CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

This chapter gives a sharp picture of the actual work done in this thesis. This work includes a brief description of every little contribution of this work such as lexicon, knowledge base, proposed CLSA algorithm, datasets, optimal preprocessing and feature or concept extraction techniques, different algorithm evaluation etc. This chapter also includes some recommendations for future researchers. Finally, figure out some limitations of this research.

6.1 Conclusions

This research disseminated significant knowledge towards understanding the problems related to NLP, such as SA, especially multilingual sentiment analysis and their probable solution. This understanding and solution will provide considerable support for the research in this field. Towards the solution to MLSA problems, the initiatives taken in this research will make the researches in this field easier than earlier. This research will facilitate the researchers of this field with some resources and optimal techniques that will help to conclude a similar study more efficiently and faster. In short, this research has practically presented lexical resources and techniques to the beneficiaries. At this conclusive stage, it could be claimed that, each planned objective has been achieved and the achievements along with the objectives are described below:

RO1- *To create a Bengali knowledge base and polarity lexicon about the English knowledge base and polarity lexicon SenticNet 5.*

RO2- *To propose an algorithm for concept-level multilingual sentiment analysis.*

The above objectives are achieved by creating two Bengali lexical resources such as knowledge base (BanglaSenticNet) with 30000 concepts and polarity lexicon with 72433 concepts concerning English lexical resources such as SenticNet 5 and polarity lexicon, respectively (details in Section 3.2, 3.3, and 5.1), and by proposing an algorithm for concept-based multilingual sentiment analysis (details in Section 3.4 and 5.1). These lexical resources can be thought as the most significant resource available for the Bengali language to the best knowledge. The study contemplates that it will provide a noteworthy contribution to NLP and a broad range of Bengali SA applications such as MLSA. Besides, the proposed algorithm will help the researchers analyze the text from different languages using concepts. Moreover, the algorithm has successfully evaluated the sentiment using both English and Bengali knowledge bases and polarity lexicons. This algorithm could be considered one of the significant contributions of this research as existing research lacks such custom algorithms. This research would like to suggest using MCSAlgo algorithm for concept-level multilingual sentiment analysis as this algorithm provides better classification accuracy over NB and SVM. Besides, this research recommends to use the created lexical resources as those found worthy in the scale of performance once compared with state-of-art research.

RO3- *To create well-form Bengali and English datasets on students' feedback.*

This research has created two English and one Bengali dataset with the data collected from the students' feedback in social media (details in Section 3.5 and 5.2). These datasets are tested on the knowledge bases, polarity lexicons, and proposed algorithm. This research got trustworthy results for the created datasets on the scale of

accuracy using NB, SVM, LSTM, and MCSAlgo. The validation results concerning other datasets (English (IMDB) and Bengali (ABSA-Cricket)) and state-of-art research shows, these datasets outperforms on the scale of accuracy. Therefore, it could be recommended that these datasets could be further investigated and help the related organizations when needed.

RO4- *To analyze and evaluate the best among different feature(s) and concept extraction techniques with their combinations for multilingual sentiment analysis.*

This thesis has evaluated best between feature and concept-based extraction techniques for multilingual sentiment analysis to help researchers analyze the text from different languages using both features and concepts (details in Sections 3.6 and 5.3). The extraction techniques are TD - IDF (Simple), unigram, bigram, trigram, parts of speech, and concepts. The experiment result prevails that applying the features and concepts in combinations enhances the performance largely. It is observed from the result that increasing the data sample size in experimentation increases the performance. In addition, the concept-based approach is found better than the feature-based approach for Bengali data sets; however, for large datasets, this gap became too minor. It was found in some cases, these extraction approaches are interchangeably better, and the study recommends adopting both approaches at a time for better performance.

RO5- *To examine the preprocessing techniques with different combinations for multilingual sentiment analysis and search optimal one.*

This thesis has tested the data sets for finding optimal preprocessing techniques or their combinations, intending to help the researchers save their preprocessing time (details in Sections 3.7 and 5.4). It is found that applying all the techniques in combination has produced better classification accuracy with NB, SVM,

LSTM and found LSTM as the best classifier. The testing results prevail that the performance increases once different preprocessing techniques were applied in combinations. For example, the result achieved using simply tokenization improves once other techniques such as stemming were added to it, and so on. This result was found to vary for some preprocessing combinations, such as punctuation removal and stop word deduction, and this happens as these techniques deduct some essential words that may express sentiments. Even then, the study advises using as many preprocessing combinations as possible because these combinations improve the performance in maximum cases.

