President Botrick: An Analysis of Deep Learning-Based Conversational AI Models to Identify and Create Influential Political Speeches

Ada Defne Tur[1], **Julia Hirschberg**[2]

[1]McGill University, Montreal, QC, Canada
[2]Columbia University, New York, NY, USA

Abstract

This paper explores the defining qualities of language that are considered influential and charismatic in the context of political speech. Transformer-based models have shown to be efficient in analyzing contextual clues and generating coherent texts in a variety of domains. With limited research in the identification and exploration of the replication of persuasion in natural human language and generation of influential speech, we seek to analyze the aspects of public speech that are deemed persuasive and impactful, and generate text accordingly. We propose a two-part experiment: First, we train a BERT-based encoder to weigh segments of speech in order to predict its influence on an audience; second, we train a GPT-based decoder to use an established understanding of persuasion to generate new political speech. We show that, using these models, a speech can be created that mimics the natural language habits of prominent political figures.

Introduction

Artificial intelligence has been useful recently in identifying traits in speakers that characterize them as charismatic and compelling, and utilizing such technologies can prove to be essential in understanding how politicians assert their influence in their speeches. In this paper, we investigate the components of language that make it persuasive and influential to an audience; particularly, how speech can provoke positive audience reaction and engagement, and how this information can be used to identify and generate influential speech. As an initial step in this research, we are focusing solely upon influential speech as implicitly annotated by the applause of the audience in public political speech corpora and we are only considering natural language, ignoring acoustic signals which may be impactful as well.

While the definition and identification of charisma and persuasion have been studied earlier (c.f. Rosenberg and Hirschberg (Rosenberg and Hirschberg 2009) and Guerini *et al.* (Guerini, Özbal, and Strapparava 2015, among others)), the use of natural language processing in enhancing and improving political speech is minimal, as the primary goal of machine learning has been for data analyses on campaign audiences and trend predictions for related AI efforts for public political speeches[1]. However, recent advances enable deep learning to be used to identify aspects of political speech that can have a significant effect on an audience, and how politicians can make use of this technology to improve their speeches.

A great majority of research in this area pertains to predicting and identifying charisma. Primarily, the work of Hirschberg and Rosenberg (Rosenberg and Hirschberg 2009) in characterizing political speech and using statistical analyses on raters' impressions of it is an important advancement in how modeling components of natural language can be effective in determining its impact. Building on this work, Guerini's study (Guerini, Özbal, and Strapparava 2015) on the same corpora used in this study showed how, as computational models advance, methods of identifying persuasive components of language are reformed as well.

To tackle the task at hand, we used a combination of the transformer based BERT (Devlin et al. 2018) and GPT (Radford et al. 2018) models. The BERT-based neural model was utilized to take transcribed speech, generate utterance embeddings, and make predictions on if audience applause is present in each segment. The GPT model's purpose was to generate new segments of text with the goal of provoking audience applause in mind. Based on results of the combined BERT based neural model, the algorithm shows a profound understanding of why certain segments in political speeches are applauded more often than others, and therefore can predict which parts of speeches are more persuasive and impactful than others. Furthermore, upon analyzing generated text from the GPT model, we can also identify levels of coherence, eloquence, and sense in these texts. A high level overview of this work can be seen in Figure 1.

Our novel contributions in this work are the usage of transformer-based models to not only analyze influential speech, but to also imitate it when generating new influential text. We show that a BERT-based classifier can identify the utterances which are applauded with 83% accuracy. Then, a GPT-based generator was fine-tuned using applauded segments (not overlapping with the training set of the classifier). 54% of the synthetic speech generated by this model is classified as applause by the classifier - nearly double the

[1]https://www.csis.org/blogs/technology-policy-blog/evolving-role-artificial-intelligence-and-machine-learning-us-politics

PART 1 Classification	PART 2 Generation
• Identifying if a speech is impactful or charismatic • Bidirectional Encoder Representations from Transformers (BERT) used for vectorization (called embeddings) • Embeddings fed into an artificial neural network for binary classification	• Creating an influential speech based on findings of Part 1 • Fine-tune a Generative Pre-trained Transformer (GPT) model to generate text • Model will generate next word given previous words in order to have an 'impactful sentence' • Can generate from scratch or from seed words or sentences

Figure 1: A high level overview.

average applause rate in the data: 29%. Furthermore, we conducted a linguistic inquiry and word count (LIWC) analysis (Pennebaker, Francis, and Booth 2001) of the performance of the generative model alongside existing political speech using LIWC-22 (Boyd et al. 2022).

