Should I *eat* it or *wear* it?
Enriching Textual Representation through its Groundings

Somak Aditya & Animesh Mukherjee

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur, India

saditya@cse.iitkgp.ac.in, animeshm@cse.iitkgp.ac.in

1 Introduction

Our goal is to learn textual representation that enable us to reason about functionality and affordability of entities represented in text. Imagine, for a sentence "an apple in the tree", we want to ask the question *what actions can be performed on "apple"* [6]. The *apple* can be eaten, *apple* can be thrown and caught, *apple* can be rolled as it is round. However, for the context "the Apple logo", we cannot eat, catch or throw the "Apple". There can be common actions such as humans can draw pictures of both. In most downstream tasks such as translation and question-answering about text, we need to utilize such information. For example, "The Apple logo is inspired by the apple. Steve Job liked to eat it often". For correct translation, the "it" needs to be resolved to the second "apple". Moreover, there are many useful properties (i.e., affordances) that may not be deducible only from text; for example, to infer the actions *throw* and *catch* for the *apple*, we need to additionally be aware of the ability of humans (and some other living beings) being able to throw and catch some objects.

The problem of object affordances has been well studied in Computer Vision and Robotics [7, 10, 21]. However, infusing such knowledge while learning textual representation is hard; as in NLP, we lack *groundings* of the surface form text. Current state-of-the-art transformers-based large pretrained language models (PTLM) learns contextual representation from massive amounts of unlabeled text and are shown to perform impressively in downstream NLU tasks. In parallel, a growing body of literature shows that PTLMs fail inconsistently and non-intuitively, showing a lack of reasoning and *grounding*. Such inconsistent and unpredictable failure modes of neural networks have prompted DARPA to propose the "AI Next" campaign. In NLP, these results prompted authors in [2] to reiterate the gap between *form* and *meaning*. Using the octopus test as a thought experiment, authors argue that language models which are exposed to only text (surface form) will never truly understand *meaning*, as PTLMs are unaware of true groundings of the surface text. Our goal is to precisely address the gap, that is to enrich textual representation through its groundings in other modalities (especially images and/or videos). Recently, the community has started exploring multimodal and multilingual representations [1]. However, there is no clear consensus about how to reason in the combined embedding space and what can we reason about. In this work, we want to concentrate on enriching entity representations in text, that enable reasoning with functionality and affordability. We take our inspiration from Minsky:

> *. .. it is not enough to classify items of information simply in terms of the features or structures of those items themselves. This is because we rarely use a representation in an intentional vacuum, but we always have goals - and two objects may seem similar for one purpose but different for another purpose. Consequently, we must also take into account the functional aspects of what we know, and therefore we must classify things (and ideas) according to what they can be used for, or which goals they can help us achieve. Two armchairs of identical shape may seem equally comfortable as objects for sitting in, but those same chairs may seem very different for other purposes, for example, if they differ much in weight, fragility, cost, or appearance .... In each functional context we need to represent particularly well the heuristic connections between each object's internal features and relationships, and the possible functions of those objects.*

Our specific goal is not to simply *classify* items, but rather to learn representation that factors in such discriminatory information and reason with such information for downstream tasks such as multi-lingual translation. There has been limited work that focuses on textual representation that encode functionalities of entities or objects represented by the nouns and noun phrases in the text. [21] collects affordance relationships in a structured knowledge graph and shows how reasoning about it can help downstream applications. In Vision and Robotics, authors [3] have explored object recognition through reasoning about its functionalities. However, the scope of such work has been limited to a well-defined closed set of actions and objects [15], and in one high-resource language, i.e., English.

---

[1] https://www.microsoft.com/en-us/research/project/project-florence-vl/
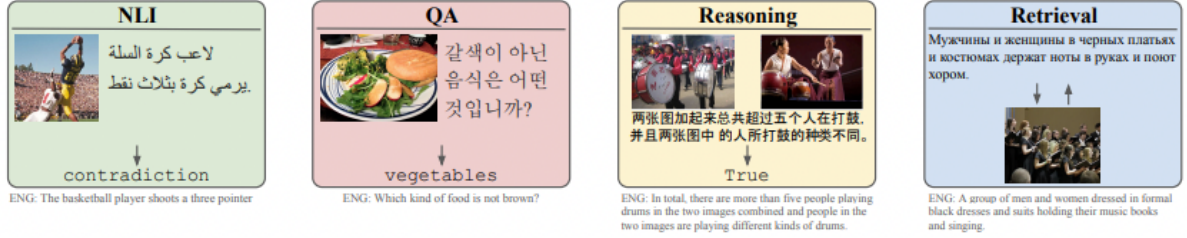
Figure 1: (Image from [4]). Authors collect a set of image-text pairs for 4 NLU tasks over 20 different languages. In these tasks, the text (for example premise in NLI) refers to the image; and hence they are termed as Image-grounded NLU tasks.

2 Project details

Our project is divided into two parts i) <u>learning textual representation that infuses knowledge of entity affordances and functionalities</u>, and ii) <u>reasoning with such representation</u>. We will evaluate such reasoning ability through defining novel auxiliary tasks, and utilizing mainstream NLU tasks.

2.1 Enriching Entity-Centric Textual Representations from its *Groundings*

Thanks to advances in Deep Learning, the NLP community can learn text representation by utilizing vast amounts of unlabeled text, which is abundantly available from the Web. <u>Our goal is to utilize both text and its groundings (in images or videos)</u>. Here, the fundamental challenge is to discover corresponding groundings of text in images or videos. Parallely, a large number of image-text corpora also exists and are often used to test image captioning and visual question answering, such as Flickr8k, Flickr30k [20], MS-COCO [14], and Visual Genome [11]. While such collections are enormous, they suffer from two problems: 1) lack of linguistic diversity (as they are mostly simple written captions by crowdsource workers), and 2) they are all in one language namely English.

Recently, the community has made huge strides in curating multilingual multimodal datasets such as WIT [18], IGLUE [4] (Figure 1). <u>The WIT dataset is based on entire Wikipedia (involving named entities such as person, location, organization).</u> We believe that learning and evaluating such representation is going to be incredibly hard (as human personalities or organizations have many more dimensions than common everyday objects). In our work, we want to focus on <u>"common objects"</u>, therefore focusing on a subset of WIT and similar parallel multilingual datasets as in [8]. <u>IGLUE is an interesting direction where, explicitly image-grounded natural language understanding capability is evaluated</u>. However, in many cases, complimentary situational facts (different from external or commonsense knowledge) are only available in images. In many NLU tasks, this is not the case. We rather require a sort of ontological and commonsense understanding of properties of entities, among which functionality and affordances constitute an important chunk. This is why we look towards <u>enhancing primarily multilingual corpora with affordance annotations</u>. Next we describe the plan for annotation.

Annotation Plan. We plan to <mark>extend multilingual datasets such as XGLUE [12] with affordance annotations and image *groundings*.</mark> Given a piece of text (premise or a paragraph), we will query the set of affordances of each object in the text. We will consider <u>each noun phrase a candidate entity[2]</u>. <u>Given the knowledge base of object affordances by authors in [21] and utilizing the learnt rules in Markov Logic Network, we will predict top 10 candidate affordance relations for each candidate object</u>. Since XGLUE is a collection of multiple datasets, we will use <u>XNLI and PAWS-X</u> (total of 150K), as they are available in most languages. We will also use the <u>parallel dataset by [8]</u>, where 10K words in 10 languages are given alongside images. For a total 800K entities (if we assume 5 entities per example in XGLUE), we will ask the turkers to tick the correct affordances out of the top 10 candidate affordance relations. We will also provide top 5 images, using image retrieval from the parallel image-text corpora (MS-COCO, Visual Genome and IGLUE). The turker will be given USD 0.10 for each HIT. This exercise will amount a total of USD 8k (close to INR 6Lacs). For overheads involving pilots and using multiple turkers, we envision a total of USD 15k (close to 11Lacs) to be the budget for this endeavor.

Continued Multi-objective Pre-training. <u>Instead of learning representations from scratch, our goal is to enrich the learnt representations by current SOTA PTLMs such as RoBERTa$_{LARGE}$, GPT-3, Turing models through using multi-lingual multimodal datasets.</u> In this manner, our technique would be useful to many current universal representation models for text (such as UniLM). During the training, our main technical focus would be to learn robust representation for rare objects/entities and objects that are represented in text by highly polysemous nouns. We also plan to compare with vision-and-language (VL) Transformers architectures. In this case, the objective is to learn a distribution $P_\phi(x_i, v_i)$, that maximizes the joint probability of the correct grounded image patch v_i and a corresponding noun-phrase x_i (or a correct grounded image v and a sentence x). $P_\phi(.,.)$ can be modeled by a function $f_\phi(.,.)$, that can take image and text feature representation as inputs. VL models learns such feature representations

---

[2] Currently we plan to convert named entities into abstract concepts, such as person, and location.

through a single network. However, recent work shows several limiting factors such as capturing less syntactic information than language-only models, over-reliance on linguistic cues. Hence, presently, we will focus on improving language-only Transformers, while encoding image information using vision-only deep learning models (such as ViT, BeiT, ConvNext).

2.2 Reasoning with Entity-centric Textual Representations

During the word2vec era of word representations, an interesting phenomena was being able to do clear vector space arithmetic to solve word analogy tasks. These word analogy tasks often captured interesting hidden dimensions captured by word embeddings (for example city-capital relationships, and gender). As the community moved towards contextual representations, such word analogy tasks were mainly replaced by cloze-style tasks and probing tasks. However, at the word or entity-level, it has been hard to propose tasks that tests for such auxiliary fine-grained properties of the representation. This is partly because of the fact that such pre-trained representation often is learnt in an "*intentional vacuum*", where the *intention* is defined later by the downstream task in the finetuning stage. The PIs of this project would like to take a slightly modified stance. We believe that apart from learning representation of words only *from its context (neighboring words)*, we should enrich the representation through its groundings in the real-world (which defines a more refined *intention* as well). To properly test such an extension, we need to define additional auxiliary tasks for the pretrained embeddings themselves.

Defining Auxiliary Grounding Tasks. In this project, we take our inspiration from [21] (refer Figure 2). We plan to propose a series of increasingly complicated tasks that tests how the learnt representations encode appropriate *groundings* of the entities and the relations mentioned in the text. Here, we will consider the following tasks.

- GRAFFORDQA: Given a piece of text, the task is to answer a natural language question about affordances and functionalities of entities (such as *can apple be rolled*). To this end, we plan to annotate a parallel multi-lingual corpora (as described before), with such object affordances and functionalities by asking specific questions to the crowd-workers, such as "is $\langle noun \rangle_i$ edible?" for each noun phrase in a sentence. We will collect the set of possible functionalities from the Knowledge Base learnt by [21].

- GRANALOGY: The above task can be made more complex by creating different analogy examples along different functionality (or relation) dimensions. An example analogy question would be "apple:watermelon::tennisball:?" for a piece of text "an apple in the tree". We plan to expand the set of analogy questions with Junior-level problem sets from Linguistic Olympiad[3].

- AFFORDANCEKB: As authors in [21] provide an affordability knowledge base, we will use the learnt enriched representation to test link completion accuracy on the knowledge base. In this case, we will convert knowledge graph triplet into sentences. For an edge $\langle u, r, v \rangle$, using sentences from neighboring edges as context, we will mask one of the components of the triplet and predict it.

In short, our goal will be to learn representations that achieve state-of-the-art in the above tasks alongwith the downstream NLU tasks such as GLUE and SuperGLUE benchmarks.

Learning Functionality-Enriched Representation We will create baselines by employing continued pretraining on the proposed parallel multilingual dataset (and dataset in [8]) using supervision from the corresponding tasks. Across many phenomena, researchers have shown that continued pretraining of language models [13] have improved the model's ability to absorb new information (such as commonsense knowledge) while retaining important old linguistic knowledge learnt during pretraining. The PI has explored neuro-symbolic and sub-symbolic representations from his doctoral days. While purely symbolic representation (as in First Order Logic) and probabilistic logical representation (as in Probabilistic Soft logic) have been explored in the context of vision and NLP, success has been limited as coming up with universal representation language is difficult. In fact, this has prompted research in two directions: 1) natural logic and quasi-natural representations (MetaQNL), and 2) vector symbolic architectures (VSAs). Through natural logic and MetaQNL, we can define rules in parametrized natural language.

Provided question-answer annotations as proposed for our dataset creation, commonsense rules over affordances can be learnt in MetaQNL (similar to the ones learnt for a structured knowledge graph in [21]). As such rules are often noisy and incomplete, researchers have proposed different ways to learn from noisy rules. Authors in [1] propose to model a rule's behavior by updating a rule network that learns a constraint over a subset of the data (subset of examples where a rule applies). In parallel, they run a representation network (similar to PTLMs) and optimizes on the combined objectives from original supervision and rule-wise supervision on a subset of data.

Similarly, VSAs define "a set of carefully designed operators on high-dimensional spaces in order to perform symbolic computations with large numerical vectors" [16]. An interesting property of such operators are binding and unbinding operations that closely resembles concepts of compositionality that natural language inherently encodes. We will also encode objectives that are inspired from such operations.

