# Supplementary Material:
# Interpreting Neural Networks to Improve Politeness Comprehension

**Malika Aubakirova**
University of Chicago
`aubakirova@uchicago.edu`

**Mohit Bansal**
UNC Chapel Hill
`mbansal@cs.unc.edu`

## 1 Supplementary Material

This is supplementary material for the main paper, where we present more analysis and visualization examples, and our dataset and training details.

## 2 Activation Clusters

### 2.1 Rediscovering Existing Strategies

**Gratitude (+)** Respect and appreciation paid to the listener. Activation cluster examples: "{*thanks for the heads up*"; "*thank you very much for this kind gesture*"; "*thanks for help!*"}

**Greeting (+)** A welcoming message for the converser. Activation cluster examples: {"*hey, long time no seeing! how's stuff?*"; "*greetings, sorry to bother you here...* "}

**Positive Lexicon (+)** Expressions that build a positive relationship in the conversation and contain positive words from the sentiment lexicon, e.g., *great*, *nice*, *good*. Activation cluster examples: {"*your new map is a great*"; "*very nice article*"; "*yes, this is a nice illustration. i 'd love to...*"}

**Counterfactual Modal (+)** Indirect strategies that imply a burden on the addressee and yet provide a face-saving opportunity of denying the request, usually containing hedges such as *Would it be.../Could you please*. Activation cluster examples: {"*would you be interested in creating an infobox for windmills...?*; "*would you mind retriveing the bibliographic data?*"}

**Deference (+)** A way of sharing the burden of a request placed on the addressee. Activation cluster examples: {"*nice work so far on your rewrite...*", "*hey, good work on the new pages...*", "*good point for the text...*", "*you make some good points...*"}

**Direct Question (-)** Questions imposed on the converser in a direct manner with a demand of a factual answer. Activation cluster examples: {"*why would one want to re-create gnaa?*"; "*what's with the radio , and fist in the air?*"; "*what level warning is appropriate?*"}

### 2.2 Extending Existing Strategies

**Counterfactual Modal (+)** Sentences with *Would you/Could you* get grouped together as expected; but in addition, the cluster contains requests with *Do you mind*. Activation cluster examples: "{*do you mind having another look?*"; "*do you mind if i migrate these to your userspace for you?*"}

**Gratitude (+)** Our CNN learns a special shade of gratitude, namely it distinguishes a cluster consisting of the bigram *thanks for*. Activation cluster examples: "*thanks for the good advice.*"; "*thanks for letting me know.*"; "*fair enough, thanks for assuming good faith*"}

**Indicative Modal (+)** The same neuron as for counterfactual modal cluster above also gets activated on gapped 3-grams like *Can you ... please?*, which presumably implies that the combination of a later *please* with future-oriented variants *can/will* in the request gives a similar effect as the conditional-oriented variants *would/could*. Activation cluster examples: "*can this be reported to london grid,* *please?*"; "*can you delete it again, please?*"; "*good start . can you add more, please?*"}

### 2.3 Discovering Novel Strategies

**Indefinite Pronouns (-)** Danescu-Niculescu-Mizil et al. (2013) distinguishes requests with first and second person (plural, starting position, etc.). However, we find activations that also react to indefinite pronouns such as *something/somebody*. Activation cluster examples: {"*am i missing something here?*"; "*he 's gone. was it something i said?*"; "*you added the tag and then mentioned it on talk- you did not gain consensus first or even wait for anyone to discuss it.* "; "*but how can something be both correct and incorrect*" }

**Punctuation (-)** Though non-characteristic in direct speech, punctuation appears to be an important special marker in online communities, which in some sense captures verbal emotion in text. One of our neuron clusters gets activated on question marks "**???**" and one on ellipsis "**...**". Activation cluster examples of question marks: {"*now???*"; "*original article????*"; "*helllo?????*"} Activation cluster examples of ellipsis: {"*ummm , it 's a soft redirect. a placeholder for a future page* **...** *is there a problem ?*"; "*Indeed* **...** *the trolls just keep coming.*"; "*I can't remember if i asked/checked to see if it got to you? so* **...** *did it ?*"}

## 3 First Derivative Saliency

In Fig. 1, we show some additional examples of saliency heatmaps. In the first heatmap, we see a clear example of the Positive Lexicon politeness strategy. The key *great* captures most of the weight for the final decision making. Note that, in particular, the question mark in this case provides no influence. Contrast that to the second figure, which echos back the proposed negative politeness strategy on punctuation from Section 6.1.3. Initial question marks give a high influence in magnitude for the negative predicted label. In the third example, we see that these punctuation markers still provide a lot of emphasis. For instance, other words such as *really*, *successful* and a personal pronoun *I* have very little impact. Overall, this exemplifies Direct Question strategy since most of the focus is on *why*.

As was noted in the embedding space transformations discussion, the Gratitude key *thanks* with a preposition *for* has a much stronger polarity than other positive politeness keys in the fourth heatmap. Indeed, *can you please* does not nearly provide as much value. In the fifth heatmap, the sensitivity of the final score comes more from the greeting, namely *hi*, as compared to the phrase *can you please tell me* or positive lexicon *very nice*. These results match the politeness score results in Table 3 of Danescu-Niculescu-Mizil et al. (2013), where the Greeting strategy has a score of 0.87 compared to 0.49 for Please strategy and 0.12 for Positive Lexicon strategy. The sixth and last heatmap demonstrates the contribution of indefinite pronouns. In this case phrase *am I missing something* with the focus on the latter two words decides the final label prediction.

