RESEARCH STATEMENT

Vivek Kulkarni, Stanford University

My research seeks to develop models that incorporate the social/human context of language to enable human-centric, robust, fair natural language processing applications. Natural Language Processing (NLP) applications are now ubiquitous and used by millions of individuals world-wide. Yet, these applications are overwhelmingly ***brittle,* and *biased*.** For example, the accuracy of syntactic parsing models drops by at-least 20 points on African American Vernacular English when compared to Standard English. Sentiment analyzers fail on language from different time periods, question answering systems fail on British English, conversational assistants struggle to interact with millions of old age people who have speech disabilities and hate speech detection systems are biased and more likely to incorrectly classify language from specific demographics as offensive. In short, NLP models and applications work well only for a minority of the population, effectively excluding a significant majority.

I argue that this crippling brittleness of NLP models stems from treating language as if it were devoid of human and social context. Language is fundamentally a human endeavor, where social context and human factors are critical to language understanding – context that should be modeled to make NLP **robust, human-centric, and socially aware**. I propose methods to reliably detect linguistic variation in a variety of social contexts and build models that are robust to such variation [1, 2, 3]. Similarly, I develop methods to add the crucially missing but critical **"human element" in NLP** models exploiting insights from inter-disciplinary sciences including sociolinguistics, and psychology [4, 5].

Finally**, I develop NLP methods to uncover social biases.** Language and society are inextricably linked. By analyzing language use at scale, we can uncover societal biases, and reveal potential mechanisms contributing to such biases. In this vein, leveraging my expertise in NLP and computational social science I have developed NLP methods to analyze hate speech [6], detect political ideology of news articles [7], uncover gender biases in media coverage [8], as well as in academia [9]. I elaborate on some of my research findings and outline my future agenda below.

Making NLP models robust to language variation

Language varies across many dimensions: (a) time, geography, domains and (b) human factors like age, gender. Yet, a surprising number of modern NLP models are trained on language from a specific context (like language from the Wall Street Journal in the 1980's). It is no surprise therefore, that these models are brittle and perform poorly in a different context (like African American Vernacular English). Naïvely training specific models on every possible subset is neither optimal nor practical. *So, how do we make NLP models robust to linguistic variation (to both social contexts and human factors)?*

I investigate two specific approaches: (a) learning context-aware linguistic representations that reliably encode linguistic variation (across dimensions like time, geography, and domain) and (b) incorporating human factors (both observed and latent) to better guide models.

I propose methods based on representation learning and robust change-point detection to learn context-sensitive word representations (embeddings) that reliably encode linguistic variation across many contexts [1, 2, 3]. For example, my methods detect that *"lean"* acquired a positive sentiment over the last century (see Figure 1). In the 1900's *"lean"* was associated with *"weak"*, but in the 2000's it is associated with *"lean muscle"* and *"strength"* – a nuance that is critical to capture for accurate sentiment classification.

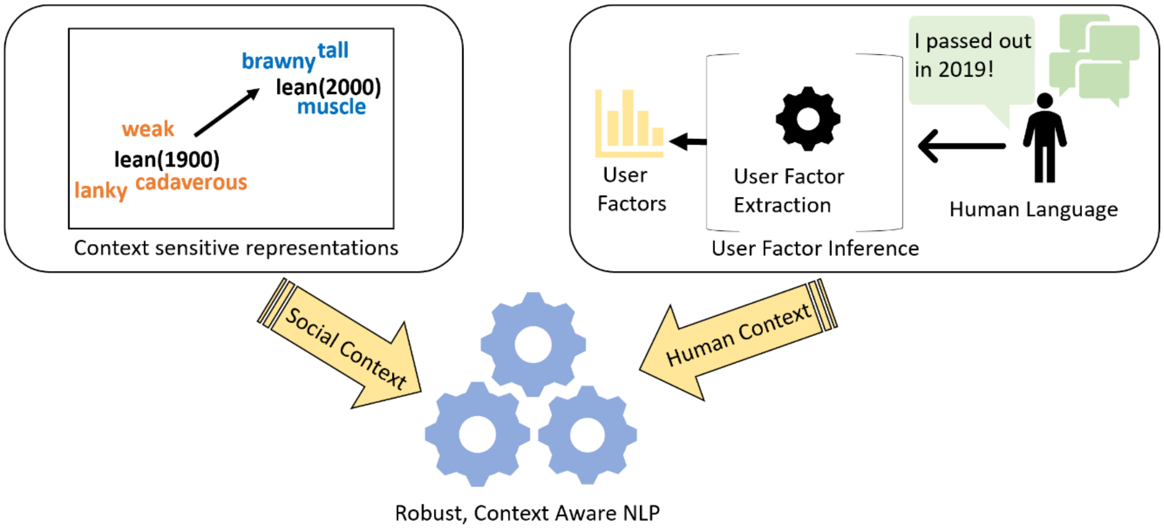

Figure 1: *Context sensitive representations model linguistic variation in varying social contexts while latent human factors inferred directly from language incorporate human context.*