In the next sections, we provide information on related work. We then describe the data we use for political public speech in Section . We next present our approach using transformer-based models for classification in Section and generation in Section . Section provides experimental results before we conclude.

Related Work

Rosenberg and Hirschberg (Rosenberg and Hirschberg 2009) introduced the pioneering advancement in understanding and identifying charisma in text and speech through a statistical approach in charisma detection. They collected American political speech segments and asked raters to rate the segments on charisma and 26 additional speaker traits, finding that charismatic speakers used longer sentences, more first-person plural and third-person singular pronouns, more repetitions and complex words; acoustic-prosodic correlates of charismatic speech were higher in pitch, faster, and louder, with more variation in intensity. Using the Pearson's R test and ANOVA to analyze the correlation between charisma and natural language attributes, Rosenberg and Hirschberg modeled the parallels between perceived charisma and spoken and transcribed text. The results of the study showed that speeches that were tagged by annotators as charming, persuasive, and enthusiastic had the highest correlation with being considered charismatic. A more recent study from Hirschberg (Yang et al. 2020) analysed charisma in the context of a gender-balanced experimentation using a statistical approach. Studying the gender of a speaker, as well as the academic background, speech preference, and personal characteristics of raters, conclusions can be drawn about how charismatic someone can be perceived to be, separate from their actual speaking techniques.

Tsai's study (Tsai 2015) comparing how speech between TED speakers and professors are distinct was a relevant step forward in statistical analyses on prosody and spoken attributes. The study aimed to classify an audio segment as either a TED talk or a university lecture, and utilized the Adaboost classifier on three different categories of speech: pitch-related features, energy-related features, and a combination of the two. Results showed that TED speakers tended to deliver speeches with less silence, deeper voices, and more energy, while also concluding that speakers that prioritized the energy of their delivery had given more impactful talks.

More recently, Tanveer et al. (Tanveer et al. 2018) conducted an experiment on the trajectories of TED talks and how this affects audience ratings of the talk through an analysis of speech patterns and audience impact. Similar to Rosenberg and Hirschberg's study, this experiment used ratings and tags such as 'inspiring', 'fascinating', and 'informative', and analyzed the correlation between such ratings and the view count of each talk in their corpus. The experiment used a logistic regression classifier and a support vector machine (SVM). This study found that many statistically significant patterns existed in narrative trajectories, and this in turn effected how an audience perceived a talk.

Guerini *et al.* (Guerini, Özbal, and Strapparava 2015) introduced the CORPS corpus, covering thousands of political speeches from notable speakers, in addition to using an SVM classifier in order to predict an audience reaction from the data. Guerini's analysis also utilized various characteristics of language, including rhyme, alliteration, plosive, and homogenity, and obtained accuracy of about 74%. Gillick and Bamman (Gillick and Bamman 2018) employed the first neural methods for applause detection in CORPS using feed forward networks and long short-term memory (LSTM) networks. Note that this work was done before the era of foundational models such as BERT, and the networks were trained from scratch. In related research, Liu *et al.* (Liu et al. 2017) focused on applause in TED talks using rhetorical devices, such as linguistic style, emotional expression, phonetic structure, projection of names, gratitude expression, rhetorical questions, and applause-seeking expression. The study used C4.5, logistic regression, Naive Bayes, and SVM models and had results from 50.1% to 71.9%. In our work, we focus on detecting segments that can be applauded using the text of these speeches and training models that can generate language that will be applauded.

Data

We utilize a corpus of 3618 political speeches from the CORPS dataset (Guerini, Strapparava, and Stock 2008). [2], in which speeches are each transcribed from spoken events and tagged with audience reactions, such as applause, laughter, and comments. There are 118,909 segments in the dataset which we are using for our experiments. We collected these using on paragraphs from the transcribed speeches. Each segment can include one or more sentences, and we truncated segments to contain at most 200 tokens.