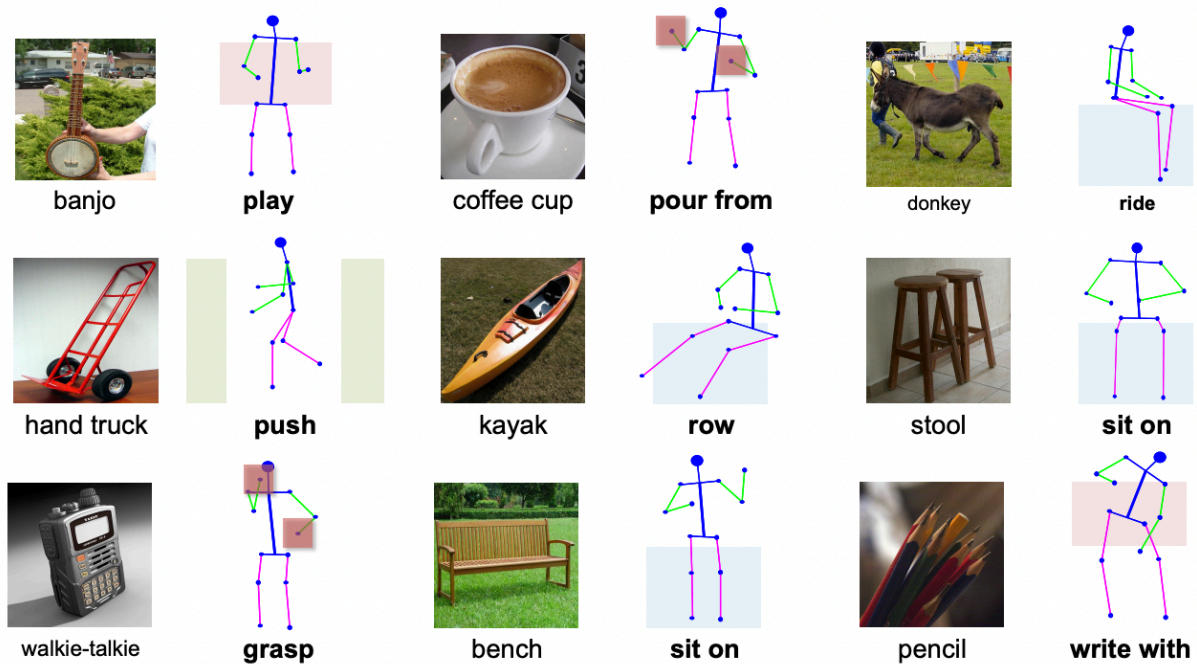---

[3] https://ltrc.iiit.ac.in/nlpmt/plo/#/resource/

Figure 2: (Image from [21]). The figure shows affordances of the corresponding objects in the image, i.e., how a human can interact with the object. In our project, we plan to annotate text with similar functionalities of the entities mentioned in the text.

2.3 Broader Impact

Alongwith the broader academic impact (as described in §1), we also consider Product Impact. Universal Language Models (ULM) such as Turing is expected to become the backbone of many NLP services across Microsoft Products, such as Translation, LUIS API, and even Bing search. It is important that, these *user-facing* systems do not inherit incosnistencies and unpredictable failure modes present in the ULMs. The PI (Dr. Somak Aditya), during his postdoctoral tenure at Microsoft Research India has explored such inconsistencies with his colleagues at Microsoft. During his tenure, he was involved in the LITMUS project (especially the LITMUS deductive part) [17] for robust evaluation of language models. Through LITMUS and tools from other published research work [5, 9, 19], PI and colleagues have discovered many interesting shortcomings across different reasoning dimensions and across languages. Many such shortcomings can be traced back to the lack of *grounding* and proper reasoning. While *grounding* is a hard problem, we aim to solve a part of it by learning robust representation for entities (objects) mentioned in text. We believe, this in turn will benefit the ULM backbones, which will directly benefit the services it feeds to.

2.4 Timeline

1. Multilingual Dataset Annotation with Functionalities – **4 months**

2. Using MetaQNL based Rule-learning and VSA operations Enriching Grounded Representations – **4 months**

3. Experimenting on High-resource and Low-resource languages of interest– **4 months**

2.5 All Other Details

1. Project Area: Learning multilingual Embeddings

2. Microsoft contact: PI Prof. Somak Aditya has done his postdoctoral research in Microsoft Research India under the able supervision of Dr. Monojit Choudhury (presently Principal and Applied Data Scientist, Turing India). Dr. Somak Aditya has also extensively interacted with Dr. Sandipan Dandapat of STCI exploring opportunities of plugging his research into Bing and the offensive team. Currently the effort is being led by Dr. Sunayana Sitaram. Co-PI Prof. Animesh Mukherjee has extensively collaborated with Dr. Monojit Choudhury from 2006.

3. Overheads charged: 20% of the approved budget.

4. Total Budget: INR 25 Lacs (USD 32.8k).

5. Students: Several excellent BTech/MTech students are working with the PI (Abhishek Kumar, BTech Computer Science and Engineering; Manav Kapadnis, Dual Degree, Electrical Engineering). The PI and Co-PI are planning to take a PhD student who will join this autumn (August). Potentially, Sayan Sadhukhan (currently pursuing his MS in UBC, Canada is expected to join us).

Please see Table 1 for the budget breakup for this project.

Categories	Budget (INR)	Comment
Manpower	5 Lacs	Research Scholar
Contingency	11 Lacs	Annotation
Travel	4 Lacs	Conference
Equipment	5 Lacs	GCP Credits
Total	25 Lacs	

Table 1: Overall Budget and Breakup

References

[1] Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, and Sunita Sarawagi. Learning from rules generalizing labeled exemplars. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[2] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics.

[3] Kevin Bowyer, Melanie Sutton, and Louise Stark. Object recognition through reasoning about functionality: A survey of related work. *Object Categorization: Computer and Human Vision Perspectives*, page 129, 2009.

[4] Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulic. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. *CoRR*, abs/2201.11732, 2022.

[5] Pratik Joshi, Aditya, Somak, Aalok Sathe, and Monojit Choudhury. TaxiNLI: Taking a Ride up the NLU Hill. In *CoNLL*, pages 41–55, November 2020.

[6] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.

[7] Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10):1775–1789, 2009.

[8] John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. Learning translations via images with a massively multilingual image dataset. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2566–2576. Association for Computational Linguistics, 2018.

[9] Karthikeyan K, Aalok Sathe, Somak Aditya, and Monojit Choudhury. Analyzing the effects of reasoning types on cross-lingual transfer performance. In *Multilingual Representation Learning Workshop*. EMNLP, Nov 2021.

[10] Hedvig Kjellström, Javier Romero, and Danica Kragic. Visual object-action recognition: Inferring object affordances from human demonstration. *Comput. Vis. Image Underst.*, 115(1):81–90, 2011.

[11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

[12] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online, November 2020. Association for Computational Linguistics.

[13] Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1274–1287. Association for Computational Linguistics, 2021.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[15] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[16] Kenny Schlegel, Peer Neubert, and Peter Protzel. A comparison of vector symbolic architectures. *CoRR*, abs/2001.11797, 2020.

[17] Anirudh Srinivasan, Gauri Kholkar, Rahul Kejriwal, Tanuja Ganu, Sandipan Dandapat, Sunayana Sitaram, Balakrishnan Santhanam, Aditya, Somak, Kalika Bali, and Monojit Choudhury. Litmus predictor: An ai assistant for building reliable, high-performing and fair multilingual nlp systems. In *AAAI Demonstrations 2022*. AAAI, 2022.

[18] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Mike Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, 2021.

[19] Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. LoNLI: An Extensible Framework for Testing Diverse Logical Reasoning Capabilities for NLI. 2021.

[20] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014.

[21] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV (2)*, volume 8690 of *Lecture Notes in Computer Science*, pages 408–424. Springer, 2014.

Somak Aditya

Assistant Professor
Computer Science, IIT Kharagpur
Kharagpur, WB

@ Email: saditya@cse.iitkgp.ac.in
% Website: http://bit.ly/2AxUuhX

RESEARCH INTERESTS

My current interests revolve around developing hybrid statistical and symbolic models that can perform higher-level reasoning by leveraging background knowledge. I am deeply motivated by interdisciplinary efforts and I am a believer of the deep potential impacts of the joint efforts of symbolic reasoning and probabilistic modeling in the field of Artificial Intelligence.

EDUCATION

Doctor of Philosophy, Computer Science   **Arizona State University**, Tempe, AZ
Dissertation Title: Knowledge and Reasoning for Image Understanding   2014-2018
Advisors: Prof. Chitta Baral, Dr. Yezhou Yang
GPA 4.0 (out of 4.0)

Master of Engineering, Computer Science   **Indian Institute of Science**, Bangalore
Dissertation Title: Generic Incremental K-Means Clustering   2009-2011
Advisor: Professor M. Narasimha Murty
GPA 7.3 (out of 8.0), 2nd in Class

Bachelor of Engineering, Computer Science   **Jadavpur University**, Kolkata
CGPA 8.87 (out of 10), 7th in Class   2005-2009

PROFESSIONAL EXPERIENCE

Nov 2021 – Present	Assistant Professor, IIT KGP Computer Science
Feb 2020 – Nov 2021	Postdoctoral Researcher, Microsoft Research India
Sep 2018 – Feb 2020	Research Scientist, Adobe Research India
May 2017 – Aug 2017	Research Intern, Verisk Analytics, New Jersey
May 2015 – Aug 2015	Research Intern, IBM Research Lab, India
Dec 2011 – Jul 2014	Specialist Software III, Strand Life Sciences Pvt. Ltd., India
Jul 2011 – Nov 2011	Senior Software Engineer, Yahoo R&D India

TEACHING EXPERIENCE

Courses Taught, Department of Computer Science, IIT Kharagpur

- CS60092 Information Retrieval   Spring 2021-22
- CS60010 Deep Learning (PGDBA)   Spring 2021-22
- Programming and Data Structures Lab   Spring 2021-22

Teaching Assistant, Department of Computer Science, Arizona State University 2014 – 2016

- CSE-576 Natural Language Processing   Fall 2015 and 2016
- CSE-471 Introduction To Artificial Intelligence   Spring 2016
- CSE-310 Data Structures and Algorithms   Spring 2015

INTERNSHIPS

Cognitive Analytics, Verisk Analytics, New Jersey   May 2017-Aug 2017
Cognitive Analytics and Machine Learning Research Intern
Worked under Dr. Maneesh Singh, director of Image and Video Analytics, JDE, Verisk Analytics. Taking a detour from my PhD research, we explored Generative Adversarial Networks, that can utilize knowledge from pre-trained experts, which is aimed to generate complex images that depict situations (objects, actions and objects interacting through actions). I also assisted in several other projects (Visual Semantic Role Labeling, Situational Image Generation using Blender). I also took part in screening incoming Ph.D. candidates for full-time and intern positions.

IBM India Research Labs, Delhi   May 2015-Aug 2015
Research Intern, Cognitive System Solutions
Worked in Important Event and Entity Detection from Financial News Articles. I successfully created a reasoning framework based on an automatically constructed Knowledge Base of Financial Terms from Investopedia and manually created a Gold-standard for evaluating the system.

ACADEMIC ACHIEVEMENTS

- **Assistantship 2018** The research assistantship is gifted by Verisk Analytics based on the internship.

- **Fellowship 2016, 2017, 2018** Awarded Partial University Graduate Fellowship from Arizona State University. ($2k for Spring-2016, $2.5k for Spring-2017, $4k for Spring-2018)

- **Fellowship 2014** Awarded CIDSE Doctoral Fellowship from Arizona State University. ($15000 for 2014-2015. $30000 in total with assistantships.)

- **Fellowship 2009-11** Awarded MHRD Scholarship for qualifying GATE, 2009 (Rs. 8000 p.m.)

- 🔖 **GATE 2009** AIR (All India Rank) 15th in Graduate Aptitude Test in Engineering (GATE) 2009, entrance exam of the IISc & IITs, in Computer Science & Engineering.
- 🔖 **WBJEE, 2005** Ranked 105th in West Bengal Joint Entrance Examination, 2005 for Engineering Entrance (out of over 50000 candidates)

GRANTS

- 🔖 Google Cloud research credits worth USD 5k by Google Research India.

Patents

1. Somak Aditya and Atanu Sinha. Creating a knowledge graph based on text-based knowledge corpora. 2021 (Filed 2019, *Pending USPTO 16656163*)
2. Somak Aditya, Kushal Chawla, Sharmila Reddy Nangi, Abhinav Mishra, Bhavy Khatri and Pranil Joshi. Predicting joint intent-slot structure. 2021 (Filed 2019, *Pending USPTO 16797164*)

Publications

Journals

[1] Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermüller. "Image Understanding using Vision and Reasoning through Scene Description Graph". In: *Computer Vision and Image Understanding (CVIU)* (2017). ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2017.12.004.

Conference Proceedings

[1] Karthikeyan K, Shaily Bhatt, Pankaj Singh, Somak Aditya, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhary. "Multilingual CheckList: Generation and Evaluation". In: arXiv, 2022. DOI: 10.48550/ARXIV.2203.12865.

[2] Anirudh Srinivasan, Gauri Kholkar, Rahul Kejriwal, Tanuja Ganu, Sandipan Dandapat, Sunayana Sitaram, Balakrishnan Santhanam, Somak Aditya, Kalika Bali, and Monojit Choudhury. "LITMUS Predictor: An AI Assistant for Building Reliable, High-Performing and Fair Multilingual NLP Systems". In: *AAAI Demonstrations 2022*. AAAI. 2022.

[3] Vishesh Agarwal, Somak Aditya, and Navin Goyal. "Analyzing the Nuances of Transformers' Polynomial Simplification Abilities". In: *MATH-AI Workshop*. ICLR. May 2021.

[4] Karthikeyan K, Aalok Sathe, Somak Aditya, and Monojit Choudhury. "Analyzing the Effects of Reasoning Types on Cross-Lingual Transfer Performance". In: *Multilingual Representation Learning Workshop*. EMNLP. Nov. 2021.

[5] Pratik Joshi*, Somak Aditya*, Aalok Sathe*, and Monojit Choudhury. "TaxiNLI: Taking a Ride up the NLU Hill". In: *CoNLL*. Nov. 2020, pp. 41–55.

[6] Somak Aditya and Atanu Sinha. "Uncovering Relations for Marketing Knowledge Representations". In: *StarAI Workshop, AAAI 2020*. 2020.

[7] Somak Aditya, Rudra Saha, Yezhou Yang, and Chitta Baral. "Spatial knowledge distillation to aid visual reasoning". In: *2019 IEEE WACV*. IEEE. 2019, pp. 227–235.

[8] Somak Aditya, Yezhou Yang, and Chitta Baral. "Integrating Knowledge and Reasoning in Image Understanding". In: *IJCAI, 2019*. 2019.

[9] Somak Aditya, Yezhou Yang, and Chitta Baral. "Explicit Reasoning over End-to-End Neural Architectures for Visual Question Answering". In: *AAAI 2018*. 2018.

[10] Somak Aditya, Yezhou Yang, Chitta Baral, and Yiannis Aloimonos. "Combining knowledge and reasoning through probabilistic soft logic for image puzzle solving". In: *UAI 2018*. 2018, pp. 238–248.

[11] Somak Aditya. "Explainable Image Understanding Using Vision and Reasoning". In: *AAAI 2017 Doctoral Consortium*. AAAI Press, 2017, pp. 5028–5029.

[12] Somak Aditya, Chitta Baral, Yezhou Yang, Yiannis Aloimonos, and Cornelia Fermuller. "DeepIU: An Architecture for Image Understanding". In: *Advances of Cognitive Systems*. 2016.

[13] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. "Visual Commonsense for Scene Understanding Using Perception, Semantic Parsing and Reasoning". In: *2015 AAAI Spring Symposium Series*. 2015.

