

**COMPREHENSIVE EVALUATION OF INDIAN GOODS AND
SERVICES TAX ROLLOUT USING DEEP LEARNING AND
MACHINE LEARNING OF SOCIAL MEDIA IMPRESSIONS
AND INFORMATION SYSTEM SUCCESS MODEL**

Pankaj Dikshit



**AMAR NATH AND SHASHI KHOSLA SCHOOL OF
INFORMATION TECHNOLOGY**

INDIAN INSTITUTE OF TECHNOLOGY DELHI

October 2022

© Indian Institute of Technology Delhi (IITD), New Delhi, 2022

**COMPREHENSIVE EVALUATION OF INDIAN GOODS AND
SERVICES TAX ROLLOUT USING DEEP LEARNING AND
MACHINE LEARNING OF SOCIAL MEDIA IMPRESSIONS
AND INFORMATION SYSTEM SUCCESS MODEL**

by

Pankaj Dikshit

Amar Nath and Shashi Khosla School of Information Technology

Submitted

**in fulfilment of the requirements of the degree of Doctor of Philosophy
to the**



INDIAN INSTITUTE OF TECHNOLOGY DELHI

October 2022

“Dedicated to my family (Ila, Pulkit and Saakshi) and my parents for the inspiration, devotion and constant presence!”

Certificate

This is to certify that the thesis titled **Comprehensive Evaluation of the Indian Goods and Services Tax Rollout using Deep Learning and Machine Learning of Social Media Impressions and Information System Success Model** being submitted by **Mr. Pankaj Dikshit** (2015ANZ8479) for the award of **Doctor of Philosophy** in Information Technology is a record of bona fide work carried out by him under our guidance and supervision at the Amar Nath and Shashi Khosla School of Information Technology, Indian Institute of Technology, Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Dr B. Chandra

Adjunct Professor
School of Information Technology
Indian Institute of Technology
Delhi

Date:

Dr M. P. Gupta

Professor
Department of Management Studies
Indian Institute of Technology
Delhi

Date:

Acknowledgements

At the outset, I express my sincere gratitude to my research guides Professor B Chandra and Professor M P Gupta, of School of IT and the Department of Management Studies respectively for their constant guidance and motivation in the conduct of this research.

I am extremely thankful to Professor B Chandra who has devoted extensive time at each step to explain the deep learning and machine learning techniques and apply them in an innovative and unique manner for conducting this research of the GST System from multiple perspectives. Her depth of knowledge and vast experience in the subject of machine learning, deep learning and data mining immensely contributed to the depth of research. Her vision, inspiration and constant encouragement was key and fundamental to the fructification of the theses. Her effort in the conduct of this research is unparalleled.

I would like to pay my special regards and thank Professor M P Gupta and Professor Arpan Kumar Kar who always gave their time and guidance and invaluable advice in the conduct of the research.

I would also take the opportunity to express my thanks to Goods and Services Tax Network (GSTN), my place of work, for allowing me the time and opportunity to conduct research of this interesting subject from School of Information Technology, IIT, Delhi.

Besides my advisors, I would like to thank the Student Research Committee members, Professor Prem K Kalra and Professor Aaditeshwar Seth and Professor Aparna Mehra for their insightful comments and encouragement.

I express my gratitude to Professor Kolin Paul, the Head of School of IT, for all the cooperation.

New Delhi

Pankaj Dikshit

Abstract

The GST tax regime was rolled out in June 2017. This was a big-bang transformation to have a unified tax system for the entire country. Whenever such a new tax law is introduced, certain section of the users has satisfaction whereas others have grievances. They express their feelings on social media like Twitter and Facebook. In order to have a meaningful and effective transformation, the sentiments of the taxpayers have to be taken into account. It is all the more important to look into greater detail regarding terms they use in the tweets.

This thesis aims at providing detailed analysis of the newly introduced GST system using machine learning and deep learning techniques in order to improve the system. The thesis also aims to evaluate the success of the GST system.

Tweets by the taxpayers with respect to the GST regime and the system collected from the year June 2017 to May 2020 formed the database for this analysis. The entire analysis is in five different modules. In the first module sentiment analysis has been carried out using deep learning techniques viz. CNN, LSTM and bi-directional LSTM with different context-based embeddings like BERT, ROBERTA and Universal sentence encoder. Prediction of sentiments for the last has been carried out based on training data of tweets of previous 24 months. In the second module a vivid analysis has been portrayed to show how the attention weights of the key words appearing in the tweets (related to GST) vary over the quarters in the three-year period. This has been achieved using attention based bidirectional LSTM model. Thorough analysis has been carried out to show how the attention weights of key words that exceed 2 sigma limits contribute towards the positive and negative sentiment towards GST.

