Legend:
! key point
≠ potentially disagree
* more sources/info

The Social Impact of Natural Language Processing

ie "ethics guidelines are needed in NLP research!"

Dirk Hovy
Center for Language Technology
University of Copenhagen
Copenhagen, Denmark
dirk.hovy@hum.ku.dk

Shannon L. Spruit
Ethics & Philosophy of Technology
Delft University of Technology
Delft, The Netherlands
s.l.spruit@tudelft.nl

Abstract

Medical sciences have long since established an ethics code for experiments, to minimize the risk of harm to subjects. Natural language processing (NLP) used to involve mostly anonymous corpora, with the goal of enriching linguistic analysis, and was therefore unlikely to raise ethical concerns. As NLP becomes increasingly wide-spread and uses more data from social media, however, the situation has changed: the outcome of NLP experiments and applications *can* now have a direct effect on individual users' lives. Until now, the discourse on this topic in the field has not followed the technological development, while public discourse was often focused on exaggerated dangers. This position paper tries to take back the initiative and start a discussion. We identify a number of social implications of NLP and discuss their ethical significance, as well as ways to address them.

* lots of this ethics research now comes out of the Oxford Internet Institute.

1 Introduction

After the Nuremberg trials revealed the atrocities conducted in medical research by the Nazis, medical sciences established a set of rules to determine whether an experiment is ethical. This involved incorporating the principles of biomedical ethics as a *lingua franca* of medical ethics (Beauchamp and Childress, 2001).

These guidelines were designed to balance the potential value of conducting an experiment while preventing the exploitation of human subjects. Today, any responsible research institution uses these—or comparable—criteria to approve or reject experiments before any research can be conducted. The administrative body governing these decisions is the Institutional Review Board (IRB).

IRBs mostly pertain to experiments that directly involve human subjects, though, and so NLP and other data sciences have not employed such guidelines. Work on existing corpora is unlikely to raise any flags that would require an IRB approval.[1]

Data sciences have therefore traditionally been less engaged in ethical debates of their subject, even though this seems to be shifting, see for instance Wallach (2014), Galaz et al. (2015), or O'Neil (2016). The public outcry over the "emotional contagion" experiment on Facebook (Kramer et al., 2014) further suggests that data sciences now affect human subjects in real time, and that we might have to reconsider the application of ethical considerations to our research (Puschmann and Bozdag, 2014). NLP research not only involves similar data sets, but also works with their content, so it is time to start a discussion of the ethical issues specific to our field.

* eg we can gather intell. on how to "push peoples' buttons" by trawling social media to see what clickbait works

Much of the ethical discussion in data sciences to date, however, has centered around privacy concerns (Tse et al., 2015). We do not deny the reality and importance of those concerns, but they involve aspects of digital rights management/access control, policy making, and security, which are not specific to NLP, but need to be addressed in the data sciences community as a whole. Steps towards this have been taken by Russell et al. (2015).

Instead, we want to move beyond privacy in our ethical analysis and look at the wider **social impact** NLP may have. In particular, we want to explore the impact of NLP on social justice, i.e., equal opportunities for individuals and groups (such as minorities) within society to access resources, get their voice heard, and be represented in society.

! !

[1] With few exceptions, such as dialogue research (Joel Tetreault, pers. comm.)

* Holy smokes, this is very recent its surprising that more ethics talk hasn't occured!

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 591–598,
Berlin, Germany, August 7-12, 2016.

Our contributions We believe ethical discussions are more constructive if led by practitioners, since the public discussion of ethical aspects of IT and data sciences is often loaded with fear of the unknown and unrealistic expectations. For example, in the public discourse about AI (Hsu, 2012; Eadicicco, 2015; Khatchadourian, 2015), people either dismiss the entire approach, or exaggerate the potential dangers (see Etzioni (2014) for a practioner's view point). This paper is an attempt to take back the initiative for NLP.

At the same time, we believe that the field of ethics can contribute a more general framework, and so this paper is an interdisciplinary collaboration between NLP and ethics researchers.

To facilitate the discussion, we also provide some of the relevant **terminology** from the literature on ethics of technology, namely the concepts of *exclusion, overgeneralization, bias confirmation, topic under- and overexposure*, and *dual use*.

