Part I

Prediction
Using patterns in the data

prediction

In the previous part, we looked at ways of exploring textual data, to find out what kind of structures are in there. We will build on some of that knowledge in this part, but we will use it to infer information for new, unseen data using **prediction**.

Prediction is useful for inferring both linguistic structure (parts of speech, syntax, NER, discourse structure) as well as a multitude of social constructs signaled in language. Sentiment is probably the most well-known one, but there is virtually no limit in the kind of things we can predict from text: power, trust, misogyny, age, gender, etc.

classes

Prediction is a core part of machine learning, and essentially involves showing the computer lots of examples of inputs (i.e., in our case documents) and the correct output label for them (for example, *spam* or *no spam*, or *positive*, *negative*, *neutral*). Note that we can have any number of possible output labels (**classes**).

regression

If instead of classes, we predict a continuous numerical value, we refer to it as **regression**. However, regression is practically never used in NLP. In social sciences, there are some cases where we would like to do regression: the principles described below are largely analogous for that case, but we will not cover it in detail.

Predictive modeling is a very useful application for social sciences, which allows us to infer variables of interest (power relations, buying intent, etc.) and complete missing variables (e.g., age, gender) based on language. The principles behind these models are fairly simple, and with sufficient data and more and more powerful models, it is easy to get carried away in the possibilities.

We will look at the basic principles behind classification, how to set up experiments in a realistic way (see page **??** and **??**) to evaluate the performance (see page **??**), test the significance of our performance results (see page **??**), and prevent overfitting (see page **??**). We will then look at how to improve performance of the predictive models with a variety of techniques (see page **??**).

In the last few chapters of the books, we will take a look at neural networks. This is a booming research area (even though the basic idea is already from the 1950s), so we will only cover the basics (see page **??**), as well as a number of model architectures that have shown themselves especially useful for text analysis (see pages **??**, **??**, **??**) and some additional architectural features that have proven useful (see page **??**).

However, since language is one of the most individual human capabilities, it also reveals a lot about the person using it. The predictive models we build have the possibility to expose people's profiles and characteristics without them being aware of it. With great power, there comes great responsibility, so before we delve deeper into the algorithmic side of prediction, let's look at the ethical aspects of it.

CHAPTER I.1

Ethics, Fairness, and Bias

You have a data set with survey answers, and for some of the participants, you have information about their age and gender. However, you ran several trials, and not in all of them were people required to provide this demographic information, so some people did not supply it. It then turns out that controlling for gender would be a very useful thing to do for your study. You decide to train a classifier on the subset of the data with gender information available, and then apply it to the rest of the participants in order to impute the missing demographic information. Your classifier performs well, and you complete the study, now able to control for the effect of gender.

As you look at the data again, you realize that it has a bias: most of the participants who provided gender information were women, and it turns out that the classifier is much better at identifying women. You now suspect that the gender estimate you have for the entire data set is even more skewed towards predicting female participants. You realize that you have essentially created a tool that can infer gender where people have not provided it. Maybe they just forgot, but maybe they did not want to reveal this information. Is it ok to still use your tool? Obviously, you are only interested in answering a scientific question, but you realize that your tool could also be used by people with much less honorable motives.

With great (predictive) power comes great responsibility, and especially when working with language, there are a number of ethical questions that arise. There are no hard and fast rules for everything, and the topic is still evolving, but a number of topics have emerged so far. While the notes here are necessarily incomplete, they are meant as a starting point on the issue.[1]

Originally, machine learning and NLP were about solving fairly artificial problems on small data sets, with the promise of doing it on larger data at some later point. And while there has always been a certain amount of skepticism and worry about AI's power, these worries were largely theoretical: there was not enough data and computational power to actually make an impact on people's life.

With the recent availability of large amounts of data, and the ubiquitous application of ML, this point has finally arrived. With the focus on making useful tools, empiricism has moved the needle away from explanatory and descriptive models (which we could

[1] This chapter is based on the work by Hovy and Spruit (2016), as well as the ACL workshops on Ethics in NLP.

analyze to understand *why* it returned a certain result), towards predictive models that are hard to analyze, but produce excellent predictions.

