# Power Networks: A Novel Neural Architecture to Predict Power Relations

**Michelle Lam**[*1], **Catherina Xu**[*1], **Angela Kong**[1], and **Vinodkumar Prabhakaran**[1]

[1]Department of Computer Science, Stanford University, Stanford, CA, USA.
{mlam4, yuex, akong2, vinod}@cs.stanford.edu

## Abstract

Can language analysis reveal the underlying social power relations that exist between participants of an interaction? Prior work within NLP has shown promise in this area, but the performance of automatically predicting power relations using NLP analysis of social interactions remains wanting. In this paper, we present a novel neural architecture that captures manifestations of power within individual emails which are then aggregated in an order-preserving way in order to infer the direction of power between pairs of participants in an email thread. We obtain an accuracy of 80.4%, a 10.1% improvement over state-of-the-art methods, in this task. We further apply our model to the task of predicting power relations between individuals based on the entire set of messages exchanged between them; here also, our model significantly outperforms the 70.0% accuracy using prior state-of-the-art techniques, obtaining an accuracy of 83.0%.

## 1 Introduction

With the availability and abundance of linguistic data that captures different avenues of human social interactions, there is an unprecedented opportunity to expand NLP to not only understand language, but also to understand the *people* who speak it and the *social relations* between them. Social power structures are ubiquitous in human interactions, and since power is often reflected through language, computational research at the intersection of language and power has gained interest recently. This research has been applied to a wide array of domains such as Wikipedia talk pages (Strzalkowski et al., 2010; Taylor et al., 2012; Danescu-Niculescu-Mizil et al., 2012; Swayamdipta and Rambow, 2012), blogs (Rosenthal, 2014) as well as workplace interactions (Bramsen et al., 2011; Gilbert, 2012; Prabhakaran, 2015).

The corporate environment is one social context in which power dynamics have a clearly defined structure and shape the interactions between individuals, making it an interesting case study on how language and power interact. Organizations stand to benefit greatly from being able to detect power dynamics within their internal interactions, in order to address disparities and ensure inclusive and productive workplaces. For instance, (Cortina et al., 2001) reports that women are more likely to experience incivility, often from superiors. It has also been shown that incivility may breed more incivility (Harold and Holtz, 2015), and that it can lead to increased stress and lack of commitment (Miner et al., 2012).

Prior work has investigated the use of NLP techniques to study manifestations of different types of power using the Enron email corpus (Diesner et al., 2005; Prabhakaran et al., 2012; Prabhakaran and Rambow, 2013; Prabhakaran and Rambow, 2014). While early work (Bramsen et al., 2011; Gilbert, 2012) focused on surface level lexical features aggregated at corpus level, more recent work has looked into the thread structure of emails as well (Prabhakaran and Rambow, 2014). However, both (Bramsen et al., 2011; Gilbert, 2012) and (Prabhakaran and Rambow, 2014) group all messages sent by an individual to another individual (at the corpus-level and at the thread-level, respectively) and rely on word-ngram

---

* Authors (listed in alphabetical order) contributed equally.

based features extracted from this concatenated text to infer power relations. They ignore the fact that the text comes from separate emails, and that there is a sequential order to them.

In this paper, we propose a hierarchical deep learning architecture for power prediction, using a combination of Convolutional Neural Networks (CNN) to capture linguistic manifestations of power in individual emails, and a Long Short-Term Memory (LSTM) that aggregates their outputs in an order-preserving fashion. We obtain significant improvements in accuracy on the corpus-level task (82.4% over 70.0%) and on the thread-level task (80.4% over 73.0%) over prior state-of-the-art techniques.

## 2   Data and Problem Formulation

We use the version of the Enron Email corpus released by Agarwal et al. (2012) that captures the organizational power relation between 13,724 pairs of Enron employees, in addition to the reconstructed thread structure of email messages added by Yeh and Harnly (2006). We mask greetings and signature lines in the email content to prevent our model from being biased by the roles held by specific employees.

| Entity type | # of Pairs | Train | Dev | Test |
|---|---|---|---|---|
| Per-Thread | 15,048 | 7,510 | 3,578 | 3,960 |
| Grouped | 3,755 | 2,253 | 751 | 751 |

Table 1: Data instance statistics by problem formulation.