[14] Somak Aditya, Chitta Baral, Nguyen H Vo, Joohyung Lee, Jieping Ye, Zaw Naung, Barry Lumpkin, Jenny Hastings, Richard Scherl, Dawn M Sweet, et al. "Recognizing social constructs from textual conversation". In: *NAACL*. 2015.

[15] Arpit Sharma, Nguyen H Vo, Somak Aditya, and Chitta Baral. "Towards addressing the winograd schema challenge: building and using a semantic parser and a knowledge hunting module". In: *IJCAI*. AAAI Press. 2015, pp. 1319–1325.

[16] Arpit Sharma, Nguyen Vo, Somak Aditya, and Chitta Baral. "Identifying various kinds of event mentions in k-parser output". In: *EVENTS Workshop, NAACL*. 2015, pp. 82–88.

Informal Publications

[1] Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. "Trusting RoBERTa over BERT: Insights from CheckListing the Natural Language Inference Task". 2021.

[2] Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. "LoNLI: An Extensible Framework for Testing Diverse Logical Reasoning Capabilities for NLI". 2021.

[3] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloi-monos. "From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge". 2015.

INVITED TALKS & SEMINARS

Evolution of Representation and Reasoning in the Era of Deep Learning *2022*
Rakuten Institute of Technology

Inspirations from Logic and Language to (Re)Evaluate NLP Systems *2021*
Microsoft Research India Speech & NLP Group

The Pervasiveness of Reasoning from Images to Text *2021*
MILA Philosophy of NLP Reading Group

The Interplay between Deep Learning, Logic, and Reasoning *2021*
CSIR and IEEE SPS Deep Learning Summer School

Explicit Knowledge and Reasoning in the era of Deep Learning *2021*
IIT Hyderabad CSE+AI Faculty Candidate Seminar

Explicit Knowledge and Reasoning in the era of Deep Learning *2021*
IIT Kharagpur CSE+AI Faculty Candidate Seminar

CONFERENCE PRESENTATIONS

TaxiNLI: Taking a ride up the NLU Hill *2020*
CoNLL 2020 Talk

Integrating Knowledge and Reasoning in Image Understanding *2019*
Macau, China, IJCAI 2019 Talk and Poster

Combining Knowledge and Reasoning through Probabilistic Soft Logic for Image Puzzle Solving *2018*
Monterey, California, UAI 2018 Talk and Poster

Explicit Reasoning over End-to-End Neural Architectures *2018*
Hilton, New Orleans, Louisiana, AAAI 2018 Technical Talk

Explicit Reasoning over End-to-End Neural Architectures *2018*
Arizona State University, Tempe, USA, 2018 Southwest Robotics Symposium

Explainable Image Understanding using Vision and Reasoning *2016*
Hilton Union Square, San Fransisco, CA, USA, AAAI 2017 Doctoral Consortium

Visual Commonsense for Scene Understanding Using Perception, *2015*
Semantic Parsing and Reasoning
Stanford University, California, USA, 2015 AAAI Spring Symposium Series

PROFESSIONAL SERVICES

- Organizer

- CVPR 2022: Open-Domain Retrieval Under a Multi-Modal Setting* (proposal accepted). (with colleagues from Arizona State University, FAIR, DeepMind, Microsoft Azure, IDIAP)
- IndoML 2021: Second Edition of Indian Symposium on Machine Learning (with colleagues from IIT Gandhinagar, IIT KGP, Google Research, Univ. Warwick; http://indoml.in/).
- CIKM 2021: Knowledge Injection in Neural Networks (with colleagues from Intel Labs, University of College London, Arizona State University; https://sites.google.com/view/kinn2021/)
- IJCAI 2021: Is Neuro-symbolic SOTA still a myth for Natural Language Inferencing (with colleagues from IBM Research, KU Leuven, MSR India, UT Austin; https://nsnli.github.io/)
- KR 2018: Integrating learning of Representations and models with deductive Reasoning that leverages Knowledge (https://sites.google.com/view/r2k2018/home)

- Panelist: IJCAI 2019 Doctoral Consortium Career Panel
- PC Member: EMNLP, ACL, AAAI, IJCAI (2020, 2021, 2022); NAACL, EACL (2021), IEEE MASS 2022
  - Workshops: MathAI4ED Workshop NeurIPS 2021, MathAI Workshop ICLR 2021, LANTERN-COLING (2020, 2021), Cognitive Vision 2019 (ACS at MIT)
- Conference Reviewer: ICRA (2020), IJCAI (2017, 2016)
- Journal Reviewer: IEEE TIP, AIJ, CVIU, The Visual Computer, Robotics and Autonomous Systems (RAS), Pattern Recognition, Neurocomputing

PROJECTS (During PostDoc)

Natural Language Inferencing: Evaluation and Enhancement
Advisors: Dr. Monojit Choudhury   Microsoft Research India
The aim is to propose fine-grained evaluation of Natural Language Inference systems across well-defined Linguistic and Logical capabilities. Through extensive evaluation, we aspire to expose if Neural Network-based systems trained using the current data-driven paradigm is fundamentally insufficient across certain reasoning capabilities, and whether Neuro-symbolic mechanisms can step in to fill the void.

Interactive Evaluation and Explanation
Collaborators: Dr. Monojit Choudhury, Dr. Amit Deshpande   Microsoft Research India
Interactive prover systems provide a framework where verifier and prover interact to find a proof. Such interactive protocols may inspire human-machine interactions to find related errors of language models based on a set of given errors, behavioral performance (such as CheckList report-card) and the model API. Similar interaction can be enabled to find post-hoc example-based explanations.

Symbolic Reasoning abilities of Deep Neural Networks
Collaborators: Dr. Navin Goyal   Microsoft Research India

Automated Theorem Proving and symbolic Mathematics tasks presents very specific well-defined challenges. For example, a step in an integration may involve finding a sub-expression to operate, finding the next operation, and executing the operation on a mathematical expression to output new expression. Transformers and Graph Neural Nets have performed impressively, but there is to much to learn about its strengths and shortcomings. The goal here is to extensively evaluate the state-of-the-art architectures on synthetic symbolic tasks and propose informed enhancements based on the evaluation.

PROJECTS (During PhD)

Understanding images and videos using a combination of Computer Vision, Statistical and Logical Reasoning
Advisors: Prof. Chitta Baral, Dr. Yezhou Yang   Arizona State University
Collaborator: Prof. Yiannis Aloimonos   University of Maryland, College Park
i) Solving Image Riddles using Vision and Reasoning, ii) Visual Question Answering using Deep Learning and Probabilistic Soft Logic, iii) Image Caption Generation Through Knowledge Graphs from Images, iv) Integrating Spatial Knowledge for Visual Reasoning

Compositional and Generative modeling of Images from text
Advisors: Dr. Yezhou Yang   Arizona State University
Collaborator: Dr. Maneesh Kr. Singh   Verisk Analytics, New Jersey
The goal of the project is to learn a generative model of natural images that abides by properties of compositionality with respect to an input condition (text or features).

Building a Knowledge Parser for Formal Representation of English Text
Advisor: Prof. Chitta Baral   Arizona State University
Parsing Natural Language Text and Creating a Knowledge Graph using a combination of Rule-based and Machine Learning Techniques.

Integrating Statistical and Logical Approaches in Recognizing Social Constructs from Textual Conversation
Advisor: Prof. Chitta Baral   Arizona State University
Our goal was to recognize high-level social constructs such as Leadership and Status from textual conversation using an approach that integrates statistical methods and symbolic logic based methods.

REFERENCES

Prof. Chitta Baral
Professor & Chair
Arizona State University.
✉ chitta@asu.edu

Prof. Yezhou Yang
Assistant Professor
Arizona State University.
✉ yz.yang@asu.edu

Prof. Yiannis Aloimonos
Professor
University of Maryland, College Park.
✉ yiannis@cs.umd.edu

Dr. Monojit Choudhury
Principal Researcher
Microsoft Research India
✉ monojitc@microsoft.com

CV of Animesh Mukherjee

1. Name of the Nominee: Animesh Mukherjee

2. Current designation and affiliation: Associate Professor

3. Area of Specialization: Computer Science and Engineering

4. Current Nationality: Indian

5. Date of Birth: 21/12/1979

6. Gender: Male ✓   Female   ☐

7. Contact Address:

Tel.__03222283472_____Mobile__9433124004_____FAX____

Email___animeshm@gmail.com_____

8. Educational Qualifications (from Bachelor degree onwards)

S.No.	Degree	University	Year	Subjects	Percentage/GPA
1	**BTech**	**VU**	**2003**	**CSE**	**89.8 (Rank 1)**
2	**MTech**	**IITKGP**	**2005**	**CSE**	**9.43 (Rank 4)**
3	**PhD**	**IITKGP**	**2009**	**CSE**	**NA**

9. Details of professional training and research experience:

S.No.	Place of Training	Position	Period	Country
1	**ISI Foundation**	**PostDoc Researcher**	**2009-2010**	**Italy**

10. Details of employment/position held in last 10 years:

S.No.	Employed at	Position	Period (From-To)	Country
1	**IIT Kharagpur**	**Asst. Prof**	**2011 - 2016**	**India**
2	**IIT Kharagpur**	**Asso. Prof**	**2016 -**	**India**

11. Professional awards/recognition/fellowship (e.g. Swarnjayanti/Bhatnagar Prize, Fellow of INAE, INSA etc.):

S.No.	Name	Year	Donor organization	
1.	**ICADL 2021 Best student paper award**	**2021**	**ICADL**	
2.	**Ted Nelson Newcomer Best Paper Award**	**2021**	**ACM**	
3.	**A K Singh Chair**	**2020**	**IIT Foundation, USA**	
4.	**Data grant of 1trillion URLs**	**2020**	**Facebook**	

5.	**Ethics in AI research award**	**2019**	**Facebook**	
6.	**Best Paper Honorable Mention**	**2019**	**ACM WebSci**	
7.	**Google course award**	**2019**	**Google**	
8.	**Humboldt fellowship for experienced researchers**	**2017**	**Humboldt**	
9.	**GYTI award**	**2017**	**SRISTI**	
10.	**Best Paper Honorable Mention**	**2016**	**ACM CSCW**	
11.	**IBM Faculty award**	**2015**	**IBM**	
12.	**INSA YS Medal**	**2014**	**INSA**	
13.	**Simons Associate**	**2013**	**ICTP**	
14.	**INAE YE award**	**2012**	**INAE**	
15.	**MSR PhD fellowship**	**2006**	**MSR**	
16.	**ISCA YS award**	**2006**	**ISCA**	

12. List of peer reviewed publications. I am listing the Conferences before the Journals since in Computer Sciences publications in top tier conferences are often more prestigious than publications in Journals.

Conferences

1. Dash, A., Chakraborty, A., Ghosh, S, *Mukherjee, A.* and Gummadi, K. (2022). Alexa, in you, I trust! Fairness and Interpretability Issues in E-commerce Search through Smart Speakers. In *The WebConf 2022*, Virtually (online). **[CORE RANK: A*]**
2. Das, M., Banerjee, S., and *Mukherjee, A.* (2022). Data Bootstrapping Approaches to Improve Low Resource Abusive Language Detection for Indic Languages .In *ACM Hypertext 2022*, Bacelona, Spain.
3. Dutta, P., Chakraborty, S., Roychowdhury, S. and *Mukherjee, A.* (2022). CRUSH: Contextually Regularized and User anchored Self-supervised Hate speech Detection. In *NAACL 2022 (Findings)*, Seattle, Washington. **[CORE RANK: A*]**

4. Das, M., Saha, P., Mathew, B. and *Mukherjee, A.* (2022). HateCheckHIn: Evaluating Hindi Hate Speech Detection Models. In *LREC 2022*, Marseille, France.

5. Jaiswal, S. and *Mukherjee, A.* (2022). Marching with the Pink Parade: Evaluating Visual Search Recommendations for Non-binary Clothing Items. In *CHI 2022* (case studies), New Orleans and online. **[CORE RANK: A*]**

6. Jaiswal, S., Duggirala, K., Dash, A. and *Mukherjee, A.* (2022). Two-Face: Adversarial Audit of Commercial Face Recognition Systems. In *ICWSM 2022*, Atlanta, Georgia and online.

7. Das, P., Seelaboyina, S. B., Reddy, B. P., Sarkar, S. and *Mukherjee, A.* (2022). Quality Change: norm or exception? Measurement, Analysis and Detection of Quality Change in Wikipedia. In *CSCW 2022*, Virtually (online mode). **[CORE RANK: A]**

8. Chowdhury, A., Srinivasan, S., Bhowmick, S., *Mukherjee, A.* and Ghosh, K. (2021). Constant Community Identification in Million Scale Networks Using Image Thresholding Algorithms. In *ASONAM 2021*, (short paper), Virtually (online mode).

9. Nag, A., Samanta, B., *Mukherjee, A.*, Ganguly, N. and Chakrabarti, S. (2021). A Data Bootstrapping Recipe for Low-Resource Multilingual Relation Classification. In *CoNLL 2021*, Virtually (online mode). **[CORE RANK: A]**

10. Das, P., Reddy, B. P., Chakraborty, D., Sarkar, S. and *Mukherjee, A.* (2021). When expertise gone missing: Uncovering the loss of prolific contributors in Wikipedia. In *ICADL 2021*, Virtually (online mode). (**Best student paper award**).

11. Das, M., Saha, P., Dutt, R., Goyal, P., *Mukherjee, A.* and Mathew, B. (2021). You too Brutus! Trapping Hateful Users in Social Media: Challenges, Solutions and Insights. In *The Hypertext 2021*, Virtually (online mode). (**Ted Nelson best newcomer paper award**).

12. Bansal, S., Garimella, V., Suhane, A. and *Mukherjee, A.* (2021). Debiasing Multilingual Word Embeddings: A Case Study of Three IndianLanguages. In *Hypertext 2021*, Virtually (online mode).

13. Saha, P., Mathew, B., Garimella, K. and *Mukherjee, A.* (2021). "Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups. In *The WebConf 2021*, Ljubljana, Slovenia. **[CORE RANK: A*] Media Coverage:** The Wire, ToI, The Print, ClarionIndia, The Siasat Daily, India Today, Newsmeter, Sabrang India.

14. Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P. and *Mukherjee, A.* (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *AAAI 2021*, Virtually (online mode).**[CORE RANK: A*]**

15. Dash, A., Chakraborty, A., Ghosh, S., *Mukherjee, A.* and Gummadi, K. P. (2021). When the Umpire is also a Player: Bias in Private Label Product Recommendations on E-commerce Marketplaces. In *ACM FAccT 2021*, Virtually (online mode).

