type to the field of NLP tasks in Bengali. Having a number of limitations, it requires a lot of improvement to achieve the desired abstraction level of a concept. Besides the sentiment analysis, the concept based analysis can be implemented in NLP tasks like opinion mining, document summarizing, subjectivity detection, etc.

References

[1] M. Z. Asghar, A. Khan, S. Ahmad, M. Qasim, and I. A. Khan, "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme," *PloS one*, vol. 12, no. 2, p. e0171649, 2017.

[2] S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 45–63, 2014.

[3] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, "AffectiveSpace: blending common sense and affective knowledge to perform emotive reasoning," *WOMSA at CAEPIA, Seville*, pp. 32–41, 2009.

[4] E. Cambria and A. Hussain, *Sentic computing: Techniques, tools, and applications*, vol. 2. Springer Science & Business Media, 2012.

[5] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60, 2014.

[6] M.-C. De Marneffe and C. D. Manning, "The stanford typed dependencies representation," in *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pp. 1–8, Association for Computational Linguistics, 2008.

[7] C. Strapparava *et al.*, "Wordnet affect: an affective extension of wordnet.," in *Lrec*, vol. 4, pp. 1083–1086, Citeseer, 2004.

[8] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[9] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in *Lrec*, pp. 2200–2204, 2010.

[10] G. F. Simons and C. D. Fennig, "Ethnologue: Language of the World." https://www.ethnologue.com/ethnoblog/gary-simons/welcome-21st-edition, Feb. 2018. (last accessed <Jan 19, 2019>).

[11] L. Mehedy, S. N. Arifin, and M. Kaykobad, “Bangla syntax analysis: A comprehensive approach,” in *Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh*, pp. 287–293, 2003.

[12] G. Qiu, B. Liu, J. Bu, and C. Chen, “Expanding domain sentiment lexicon through double propagation,” in *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[13] N. S. Dash and L. Ramamoorthy, “Corpus and technical termbank,” in *Utility and Application of Language Corpora*, pp. 173–191, Springer, 2019.

[14] S. Sun, C. Luo, and J. Chen, “A review of natural language processing techniques for opinion mining systems,” *Information Fusion*, vol. 36, pp. 10–25, 2017.

[15] B. Agarwal, S. Poria, N. Mittal, A. Gelbukh, and A. Hussain, “Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach,” *Cognitive Computation*, vol. 7, no. 4, pp. 487–499, 2015.

[16] T. Mullen and N. Collier, “Sentiment analysis using support vector machines with diverse information sources,” in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[17] C. Whitelaw, N. Garg, and S. Argamon, “Using appraisal groups for sentiment analysis,” in *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 625–631, ACM, 2005.

[18] V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives,” in *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pp. 174–181, Association for Computational Linguistics, 1997.

[19] A. Esuli and F. Sebastiani, “Determining the semantic orientation of terms through gloss classification,” in *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 617–624, ACM, 2005.

[20] C. H. E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.

[21] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “Affectnet: A database for facial expression, valence, and arousal computing in the wild,” *arXiv preprint arXiv:1708.03985*, 2017.

[22] H. Saggionα and A. Funk, "Interpreting SentiWordNet for opinion classification," in *Proceedings of the seventh conference on international language resources and evaluation LREC10*, 2010.

[23] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, ACM, 2004.

[24] A.-M. Popescu, B. Nguyen, and O. Etzioni, "Opine: Extracting product features and opinions from reviews," in *Proceedings of HLT/EMNLP on interactive demonstrations*, pp. 32–33, Association for Computational Linguistics, 2005.

[25] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[26] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417–424, Association for Computational Linguistics, 2002.

[27] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 international conference on web search and data mining*, pp. 231–240, ACM, 2008.

[28] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86, Association for Computational Linguistics, 2002.

[29] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of Sentimental Reviews Using Machine Learning Techniques," *Procedia Computer Science*, vol. 57, pp. 821–829, 2015.

[30] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and SVMperf," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1857–1863, 2015.

