Returning the N to NLP: Towards Contextually Personalized Classification Models

Lucie Flek
Mainz University of Applied Sciences
Germany
lucie.flek@hs-mainz.de

Abstract

Most NLP models today treat language as universal, even though socio- and psycholingustic research shows that the communicated message is influenced by the characteristics of the speaker as well as the target audience. This paper surveys the landscape of personalization in natural language processing and related fields, and offers a path forward to mitigate the decades of deviation of the NLP tools from sociolingustic findings, allowing to flexibly process the "natural" language of each user rather than enforcing a uniform NLP treatment. It outlines a possible direction to incorporate these aspects into neural NLP models by means of socially contextual personalization, and proposes to shift the focus of our evaluation strategies accordingly.

1 Introduction

Our language is influenced by one's individual characteristics as well as by the affinity to various sociodemographic groups (Bucholtz and Hall, 2005; McPherson et al., 2001; Eckert and McConnell-Ginet, 2013). Yet the majority of NLP models today treats language as universal, acknowledging that words have different meanings in different semantic context, but typically assuming that this context has the same meaning for everyone. In this paper, I propose that our focus shifts towards interpreting the language together with its user-dependent, contextual personal and social aspects, in order to truly process the "natural" language of a user. I outline a possible direction to incorporate these aspects into neural NLP models, and suggest to adjust our evaluation strategies.

The paper is structured with the following aims in mind: Sec. 2 provides historical context, seeking evidence on personalization needs. Sec. 3 reviews existing personalization work, as the personalization efforts and success stories are scattered across contributions to various applied tasks. Sec. 4 contemplates on how NLP personalization could be adopted as a process of several stages. Sec. 5 outlines an implementation proposal on contextually personalized classification models, building upon flexible, socially conditioned user representations. Sec. 6 proposes novel evaluation approaches reflecting the benefit of personalized models. Finally, Sec. 7 opens the discussion on ethical aspects, non-personalizable NLP tasks, and the role of industry in personal data collection and protection.

2 Historical context

Since 1990s, with the rise of so-called *empirical* or *statistical NLP* area (Manning et al., 1999; Brill and Mooney, 1997), the focus on frequently appearing phenomena in large textual data sets unavoidably led to NLP tools supporting "standard English" for generic needs of an anonymous user. An NLP tool - whether e.g. a POS tagger, dependency parser, machine translation model or a topic classifier - was typically provided as one trained model for one language (Toutanova et al., 2003; Klein and Manning, 2003; Morton et al., 2005), or, later on, for major underperforming domains, such as Twitter (Gimpel et al., 2011). However, enforcing artificial domain boundaries is suboptimal (Eisenstein, 2013). Neglecting the variety of users and use cases doesn't make the tools universally applicable with the same performance - it only makes our community blind to the built-in bias towards the specifics of user profiles in training data (Hovy, 2015; Tatman, 2017).

Meanwhile, in the information retrieval area, personalization has been incorporated from the early days - it is a long accepted paradigm that different users with different information needs might search for that need using the same query (Verhoeff et al., 1961) and that individual information needs evolve (Taylor, 1968). With the rising popularity

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7828–7838
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

of search engines in 1990s, the need for personalization in the interpretation of the query becomes obvious (Wilson, 1999). Exploiting logs of user search interactions allowed personalization at scale (Carbonell and Goldstein, 1998; Sanderson and Croft, 2012). In 2000s, it became acceptable to personalize search results using implicit information about user's interests and activities, e.g. leveraging browsing history or even e-mail conversations (Teevan et al., 2005; Dou et al., 2007; Matthijs and Radlinski, 2011). Today, hardly any of us can imagine that searching e.g. for *pizzeria* from our cell phone would return the same list of results for everyone no matter our location.

