Automated Emotion Detection in State of the Union Speeches

Arjun Nair
University of Illinois at Urbana-Champaign
arjunvn2@illinois.edu

Abstract

Computational emotion detection is an important next step in improving artificial intelligence and data analytics systems. While great advances have been made in recent years regarding emotion detection on various domains of interest, especially tweets, very little work has been done on emotion detection in political speeches. We present the first publicly available emotion-annotated corpus of sentences in political speeches, constructed using a Best-Worst Scaling methodology in order to ensure that ratings can be compared across annotators. We then test a variety of state-of-the-art emotion detection systems, such as IBM Watson, IMS EmoInt, and SeerNet, on our corpus and show that they do not perform nearly as well as they do on other domains, highlighting the need for domain-specific emotion detection systems in the future. Finally, we perform a qualitative analysis of various sentences that the emotion detection systems failed on in order to uncover the grammatical and lexical factors that are contributing to the inability of these systems to adapt to this new domain.

1 Introduction

Emotion detection is a problem in machine learning and affective computing that has been studied largely from the lens of sentiment analysis, the classification of text as having either positive or negative polarity. Only recently have researchers begun to design machine learning models that incorporate a more nuanced definition of emotion that goes beyond sentiment and measures other dimensions, such as joy, sadness, and anger (Mohammad and Bravo-Marquez, 2017; Wang et al., 2016). This allows emotion detection systems to distinguish between a number of affective states and predict the presence of secondary emotions, such as contempt or jealousy, by modeling them as combinations of basic emotional intensities (Gu et al., 2019).

Because emotion detection, beyond sentiment analysis, is still a very new research area, there are several domains to which it has yet to be fully applied. Political speeches are one such example; despite the wide array of historical and contemporary political speech corpora available to the public, it does not seem that the machine learning research community is moving to develop emotion recognition systems that are specifically tailored towards this domain. The only previous work, to the best of our knowledge, that focuses on the application of emotion detection systems specifically to the domain of political speeches only measures sentiment and ignores other emotional dimensions (Rheault et al., 2016).

Although researchers have access to political speech transcripts as a matter of public record, the task of manually annotating these speeches has prevented the ML community from developing models specifically suited for this domain. A previous study attempted to bypass this manual annotation step by building a corpus of political speeches annotated by audience applause and laughter (Guerini et al., 2013); however, the reactions of a biased political audience can tell a machine learning model very little about the actual content of a speech, emotional or otherwise.

In this paper, we present the first publicly available gold standard of emotional content in political speeches, annotated across six dimensions: sentiment, joy, anger, sadness, fear, and disgust. The purpose of this corpus is twofold. First, it provides a proof of concept for the use of Best-Worst Scaling, a technique developed by Louviere et al. (2015), in constructing emotional intensity and sentiment annotations for sentences in political speeches. Second, by comparing the human-made annotations in this corpus against predictions made

by generalized emotion detection systems, such as IBM Watson, it showcases the need for developing domain-specific emotion detection systems in the future. We urge the readers to not use this corpus to directly train machine learning models, as the sample size is too small (128 sentences) and only one human annotator was used. However, as shown by Kiritchenko and Mohammad (2017), the Best-Worst Scaling methodology we use to contruct the corpus is able to reliably compare the emotional content of different sentences within the corpus itself; thus, if more annotators are recruited in the future to construct a larger corpus, we may be able use this new dataset to train domain-specific machine learning models that significantly outperform existing state-of-the-art emotion detection systems when tested on political speeches.

2 Corpus Design and Methodology

This corpus was initially designed as part of a correlational study between the emotions found in State of the Union speeches and the general happiness of the American population. When it was later discovered, in the process of analyzing the corpus, that systems such as IBM Watson and IMS EmoInt performed poorly when predicting the emotional ratings of sentences in this domain, we pivoted towards an analysis of how current state-of-the-art emotion detection systems perform when tested on political speeches. Due to the original goal of the research project, the corpus is drawn from a total of 32 State of the Union speeches, one for each year preceding the General Social Survey, which collects several demographic metrics including happiness. Four sentences were randomly drawn from each State of the Union speech to make up a total of one hundred and twenty eight sentences in the corpus.