These context-sensitive word embeddings reliably detect instances of linguistic variation and provide explicit cues to NLP applications. For example, detecting named entities in text from various domains (Sports, or Finance) is an important task in NLP. I show that learning domain specific word embeddings using my methods improves the performance of such named entity detection systems [3]. My contributions have been highly influential scientifically (**>230 citations**), are widely used in research labs (Yahoo! Research), and received extensive media coverage (**MIT Tech Review, ACM Tech Review, VICE**).

Next**,** I explore incorporating human/user factors (observed or latent) into NLP models. For example, in an open-domain question answering setting, the question "*how to remove parent from children?*" has two very different interpretations depending on whether the person interacting is a software engineer (who is manipulating tree structures) or not. Similarly, for sentiment analysis, the statement *"I passed out in 2019"* can have positive or negative sentiment depending on whether the speaker is a 21-year-old Indian student (where *pass out* means *to graduate*) or an 85-year-old American (where *pass out* means *to faint*). It is difficult to identify what specific human attributes need to be modeled, given new downstream NLP applications. In my work, I therefore adopt a different approach -- "*Can we computationally learn a small set of broad dimensions that capture meaningful differences among users directly from their prior language use, that generalize well to a variety of new downstream tasks?"* Leveraging collaborations with psychologists, I proposed a method to infer *"user factors"* (*see Figure 1*) directly from their language [4]. These factors correlate with survey based Big5 personality factors and capture meaningful differences among people. They are also predictive of a variety of *unforeseen outcomes*. Finally, we show how to incorporate these factors in NLP models to boost performance of models on tasks like sentiment classification and sarcasm detection [5].

NLP for uncovering social bias

Language is a mirror into society and thus reflects societal biases. This is particularly problematic for machine learning models that not only encode these biases but can also perpetuate and amplify them. Moreover, millions of people use social media platforms where abusive and prejudiced language can significantly harm individuals and society. Consequently, it is critical to develop NLP models to reliably detect and precisely characterize societal biases and prejudices. In my research, I investigate two manifestations of such biases (a) hate speech and (b) political slant of online news articles. Specifically, we reveal nuances in hate speech and show that it can be directed or generalized, and that these differ in their linguistic properties (including lexical diversity and semantic framing) [6]. Similarly, I also propose a new deep-learning model for political ideology detection of online news articles that leverages non-textual cues in news articles (like their hyper-link structure) in addition to textual features for improved detection of political slant of news articles [7].

Collaborating with sociologists at McGill University, I developed and applied NLP tools for gender classification, entity detection, and sentiment analysis to track media coverage of entities. We observed that despite an increased participation of women in social life, media coverage is disproportionately skewed towards males, and find that societal level inequalities are the largest factor contributing towards this biased coverage [8]. This work was published in **American Sociological review**, the top journal in Sociology, won the **CITAMS Best Paper Award** and received media coverage in **The Guardian**.

At Stanford, I have collaborated with social scientists at the Department of Education to develop NLP methods to detect and characterize new scientific innovations introduced in academic literature from a large-scale analysis of millions of academic texts. These NLP methods enable us to investigate the link between author demographics, innovation patterns and subsequent academic impact thus uncovering factors potentially explaining the under-representation of diverse individuals in academia [9].

Research Agenda

In my research so far, I have focused on human centric NLP and NLP for uncovering social biases. Looking into the future, I seek to deepen my focus on Human Centric NLP and broaden the scope of my research in NLP for uncovering social biases. I discuss specific research directions below.

