

**MULTILINGUAL SENTIMENT ANALYSIS  
FROM STUDENTS FEEDBACK: OPTIMAL  
TECHNIQUES AND RESOURCES FOR BENGALI  
LINGUAL DATA**

**MOHAMMAD AMAN ULLAH**

**UNIVERSITI SAINS ISLAM MALAYSIA**

**MULTILINGUAL SENTIMENT ANALYSIS FROM STUDENTS  
FEEDBACK: OPTIMAL TECHNIQUES AND RESOURCES FOR  
BENGALI LINGUAL DATA**

Mohammad Aman Ullah

Thesis submitted in partial fulfilment for the degree of  
DOCTOR OF PHILOSOPHY IN  
SCIENCE AND TECHNOLOGY

UNIVERSITI SAINS ISLAM MALAYSIA

July 2021



UNIVERSITI SAINS ISLAM MALAYSIA

جامعة العلوم الإسلامية الماليزية  
ISLAMIC SCIENCE UNIVERSITY OF MALAYSIA

### AUTHOR DECLARATION AND COPYRIGHT

Author's Full Name: MOHAMMAD AMAN ULLAH

Student's Number: 4150133

Title: MULTILINGUAL SENTIMENT ANALYSIS FROM STUDENTS  
FEEDBACK: OPTIMAL TECHNIQUES AND RESOURCES FOR  
BENGALI LINGUAL DATA

Academic Session: ACADEMIC SESSION II 2020/2021 (A202)

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledge.

I hereby declare that the work in this thesis as:

- CONFIDENTIAL** Contains confidential information under the Official Secret Act 1972.\*
- RESTRICTED** Contains restricted information as specified by the organization where research was done.\*\*
- OPEN ACCESS** I agree that my thesis to be published as online open access (full text).

I acknowledged that Universiti Sains Islam Malaysia (USIM) reserves the right as follows:

1. The thesis is solely owned by Universiti Sains Islam Malaysia as stated in the Universiti Sains Islam Malaysia Intellectual Property Policy.
2. The library of Universiti Sains Islam Malaysia has the right to publish my thesis as online open access (fulltext) and make copies for the purpose of research or teaching and learning only.

(Signature of Student)

EG0221337

(MyKAD No./ Passport No.)

Date: 30/5/2021

(Signature of Supervisor)

DR. NORHIDAYAH AZMAN

(Name of Supervisor)

Date: 17 Jul 2021

Notes: \*If the thesis is **confidential**, please attach with the letter from the organization with period and reasons for confidentiality.  
\*\*The **restricted** thesis will be published as online access (fulltext) after 3 years from the date produced.

## ACKNOWLEDGEMENTS

I am grateful to almighty Allah SWT, and His Messenger Muhammad SAW for endless spiritual support in completing my work plan. This Ph.D. journey is a great experience and would not have made it possible without the continuous support, guidance, and diligence of many intelligent people. In this regard, I am very much thankful and show special gratitude to my supervisors **Dr. Norhidayah Binti Azman, Dr. Zulkifly Bin Mohd Zaki, and Dr. Md. Monirul Islam** for their insightful comments, patience, motivation, immense knowledge, and encouragement throughout the journey. I also would like to thank their family members. My special thanks to the Dean of the Faculty for the necessary cooperation. I would also like to express my gratitude to the examiners.

I especially thank my father, mother, brothers, and sisters. My careful parents have sacrificed their lives for me and gave unconditional love and care. I would not have made it this far without the support and guidance from them. I always have my family with me when times are tough. I love them so much. Special thanks to my father and mother-in-law, who have always been supportive and caring in completing this study. I have found one person too caring and loving in these five years is my soul mate Nishat Sultana. She had sacrificed a lot and stuck my side when I was depressed on both academic and personal issues. She appreciates all my good works and keeps encouraging me for new works. I thank her for being with me in my good and bad times and strengthening the commitment to live life to the fullest. To my kids, Ashrafus Sadat Aas and Afrah Binte Aman, you are my inspiration to achieve greatness. You have made me stronger with all the brightness, smile, and love. You have helped me reduce all the stress and anxiety, thereby, continue my work with hope and energy. This work is dedicated to my family.