2 Does NLP need an ethics discussion?

As discussed above, the makeup of most NLP experiments so far has not obviated a need for ethical considerations, and so, while we are aware of individual discussions (Strube, 2015), there is little discourse in the community yet. A search for "*ethic**" in the ACL anthology only yields three results. One of the papers (McEnery, 2002) turns out to be a panel discussion, another is a book review, leaving only Couillault et al. (2014), who devote most of the discussion to legal and quality issues of data sets. We know social implications have been addressed in some NLP curricula,[2] but until now, no discipline-wide discussion seems to take place.

The most likely reason is that NLP research has not directly involved human subjects.[3] Historically, most NLP applications focused on further enriching existing text which was not strongly linked to any particular author (newswire), was usually published publicly, and often with some temporal distance (novels). All these factors created a distance between text and author, which prevented the research from directly affecting the authors' situation.

This situation has changed lately due to the increased use of social media data, where authors are current individuals, who can be directly affected by the results of NLP applications. Couillault et al. (2014) touch upon these issues under "traceability" (i.e., whether individuals can be identified): this is undesirable for experimental subjects, but might be useful in the case of annotators.

Most importantly, though: the subject of NLP—language—is a *proxy for human behavior*, and a strong signal of individual characteristics. People use this signal consciously, to portray themselves in a certain way, but can also be identified as members of specific groups by their use of subconscious traits (Silverstein, 2003; Agha, 2005; Johannsen et al., 2015; Hovy and Johannsen, 2016).

Language is always *situated* (Bamman et al., 2014), i.e., it is uttered in a specific situation at a particular place and time, and by an individual speaker with all the characteristics outlined above. All of these factors can therefore leave an imprint on the utterance, i.e., the texts we use in NLP carry *latent* information about the author and situation, albeit to varying degrees.

This information can be used to predict author characteristics from text (Rosenthal and McKeown, 2011; Nguyen et al., 2011; Alowibdi et al., 2013; Ciot et al., 2013; Liu and Ruths, 2013; Volkova et al., 2014; Volkova et al., 2015; Plank and Hovy, 2015; Preotiuc-Pietro et al. 2015a; Preoţiuc-Pietro et al., 2015b), and the characteristics in turn can be detected by and influence the performance of our models (Mandel et al., 2012; Volkova et al., 2013; Hovy, 2015).

As more and more language-based technologies are becoming available, the ethical implications of NLP research become more important. What research is carried out, and its quality, directly affect the functionality and impact of those technologies.

The following is meant to start a discussion addressing ethical issues that can emerge in (and from) NLP research.

3 The social impact of NLP research

We have outlined the relation between language and individual traits above. Language is also a political instrument, though, and an instrument of power. This influence stretches into politics and everyday competition, for example for turn-taking (Laskowski, 2010; Bracewell and Tomlinson, 2012; Prabhakaran and Rambow, 2013; Prab-

[2] Héctor Martínez Alonso, personal communication

[3] Except for annotation: there are a number of papers on the status of crowdsource workers (Fort et al., 2011; Pavlick et al., 2014).Couillault et al. (2014) also briefly discuss annotators, but mainly in the context of quality control.

hakaran et al., 2014; Tsur et al., 2015; Khouzami et al., 2015, inter alia), .

The mutual relationships between language, society, and the individual are also the source for the societal impact factors of NLP: failing to recognize group membership (Section 3.1), implying the wrong group membership (see Section 3.2), and overexposure (Section 3.3). In the following, we discuss sources of these problems in the data, modeling, and research design, and suggest possible solutions to address them.

3.1 Exclusion

As a result of the situatedness of language, any data set carries a **demographic bias**, i.e., latent information about the demographics in it. Overfitting to these factors can have have severe effects on the applicability of findings. In psychology, where most studies are based on western, educated, industrialized, rich, and democratic research participants (so-called WEIRD, Henrich et al. (2010)), the tacit assumption that human nature is so universal that findings on this group would translate to other demographics has led to a heavily biased corpus of psychological data. In NLP, overfitting to the demographic bias in the training data is due to the *i.i.d.* assumption. I.e., models implicitly assume all language to be identical to the training sample. They therefore perform worse or even fail on data from other demographics.

Potential consequences are **exclusion** or demographic misrepresentation. This in itself already represents an ethical problem for research purposes, threatening the universality and objectivity of scientific knowledge (Merton, 1973). These problems exacerbate, though, once they are applied to products. For instance, standard language technology may be easier to use for white males from California (as these are taken into account while developing it) rather than women or citizens of Latino or Arabic descent. This will reinforce already existing demographic differences, and makes technology less user friendly for such groups, cf. authors like Bourdieu and Passeron (1990) have shown how restricted language, like class specific language or scientific jargon, can hinder the expression of outsiders' voices from certain practices. A lack of awareness or decreased attention for demographic differences in research stages can therefore lead to issues of exclusion of people along the way.