In the third module it has been the attempted to find the importance of pairs of terms appearing in the tweets using a hybrid approach of co-occurrence graph and attention based deep learning technique. Co-occurrence graphs have been constructed pertaining to different categories in GST. The frequency of edges (in the co-occurrence graphs) between the key words and special terms given by GST formed the basis for finding the important pairs of terms occurring in the tweets. Attention weights for these important pairs have been evaluated using bidirectional LSTM model. Finally, a new measure called ‘attention factor’ has been coined, which is a product of the frequency of occurrence of the pairs determined using co-occurrence graphs and the attention weight determined using attention based bidirectional LSTM model.

The attention factor gives the highly important pair of terms occurring with the keywords in the tweets.

A helpdesk was established by GST to answer the queries posed by the taxpayers. This was a manual system bringing out the need to automate the question answer system. In the fourth module clustering techniques have been used with context-based embeddings like BERT and ROBERTA. Clustering of embeddings of questions have been carried out using different distance measures and three choices of answers have been provided for each question using similarity score. The entire analysis has also been carried out by selecting important terms from the questions using inverse-document frequency technique.

In the fifth module, the success of the GST e-governance information system has been evaluated using modified De Lone and McLean information system success model. The reliability analysis along with convergent and divergent validity is carried out followed by structured equation modeling. Regression and path analysis is also performed to evaluate the success of the GST e-governance system.

सारांश

जीएसटी कर व्यवस्था जून 2017 में लागू की गई थी। पूरे देश के लिए एक एकीकृत कर प्रणाली रखने के लिए यह एक बड़ा बदलाव था। जब भी इस तरह का कोई नया कर कानून पेश किया जाता है, तो उपयोगकर्ताओं के कुछ वर्ग को संतुष्टि मिलती है जबकि अन्य को शिकायतें होती हैं। वे ट्विटर और फेसबुक जैसे सोशल मीडिया पर अपनी भावनाओं को व्यक्त करते हैं। एक सार्थक और प्रभावी परिवर्तन के लिए, करदाताओं की भावनाओं को ध्यान में रखा जाना चाहिए। ट्वीट्स में उपयोग किए जाने वाले शब्दों के बारे में अधिक विस्तार से देखना अधिक महत्वपूर्ण है।

इस थीसिस का उद्देश्य सिस्टम में सुधार के लिए मशीन लर्निंग और डीप लर्निंग तकनीकों का उपयोग करके नई शुरु की गई जीएसटी प्रणाली का विस्तृत विश्लेषण प्रदान करना है। थीसिस का उद्देश्य जीएसटी प्रणाली की सफलता का मूल्यांकन करना भी है।

जीएसटी व्यवस्था और वर्ष जून 2017 से मई 2020 तक एकत्र की गई प्रणाली के संबंध में करदाताओं द्वारा किए गए ट्वीट ने इस विश्लेषण के लिए डेटाबेस का गठन किया। पूरा विश्लेषण पांच अलग-अलग मॉड्यूल में है। पहले मॉड्यूल में भावना विश्लेषण को सीएनएन, एलएसटीएम और द्वि-दिशात्मक एलएसटीएम जैसे गहरे सीखने की तकनीकों का उपयोग करके किया गया है, जिसमें बर्ट, रॉबर्टा और यूनिवर्सल वाक्य एन्कोडर जैसे विभिन्न संदर्भ-आधारित एम्बेडिंग शामिल हैं। पिछले 24 महीनों के ट्वीट्स के प्रशिक्षण डेटा के आधार पर अंतिम के लिए भावनाओं का पुनरुत्पादन किया गया है। दूसरे मॉड्यूल में एक ज्वलंत विश्लेषण को यह दिखाने के लिए चित्रित किया गया है कि ट्वीट्स (जीएसटी से संबंधित) में दिखाई देने वाले प्रमुख शब्दों के ध्यान वजन तीन साल की अवधि में तिमाहियों में कैसे भिन्न होते हैं। यह ध्यान आधारित द्विदिश LSTM मॉडल का उपयोग करके प्राप्त किया गया है। यह दिखाने के लिए गहन विश्लेषण किया गया है कि 2 सिग्मा सीमा से अधिक प्रमुख

