Statement of Intent
Zahra Iman

Every day, the world population is moving from rural to urban locations. With the current growth of urbanization, developing more intelligent strategies for cities to deal with various aspects of urban life such as wellbeing, social mobility, and economic growth is a global imperative. This requires new strategies to leverage advances in data analysis that I want to research during my PhD.

In order to understand the needs and preferences of society and provide people with better healthcare, transportation, and government services, we need to exploit all available data. Social networks such as Twitter and Facebook have millions of users from diverse backgrounds, who share news, events, opinions, and facts every second of every day. For example, they tweet about localized urban issues such as potholes, car accidents, traffic jams, and public transport as well as family, education, health and sports events. They also tweet about wider-ranging national and global issues such as natural disasters, epidemics, and politics. Hence, social media provides a rich perspective on humanity ranging from minor, localized personal observations to topics of global concern. In order to deal with global challenges, improve emergency management, and achieve a higher quality of life, there is a need to capture and make use of this massive amount of information. To this end, I propose the research topic of "Learning Social Media Sensors" as a novel tool to help achieve these goals.

In the abstract, sensors detect measurable quantities in our environment. Expanding on this definition, social media can be used as a sensor to detect trending news (celebrity death), real-time events (natural disasters or epidemics), sentiment and opinions (political orientation of society), and preferences and traits (stock market prediction). However, designing methodologies and extracting features that are robust to highly dynamic changes in social media, with various forms of expression e.g., informal, short, unstructured texts, written by individuals from different educational levels, and with large volumes of extraneous material mixed is a difficult task. In light of these technical difficulties, designing sensors to detect specific phenomena requires highly specialized research to learn to extract targeted information for a variety of individual information needs.

Unfortunately, the existing tools for searching Twitter do not facilitate targeted information extraction covering individualized information needs. Currently, Twitter only provides search through hashtags or keywords focusing on exact matches ordered largely by recency. This search critically fails to capture the underlying query intent when non-exact matches or more informative historical content may be more useful to the user than the most recent, exact matches. To build a more flexible search tool for Twitter, capable of learning a user's information needs from a small set of examples and generalizing to broader related content matching those information needs (including future events that could not have been anticipated at training time), I began to research the topic of learning social media sensors during my graduate studies at Oregon State University. This research has involved a variety of technical contributions ranging from designing novel paradigms for machine learning with social media through to the implementation of scalable Big Data frameworks leveraging Apache Spark on the Amazon EC2 Compute Cloud to analyze over 40 TB of Twitter data in short periods of time.

Statement of Intent
Zahra Iman

One of the key challenges of learning social media sensors is providing the appropriate training data for the model to learn. Considering the scale of the dataset and the diversity of information needs on Twitter, it is simply not possible (even with crowdsourcing) to manually annotate millions of Tweets and decide whether they are related to a specific topic. Thus in our work, we focused on Twitter hashtags as surrogates for topics and leveraged Snowball sampling methods originating in sociology and statistics to automatically generate labeled data and learn a user's information needs from minimal examples.

One of the additional challenges in this work, and in general when dealing with Big Data, is how to learn when the number of potential features (in the tens of millions) far exceed the number of positive examples (typically in the tens of thousands). This has analogues with computational biology where one must learn which gene sequences are predictive of human traits and diseases given relatively few data samples. The key to narrowing down the feature set to a tractable learning problem lies in leveraging knowledge of the social network and the social process of generating topical content. Leveraging such models to resolve these machine learning problems is the primary focus of my current research.

In future research, I intend to continue working on learning social media sensors by refining current learning paradigms to improve their usability. For example, one research direction is to work on real-time search within the highly limited Twitter search API, as opposed to offline learning and search with local batch content. Another research direction is to improve tweet ranking functions by focusing on identifying trendsetting tweets and users that have, or are likely to, lead to viral information cascades. Current research enables these further extensions, however these further extensions require intensive research themselves covering better learning methodologies and better annotation strategies such as taking advantage of semi-supervised learning techniques such as co-training.

Altogether, this proposed research provides insights into how to take advantage of large amounts of noisy, dynamic, and unstructured data provided by social media to create smarter cities with improved awareness of public needs ranging from urban infrastructure to emergency response to healthcare.

During my current research studies at Oregon State, I became familiar with the Information Engineering research area in the department of Mechanical and Industrial Engineering at the University of Toronto due to my PhD supervisor Dr. Scott Sanner's move there on April 1, 2016. The Information Engineering area, with the focus of using information technology to help people and organizations with innovation, overlaps strongly with my future research interests. My passion is researching methods to provide a better quality of life using Big Data. To this end, I have extensive experience and skills from various projects and courses taken during my current graduate studies. I believe that I will be a good fit to this group and I look forward to continuing my research along these lines with Dr. Sanner at the University of Toronto.