The area of recommendation systems has followed the IR trends, with more emphasis on the social than the personal component. Already early GroupLens Usenet experiments (Miller et al., 1997; Resnick et al., 1994) have shown the effectiveness of personalized article recommendations via collaborative filtering. Acknowledging the potential of personalizing via similar or related users, the focus moved towards exploiting information from user's social networks (Guy et al., 2010; De Francisci Morales et al., 2012; Guy et al., 2009).

Similar developments are emerging for example in the area of personalized language models (Ji et al., 2019; Wen et al., 2012; Yoon et al., 2017; McMahan et al., 2017), which are largely used e.g. in predictive writing, and in natural language generation (Oraby et al., 2018; Harrison et al., 2019), aiming e.g. at selecting and preserving a consistent personality and style within a discourse.

Drawing inspiration from these areas, I argue it is natural for users to expect personalized approaches when an NLP system attempts to interpret their language, i.e., attempts to assign any label to a provided text segment, whether it is, e.g., a sentiment of their sentence, a part-of-speech of a word they used, a sense definition from a knowledge base, or even a translation. As I discuss in the following section, already basic personal information has been shown to be relevant for the system accuracy.

3 User traits and NLP models

Inferring user traits We adjust our language with respect to the sociodemographic group we feel related to (McPherson et al., 2001; Bucholtz and Hall, 2005; Holmes and Meyerhoff, 2008; Eckert, 2012). This language adjustment can be, in turn, used in NLP algorithms to infer a range of individual user traits. Experiments have been conducted with estimating variables such as age (Rao et al., 2010; Nguyen et al., 2011), gender (Burger et al., 2011; Bamman et al., 2014; Sap et al., 2014), geolocation (Eisenstein et al., 2010), political preferences (Volkova et al., 2014), socio-economic status (Preoţiuc-Pietro et al., 2015), impact (Lampos et al., 2014), and a range of psychological traits and issues (Schwartz et al., 2013; Park et al., 2015; Sumner et al., 2012; Guntuku et al., 2017; Coppersmith et al., 2014). While most of the above-listed experiments have been conducted on Twitter, a variety of other datasets have been used, including phone conversations (Mairesse et al., 2007; Ivanov et al., 2011), blogs (Mukherjee and Liu, 2010; Schler et al., 2006), Facebook (Markovikj et al., 2013), or YouTube (Filippova, 2012). Human judges show surprisingly inferior performance on user profiling tasks, grounding their judgement in topical stereotypes (Carpenter et al., 2017). However, albeit more accurate thanks to capturing stylistic variation elements, statistical models are prone to stereotype propagation as well (Costa-jussà et al., 2019; Koolen and van Cranenburgh, 2017).

While many experiments have been conducted using discrete variables for demographics and personality, real-valued continuous representation are preferable (Lynn et al., 2017). Numerous researchers have been pointing out that it would be more meaningful to create models building on recent developments in sociolinguistics, i.e. treating demographic variables as fluid and social, e.g. modeling what influences speakers to show more or less of their identity through language, or jointly modeling variation between and within speakers (Eckert and McConnell-Ginet, 2013; Nguyen et al., 2014; Bamman et al., 2014; Eisenstein, 2013).

Improving NLP tasks with user traits Actively accounting for sociodemographic factors in text classification models leads to improved performance across NLP applications. So far, such studies have being conducted most prominently for English language, using age and gender variables, with the most focus on sentiment analysis tasks (Volkova et al., 2013; Hovy, 2015; Lynn et al., 2017; Yang and Eisenstein, 2017). Other explored tasks include topic detection, part-of-speech tagging (Hovy, 2015), prepositional phrase attachment, sarcasm detection (Lynn et al., 2017), fake news detection (Long et al., 2017; Potthast et al., 2018), or detection of mental health issues (Benton et al., 2016). Apart from demographic variables, personality traits play a role as well - e.g. in stance detection (Lynn et al., 2017), sarcasm detection, opinion change prediction (Lukin et al., 2017), prediction of regional life satisfaction or mortality rate (Zamani et al., 2018). NLP models can also improve by exploiting user's past context and prior beliefs, e.g. for sarcasm (Bamman and Smith, 2015), stance prediction (Sasaki et al., 2018), persuasion (Durmus and Cardie, 2018) or conversation re-entry (Zeng et al., 2019). Methods used to incorporate the social and psychological variables to models are discussed in Sec. 5.