16. Hazra, R., Aggarwal, H., Goyal, P., *Mukherjee, A.* and Chakrabarti, S. (2021). Joint Autoregressive and Graph Models for Software and Developer Social Networks. In *ECIR 2021*, Virtually (online mode).**[CORE RANK: A]**

17. Reddy, B. P., Seelaboyina, S. B., Sarkar, S. and *Mukherjee, A.* (2020). NwQM: A neural quality assessment framework for Wikipedia. In *EMNLP 2020*, Virtually (online mode).

18. Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P. and *Mukherjee, A.* (2020). Hate begets Hate: A Temporal Study of Hate Speech. In *ACM CSCW 2020*, Virtually (online mode). **[CORE RANK: A]**

19. Aluru, S. S., Mathew, B., Saha, P. and *Mukherjee, A.* (2020). A Deep Dive into Multilingual Hate Speech Classification. In *ECML-PKDD 2020*, Ghent Belgium. **[CORE RANK: A]**

20. Atharva Vyas, Heer Ambavi, Sayantan Adak, Shivam Patel, Pritam Kadasi, Mayank Singh and Animesh Mukherjee (2020). Gandhipedia: A one-stop AI-enabled portal for browsing Gandhian literature, life-events and his social network. In *JCDL 2020* (demo track), Xi'an, China. **[CORE RANK: A*]**

21. Bansal, S., Garimella, V., Suhane, A., Patro, J. and *Mukherjee, A.* (2020). Code-switching patterns can be an effective route to improve performance of downstream NLP applications: A case study of humour, sarcasm and hate speech detection. In *ACL 2020* (short paper), Seattle, Washington. **[CORE RANK: A*]**

2. Chakraborty, S., Goyal, P. and *Mukherjee, A.* (2020). Aspect-based Sentiment Analysis of Scientific Reviews. In *JCDL 2020*, Xi'an, China. **[CORE RANK: A*]**

3. Hazra, R., Aryan, Aggarwal, H., Marsili, M. and *Mukherjee, A.* (2020). Characterising authors on the extent of their paper acceptance: A case study of the Journal of High Energy Physics, In *JCDL 2020*, Xi'an, China. **[CORE RANK: A*]**

4. Jalal, J., Singh, M., Pal, A., Dey, L. and *Mukherjee, A.* (2020). Identification, Tracking and Impact: Understanding the trade secret of catchphrases. In *JCDL 2020*, Xi'an, China. **[CORE RANK: A*]**

5. Mathew, B., Kumar, N., Goyal, P. and *Mukherjee, A.* (2020). Interaction dynamics between hate and counter users on Twitter. In *CoDS-COMAD (20)*, (research track), Hyderabad, India.

6. Mathew, B., Maity, S. K., Goyal, P. and *Mukherjee, A.* (2020). Competing Topic Naming Conventions in Quora: Predicting Appropriate Topic Merges and Winning Topics from Millions of Topic Pairs. In *CoDS-COMAD (20)*, (research track), Hyderabad, India.

7. Singh, M., Pal, A., Dey, L. and *Mukherjee, A.* (2020). Innovation and Revenue: Deep Diving into the Temporal Rank-shifts of Fortune 500 Companies. In *CoDS-COMAD (20)*, (industry track), Hyderabad, India.

8. Patro, J., Bansal, S. and *Mukherjee, A.* (2019). A deep-learning framework to detect sarcasm targets. In *EMNLP (19)*, (short paper), Hong Kong. **[CORE RANK: A]**

9. Hazra, R., Singh, M., Goyal, P., Adhikari, B. and *Mukherjee, A.* (2019). The rise and rise of interdisciplinary research: Understanding the interaction dynamics of three major fields -- Physics, Mathematics & Computer Science. In *ICADL (19)*, Kuala Lumpur, Malayasia. **[CORE RANK: A]**

10. Mathew, B., Dutt, R., Maity, S. K., Goyal, P. and *Mukherjee, A.* (2019). Deep Dive into Anonymity: Large Scale Analysis of Quora Questions. In *SocInfo (19)*, Doha, Qatar.

11. Patro, J., Baruah, S., Gupta, V., Choudhury, M., Goyal, P. and *Mukherjee, A.* (2019). Characterizing the spread of exaggerated health news content over social media. In *ACM Hypertext and Social Media (19)*, (poster paper), Hof, Germany. **[CORE RANK: A]**

12. Jana, A., Puzyrev, D., Panchenko, A., Goyal, P., Biemann, C. and *Mukherjee, A.* (2019). On the Compositionality Prediction of Noun Phrases using Poincaré embeddings. In *ACL (19)*, Florence, Italy. **[CORE RANK: A*]**

13. Sarkar, S., Reddy, B. P., Sikdar, S. and *Mukherjee, A.* (2019). StRE: Self Attentive Edit Quality Prediction in Wikipedia. In *ACL (19)*, Florence, Italy. **[CORE RANK: A*]**

14. Matthew, B., Dutt, R., Goyal, P. and *Mukherjee, A.* (2019). Spread of hate speech in online social media. In *ACM WebSci (19)*, Boston, USA. (**Best Paper, Honourable Mention Award**).

15. Patro, J., Choudhary, N., Chittora, K. and *Mukherjee, A.* (2019). A deep learning approach to detect emotions in the dialog utterances. In *SemEval (19)*, (poster paper), Minneapolis, USA.

16. Matthew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal P. and *Mukherjee, A.* (2019). Thou shalt not hate: Countering online hate speech. In *ICWSM (19)*, Munich, Germany.

17. Singh, M., Sarkar, R., Vyas, A., Goyal, P., *Mukherjee, A.* and Chakrabarti, S. (2019). Automated Early Leaderboard Generation From Comparative Tables. In *ECIR (19)*, Cologne, Germany. **[CORE RANK: A]**

18. Maity, S.K., Panigrahi, A., Ghosh, S., Banerjee, A., Goyal, P. and *Mukherjee, A.*. (2019). DeepTagRec: A Content-cum-User based Tag Recommendation Framework for Stack Overflow. In *ECIR (19)*, (short paper), Cologne, Germany. **[CORE RANK: A]**

19. Dash, A., *Mukherjee, A.* and Ghosh, S. (2019). A Network-centric Framework for Auditing Recommendation Systems. In *IEEE Infocom (19)*, Paris, France. (**Best In-session Presentation Award**). **[CORE RANK: A*]**

20. Jana, A., *Mukherjee, A.* and Goyal. P. (2019). Detecting Reliable Novel Word Senses: A Network-Centric Approach. In *ACM SAC (19) -- Knowledge and Language Processing track*, Limassol, Cyprus.

21. Maity, S. K., Chaudhuri, A. and *Mukherjee, A.* (2019). "Woman-Metal-White vs Man-Dress-Shorts": Combining Social, Temporal and Image Signals to Understand Popularity of Pinterest Fashion Boards. In *ICWSM (19)*, Munich, Germany.

22. Santosh, T. Y. S. S., Gambhir, V. and *Mukherjee, A.* (2018). Deep learning for social media health text classification. In *Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task, EMNLP(18)*, Brussels, Belgium.

23. Maity, S. K., Chakraborty, A., Goyal, P. and *Mukherjee, A.* (2018). Opinion Conflicts: An Effective Route to Detect Incivility in Twitter. In *CSCW (18)*, New York City's Hudson River. **[CORE RANK: A]**

24. Sarkar, S., Bhowmick, S. and *Mukherjee, A.* (2018). On Rich Clubs of Path-Based Centralities in Networks. In *CIKM (18)*, Lingotto, Turin, Italy. **[CORE RANK: A]**

25. Sarkar, S., Bhagwat, A. and *Mukherjee, A.* (2018). Core2Vec: A core-preserving feature learning framework for networks. In *IEEE/ACM ASONAM(18)*, Barcelona, Spain. (short paper)

26. Joshi, T., *Mukherjee, A.* and Ippadi, G. (2018). One size does not fit all: Predicting product returns in e-commerce platforms In *IEEE/ACM ASONAM(18)*, Barcelona, Spain. (poster)

27. Jana, A., Kanojiya, P., Goyal, P., and *Mukherjee, A.* (2018). WikiRef: Wikilinks as a route to recommending appropriate references for scientific Wikipedia pages. In *COLING (18)*, Santa Fe, New Mexico, USA. **[CORE RANK: A]**

28. Maity, S. K., Mullick, A., Ghosh, S., Kumar, A., Dhamnani, S., Bahety, S., and *Mukherjee, A.* (2018). Understanding Psycholinguistic Behavior of Predominant Drunk Texters in Social Media. In *IEEE ISCC Workshops - ICTS4eHealth (18)*, Natal, Brazil.(**BEST PAPER AWARD**)

29. Singh, M., Dogga, P., Patro, S., Barnwal, D., Dutt, R., Haldar, R., Goyal, P. and *Mukherjee, A.* (2018). <u>CL Scholar</u>: The ACL Anthology Knowledge Graph Miner. In *NAACL (18) demo track*, New Orleans, Louisiana. **[CORE RANK: A]**

30. Sikdar, S., Chakraborty T., Sarkar, S., Ganguly, N. and *Mukherjee, A.* (2018). ComPAS: Community Preserving Sampling for Streaming Graphs. In *AAMAS (18)*, Stockholm, Sweden. **[CORE RANK: A*]**

31. Maity, S. K., Gajula, R. and *Mukherjee, A.* (2018). Why Did They #Unfollow Me? Early Detection of Follower Loss on Twitter. In *ACM GROUP (18)*, (poster highlight), Florida, USA.

32. Maity, S. K., Kumar, A., Choudhary, V., Mallick, A., and *Mukherjee, A.* (2018). Understanding Book Popularity on Goodreads. In *ACM GROUP (18)*, (poster highlight), Florida, USA.

33. Patro, J., Samanta, B., Singh, S., Basu, A., Mukherjee, P., Choudhury, M., *Mukherjee, A.* (2017). All that is English may be Hindi: Enhancing language identification through automatic ranking of the likeliness of word borrowing in social media. In *EMNLP (17)*, Copenhagen, Denmark. **[CORE RANK: A]**

34. Singh, M., Sarkar, R., Goyal, P., *Mukherjee, A.* and Chakrabarti, S. (2017). Relay-Linking Models for Prominence and Obsolescence in Evolving Networks. In *ACM SIGKDD (17)*, Halifax, Canada. **[CORE RANK: A*]**

35. Maity, S. K., Panigrahi, A., and *Mukherjee, A.* (2017). Book Reading Behavior on Goodreads Can Predict the Amazon Best Sellers. In *ACM/IEEE ASONAM (17)*, (short paper), Sydney, Australia.

36. Singh, M., Dan, S., Agarwal, S., Goyal, P. and *Mukherjee, A.* (2017). AppTechMiner: Mining Applications and Techniques from Scientific Articles. In *ACM/IEEE JCDL 2017* workshop *WOSP (17)*, Toronto, Canada.

37. Mathew, B., Maity, S. K., Sarkar, P., *Mukherjee, A.* and Pawan Goyal (2017). Adapting predominant and novel sense discovery algorithms for identifying corpus-specific sense differences. In *ACL 2017* workshop *TextGraphs-11*, Vancouver, Canada.

38. Sikdar, S., Marsili, M., Ganguly, N., *Mukherjee, A.* (2017). Influence of Reviewer Interaction Network on Long-term Citations: A Case Study of the Scientific Peer-Review System of the Journal of High Energy Physics. In *ACM/IEEE JCDL (17)*, Toronto, Canada. **[CORE RANK: A*]**

39. Jana, A., Mooriyath, S., *Mukherjee, A.*, and Goyal, P. (2017). WikiM: Metapaths based Wikification of Scientific Abstracts. In *ACM/IEEE JCDL (17)*, Toronto, Canada. **[CORE RANK: A*]**

40. Singh, M., Niranjan A., Gupta, D., Bakshi, N. A., *Mukherjee, A.*, and Goyal, P. (2017). Citation sentence reuse behavior of scientists: A case study on massive bibliographic text dataset of computer science. In *ACM/IEEE JCDL (17)*, Toronto, Canada (short paper). **[CORE RANK: A*]**

41. Singh, M., Jaiswal, A., Priya Shree, Pal, A., *Mukherjee, A.*, and Goyal, P. (2017). Understanding the Impact of Early Citers on Long-Term Scientific Impact. In *ACM/IEEE JCDL (17)*, Toronto, Canada. **[CORE RANK: A*]**

42. Maity, S. K., Kharb, A. and *Mukherjee, A.* (2017). Language Use Matters: Analysis of the Linguistic Structure of Question Texts can Characterize Answerability in Quora In *ICWSM (17)*, (poster paper), Montreal, Canada.

43. Maity, S. K., Chakraborty, A., Goyal, P. and ***Mukherjee, A.*** (2017). Detection of sockpuppets in social media. In *CSCW (17) poster highlight*, Portland, OR.

44. Maity, S. K., Jha, C., Kumar, A., Sengupta, A., Modi, M. and ***Mukherjee, A.*** (2017). A Large-scale Analysis of the Marketplace Characteristics in Fiverr, In *HICSS-50*, Hilton Waikoloa Village, Hawaii. **[CORE RANK: A]**

45. Singh, M., Barua, B., Palod, P., Garg, M., Satapathy, S., Bushi, S., Ayush, K., Rohith, K. S., Gamidi, T., Goyal, P. and ***Mukherjee, A.*** (2016) OCR++: A Robust Framework For Information Extraction from Scholarly Articles. In *COLING (16)*, Osaka, Japan. Live System **[CORE RANK: A]**

46. Ufimtsev, V., Sarkar, S., ***Mukherjee, A.*** and Bhowmick, S. (2016). Understanding Stability of Noisy Networks through Centrality Measures and Local Connections, In *CIKM (16)*, (extended short paper), Indianapolis, IN. **[CORE RANK: A]**

47. Sikdar, S., Marsili, M, Ganguly, N. and ***Mukherjee, A.*** (2016). Anomalies in the peer-review system: A case study of the journal of High Energy Physics, In *CIKM (16)*, (extended short paper), Indianapolis, IN. **[CORE RANK: A]**

48. Sarkar, S., Kumar, S., Bhowmick, S., and ***Mukherjee, A.*** (2016). Sensitivity and Reliability in Incomplete Networks: Centrality Metrics to Community Scoring Functions, In *ASONAM (16)*, (short paper), San Francisco, CA. (Extended version invited as a book chapter in *Lecture Notes on Social Networks*, Springer).