It now turns out that one of the reasons why models have gotten so good at prediction is because they are sometimes picking up on things that they were not meant to exploit: embedding models reflect ethnic and gender stereotypes (Bolukbasi et al., 2016), bail decision predictions are majorly influenced by the defendants' ethnicity (Angwin et al., 2016), automatic captioning and smart speakers do not work for people with non-standard varieties (Tatman, 2017; Harwell, 2018), and skin cancer detectors work only on white skin (Adamson and Smith, 2018).

bias

All of these unintended consequences are examples of **bias**: a systematic difference from the truth. These biases can arise from the data, the models, or the research design itself.

With language data, demographic biases are very strong, since we use language to consciously and subconsciously signal who we are. And so language reflects a lot of information about our age, gender, ethnicity, region, personality, and even things like our profession and income bracket.

privacy

This leads to a second problem with better and better predictive models, namely **privacy**: at this point, we can use NLP systems that exploit the signals in our language use to predict all of the above features: people's age (Rosenthal and McKeown, 2011; Nguyen et al., 2011), gender (Alowibdi et al., 2013; Ciot et al., 2013; Liu and Ruths, 2013), personality (Park et al., 2015), job title (Preoţiuc-Pietro et al., 2015a), income (Preoţiuc-Pietro et al., 2015b), and much more (Volkova et al., 2014, 2015).

dual use

Being able to predict arbitrary demographic and socio-cultural attributes puts practitioners in a moral quandary: on the one hand, we want the best tools to answer our questions and make generalizable claims about the world. On the other hand, we do not want to develop tools that could be misappropriated for nefarious goals. This is the third problem, called **dual use**.

1. Sources of Bias

selection bias
annotation bias

1.1. Data. There are two ways in which data can introduce bias into our work: through the selection of a demographically not-representative data set (**selection bias**) or through annotation decisions made by the coders of our data (**annotation bias**).

demographic bias

When choosing a text data set to work with, we are also making decisions about the demographic groups represented in the data. As a result of the demographic signal present in language, any data set carries a **demographic bias**, i.e., latent information about the demographic groups present in it. As humans, we would not be surprised if someone who grew up hearing only their dialect would have trouble understanding other people. So if our data set is dominated by the "dialect" of a certain demographic group, we should not be surprised that our models have problems understanding others.

Bias is not necessarily a problem a priori: most data sets have some kind of built-in bias, and in many cases, it is benign. It becomes problematic when these bias negatively affect certain groups, or advantage others. On a biased data sets, statistical models will overfit to the presence of certain linguistic signals that are particular to the dominant group. As a result, the model will work less well for other groups, i.e., it leads to **exclusion** of demographic groups.

exclusion

Concretely, the consequences of exclusion for NLP research have recently been shown by Hovy and Søgaard (2015) and Jørgensen et al. (2015): POS models have a significantly lower accuracy for young people and ethnic minorities vis-à-vis the dominant demographics in the training data. Apart from exclusion, these models will pose a problem for future research: given that a large part of the world's population is currently under 30,[2] such models are bound to degrade even more over time, and ultimately not meet the needs of its users.

This also has severe ramifications for the general applicability of any findings using these tools. In psychology, most studies are based on college students, and therefore hold only for a particular demographic: western, educated, industrialized, rich, and democratic research participants (so-called WEIRD, Henrich et al. (2010)). Assuming that findings from this group would translate to other demographics has proven wrong and led to a heavily biased corpus of psychological data and research.

Counter measures. Potential counter-measures to demographic selection bias can be as simple as post-stratification: downsampling the over-represented group in the training data to even out the distribution until is reflects the true distribution. Mohammady and Culotta (2014) have shown existing demographic statistics can be used as supervision. In general, measures to address overfitting or imbalanced data can be used to correct for demographic bias in data. In general, avoiding biased selections is even better, so when creating new data sets, NLP researchers have been encouraged to provide a **data statement** (Bender and Friedman, 2018). This includes various aspects of the data collection process and the underlying demographics. As a useful side effect, it forces us to consider how our data is made up, and provides future researchers a way to assess the effect of any bias they might notice when using the data.

