IITB-Monash Research Academy

Indian Institute of Technology Bombay | MONASH University

An Indian-Australian research partnership

Project Title:	**Artificial Intelligence for Stronger Indian Democracy: Advances in Machine Learning and Natural Language Processing for Analysing Political Text**	
Project Number	**IMURA0867**	
Monash Main Supervisor (Name, Email, Phone)	Prof. Robert Thomson robert.thomson@monash.edu	*Full name, Email*
Monash Co-supervisor(s) (Name, Email, Phone)	Prof. Wray Buntine Wray.Buntine@monash.edu	
Monash Head of Dept/Centre (Name, Email)	Prof. Dharmalingam Arunachalam dharma.arunachalam@monash.edu	*Full name, email*
Monash Department:	School of Social Sciences and Faculty of Information Technology	
Monash ADGR (Name, Email)	Prof. Rita Wilson Rita.wilson@monash.edu	*Full name, email*
IITB Main Supervisor (Name, Email, Phone)	Prof. Shivaram Kalyanakrishnan shivaram@cse.iitb.ac.in	*Full name, Email*
IITB Co-supervisor(s) (Name, Email, Phone)	Prof. Preethi Jyothi pjyothi@cse.iitb.ac.in	*Full name, Email*
IITB Head of Dept (Name, Email, Phone)	Prof. Umesh Bellur Umesh.bellur@iitb.ac.in	*Full name, email*
IITB Department:	Department of Computer Science and Engineering	

Research Clusters:

Highlight which of the Academy's CLUSTERS this project will address? *(Please nominate JUST **one.** For more information, see www.iitbmonash.org)*	
1	*Material Science/Engineering (including Nano, Metallurgy)*
2	*Energy, Green Chem, Chemistry, Catalysis, Reaction Eng*
3	*Math, CFD, Modelling, Manufacturing*
4	**CSE, IT, Optimisation, Data, Sensors, Systems, Signal Processing, Control**
5	*Earth Sciences and Civil Engineering (Geo, Water, Climate)*
6	*Bio, Stem Cells, Bio Chem, Pharma, Food*
7	*Semi-Conductors, Optics, Photonics, Networks, Telecomm, Power Eng*
8	*HSS, Design, Management*

Research Themes:

Highlight which of the Academy's Theme(s) this project will address? *(Feel free to nominate more than one. For more information, see www.iitbmonash.org)*	
1	***Advanced computational engineering, simulation and manufacture***
2	
3	*Infrastructure Engineering*
4	*Clean Energy*
5	*Water*
6	*Nanotechnology*
7	*Biotechnology and Stem Cell Research*
8	***Humanities and social sciences***
	Design

The research problem

Define the problem

India is the world's largest democracy and is increasingly technologically sophisticated. A prominent feature of modern technologically turbo-charged election campaigns is that parties' communication with voters is increasingly fragmented and targeted. Fragmented and targeted election campaigns with massive flows of information and misinformation pose a challenge for citizens and analysts. For democracy to function effectively, political parties must offer clear choices to voters during election campaigns. However, it is now much harder for citizens to keep track of what parties are promising, which threatens the quality of democratic representation. It also challenges established traditional research methods for studying parties' campaign promises, which are mainly qualitative. New analytical tools based on the Artificial Intelligence (AI) subfields of machine learning and natural language processing have the

potential to strengthen Indian democracy. AI-powered tools enable analysts to examine parties' campaign promises in large amounts of text and speech. This could be of significant benefit to citizens, who will receive greater clarity on the choices that parties are offering. These existing and new methods are highly relevant to research on text and speech in a wide range of social science fields.