I would like to thanks funding bodies, all my teachers, colleagues (for being with me throughout the Ph.D. journey with support, cooperation, and inspiration), classmates, friends, research participants, editors/proofreaders, and librarians. Besides, special thanks are due to all the students taking part in the study. Finally, I am thankful to the Universiti Sains Islam Malaysia (USIM) and International Islamic University Chittagong (IIUC) for creating the opportunity of a Ph.D. through a memorandum of understanding. This opportunity has leveraged me to work with some of the brightest scholars and the resources to achieve great success. Many many thanks for this opportunity.

## ABSTRAK

Dalam era kemajuan teknologi maklumat pada masa kini, penglibatan individu di media sosial semakin meningkat setiap hari dan penghasilan teks dalam pelbagai bahasa seperti Bahasa Inggeris dan Bengali semakin banyak disediakan. Penghasilan teks ini mengandungi maklumat berguna yang hanya boleh diperolehi melalui analisis sentimen. Kebanyakan analisis sentimen pada masa kini dilakukan dalam satu bahasa sahaja iaitu Bahasa Inggeris. Selain itu, analisis sentimen yang hanya menggunakan satu bahasa boleh menyebabkan berlakunya kehilangan maklumat berguna yang ditulis dalam bahasa lain. Oleh itu, analisis sentimen pelbagai bahasa adalah penting dan beberapa alat dan teknik bagi tujuan ini telah dibangunkan sehingga hari ini. Proses menganalisis teks dalam pelbagai bahasa menggunakan alat dan teknik tertentu bagi mendapatkan maklumat berguna dikenali sebagai analisis sentimen pelbagai bahasa. Latar belakang kajian dan kajian literatur menunjukkan bahawa terdapat kekurangan reka bentuk umum, teknik prapemprosesan terbaik atau gabungan teknik-teknik tersebut, dan teknik pengekstrakan konsep atau gabungannya bagi tujuan analisis sentimen pelbagai bahasa yang berasaskan ciri dan konsep. Soroton literatur terdahulu juga jelas membuktikan bahawa bidang ini kehilangan asas pengetahuan atau leksikon polariti yang mencukupi untuk menguasai teks Bengali. Bagi merapatkan jurang ini, tesis ini mencadangkan reka bentuk analisis sentimen pelbagai bahasa dengan menggunakan pendekatan berasaskan ciri dan konsep. Kajian ini dapat menyumbang kepada asas pengetahuan Bahasa Bengali (BanglaSenticNet) dengan 30000 konsep, 150000 semantik konsep-konsep tersebut, 72433 konsep leksikon polariti dan juga algoritma bagi analisis sentimen pelbagai bahasa berasaskan konsep. Bagi menguji reka bentuk, asas pengetahuan, leksikon polariti, dan algoritma yang telah dicadangkan, kajian ini telah menyediakan dua set data Bahasa Inggeris dan satu set data Bahasa Bengali menggunakan data yang diperolehi daripada maklum balas pelajar di sosial media, dan satu set data garis dasar Bahasa Inggeris dan Bahasa Bengali masing-masing telah dipertimbangkan bagi tujuan pengesahan. Hasil kajian yang boleh dipercayai dengan ketepatan pengelasan yang baik telah diperolehi menggunakan alatan ini. Kajian ini kemudiannya menguji set data bagi mendapatkan prapemprosesan dan teknik pengekstrakan konsep, ciri atau gabungan mereka yang optimum. Oleh itu, kajian ini mendapati bahawa penerapan teknik prapemprosesan dapat menghasilkan ketepatan yang lebih baik, dan penerapan ciri trigram dan konsep pada satu masa dapat menghasilkan ketepatan pengelasan yang baik. Kajian ini menguji kesemua kes menggunakan Naïve Bayes (NB), Support Vector Machine (SVM), dan Artificial Neural Network - Long Short Term Memory (ANN-LSTM). ANN-LSTM didapati menjadi pengklasifikasi yang terbaik dalam kajian ini. Kajian ini dapat membantu para penyelidik dalam bidang ini untuk menggunakan alatan optimum dan membantu dalam membuat kesimpulan kajian serupa dengan lebih pantas dan mudah. Algoritma yang dicadangkan dalam kajian ini digunakan untuk menguji sumber leksikal seperti BanglaSenticNet, dan prestasinya didapati lebih baik berbanding NB dan SVM dari segi ketepatan. Kajian pada masa akan datang boleh memperluaskan lagi asas pengetahuan dan leksikon polariti dan boleh diuji dengan menggunakan algoritma pengelasan yang lain dalam domain yang berbeza.