data statement

In order to prevent the effect of annotation bias, we can use **annotation models** (Hovy et al., 2013; Paun et al., 2018) that help us find biased annotators, and we can account for the human disagreement between labels in the update process of our models (Plank et al., 2014).

annotation models

1.2. Models. Another source of biased predictions can be the tendency of statistical models for **overamplification**, i.e., the tendency of a model to rely on small differences between subjects to satisfy the objective function and make good predictions. Unchecked, the model can maximize its goal by amplifying the difference when predicting new data, and creating an imbalance that is much larger than in the original data. Yatskar et al. (2016) have shown that in an image captioning data set where 58% of the

overamplification

[2] http://www.socialnomics.net/2010/04/13/over-50-of-the-worlds-population-is-under-30-social-media-on-the-rise/

captions for pictures of a person in a kitchen mentioned women. A small, but noticeable difference, which could be corrected by sampling. However, as Zhao et al. (2017) showed, when a standard statistical model was trained on this slightly biased data, it ended up predicting the gender of the person in a kitchen picture to be a woman in 63% of the cases.

The cost of these false positive predictions seems low: a user might be puzzled or amused when seeing a mislabeled image, or when receiving an email addressing them with the wrong gender. However, relying on models that produce false positives may lead to **bias confirmation** and **overgeneralization**. And while we might be amused by some errors, would we accept the same false positives in a system that was used to predict sexual orientation or religious views, rather than age or gender? For any prediction task, this is just a matter of changing the target variable and finding some data.

bias confirmation
overgeneralization

Another problem of overamplification is the proliferation of stereotypes: Rudinger et al. (2018) found that coreference resolution systems (which link a pronoun to the noun they refer to) were biased by gender. In the sentence "The surgeon could not operate on her patient: it was her son", the model does not link "surgeon" and "her". The cause of this is presumably biased training data, but the effect feeds into gender stereotypes. Similarly, Kiritchenko and Mohammad (2018) showed that sentiment analysis models changed their scores for the same sentences when only replacing a female with a male pronoun ("She/He made me feel afraid"), or when changing a typically "white" name with a typically "black" name ("I made Heather/Latisha feel angry"), in both cases giving higher scores for the second case.

Counter measures. To address the overgeneralization of models, we can ask ourselves "would a false answer be worse than no answer?" Instead of taking a *tertium non datur* approach to classification, where a model has to (and will) produce *some* answer, we can use dummy variables that say "unknown". We can also use measures such as error weighting, incurring a higher penalty if the model make mistakes on the smaller calls, as well as confidence thresholds below which we do not assign a label.

adversarial learning

Recently, (Li et al., 2018) have shown that **adversarial learning** (a special architecture in neural networks) can help not only reduce the effect of predictive biases, but can actually also help to improve performance of the models.

1.3. Research Design. As we have seen in previous chapters (and as this book has further advanced), most NLP research is done on English, and generally tends to focus on Indo-European data/text sources, rather than small languages from other language groups, for example in Asia or Africa. Even if there is a potential wealth of data available from other languages, most NLP tools are geared towards English (Schnoebelen, 2013; Munro, 2013).

underexposure

This **underexposure** for other languages creates an imbalance in the available amounts of labeled data, and proliferates itself: because most of the existing labeled data covers only a small set of languages, most of the research is focused on those languages, and so more resources are created for those languages. This dynamic makes it more difficult for new research on smaller languages, and it naturally directs new researchers towards the existing ones. The focus on English may therefore be self-reinforcing: the existence of off-the-shelf tools for English makes it easy to try new ideas, while it requires a much higher cost to start exploring other languages, in terms of data annotation, basic models, and other resources.

There is little in the way of semantic or syntactic resources for many languages. In a random sample of Tweets from 2013,[3] we found 31 different languages: there were no treebanks for 11 of them, and even fewer semantically annotated resources like WordNets. Consequently, researchers are less likely to work on them.