49. Maity, S. K., Sarda, C., Chaudhary, A., Patil, A., Kumar, S., Mondal, A. and ***Mukherjee, A.*** (2016). WASSUP? LOL : Characterizing Out-of-Vocabulary Words in Twitter. In *CSCW (16) poster highlight*, San Francisco, CA. **[CORE RANK: A]**

50. Sikdar, S., Chakraborty, A., Chaudhary, A., Kumar, G., Kumar, S., Patil, A., Ganguly, N. and ***Mukherjee, A.*** (2016). Identifying and Characterizing Sleeping Beauties on YouTube. In *CSCW (16) poster highlight*, San Francisco, CA. **[CORE RANK: A]**

51. Prasad, U., Kumari, N., Kumar, M., Ganguly, N. and ***Mukherjee, A.*** (2016). The role of outsiders in consensus formation: A case study of Yelp. In *CSCW (16) poster highlight*, San Francisco, CA. **[CORE RANK: A]**

52. Maity, S. K., Saraf, R. and ***Mukherjee, A.*** (2016). #Bieber + #Blast = #BieberBlast: Early prediction of popular hashtag compounds. In *CSCW (16)*, San Francisco, CA. (CSCW Honorable Mention, MIT Tech Review best of the rest of arXiv. Synopsis. Other Comments.) **[CORE RANK: A]**

53. Chakraborty, T., Krishna, A., Singh, M., Ganguly, N., Goyal, P. and ***Mukherjee, A.*** (2016). FeRoSA: A Faceted Recommendation System for Scientific Articles. In *PAKDD (16)*, Auckland, New Zealand. (Working Prototype) **[CORE RANK: A]**

54. Maity, S. K., Ghuku, B., Upmaynu, A. and ***Mukherjee, A.*** (2016). Out of vocabulary words decrease, running texts prevail and hashtags coalesce: Twitter as an evolving sociolinguistic system. In *HICSS (16)*, Kauai, Hawaii. (This paper featured as MIT Tech Review best of the rest of arXiv) **[CORE RANK: A]**

55. Singh, M., Patidar, V., Kumar, S., Chakraborty, T., ***Mukherjee, A.*** and Goyal, P. (2015). The role of citation context in predicting long-term citation profiles: an experimental study based on a massive bibliographic text dataset. In *CIKM (15)*, Melbourne, Australia. **[CORE RANK: A]**

56. Chakraborty, T., Patranabis, S., Goyal, P. and *Mukherjee, A.* (2015). On the formation of circles in co-authorship networks. In *KDD(15)*, Sydney, Australia. (Video) **[CORE RANK: A*]**

57. Singh, M., Chakraborty, T., Goyal, P. and *Mukherjee, A.* (2015). ConfAssist: A Conflict resolution framework for assisting the categorization of Computer Science conferences. In *JCDL (15)*, Tennessee, USA (poster). **[CORE RANK: A]**

58. Maity, S. K., Singh, J. S. and *Mukherjee, A.* (2015). Analysis and prediction of question topic popularity in community QA sites: A case study of Quora. In *ICWSM (15)*, Oxford, UK.

59. Maity, S. K., Gupta, A., Goyal, P. and *Mukherjee, A.* (2015). A stratified learning approach for predicting the popularity of Twitter Idioms. In *ICWSM (15)*, Oxford, UK (poster).

60. Sikdar, S., Bodych, M., Maiti, R. R., Paria, B., Ganguly, N., Krueger, T. and *Mukherjee, A.* (2015). On the broadcast of segmented messages in dynamic networks. In *IEEE NetSciCom(15)* (co-located with *IEEE INFOCOM*), Hong Kong, China.

61. Chakraborty, T., Ganguly, N. and *Mukherjee, A.* (2014). Automatic Classification of Scientific Groups as Productive: An Approach based on Motif Analysis. In *ASONAM(14)*, China.

62. Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N. and *Mukherjee, A.* (2014). Towards a Stratified Learning Approach to Predict Future Citation Counts. In *JCDL(14)*, UK. **[CORE RANK: A*]**

63. Chakraborty, T., Srinivasan, S., Ganguly, N., *Mukherjee, A.* and Bhowmick, S. (2014). On the permanence of vertices in network communities. In *KDD(14)*, New York, USA. **[CORE RANK: A*]**

64. Chakraborty, T., Tammana, V., Ganguly, N. and *Mukherjee, A.* (2014). Analysis and modeling of lowest unique bid auctions. In *ASE Socialcom(14)*, Stanford, USA.

65. Maity, S. K., Manoj, T. V. and *Mukherjee, A.* (2014). Opinion dynamics in correlated time-varying social networks. In *ASE Socialcom(14)*, Stanford, USA.

66. Mitra, S., Mitra, R., Riedl, M., Biemann, C., *Mukherjee, A.* and Goyal, P. (2014). That's sick dude!: Automatic identification of word sense change across different timescales. In *ACL(14)*, Baltimore, USA. **[CORE RANK: A*]**

67. Quasthoff, U., Mitra, R., Mitra, S., Eckart, T., Goldhahn, D., Goyal, P. and *Mukherjee, A.* (2014). Large Web Corpora of High Quality for Indian Languages. LREC Workshop on *Indian Language Data: Resources and Evaluation*, Reykjavik, Iceland (poster).Maity, S. K., Porwal, A. and *Mukherjee, A.* (2013). Understanding how learning affects agreement process in social networks, In *ASE/IEEE Socialcom(13)*, Washington DC, USA.

68. Chakraborty, T., Kumar, S., Reddy, M. D., Kumar, S., Ganguly, N. and *Mukherjee, A.* (2013). Automatic Classification and Analysis of Interdisciplinary Fields in Computer Sciences, In *ASE/IEEE Socialcom(13)*, Washington DC, USA.

69. Chakraborty, T., Sikdar, S., Tammana, V., Ganguly, N. and *Mukherjee, A.* (2013). Computer Science Fields as Ground-truth Communities: Their Impact, Rise and Fall, In *IEEE/ACM ASONAM(13)*, Niagra Falls, Canada.

70. Ghosh, S., Srivastava, H., Agarwal, S. and *Mukherjee, A.* (2013). Runtime delays in Indian railways: Is traffic the cause?, In *ACM DEV(13)* (poster), Bangalore, India.

71. Maity, S. K. and *Mukherjee, A.* (2012). Understanding how dominance affects the emergence of agreement in a social network: The case of Naming Game, In *ASE/IEEE Socialcom(12)*, Amsterdam, The Netherlands.

72. Saha, S., Ganguly, N., and *Mukherjee, A.* (2012). Information Dissemination Dynamics in Delay Tolerant Network: A Bipartite Network Approach, In *ACM MobiOpp(12)*, Zurich, Switzerland.

73. Choudhury, M., Chatterjee, D., and *Mukherjee, A*. (2010). Global topology of word co-occurrence networks: Beyond the two-regime power-law*, In the proceedings of COLING(10)*, Beijing, China. **[CORE RANK: A]**

74. Biemann, C., Choudhury, M., and *Mukherjee, A*. (2009). Syntax is from Mars while Semantics from Venus! Insights from Spectral Analysis of Distributional Similarity Networks*, In the proceedings of ACL(09)*, (*short paper*), Suntec, Singapore. **[CORE RANK: A*]**

75. *Mukherjee, A.*, Choudhury, M., and Kannan, R. (2009). Discovering Global Patterns in Linguistic Networks through Spectral Analysis: A Case Study of the Consonant Inventories*, In the proceedings of EACL(09)*, Athens, Greece *http://arxiv.org/abs/0901.2216* **[CORE RANK: A]**

76. Choudhury, M., *Mukherjee, A.*, Garg, A., Jalan, V., Basu, A., and Ganguly, N. (2009). Language Diversity across the Consonant Inventories: A Study in the Framework of Complex Networks, *In the proceedings of EACL(09)* workshop on *Cognitive Aspects of Computational Language Acquisition*, Athens, Greece *http://arxiv.org/abs/0904.1289*

77. *Mukherjee, A.*, Choudhury, M., Basu, A., and Ganguly, N. (2008). Modeling the Structure and Dynamics of the Consonant Inventories: A Complex Network Approach , *In the proceedings of COLING(08)*, Manchester, UK *http://arxiv.org/abs/0806.0923* **[CORE RANK: A]**

78. Nath, J., Choudhury, M., *Mukherjee, A.*, Biemann, C., and Ganguly, N. (2008) Unsupervised Parts-of-Speech Induction for Bengali, *In the proceedings of LREC*, Morocco, Marrakech (*pdf*)

79. *Mukherjee, A.*, Choudhury, M., Basu, A., and Ganguly, N. (2007). Redundancy Ratio: An Invariant Property of the Consonant Inventories of the World's Languages, *In the proceedings of ACL(07)*, Prague, Czech Republic *(pdf)* **[CORE RANK: A*]**

80. *Mukherjee, A.*, Choudhury, M., Basu, A., and Ganguly, N. (2007). Emergence of Community Structures in Vowel Inventories: An Analysis based on Complex Networks, *In the proceedings of ACL-SIGMORPHON9(07)*, Prague, Czech Republic *(pdf)*

81. Choudhury, M., Thomas, M., *Mukherjee, A.*, Basu, A., and Ganguly, N. (2007). How Difficult is it to Develop a Perfect Spell-checker? A Cross-linguistic Analysis through Complex Network Approach , *Proceedings of TEXTGRAPHS-2, HLT/NAACL(07)*, Rochester, New York *http://arXiv.org/abs/physics/0703198*

82. Choudhury, M., *Mukherjee, A.*, Basu, A., and Ganguly, N. (2006). Analysis and Synthesis of the Distribution of Consonants over Languages: A Complex Network Approach, *Proceedings of COLING-ACL(06)*, Sydney, Australia *http://arXiv.org/physics/0606131* **[CORE RANK: A*]**

83. **Mukherjee, A.**, Bhattacharya, S., Halder, P.K., & Basu, A. (2005). A Virtual Predictive Keyboard as a Learning Aid for People with Neuro-motor Disorders, *Proceedings of the 5th IEEE International Conference on Advanced Learning Technologies, (ICALT - 2005)*, Kaohsiung, Taiwan *(doc)*

Journals

1. Dash, A., Chakraborty, A., Ghosh, S, **Mukherjee, A.** and Gummadi, K. (2022). FaiRIR: Mitigating Exposure Bias from Related Item Recommendations in Two-Sided Platforms. In *IEEE TCSS*, (accepted).
2. Adak, S., Chakraborty, S., Das, P., Das, M., Dash, A., Hazra, M., Mathew, B., Saha, P., Sarkar, S. and **Mukherjee, A.** (2021). Mining the Online Inphosphere: A survey. In *WIREs Data Mining and Knowledge Discovery journal*, (in press).
3. Sarkar, S., Bhagwat, A. and **Mukherjee, A.** (2021). A core-periphery structure based network embedding approach. In *Journal of Social Network Analysis and Mining (SNAM)*, (in press).
4. Das, M., Mathew, B., Saha, P., Goyal, P. and **Mukherjee, A.** (2020). Hate speech in online social media. In *ACM SIGWEB Newsletters*, **4**:1--8.
5. Sikdar, S., **Mukherjee, A.** and Marsili, M. (2020). Unsupervised Ranking of Clustering Algorithms by INFOMAX. In *PLoS ONE*, **10**:e0239331.
6. Singh, M., Pandey, P. K., Goyal, P., **Mukherjee, A.** and Chakrabarti, S. (2020). Analysis of Reference and Citation Copying in Evolving Bibliographic Networks. In *Journal of Informetrics*, **14**(1), 101003. **[IF: 3.88]**
7. Jana, A., **Mukherjee, A.** and Goyal. P. (2020). Network Measures: A New Paradigm Towards Reliable Novel Word Sense Detection. *In Journal of Information Processing and Management*, 57(6), 102173. **[IF: 3.892]**
8. Maity, S. K., Panigrahi, A., and **Mukherjee, A.** (2018). Analyzing Social Book Reading Behavior on Goodreads and how it predicts Amazon Best Sellers. In *Influence and Behavior Analysis in Social Networks and Social Media (LNSN)*, Kaya, M. and Alhajj R. (eds.), Springer, ISBN: 978-3-030-02592-2.
9. Sarkar, S., Kumar, S., Bhowmick, S. and **Mukherjee, A.** (2018). Centrality and Community Scoring Functions in Incomplete Networks: Their Sensitivity, Robustness, and Reliability. In *Machine Learning Techniques for Online Social Networks*, Ozyer T. and Alhajj R. (eds.), Springer, 135--154, ISBN: 978-3-319-89931-2.
10. Maity, S. K., Kharb, A. and **Mukherjee, A.** (2018). Analyzing Linguistic Structure of Question Texts to Characterize Answerability in Quora. *IEEE Transactions on Computational Social Systems*, **5**(3), 816--828. **[IF: 3.29]**
11. Sarkar, S., Sikdar, S., Bhowmick, S. and **Mukherjee, A.** (2018). Using Core-Periphery Structure to Predict High Centrality Nodes in Time-Varying Networks, *Data Mining and Knowledge Discovery Journal (ECML-PKDD 2018 journal track)*, **32**(5), 1368--1396.**[IF: 3.16]**
12. Sikdar, S., Tehria, P., Marsili, M., Ganguly, N. and **Mukherjee, A.** (2018). On the effectiveness of the scientific peer-review system: a case study of the Journal of High Energy Physics, *International Journal on Digital Libraries*, <u>DOI</u>, 1--15.

13. Chakraborty, T., Dalmia, A., *Mukherjee, A.* and Ganguly, N. (2017). Metrics for community analysis: A survey. *ACM Computing Surveys*, **50**(4), 54:1--37. **[IF: 6.13]**

14. Bodych, M., Ganguly, N., Krueger, T., *Mukherjee, A.*, Siegmund-Schultze, R. and Sikdar S. (2017). Threshold-based epidemic dynamics in systems with memory. *Europhysics Letters*, **116**(4), 48004(1--5). **[IF: 1.56]**

15. Sur, S., Ganguly, N. and *Mukherjee, A.* (2016). Brokerage based attack on real world temporal networks. *Journal of Network Science*, **4**(4), 446--459.