Improving NLP tasks with social graphs An emerging line of research makes use of social interactions to derive information about the user - representing each user as a node in a social graph and creating low dimensional user embeddings induced by neural architecture (Grover and Leskovec, 2016; Qiu et al., 2018). Including network information improves performance on profiling tasks such as predicting user gender (Farnadi et al., 2018) or occupation (Pan et al., 2019), as well as on detecting online behavior such as cyberbullying (Mathur et al., 2018), abusive language use (Qian et al., 2018; Mishra et al., 2018) or suicide ideation (Mishra et al., 2019).

4 NLP personalization as a process

From the user experience perspective, personalization of NLP tools could be divided into three steps.

Explicit input. In the first step, user is allowed to provide personal information for the NLP components explicitly. The depth of information provided can vary from specifying own age to taking personality questionnaires. This user behavior is somewhat similar to subscribing to topics of interest for personalized newsletters - user has a full control over the level of customization. However, results of increasing the burden on the user can be inferior to implicit inference (Teevan et al., 2005).

Implicit inference. More conveniently, personal information about the user can be inferred implicitly by the system, as demonstrated e.g. by the models discussed in section 3. The result of such inference can be either a set of explicit labels, or latent user representation capturing similar information in a larger number of data-driven dimensions. For the user, such personalization might currently feel intrusive in the context of an NLP system, however, in many related research areas the user expectations are already altered (cf. Sec. 2).

Contextualized implicit inference. In the third step, personalization includes also an intrauser modeling of different individual contexts based on user's communication goals. This reflects the social science argument that an identity is the product rather than the source of linguistic and other semiotic practices, and identities are relationally constructed through several, often overlapping, aspects of the relationship between self and other, including similarity/difference, genuineness/artifice and authority/delegitimacy (Bucholtz and Hall, 2005). This approach is also aligned with NLP findings on social power in dialogue (Bracewell et al., 2012; Bramsen et al., 2011; Prabhakaran et al., 2012). Such solution can be perceived less invasive by the users, as the contextual adaptation may diminish the otherwise built-in stereotypes of language use (e.g. some users may prefer to use more emotionally charged words in private social contexts, but not necessarily in professional conversations).

5 Methods of incorporating psychosocial profiles into NLP models

Early experiments used basic demographic variables directly as input features in the model (Volkova et al., 2013). Hovy (2015) uses age and gender as modifying factors for the input word embeddings. In a similar manner, Lynn et al. (2017) uses a multiplicative compositional function to combine continuous user trait scores, inferred via factor analysis, with original feature values, augmenting the feature set so that each feature exists with and without the trait information integrated. Benton et al. (2017) use age and gender as auxiliary tasks in a multitask learning setup for psychological labeling of users. Zamani and Schwartz (2017) apply a residualized control approach for their task, training a language model over the prediction errors of the model trained on sociodemographic variables only. Later they combine it with the factor analysis approach (Zamani et al., 2018). Benton et al. (2016) learns user representations by encoding user's social network as a vector, where users with similar social networks have similar vector representations. A commonly used technique is to define the "context" for each node, for example by random walks, and train a predictive model to perform context prediction.Similar network-based learning is employed in node2vec (Grover and Leskovec, 2016).

Yang and Eisenstein (2017) propose to use neural attention mechanisms in a social graph over followers, mentions and retweets, to leverage linguistic homophily.