The entire corpus consists of speeches from the 20^{th} and 21^{st} century, with over 100 different speakers, primarily previous United States presidents and other political figures. Each speech transcription also provides information about the speech in a brief description, including the event the speech was given in, the date, the speaker, the title, and the

[2]This corpus is available for researchers at the following link: http://hlt-nlp.fbk.eu/corps

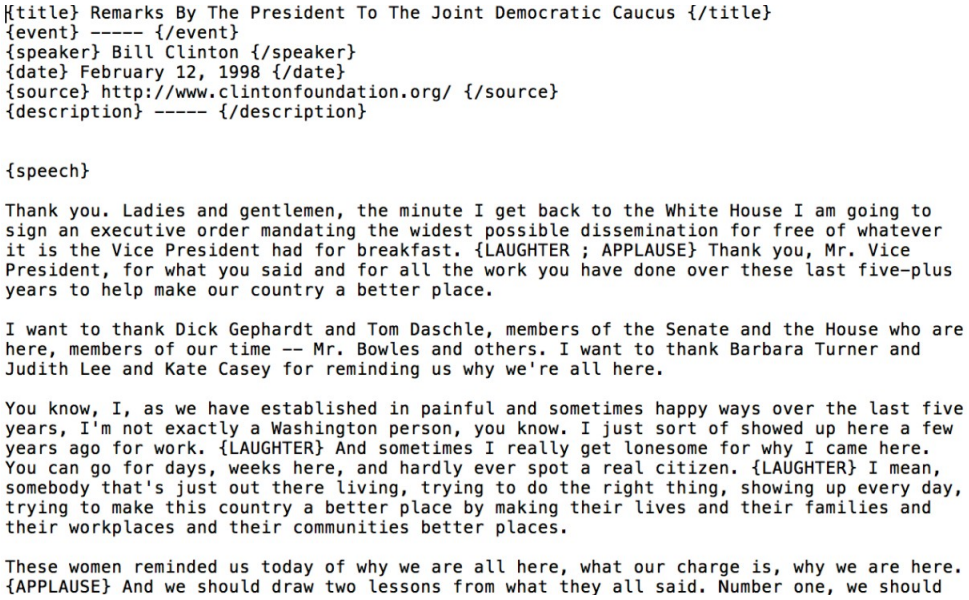

Figure 2: Example speech from the CORPS dataset

source of the transcript, as shown in Figure 2. Each speech, on average, contains 13 instances of applause, and 29% of the segments in the data were applauded. This dataset is compatible with our intended goal of building a political speech model using a variety of influential figures, along with thousands of examples of segments that provoked audience interaction; we will also use it to generate political text from our analysis of the data. Although applause, laughter, and audience comments are not the sole indicators of persuasion and impact, a stable generative model can be developed using this information.

We first divided the data in half: one half is reserved for applause classification and the second half for generation. For the classification portion, we divided the dataset such that the testing and validation sets consisted of 5000 segments each, and all remaining segments were used for training.

Data pre-processing was conducted in order to prepare each speech for our experiment, through the removal of extraneous details and by categorizing data by audience reaction. For instance, details about event, date, speaker, and sourcing were not necessary for the task, so segments of each speech were extracted without this information. Additionally, all tags were removed, and segments with audience tags were marked separately from those without, such that training, evaluation, and testing data could be prepared.

Classification

The classification portion of our experiments used the BERT model (Devlin et al. 2018) as an encoder to assist in the prediction of persuasive sentences, in addition to an artificial neural network, using audience applause as the primary metric for influence. Based on the information we identify on influence from the classification task, we can create, evaluate and improve the generation portion of our experiments.

Our BERT models are trained in a self-supervised way, relying on masked language model objective, where the model is trying to predict a masked word in a sentence. BERT provides contextual embeddings for a given input text, and BERT-based classifiers are shown to be state-of-the-art for many encoder tasks and are now standard for text classification (Li et al. 2022, among others). During fine-tuning, on top of the publicly available BERT model, a linear layer for binary classification is added for the task of applause detection for each sentence using the pooled output on the [CLS] token, as shown in Figure 3. Uncased base BERT model is used from the Huggingface Transformers library[3]. As the optimizer we used AdamW for decoupled weight decay regularization (Loshchilov and Hutter 2017) with a learning rate of 5e-5 and epsilon of 0.1, using Cross-Entropy Loss and 10% dropout.

