Research Statement

Naushad UzZaman
http://www.cs.rochester.edu/u/naushad

We, as humans, are collectively generating an unprecedented amount of data everyday. However, none of us can manually process this exponentially growing data fast enough to benefit from its existence. Researchers in the industry are building technology for using Big Data to better serve their customers. For example, Facebook, Amazon, Netflix and other companies use the data to suggest friends, products and movies to their users; Google and similar companies are providing personalized search results, targeted advertising, and many other services. My research interest is to utilize this impact of Big Data as it is used in the industry, to attempt to solve social problems.

While pursuing my PhD, I implemented a state-of-the-art temporal information extraction system for temporal evaluation shared task - *TempEval-2*. I also proposed a metric to evaluate temporal information which has been used in premier temporal evaluation challenges. Consequently, I had organized the temporal evaluation shared task - *TempEval-3*. Additionally, I have substantial amount of industry experiences in Big Data related research.

From my previous experiences and current interests, I see myself doing interdisciplinary research with researchers from Computer Science, Medical Informatics, Political Science, Social Science, Economics, Business and other disciplines. Some of the problems that I would like to work on are, predicting future trends using social media, automatically understanding the demographic information of social media users, clustering and summarizing tweets on a specific topic, and automatically finding what is missing in the social media discussions.

Research Accomplishments

Big Data/Data Science Related Research Accomplishments.

- **Explore the possibilities of using query logs and twitter for data mining at Nuance**:
  - Nuance gets millions of unrestricted text queries every day for personal assistant applications. It's a gold mine if we can automatically extract useful information from it. I implemented a system to automatically extract named entities from Nuance query logs and twitter by using a very small number of named entities (around 100) as seed examples, then learn the distinctively important common patterns for each named entity classes (e.g. Actor, Artist, etc.) and finally extracting named entities (e.g. Tom Cruise, Adele, etc.) from a large amount of unrestricted texts from our query logs and twitter. This approach enables us to automatically extract named entities, and also the popularity information of named entities.
  - I prototyped a system that can automatically cluster similar queries and entities. Instead of clustering based on words in the query, I clustered based on the entities being used in the queries. As a result queries like "find me the nearest <entity>" and "<entity> hours" can be clustered together, because they share similar entities, such as, Walmart, Walgreens, McDonald's, BestBuy, etc. I also explored the hierarchical clustering; the idea is, we can have all locations in one big cluster, then Countries in another sub-cluster, then Cities in another sub-cluster, etc, automatically. These cluster information can also be used as features for different classifiers in the system pipeline.
  - Future personal assistants need to infer queries like "Show me movies by the actress who tripped at Oscar". We can parse the query to understand "who tripped at Oscar" is the description of an actress. However, from the movie descriptions of existing movie data sources, such as IMDB[1], we cannot extract Jennifer Lawrence and others from the description "who tripped at Oscar". The reason is, movie descriptions do not include new and dynamic information. On the other hand, social media users discuss these dynamic events. I implemented a prototype to show that we can answer these questions very easily from Twitter when the existing data sources fail.

[1] www.imdb.com

- **Game Prediction with Social Media at Yahoo! Research, Barcelona**: Can we predict game outcome from social media? I explored the use of social media for game prediction [1]. Our system reliably extracted predictions from tweets, identified the predictive power of individual users, ranked predictions and aggregated them to find the most likely outcome of upcoming games. We evaluated our results on the 2010 FIFA World Cup tournament using a corpus of over 1.8 billion tweets containing over hundred-fifty thousand predictions. A key benefit of our system is that it does not rely on domain dependent knowledge for prediction, rather extracts people's predictions from social media. As a result, the framework can be used to predict the outcome of other sport events, elections, product release dates and any other future events discussed in the social media.

- **Information Extraction on Demand at Microsoft Medical Media Lab**: Can we build a system that can dynamically learn patterns from a few examples and extract pattern instances? I implemented an Information Extraction tool that can learn different patterns from user's examples and extract the pattern instances from natural language texts in medical documents, e.g. release notes, radiology reports, etc. For pattern matching, I used the medical ontology (UMLS) to understand medical patterns like diseases, medicines, etc. and other general language features. My implemented tool was shared with the clinical staff at Washington Hospital Center and they found it useful.