However, the user modeling approaches discussed so far focus on finding one representation for one user. A modern, personalized NLP system shall be able to capture not only the inherent semantic aspects of the analyzed discourse together with the latent vectorial representations of user characteristics, but also contextual user profiles based on an identity sought in their current social microenvironment. A strengthened industry-academia cooperation is crucial in such data collection (more on this in Sec. 7). Assuming the access to a larger online history of each user, we could draw a parallel to the design of the contextual word embeddings (Peters et al., 2018; Howard and Ruder, 2018; Devlin et al., 2019), which train neural networks as language models, then use the context vectors provided for each word token as pretrained word vectors. With an increasing number of online corpora containing user metadata, we can use recurrent or attentive neural networks to create large-scale social representations of users in a similar manner, allowing multiple pretrained "senses" of each user identity - vector representations of user conversational styles, opinions, interests, etc., treating those representations as dynamically changing in different social contexts. These representations can be then matched to new users based on the sparse linguistic, sociodemographic, psychological, and network information available, and fine-tuned on the context of a given task in a given social microenvironment, e.g. based on the stable part of the personal vectorial representation of the other users present in the conversation.

6 Evaluation

Currently, most of the NLP ground truth exists in the vacuum, "for everyone". Our systems typically use labels obtained as an average or majority vote provided by a number of impersonated annotators, even for tasks where they highly disagree (Waseem, 2016; Stab and Gurevych, 2014). As pointed out in Bender and Friedman (2018), we rarely get to know anything about the people other than if they were "expert"[1]. If we truly aim at personalizing NLP systems, the first step is understanding who the recipients of our system decisions are. In contrast to IR, where the user of the interpreted result is normally the author of the query, in NLP the use cases vary. For example, rather than merely labeling a piece of text as a "sarcasm", we shall ask *(A) Did the author mean this statement as sarcasm? (B) Was this understood by others as sarcasm? What kind of users interprets this statement as sarcasm?*

In the tasks of type A, it is sensible to ask the authors themselves about the intended label (e.g. *Are we correct this was a joke / positive review / supportive argument?*. We shall further assess the value of the system personalization. E.g. a user may prefer a model that correctly interprets her sarcasm even when most annotators typically don't recognize it. We can take inspiration from subjective measures used in evaluating spoken dialogue systems, such as A/B testing (Kohavi et al., 2014), customer satisfaction (Kelly et al., 2009; Kiseleva et al., 2016) or interestingness (Harrison et al., 2019; Oraby et al., 2018).

Yet most of the tasks are of type B, where we implicitly try to label how a piece of text is perceived by others (e.g. hate speech, assertiveness, persuasiveness, hyperpartisan argumentation). Given that these "others" vary in their judgments (Kenny and Albright, 1987) and this variation is informative for NLP models (Plank et al., 2014; Chklovski and Mihalcea, 2003), I suggest we start caring in NLP explicitly about who these "others" are, and evaluate our models with respect to labels assigned by defined target groups of users (e.g. with regards to sociodemographics, personality, expertise in the task) rather than one objective truth. Initial exploration of this area has been started e.g. for perceived demographics (Volkova and Bachrach, 2016; Carpenter et al., 2017) and natural language inference (Pavlick and Kwiatkowski, 2019).

7 Ethical considerations

The ability to automatically approximate personal characteristics of online users in order to improve language understanding algorithms requires us to consider a range of ethical concerns.

Unfair use prevention It is almost impossible to prevent abuse of once released technology even when developed with good intentions (Jonas, 1983). Hence it may be more constructive to strive for an informed public, addressing the dual use danger with a preemptive disclosure (Rogaway, 2015; Hovy and Spruit, 2016) - letting potential abusers know that certain illegal and unethical purposes of

[1] read: undergrad students vs. lab colleagues

using personalized models are not supported, and letting potential users know about the risk. For example the European Ethics Guidelines for Trustworthy AI foresee that “Digital records of human behaviour may allow AI systems to infer not only individuals’ preferences, but also their sexual orientation, age, gender, religious or political views.” and claim that “it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them.”