## ABSTRACT

In this age of information technology, the individuals in social media are generating vast amounts of helpful multilingual (such as English, Bengali, etc.) data. Sentiment analysis is an approach that could help in organizational and individual decision-making using those data. However, most of today's sentiment analysis is done in a single language, mainly in English thus creating the chance to miss helpful information written in other languages. So, the multilingual sentiment analysis (MLSA) became essential; however, MLSA research faces some problems. Literature review shows existing lexical and knowledge resources are not concept-based, and primarily in English that overlooked the research in a resource-poor language like Bengali. Moreover, the studies that have been done so far have mainly used standard algorithms like Naïve Bayes (NB), Support Vector Machine (SVM), Long Short Term Memory (LSTM) etc., ignoring concept-level MLSA algorithms. Besides, this research has found an insufficient number of student feedback datasets, especially in Bengali. The literature also reveals that preprocessing, feature extraction, and concept extraction techniques need in-depth investigation for sentiment analysis concerning their applications (sole or combination), the effect of using with knowledge bases, algorithms, and datasets of different languages. Therefore, this thesis has contributed a Bengali knowledge base (BanglaSenticNet) of 30,000 concepts and almost 150,000 semantics of those concepts to mitigate the gaps. This research also developed a Bengali polarity lexicon with 72,433 concepts and proposed an algorithm for concept-level MLSA (MCSAlgo). Besides, this research has created two English, and one Bengali students feedback dataset using data from social media. The above knowledge base, polarity lexicon, and algorithm are tested using these datasets and validated the results using both baseline datasets (English and Bengali) and standard algorithms (NB, SVM, and LSTM). The research attained a trustworthy result with good classification accuracy using these resources and algorithms. The MCSAlgo is specially applied to test BanglaSenticNet and Bengali polarity lexicon, and its performance is found to be better than NB and SVM in terms of accuracy, recall, precision, and F-score. These lexical and knowledge resources can be considered the most significant resources available for the Bengali sentiment analysis to the best knowledge. This thesis then tested the data sets to find optimal preprocessing, feature, and concept extraction techniques or their combinations and found that applying them in predefined combinations produced better accuracy with NB, SVM, and LSTM, whereas the LSTM classifier outperforms. Moreover, comparative studies with related works show that the experimental results of current research outperform on the scale of accuracy. It is contemplating that this research will provide a noteworthy contribution and facilitate the researchers of this field with some optimal techniques and resources to conclude similar research more quickly, effectively, and efficiently. Future studies may enlarge the knowledge bases and polarity lexicons and be tested with many other classification algorithms and different domains.