Project aims

Define the aims of the project

The project will be supported by and aims to make technological and methodological contributions to two established international research networks. The first research network is the Comparative Party Pledges Project (CPPP), the world's largest comparative research program devoted to the study of campaign promises using mainly established qualitative research methods. The most recent research milestones of the CPPP are the publication of the twelve-country comparative study on the making, breaking and keeping of over 20,000 campaign promises in the form of a book published by the University of Michigan Press (Naurin, Royed and Thomson eds. 2019) and an article in the *American Journal of Political Science* (Thomson, Naurin, Royed et al. 2017), which is the highest-ranked journal by impact factor in political science. The CPPP network currently includes 27 researchers who are studying campaign promises in 20 countries with some of these country-studies still underway, and several additional country-studies being planned. The 20 countries currently included in the CPPP network are: Australia, Austria, Bulgaria, Canada, Denmark, France, Germany, Greece, India, Ireland, Italy, Nepal, the Netherlands, Norway, Portugal, Spain, Sweden, the UK, Ukraine and the USA. The work on campaign promises in India began recently with two PhD projects, one at Monash and one at the IITB-Monash Academy and the IITB Faculty of Humanities and Social Sciences.

The first high-level aim of the project are to engage with the CPPP approach to analysing campaign promises, particularly in the Indian context, and to develop automated AI-powered methods that complement and enrich the CPPP approach.

The second research network is the Mixed-Methods for Analysing Communication (MiMAC) project. MiMAC is a group of Political Scientists and Computer Scientists focused on developing and integrating qualitative and quantitative methods for analysing textual data. Like the present project, MiMAC includes Computer Scientists with a focus on AI, machine learning and NLP, and develops innovative methods for analysing political texts. However, it does not currently include a focus on Indian politics. MiMAC is led by Thomson (Monash) and Prof. Naurin (Gothenburg) and includes Computer Scientists in several universities in Australia (Monash), Europe (Gothenburg, Amsterdam, Bologna, Hertie) and the United States (UCSD). MTAP is supported by a grant of over US$1m from a Swedish Foundation (Riksbankensjubileumfond) starting in January 2020.

The second high-level aim of the project is to engage with MiMAC, particularly with a view to tailoring the technology being developed to the Indian context.

More specific methodological research objectives, which this project will pursue are:

- Develop new methods that automate the identification and comparison of Indian parties' promises and the ways in which they are presented during election campaigns.
- Develop new methods that automate the classification of units of text and speech from Indian election campaigns into policy themes and issues.

In pursuing these aims, the project will advance text analysis methods that have been developed in the field of Natural Language Processing. Broadly speaking, text analysis methods come in two forms: unsupervised exploratory analysis, which discovers structures in the data; and supervised classification, which allocates text in relation to categories or scales that are determined by the analyst.

Unsupervised exploratory methods include topic models, which were introduced by Blei et al. (2013) as Latent Dirichlet Allocation (LDA). The intuition behind LDA is that documents contain several topics to different degrees, and that specific words belong to these topics to different degrees. LDA estimates the probability that groups of documents or words refer to the same or different topics. Topic models are powerful tools for discovering broad trends in the data, and have been successfully used in a wide range of social science research. Topic models uncover a wide range of topics and require careful interpretation and analysis. Methods exist to include external knowledge, either to guide the algorithms to ignore known dead ends, or to include prior beliefs and knowledge about the issues at hand and some of the likely keywords (Jagarlamudi et al., 2012).

Research opportunities in relation to the use of unsupervised exploratory methods include the application of these methods to identify patterns in the words used to characterize promises. This will enable analysts to identify similarities and differences between the campaign promises that are presented by different parties or by the same party at different times.

Supervised text classification methods rely on correlations between certain words or expressions that signal the constructs we are interested in. One of the widely known applications of this method is sentiment analysis (Pang et al. 2002), where the task is to identify whether a text is positive, negative, or neutral. This research has been taken further, to include whether a text is positive, negative, or neutral with respect to a specific target (Jiang et al. 2011).

Recently, a set of methods, called word-embeddings (Mikolov et al. 2013) have gained popularity. These methods project words into a high dimensional space reflecting semantic similarities through the

distance between words. These methods have been shown to capture societal biases, judgments and ideological positions (Bhatia 2017; Garg et al. 2018; Nanni et al. 2018). Word embedding methods are also often deployed for text pre-processing in preparation for natural language processing, particularly deep neural network systems. Embedding approaches have also been developed for multimodal input that combine, for example, textual and visual information.