Concretely, the consequences of exclusion for NLP research have recently been pointed out by Hovy and Søgaard (2015) and Jørgensen et al. (2015): current state-of-the-art NLP models score a significantly lower accuracy for young people and ethnic minorities vis-à-vis the modeled demographics.

Better awareness of these mechanism in NLP research and development can help prevent problems further on. Potential counter-measures to demographic bias can be as simple as downsampling the over-represented group in the training data to even out the distribution. The work by Mohammady and Culotta (2014) shows another approach, by using existing demographic statistics as supervision. In general, measures to address overfitting or imbalanced data can be used to correct for demographic bias in data.

3.2 Overgeneralization

Exclusion is a side-effect of the data. **Overgeneralization** is a modeling side-effect.

As an example, we consider automatic inference of user attributes, a common and interesting NLP task, whose solution also holds promise for many useful applications, such as recommendation engines and fraud or deception detection (Badaskar et al., 2008; Fornaciari and Poesio, 2014; Ott et al., 2011; Banerjee et al., 2014).

The cost of false positives seems low: we might be puzzled or amused when receiving an email addressing us with the wrong gender, or congratulating us to our retirement on our 30th birthday.

In practice, though, relying on models that produce false positives may lead to bias confirmation and overgeneralization. Would we accept the same error rates if the system was used to predict sexual orientation or religious views, rather than age or gender? Given the right training data, this is just a matter of changing the target variable.

To address overgeneralization, the guiding question should be "would a false answer be worse than no answer?" We can use dummy variables, rather than take a *tertium non datur* approach to classification, and employ measures such as error weighting and model regularization, as well as confidence thresholds.

3.3 The problem of exposure

Topic overexposure creates biases Both exclusion and overgeneralization can be addressed algo-

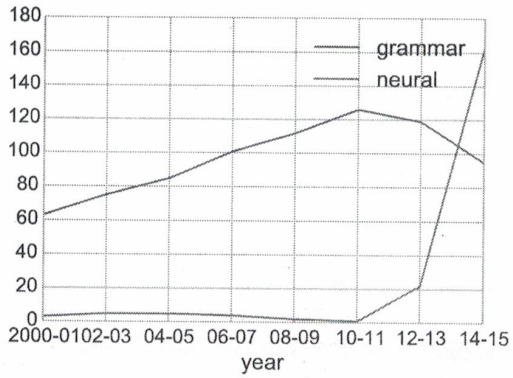

Figure 1: ACL title keywords over time

rithmically, while **topic overexposure** originates from research design.

In research, we can observe this effect in waves of research topics that receive increased mainstream attention, often to fall out of fashion or become more specialized, cf. ACL papers with "grammars" vs. "neural" in the title (Figure 1).

Such topic overexposure may lead to a psychological effect called **availability heuristic** (Tversky and Kahneman, 1973): if people can recall a certain event, or have knowledge about specific things, they infer it must be more important. For instance, people estimate the size of cities they recognize to be larger than that of unknown cities (Goldstein and Gigerenzer, 2002).

However, the same holds for individuals/groups/characteristics we research. The heuristics become ethically charged when characteristics such as violence or negative emotions are more strongly associated with certain groups or ethnicities (Slovic et al., 2007). If research repeatedly found that the language of a certain demographic group was harder to process, it could create a situation where this group was perceived to be difficult, or abnormal, especially in the presence of existing biases. The confirmation of biases through the gendered use of language, for example, has also been at the core of second and third wave feminism (Mills, 2012).

Overexposure thus creates biases which can lead to discrimination. To some extent, the frantic public discussion on the dangers of AI can be seen as a result of overexposure (Sunstein, 2004).

There are no easy solutions to this problem, which might only become apparent in hindsight. It can help to assess whether the research direction of a project feeds into existing biases, or whether it overexposes certain groups.

Underexposure can negatively impact evaluation. Similar to the WEIRD-situation in psychology, NLP tends to focus on Indo-European data/text sources, rather than small languages from other language groups, for example in Asia or Africa. This focus creates an imbalance in the available amounts of labeled data. Most of the exisitng labeled data covers only a small set of languages. When analyzing a random sample of Twitter data from 2013, we found that there were no treebanks for 11 of the 31 most frequent languages, and even fewer semantically annotated resources (the ACE corpus covers only English, Arabic, Chinese, and Spanish).[4]

Even if there is a potential wealth of data available from other languages, most NLP tools are geared towards English (Schnoebelen, 2013; Munro, 2013). The prevalence of resources for English has created an **underexposure** to typological variety: both morphology and syntax of English are global outliers. Would we have focused on n-gram models to the same extent if English was as morhpologically complex as, say, Finnish?