Generation

The second part of the study focuses on generative modeling. A GPT-2-Medium model from the Huggingface Transformers library[4] was fine-tuned as the decoder to generate entirely new text to be used based on our understanding of influence from the models we created for classification. The goal of the generation task was to examine the ability of the model to replicate a political speech, given thousands of example speeches from prior presidents, leaders, and speakers. A generative pre-trained transformer model, or GPT-2, was the key for generation, as the model's architecture deemed it ideal for the task. The GPT-2 model is an autoregressive model that utilizes self-attention, such that the attention layer only attends to earlier positions in an output

[3]https://huggingface.co/bert-base-uncased

[4]https://huggingface.co/gpt2-medium

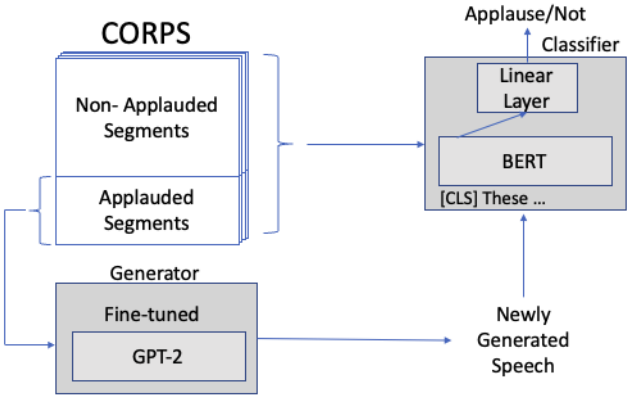

Figure 3: Model architecture.

sequence using the causal language model objective (in contrast to masked language model objective that BERT uses).

During the generation process, nucleus sampling (Holtzman et al. 2019) (with a threshold probability of 0.9) was employed with beam search, again using the Huggingface Transformers library[5]. We also investigated using top-K sampling, but found that, especially with new topics having unknown words, the outputs were sub-optimal. The model we used is fine-tuned for 10 epochs, again using AdamW as the optimizer, with 1e-4 as the learning rate and epsilon of 1e-8. Figure 3 shows the overall architecture in which the generated speech is fed into the classifier model for objective evaluation.

Experimental Results and Analysis

The first portion of the experiment achieved an accuracy of over 83% on our test set, showing that the model had developed an understanding of the patterns in language that result in engagement and persuasion and their link to audience applause. Similarly, a previous experiment, which consisted of a support-vector machine-based approach on a comparable task using the same CORPS dataset, achieved an accuracy between 73% to 74%, depending on a variety of phonetic features analyzed, including rhyme, plosive, homogeneity, and alliteration (Guerini, Özbal, and Strapparava 2015) However, a different aspect of this work was the use of 10-fold cross validation conducted on four primary datasets, whereas the current study used a fixed split for the data.

A common mistake that the classification model tends to make involves the presence of repetitive devices in speech. Below is an example of a segment that the model predicted would receive audience applause, which it did not (perhaps because of the repeated words, "*but*" and "*let*") and another segment which the model believed would receive no applause, although it actually did.

- **False Alarm:** "*... The country needs a surgeon general. Thanks to Senator Kennedy, the chairman of the committee, they went back and had the hearing. He told them they were going to stay there till kingdom come till they finished. But if somebody wants to vote against her, let them vote. But let's get on with it.*"
- **Miss:** "*... We have put forward a program which will open the doors of college education to all Americans, just like I promised in the campaign - lower interest loans; better repayment terms; and giving tens of thousands of Americans a chance to pay their college loans by serving their communities here at home, by working to make their communities a better place.*"

[5]https://huggingface.co/docs/transformers/model_doc/gpt2

	Applause-Based Model	Non-Applause-Based Model
No Context	54%	11%
Frequently Seen Context	47%	14%
Rarely Seen Context	41%	7%

Table 1: Results from the applause and non-applause models and various forms of context.

Influential Language Generation

The second part of our work studies how to generate speeches from fine-tuned generation models; we call our generative model "President Botrick". We performed our evaluation of these "Botrick" speeches using two methods, manual and automated. In the automated method, we used the classifier trained in the first part of the project to evaluate the generated segments to provide a more objective evaluation, free from human bias (e.g., political beliefs). It should be noted that 29% of the training set segments are applauded. However, 300 generated segments from President Botrick, when sent through the same process, have 54% applause rate. Therefore, the understanding of influence we developed had a noticeable effect on the speeches the model was able to generate. To further confirm this, we fine-tuned a second generator using the *non*-applauded segments. The ratio of generated segments classified as applause dropped to 11% using this model, verifying the power of the applause model. We have also run a longest-common-substring algorithm on the generated segments and found that the average maximum overlap of a generated segment with an existing training data segment is 7.2%, showing that our models generated novel segments, rather than copying the training data. In another experiment, we seeded the models with 2 sets of phrases, one with frequently occurring phrases, such as "*this election*" or "*our party*, and another with rarely or never occurring phrases such as "*penguin migrations*" or "*white bear*". We provided the seed phrases to the model as segments to complete. For these newly generated segments, we found that, while the ratio of segments which are classified to be applauded reduced to 47% and 41% for these two sets of phrases, as shown in Table 1, these ratios are still significantly above the chance ratio.

