Do you know who I am? A neural network model for the discovery of status norms in socio-textual data

Final Project

Bogdan State
Stanford University
bstate@stanford.edu

ABSTRACT

Social norms are complex rules which govern human interaction. Despite being essential elements of social theory, the quantitative identification of social norms has proven to be an elusive task. Increasing availability of large social datasets, coupled with advances in the field of artificial neural networks and in scientific computation are revolutionizing fields as diverse as computer vision, machine translation, or robotics. Because of their ability to capture complex relationships between many variables, artificial neural networks appear particularly well-suited for the problem of social norm extraction. This paper provides an example of how graph and text embeddings could be used to investigate the complex rules that dictate how people behave.

ACM Reference Format:
Bogdan State. 2018. Do you know who I am? A neural network model for the discovery of status norms in socio-textual data: Final Project. In *Proceedings of Stanford CS230 (CS230).* ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.475/123_4

1 PRIOR WORK

A norm is a socially-standardized rule that dictates how an agent in a particular setting should respond to a set of social inputs [see 2, 26].[1] The study of social norms is complicated immensely by their distributed nature however. Every individual has their own understanding of how they are supposed to act in a particular social setting. In the theoretical sense of the concept, we can think of a social norm as a computer program [e.g. 11], although each individual carries their own slightly different version of that program. Further complicating things is the fact that individuals oftentimes have multiple programs available from which to choose [e.g. 4]. For instance, the same person may alternatively respond to a beggar's entreaties following the norm of pious charity, or that of firm civic spirit, resulting in widely different outcomes for the beggar. To make things even more complex, norms aggregate into larger institutions which impose a set of rules on a multitude of social situations. How norms and institutions function, emerge and evolve is arguably *the* question for social scientists to answer in the quest for a science of society as a system, rather than a sum of individuals.

Documenting social norms remains an effectively inductive task at this point, a problem which can be approached mostly through ethnography, surveys, or social psychology experiments. Living with a group of people for a year and methodically observing them through ethnography is arguably the best means for the researcher to acquire detailed information about their system of norms, but is also an inherently limited task given the amount of time investment required. Additionally, surveys can be used to probe attitudes about broad subjects, and social psychology experiments can provide behavioral data on particularly important questions. But none of these methods offers a scalable way to summarize the norms of a group, organization, or society. Having a general-purpose way to extract social norms from empirical data would be a very important step in articulating a science of norms and institutions, itself the likeliest answer to the still-unsolved *micro-to-macro* problem long identified by [5].

Until recently data from which social norms could be extracted quantitatively was extremely rare, and collecting extensive records of organizations or societies was but a happy accident. This has changed with the data revolution, spurred by the ubiquitous availability of very large datasets related to computer-mediated communication. A second development has been the emergence of cloud computing to provide the infrastructure for the large-scale processing of very large datasets. These two evolutions have allowed for the extraction of a number of simple linguistic norms, for instance around power and status differentials [6, 7, 28]. Computational social science tools have also allowed for the extraction broad domains of social interaction – usable as the coarse building blocks of social norms – from user wall postings [1]. A novel, keystone development has been the progress made in the development of machine learning algorithms, particularly the emergence of powerful, scalable neural network algorithms – especially in the deep learning sub-discipline [17] – and the creation of multiple toolkits to build them (e.g. TensorFlow, Torch or Theano).

The extraction of social norms may ultimately aid in the understanding of large-scale social interaction as a form of computation. The analytical potential of this idea is already visible in [8–10], who shows the likely existence of non-finite-state computation in human social systems. And even though a quantitative understanding of norms has immense theoretical merits, awareness of the rules of social interaction may also help with multiple practical problems. Normative awareness could help build better virtual assistants, by making them more attuned to their human conversational

[1] This section is a heavily reworked version of the postdoctoral application which formed the basis for this project, as detailed in my project proposal and submission.

CS230, February 2018, Stanford, California USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

partners' behavioral expectations. Or normative awareness could improve recommendation and matching algorithms: it would likely be a helpful thing to understanding when a "five-star" rating really equals excellent service or is just the result of the rater conforming to a behavioral expectation as a matter of course (e.g., in North America, giving anything less than 5 stars on Uber is generally not considered to be justified, unless the customer is truly dissatisfied with the service).