Finally, after applying different polarity lexicons and knowledge bases, algorithms, feature and concept extraction techniques, and preprocessing techniques on above mention datasets reveal that the performance highly depends on the domain of study. Not all preprocessing techniques and their combinations; feature and concept extractions and their combinations, polarity lexicons, knowledge bases, and algorithms are suitable for every domain, and instead, if appropriate techniques and resources are not adopted, the performance will surely decrease.

6.2 Future Directions

The research in this field progresses exponentially; thereby, the experimental results from research to research are constantly changing. Likewise, the results and conclusions drawn in this thesis will also vary if the appropriate tools, techniques, and resources could be adopted. Some of the future extension of this research is listed below:

1. Future studies may enlarge the knowledge bases and polarity lexicons, especially for the Bangla language with more concepts. For instance, the newly created Bengali knowledge base and polarity lexicon have 180000 and 72433 concepts.
2. The above resources may be tested on new data sets or enriched the contents of existing data sets.
3. This research worked with one rare resource language. However, future research may work with other rare resource languages like Urdu and Hindi.
4. The proposed algorithm for concept-based MLSA could be made generalized for all types of data. For example, image, audio, and video data.
5. Several other machine learning algorithms such as multilayer perceptron, CNN, RF, Bi-LSTM could be adopted for the same type of analysis.
6. The same techniques, approaches, algorithms, etc., could be adopted for multimodal sentiment analysis and other subtypes of SA.
7. The preprocessing combination could be extended with a few more combinations. Here, seven preprocessing techniques are adopted. However, some other techniques, such as emoticon removal and unknown word removal techniques, could improve the performance.
8. The resources such as BanglaSenticNet and Bangla polarity lexicon could be applied to other branches of NLP, such as speech recognition and personality detection.

6.3 Limitations of Study

This study tried to fill the existing research gaps; however, single research is not possible to fill all the gaps. Therefore, this research only narrows down some gaps and poses the following limitations:

1. The Bengali knowledge base and polarity lexicon contain fewer concepts than the English knowledge base and polarity lexicon.
2. The knowledge sources are language-dependent; for instance, only Bengali lingual data is supported.
3. The experimentations adopted are domain dependant, not generous. For example, three domains are considered such as student feedback, cricket, and movie review.
4. Only one rare resource language, such as Bengali, is considered.
5. The proposed algorithm for concept-based MLSA has only considered text and ignored multimodal data.
6. Only a few machine learning algorithms such as NB, SVM, and LSTM have been adopted.
7. All the experimentations have considered only text and ignored multimodal data.
8. The resources such as BanglaSenticNet and Bangla polarity lexicon are only applied to SA and ignored other branches of NLP.

6.4 Summary

This chapter has given a brief description of the contribution of this research. The contribution of this thesis includes resource creation, proposing algorithms,

optimal techniques (preprocessing, feature and concept) determination. This research has created two Bengali lexical resources, such as a knowledge base (BanglaSenticNet) with 30000 concepts and a polarity lexicon with 72433 concepts, and an algorithm for concept-based multilingual sentiment analysis. To test the knowledge base, polarity lexicon, and algorithm, this research has created two English and one Bengali dataset with the data collected from the students' feedback in social media; Also, one English (IMDB) and one Bengali (ABSA-Cricket) baseline dataset is considered for validation.

This research has presented the best extraction techniques among TD - IDF (Simple), unigram, bigram, trigram, parts of speech, and concepts for multilingual sentiment analysis. This will help researchers analyze the text from different languages with the use of both features and concepts. Finally, the dataset was tested for finding optimal preprocessing techniques or their combinations, intending to help the researchers save their preprocessing time.

This chapter has given a brief idea of this research's future extension possibilities; some of them enrich datasets, lexicons, and knowledge bases; testing the algorithms, techniques, and resources in other domains, applying the resources and tools in other lingual data etc. However, even then, the research is posing some limitations. This chapter highlighted some of the limitations of this research. In the next section, the bibliography of this thesis is presented.