Prior work on NLP approaches to predict power in organizational email has used two different problem formulations — *Per-Thread* and *Grouped*. We investigate both formulations in this paper. Table 1 shows the number of data instances in each problem formulation.

**Per-Thread**: This formulation was introduced by Prabhakaran and Rambow (2014) in which, for a given thread *t* and a pair of related interacting participant pairs (*A*, *B*), the direction of power between *A* and *B* is predicted (where the assignment of labels *A* and *B* is arbitrary). The participants in these pairs are 1) *interacting*: at least one message exists in the thread such that either *A* is the sender and *B* is a recipient or vice versa, and 2) *related*: *A* and *B* are related by a dominance relation (either superior or subordinate) based on the organizational hierarchy. As in (Prabhakaran and Rambow, 2014), we exclude pairs of employees who are peers, and we use the same train-dev-test splits so our results are comparable.

**Grouped**: Here, we group all emails *A* sent to *B* across all threads in the corpus, and vice versa, and use these sets of emails to predict the power relation between *A* and *B*. This formulation is similar those in (Bramsen et al., 2011; Gilbert, 2012), but our results are not directly comparable since, unlike them, we rely on the ground truth of power relations from (Agarwal et al., 2012); however, we created an SVM model that uses word-ngram features similar to theirs as a baseline to our proposed neural architectures.

## 3   Methods

The inputs to our models take on two forms: **Lexical features:** We represent each email as a series of tokenized words, each of which is represented by a 100-dimensional GloVe vector pre-trained on Wikipedia and Gigaword (Pennington et al., 2014). We cap the email length at a maximum of 200 words. **Non-lexical features:** We incorporate the structural non-lexical features identified as significant by Prabhakaran and Rambow (2014) for the *Grouped* problem formulation. We used (1) the average number of recipients and (2) the average number of words in each email for each individual; these features were concatenated into a single input vector. We investigate the following three network architectures, in increasing order of complexity, to train our model:
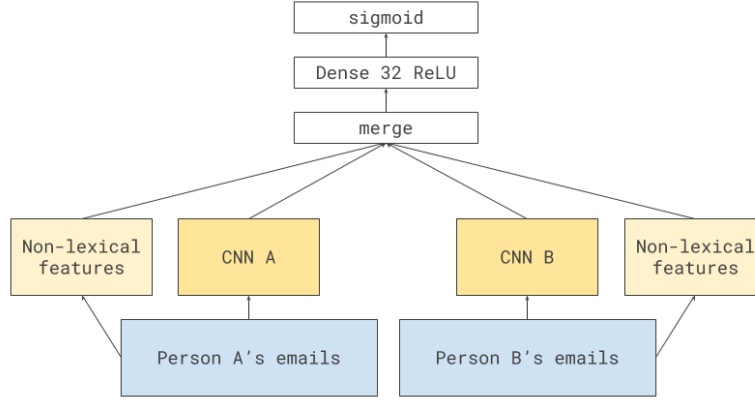
Figure 1: Batched emails (Batched-CNN)