The algorithms presented in this paper focus on a common set of norms related to status, a concept often operationalized as either esteem [30] or competence [25]. Status norms are ubiquitous in social life, and often shape communication decisions ranging from the choice of communication medium to phonetic or stylistic aspects of communication. The algorithm presented here deals with the extraction of status norms from textual interactions in online communities, an increasingly common setting in which social life takes place in the digital age.

2 DATA AND METHODS

To investigate social norms in a quantitative manner I focus on the Ubuntu Dialogue Corpus [19], a dataset comprising many years of Internet Relay Chat (IRC) dialogue between users and developers of Ubuntu, a popular Linux distribution. I downloaded A dataset of 22 million Internet Relay Chat (IRC) exchanges containing 326 million words in total and covering 8 years of activity (from 2004 to 2012) in the main Ubuntu chat room. The dataset has been previously used successfully for Natural Language Applications, particularly dialog prediction [e.g. 15, 20].

The Ubuntu dialog corpus also captures a rich history of an active and extremely successful open-source community which has coalesced around what is arguably (as of 2018) the most popular Linux distribution. Because they create real economic goods without tangible, immediate compensation for contributors, open source communities pose a fundamental *free-rider problem* where strict economic logic would seem to exclude their existence [18]. Social norms around the conferral of status have been advanced as an explanation for the continuing vitality of communities based on the logic of *gift exchange*, as is the case with the Ubuntu user community [31]. Because the Ubuntu Dialog Corpus captures not just a simple collection of speech acts, but the life of an online community, we can expect the normative aspects that have provided the deep structure to social interaction in the chat room to emerge over time.

2.1 Pre-Processing

1.96 million dyads were extracted from the 22 million IRC exchanges which were downloaded. Each message was assigned an ego (person speaking) and an alter (recipient of message), using the following heuristics

- "Ego" was extracted from the IRC user name.
- User names followed by underscores were all assigned to the same Ego.
- "Alter" was extracted from the first word in the chat message, where available. Because the main Ubuntu chat room is high traffic, users often prefixed their references to other conversation partners with their names.

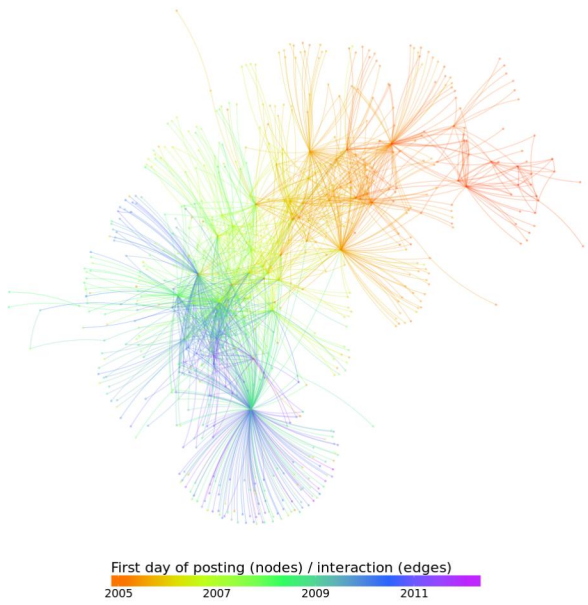

Figure 1: Interactions on Ubuntu IRC, 2004-2012.
Note: Color gradient indicates time of first contribution (for users) or time of first interactions (for edges). Only pairs of users interacting on more than 10 distinct days included in plot.

- If no information was available in the message for a user's current conversation partner, I assigned the user's prior conversation partner in the chat room on the day (if existing) as Alter.
- Opening messages (usually questions addressed to no one in particular, e.g. "Hey everyone, I have a question about driver X") were assigned to all subsequent conversations involving Ego.

2.2 Overview of Social Structure

Figure 1 shows a plot of the 323,237 dyads (between 1,681 total users), which recorded interactions on more than 10 calendar days during the observation window. The graph's nodes and edges are colored according to the first time of interaction (for dyads) and the first time of contribution to the IRC (for users). The graph's coloring evinces multiple "cohorts" of users, as new heavy contributors appear to join and leave the community constantly throughout its life. The individuals displayed in this graph are also likely to be the repository of the community's social norms, the stalwart contributors who, like virtually any cohesive group of individuals are likely to have developed a set of rules of interaction.