Each sentence had to be annotated across six emotion dimensions - sentiment, joy, anger, sadness, fear, and disgust; additionally, we wanted to ensure high comparability between human and machine annotation ratings and thus made use of the Best-Worst Scaling technique to create real-valued scores for each dimension from [-1, 1] such that each item's score on each dimension was determined based on its ranking relative to other items on the same dimension (Louviere et al., 2015). A trained human annotator was tasked with ranking 256 4-tuples of sentences from the corpus on each of the six dimensions, producing a total of 1536 annotations. These 4-tuple rankings were then converted to real-valued scores for each item on each each dimension using a Best-Worst Scaling conversion script developed by Kiritchenko and Mohammad (2016).

Next, we selected a set of three state-of-the-art emotion detection systems to serve as machine annotators. We will briefly describe each system.

IBM Watson NLU: IBM Watson, in its current form, is a set of natural language processing tools with a wide range of applications from business analytics to technical support. We will be focusing on IBM Watson Natural Language Understanding (NLU), a tool which, among other functionalities, can predict the sentiment, joy, anger, sadness, disgust, and fear in a sentence on a real-valued scale from [-1, 1] for sentiment and [0, 1] for emotional intensity.

IMS EmoInt: IMS EmoInt is an emotion detection system that emerged from the WASSA-2017 shared task on emotion intensity, winning second place in the official competition (Mohammad and Bravo-Marquez, 2017). It predicts anger, fear, disgust, and joy on a scale of [0,1] and uses a CNN-LTSM regression model in addition to various lexical featuresets in order to form predictions (Köper et al., 2017).

SeerNet: SeerNet is another emotion detection system from the WASSA-2017 shared task on emotion intensity, winning third place in the official competition. Like IMS EmoInt, it predicts anger, fear, disgust, and joy on a scale of [0,1]; however, it does so by taking the best-performing systems trained on a combination of lexical, syntactic, and word embedding features and combining them all together to form an ensemble model that performs better than each of them individually (Duppada and Hiray, 2017).

Due to the use of Best-Worst Scaling, the internal rating scale of each individual emotion detection system does not matter as long as each item's rating can be compared across a single system. We repeated the annotation process once for each emotion detection system, using it as a machine annotator that would make predictions for each item in a given 4-tuple and assign "best" and "worst" ratings to the items with the highest and lowest scores

	Cohen's Kappa	Average Agreement	Pearson's r	Assessment
IBM Watson				
Valence	0.232	0.492	0.440	fair
Sadness	-0.025	0.328	0.0230	no agreement
Joy	0.436	0.625	0.755	moderate
Disgust	0.294	0.531	0.607	fair
Fear	0.283	0.523	0.616	fair
Anger	0.238	0.500	0.405	fair
IMS EmoInt				
Sadness	0.019	0.352	-0.028	none to slight
Joy	0.306	0.539	0.699	fair
Fear	0.213	0.477	0.431	fair
Anger	-0.058	0.297	-0.149	no agreement
SeerNet				
Sadness	-0.043	0.313	-0.089	no agreement
Joy	0.305	0.539	0.564	fair
Fear	0.164	0.445	0.378	none to slight
Anger	-0.013	0.336	-0.002	no agreement

Table 1: Agreement and correlation metrics comparing machine predictions with human annotations. Assessments are based on Cohen (1960)'s proposed interpretation of kappa.

respectively on each dimension. It is important to note that IMS EmoInt and SeerNet do not measure sentiment and disgust, so they could not perform annotations for those two dimensions. These Best-Worst machine ratings were finally converted into real-valued scores that could be compared across human and machine annotators. These scores were then categorized into three intervals: positive for scores in the range (0.33, 1], neutral for scores in the range (-0.33, 0.33], and negative for scores in the range [-1, -0.33]. Tests of agreement were performed between the human annotator's categorized scores and those of each machine annotator in order to determine the validity of each system when tested on the domain of political speeches. Additionally, linear regression analyses were performed to calculate the Pearson correlation coefficient between the (uncategorized) scores of the human annotator and those of each machine annotator.