[31] J. Zhao, K. Liu, and G. Wang, "Adding redundant features for CRFs-based sentence sentiment classification," in *Proceedings of the conference on empirical methods in natural language processing*, pp. 117–126, Association for Computational Linguistics, 2008.

[32] E. Breck, Y. Choi, and C. Cardie, "Identifying Expressions of Opinion in Context," in *IJCAI*, vol. 7, pp. 2683–2688, 2007.

[33] B. G. Patra, S. Mandal, D. Das, and S. Bandyopadhyay, "Ju_cse: A conditional random field (crf) based approach to aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 370–374, 2014.

[34] Y. Xiang, H. He, and J. Zheng, "Aspect Term Extraction Based on MFE-CRF," *Information*, vol. 9, no. 8, p. 198, 2018.

[35] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decision Support Systems*, vol. 115, pp. 24–35, 2018.

[36] H. Liu and P. Singh, "ConceptNet—a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[37] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cognitive behavioural systems*, pp. 144–157, Springer, 2012.

[38] A. Björkelund, B. Bohnet, L. Hafdell, and P. Nugues, "A high-performance syntactic and semantic dependency parser," in *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pp. 33–36, Association for Computational Linguistics, 2010.

[39] J. Hajič, M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, and J. Štěpánek, "The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–18, Association for Computational Linguistics, 2009.

[40] T. Koo, X. Carreras, and M. Collins, "Simple semi-supervised dependency parsing," *Proceedings of ACL-08: HLT*, pp. 595–603, 2008.

[41] C. D. Manning, "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?," in *International conference on intelligent text processing and computational linguistics*, pp. 171–189, Springer, 2011.

[42] A. Das and S. Bandyopadhyay, "Sentiwordnet for Bangla," *Knowledge Sharing Event-4: Task*, vol. 2, pp. 1–8, 2010.

[43] D. Das and S. Bandyopadhyay, "Developing Bengali WordNet affect for analyzing emotion," in *International Conference on the Computer Processing of Oriental Languages*, pp. 35–40, 2010.

[44] A. Das and S. Bandyopadhyay, "Opinion-polarity identification in Bengali," in *International Conference on Computer Processing of Oriental Languages*, pp. 169–182, 2010.

[45] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Towards universal paraphrastic sentence embeddings," *arXiv preprint arXiv:1511.08198*, 2015.

[46] "Oxford Learner's Dictionaries." https://www.oxfordlearnersdictionaries.com.

[47] N. D. Goodman, J. B. Tenenbaum, J. Feldman, and T. L. Griffiths, "A rational analysis of rule-based concept learning," *Cognitive science*, vol. 32, no. 1, pp. 108–154, 2008.

[48] A. Polguère *et al.*, *Dependency in linguistic description*, vol. 111. John Benjamins Publishing, 2009.

[49] Robin, "Article on Natural Language Processing." http://language.worldofcomputing.net/category/tokenization, Nov. 2009. (last accessed <Jan 19, 2019>).

[50] A. Ekbal, R. Haque, and S. Bandyopadhyay, "Bengali part of speech tagging using conditional random field," in *Proceedings of the seventh International Symposium on Natural Language Processing, SNLP-2007*, 2007.

[51] T. Koo, X. Carreras, and M. Collins, "Simple Semi-supervised Dependency Parsing," in *Proceedings of ACL-08: HLT*, Association for Computational Linguistics, 2008.

[52] C. Fellbaum, "Wordnet," in *Theory and applications of ontology: computer applications*, pp. 231–243, Springer, 2010.

[53] E. Cambria, J. Fu, F. Bisio, and S. Poria, "Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis.," in *AAAI*, pp. 508–514, 2015.

[54] N. Wang, P. Wang, and B. Zhang, "An improved TF-IDF weights function based on information theory," in *2010 International Conference on Computer and Communication Technologies in Agriculture Engineering*, vol. 3, pp. 439–441, IEEE, 2010.

[55] K. M. Ting, "Precision and recall," *Encyclopedia of machine learning*, 2010.