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

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 शुरू 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

List of Tables

Chapter 5

List of Figures