शब्दों के ध्यान भार जीएसटी के प्रति सकारात्मक और नकारात्मक भावना की ओर कैसे योगदान करते हैं।

तीसरे मॉड्यूल में यह सह-घटना ग्राफ और ध्यान आधारित गहरी सीखने की तकनीक के हाइब्रिड दृष्टिकोण का उपयोग करके ट्वीट्स में दिखाई देने वाले शब्दों के जोड़े के महत्व को खोजने का प्रयास किया गया है। जीएसटी में विभिन्न श्रेणियों से संबंधित सह-घटना ग्राफ का निर्माण किया गया है। प्रमुख शब्दों और जीएसटी द्वारा दिए गए विशेष शब्दों के बीच किनारों की आवृत्ति (सह-घटना ग्राफ में) ने ट्वीट्स में होने वाले शब्दों के महत्वपूर्ण जोड़े को खोजने के लिए आधार बनाया। इन महत्वपूर्ण जोड़ों के लिए ध्यान वजन द्विदिश LSTM मॉडल का उपयोग करके मूल्यांकन किया गया है। अंत में, 'ध्यान कारक' नामक एक नया उपाय गढ़ा गया है, जो सह-घटना रेखांकन का उपयोग करके निर्धारित जोड़े की घटना की आवृत्ति का एक उत्पाद है और ध्यान आधारित द्विदिश एलएसटीएम मॉडल का उपयोग करके निर्धारित ध्यान वजन है। ध्यान कारक tweets में खोजशब्दों के साथ होने वाली शर्तों की अत्यधिक महत्वपूर्ण जोड़ी देता है।

करदाताओं द्वारा पूछे गए प्रश्नों का उत्तर देने के लिए जीएसटी द्वारा एक हेल्पडेस्क की स्थापना की गई थी। यह एक मैनुअल प्रणाली थी जो प्रश्न उत्तर प्रणाली को स्वचालित करने की आवश्यकता को बाहर लाती थी। चौथे मॉड्यूल में क्लस्टरिंग तकनीकों का उपयोग बर्ट और रॉबर्टा जैसे संदर्भ-आधारित एम्बेडिंग के साथ किया गया है। विभिन्न दूरी उपायों का उपयोग करके प्रश्नों के एम्बेडिंग का क्लस्टरिंग किया गया है और समानता स्कोर का उपयोग करके प्रत्येक प्रश्न के लिए उत्तरों के तीन विकल्प प्रदान किए गए हैं। व्युत्क्रम-दस्तावेज़ आवृत्ति तकनीक का उपयोग करके प्रश्नों में से महत्वपूर्ण शब्दों का चयन करके भी पूरा विश्लेषण किया गया है।

पांचवें मॉड्यूल में, जीएसटी ई-गवर्नेंस सूचना प्रणाली की सफलता का मूल्यांकन संशोधित डी लोन और मैकलीन सूचना प्रणाली की सफलता मॉडल का उपयोग करके किया गया है। अभिसरण और

अलग-अलग वैधता के साथ विश्वसनीयता विश्लेषण संरचित समीकरण मॉडलिंग के बाद किया जाता है। प्रतिगमन और पथ विश्लेषण भी जीएसटी ई-गवर्नेंस प्रणाली की सफलता का मूल्यांकन करने के लिए किया जाता है।

Contents

	Page
Chapter 1 : Introduction	1
1.1 Background and Introduction to the work	1
1.2 Literature Survey	4
1.2.1 Sentiment analysis using Deep learning techniques with different embeddings.	4
1.2.2 Literature Survey on Co-occurrence analysis using Attention Weights	6
1.2.3 Question Answer Systems	6
1.2.4 Literature study on Evaluation of the Success of e-governance Information Systems	8
1.2.5 Key areas of research in this Thesis related to GST	8
1.2.6 Summary of the Work done in different Chapters	9
Chapter 2: Sentiment Analysis using Deep Learning techniques with different Embeddings using Attention Weights and assessing behavior of key terms in social media	15
2.1 Introduction	15
2.2 Brief Summary of LSTM, Bi-LSTM and CNN Deep Learning techniques	16
2.2.1 LSTM and Bi-LSTM	16
2.2.2 Brief description of CNN	18
2.3 BERT, ROBERTA and Universal Sentence Encoder embeddings	20
2.4 Sentiment Analysis of Tweets from Twitter related to GST	21
2.5 Attention weight based analysis of tweets to study the behaviour of key terms on quarterly basis	25
2.5.1 Quarterly prediction of sentiments	25
2.5.2 Attention-weight based analysis	27
2.6 Contribution of key words towards Positive and Negative Sentiments of Tweets	37
2.7 Conclusions	48
Chapter 3: Co-occurrence analysis of terms related to Goods and Services Tax using Attention Factor	49
3.1 Introduction	49
3.2 Methodology	50
3.3 Detailed Analysis using co-occurrence graph	51
3.3.1 Construction of co-occurrence graphs and analysis	51
3.3.2 Summarised analysis of frequency of edges in the co-occurrence graphs	65
3.4 Attention weight based analysis	72