**Approach 1: Batched emails (Batched-CNN).** In this model (see Figure 1), all of *A*'s emails to *B* are batched and fed into a Convolutional Neural Network (CNN), and the same operation is performed for *B*'s emails to *A*. The format of this input is described earlier in this section. This representation can be thought of as a neural equivalent of the SVM-based approaches in prior work, since they merge together all emails in either direction as a single unit. Then, the output of these two CNNs is merged with the non-lexical features from *A*'s emails and *B*'s emails, passed through a dense layer with rectified linear unit (ReLU) activation, and fed to a sigmoid classifier that predicts the power relation between *A* and *B*.
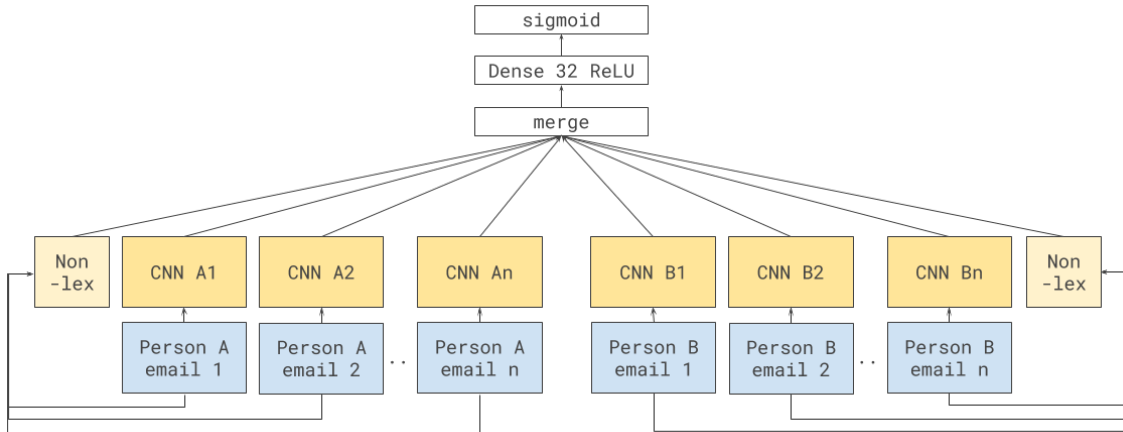


Figure 2: Separated emails (Separated-CNN)

**Approach 2: Separated emails (Separated-CNN).** In this model (see Figure 2), we capture the essence of individual emails by separating them in the model input. As in Batched-CNN, we separate *A*'s and *B*'s emails, but here we feed *each* email as input to a CNN. The motivation here is to first capture local patterns from individual emails. We then merge the output of these CNNs with the non-lexical features from *A*'s and *B*'s emails, pass this to a dense layer with ReLU activation, and pass the result to a sigmoid classifier that predicts the power relation.

**Approach 3: Sequential emails (Sequential-CNN-LSTM).** Finally, we use a third model where we account for the temporal order of emails, which may be important in the case of the *Per-Thread* formulation. In this model (see Figure 3), we separate each individual's emails, feed each email to a CNN, and pass the sequence of CNN outputs for each email to a Long Short-Term Memory network (LSTM) for that individual. We then merge the resulting output of the two LSTMs with the non-lexical features from each individual's emails, pass it on to a dense layer with ReLU activation, and then to a sigmoid classifier for the final prediction.
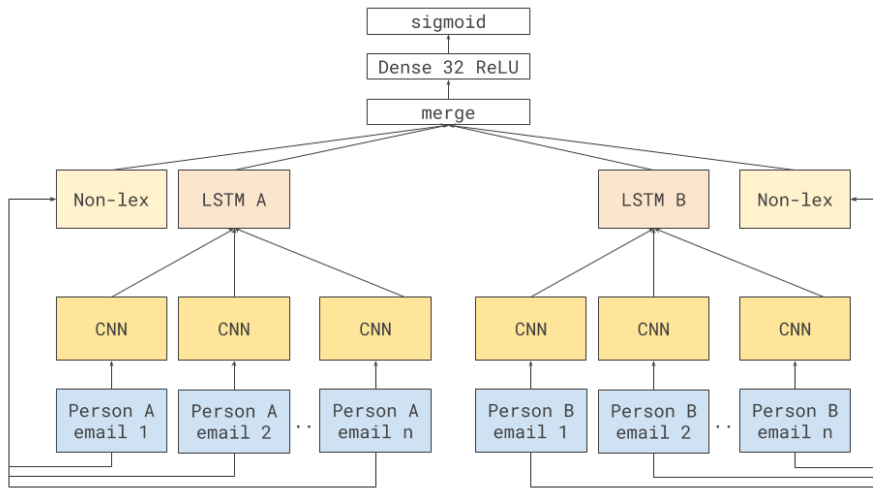
99

Figure 3: Sequential emails (Sequential-CNN-LSTM)

## 4 Experiments and Results

We use support vector machine (SVM) based approaches as our baseline, since they are the state-of-the art in this problem (Prabhakaran and Rambow, 2014; Bramsen et al., 2011; Gilbert, 2012). We use the performance reported by (Prabhakaran and Rambow, 2014) using SVM as baseline for the *Per-Thread* formulation (using the same train-dev-test splits) and implemented an SVM baseline for the *Grouped* formulation (not directly comparable to performance reported by (Bramsen et al., 2011; Gilbert, 2012)).

