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
جَامِعَة العُلُوم الإِسْلَامِيَة المَالِيزِيَة
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:

☐	**CONFIDENTAL**	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:

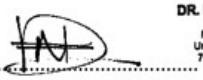
1. The thesis is solely owned by Universiti Sains Islam Malaysia as stated in the Universiti Sains Islam Malaysia Intellectual Property Policy.
2. The library of Universiti Sains Islam Malaysia has the right to publish my thesis as online open access (fulltext) and make copies for the purpose of research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

(Signature of Student)

EG0221337

(MyKAD No./ Pasport No.)

Date: 30/5/2021



DR. NORHIDAYAH AZMAN
Pensyarah Kanan
Fakulti Sains Dan Teknologi
Universiti Sains Islam Malaysia
71800 Nilai, Negeri Sembilan
Tel : 06-798 8779

(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 diperoleh 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 diperoleh 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 ألاف) مفهوم وما يقرب من 150000 (مائة وخمسين ألف) دلالة لهذه المفاهيم

لقد طوّر هذا البحث أيضًا معجمًا للقطبية البنغالية مع 72433 (اثنين وسبعين .لتخفيف الفجوات على مستوى المفهوم MLSA ألف بعد أربعمائة وثلاثة وثلاثين) مفهومًا واقترح خوارزمية لـ إلى جانب ذلك، أنشأ هذا البحث بيانين اثنين باللغة الإنجليزية والبنغالية باستخدام .(MCSAlgo) بيانات الطلاب من وسائل التواصل الاجتماعي. يتم اختبار قاعدة المعرفة المذكورة أعلاه ومعجم القطبية والخوارزمية باستخدام مجموعات البيانات هذه وتحققت صحة النتائج باستخدام كل من مجموعات أثبت 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 أداؤها (BanglaSenticNet) ومعجم القطبية البنغالية خصيصًا لاختبار(MCSAlgo) 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 مجموعاتها ووجدت أن (NB), (SVM), (LSTM) تطبيقها في مجموعات محددة مسبقًا ينتج دقة أفضل مع ، بينما يتفوق 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 جديرة 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

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

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