3 Results

Overall, the three state-of-the-art emotion detection systems tested in this paper performed very poorly on the domain of political speeches. Not a single system achieved a substantial amount of agreement with the human annotator, as defined by Cohen (1960), on any emotional dimension.

3.1 IBM Watson NLU

IBM Watson NLU performed best on the dimension of joy, in which it showed a moderate amount of agreement with the human annotator. It had a fair amount of agreement for sentiment, disgust, fear, and anger and had slight disagreement for sadness. Although the specifics of Watson's model are unknown to the broader research community, IBM claims, based on internal studies, that its emotion detection system achieves higher F1 scores than Wang and Pal (2015)'s model, which has average F1 scores of 0.63 on SemEval (Strapparava and Mihalcea, 2007) and 0.74 on ISEAR (Abdel Razek and Frasson, 2017). IBM Watson NLU did not have an average agreement greater than or equal to 0.63 on even a single emotional dimension when tested on this corpus, indicating that this emotion detection system may not be well-suited for this domain.

3.2 IMS EmoInt

Overall, IMS EmoInt performed worse than IBM Watson across the board, with lower agreement on every dimension except sadness, which had a marginally higher Cohen's kappa. There was fair agreement for both joy and fear, slight disagreement on anger, and very little, if any, agreement on sadness. When tested against the Tweet Emotion

Intensity Dataset, an annotated corpus of tweets, IMS EmoInt achieved an average Pearson correlation of r = 0.722 against the gold standard (Mohammad and Bravo-Marquez, 2017); in contrast, when tested against our corpus of political speeches, its correlation with the human annotator was only r = 0.238. It is clear from this analysis that the effectiveness of IMS EmoInt in predicting the emotions of tweets does not carry over to the domain of political speeches, at least when it comes to those made at the State of the Union.

3.3 SeerNet

SeerNet performed about as poorly as IMS EmoInt, with slight disagreement on sadness and anger, little to no agreement on fear, and only a fair amount of agreement on joy. As with IMS EmoInt, there was a severe drop in Pearson's r when switching from the Twitter Emotion Intensity Dataset to our corpus, from r = 0.708 down to r = 0.213.

4 Discussion

Not a single emotion detection system tested in this experiment achieved satisfactory performance on our corpus. In this section, we perform a qualitative analysis of cases where the human annotator and one or more of the machine annotators disagreed in order to determine what factors may be causing emotion detection systems to fail in the domain of political speeches.

Take, for example, this sentence from President Trump's 2017 State of the Union Address:

American footprints on distant worlds are not too big a dream.

The human annotator, through a series of Best-Worst rankings, rated this sentence as extremely positive, with an overall sentiment score of 0.875 on a scale of [-1, 1]; additionally, they rated the sentence as having a somewhat high amount of joy and little to no sadness, fear, disgust, or anger. IBM Watson, on the other hand, rated this sentence as moderately negative, although it predicted the emotional intensities found in the text with relatively high accuracy.

In the historical and geopolitical context of American space exploration, it is easy to see why the human annotator rated this sentence as highly positive. The imagery of footprints on distant worlds harkens back to the Moon landing, considered one of America's greatest accomplishments, and the sentence goes on to hint at the possibility of even greater feats occurring in the future. To an audience of American citizens, this is an explicitly positive statement; however, to a machine that relies largely on lexicon-based methodology, *footprints* and *distant worlds* mean very little out of context. The word *dream* might have potentially added some positivity to the sentence; however, the negation *not* that came shortly before it may have confused the model. An effective domain-specific emotion recognition system for political speeches will need to consider a different set of evocative terminology than one that deals with other domains; the words *American*, *footprints*, and *dream* are all examples of terms that make a greater emotional contribution than usual when used in political speeches.