For each of our neural net models, we trained for 30-70 epochs until the performance on the development set stopped improving, in order to avoid overfitting. We used Hyperas to tune hyperparameters on our development dataset for the same set of parameter options for each task formulation, varying activation functions, hidden layer size, batch size, dropout, number of filters, and number of words to include per email.[1]

Table 2 presents the accuracy obtained using different models. All of our models significantly outperformed the SVM baseline in both task formulations. In the *Per-Thread* formulation, we obtained the best accuracy of 80.4% using the Sequential-CNN-LSTM approach, compared to the 73.0% reported by (Prabhakaran and Rambow, 2014). This is also a marked improvement over the simpler Batched-CNN and Separated-CNN models. This suggests that both temporal and local

| Model | Per-Thread | Grouped |
|---|---|---|
| SVM Baseline | 73.0 | 70.0 |
| Batched-CNN | 78.7 | 82.0 |
| Separated-CNN | 79.8 | **83.0** |
| Sequential-CNN-LSTM | **80.4** | 82.4 |

Table 2: Accuracy obtained using different models SVM Baseline: (Prabhakaran and Rambow, 2014)

email features aid in the power prediction task within the *Per-Thread* formulation. In the *Grouped* formulation, the Separated-CNN model obtained the best accuracy of 83.0%, outperforming the Sequential-CNN-LSTM accuracy of 82.4%. We hypothesize that this is because the grouped formulation does not inherently have a temporal structure between emails, unlike the thread formulation where Sequential-CNN-LSTM is able to tap into the temporal structure.

Table 3 presents a few emails from our corpus, along with the true and predicted labels for the power relation between their sender and recipient(s). Our model seems to pick up on linguistic signals of lack of power such as relinquishing agency (*let me know who you'd like us to work with*), and status reports (*model is nearly completed*), as well as overt displays of power such as *I personally would like to see the*

---

[1] https://github.com/maxpumperla/hyperas

*results* and *we need to end all payments*. On the other hand, the model picks up on the phrasing in *don't use the ftp site* as displaying superiority while the superiority displayed here may have been derived from the expertise the subordinate has in file-transfer protocols. Similarly, the model may have misunderstood the overtly polite phrasings in the last email sent by a superior to be subordinate-like behavior. This sheds light on an important challenge in this task: superiors don't express their superiority in all emails, and subordinates may sometimes display power derived from other sources such as expertise. In such cases where text features alone are not informative enough, signals from additional non-lexical features may be key to accurate classification.

| Text extracted from email | Actual | Predicted |
|---|---|---|
| Let me know who you'd like us to work with in your group. The Adaytum planning model is nearly completed. | Subordinate | Subordinate |
| Vince is hosting on Wharton and a project they are doing for us, I personally would like to see the results of that before doing more with Wharton. | Superior | Superior |
| We need to end all payments as of December 31, 2001. | Superior | Superior |
| Don't use the ftp site while I am away [...] I will check my messages when I return. | Subordinate | Superior |
| Here is the draft letter for your consideration. Please distribute this to the members of the legal team at Enron. Thank you for your assistance, and have a happy holiday season. | Superior | Subordinate |

Table 3: Example power labels from Separated-CNN on the Grouped formulation. For *correct* labels, text segments that may signal the power relation are underlined in green; for *incorrect* labels, potentially confusing power signals are underlined in red. (text segments chosen based on our qualitative judgment).

## 5 Concluding Discussions

In this paper, we investigated the intersection between language and power in the corporate domain via neural architectures grounded in an understanding of how expressions of power unfold in email. Our Sequential-CNN-LSTM model, which utilizes an LSTM to capture the temporal relations underlying per-email features, achieved 80.4% accuracy in predicting the direction of power between participant pairs in individual email threads, which is a 10.1% accuracy improvement over the state-of-the-art approach (Prabhakaran and Rambow, 2014). Our Separated-CNN model also obtains an accuracy of 83.0% in predicting power relations between individuals based on the entire set of messages exchanged between them, a significant boost over 70.0% accuracy obtained using traditional methods. We also present a qualitative error analysis that sheds light on the patterns that confuse the model.