As a second method for evaluation, we manually annotated pairs of segments generated by the applause and non-applause models using the rarely-occurring phrases. An annotator was shown 50 pairs of segments, where each pair

	Average General	Average for Speeches	CORPS Applauded Segments	CORPS Non-Applauded Segments	Generated Segments
Affect	4.96	4.27	5.75	4.91	5.83
Authentic	55.84	29.97	38.72	47.41	17.13
Clout	33.95	55.10	85.59	80.17	72.36
Conflict	0.06	0.34	0.34	0.29	0.45
Culture	0.37	3.06	3.34	2.61	3.02
Emotion	1.16	0.85	1.28	1.10	1.85
Future Focus	0.84	1.52	2.01	1.92	1.65
Past Focus	4.85	2.82	3.23	4.12	2.64
I-words (I, Me, My)	5.93	2.25	2.63	2.37	3.36
Polite	0.12	0.43	0.91	0.37	0.13
Politics	0.16	2.55	2.87	2.06	2.35
Power	1.24	3.21	3.05	2.47	2.83
Positive Tone	4.13	3.01	4.55	3.69	4.50
We	0.31	1.95	3.16	3.07	2.87

Table 2: Results of the LIWC comparison between a general average for all text, a general average for speech, the average for applauded segments of the CORPS data, the average for the non-applauded segments of the CORPS data, and the average for generated lines from the model.

had one segment generated using the applause model and one generated using the non-applause model. The human annotator was asked to select the segment that they believed would have invoked applause from the audience. The human annotation accuracy, that is the ratio of times the human annotator chose the segment generated by the applause model, was found to be 66%, higher than the chance probability of 50%.

Analysis of Generated Speeches The generated speeches displayed good coherence and sense such that the imitation of notable political figures is noticeable. Here is a representative example of a generated segment[6]: "*You can't just tell people that we're going to ignore them. You can't just tell people that we won't care whether they vote. You can't just tell them that we'll ignore them. If they vote, the politics will work. And that's the point.*".

Anaphora is a strategy used in many impactful speeches, such as Martin Luther King Jr.'s 'I Have A Dream': "*So let freedom ring from the prodigious hilltops of New Hampshire. Let freedom ring from the mighty mountains of New York. Let freedom ring from the heightening Alleghenies of Pennsylvania...*" The anaphora in the model-generated example above, relying on the long phrase "*you can't just tell people that*", actually does not appear in the training data as such. The closest segment is from an Obama speech from 2010 which says "*And we can't just declare that our kids need to get more exercise when they don't have parks to play in or safe streets to walk on. APPLAUSE We can't just tell folks to put more fruits and vegetables on the dinner table when many a family lives miles from the nearest grocery store. APPLAUSE.*".

Manual checking of the generated segments shows that they are composed of phrases from the applauded and non-applauded portions of the data put together into semantically coherent sentences and are characteristically representative of both portions of the data. For example, the applause model consistently generates segments with repetitions and motivational endings, such as "*We will prosper, we will win*", whereas the non-applause segments are usually declarative and dull, such as "*We can only assume that our children are taught the meaning of freedom.*". Below, we provide examples of generated segments using frequent and infrequent contexts with the models trained with applause and non-applause portions of the data.

Frequent context, applause: "*The world has suffered enough of these fanatics, and the fanatics who still try to kill innocent people who would hide in a cave they could not hide from the United States and the coalition forces.*" **Infrequent context, applause:** "*Urban planning strategies and urban policy are also crucial for the long-term sustainability of our economy. We're reinvesting in our people and our infrastructure, especially in upgrading our nation's airports. We're making unprecedented progress on the economy.*"

Frequent context, non-applause: "*The United States must lead the fight for freedom and security for our own people. We must remind ourselves that freedom and security depend on the advance of human ideals and human dignity.*"