2.3 Socio-Liguistic Aspects of Textual Embeddings

?? To facilitate conversation modeling, embedding vectors of 100 real numbers were obtained for 287,346 tokens occurring more than 5 times in the text. The embeddings were trained using fastText [14], using word unigrams and size 3-5 character skipgrams

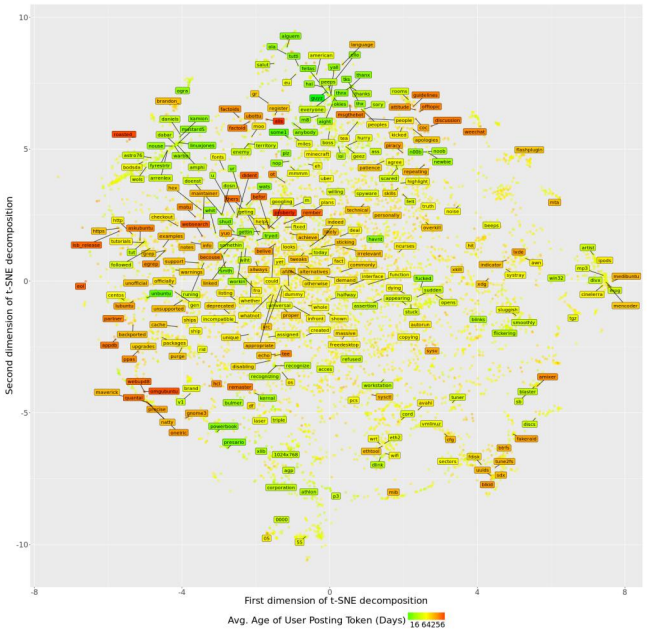

Figure 2: 2-d t-SNE decomposition of fastText embeddings.
Note: Color gradient indicates average age of user posting a token.

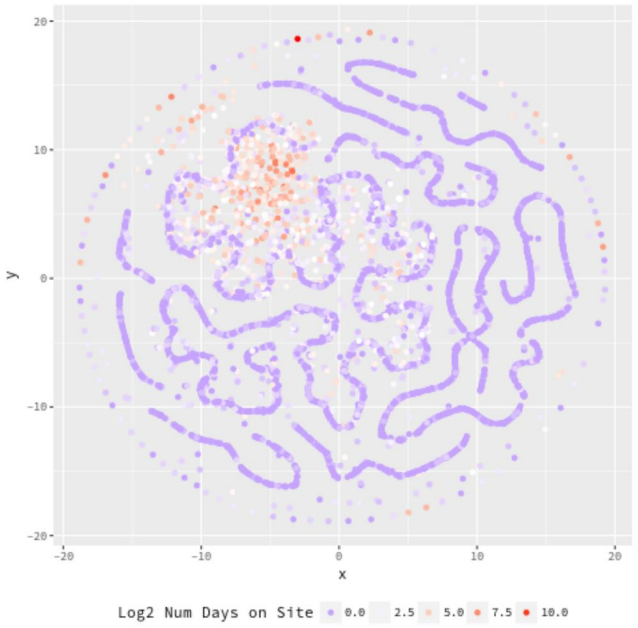

Figure 3: t-SNE decomposition of social graph embeddings.
Note: Color gradient indicates average age of user posting.

(different parametrizations were attempted without substantial improvements to the model fit). Figure 2 shows the 2-dimensional t-SNE embeddings [22] of the size-100 fastText vectors associated with the most common 10,000 tokens. The average user age at the time of usage was computed for each token, and 300 words were selected at random to obtain a roughly-balanced mix of tokens used by new and old users. The visualization shows a detailed view of the technical complexities of the Ubuntu world, but also captures a set of words relevant to social interaction in particular. These words are found in the upper-middle part of the graph, where newcomer (green) tokens such as "n00b" or "thx" cluster together with words usually employed by seasoned users, such as "attitude" or "offtopic."