16. Singh, M., Chakraborty, T., *Mukherjee, A.* and Goyal, P. (2016). Is this conference a top-tier? ConfAssist: An assistive conflict resolution framework for conference categorization. *Journal of Informetrics*, **10**(4), 1005--1022. **[IF: 3.88]**

17. Chakraborty, T., Srinivasan, S., Ganguly, N., *Mukherjee, A.* and Bhowmick, S. (2016). Permanence and Community Structure in Complex Networks. *ACM TKDD*, **11**(2), 14:1--14:34. (Included in <u>ACM's 21st Annual Best of Computing</u>) **[IF: 1.49]**

18. Chakraborty, T., Kumar, S., Ganguly, N., *Mukherjee, A.* and Bhowmick, S. (2016). GenPerm: A Unified Method for Detecting Non-overlapping and Overlapping Communities. *IEEE TKDE*, **28**(8): 2101--2114. **[IF: 3.86]**

19. Sikdar, S., Ganguly, N., *Mukherjee, A.* (2016) Time series analysis of temporal networks. *EPJB Topical Issue on Temporal Network Theory and Applications*, **89**(1), 1--11. **[IF: 1.44]**

20. Gupta, A. K., Bhattacharya, I., Banerjee, P. S., Mandal, J. K. and *Mukherjee, A.* (2015) DirMove: direction of movement based routing in DTN architecture for post-disaster scenario. *Wireless Network Journal*, **22**(3), 723--740. **[IF: 2.4]**

21. Chakraborty, T., Ganguly, N., *Mukherjee, A.* (2015) An author is known by the context she keeps: significance of network motifs in scientific collaborations. *Journal of Social Network Analysis and Mining*, 5:16.

22. Mitra, S., Mitra, R., Maity, S. K., Riedl, M., Biemann, C., Goyal, P. and *Mukherjee, A.* (2015). An automatic approach to identify word sense changes in text media across timescales. *Journal of Natural Language Engineering*, Special Issue on *Graphs for NLP*, **21**(5), 773--798. **[IF: 1.22]**

23. Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N. and *Mukherjee, A.* (2015). On the categorization of scientific citation profiles in computer sciences. *Communications of the ACM*, **58**(9), 82--90. **[IF: 3.06]**

24. Chakraborty, T., Tammana, V., Ganguly, N. and *Mukherjee, A.* (2015). Understanding and Modeling Diverse Scientific Careers of Researchers. *Journal of Informetrics*, **9**(1), 69--78. **[IF: 3.88]**

25. Sur, S., Ganguly, N. and *Mukherjee, A.* (2015). Attack tolerance of correlated time-varying social networks with well-defined communities. *Physica A: Statisical Mechanics and its Applications*, **420**, 98--107. **[IF: 2.13]**

26. Ganguly, N., Krueger, T., *Mukherjee, A.* and Saha, S. (2014). Epidemic spreading through direct and indirect interactions. *Phys. Rev. E.*, **90**, 032808. **[IF: 2.35]**

27. Maity, R. R., Mallya, A., *Mukherjee, A.* and Ganguly, N. (2014). Understanding the Correlation of the Properties of Human Movement Patterns. *Advances in Complex Systems*, **17**(6),1450019. **[IF: 1.22]**

28. Saha, S., Ganguly, N., *Mukherjee, A.* and Krueger, T. (2014). Intergroup networks as random threshold graphs. *Phys. Rev. E.*, **89**, 042812. **[IF: 2.35]**

29. Chakraborty, T., Sikdar, S., Ganguly, N. and *Mukherjee, A.* (2014). A Quantitative Route to the Rise and Fall of Scientific Research. *Journal of Social Network Analysis and Mining*, 4:187.

30. Bhattacharjee, B., Manna, S. S. and *Mukherjee, A.* (2013). Information sharing and sorting in a community. *Phys. Rev. E.*, **87**(6), 062808. [**IF: 2.35**]

31. Chakraborty, T., Srinivasan, S., Ganguly, N., Bhowmick, S. and *Mukherjee, A.* (2013). Constant communities in complex networks. *Nature Scientific Reports*, **3**, 1825. [**IF: 4.01**]

32. Ganguly, N., Saha, S., Maiti, A., Agarwal, S., Peruani, F. and *Mukherjee, A.* (2013). Effect of attachment strategies on bipartite networks. *European Phys. Jour. B.*, **86**, 287. [**IF: 1.44**]

33. Maity, S. K., *Mukherjee, A.*, Tria, F. and Loreto, V. (2013). Emergence of fast agreement in an overhearing population: The case of the naming game. *Euro. Phys. Lett.*, **101**, 68004. [**IF: 1.88**]

34. Ghosh, S., Saha, S., Srivastava, A., Kuerger, T., Ganguly, N. and *Mukherjee, A.* (2013). Understanding Evolution of Inter-Group Relationships using Bipartite Networks. *IEEE Journal on Selected Areas in Communications*, Special Issue on *Emerging Technologies in Communications*, **31**(8), 1--11. [**IF: 7.17**]

35. *Mukherjee, A.*, Choudhury, M., Ganguly, N. and Basu, A. (2013). Language Dynamics in the Framework of Complex Networks: A Case Study on Self-organization of the Consonant Inventories. In *Cognitive Aspects of Computational Language Acquisition*, Poibeau, T., Villavicencio, A., Korhonen, A. and Alishahi, A. (eds.), Springer, 51--78, ISBN 978-3-642-31863-4.

36. Maity, S. K., Venkat, M. T. and *Mukherjee, A.* (2012). Opinion formation in time-varying social networks: The case of the naming game. *Phys. Rev. E*, **86**, 036110. [**IF: 2.35**]

37. Loreto, V., *Mukherjee, A.* and Tria, F. (2012). On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences* (*PNAS*), **109**(18), 6819--6824. Full version.
Media Coverage: PNAS Press Highlights, Scientific American, Live Science, MedicalXpress, The Hindu, Business Line, Times of News, BBC Future, Prospect Magazine (UK), Yahoo! News, Deccan Herald, MSN, Homunculus, Doctordisruption, Galileo Net, Scienza e Tecnica, Chinese EurekAlert, De Standaard, Design by Coffee, Blitz quotidiano, Wired (Italian). This work now has a Wikipedia entry in the page for Color term. [**IF: 9.58**]

38. *Mukherjee, A.*, Loreto, V. and Tria, F. (2012). Why are basic color names "basic"?, *Topical Issue on Language Dynamics, Advances in Complex Systems*, **15**(3,4), 1150016 (1--13). [**IF: 1.22**]

39. Tria, F., *Mukherjee, A.*, Baronchelli, A., Puglisi, A. and Loreto, V. (2011). A fast no-rejection algorithm for the Category Game, *Journal of Computational Science* , **2**(4), 316--323, *http://arxiv.org/abs/1012.3583*

40. *Mukherjee, A.*, Choudhury, M. and Ganguly, N. (2011). Understanding how both the partitions of a bipartite network affect its one-mode projection, *Physica A: Statisical*

Mechanics and its Applications, **390**(20), 3602--3607 http://arxiv.org/abs/1105.3902. [**IF: 2.13**]

41. Ghosh, S., Banerjee, A., Sharma, N., Agarwal, S., Ganguly, N., Bhattacharya, S. and ***Mukherjee, A.*** (2011). Statistical Analysis of the Indian Railway Network: A Complex Network Approach, *Acta Physica Polonica B Proceedings Supplement*, **4**(2), 123--137, (pdf).

42. Loreto, V., Baronchelli, A., ***Mukherjee, A.***, Puglisi, A. and Tria, F. (2011). Statistical physics of language dynamics, *Journal of Statistical Mechanics: Theory and Experiment*, P04006. [**IF: 2.37**]

43. ***Mukherjee, A.***, Tria, F., Baronchelli, A., Puglisi, A. and Loreto, V. (2011). Aging in language dynamics, *PLoS ONE* **6**(2): e16677. [**IF: 2.77**]

44. Pal, R., ***Mukherjee, A.***, Mitra, P., and Mukherjee, J. (2010). Modelling Visual Saliency Using Degree Centrality, *IET Computer Vision* **4**(3), 218--229. [**IF: 1.64**]

45. ***Mukherjee, A.***, Choudhury, M., Basu, A., and Ganguly, N. (2010). Modeling the Redundancy of Human Speech Sound Inventories: An Information Theoretic Approach, *Journal of Quantitative Linguistics*, **17**(4), 317--343.

46. Choudhury, M., Ganguly, N., Maiti, A., ***Mukherjee, A.***, Brusch, L., Deutsch, A., and Peruani, F. (2010). Modeling discrete combinatorial systems as Alphabetic Bipartite Networks: Theory and Applications *Phys. Rev. E,* **81** 036103 *http://arxiv.org/abs/0811.0499* [**IF: 2.35**]

47. ***Mukherjee, A.***, Choudhury, M., Basu, A., and Ganguly, N. (2009). Self-organization of the Sound Inventories: Analysis and Synthesis of the Occurrence and Co-occurrence Networks of Consonants, *Journal of Quantitative Linguistics,* **16**(2), 157--184 *http://arXiv.org/physics/0610120*

48. Choudhury, M., and ***Mukherjee, A.*** (2009). The Structure and Dynamics of Linguistic Networks. In *Dynamics on and of Complex Networks: Applications to Biology, Computer Science, Economics, and the Social Sciences*, Ganguly, N., Deutsch, A., and Mukherjee, A., (eds.), Birkhauser, Springer, Boston, 145--166, ISBN: 978-0-8176-4750-6 *http://www.springerlink.com/content/x0g78760w753p781/*

49. ***Mukherjee, A.***, Choudhury, M., RoyChowdhury, S., Basu, A., and Ganguly, N. (2008). Rediscovering the Co-occurrence Principles of the Vowel Inventories: A Complex Network Approach, *Advances in Complex Systems,* **11**(3), 371--392 *http://arxiv.org/abs/physics/0702056*

50. ***Mukherjee, A.***, Chakraborty, K., and Basu, A.(2008). SweepSticks - An Adaptive Virtual Mouse for People with Neuro-Motor Disorders, *Assistive Technology Journal of the Rehabilitation Engineering Society of North America,* **20**(2), 111--124 *(doc)* [**IF: 0.65**]

51. Choudhury, M., Saraf, R., Jain, V., ***Mukherjee, A.***, Sarkar, S., and Basu, A. (2007). Investigation and Modeling of the Structure of Texting Language, *Special Issue of the International Journal of Document Analysis and Recognition on Noisy Text Analytics*, Springer, **10**(3-4)*,* 157--174 *(pdf)* [**IF: 0.84**]

52. Peruani, F., Choudhury, M., ***Mukherjee, A.***, and Ganguly, N. (2007). Emergence of a non-scaling degree distribution in bipartite networks: a numerical and analytical study, *Euro. Phys. Letters*, **79**(2), 28001 *http://arXiv.org/abs/cond-mat/0703634* [**IF: 1.88**]

53. *Mukherjee, A.*, Choudhury, M., Basu, A., and Ganguly, N. (2007). Modeling the Co-occurrence Principles of the Consonant Inventories: A Complex Network Approach, *Int. Jour. of Modern Phy. C*, **18**(2), 281--295 *http://arXiv.org/physics/0606132* **[IF: 1.017]**

13. Collaborations

- **Matteo Marsili, ICTP, Trieste, Italy**

This collaboration started with the award of Simons Associateship to Dr. Mukherjee in 2014. Since then Dr. Mukherjee has been actively collaborating with Dr. Marsili on various research problems pertaining to (i) evaluation of data clustering algorithms for high dimensional data (ii) large-scale analysis of peer review systems. The collaboration has also got further extended to the Media Lab researchers at SISSA who have generously provided various datasets to run the experiments. The PhD student involved in this project received the prestigious STEP fellowship to visit ICTP in 2015 and 2016 to continue this research. The collaboration has resulted in two top-tier conference papers – ACM CIKM 2016 and JCDL 2017 (the work received a best paper nomination). Two more works are in submission.

- **Krishna Gummadi, MPI-SWS, Germany**

This collaboration started in 2018 and centers around building fair and ethical machine learning algorithms. The main focus is to mitigate exposure bias in two sided recommendation systems. Many items on various online portals get undue exposure by virtue of their relatedness to other items in the platform. In this work we attempt to devise new recommendation algorithms to bring the deserved and observed of items as close as possible.

- **Chris Biemann, University of Hamburg, Germany**

This collaboration started with a DAAD-IIT faculty exchange program jointly awarded to Dr. Mukherjee and Dr. Biemann in 2013. The two have been closely collaborating on automatic detection of word sense changes from time-stamped corpora. The collaboration has led to the success of a DST-DAAD joint project proposal and two top-tier publications – ACL 2014 and JNLE 2015. Recently, Dr. Mukherjee and Dr. Biemann have started work on co-hyponymy detection in natural language text and their evolution over time. Dr. Mukherjee has been awarded the prestigious Humboldt Fellowship for Experienced Researchers to continue this new thread of collaboration with Dr. Biemann at the University of Hamburg.

- **Sanjukta Bhowmick, University of North Texas**

Dr. Bhowmick and Dr. Mukherjee have been consistently collaborating for more than five years now. This collaboration has, to its credit, top-notch publications like Nature Scientific Reports, ACM SIGKDD, ACM CIKM, ACM TKDD, IEEE TKDE, ASONAM. The collaboration has also resulted in the production of an excellent quality Ph.D. thesis of Dr. Tanmoy Chakraborty. The thesis has attracted many notable awards

like the INAE Best Dissertation Award, XRCI Best Dissertation Award etc. Both the collaborators have been actively working on noisy incomplete network analysis and stability of networks for the past one year with a new Ph.D. student. The current proposal is based on the problems to be solved in this thesis and the work has already started receiving appreciation in the community resulting in two notable publications in ASONAM 2016 and CIKM 2016.

- **Tyll Krueger, Wroclaw University of Science and Technology, Poland**

This collaboration started with the visit of Dr. Krueger to IIT Kharagpur in 2012. Dr. Mukherjee and Dr. Krueger have jointly studied various theoretical aspects pertaining to (i) evolution and (ii) diffusion in large networks. The two researchers have jointly come up with new ideas of segmented message diffusion and growth of threshold random graphs. This work resulted in a successful Indo-Polish bilateral researcher exchange project supported by the governments of both the countries. Various PhD students from both the groups have been working on these problems resulting in publications in leading physics and computer science journals including Physical Review E, Europhysics Letters, IEEE JSAC.