## نبذة مختصرة

يقوم الأفراد بتوليد كميات هائلة من البيانات المفيدة بعدة لغات في وسائل التواصل الاجتماعي (مثل تحليل المشاعر هو منهج يمكن أن الإنجليزية والبنغالية وغيرها) في هذا العصر - عصر التقنية والإعلام. يساعد في اتخاذ القرارات التنظيمية والفردية باستخدام هذه البيانات. ومع ذلك، يتم إجراء معظم تحليل المشاعر اليوم بلغة واحدة، وبشكل رئيسي باللغة الإنجليزية مما يخلق فرصة لتفويت المعلومات المفيدة ضروريًا؛ ومع ذلك، (MLSA) المكتوبة بلغات أخرى. لذلك، أصبح تحليل المشاعر بعدة اللغات بعض المشاكل. مراجعة الأدب تظهر أن الموارد المعجمية والمعرفية الحالية ليست MLSA تواجهت أبحاث قائمة على المفاهيم، وبشكل أساسي باللغة الإنجليزية التي أغفلت البحث بلغة فقيرة الموارد مثل مع ذلك، أن الدراسات التي تم إجراؤها استخدمت حتى الآن بشكل أساسي الخوارزميات البنغالية وما إلى (LSTM) والذاكرة طويلة المدى (SVM) دعم آلة المتجهات، و (NB) القياسية مثل نايف بايز إلى جانب ذلك، هذا البحث وجد عددًا غير على مستوى المفهوم. MLSA ذلك، متجاهلة خوارزميات كافٍ من مجموعات بيانات ملاحظات الطلاب، خاصة باللغة البنغالية. تكشف الأدبيات أيضًا أن تقنيات المعالجة المسبقة واستخراج الميزات واستخراج المفاهيم تحتاج إلى تحقيق متعمق لتحليل المشاعر فيما يتعلق بتطبيقاتها (الفردية أو الجماعية)، وتأثير الاستخدام مع قواعد المعرفة والخوارزميات ومجموعات (BanglaSenticNet) لذلك، ساهمت هذه الأطروحة بقاعدة المعرفة البنغالية. البيانات من لغات مختلفة من 30000 (ثلاثة آلاف) مفهوم وما يقرب من 150000 (مائة وخمسين ألف) دلالة لهذه المفاهيم لقد طوّر هذا البحث أيضًا معجمًا للقبطية البنغالية مع 72433 (اثنين وسبعين). لتخفيف الفجوات على مستوى المفهوم MLSA ألف بعد أربع مائة وثلاثة وثلاثين) مفهومًا واقترح خوارزمية ل إلى جانب ذلك، أنشأ هذا البحث بيانين اثنين باللغة الإنجليزية والبنغالية باستخدام (MCSAigo). بيانات الطلاب من وسائل التواصل الاجتماعي. يتم اختبار قاعدة المعرفة المذكورة أعلاه ومعجم القبطية والخوارزمية باستخدام مجموعات البيانات هذه وتحققت صحة النتائج باستخدام كل من مجموعات أثبتت البحث (NB, SVM, LSTM) البيانات الأساسية (الإنجليزية والبنغالية) والخوارزميات القياسية نتيجة جديرة بالثقة مع دقة تصنيف جيدة باستخدام هذه الموارد والخوارزميات. يتم تطبيق ، ولقد وجد أدائها (BanglaSenticNet) ومعجم القبطية البنغالية خصيصًا لاختبار (MCSAigo) يمكن أيضًا اعتبار هذه الموارد المعجمية والمعرفية من أهم. من حيث الدقة (SVM)، (NB) أفضل من

الموارد المتاحة لتحليل المشاعر البنغالية لأفضل معرفة. اختبرت هذه الأطروحة بعد ذلك مجموعات البيانات للعثور على تقنيات المعالجة المسبقة والميزات واستخراج المفاهيم المثلى أو مجموعاتهما ووجدت أن ، بينما يتفوق (LSTM), (SVM), (NB) تطبيقها في مجموعات محددة مسبقاً ينتج دقة أفضل مع ذلك، تظهر الدراسات المقارنة مع الأعمال ذات الصلة أن النتائج التجريبية (LSTM). مصنف للبحوث الحالية تفوق في الأداء على مقياس الدقة. من المتصور أن هذا البحث سيقدم مساهمة جديدة بالملاحظة ويسهل للباحثين في هذا المجال بعض التقنيات والموارد المثلى لإتمام البحوث المماثلة بسرعة وفعالية وكفاءة. قد توسعت الدراسات المستقبلية قواعد المعرفة ومعاجم القطبية ويتم اختبارها باستخدام العديد من خوارزميات التصنيف الأخرى والمجالات المتنوعة.