2.4 Graph Embeddings

To facilitate the simultaneous modeling of social and textual data, a simpler representation was obtained for the social graph of 1.96m interactions (dyads) extracted from IRC conversations. Specifically, the graph was embedded into a 100-dimensional latent space by solving an edge reconstruction problem [3]. Negative samples of random edges were obtained through two strategies: sampling completely at random, and randomly rewiring the social network. Completely random negative edges were selected to be 3 times as numerous as randomly-rewired edges. The embeddings were trained using the Adam optimization algorithm ([16]), under L2 regularization using the 'EmbeddingsBag' module in PyTorch. Using the analytical setup described above, the edge reconstruction model achieved ROC/AUC of .818 against a held-out validation set.

Figure 3 shows a t-SNE visualization of the computed embeddings, colored according to the total time spent on the website by each user. The plot shows an interesting linear pattern, likely a result of training under L2 regularization, which revealed multiple communities of users, many of whom tended to ask questions solely in one area (e.g. NVIDIA graphics drivers, etc.). But one particularly interesting set of users emerges in the upper left corner of the graph. These are users who have spent large numbers of days contributing to the site, a group colored in red in Figure 3. This emergent clustering hints at the feasibility of using number of days spent on the site as proxy for status in the community.

2.5 Status Score from Shrunken Graph Embeddings

Having spent a large number of time in the chat room is likely to be correlated with high status in the Ubuntu IRC community, though this assumption is not without its problems. For one, coming back solely to ask more questions (or worse, to troll) is not likely to be perceived as valuable by the community as returning to provide answers. But Figure 1 reveals another important problem: some "core" users only joined the community years after its founding. Even though their contributions may become as valuable as older users', they will not be active for as many days as older users, given the issue of right-censoring in the data (i.e., we do not observe as much of the lifespan of new users as we do in the case of older users). There are multiple remedies to this problem, but one particularly convenient one emerges from the plot in Figure 3 – we can fit a model to predict the number of days a user would spend in the chat room, using the graph embeddings derived in the previous section as features. Indeed, fitting a multi-layer perceptron[2] yields

[2]The model used first- and second-order polynomials of the graph embeddings, using the following sequence of layers: Linear(200, 200), LeakyReLU(0.5), Linear(200, 200), LeakyReLU(0.5), Linear(200, 1), followed by a softplus activation.

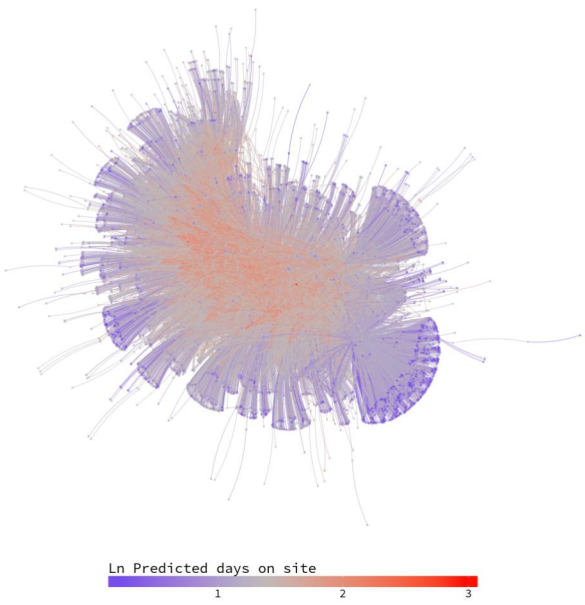

Figure 4: Social network plot, colored by status score

Note: Edges defined as repeated interactions over more than 5 calendar days during the observation window. Edges and nodes colored to reflect the fitted status score from a regression model trained on the social graph embeddings to predict days on the site. Edge colored according to the lowest-scoring node.

a reasonably good fit to the data ($R^2 = .62$).[3] Figure 4 shows what the fitted data resulting from the model looks like when plotted against the social network extracted between individuals who have interacted for more than five distinct calendar days. Despite the graph's multicohort generative process evident from Figure 1, the use of a regression model to predict graph embeddings does well at evidencing what is effectively the core of the Ubuntu IRC network.

