Statement of Research

Muhammad Haroon Shakeel
m.shakeel@lums.edu.pk

Overview

In recent years, the short text is being generated in abundance on social media. The users tend to share their life events on such platforms. Natural language processing (NLP) on such posts can provide useful insights with applications in human-centered computing, product development, managerial decision making etc. However, short length of texts make the NLP tasks challenging. Furthermore, posts on social media are often influenced by regional language, generating distinct dialects of communication that do not follow any grammatical rules and tend to code-switch between multiple languages mid utterances. Being international language, English is well-studied under the umbrella of NLP while regional languages are under-resource in terms of language models, datasets, and external knowledge base. This becomes a major bottleneck in the usefulness of several NLP tasks involving social media such as sentiment analysis, depression detection, domestic violence detection, early suicidal tendency detection to name a few.

Building on these observations, firstly, my research is focused on addressing the challenges in short text to perform NLP tasks involving human-centered computing and simultaneously focus on under-resourced languages. In recent years, deep learning has revolutionized the modeling and understanding of human language. However, its usefulness is still limited to English language in majority of the cases. The second direction of my research is to make most common human-centered computing tasks language independent. I have 6 publications in this direction. I also collaborate with *Computer Vision and Graphics Lab* at *Lahore University of Management Sciences (LUMS)* on medical imagery projects. Furthermore, I also have submitted two manuscripts for publication on efficient data analytics in collaboration with *Data Analytics Lab* at LUMS (details attached in the CV).

Research Plan

In the next five years, I would like to concentrate on three research themes: (1) Human-centered computing, (2) NLP for under-resourced languages, and (3) performing NLP tasks in language oblivious fashion.

(1) *Human-centered computing:* Majority of the information on social media is shared about an event, personality or product of interest. This information is reflexive of one's personal feelings, emotions and sentiments towards such topics, hence has many applications in human-centered computing such as depression detection, suicidal tendency detection, religious extremism detection or epidemic detection etc. I plan to continue research in this direction with focus on utilizing deep learning architectures to achieve the goals.

(2) *NLP for under-resourced languages*: Languages other than English suffer from deficiency of datasets, language models and knowledge base such as lexical databases. I plan to extend my current research work on under-resourced languages.

(3) *Language independent NLP*: In majority of the state-of-the-art algorithms for different NLP applications, the models are specifically tailored towards a particular language. However, in real setting on social media, text is generated in multiple, informal and code-switched languages where users tend to use self-created abbreviations, jargons, improper grammar. These challenges limit the usefulness of language dependent models. Thus, I plan to work on research on development of models that are language oblivious.