While there are many approaches to develop multi-lingual and cross-lingual NLP tools for linguistic outliers (Yarowsky and Ngai, 2001; Das and Petrov, 2011; Søgaard, 2011; Søgaard et al., 2015; Agić et al., 2015), there simply are more commercial incentives to overexpose English, rather than other languages. Even if other languages are equally (or more) interesting from a linguistic and cultural point of view, English is one of the most widely spoken language and therefore opens up the biggest market for NLP tools. This focus on English may be self-reinforcing: the existence of off-the-shelf tools for English makes it easy to try new ideas, while to start exploring other languages requires a higher startup cost in terms of basic models, so researchers are less likely to work on them.

4 Dual-use problems

Even if we address all of the above concerns and do not intend any harm in our experiments, they can still have unintended consequences that negatively affect people's lives (Jonas, 1984).

Advanced analysis techniques can vastly improve search and educational applications

[4]Thanks to Barbara Plank for the analysis!

(Tetreault et al., 2015), but can re-enforce prescriptive linguistic norms when degrading on non-standard language. Stylometric analysis can shed light on the provenance of historic texts (Mosteller and Wallace, 1963), but also endanger the anonymity of political dissenters. Text classification approaches help decode slang and hidden messages (Huang et al., 2013), but have the potential to be used for censorship. At the same time, NLP can also help uncovering such restrictions (Bamman et al., 2012). As recently shown by Hovy (2016), NLP techniques can be used to detect fake reviews, but also to generate them in the first place.

All these examples indicate that we should become more aware of the way other people appropriate NLP technology for their own purposes. The unprecedented scale and availability can make the consequences of NLP technologies hard to gauge.

The unintended consequences of research are also linked to the incentives associated with funding sources. The topic of government and military involvement in the field deserves special attention in this respect. On the one hand, Anderson et al. (2012) show how a series of DARPA-funded workshops have been formative for ACL as a field in the 1990s. On the other hand, there are scholars who refuse military-related funding for moral reasons.[5]

While this decision is up to the individual researcher, the examples show that moral considerations go beyond the immediate research projects. We may not directly be held responsible for the unintended consequences of our research, but we can acknowledge the ways in which NLP can enable morally questionable/sensitive practices, raise awareness, and lead the discourse on it in an informed manner. The role of the researcher in such ethical discussions has recently been pointed out by Rogaway (2015).

5 Conclusion

In this position paper, we outlined the potential social impact of NLP, and discussed ways for the practitioner to address this. We also introduced *exclusion*, *overgeneralization*, *bias confirmation*, *topic overexposure*, and *dual use*. Countermeasures for exclusion include bias control techniques like downsampling or priors; for overgeneralization: dummy labels, error weighting, or confidence thresholds. Exposure problems can only be addressed by careful research design, and dual-use problems seem hardly addressable on the level of the individual researcher, but require the concerted effort of our community.

We hope this paper can point out ethical considerations for collecting our data, designing the experimental setup, and assessing the potential application of our systems, and help start an open discussion in the field about the limitations and problems of our methodology.

[5] For a perspective in a related field see https://web.eecs.umich.edu/~kuipers/opinions/no-military-funding.html

Acknowledgements

The authors would like to thank Joel Tetrault, Rachel Tatman, Joel C. Wallenberg, the members of the COASTAL group, and the anonymous reviewers for their detailed and invaluable feedback. The first author was funded under the ERC Starting Grant LOWLANDS No. 313695. The second author was funded by the Netherlands Organization for Scientific Research under grant number 016.114.625.

References

Asif Agha. 2005. Voice, footing, enregisterment. *Journal of linguistic anthropology*, pages 38–59.

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd annual meeting of the ACL*.

Jalal S Alowibdi, Ugo A Buy, and Philip Yu. 2013. Empirical evaluation of profile characteristics for gender classification on twitter. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 365–369. IEEE.

Ashton Anderson, Dan McFarland, and Dan Jurafsky. 2012. Towards a computational history of the ACL: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21. Association for Computational Linguistics.

Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. 2008. Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

David Bamman, Brendan O'Connor, and Noah Smith. 2012. Censorship and deletion practices in Chinese social media. *First Monday*, 17(3).