3.5 Conclusions	77
Chapter 4: Question and Answer Systems	78
4.1 Introduction	78
4.2 Methodology for QA Automation	79
4.2.1 Pre-processing of the text in questions and answers	80
4.2.2 Forming embeddings of questions and answers	80
4.2.3 Clustering of Embeddings of Questions and Prediction of answers for test questions	80
4.2.4 Choosing important terms in the questions posed	81
4.2.5 Overview Clustering Algorithms	82
4.2.5.1 Hierarchical Clustering Algorithms	82
4.2.5.2 K Means Clustering	83
4.3 Prediction of answers for Sample Questions based on Similarity scores	83
4.3.1 Similarity Scores of Predicted answers of Sample Questions using all terms in the Questions	84
4.3.1.1 Clustering using Euclidean distance	84
4.3.1.2 Clustering using Cosine Distance	94
4.4 Similarity scores of predicted answers of sample questions using important terms in questions and creating embeddings with ROBERTA.	102
4.4.1 Clustering with important terms using Euclidean distance	102
4.4.2 Clustering with important terms using cosine distance	106
4.5. Detailed analysis to determine the Embedding and distance measure for clustering for best prediction of answers to questions	110
4.5.1 Clustering using Euclidean distance	110
4.5.2 Clustering using Cosine distance	112
4.5.3 Clustering using important terms in the questions	114
4.6 Conclusions	115
Chapter 5: Information System Success Assessment of the GST e-Governance Information System using DeLone and McLean ISSM	117
5.1 Introduction	117
5.2 Methodology of assessment of the GST e-governance information system	118
5.2.1 System and Information Quality => Use and User Satisfaction	119
5.2.2 Service Quality => Use and User Satisfaction	120
5.2.3 Relationship between Government Policy Intervention and System Implementation and Technostress factors	120
5.2.4 System Implementation and System Quality	122

5.2.5 Relationship between Technostress factors (Techno Overload & Techno Uncertainty) and Use and Benefits to enterprise	122
5.3 Sampling and survey methodology	123
5.4 Demography of companies	123
5.5 Analysis and Results	124
5.5.1 Reliability, Validity and Multicollinearity	124
5.5.2 Confirmatory factor analysis (CFA) and Structural equation modelling	128
5.5.3 Regression and Path Analysis	129
5.6. Discussion	131
5.6.1 Theoretical implications	132
5.6.2 Implications for policy and practice	132
5.7. Conclusions	133
References	136
List of Publications	145
Appendix A	146
Appendix B	171
Appendix C	219
Bio Data	220

List of Tables

	Page
Chapter 1	
Chapter 2	
2.1. Categories of terms and key terms related to GST	22
2.2. Month-wise prediction accuracy using CNN for June 2019 to May 2020 using different embeddings	24
2.3. Month-wise prediction accuracy using LSTM for June ‘19 to May ‘20 using different embeddings	24
2.4. Month-wise prediction accuracy using bi-LSTM for June ‘19 to May ‘20 using BERT & ROBERTA	25
2.5. Division Of Training and Testing periods.	26
2.6. Training and testing accuracy of prediction over different parts using bi-LSTM	26
2.7. Count of positive and negative sentiments of key words <i>amendment, audit, council, evasion, eway, askgst_goi, filing, tax, invoice & portal.</i> (2017 – February 2020).	38, 39, 44
2.8. Count of average sentiment of 10 key words from June 2017 to May 2020	47
Chapter 3	
3.1. Categories and key terms pertaining to GST	52
3.2. Frequency of edges between key words in E Way bill and special terms	54
3.3. Frequency of edges between key words in ‘Ease of doing Business’ category and special terms	57
3.4. Top ten frequency of edges between key words in “ <i>GST Entities</i> ” category and special terms	58
3.5. Top ten frequency of edges between key words in “ <i>Registration</i> ” category and special terms	58
3.6. Top ten frequency of edges between key words in “ <i>Returns</i> ” category and special terms	59
3.7. Top ten frequency of edges between key words in “ <i>Transition</i> ” category and special terms	60
3.8. Top ten frequency of edges between key words in “ <i>Input Credit</i> ” category and special terms	60
3.9. Top ten frequency of edges between key words in “ <i>Payment</i> ” category and special terms	61