To further our work, we plan to granularize the level at which features are learned. We hypothesize that by training a CNN on each *sentence* rather than email, the model will better capture mid-level indicators of power that occur between the word level and email level. We will also investigate ways to better incorporate structural features by accounting for their relevance to a holistic judgment of power; for example, features like gender and temporal position in a thread are more suited to merge with a higher level of the architecture like the per-individual LSTMs while features like number of email tokens are more suited to merge at the low level of the per-email CNNs. Lastly, we plan to incorporate additional datasets such as the Avocado Research Email Collection (Oard et al., 2015) to study cross-corpora performance.

## 6 Acknowledgements

# References

Apoorv Agarwal, Adinoyi Omuya, Aaron Harnly, and Owen Rambow. 2012. A Comprehensive Gold Standard for the Enron Organizational Hierarchy. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 161–165, Jeju Island, Korea, July. ACL.

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, June. ACL.

Lilia M Cortina, Vicki J Magley, Jill Hunter Williams, and Regina Day Langhout. 2001. Incivility in the Workplace: Incidence and Impact. *Journal of Occupational Health Psychology*, 6(1):64.

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*, WWW '12, New York, NY, USA. ACM.

Jana Diesner, Terrill L Frantz, and Kathleen M Carley. 2005. Communication Networks from the Enron Email Corpus It's Always About the People. Enron is no Different. *Computational & Mathematical Organization Theory*, 11(3):201–228.

Eric Gilbert. 2012. Phrases That Signal Workplace Hierarchy. *CSCW '12 Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1037–1046.

Crystal M Harold and Brian C Holtz. 2015. The Effects of Passive Leadership on Workplace Incivility. *Journal of Organizational Behavior*, 36(1):16–38.

Kathi N Miner, Isis H Settles, Jennifer S Pratt-Hyatt, and Christopher C Brady. 2012. Experiencing Incivility in Organizations: The Buffering Effects of Emotional and Organizational Support. *Journal of Applied Social Psychology*, 42(2):340–372.

Douglas Oard, William Webber, David Kirsch, and Sergey Golitsynskiy. 2015. Avocado Research Email Collection. *Philadelphia: Linguistic Data Consortium*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Vinodkumar Prabhakaran and Owen Rambow. 2013. Written Dialog and Social Power: Manifestations of Different Types of Power in Dialog Behavior. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 216–224, Nagoya, Japan, October. Asian Federation of NLP.

Vinodkumar Prabhakaran and Owen Rambow. 2014. Predicting Power Relations between Participants in Written Dialog from a Single Thread. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 339–344.

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, December. The COLING 2012 Organizing Committee.

Vinodkumar Prabhakaran. 2015. *Social Power in Interactions: Computational Analysis and Detection of Power Relations*. Ph.D. thesis, Columbia University.

Sara Rosenthal. 2014. Detecting Influencers in Social Media Discussions. *XRDS: Crossroads, The ACM Magazine for Students*, 21(1):40–45.

Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Samira Shaikh, Sarah Taylor, and Nick Webb. 2010. Modeling Socio-Cultural Phenomena in Discourse. In *Proceedings of the 23rd International Conference on COLING 2010*, Beijing, China, August. Coling 2010 Organizing Committee.

Swabha Swayamdipta and Owen Rambow. 2012. The Pursuit of Power and Its Manifestation in Written Dialog. *2012 IEEE Sixth International Conference on Semantic Computing*, 0:22–29.

Sarah M. Taylor, Ting Liu, Samira Shaikh, Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Umit Boz, Xiaoai Ren, Jingsi Wu, and Feifei Zhang. 2012. Chinese and American Leadership Characteristics: Discovery and Comparison in Multi-party On-Line Dialogues. In *ICSC*, pages 17–21.

J.Y. Yeh and A. Harnly. 2006. Email Thread Reassembly Using Similarity Matching. In *Third Conference on Email and Anti-Spam (CEAS)*, pages 27–28.