- **Data Mining Projects at Bosch RTC**: I worked on a large scale *data mining* problem for KDD cup 2008: Breast Cancer Detection[2]. I experimented with different machine learning techniques, e.g. Support Vector Machine (SVM), Decision Tree, KNN, Bayes Net, Neural Network, etc. Finally built the system using SVM and also implemented the Feature Selection technique to improve the performance and reduce the computation time. Additionally, I worked on a recommendation system for Netflix contest.

- **Bangladesh's performance in Millennium Development Goal**: After office hours, just for my personal interest, I explored World Bank data on Millennium Development Goals (MDG) indicators for Bangladesh. I tried to understand how well Bangladesh is doing in MDG goals, also from the data I calculated the most correlated indicators to find the positive and negative impacts of improving on each goals. Details about this work can be found at www.DecipherData.org.

Other Research Highlights while pursuing my PhD.

- **Temporal Information Extraction**: We implemented hybrid systems with linguistically motivated solutions and machine learning classifiers for extracting temporal information from raw text. We extracted events, temporal expressions and classified temporal relations. Our system had a state-of-the art performance in *TempEval-2* [2], [3], [4], [5].

- **Temporal Evaluation**: We proposed a new method for evaluating systems that extract temporal information from text. Our metric uses temporal closure to reward relations that are equivalent but distinct. It also measures the overall performance of systems with a single score, making comparison between different systems straightforward. Our intuitive and computationally inexpensive metric is used to evaluate participants in the premier temporal annotation shared task *TempEval-3* [2], [6], and also temporal information processing shared task on clinical data - *i2b2*[3].

- **Temporal Question Answering**: We proposed a temporal QA system that performed temporal reasoning and showed how it can be used to evaluate automated temporal information understanding [2], [7]. With temporal reasoning, our QA system can answer *list, factoid and yes/no questions. TempEval-3* organizers are discussing now to use this evaluation methodology for *TempEval-4*.

- **Multimodal Summarization (MMS)**: We worked on illustrating complex sentences as multimodal summaries combining pictures, simple sentence structure and summarized text [8]. We show that pictures alone are insufficient to help people understand most sentences, especially for readers who are unfamiliar with the domain. MMS could be used to help people with cognitive disabilities, children, older people, or people whose first language is not English [9].

[2]http://www.sigkdd.org/kdd-cup-2008-breast-cancer
[3]https://www.i2b2.org/NLP/TemporalRelations/Call.php

Future Research Directions

I see myself doing interdisciplinary research with researchers from computer science, medical informatics, political science, social science, economics, business and other disciplines. Following are a few suggested projects:

- **Prediction using Social Media**: A key benefit of our soccer game prediction system is that it does not rely on domain dependent knowledge for prediction, rather extracts people's predictions from social media. As a result, the framework can be used to predict the outcome of other sport events, elections, product release dates and any other future events discussed in the social media. This framework can extract explicit signals for different event predictions. On the other hand, social media data can be used for implicit signals for event predictions, such as, disease outbreak, political campaign outcome, etc.

  I want to explore both explicit and implicit signals for prediction of different events. I see myself working on this problem with researchers from Computer Science, Political Science, Social Science, Medical Informatics, Business School, and others.

- **Automatically deriving demographic information of Social Media users**: Most of the research based on social media has one important drawback, lacking demographic information. One approach in the literature is to use Mechanical Turk[4] to estimate the demographic information, such as, age, sex, location, from twitter profile picture. This is a great start but it is not scalable. For example, to collect the demographic information for tweets from 1 million users may cost, say $1,000, if we hire only one turker at 1 cent. Since the age (age range) is estimated from profile picture it might be appropriate to take the estimation from more than one turkers[5], which means the cost will go even higher. I want to tackle this problem of deriving demographic information as a data-driven problem with the following approaches:

  - We can predict user's location from friends' location, similarly we can use known friends demographic information (from mechanical turk) to predict unknown ones. Main hypothesis here is, friends[6] are usually from same demographics.
  - We can predict demographics from user's tweets. What we share tells a lot about who we are, where we live and what interests us.

  Combining these two approaches I want to address the problem of deriving demographic information of social media users. With this tool we can understand the demographics of social media users and we usually know the actual demographics information. We can then sample or weigh the tweets to have a similar distribution of actual demographics. All Big Data research on social media could be benefited from this work.