Incorrect and stereotypical profiling Sociodemographic classification efforts risk invoking stereotyping and essentialism. Such stereotypes can cause harm even if they are accurate on average differences (Rudman and Glick, 2012). These can be emphasized by the semblance of objectivity created by the use of a computer algorithm (Koolen and van Cranenburgh, 2017). It is important we control for variables in the corpus as well as for own interpretation biases.

Privacy protection Use of any data for personalization shall be transparent. Even public social media data shall be used with consent and in an aggregated manner, no individual posts shall be republished (Hewson and Buchanan, 2013). Regarding explicit consent, research shall take account of users’ expectations (Williams et al., 2017; Shilton and Sayles, 2016; Townsend and Wallace, 2016). Similar issue is discussed by Smiley et al. (2017) regarding NLG ethics, as NLG systems can incorporate the background and context of a user to increase the communication effectiveness of the text, but as a result may be missing alternative views. They suggest to address this limitation by making users aware of the use of personalization, similar to addressing provenance.

Role of industry and academia in user data collection Privacy and controllability is an auxiliary task to personalization and adaptation (Torre, 2009). Strictly protecting user privacy when collecting user data for model personalization is of utmost importance for preserving user trust, which is why, perhaps counter-intuitively, I encourage stronger industry-academia collaborations to facilitate a less intrusive data treatment. An inspiration can be taken from the concept of differential privacy (Dwork, 2008), applied e.g. in the differentially private language models (McMahan et al., 2017), which allow to customize for the user without incorporating her private vocabulary information into the public cloud model. Similarly, doing academic research on personalized NLP classification tasks directly within industry applications such as mobile apps with explicit user consent would enable transparent experiments at scale, being potentially more secure than gathering and manipulating one-time academic data collections offline. It may also contribute to better generalizability of the conclusions than strictly academic case studies that are typically limited in scale.

Personalization as a harmful ambiguity layer Given the field bias to reporting personalization results only when successful, no “unpersonalizable” tasks have been defined so far. With that, one question remains open - can we benefit from personalization everywhere across NLP, or are there cases where subjective treatment of a language is not desired, or even harmful? E.g., a legal text shall remain unambiguous to interpretation. On the other hand, the ability to understand it is subjective, and some users may appreciate lexical simplification (Xu et al., 2015). Are there objective NLP tasks as such, or can we segment all of those into an objective and subjective part of the application?

8 Conclusion

Building upon Eisenstein (2013); Lynn et al. (2017), and Hovy (2018), I argue that, following the historical development in areas related to NLP, users are ready also for the personalization of text classification models, enabling more flexible adaptation to truly processing their “natural” language rather than enforcing a uniform NLP treatment for everyone. Reflecting the current possibilities with available web and mobile data, I propose to expand the existing user modeling approaches in deep learning models with contextual personalization, mirroring different facets of one user in dynamic, socially conditioned vector representations. Modeling demographic and personal variables as dynamic and social will allow to reflect the variety of ways individuals construct their identity by language, and to conduct novel sociolinguistic experiments to better understand the development in online communities. I suggest to also shift the focus of our evaluation strategies towards the individual aims and characteristics of the end users of our labeling models, rather than aggregating all variations into objective truths, which will allow us to pay more attention to present social biases in our models.

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on Twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Adrian Benton, Raman Arora, and Mark Dredze. 2016. Learning multiview embeddings of twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Berlin, Germany. Association for Computational Linguistics.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

David Bracewell, Marc Tomlinson, and Hui Wang. 2012. Identification of social acts in dialogue. In *Proceedings of COLING 2012*, pages 375–390, Mumbai, India. The COLING 2012 Organizing Committee.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Portland, Oregon, USA. Association for Computational Linguistics.

Eric Brill and Raymond J Mooney. 1997. An overview of empirical natural language processing. *AI magazine*, 18(4):13–13.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics.

Jaime G Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, volume 98, pages 335–336.