- **Vittorio Loreto, Sapienza University, Rome**

Dr. Mukherjee have been collaborating with Dr. Loreto since 2009 while he was in the group of Dr. Loreto as a post-doctoral researcher in ISI Foundation, Italy. Dr. Mukherjee has been working on various issues related to human language dynamics with a special focus on emergence of naming and categorization in linguistic communities. This research has resulted in the development of a computational basis to the emergence of hierarchy in color naming patterns. A successful Indo-Italian bilateral researcher exchange project supported by both the governments has resulted from this project. Many significant publications including PNAS, PLoS One, Europhysics Letters, Journal of Statistical Mechanics have come out of this collaboration.

- **Soumen Chakrabarti, IIT Bombay**

This collaboration started in 2016 to jointly study the dynamics of scientific citations. Dr. Chakrabarti and Dr. Mukherjee have together designed a new model of citation growth that is able to explain both entrenchment as well as obsolescence in a single model of evolution of citation network. Earlier models would mostly explain obsolescence by fitting exponential functions with either global parameters or per-node parameters. However, they could not explain where the diverted attention would go in the network when a certain paper is undergoing obsolescence. In this work, Dr. Chakrabarti and Dr. Mukherjee show the evidence of "relaying" of citations from older papers to newer papers, i.e., how younger papers would "steal" citations from their older neighbors undergoing obsolescence. This work has been accepted to appear in AC SIGKDD 2017, which is the premium conference in knowledge discovery and data mining.

- **Monojit Choudhury, Microsoft Research Lab, India**

Dr. Choudhury and Dr. Mukherjee have been collaborating since 2006 while both of them were still PhD students. They have together significantly contributed to the literature of language evolution and change with publications in top-tier CS conferences like ACL, COLING, EACL and top physics journals like Physical Review E, Advances in Complex Systems, Europhysics Letters and Physcia A.

- **Lipika Dey and Indrajit Bhattacharya, TCS Innovation Labs, Kolkata**

Recently, Dr. Mukherjee's continuous efforts in citation analysis got noted by one of the major Indian tech giant – Tata Consultancy Services. Consequently the company has engaged into active research collaboration with Dr. Mukherjee and his scholars. Three research scientists (Dr. Dey, Dr. Bhattacharya and Dr. Pal) from TCS innovation labs Kolkata and Delhi are actively collaborating on (i) analysis of patent citations and the growth of industry leaders, (ii) concept mining from citation networks for enhanced future co-authorship prediction and (iii) link dynamics in Wikipedia articles.

- **Shourya Roy, Flipkart Inc.**

In this collaboration, we study the dynamics of product returns on e-commerce portals like Flipkart. Using a one-year long longitudinal data, we model the problem fusing concepts from network science and machine learning that allow us to predict if a user shall be returning a product of certain brand and size. The key idea is to cluster products into groups and interpolate the return probability of a query (user, brand, size) by the overall return probability of the cluster to which this query belongs to. Our approach achieves an improvement of 20% over the vanilla baseline.

- **Niyathi Chaya, Adobe Research.**

In this collaboration, we study how we can automatically identify if a statement is an exaggerated version of another. We use a sequence labeling approach to identify relationship phrases and compare them to perform strength classification and finally exaggeration identification.

14. Media Attention

- The 2012 PNAS paper: PNAS Press Highlights, Scientific American, Live Science, MedicalXpress, The Hindu, Business Line, Times of News, BBC Future, Prospect Magazine (UK), Yahoo! News, Deccan Herald, MSN, Homunculus, Doctordisruption, Galileo Net, Scienza e Tecnica, Chinese EurekAlert, De Standaard, Design by Coffee, Blitz quotidiano, Wired (Italian). This work now has a Wikipedia entry in the page for Color term.
- The TKDD 2016 paper: Included in ACM's 21st Annual Best of Computing.
- The CSCW 2016 paper: Received the best paper honorable mention, was included in MIT Tech Review best of arXiv.
- The HICSS 2016 paper: Included in the MIT Tech Review best of arXiv.

- The work on disaster management has been recently covered by Business Standard and Financial Express.
- Our work on product returns with Flipkart features in Economic Times.
- The JCDL 2017 work on analysis of scientific peer review system received a best paper nomination.
- The ASONAM 2013 paper on dynamics of popularity of CS fields received a best paper nomination.
- Our work on analyzing book reading behavior got covered in TechXplore, geek.com and phys.org.
- Our work on incivility detection got covered in TechXplore.
- Our work on hate mining and misogyny detection was included in the MIT Tech Review best of arXiv.
- Our WebSci 2019 paper receives the best paper honorable mention.
- Our SocInfo 2019 paper gets the best paper nomination.
- Our work on targeted media bias with Facebook gets media attention in Financial Times, Hindustan Times, Business Standard, NDTV etc.
- Our work on Gandhipedia gets huge media coverage - The Hindu BusinessLine, The Hindu, Careers360, The Week, India Today (Web), India Today, Deccan herald, News18, Hindustan Times, Indian Express, NDTV, Jagran, IndiaTV, OrissaPost, Business Standard, The Statesman Financial Express.
- Our work on explainable hate speech covered in technology.org and newshive.

15. Invited Lectures (recent)

   Invited lecture at ICTP-CSRC Workshop on Big data, Machine Learning and Complexity research, 2020, Online.
   Invited talk at Machine Learning Prospective for Complex Networks workshop in CCS 2020, Online.
   Several FDP lectures on machine learning, data science and cyber-security at different Indian colleges including Kolaghat Engg. college, Guru Nanak Dev Engg. college, NIT Jalandhar, KIIT Bhubaneswar, JIET Jaipur.
   Invited lecture at MIU, ISI Kolkata on its foundation day, 2019.
   Invited lecture at PReMI 2019, Tezpur, Assam, 2019.
   ACM DSP lecture at UEM Jaipur, 2019.
   Invited panel lecture at Facebook, Delhi 2019.
   Invited lecture at NIST, Behrampur, Odisha 2019
   Invited lecture at UCET, VBU, Hazaribag 2019.
   Invited lecture at WebSci workshop at IIT Madras, 2019.
   Invited lecture at University of Hamburg, Germany, 2018.

Invited lecture at MPI SWS, Kaiserslautern, Germany, 2018.

Invited speaker at the 2017 Microsoft Research India event on "AI for Social Good".

Invited speaker at the 2017 IISc workshop on User Generated Content Analysis.

Invited to attend and collaborate at the strategy workshop 2017 on "Computation, Sciences and Society" held at Mysore Park Infosys campus. This workshop was organised by K Vijay Raghavan, Kris Gopalakrishnan, Manindra Agrawal, Nisheeth Vishnoi and Somenath Biswas.

Invited for the early career experience talk at the India-KDD event of ACM SIGKDD 2016 at San Francisco. This is a flagship program of the Indian data mining and data science community organized as a session of the premium data mining conference ACM SIGKDD. (https://ikdd.acm.org/kdd2016/)

Invited speaker for the Machine Learning Data Science event of Microsoft (MLDS 2016) (http://indiamlmeet.azurewebsites.net/#agenda)

Invited speaker at the IWML 2016 workshop at IIT Kanpur. This has been the flagship ML conference in India.

Invited panelist at the Academic Research Summit jointly organized by ACM India and Microsoft Research.(https://www.microsoft.com/en-us/research/event/academic-research-summit/?from=http%3A%2F%2Fresearch.microsoft.com%2Fen-us%2Fevents%2Facadsummit%2F).

Invited to participate and talk on the event of Language Day 2016 jointly organized by ICTP, Trieste, Italy and SISSA, Trieste, Italy.

16. Sponsored projects

- **Project name**: Leveraging Bipartite Networks to Investigate the Dynamical Properties of Socio-technical Systems. Agency: Samsung GRO. (This grant was received through a world-wide highly competitive competition that we participated.). Budget: ~USD 70K. Status: Ongoing.

- **Project name:** New perspectives for computational social science. Agency: Exchange of Researchers within the frame of the Executive Program of Scientific and Technological Cooperation, between the Republic of India and the Italian Republic. (This project was in continuation of my post-doctoral research with Prof. Vittorio Loreto. One of the remarkable achievements of the work is our joint paper in the Proceedings of the National Academy of Sciences, USA). Budget: ~USD 7K. Status: Completed (2012 – 2014).

- **Project name:** ICT for computational social science. Agency: Fast Track Scheme for Young Scientists, Department of Science & Technology, Government of India.

(As a part of this project we developed various web experiments to study the dynamics of human languages.) Budget: ~USD 21K. Status: Completed (2012 - 2014).

- **Project name:** Post-disaster situation analysis and resource management using delay-tolerant peer-to-peer wireless networks. Agency: ITRA, DIT. (This project was obtained through a very competitive selection process. It was a collaborative effort of six institutions including IIT Kharagpur, IIM Kolkata, IIEST Shibpur, NIT Durgapur, Kalyani Government Engg. College and Heritage Institute of Technology. This project was massive appreciated because of it direct relevance to various government initiatives. Recently, there has been media coverage by Business Standard (http://www.business-standard.com/article/current-affairs/human-mobility-powers-wireless-communication-to-aid-disaster-relief-117071500396_1.html) and Financial Express (http://www.financialexpress.com/industry/technology/human-mobility-powers-wireless-communication-for-disaster-relief/764520/)). Budget: ~USD 138K. Status: Completed (2013 - 2018).

- **Project name:** Data driven approaches for inferring opinion dynamics on social networks. Agency: Indo-Polish bilateral exchange, DST, India. Budget: ~USD 20K. Status: Completed (2015 - 2017).

- **Project name:** Large-scale text analytics to identify temporal and media specific scopes of words. Agency: DST-DAAD PPP. Budget: ~USD 20K. Status: Ongoing.

- **Project name:** Permanence and community structure analysis. Agency: Microsoft Research India. Budget: ~USD 4K (unrestricted gift). Status: Ongoing.

- **Project name:** Algorithms for computational social science. Agency: Yahoo! Labs India. Budget: ~USD 2.5K (unrestricted gift). Status: Ongoing.

- **Project name:** Social recommendation in e-commerce. Agency: Flipkart. Budget: ~USD 7.5K + 5K + 2K (unrestricted gift three times). Status: Ongoing.

- **Project name:** Exaggeration identification in science news. Agency: Adobe Research. Budget: ~USD 25K (unrestricted gift). Status: Ongoing.

- **Project name:** Targeted bias in Indian media outlets. Agency: Facebook Research. Budget: ~USD 30K (unrestricted gift). Status: Ongoing.

- **Project name:** Behavioral modeling in multi-sensor environments. Agency: TCS Innovation Labs. Budget: ~USD 100K (unrestricted gift). Status: Ongoing.

- **Project name:** Gandhipedia, an one-stop AI enabled portal for searching Gandhian literature. Agency: Ministry of Culture, GoI. This is a project of national importance and and appears as a budget item of 2019-2020 https://www.outlookindia.com/newsscroll/gandhipedia-being-developed-as-social-media-portal-culture-ministry/1569914?fbclid=IwAR2Dwxb5nxugPVLiFj8cnTWayDeDGGgi9rflL0F5znCwqXomykBw_UBlNP0, https://www.hindustantimes.com/budget/budget-2019-encyclopedia-on-mahatma-gandhi-gandhipedia-proposed/story-qDmJHtYIQ2MUKCqYG51hEO.html

  Budget: ~USD 697K. Status: On-going

****

<u>Significant Achievements</u>

Dr. Mukherjee has made fundamental contributions to the growing field of computational social science for the last few years. His main research centers around applying artificial intelligence, machine learning and network science to understand human languages. He has been studying the language of social media in depth looking at various aspects including (i) evolution of usage of topics/hashtags in Twitter/Quora, (ii) evolution of writing style in Twitter (iii) abusive behavior detection in Twitter (iv) spread of hate speech in online social media and methods to dilute their negative effect (v) targeted bias in news text. **Some of these works have received best paper honorable mention awards at various top CS conferences (ACM CSCW 2016, ACM WebSci 2019) and many media coverages.** Dr. Mukherjee has a strong collaboration with the University of Hamburg with joint objectives of studying various linguistic properties of hate speech. **This collaboration is facilitated by the prestigious Humboldt Fellowship for Experienced Researchers which Dr. Mukherjee received in 2017.** Another area of research that Dr. Mukherjee is actively working on is the analysis of scientific text documents. He has developed working systems like FeROSA and OCR++ tools for information extraction from scientific documents. **The OCR++ tool received the prestigious Gandhian Young Technological Innovation appreciation in 2017.** Very recently, Dr. Mukherjee has started taking active interest in ethical and moral aspects of AI algorithms. He has been working on bias identification in media and fairness of recommendation algorithms. **He received the very competitive Ethics in AI research award from Facebook to pursue this line of research. He also developed a first of its kind full semester PG elective course (http://cse.iitkgp.ac.in/~animeshm/course_aieth2020.html) on this subject which received a**

lot of student attention. He is presently leading the Gandhipedia project (see: https://kgpchronicle.iitkgp.ac.in/gandhipedia/) which is of national importance and is a budgeted item of the Union ministry (https://tinyurl.com/y4cdnjvj). The project aims to digitize and index all works of Gandhi in a fully searchable format and to spread Gandhian thoughts across social media through AI-enabled social bots. The indexed data is being used to reconstruct the social network of Gandhi. The first version of the portal is up and running here: http://gandhipedia150.in/ The social network is drawn over different timepoints of Gandhi's life. The social bot https://twitter.com/Gandhipedia150 posts quotes by the Mahatma, pictures from different timepoints of his life. The project needs strong leadership and technical skills to manage a big task force.

****

<u>Supporting evidences of contributions to AI</u>

A unique feature of Dr. Mukherjee's career path is that he has always thrived to publish in the top-notch AI forums. His PhD work featured in the topmost conference in computational linguistics -- ACL'06, ACL'07, COLING'08 and EACL'09. These works have played a pioneering role in bringing the area of human language dynamics to mainstream CS. His post-doc research got a place in the PNAS, USA which is one of the most prestigious forums that cuts across the boundaries of science and engineering. Dr. Mukherjee has been constantly trying to disseminate the same culture among his PhD students.

Students working in citation analysis repeatedly publish their works in the most competitive AI/ML/DM forums – KDD'15, JCDL'15, CIKM'15, KDD'17, JCDL'17, ECIR'19, NAACL'18 demo, JCDL'20. Students working in language/social media only publish in the premium NLP/Social Computing forums like ACL'14, ICWSM'15, CSCW'16, ICWSM'17, EMNLP'17, CSCW'18, ICWSM'19, ACL'19, EMNLP'19, ECML-PKDD'20, ACL'20, EMNLP'20, CSCW'20 and AAAI'21. In core network science students only publish in premium DM conferences like KDD'14, CIKM'16, JCDL'17, CIKM'18, ECML-PKDD'18, Infocom'19 and high impact journals like ACM TKDD and Nature Scientific Reports. This itself is a testament of the superlative quality of engineering contributions made by Dr. Mukherjee. This culture has brought many notable student achievements of Dr. Mukherjee including (i) the Google, Microsoft and TCS PhD fellowships (ii) INAE & XRCI doctoral dissertation award, (iii) Ramanujan fellowship (iv) Early Career award (v) participation in Heidelberg Laureate forum etc.