## TABLE OF CONTENTS

CONTENT	PAGE
AUTHOR DECLARATION	ii
ACKNOWLEDGEMENTS	iii
ABSTRAK	iv
ABSTRACT	v
نبذة مختصرة	vi
TABLE OF CONTENTS	viii
LIST OF TABLES	xii
LIST OF FIGURES	xv
LIST OF APPENDICES	xvi
LIST OF ABBREVIATIONS	xvii
CHAPTER 1: INTRODUCTION	1
1.1 Introduction	1
1.2 Research Background	4
1.2.1 Sentiment Analysis	5
1.2.2 Multilingual Sentiment Analysis	5
1.2.3 Resources, Algorithms and Techniques for Sentiment Analysis	6
1.3 Problem Statement	9
1.4 Research Questions	12
1.5 Research Objectives	12
1.6 Research Significance	13
1.7 Research Scope	14
1.8 Research Contribution	15
1.9 Summary	16
CHAPTER 2: LITERATURE REVIEW	19
2.1 Natural Language Processing	20
2.1.1 Syntax	22
2.1.1.1 Parts-of-Speech Tagging	22
2.1.1.2 Parsing	23
2.1.1.3 Sentence Breaking or Sentence Boundary Disambiguation	23
2.1.1.4 Stemming	23
2.1.1.5 Word Segmentation	23
2.1.1.6 Terminology Extraction	24
2.1.2 Semantics	24
2.1.2.1 Named Entity Recognition	24
2.1.2.2 Word Sense Disambiguation	25

2.2	Sentiment Analysis Overview	25
2.2.1	Levels of Sentiment Analysis	29
2.2.1.1	Document-Level Sentiment Analysis	29
2.2.1.2	Sentence-Level Sentiment Analysis	29
2.2.1.3	Comparative Sentiment Analysis	30
2.2.1.4	Aspect or Feature-Based Sentiment Analysis	31
2.2.1.5	Concept-Level Sentiment Analysis	34
2.3	Bengali Language	40
2.3.1	Division of Bangla Vocabulary	41
2.3.2	Bengali Language Problems Related to Sentiment Analysis	43
2.3.3	Bengali Sentiment Analysis	44
2.4	Multilingual Sentiment Analysis	47
2.5	Lexicons and Knowledge Bases	51
2.5.1	SenticNet	55
2.6	Sentiment Analysis Algorithms	61
2.6.1	Naive Bayes	63
2.6.2	Support Vector Machine	67
2.6.3	Neural Networks and Deep Learning	72
2.6.3.1	Convolutional Neural Networks	73
2.6.3.2	Recursive Neural Networks	74
2.6.3.3	Recurrent Neural Networks	75
2.6.3.4	Long Short-Term Memory Networks	75
2.6.4	Multilingual Concept-level Sentiment Analysis Algorithms	79
2.7	Sentiment Analysis of Student Feedback Datasets	88
2.8	Feature Extraction	92
2.9	Preprocessing Techniques	98
2.10	Natural Language Processing Tools	106
2.11	Research Gap	108
2.12	Summary	110
CHAPTER 3: METHODOLOGY		112
3.1	Methodology	112
3.2	Knowledge Base Creation	117
3.2.1	Concept Extraction	122
3.2.2	Estimating Concepts Score	125
3.2.3	Sentiment Matrix Formation and Clustering	127
3.2.4	Final Concept Score Determination	128
3.2.5	Knowledge Base Evaluation	129
3.3	Lexicon Creation	129
3.4	Proposed Algorithm for Multilingual Sentiment Analysis at Concept Level	131
3.5	Dataset Creation from Multilingual Data	133
3.6	Feature or Concept Extraction	138
3.7	Preprocessing	141
3.8	Performance Evaluation	144
3.8.1	Model Evaluation Metrics	145
3.8.1.1	Confusion Matrix	145
3.8.1.2	Accuracy	145
3.8.1.3	Recall or Sensitivity	146