2.6 A Status Sequence Model for Textual Interaction

The foregoing discussion demonstrates how we can expect to encounter instantiations of social norms associated both with status giving (i.e., gratitude for help) and status seeking (i.e., appeal to chat room rules). One's prior amount of involvement in the community can thus be expected to be a major factor influencing behavior in the chat room. Specifically, we can expect both interaction partners' age on the site and level of contribution to play a major role in their responses to each other. One difficulty in operationalizing this observation is that status norms operate at a very deep level in the structure of communication, and their presence may not necessarily be detectable without additional context. For instance, the mundane decision to express gratitude as "Thanks.", "Thank You!" or "Thank You SO Much!!!" could connote status norm (where the speaker asserts higher, equal, and lower status, respectively), but it could also be the result of random idiosyncratic circumstances.

This dependency on context is not unlike the one seen in machine translation, query understanding or question-and-answering. Recurrent neural networks (RNNs) have had a transformative effect on all these fields enabling researchers to break countless performance records, most notably for this problem in the area of dialog generation [29]. One particularly expressive model is the Long Short-Term Memory (LSTM) [13]. Much like Hidden Markov Models (HMMs), LSTMs also rely on a hidden state vector to encode latent features that are updated as a sequence is read. An innovation of the LSTM architecture is the addition of a "forget gate" which effectively removes from the hidden state vector information that the model comes to consider superfluous to the optimization problem. This feature has made LSTMs particularly effective in problems with long-range dependencies within sequences, a common feature of text modeling. LSTMs have also shown promising results in the modeling of affect and emotion in in-person dyadic interaction [23], another problem closely related to that of status detection. The bidirectional LSTM (BLSTM) architecture has proven to be particularly useful for sequence classification problems [12], because it allows for gradients to flow in both directions when processing the sequence.

Model Architecture. A BLSTM model was designed to predict the status score[4] of both ego (the speaker) and alter (their partner), a BLSTM model was thus trained on the 22m utterances extracted from the Ubuntu IRC corpus. The model uses as its input the sequence of pre-trained embeddings (of size 100) at the token level obtained from fastText (as described in Section **??**). After each sequence is consumed, the BLSTM model's last hidden layers (from the left-to-right and right-to-left passes) are both averaged into a single input vector, which is then fed into a fully-connected multilayer perceptron, which includes a Dropout layer [27][5], followed by a linear layer with 30 outputs, a LeakyReLU [21] activation, and a final linear layer with two outputs.

Model Training. The multi-objective regression model thus obtained was trained under Mean-Squared Error loss using the Adam [16] optimization algorithm. The model training also allowed for gradient updates on the embedding layer, a modeling choice which was observed to significantly improve performance. The use of a fully-connected MLP on the hidden layer of the LSTM was inspired by similar open-source implementations of LSTM classifications[6]. The model was implemented and trained using PyTorch[7] with the help of an NVIDIA 1080Ti GPU unit. Given the large size of the dataset (22m rows), the model was observed to yield the best performance after only two epochs of training, which completed in under 30 minutes. When evaluated against a held-out dataset the model achieved coefficients of determination (R^2) of .25 for ego's status score and .19 for alter's status score. This was a significant improvement compared to a naive bag-of-words model, which achieved maximum R^2 of .22 for ego and .12 for alter.

Qualitative Evaluation. The multiple models presented in this paper have shown progressively worse fits when evaluated in a classic supervised learning framework. It bears repeating, however,

[3] It bears emphasizing that the goal of this model fit is not a perfect fit but denoising the data.

[4] This score was extracted from graph embeddings as per the previous section.

[5] The intuition for Dropout came from promising results for its use in LSTMs, e.g. [24], as well as empirically-observed improvements in model fit.

[6] e.g. https://github.com/jiangqy/LSTM-Classification-Pytorch.

[7] The implementation is available at https://github.com/bogdanstate/social-norms/model.