- **Clustering and Summarization of tweets on a specific topic**: Twitter is not just about understanding overall positive or negative sentiment about something or predicting some event. When we want to explore crowd's opinion about a topic, just extracting some tweets and photos are not enough. I want to explore building clustering and summarization tools for tweets. For a given topic, this tool would collect all social media discussions (tweets, Facebook posts, blog posts, etc.) on that topic, automatically cluster them in different sub-topics and show the representative tweets and the number of tweets as the summary.

- **What is missing in the social media discussions?**: We get most of our information now from social media. Researchers are building tools to find the most popular news in the social media to help us discover these information. This is very useful, however, the problem is, there are many important issues that are not being discussed in the social media. Let's take cancer as an example. We have facts from public data that lung cancer is killing the maximum number of people. If we analyze the social media data, we would have the impression that breast cancer kills the maximum number of people. It is really great that breast cancer gets a lot of social media discussions and it is creating awareness for this problem, but other important problems, such as, lung cancer, are being neglected. With location and demographic information, we can find out which problems are being neglected in a particular location or demographics.

[4] https://www.mturk.com
[5] Mechanical Turk workers.
[6] If both users follow each others can be considered as friends in Twitter.

Federal government, local government and anyone in the community can understand what needs to be discussed and being neglected. This approach is not just for cancer, it can be extended to social issues, political issues or any other issues.

- **Data science on health care and development sector**: Government, Local governments, Non Government Organizations (NGOs) and many other organizations, such as, World Bank, Health Data Consortium, have lots of data that they do not use effectively. If we can do analytics on these historical data, these organizations and other beneficiaries can make better decisions. A few sample applications based on World Bank data could be (concrete examples given for better understanding):
  - From historical data of Bangladesh and other countries, predict the future in the financial sector for Bangladesh.
  - Find out, 10 years back which countries were very similar to current Bangladesh. What are the changes after 10 years and what are the impacts in those countries. These facts can help governments to make necessary plans with specific pros and cons.
  - Which countries had potential but progress went downhill, finding from data what could be the reason for going downhill.

  Many large organizations or countries could be doing something very similar already. However, having access to a tool doing these analytics, many NGOs or small countries can be hugely benefited. I see myself working with researchers from Economics, Medical Informatics, Political Science, Social Science, etc, to create these analytics on different data sources.

To summarize, with my interests and experiences, I would like to apply Big Data analysis in the academia to find solutions to problems around us.

References

[1] Naushad UzZaman, Roi Blanco and Michael Matthews. "TwitterPaul: Extracting and Aggregating Twitter Predictions". *CoRR arXiv:1211.6496*, 2012. http://arxiv.org/abs/1211.6496.

[2] Naushad UzZaman. "Interpreting the Temporal Aspects of Language". *PhD Thesis, Department of Computer Science, University of Rochester*, July 2012.

[3] Naushad UzZaman and James F. Allen. "TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text". *International Workshop on Semantic Evaluations (SemEval-2010), Association for Computational Linguistics (ACL)*, Sweden, July 2010.

[4] Naushad UzZaman and James F. Allen. "Event and Temporal Expression extraction from raw text: first step towards a temporally aware system". *International Journal of Semantic Computing*, 2011.

[5] Naushad UzZaman and James F. Allen. "Extracting Events and Temporal Expressions from Text." *Fourth IEEE International Conference on Semantic Computing (IEEE ICSC2010)*, Pittsburgh, USA, September 2010.

[6] Naushad UzZaman and James F. Allen. "Temporal Evaluation". *Proc. of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Short Paper)*, Portland, Oregon, USA, June 2011.

[7] Naushad UzZaman, Hector Llorens and James Allen. "Evaluating Temporal Information Understanding with Temporal Question Answering". *Proceedings of IEEE International Conference on Semantic Computing, Italy, September 2012.*

[8] Naushad UzZaman, Jeffrey P. Bigham and James F. Allen. "Multimodal Summarization of Complex Sentence". *Proc. of International Conference on Intelligent User Interfaces (IUI)*, Palo Alto, California, 2011.

[9] Naushad UzZaman, Jeffrey P. Bigham and James F. Allen. "Multimodal Summarization for people with cognitive disabilities in reading, linguistic and verbal comprehension". *2010 Coleman Institute Conference*, Denver, CO, 2010.

[10] Naushad UzZaman, Jeffrey P. Bigham and James F. Allen. Pictorial Temporal Structure of Documents to Help People who have Trouble Reading or Understanding. *International Workshop on Design to Read, ACM Conference on Human Factors in Computing Systems (CHI)*, Atlanta, GA, April 2010.