Neural networks are by far the most successful algorithmic model in the family of supervised text classification methods. However, they require large amounts of data and careful development to prevent social biases (Hovy and Spruit 2016). Bilbao-Jayo and Almeida (2018) used this approach train an algorithm with data on how human coders had previously coded a large amount of text in election manifestos. The system "learned" from this human coding, and was then able to replicate human coders' decisions in a new set of manifestos. There are a wide range of applications, for instance using text classification to identify characteristics of the authors of texts, including personality types (Schwartz et al. 2013; Plank and Hovy 2015) and political preferences (Volkova et al. 2014). Multimodal deep learning systems can integrate information from image data and text data in a single neural network (Sung et al. 2017).

Neural networks-based applications present a vast range of new research opportunities in relation to campaign promises:

- Develop a sentiment analysis method to identify and compare the sentiment in relation to specific campaign promises. This enables researchers to compare the extent to which specific promises are featured positively or negatively in a range of media texts throughout election campaigns.
- Apply and refine the same models to identify election pledges in large amounts of text from parties' manifestos and campaign information.

Expected outcomes

Highlight the expected outcomes of the project including likelihood of patents

The expected outcomes are:

- New evidence and knowledge regarding a key stage of the democratic process in India.
- New AI-powered tools for analysing campaign promises in the Indian context.
- Three or more international peer-reviewed publications on AI-tools for text analysis and campaign promises in Indian politics. We expect most of these publications to be in computer science journals, but one or more may be in a political science journal. The details of the requirements of this multidisciplinary project will be agreed at an early stage of the project, and the Academy is ideally placed for such multidisciplinary research.
- A co-authored book chapter in a volume with a major university press. The CPPP is currently developing plans for a second volume that will include a chapter that compares pledge making and fulfilment in distinct political cultures. There will also be opportunities to contribute to publications with the MiMAC group.
- A Ph.D. thesis.

How will the project address the Goals of the above Themes?

Describe how the project will address the goals of one or more of the 6 Themes listed above.

In the theme of Advanced Computational Engineering, the project contributes by developing and applying new technologies in the Artificial Intelligence subfields of machine learning and natural language processing.

In the theme of Humanities and Social Sciences, this project contributes by developing technology that empowers social scientists to conduct more powerful analyses of Indian democracy.

Capabilities and Degrees Required

List the ideal set of capabilities that a student should have for this project. Be as specific or as general as you like. These capabilities will be input into the online application form and students who opt for this project will be required to show that they can demonstrate these capabilities.

Knowledge of and interest in Indian politics is required. While a formal training in Indian politics and social sciences more generally is desirable, it is not a pre-requisite.

Students should ideally have an undergraduate degree in Computer Science or allied disciplines including

Electrical Engineering. A masters degree would be an advantage. Knowledge of the Artificial Intelligence, machine learning and Natural Language Processing are required.

Potential Collaborators

Please visit the IITB website www.iitb.ac.in OR Monash Website www.monash.edu to highlight some potential collaborators that would be best suited for the area of research you are intending to float.

This project opens opportunities for collaboration with several ongoing research networks outside and inside the IMA:

- The Comparative Party Pledges Project, which is led by researchers at Monash and Gothenburg in Sweden and involves over 20 international research partners across North America and Europe.
- The Mixed Methods for Analysing Communication project, which involves Political Scientists and Computer Scientists at Monash, Gothenburg, Hertie Berlin, Amsterdam and Bologna.
- The AI Horizons project, which includes collaboration between IITB researchers and IBM: http://www.iitb.ac.in/en/story/iit-bombay-and-ibm-team-to-accelerate-ai-research-india
- The machine learning group at Monash: https://www.monash.edu/it/our-research/strengths/data-science/machine-learning

Select up to **(4)** keywords from the Academy's approved keyword list **(available at http://www.iitbmonash.org/becoming-a-research-supervisor/)** relating to this project to make it easier for the students to apply.

Data Science, optimisation, algorithms (6)
Natural Language Processing (29)
Humanities (36)