Jordan Carpenter, Daniel Preotiuc-Pietro, Lucie Flekova, Salvatore Giorgi, Courtney Hagan, Margaret L Kern, Anneke EK Buffone, Lyle Ungar, and Martin EP Seligman. 2017. Real men don't say "cute" using automatic language analysis to isolate inaccurate aspects of stereotypes. *Social Psychological and Personality Science*, 8(3):310–322.

Timothy Chklovski and Rada Mihalcea. 2003. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *Proceedings of RANLP 2003*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Marta R Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster. 2019. Proceedings of the first workshop on gender bias in natural language processing. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. 2012. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 153–162. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web*, pages 581–590. ACM.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287. Association for Computational Linguistics.

Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. 2018. User profiling through deep multimodal fusion. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 171–179, New York, NY, USA. ACM.

Katja Filippova. 2012. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1478–1488, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA. ACM.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Ido Guy, Naama Zwerdling, David Carmel, Inbal Ronen, Erel Uziel, Sivan Yogev, and Shila Ofek-Koifman. 2009. Personalized recommendation of social software items based on social relations. In *Proceedings of the third ACM conference on Recommender systems*, pages 53–60. ACM.

Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. 2010. Social media recommendation based on people and tags. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201. ACM.

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn Walker. 2019. Maximizing stylistic control and semantic accuracy in NLG: Personality variation and discourse contrast. *arXiv preprint arXiv:1907.09527*.

Claire Hewson and Tom Buchanan. 2013. Ethics guidelines for internet-mediated research. The British Psychological Society.

J Holmes and M Meyerhoff. 2008. The handbook of language and gender (vol. 25). *Hoboken, NJ: Wiley*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 42–49.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Alexei V Ivanov, Giuseppe Riccardi, Adam J Sporka, and Jakub Franc. 2011. Recognition of personality traits from human spoken conversations. In *Twelfth Annual Conference of the International Speech Communication Association*.

Shaoxiong Ji, Shirui Pan, Guodong Long, Xue Li, Jing Jiang, and Zi Huang. 2019. Learning private neural language modeling with attentive aggregation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Hans Jonas. 1983. Das prinzip verantwortung. versuch einer ethik für die technologische zivilisation. *Zeitschrift für Philosophische Forschung*, 37(1):144–147.

Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, 3(1–2):1–224.

David A Kenny and Linda Albright. 1987. Accuracy in interpersonal perception: a social relations analysis. *Psychological bulletin*, 102(3):390.

Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 121–130. ACM.

Dan Klein and Christopher D Manning. 2003. Fast exact inference with a factored model for natural language parsing. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 3–10. MIT Press.

Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. 2014. Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1857–1866. ACM.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterising user impact on Twitter. In *14th conference of the European chapter of the Association for Computational Linguistics 2014, EACL 2014*, pages 405–413.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H Andrew Schwartz. 2017. Human centered nlp with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Christopher D Manning, Christopher D Manning, and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

Dejan Markovikj, Sonja Gievska, Michal Kosinski, and David J Stillwell. 2013. Mining facebook data for predictive personality modeling. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Nicolaas Matthijs and Filip Radlinski. 2011. Personalizing web search using long term browsing history. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 25–34. ACM.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2017. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Bradley N. Miller, John T. Riedl, and Joseph A. Konstan. 1997. Experiences with grouplens: Marking usenet useful again. In *Proceedings of the Annual Conference on USENIX Annual Technical Conference*, ATEC '97, pages 17–17, Berkeley, CA, USA. USENIX Association.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019. SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 147–156, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Morton, Joern Kottmann, Jason Baldridge, and Gann Bierner. 2005. Opennlp: A java-based nlp toolkit. In *Proc. EACL*.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, pages 207–217. Association for Computational Linguistics.

Dong Nguyen, Noah A Smith, and Carolyn P Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123. Association for Computational Linguistics.

Dong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Shereen Oraby, Lena Reed, Shubhangi Tandon, TS Sharath, Stephanie Lukin, and Marilyn Walker. 2018. Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 180–190.