3.8.1.4	Precision	146
3.8.1.5	F-score	146
3.9	Tools	149
3.10	Summary	149
CHAPTER 4: DATA ANALYSIS AND INTERPRETATION		152
4.1	Knowledge Bases	152
4.1.1	Research Question	152
4.1.2	Collecting Data for Knowledge Base and Polarity Lexicon	153
4.1.2.1	Data Acquisition	153
4.1.2.2	Born-digital Data	153
4.1.3	Data Preprocessing	154
4.1.4	Quality Data	154
4.1.4.1	Compiling Data	155
4.1.5	Labels and Meta Data	156
4.1.6	Data Analysis and Interpretation for the Knowledge Base and Polarity Lexicon	156
4.2	Data Sets	163
4.2.1	Research Question	163
4.2.2	Collecting Data for Dataset	163
4.2.3	Organizing Data	165
4.2.3.1	Data Consolidation	165
4.2.3.2	Data Separation and Conversion	166
4.2.3.3	Language Identification and Data Separation	166
4.2.3.4	Convert to Single Language Data	166
4.2.4	Data Pre-Processing	166
4.2.4.1	Data Cleaning	166
4.2.4.2	Data Transformation	167
4.2.4.3	Data Reduction	167
4.2.5	Data Annotation	167
4.2.6	Method Selection	168
4.2.7	Data Analysis and Interpretation	168
4.3	Summary of Data Analysis and Interpretation	171
CHAPTER 5: RESULTS AND DISCUSSION		174
5.1	Performance Analysis of Bengali Knowledge Base (BanglaSenticNet), Polarity Lexicon, and Proposed Algorithm (MCSAlgo)	174
5.1.1	Comparing Performance of BanglaSenticNet, Bangla Polarity Lexicon, and MCSAlgo with State-of-the-Art Research	179
5.2	Performance Evaluation of Created Bengali and English Datasets with Baseline Datasets	182
5.2.1	Comparison with State-of-the-Art Research on Datasets	182
5.3	Testing with Different Feature and Concept Extraction Techniques	184
5.3.1	Comparison of Feature and Concept Extraction Techniques Performance on Different Data Sizes	184
5.3.2	Experiment with Different Feature and Concept Extraction Techniques on Whole English Datasets	186

5.3.3	Average Performance of Feature and Concept-Based Approach on English and Bengali Datasets	190
5.3.4	Maximum Performance of Feature and Concept-Based Approach on English and Bengali Datasets	193
5.3.5	Comparison with State-of-Art Feature and Concept Extraction Technique Research	194
5.4	Testing with Different Preprocessing Techniques	197
5.4.1	Experiment with the Different Preprocessing Combination on English and Bengali Datasets	197
5.4.1.1	Experiment with No Preprocessing and Applying All Pre-Processing Techniques at Once	198
5.4.1.2	Experiment with Single Preprocessing Technique	201
5.4.1.3	Experiment with Two Preprocessing Technique Combinations	205
5.4.1.4	Experiment with Three Preprocessing Technique Combinations	207
5.4.1.5	Experiment with Four Preprocessing Technique Combinations on English Datasets	212
5.4.1.6	Experiment with Five Preprocessing Technique Combinations	216
5.4.1.7	Experiment with Six Preprocessing Technique Combinations	218
5.4.1.8	Average Performance on Different Preprocessing Techniques for English and Bengali Datasets	222
5.4.1.9	Maximum Performance on Different Types of Preprocessing Techniques for English and Bengali Datasets	224
5.4.1.10	Comparison with State-of-Art Research That Emphasized Preprocessing	226
5.5	Results Summary and Discussion	227
5.5.1	Results Summary and Discussion on the Performance of BanglaSenticNet, Polarity Lexicon, and MCSAlgo	227
5.5.2	Results Summary and Discussion on the Performance of Different Datasets	229
5.5.3	Results Summary and Discussion of Performance on Feature and Concept Extraction Technique and Their Combinations	230
5.5.4	Results Summary and Discussion for Preprocessing Technique Combinations Using English and Bengali Data	233
5.6	Results and Discussion Chapter Summary	238
CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS		240
6.1	Conclusions	240
6.2	Future Directions	243
6.3	Limitations of Study	245
6.4	Summary	245
REFERENCES		247
APPENDICES		280