3.10. Top ten frequency of edges between key words in “ <i>Refund</i> ” category and special terms	62
3.11. Top ten frequency of edges between key words in “ <i>Audit</i> ” category and special terms	62
3.12. Top ten frequency of edges between key words in “ <i>Services</i> ” category and special terms	63
3.13. Top ten frequency of edges between key words in “ <i>Assessment</i> ” category and special words	63
3.14. Top ten frequency of edges between key words in “ <i>Tax</i> ” category and special terms	64
3.15. Summary of association of special words with each category: of 2017	65
3.16. Summary of association of special words with each category: of 2018	67
3.17. Summary of association of special words with each category: of 2019	69
3.18. Attention factor based list of key word pairs for 2017	72
3.19. Attention factor based list of key word pairs for 2018	73
3.20. Attention factor based list of key word pairs for 2019	75
 Chapter 4	
4.1. Five test questions posed in November and the corresponding manual answer	85
4.2. Comparative performance on the basis of similarity scores of Hierarchical and K means clustering (with Euclidean distance for BERT embedding) for answers of November with clustering of questions of October	85
4.3. Comparative performance on the basis of similarity scores of Hierarchical and K means Clustering (with Euclidean distance for ROBERTA embedding) for answers of October with clustering of questions of November	87
4.4. Five test questions posed in October and the corresponding manual answer	90
4.5. Comparative performance on the basis of similarity scores of Hierarchical and K means Clustering (with Euclidean distance for BERT embedding) for answers of October with clustering of questions of November	90
4.6. Comparative performance on the basis of similarity scores of Hierarchical and K means Clustering (with Euclidean distance for ROBERTA embedding) for answers of November with clustering of questions of October	92

4.7. Comparative performance on the basis of similarity scores of Hierarchical and K means clustering (with cosine distance for ROBERTA embedding) for answers of October with clustering of questions of November	95
4.8. Comparative performance on the basis of similarity scores of Hierarchical and K means clustering (with cosine distance for BERT embedding) for answers of October with clustering of questions of November	97
4.9. Comparative performance on the basis of similarity scores of Hierarchical and K means Clustering (using cosine distance for ROBERTA embedding) for answers of November with clustering of questions of October	98
4.10. Comparative performance on the basis of similarity scores of Hierarchical and K means Clustering (using cosine distance for BERT embedding) for answers of November with clustering of questions of October	100
4.11. Comparative performance on the basis of similarity scores of Hierarchical and K means clustering with important terms (using Euclidean distance with ROBERTA embedding) for answers of November with clustering of questions of October	102
4.12. Comparative performance on the basis of similarity scores of Hierarchical and K means Clustering with important terms (using Euclidean distance with ROBERTA embedding) for answers of October with clustering of questions of November	104
4.13. Comparative performance on the basis of similarity scores of Hierarchical and K means clustering with important terms (by using cosine distance) for answers of November with clustering of questions of October	106
4.14. Comparative performance on the basis of similarity scores of Hierarchical and K means with important terms (by using cosine distance) for answers of October with clustering of questions of November	108
4.15. Count of answers with similarity scores > 0.8 and < 0.6 : Clustering using Euclidean distance	111
4.16. Count of answers with similarity scores > 0.8 and < 0.6 : Clustering using Cosine distance	112
4.17. Count of answers with similarity scores > 0.8 and < 0.6 : Clustering with important terms using Euclidean distance	114
4.18. Count of answers with similarity scores > 0.8 and < 0.6 : Clustering with important terms using Cosine distance	114