Jiaqi Pan, Rishabh Bhardwaj, Wei Lu, Hai Leong Chieu, Xinghao Pan, and Ni Yi Puay. 2019. Twitter homophily: Network based prediction of user's occupation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2633–2638, Florence, Italy. Association for Computational Linguistics.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Who's (really) the boss? perception of situational power in written interactions. In *Proceedings of COLING 2012*, pages 2259–2274, Mumbai, India. The COLING 2012 Organizing Committee.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764, Beijing, China. Association for Computational Linguistics.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123, New Orleans, Louisiana. Association for Computational Linguistics.

Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 2018. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 459–467, New York, NY, USA. ACM.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186. ACM.

Phillip Rogaway. 2015. The moral character of cryptographic work. *IACR Cryptology ePrint Archive*, 2015:1162.

Laurie A Rudman and Peter Glick. 2012. *The social psychology of gender: How power and intimacy shape gender relations*. Guilford Press.

Mark Sanderson and W Bruce Croft. 2012. The history of information retrieval research. *Proceedings of the IEEE*, 100(Special Centennial Issue):1444–1451.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar. Association for Computational Linguistics.

Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki, and Kentaro Inui. 2018. Predicting stances from social media posts using factorization machines. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3381–3390, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Katie Shilton and Sheridan Sayles. 2016. " we aren't all going to be on the same page about ethics": Ethical practices and challenges in research on digital and social media. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1909–1918. IEEE.

Charese Smiley, Frank Schilder, Vassilis Plachouras, and Jochen L. Leidner. 2017. Say the right thing right: Ethics issues in natural language generation systems. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 103–108, Valencia, Spain. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510.

Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J Park. 2012. Predicting dark triad personality traits from Twitter usage and a linguistic analysis of tweets. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 386–393. IEEE.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59.

RS Taylor. 1968. Question-negotiation and information-seeking in libraries (vol. 29): College and research libraries.

Jaime Teevan, Susan T Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456. ACM.

Ilaria Torre. 2009. Adaptive systems in the era of the semantic and social web, a survey. *User Modeling and User-Adapted Interaction*, 19(5):433–486.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.

Leanne Townsend and Claire Wallace. 2016. Social media research: A guide to ethics. *Aberdeen: University of Aberdeen*.

J Verhoeff, William Goffman, and Jack Belzer. 1961. Inefficiency of the use of boolean functions for information retrieval systems. *Communications of the ACM*, 4(12):557–558.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1567–1578, Berlin, Germany. Association for Computational Linguistics.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 186–196, Baltimore, Maryland. Association for Computational Linguistics.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 505–510, Sofia, Bulgaria. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Tsung-Hsien Wen, Hung-Yi Lee, Tai-Yuan Chen, and Lin-Shan Lee. 2012. Personalized language modeling by crowd sourcing with social network data for voice access of cloud applications. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 188–193. IEEE.

Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168.

Tom D Wilson. 1999. Models in information behaviour research. *Journal of documentation*, 55(3):249–270.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

Seunghyun Yoon, Hyeongu Yun, Yuna Kim, Gyu-tae Park, and Kyomin Jung. 2017. Efficient transfer learning schemes for personalized language modeling using recurrent neural network. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Mohammadzaman Zamani and H. Andrew Schwartz. 2017. Using twitter language to predict the real estate market. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 28–33, Valencia, Spain. Association for Computational Linguistics.

Mohammadzaman Zamani, H. Andrew Schwartz, Veronica Lynn, Salvatore Giorgi, and Niranjan Balasubramanian. 2018. Residualized factor adaptation for community social media prediction tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3560–3569, Brussels, Belgium. Association for Computational Linguistics.

Xingshan Zeng, Jing Li, Lu Wang, and Kam-Fai Wong. 2019. Joint effects of context and user history for predicting online conversation re-entries. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2809–2818, Florence, Italy. Association for Computational Linguistics.