Dr. Mukherjee has extensive collaborations with academia and industry including ICTP, Trieste, University of Hamburg, University of North Texas, MPI-SWS, Germany, IIT Bombay, Microsoft Research India, Flipkart, TCS Innovation Labs, Adobe Research India, Facebook

Research. Dr. Mukherjee's work has been covered by various media outlets like MIT Tech Review, BBC Future, TechXplore, Geeg.com, phys.org and Scientific American.

Dr. Mukherjee regularly serves as technical program committee members of many top-tier conferences (JCDL, EMNLP, COLING, WWW, IJCAI, NAACL-HLT, ACL, SDM, AAAI, PAKDD) and referee to various high impact journals (PNAS, Nature Sci. Rep, IEEE TKDE, Physica A, ACM TKDD, ECML-PKDD journal track, IEEE TKDE,). He was the Program Chair of COMSNETS 2018 and would be the General Chair of COMSNETS 2019. He was also the organiser of IndoML 2020 which was themed around AI and Science. In order to sensitize this area he has delivered courses like Social Computing, Complex Networks and Economic and Financial Network Analytics and given talks and tutorials in the country and abroad.

Dr. Mukherjee's primary contributions to the field of AI/NLP/ML/DM are listed below. Tool and application development is an integral part of many of the projects taken up by Dr. Mukherjee.

Development of Gandhipedia: Recently, Dr. Mukherjee has been appointed the PI for the development of Gandhipedia portal to commemorate the 150th birth year of Mahatma Gandhi. This is a project of national importance and was an item in the 2019 budget of GoI (https://www.hindustantimes.com/budget/budget-2019-encyclopedia-on-mahatma-gandhi-gandhipedia-proposed/story-qDmJHtYIQ2MUKCqYG51hEO.html). The key idea is to use the advanced AI and data analytics techniques to decode Gandhian literature and use it to (i) investigate the social network of the Mahatma over time and (ii) post Mahatma's quotes on social media platforms (e.g., Twitter) to control hate spread specifically during disaster situations and to reduce the unnecessary panic caused in such situations. A first version of the portal can be found here: http://gandhipedia150.in/. This is a unique attempt to connect the Indian population (especially the youth) with Gandhian thoughts and philosophies. A screenshot of the front page of the portal is presented below.
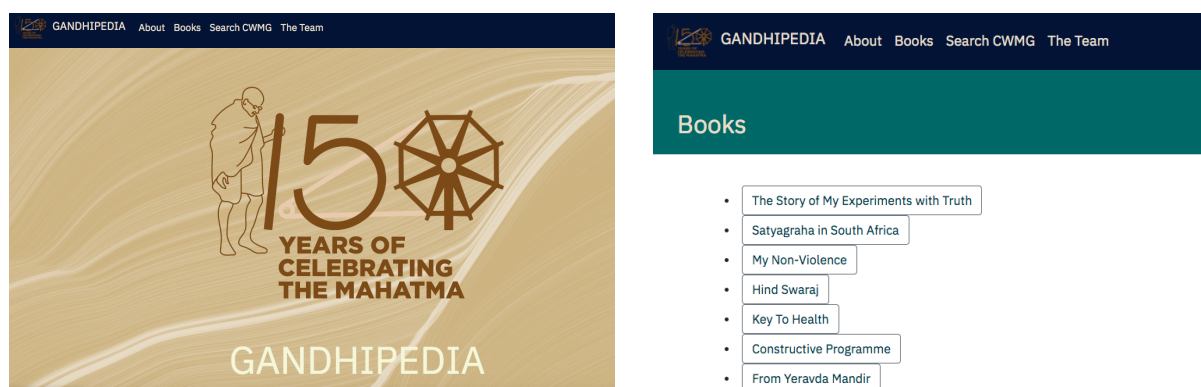
One can use this portal to find information including important people in various phases of his life, important places he visited as he grew up and the social groups that he was connected to at any particular phase of his life. An example of a social group of the Mahatma during his early

life is shown below. Each bubble on the network can be clicked to get more information about the person mentioned in the bubble. An example search page is presented below. **The first version of the portal will be presented as a demo in JCDL 2020 which is an A* conference in Computer Sciences (digital library domain).**
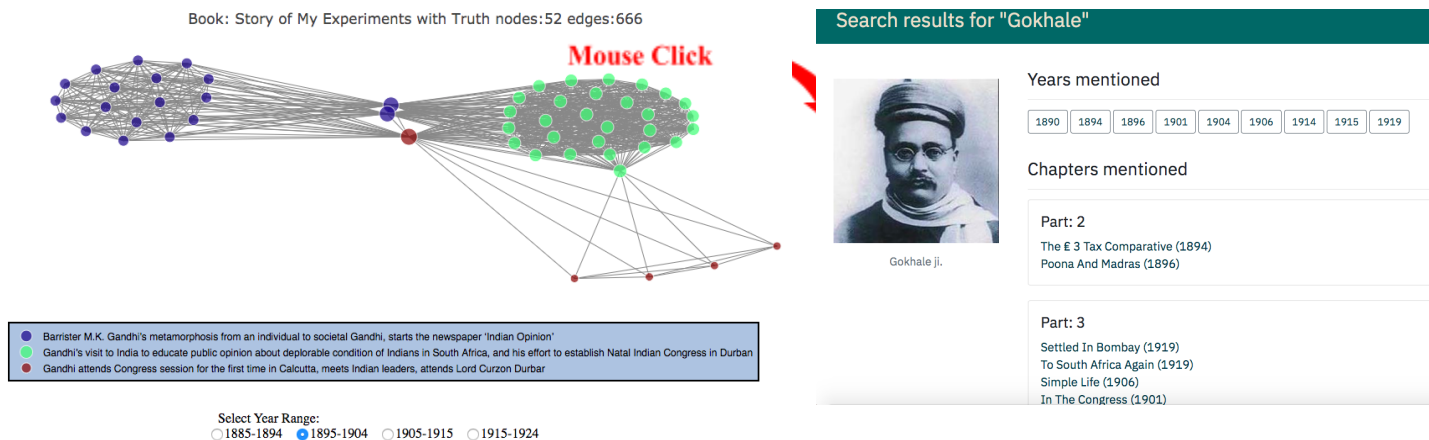
Targeted bias in Indian media outlets: Fake news has become a major point of concern in India owing to the explosive growth in the total number of smartphone users and a massive increase in the number of overall Internet users. On the other hand, the rise of social media has democratized content creation and it is more difficult than ever to discriminate between what is real and what is fake. This liberalization has led to a new kind of societal evolution. Rise of data analytics with cheap and powerful computing facilities holds another threat to democracy in a liberal country making targeted content creation easy and currently the most famous index that is followed across the globe, the Press Freedom Index, by its very definition, is unable to capture such biases as these biases arise due to abuse of liberty by a section of the media. A major influence of fake news and media bias, apart from the systematic decay of liberal democratic institutions, is on elections. There have been accusations of influencing the media and influence on the spread dynamics of news by all major party leaders in India (among other countries like the USA, Canada etc.) especially in time of elections and despite strict regulations enforced by the Election Commission, it is difficult to take definite measures when the crimes are abstract (especially in case of biased news) and sparse, and quantification is impossible in the absence of any automatic mechanism.

Mukherjee and his group set out in leveraging available information on online published media to identify bias in the news media articles. One of the major challenges in this work is formulating something as abstract as "bias" in a quantifiable manner. They have collected a huge dataset for 3 major newspapers in India (namely TOI, The Hindu and India Today) for the last 20 years and measured bias on 3 different metrics.

      a. Coverage bias - formulated as the ratio of the number of mentions of terms pertaining to two major political parties. Later when we deal with more than two
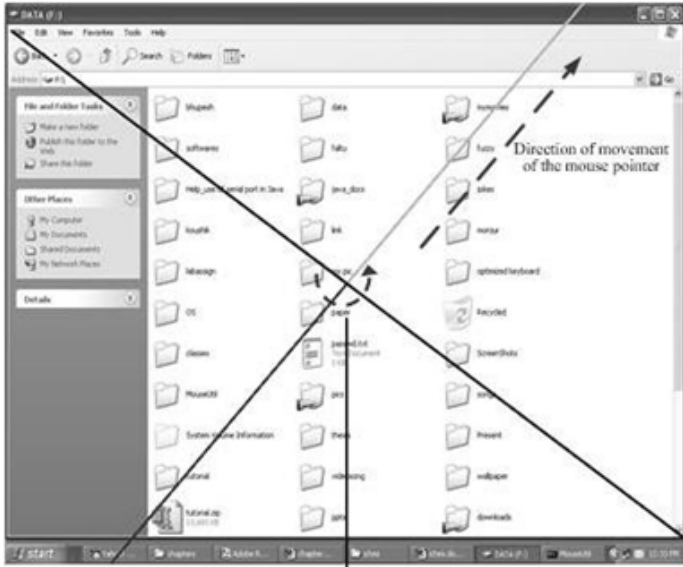
parties, we plan to use the inverse of entropy in the distribution of words pertaining to each political party.

    b. Word choice bias - formulated as the ratio of the positive and the negative words in case of each political party. For clarity, we plotted the ratio of word choice bias for both political parties in our work. We used the work of [Patankar et al. 2018] for the assignment of positive/negative sentiment scores to words.

    c. Topic choice bias - formulated as the divergence score from the aggregate topic distribution for each newspaper. In our study, we present a graphical representation of the same.
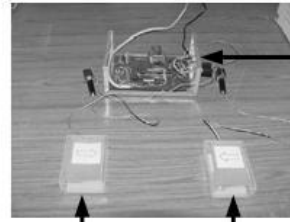
They have observed how certain bias factors peak up during election times. They also show how biased news get larger shares on social media like Twitter thus allowing the news to penetrate deep and wide into the social network. Mukherjee and his group won the very competitive Ethics in AI research award from Facebook for this project (https://www.financialexpress.com/education-2/facebook-award-for-iit-kharagpur-faculty-for-research-in-targeted-media-bias/1720503/, https://www.hindustantimes.com/education/iit-kgp-professor-bags-facebook-award-for-research-in-targeted-media-bias/story-DLOYEniMU4pB5Lsh3GA4XK.html, https://www.theweek.in/wire-updates/national/2019/09/28/ces8-wb-iit%20kgp-award.html).

Spread of hate speech online: According to a study from the Pew Research Center 41% of American adults have experienced some form of online harassment. International Network for Hate Studies has been formed in view of the severity of the problem that reports worldwide initiatives to combat hatespeech. Detecting and containing such behavior automatically is extremely challenging due to its ambiguous nature and threats of compromise of free speech. Mukherjee has started publishing actively in this area. In CSCW 2018 [Maity et al. 2018] he along with his students showed how tracking opinion conflicts between the instigator and the target can help in early detection of incivil behavior. This work was covered in TechXplore. In another work recently accepted in ICWSM 2019 [Mathew et al. 2019(a)], he studies how hatespeech can be countered using effective counterspeech across Youtube hate videos. In similar lines, he studied [Mathew et al. 2019(b)] how hate and counter tweets interact on Twitter. Using a huge dataset collected from gab.com, he and his team also studied [Mathew et al. 2019(c)] how hate speech spreads in online media. They showed that hate content spreads significantly much faster and there are special clans of dedicated users responsible for this. This work received the Best Paper Honourable Mention at the ACM WebSci 2019. Finally, his recent work [Mathew et al. 2019(d)] on misogyny identification was adjudged as MIT Tech Review's best of the Physics arXiv. Together with his students, he is also building an up-to-date reading list on all works being done in this area, which is in its infancy.

Working AI prototypes: Dr. Mukherjee has developed many working systems. In his early career he built an adaptive virtual mouse for users with neuro-motor disorders for which he received the ISCA Young Scientist Award. Some snapshots of this system called SweepSticks are presented below.
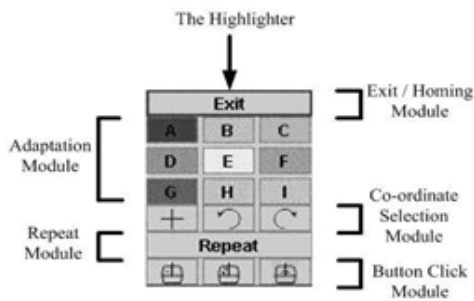
The Interface with the Computer

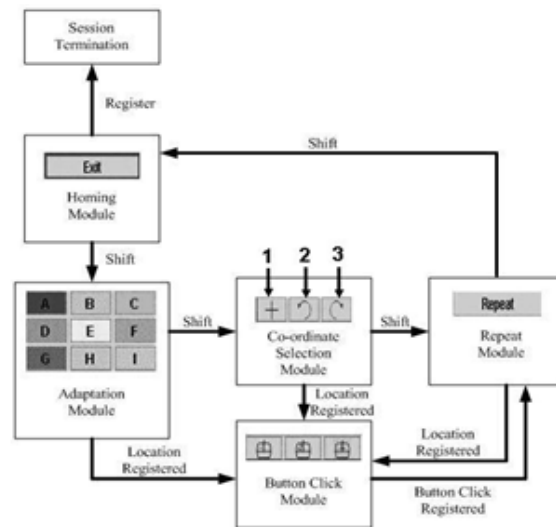The Switch Emulating the Shift Operation   The Switch Emulating the Register Operation

Direction of movement of the mouse pointer

Direction of rotation of the axes

The main

on-screen interface of SweepSticks.

Process flow of SweepSticks.

The 360 degree screen scan.   The mechanical switches to operate SweepSticks.

Recently, his team developed a faceted recommendation system – FeRoSA -- for searching scientific articles. This work was published in PAKDD 2016. In a followup project the team developed an efficient information extraction system from scientific documents – OCR++ (http://www.cnergres.iitkgp.ac.in/OCR++/home/); for this they received the very prestigious Gandhian Young Technological Innovation Award. Recently, he and his team developed advanced scholar search that support natural language queries which has been accepted in NAACL'19 demo track. A snapshot of the tool is shown below.