Conversely, the prevalence of resources for English has created an **overexposure** to this variety, even though both morphology and syntax of English are global outliers. The overexposure to English (as well as to certain research areas or methods) creates a bias described by the **availability heuristic** (Tversky and Kahneman, 1973): if people can recall a certain thing or event more easily, they infer that this thing or event must be more important, bigger, better, more dangerous, etc. For instance, people estimate the size of cities they recognize to be larger than that of unknown cities (Goldstein and Gigerenzer, 2002). The same holds true for languages, methods, and topics we research. Would we have focused on n-gram models to the same extent if English was as morphologically complex as, say, Finnish?

overexposure

availability heuristic

While there are lately many approaches to develop multi-lingual and cross-lingual NLP tools for linguistic outliers, there simply are more commercial incentives to overexpose English, rather than other languages. Even if other languages are equally 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.

Overexposure can also create or feed into existing biases: 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 been cited as being 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 ask ourselves counterfactuals: Would I research this is if the data wasn't as easily available? Would my finding still hold on another language? We can also try to assess whether the research direction of a project feeds into existing biases, or whether it overexposes certain groups.

[3] Thanks to Barbara Plank for the analysis!

2. Privacy

General Data Protection Regulation

In the wake of the Cambridge Analytica scandal, it has become apparent that our personal data is not as secret and private as we would like to think. In the wake of this, the European Parliament has enacted a law that is designed to protect privacy online: the **General Data Protection Regulation** (GDPR).

GDPR makes a number of provisions for research purposes (as opposed to commercial purposes), but it does not give out a carte blanche to be negligent with subject data. Non-protected categories can still be predicted for research, and even protected categories are ok to use, as long as they can not be used to identify individual subjects. In other words, if it becomes necessary to estimate the overall prevalence of gender or sexual orientation in the data, we can use models to infer these in aggregate. It is however not ok to use them to profile individual subjects.

federated learning

Counter measures. Protecting privacy can be helped by keeping data sources separate from each other. In order to still be able to learn from all of these sources, we can use techniques called **federated learning** (Konečnỳ et al., 2016). Coavoux et al. (2018) have shown that neural network architecture choices can help to protect the privacy of users, but there is also cautious evidence that this might not be as bullet-proof as we might want (Elazar and Goldberg, 2018).

3. Normative vs. Descriptive Ethics

normative ethics
descriptive ethics

Biased models and data sets are a nuisance when used incorrectly. However, they can also be viewed as window into the nature of society. This property illustrates an interesting distinction between **normative ethics** and **descriptive ethics**: Because language is used to express opinions, and the word embeddings capture semantic similarity, they also capture how much writers associate two terms. This association is reflected in the fact that word embeddings show a high similarity between “woman” and “homemaker”, and between “man” and “programmer”. When biased word embeddings are used as input for predictive models, the inherent bias is clearly negative (Bolukbasi et al., 2016), and therefore normatively wrong for many applications (e.g., for reviewing job candidates, where ideally, we would want all genders or ethnicities equally associated with all jobs). However, a number of social science studies quickly picked up on the insight provided by the biases contained in word embeddings. Works by Garg et al. (2018) and Kozlowski et al. (2018) have shown that it is precisely this property of word embeddings that can be used to study evolving societal attitudes over time with respect to gender roles and ethnic stereotypes, by measuring the distance between certain sets of words in different decades. Similarly, Bhatia (2017) has shown that this property of word embeddings can be used to measure people’s psychological biases and attitudes towards making certain decisions. This is therefore descriptively correct.

In a similar vein, it is both normatively and descriptively wrong that Google translate used “he” when translating the gender-neutral Turkish pronoun “o” as “he” when referring to a doctor, but as “she” when referring to a nurse. And while it is normatively wrong for Google search to suggest for the query “why are american” the autocomplete option “so fat”, it is descriptively insightful as far as stereotypes are concerned.

4. Dual Use

The ethics philosopher Hans Jonas has cautioned that any technology that is possible will also be used, for both good and for bad (Jonas, 1984). Even if we address all biases and concerns and do not intend any harm in our experiments, they can still have unintended consequences that negatively affect people’s lives. The most well-known (and extreme) example is physics, which had to confront the fact that some of their most well-intentioned findings could be (and ultimately were) used to kill.