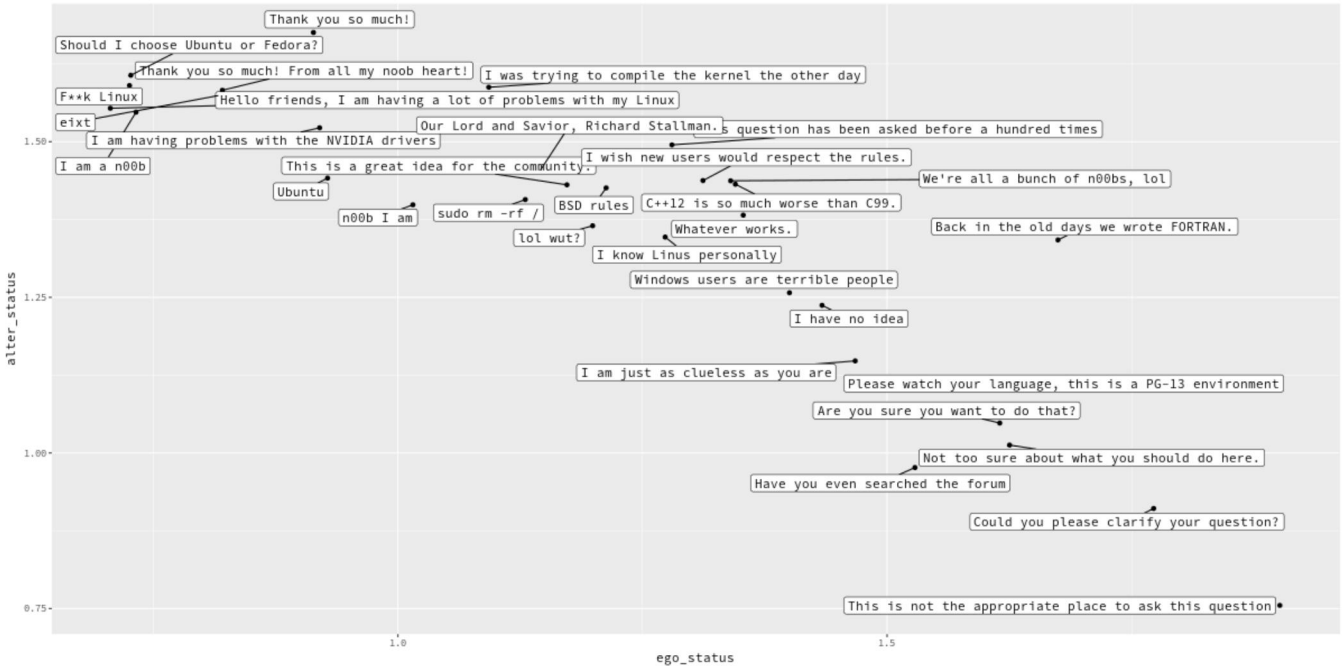

Figure 5: Example Statements and their ratings by Status Model

that understanding status norms is not a task that is easily amenable to the classic train / validation / test supervised learning setup that has proven so effective in countless machine learning problems. Status norms are ultimately latent concepts that are difficult even for humans to adjudicate properly. Additional validation work is thus called upon to validate the model's performance.

Figure 5 plots the predicted scores for a sample of example statements which were designed by the author to be potentially indicative of status norms asserted in IRC conversation on the Ubuntu forums. The model identifies offensive statements such as "F**k Linux" or misspellings of IRC commands ("eixt") as indicators of ego's inexperience and thus low status, as are self-identification statements ("I am a n00b") or general questions ("Should I choose Ubuntu or Fedora?").[8] Utterances typically made by high-status individuals, directed at lower-status ones typically focus on rule enforcement (e.g. "Please watch your language." or "This is not the appropriate place to ask this question."). Statements that contain more insider references to the Linux community (e.g. references to Linus Torvaalds or Richard Stallman, prominent personalities of the open source community, or to variants of C++ or the Linux kernel) are likely to be exchanged between mid-status peers. Finally, a statement like "Back in the old days we wrote FORTRAN." is indicative of an exchange between a very high status ego and a slightly lower-status alter – this is the sort of assertion that could serve to enforce a more fine-grained hierarchy among the already-established individuals in the community.