Infrequent context, non-applause: "*Homeless populations are kind of an invisible and hated curse that we refuse to heal, and that, frankly, seems not possible in a free country. But in a free country, too, it's easy to feel the doom and despair and hopelessness and anger when there's a surplus of hope and virtue that cannot be realized without that. We're eager for that prosperity, that I believe.*"

As a final test, we also analyzed a set of relevant attributes of a variety of texts using Linguistic Inquiry and Word Count (LIWC) features (Boyd et al. 2022)[7]. In Table 2, we draw a comparison between the applauded and non-applauded segments of the CORPS dataset, the generated lines from our model, and general averages for speeches and text, which LIWC also provides. These show great variation in a variety

[6]It should be noted that the sentiments and beliefs of all generated speeches are not representative of the beliefs of the authors.

[7]Available for researchers at https://www.liwc.app/

of features between the applauded and the non-applauded texts. Based on the relevant features, the generative model shows a high level of representation for political speeches. For instance, the generative model demonstrates a higher level of power and clout, or influence in politics than general text and speech. The same high level of power and clout can be seen from the LIWC information on the applauded segments of the CORPS dataset itself. This shows how our generative model can be closely compared with existing political speech. Furthermore, the presence of a number of I and we words, like "I", "me", "my", "we", and "us", exhibit the social aspect of political speech and the communication between a speaker and an audience. A relevant characteristic for most persuasive speeches is the presence of cultural contexts, a characteristic that is shown to be higher in the examined political texts and in the output of the generative model. Overall, because political speech tends to have a higher score in many of the features in 2, it is important for a generative model aiming to write persuasive speech to also score similarly, as "President Botrick" has shown to be capable of doing.

Conclusions and Future Work

The endless capabilities of transformer-based models present a vast future for natural language research, particularly in the understanding and imitation of human speech using deep learning. Our generated texts demonstrate a promising future for artificial intelligence to be utilized in a variety of fields, including politics. Beyond simple data analysis conducted by presidential figures, the presented technology can be useful for the organization of a candidacy and the writing of influential speeches. For our future work on persuasive speech, we plan more work with LIWC features for analysis and classification. Also, since our current models are trained using text alone, we plan more research using acoustic-prosodic features of political speeches as well. Human language not only consists of written and transcribed speech; human beings speak with their prosody, facial expression, and body gesture. Thus, persuasion is not only spoken, but also displayed in speech and visual features. Our next steps for "President Botrick" will be analyses of the acoustic-prosodic features of political speeches (pitch, energy, speaking rate and voice quality), and of the speaker's facial expressions, their body language and their body gestures. Given the success of deep learning in the field of speech and natural language understanding and generation, we believe these additions should be extremely useful.

Ethical Considerations

Despite the positive performance of such models, there are risks associated with the freedom to create any text a model wants, particularly with the potential for language models to hallucinate without safeguards. Furthermore, depending on the influences provided to such a model prior to training and generation, language produced can be morally incorrect and/or offensive, calling into account the ethics of such language generation systems.

References

Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gillick, J.; and Bamman, D. 2018. Please clap: Modeling applause in campaign speeches. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 92–102.

Guerini, M.; Özbal, G.; and Strapparava, C. 2015. Echoes of persuasion: The effect of euphony in persuasive communication. *arXiv preprint arXiv:1508.05817*.

Guerini, M.; Strapparava, C.; and Stock, O. 2008. Corps: A corpus of tagged political speeches for persuasive communication processing. *Journal of Information Technology & Politics*, 5(1): 19–32.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P. S.; and He, L. 2022. A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2): 1–41.

Liu, Z.; Xu, A.; Zhang, M.; Mahmud, J.; and Sinha, V. 2017. Fostering user engagement: Rhetorical devices for applause generation learnt from ted talks. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001): 2001.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training (2018).

Rosenberg, A.; and Hirschberg, J. 2009. Charisma perception from text and speech. *Speech Communication*, 51(7): 640–655.

Tanveer, M. I.; Samrose, S.; Baten, R. A.; and Hoque, M. E. 2018. Awe the audience: How the narrative trajectories affect audience perception in public speaking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Tsai, T. 2015. Are You TED Talk material? Comparing prosody in professors and TED speakers. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Yang, Z.; Huynh, J.; Tabata, R.; Cestero, N.; Aharoni, T.; and Hirschberg, J. 2020. What makes a speaker charismatic? producing and perceiving charismatic speech. In *Proc. 10th International Conference on Speech Prosody*, volume 2020, 685–689.