While in no way as extreme, text analysis techniques can have mixed outcomes as well. On the one hand, they can vastly improve search and educational applications (Tetreault et al., 2015), but they can alos re-enforce prescriptive linguistic norms when they work badly for non-standard language. Stylometric analysis can shed light on the provenance of historic texts (Mosteller and Wallace, 1963) and aid forensic analysis of extortion letters, but it can also endanger the anonymity of political dissenters. Text classification approaches can help decode slang and hidden messages (Huang et al., 2013), but have the potential to be used for censorship and suppression. At the same time, NLP can also help uncovering such restrictions (Bamman et al., 2012). Hovy (2016) shows that simple NLP techniques can be used to both 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. We also need to be aware that NLP research is and will be used for solely undesirable applications. Automated censorship or the measuring of party-line adherence online in order to punish dissenters are two examples. Statistical models make both of these uses possible. The unprecedented scale and availability can make the consequences of new technologies hard to gauge.

The examples show that moral considerations go beyond immediate research projects. As a practitioner, we need to be aware of this duality, and openly address it. We may not directly be held responsible for the unintended consequences of our research or products, but we should acknowledge the ways in which they can enable morally questionable or sensitive practices, raise awareness in our customers, colleagues, and students, and lead the discourse on it in an informed manner. The role of the researcher in such ethical discussions has been pointed out by Rogaway (2015), and the ethical obligations for the practicing data scientists in O’Neil (2016).

Bibliography

Adewole S Adamson and Avery Smith. 2018. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248.

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.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May*, 23.

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

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.

Sudeep Bhatia. 2017. Associative judgment and vector space semantics. *Psychological review*, 124(1):1.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender Inference of Twitter Users in Non-English Contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Maximin Coavoux, Shashi Narayan, and Shay B Cohen. 2018. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Daniel G Goldstein and Gerd Gigerenzer. 2002. Models of ecological rationality: the recognition heuristic. *Psychological review*, 109(1):75.

Drew Harwell. 2018. The accent gap. Why some accents don't work on Alexa or Google Home. *The Washington Post*.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Dirk Hovy. 2016. The Enemy in Your Own Camp: How Well Can We Detect Statistically-Generated Fake Reviews–An Adversarial Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 483–488.

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)*, volume 2, pages 591–598.

Hongzhao Huang, Zhen Wen, Dian Yu, Heng Ji, Yizhou Sun, Jiawei Han, and He Li. 2013. Resolving entity morphs in censored data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1083–1093.

Hans Jonas. 1984. *The Imperative of Responsibility: Foundations of an Ethics for the Technological Age*. (Original in German: Prinzip Verantwortung.) Chicago: University of Chicago Press.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2018. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 25–30.

Wendy Liu and Derek Ruths. 2013. What's in a name? using first names as features for gender inference in twitter. In *Analyzing Microtext: 2013 AAAI Spring Symposium*.

Sara Mills. 2012. *Gender matters: Feminist linguistic analysis*. Equinox Pub.

Ehsan Mohammady and Aron Culotta. 2014. Using county demographics to infer attributes of twitter users. In *Proceedings of the joint workshop on social dynamics and personal attributes in social media*, pages 7–16.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Robert Munro. 2013. NLP for all languages. Idibon Blog, May 22 `http://idibon.com/nlp-for-all` Retrieved May 17, 2016.

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.

Cathy O'Neil. 2016. The Ethical Data Scientist. Slate, February 4 `http://www.slate.com/articles/technology/future_tense/2016/02/how_to_bring_better_ethics_to_data_science.html` Retrieved Feb 24, 2016.

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.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

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.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. An analysis of the user occupational class through twitter content. In *ACL*.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717.

Phillip Rogaway. 2015. The moral character of cryptographic work. Technical report, IACR-Cryptology ePrint Archive.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 8–14.

Tyler Schnoebelen. 2013. The weirdest languages. Idibon Blog, June 21 `http://idibon.com/the-weirdest-languages` Retrieved May 17, 2016.

Cass R Sunstein. 2004. Precautions against what? the availability heuristic and cross-cultural risk perceptions. *U Chicago Law & Economics, Olin Working Paper*, (220):04–22.

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.

Joel Tetreault, Jill Burstein, and Claudia Leacock. 2015. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, chapter Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media (demo). In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, Austin, TX.

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 ACL*, pages 186–196.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.