3 CONCLUSIONS

"Better than human" performance has become an almost mundane feature of the steady stream of breakthroughs coming from artificial neural network research. But there is an entire class of problems for which the very notion of exceeding human performance of ill-defined or even nonsensical. These are the problems that call on artificial intelligence to understand humans as social beings. Their resolution may turn out to be even trickier than the hardest problems in computer vision or speech recognition. But as technology continues to evolve and computers become ever more embedded in the everyday, developing an understanding of the subtle rules that govern human lives is crucial for building humane AIs.

The algorithms presented in this paper represent an exercise in the extraction of meaning from a complex and noisy dataset. It is certainly neither the most theoretically satisfying nor the best practical implementation of the concepts presented here. For one, modeling entire dyads (rather than individual utterances) would likely improve the performance of the model. And taking into account the multiple temporal dimensions of the data, as well as the sparsity of certain kinds of interactions (e.g. noob-noob) would certainly make for a more expressive model. Finally, truly learning status norms would require even more in-depth investigation into the extraction of conditional rules from the sea of parameters available in a neural network. All of these tasks are promising extensions that may lead us just a bit closer to building a robust methodology for the practical identification of basic sociological concepts.

[8] In the context of the IRC chat rooms most interactions of low status individuals ("noobs") tend to be with established individuals who enjoy higher status. This serves to explain why all typical newcomer statements are identified as being directed at a higher-status individual.

4 PROJECT REPOSITORY

Implementation available at:
https://github.com/bogdanstate/social-norms/
Main model is in encoder_lstm.py.

REFERENCES

[1] Luca Maria Aiello, Rossano Schifanella, and Bogdan State. 2014. Reading the source code of social ties. In *Proceedings of the 2014 ACM conference on Web science*. ACM, 139–148.

[2] Kenneth Bettenhausen and J Keith Murnighan. 1985. The emergence of norms in competitive decision-making groups. *Administrative science quarterly* (1985), 350–372.

[3] Hongyun Cai, Vincent W Zheng, and Kevin Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques and applications. *IEEE Transactions on Knowledge and Data Engineering* (2018).

[4] Robert B Cialdini, Raymond R Reno, and Carl A Kallgren. 1990. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology* 58, 6 (1990), 1015.

[5] James S Coleman. 1994. *Foundations of social theory*. Harvard University Press.

[6] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 699–708.

[7] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078* (2013).

[8] Simon DeDeo. 2012. Evidence for Non-Finite-State Computation in a Human Social System. *arXiv preprint arXiv:1212.0018* (2012).

[9] Simon DeDeo. 2014. Group minds and the case of Wikipedia. *arXiv preprint arXiv:1407.2210* (2014).

[10] Simon DeDeo. 2015. Conflict and Computation on Wikipedia: a Finite-State Machine Analysis of Editor Interactions. *arXiv preprint arXiv:1512.04177* (2015).

[11] Joshua M Epstein. 2001. Learning to be thoughtless: Social norms and individual computation. *Computational economics* 18, 1 (2001), 9–24.

[12] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).

[15] Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. 2015. Improved deep learning baselines for ubuntu corpus dialogs. *arXiv preprint arXiv:1510.03753* (2015).

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[18] Josh Lerner and Jean Tirole. 2005. The economics of technology sharing: Open source and beyond. *Journal of Economic Perspectives* 19, 2 (2005), 99–120.

[19] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909* (2015).

[20] Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse* 8, 1 (2017), 31–65.

[21] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, Vol. 30. 3.

[22] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[23] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2, 2 (2011), 92–105.

[24] Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. 2014. Dropout improves recurrent neural networks for handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 285–290.

[25] Cecilia L Ridgeway. 1978. Conformity, group-oriented motivation, and status attainment in small groups. *Social Psychology* (1978), 175–188.

[26] Muzafer Sherif. 1936. *The psychology of social norms*. Harper.

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[28] Bogdan State, Bruno Abrahao, and Karen Cook. 2016. Power Imbalance and Rating Systems. In *Tenth International AAAI Conference on Web and Social Media*.

[29] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).

[30] Max Weber. 1978. *Economy and society*. Vol. 1. University of California Press.

[31] Robb Willer, Mathew Feinberg, Francis J Flynn, and Brent Simpson. 2011. Is generosity sincere or strategic? Altruism versus status-seeking in prosocial behavior. *Unpublished manuscript* (2011).