Next, consider this sentence from President Ford's 1976 State of the Union Address, in which he quotes former President Eisenhower:

"America is not good because it is great," the President said.

The human annotator disagreed strongly with IBM Watson, IMS EmoInt, and SeerNet on every single emotional dimension of this sentence. The machine annotators likely took the phrase *is not good* out of context, which explains the strong negative sentiment and intense negative emotions that they predicted for this sentence. The human annotator, however, took a more holistic view of the sentence, understanding that *is not good* was simply a precursor to strengthen the meaning of *it is great*, and thus rated the sentence as expressing positive sentiment, high joy, and low anger, disgust, fear, and sadness. In the future, emotion detection systems of all types, not just domain-specific ones, will need to develop more linguistic maturity so that they can untangle the wordplay found in complicated sentences such as this one.

Finally, take this sentence from President Carter's 1979 State of the Union Address:

Real per capita income and real business profits have risen substantially in the last 2 years.

The human annotator, knowing the definitions of *real per capita income* and *real business profits*, rated this sentence as extremely positive and moderately joyful. While the machine annotators were all able to detect some small amount of joy in the

sentence, they unanimously concluded that there were extremely high levels of negative emotions present in this sentence as well, and IBM Watson interpreted the overall sentiment as negative.

Understanding that an increase in income and business profits is positive requires both the ability to interpret the term *risen* as an amplifier for *real per capita income* and *real business profits* and an understanding that income and profits are regarded as a positive in the context of political speeches. Thus, the failure of machine annotators to detect the emotions present in this sentence speaks to the need for both domain-specific lexica and stronger linguistic feature analysis capabilities in future emotion detection systems.

5 Conclusion

Measures of agreement and correlation between the human and machine annotators reveal that there is a large drop in performance when generalized emotion detection systems are applied to the domain of political speeches. Further qualitative analysis of various sentences that were disagreed upon reveal that the domain-specific terminology and linguistic complexity found in political speeches present great barriers to emotion detection systems. In the future, we recommend constructing lexica and corpora specifically designed for political speeches and then using those resources to train domain-specific emotion detection systems in the future. In this manner, researchers can develop reliable models for analyzing the emotions found in political speeches and can then apply these emotion detection systems towards conducting automated analyses of the vast political and historical speech corpora available to the public.

Acknowledgments

I would like to thank Dr. Roxana Girju for advising me on this project and putting together a relevant set of papers that acquainted me with the field. I would also like to thank Dr. Svetlana Kiritchenko and Dr. Saif Mohammad from the National Research Council of Canada for developing a set of Best-Worst Scaling scripts that were used in this project.

References

Mohammed Abdel Razek and Claude Frasson. 2017. Text-based intelligent learning emotion system. *Journal of Intelligent Learning Systems and Applications*, 09:17–20.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Venkatesh Duppada and Sushant Hiray. 2017. Seernet at emoint-2017: Tweet emotion intensity estimator. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 205–211.

Simeng Gu, Fushun Wang, Nitesh P. Patel, James A. Bourgeois, and Jason H. Huang. 2019. A model for basic emotions using observations of behavior in drosophila. *Frontiers in Psychology*, 10:781.

Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava. 2013. *The New Release of CORPS: A Corpus of Political Speeches Annotated with Audience Reactions*, volume 7688, pages 86–98.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California.

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. IMS at EmoInt-2017: Emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57, Copenhagen, Denmark. Association for Computational Linguistics.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. *CoRR*, abs/1708.03700.

L. Rheault, K. Beelen, C. Cochrane, and Graeme Hirst. 2016. Measuring emotion in parliamentary debates with automated textual analysis. *PLoS ONE*, 11.

C. Strapparava and R. Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *SemEval@ACL*.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 225–230, Berlin, Germany. Association for Computational Linguistics.

Y. Wang and Aditya Pal. 2015. Detecting emotions in social media: A constrained optimization approach. In *IJCAI*.