## 4 Dataset and Training Details

We split the Wikipedia and Stack Exchange datasets of Danescu-Niculescu-Mizil et al. (2013) into training, validation and test sets with 70%, 10%, and 20% of the data respectively (after random shuffling). Therefore, the final split for Wikipedia is 1523, 218 and 436; and for Stack Exchange it is 2298, 328, and 657, respectively. We will make the dataset split indices publicly available.

We use 300-dim pre-trained `word2vec` embeddings as input to the CNN Mikolov et al. (2014), and then allow fine-tuning of the embeddings during training. All sentence tokenization is done using NLTK (Bird, 2006). For words not present in the pre-trained set, we use uniform unit scaling initialization.

We implement our model using a python version of TensorFlow (Abadi et al., 2015). Hyperparameters, e.g., the mini-batch size, learning rate, optimizer type, and dropout rate were tuned using the validation set of Wikipedia via grid search.[1] The final chosen values were a mini-batch size of 32, a learning rate of 0.001 for the Adam Optimizer, a dropout rate of 0.5, filter windows are 3, 4, and 5

---

[1] Grid search was performed over dropout rates from 0.1 to 0.9 in increments of 0.1; four learning rates from 1e-1 to 1e-4; Adam, SGD, and AdaGrad optimizers; filter windows sizes from 1 to 3-grams; and features maps ranging from 10 to 200 with an incremental step of 20.

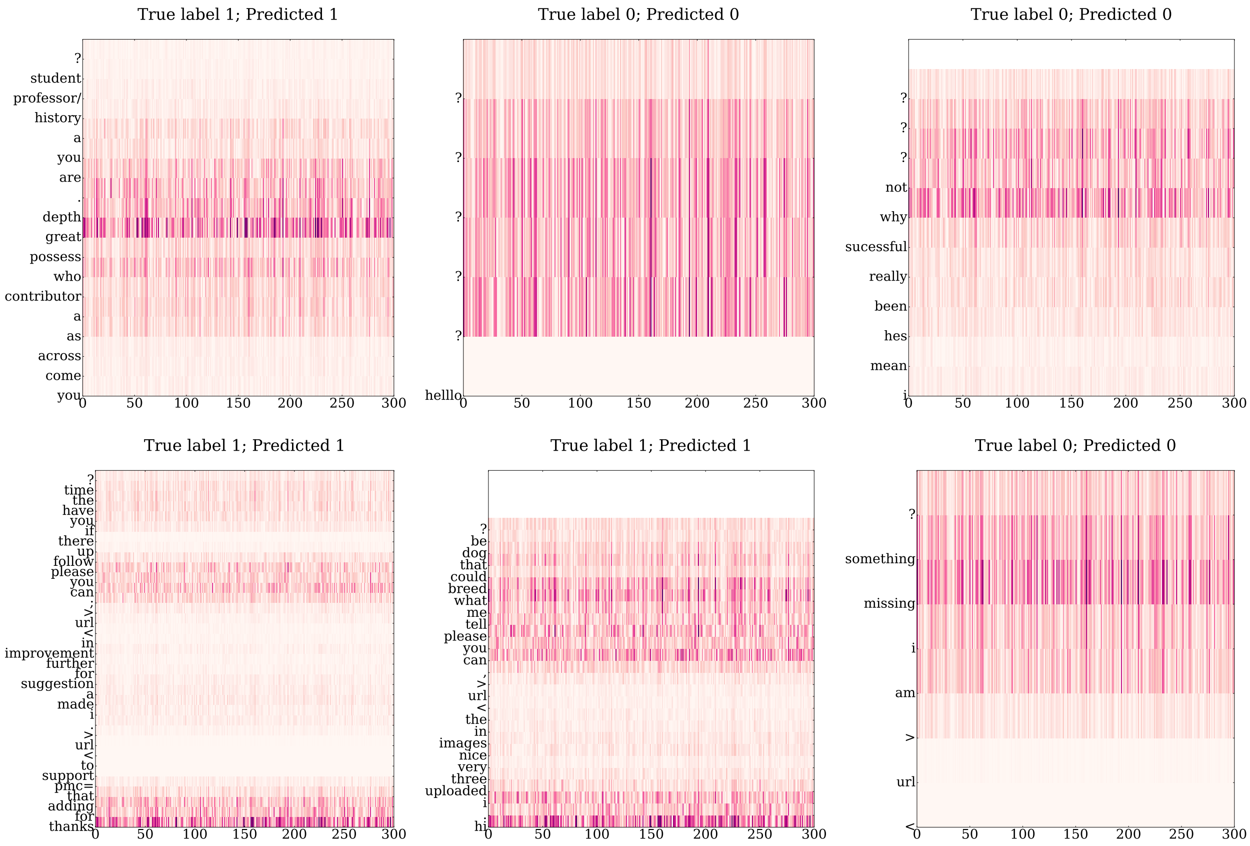

**Figure 1:** Additional saliency heatmaps for correctly classified sentences.

with 75 feature maps each, and ReLU as non-linear transformation function (Nair and Hinton, 2010). For convolution layers, we use valid padding and strides of all ones.

## References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of ACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2014. word2vec.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of ICML*, pages 807–814.