Socially Intelligent Human Centric NLP

Representation learning of contextual factors. A key piece to developing human centric, personalized, and socially aware NLP models is to learn powerful representations of social and human contexts. In the broad context in which language is situated some factors may be explicit and observed (like time, geography, domain or user demographics) while several are implicit. Such implicit factors include common sense/world knowledge, the underlying social network, and finally individual factors (like personality, socio-economic status, education level). How can we encode these contextual factors in generalizable representations that can be used on new unforeseen downstream models? I believe that it is promising to leverage new advances in representation learning (like pre-trained language models, or graph convolutional networks) for this. I have already explored learning contextual representations of entities and relations in knowledge bases using novel deep learning algorithms [10] that learn generalizable entity and relation embeddings that can be effectively finetuned for down-stream knowledge graph completion tasks. My prior work [11, 12] has also explored learning representations of nodes in social networks. I look forward to exploring such techniques to encode real world knowledge, embed the user's underlying social network, and other latent factors in downstream NLP models.

Robustness to adversarial attacks and interpretability. While human-centric NLP that robustly adapts to dynamically changing contexts can positively affect millions of people, this very flexibility and adaptability can also make it susceptible to adversarial attacks. I therefore believe it is important to investigate the robustness of NLP models to adversarial inputs especially when used in the broader social context. This is especially important in the case of dialog agents and conversational assistants who may end up responding in racist and offensive language due to interactions with an adversary. I believe that certifiable robustness to specific classes of input and better model interpretability are promising ways forward. Therefore, in the future I would like to expand my research to characterize the robustness of NLP systems to adversarial input and ideally provide certifiable guarantees for certain classes of inputs.

NLP for Science of Science

I seek to broaden my prior work on developing NLP tools for uncovering social biases in online media to developing methods for studying social fields through the lens of language. In the short term, I would like to turn my attention to the social field of Science itself. The scientific community like any other social context has its own norms, expectations and social dynamics. Characterizing the structure and dynamics of this field will have tremendous impact on uncovering social aspects that govern scientific practice thus shaping policy decisions that impact science and in turn society. Academic publications and grant applications reflect these dynamics and thus provide an excellent venue of investigation using NLP. In addition to societal implications, such a large-scale analysis of scientific literature requires development of novel NLP methods which can detect field-specific scientific innovations, various kinds of scientific claims, citation contexts, and model how prior work, and novel contributions are framed. My work with social scientists at Stanford [9] has already developed computational methods to detect revolutions in science, discovered innovation patterns of researchers, as well as uncovered mechanisms that explain persistent under-representation of diverse individuals in academia.

To conclude, in the long term I am convinced that making NLP human-centric and socially intelligent is one of the next big challenges of AI and exciting research questions lie at the intersection of natural language processing, machine learning and the inter-disciplinary sciences – a point towards which my research vision is fixed.

References

[1] V. Kulkarni, R. Al-Rfou, B. Perozzi and S. Skiena, "Statistically significant detection of linguistic change," in *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web*, 2015.

[2] V. Kulkarni, B. Perozzi and S. Skiena, "Freshman or fresher? Quantifying the geographic variation of language in online social media," in *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, 2016.

[3] V. Kulkarni, Y. Mehdad and T. Chevallier, "Domain Adaptation for Named Entity Recognition in Online Media with Word Embeddings," 2016.

[4] V. Kulkarni, M. L. Kern, D. Stillwell, M. Kosinski, S. Matz, L. Ungar, S. Skiena and H. A. Schwartz, "Latent human traits in the language of social media: An open-vocabulary approach," *PLoS ONE,* 2018.

[5] L. Veronica, S. Youngseo, K. Vivek, B. Niranjan and S. H. Andrew, "Human Centered NLP with User-Factor Adaptation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 2017.

[6] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," in *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 2018.

[7] V. Kulkarni, J. Ye and S. Skiena, "Multi-view Models for Political Ideology Detection of News Articles," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels.

[8] E. Shor, A. van de Rijt, A. Miltsov, V. Kulkarni and S. Skiena, "A Paper Ceiling: Explaining the Persistent Underrepresentation of Women in Printed News," *American Sociological Review,* 2015.

[9] B. Hofstra, S. Galvez, B. He, V. Kulkarni and M. Daniel, "Diversity Breeds Innovation With Discounted Impact and Recognition," in *(In submission)*, 2019.

[10] W. Haoyu, V. Kulkarni and W. Y. Wang, "DOLORES: Deep Contextualized Knowledge Graph Embeddings," in *In submission*, 2019.

[11] B. Perozzi, V. Kulkarni, H. Chen and S. Skiena, "Don’t walk, skip! online learning of multi-scale network embeddings," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, 2017.

[12] H. Yu, V. Kulkarni and W. Y. Wang, "MOHONE: Modeling Higher Order Network Effects in KnowledgeGraphs," in *In submission*, 2018.

[13] V. Kulkarni and W. Y. Wang, "Simple Models for Word Formation in English Slang," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.