An evaluation of sentiment models relative human coding in pragmatically-defined speech

Anonymous ACL submission

1 Introduction

Computational methods aim to closely replicate human sentiment encodings, yet the simple NLP models traditionally employed by linguists show variable success in task performance (Kansara and Sawant, 2020). This extended abstract illustrates the ways in which more sophisticated yet accessible sentence- or discourse-level techniques, such as Long-Short Term Memory (LSTM) modeling, can better match human sentiment analyses as compared to word-level techniques, such as Bag-of-Words models, by mimicking aspects of language pragmatics (Comstock, 2015; Thomas, 2014). Our aim is to encourage linguists to consider the implications of model selection for their analysis task.

2 Methods

We utilize a corpus of questions posed to Russian Presidents at the G8 and G20 press conferences from 2000 to 2019. Transcripts are sourced from the Kremlin press archive (`http://kremlin.ru`, `http://en.kremlin.ru/`). All texts are in Russian. The corpus comprised 256 questions (12338 words).

2.1 Human coding

Questions occur within a larger text referred to as a "questioning turn–QT" (Clayman et al., 2006). Human coding consisted of labeling (i) individual sentences–ST within a QT as "positive," "negative," and "neutral," and then aggregating sentence-level codes to (ii) classify the entire QT according to one of the three categories. A second analysis considered additional contextual information to subdivide each category: (i) positive politically-related questions vs. non-political human interest questions, and (ii) questions hostile towards the policy described vs. toward the lack of solidarity exhibited among summit members. Details of the coding scheme have been published (Comstock, 2023).

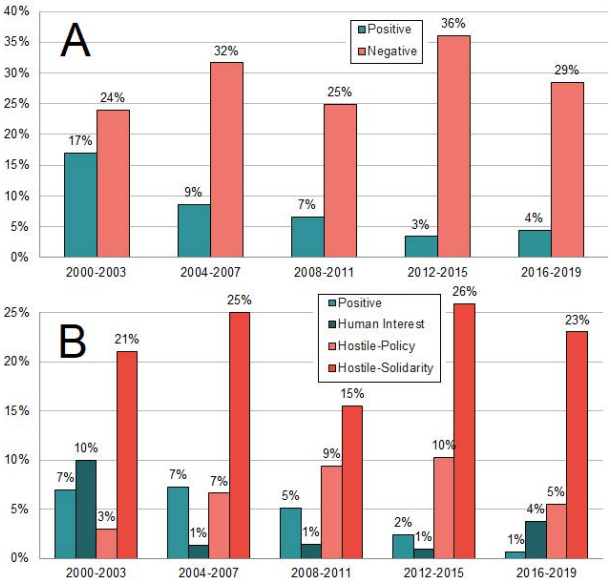

Figure 1: Human coding. (A) Simple sentiment coding. (B) Contextual sentiment coding. Axes show the percentage of correct labels for STs by presidential term.

Human coding illustrated a trend towards less positive sentiment over time (see Figure 1). Contextually-determined subcategories were well-represented, particularly in the positive category.

2.2 Matching human coding: Bag-of-Words

The Bag-of-Words (BoW) analysis used a modified Lexicoder Sentiment Dictionary to determine key sentiment word cues (Young and Soroka, 2012) from the English transcripts; afterward, words that were summit-specific (i.e., "resolution", etc.) or carried a different sentiment in Russian were removed by a professional Russian translator. Positive and negative words were divided by the total number of words in a given presidential term to determine sentiment frequency by term.

2.3 Matching human coding: Neural Network

The model and training data were adapted from an open-source Kaggle competition for sentiment

analysis of Russian news. The best-performing convolutional neural network bidirectional long-short-term memory (CNN-BiLSTM) model was used (https://www.kaggle.com/code/thehemen/cnn-bilstm-russian-news-classifier). Training data comprised 8263 excerpts of varying length from pre-classified Russian news articles. The model predicted sentiments of STs and QTs in our data, assigning a “positive,” “negative,” or “neutral” label.

3 Results

When compared to the human coding, both computational models were ineffective in matching the actual percentages of human-coded sentiment. The BoW analysis not only showed much lower percentages of captured sentiment but also did not accurately reflect the trends shown in the human coding (see Figure 2). While the LSTM analysis of STs captures "more sentiment," this may be misleading: it is ineffective in capturing the true scope of sentiment and the trends over time as compared to the QT analysis.

Figure 2: Sentiment analyses. (A) BoW analysis. (B) LSTM analysis of STs. (C) LSTM analysis of QTs. Axes show the percentage of correct labels for words, STs, or QTs by presidential term.

4 Discussion

Although more sophisticated models such as LSTM increase complexity and involve a greater learning curve to utilize, the results are markedly better at reflecting trends across time. As neither model is accurate on an item-by-item basis, LSTM models are strongly preferable to capture sentiment analysis trends. Consideration of extra contextual data appears to have boosted the QT model performance: when individual items are reviewed with their assigned labels, we see that the BoW analysis captures a large portion of "positive" and "hostile-policy" human-coded data, but not the "human interest" or "hostile-solidarity" data. The LSTM performs markedly better in accounting for these more pragmatically defined subcategories.

5 Limitations and further direction

In our extended abstract, we will provide text excerpts to illustrate how the various models focus on different pragmatic elements of the sentence. We will also have space to provide statistical analyses. LSTM performance is highly dependent on training data; using a dataset more closely related to political questioning than the given Russian media dataset might improve or change our findings. Additionally, there is a large gap in complexity between BoW and neural network analyses. Considering a wider variety of models will provide more detailed insight into which computational methods are effective for different use cases.

References

Steven E Clayman, Marc N Elliott, John Heritage, and Laurie L McDonald. 2006. Historical trends in questioning presidents, 1953-2000. *Presidential Studies Quarterly*, 36(4):561–583.

Lindy Comstock. 2023. Journalistic practice in the international press corps: Adversarial questioning of the russian president. *Journal of Language Aggression and Conflict*, 11(2):145–175.

Lindy B Comstock. 2015. Facilitating active engagement in intercultural teleconferences: A pragmalinguistic study of russian and irish participation frameworks. *Intercultural pragmatics*, 12(4):481–514.

Dhvani Kansara and Vinaya Sawant. 2020. Comparison of traditional machine learning and deep learning approaches for sentiment analysis. In *Advanced Computing Technologies and Applications: Proceedings of 2nd International Conference on Advanced Computing Technologies and Applications—ICACTA 2020*, pages 365–377. Springer.

Jenny A Thomas. 2014. *Meaning in interaction: An introduction to pragmatics*. Routledge.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.