Chapter 5	
5.1. Notifications that were issued by Government of India (from July 2017 to September 2019).	121
5.2. The industry demographics of companies who responded to the survey (as per NACE rev 2 definitions).	124
5.3. Descriptive statistics for the variables.	125
5.4. Reliability test for observed variables.	125
5.5. Measurement.	126
5.6. Discriminant validity.	127
5.7. CFA model fit indices results.	128
5.8. Structure equation model fit indices results.	129
5.9. Results of Path Analysis.	130

List of Figures

	Page
Chapter 1	1
Chapter 2	
2.1. Standard LSTM architecture	17
2.2 Bidirectional LSTM model	18
2.3 Architecture of Convolutional Neural Networks	20
2.4. Variation of attention weights over different quarters for <i>amendment</i>	28
2.5. Variation of attention weights over different quarters for <i>Audit</i>	29
2.6. Variation of attention weights over different quarters for <i>council</i>	30
2.7. Variation of attention weights over different parts for <i>evasion</i>	31
2.8 Variation of attention weights over different parts for <i>eway</i>	32
2.9. Variation of attention weights over different quarters for <i>askgst_goi</i>	33
2.10. Variation of attention weights over different quarters for <i>filing</i>	34
2.11. Variation of attention weights over different quarters for <i>tax</i>	35
2.12. Variation of attention weights over different quarters for <i>invoice</i>	36
2.13. Variation of attention weights over different quarters for <i>portal</i>	37
2.14 (a). Positive and negative sentiments of key word <i>amendment</i> .	40
2.14 (b). Positive and negative sentiments of key word <i>audit</i> .	40
2.14 (c). Positive and negative sentiments of key word <i>council</i> .	41
2.14 (d). Positive and negative sentiments of key word <i>evasion</i> .	41
2.14 (e). Positive and negative sentiments of key word <i>eway</i> .	42
2.14 (f). Positive and negative sentiments of key word <i>askgst_goi</i> .	43
2.14 (g). Positive and negative sentiments of key word <i>filing</i> .	45
2.14 (h). Positive and negative sentiments of key word <i>tax</i> .	45
2.14 (i). Positive and negative sentiments of key word <i>invoice</i> .	46
2.14 (j). Positive and negative sentiments of key word <i>portal</i> .	47
2.15. Average positive and negative sentiments of ten key words for the period June 2017 – May 2020	48
Chapter 3	
3.1. Flow chart of various steps of the methodology	51
3.2. Co-occurrence graph for E-way Bill : 2017	53
3.3. Co-occurrence graph for E-way Bill : 2018	53
3.4. Co-occurrence graph for E-way Bill : 2019	54
3.5. Co-occurrence graph for Ease of Doing Business : 2017	55
3.6. Co-occurrence graph for Ease of Doing Business: 2018	56
3.7. Co-occurrence graph for Ease of Doing Business: 2019	56
Chapter 4	

4.1. Comparison of similarity scores of predicted answers of November with clustering of questions of October - Euclidean	87
4.2 Comparison of similarity scores of predicted answers for October with clustering of questions of November – Euclidean	89
4.3. Comparison of similarity scores of predicted answers for October with clustering of questions of November – Euclidean.	92
4.4. Comparison of prediction of answers for November with clustering of questions of October - Euclidean	94
4.5. Comparison of prediction of answers for October with clustering of questions of November – Cosine Distance	96
4.6. Comparison of prediction of answers for October with clustering of questions of November – Cosine distance	98
4.7. Comparison of prediction of answers for November with clustering of questions of October - Cosine Distance	99
4.8. Comparison of prediction of answers for November with clustering of questions of October – Cosine Distance	101
4.9. Comparison of prediction of answers for November with important terms selected using IDF – Euclidean	104
4.10. Comparison of prediction of answers for October with important terms selected using IDF – Euclidean	106
4.11. Comparison of prediction of answers for November with important terms selected using IDF – Cosine	108
4.12. Comparison of prediction of answers for October with important terms selected using IDF – Cosine	110
4.13. Count of answers according to similarity scores (SS) using BERT and ROBERTA embeddings – Clustering using Euclidean Distance	111, 112
4.14. Count of answers according to similarity scores (using BERT and ROBERTA embeddings) - Clustering using Cosine Distance	113
4.15 Count of answers according to similarity scores -Clustering of questions using Euclidean and Cosine Distance by selecting important terms on the basis of IDF	115
 Chapter 5	
5.1. Research model with the congenial modifications	118
5.2. Path Analysis of Structural Model	130