## LIST OF TABLES

<b>Tables</b>	<b>Page</b>
Table 2.1: Multilingual Approaches	39
Table 2.2: Existing Lexicons and Corpora of Mono and Multilingual Sentiment Analysis	54
Table 2.3: Monolingual and Multilingual Models, Algorithms, Resources and Related Performance	85
Table 2.4: Works on Student Feedback Datasets	90
Table 2.5: Summary on Feature Extraction Task in State-of-Art Research	95
Table 2.6: Summary of Bengali SA that Applied Preprocessing Techniques	102
Table 2.7: Summary of SA Research that Applied Preprocessing Techniques	104
Table 2.8: Tools for multilingual Sentiment analysis	107
Table 3.1: Confusion Matrix	145
Table 3.2: Overview of Tools/Algorithm	149
Table 4.1: Statistics of Lexicons	158
Table 4.2: Some example of BanglaSenticNet Entries	161
Table 4.3: Statistics of Bengali and English SenticNet	161
Table 4.4: Statistics of Datasets	170
Table 5.1: Performance of BanglaSenticNet, MCSAlgo, and Relative Evaluation with Baseline	177
Table 5.2: Performance of Bengali and English Polarity Lexicon	178
Table 5.3: Significance Analysis of MCSAlgo and Other Applied Algorithms	179
Table 5.4: Performance Comparison with State-of-Art Research	181
Table 5.5: Significance Analysis of Datasets with Baseline	182
Table 5.6: Comparison with State-of-Art Research on Datasets	183
Table 5.7: Performance of Feature and Concept Extraction Techniques on Different Data Size	185

Table 5.8: Result on Different Feature and Concept Extraction Techniques on Whole English Datasets	188
Table 5.9: Result of Different Feature and Concept Extraction Techniques on Whole Bengali Datasets	189
Table 5.10: Average Performance of the Feature and Concept-Based Approach on Individual Dataset	192
Table 5.11: Average Performance of Feature and Concept-Based Approach on All datasets	193
Table 5.12: Maximum Performance of the Classifier on Feature and Concept-Based Approach	194
Table 5.13: Comparison with State-of-Art Feature and Concept Extraction Technique Research	196
Table 5.14: List of Abbreviations of Different Preprocessing Techniques	197
Table 5.15: Result of No Preprocessing and All Preprocessing Type Formation on English Datasets	200
Table 5.16: Result on No Preprocessing and All Preprocessing Type Formation Applied to Bengali Datasets	201
Table 5.17: Result of Single Preprocessing Technique on English Dataset	203
Table 5.18: Result of Single Preprocessing Technique Applied to Bengali Datasets	204
Table 5.19: Result of Two Preprocessing Technique Combinations on English Datasets	206
Table 5.20: Result of Two Preprocessing Technique Combinations Applied to Bengali Datasets	207
Table 5.21: Result of Three Preprocessing Technique Combinations on English Datasets	209
Table 5.22: Result of Three Preprocessing Technique Combinations Applied to Bengali Datasets	211
Table 5.23: Result of Four Preprocessing Technique Combinations on English Datasets	213
Table 5.24: Results of Four Preprocessing Technique Combinations Applied to Bengali Datasets	215
Table 5.25: Result of Five Preprocessing Technique Combinations on English Datasets	217

Table 5.26: Result of Five Preprocessing Technique Combinations Applied to Bengali Datasets	218
Table 5.27: Result of Six Preprocessing Technique Combinations on English Datasets	221
Table 5.28: Result of Six Preprocessing Technique Combinations Applied to Bengali Datasets	221
Table 5.29: Average Performance of Preprocessing Techniques for English Datasets	223
Table 5.30: Average Performance of Preprocessing Techniques on Bengali Datasets	224
Table 5.31: Maximum Performance of Classifiers for English Datasets	225
Table 5.32: Maximum Performance of Classifiers for Bengali Datasets	225
Table 5.33: Comparison with State-of-Art Research on Preprocessing Techniques	226

## LIST OF FIGURES

<b>Figures</b>	<b>Page</b>
Figure 2.1: Sentiment Analysis Tasks, Approaches, and Applications	26
Figure 3.1: Flow chart of this Research	116
Figure 3.2: Knowledge Base Creation Process	120
Figure 3.3: Parse Tree of the Sentence “I was going to the university”	124
Figure 3.4: Parse Tree of Sentence “আমি বিশ্ববিদ্যালয়ে যাচ্ছিলাম”	124
Figure 3.5: Data Collection Method	135
Figure 3.6: Annotated Multilingual Dataset Creation Process	136
Figure 3.7: Pre-processing Topology	144
Figure 3.8: Abstract View of Figure 3.1 Showing Resources and Techniques Evaluation Process	148
Figure 4.1: Statistics of Bengali and English Lexicons	158
Figure 4.2: Statistics of Bengali and English SenticNet	162
Figure 4.3: Statistics of Datasets	170

## LIST OF APPENDICES

<b>Appendices</b>	<b>Page</b>
Appendix 1: Certificate of translation accuracy	280
Appendix 2: Sample of translated polarity lexicon	281
Appendix 3: Sample of translated knowledge base	284
Appendix 4: Sample copy of translated facebook comments	289

## LIST OF ABBREVIATIONS

NB	Naïve Bayes
SVM	Support Vector Machine
ANN- LSTM	Artificial Neural Network- Long Short Term Memory
SA	Sentiment Analysis
OM	Opinion Mining
NLP	Natural Language Processing
EDM	Educational Data Mining
MLSA	Multilingual Sentiment Analysis
ML	Machine Learning
LSTM	Long Short Term Memory
CNN	Convolutional Neural Network
ME	Maximum Entropy
CNB	Complementary Naïve Bayes
RP	Remove Punctuation
N	Negation
RLR	Reduction of Letter Reputation
SWD	Stop Word Deduction
S	Stemming
T	Tokenization
CC	Case Conversion
TD – IDF	Term Frequency- Inverse Document Frequency
MARS	Multivariate adaptive regression splines
NLU	Natural Language Understanding
NLG	Natural Language Generation
NL	Natural Language
NER	Named Entity Recognition
OCR	Optical Character Recognition
PCFG	Probabilistic Context-Free Grammar
BOW	Bag of Words
SO	Semantic Orientations
PMI	Mutual Information
CSR	Class Sequential Rule
HMMs	Hidden Markov Model
POS	Part-of-Speech
SS-FE	Sample Selection and Feature Ensemble
SNBC	Semantic Naïve Bayes Classifier
PCA	Principal Component Analysis
MRS	Minimal Recursion Semantics
NP	Noun Phrase
DT	Decision Tree
KNN	K-Nearest Neighbor
RF	Random Forest
FFNNs	Feed Forward Neural Networks
DNN	Deep Neural Network

MLPs	Multilayer Perceptrons
BNNs	Biological Neural Networks
GRU	Gated Recurrent Unit
RvNN	Recursive Neural Network
SMT	Statistical Machine Translation
WNA	WordNet-Affect
STS	Stanford Twitter Sentiment
PTE	Partial Textual Entailment
BOC	Bag of Concepts
MASC	Multi-Domain Arabic Sentiment Corpus
NN	Neural Network
IPA	International Phonetic Alphabet
OOV	Out-of-Vocabulary
IV	In-Vocabulary
LR	Logistic Regression
SVR	Support Vector Regression
BR	Bagging Regressor
ABR	Adaboost Regressor
GBR	Gradient Boosting Regressor
XGB	Xgboost Regressor
SN	SenticNet
C	Concepts
POV	Polarity Value
BSN	Banglasenticnet
EBCP	English Bangla Concept Polarity Lexicon
SC	Semantics of Concepts
PV	Pleasantness Value
I	Intensity
AV	Attention Value
SV	Sensitivity Value
APV	Aptitude Value
PM	Primary Mood
SM	Secondary Mood
PL	Polarity Label
WS	Concepts With Semantics
WOS	Concepts Without Semantics
IIUC	International Islamic University Chittagong
API	Application Programming Interfaces
CSV	Comma Separated Value
D	Datasets
U	Unigram Dictionary
B	Bigram Dictionary