Classifying Audience Response on Political Speech

Ankita Naik arnaik

Apurva Swarnakar aswarnakar

Parag Pachpute ppachpute

Kartik Mittal kartikmittal

Chirag Goel cgoel

1 Problem statement

In 2018, a set of new state-of-the-art results were established for a variety of Natural Language Processing tasks, the majority of which can be attributed to the introduction of context aware token representations, learned from large amounts of data with Language-modeling.

It is, however, unclear to what degree the computed representations capture and encode high level syntactic or linguistic knowledge about the usage of a given token in a sentence. One way of exploring the potential of the learned representation would be through investigating the performance on a task that would require the representation to acquire some notion of linguistic units such as phrases and clauses, as well as the relationship between the linguistic units and other tokens in the model. An example of such a task is Audience Reaction to Political speeches Detection.

We aim towards building a model that generates sentence-level classification of audience reaction to transcripts of speeches. The purpose of building this model is to help humans evaluate their textual content to better evaluate the audience reaction. After evaluating the existing standard pipelines over the CORPS dataset provided by (Guerini et al., 2013), we used more cognitive features of linguistics such as humour and sarcasm in-conjunction with our best baseline model to improve overall accuracy.We also checked if BERT (Devlin et al., 2018) captures humor and sarcasm features and is there any positive/ negative effect on applause propensity by presence of humour/ sarcasm in speech.

The main contributions of this work can be summarized as follows:

1. We achieved and reported improved performance for the applause detection on CORPS political speeches dataset
2. We investigated ways of incorporating additional linguistic features into the model and explore the potential improvements resulting from the addition of such features

2 What you proposed vs. what you accomplished

The project started off with a goal to predict sentence-level audience response to political speeches. The reason for choosing this was because not much work was done on the CORPS dataset (Guerini et al., 2013) using the advanced neural network models released recently. The plan was to first implement the existing baseline model of SVM and CNN and get the same accuracy as the published papers. Once the baseline is set up, apply new neural network-based models on that dataset and get the performance. Perform result and error analysis and improve the accuracy of the results by carefully evaluating the results. Later on, use the framework on transcripts of audio or video speeches to identify audience responses.

Milestone 1:

By the time of the first milestone, we were able to get the data from the creators of the dataset, preprocess it, and structure it so that a CSV dataset is generated to be used in any model. We then implemented two of the proposed baselines of SVM and CNN and achieved comparable results as that of the published paper (accuracy difference within 3%). We then took a pre-trained BERT model from Google and applied that on our dataset. Post preliminary analysis we came to know that the results are skewed because of class imbalance (75-25). We then updated our dataset and carefully partitioned the data so that the data distribution is 50-50. Post that we again retrieved results, this time the results dropped from 89% however, they were marginally better than the baseline implementations.

Post milestone:

Through subsequent result analysis, we came to know that the BERT in its word embeddings is not able to capture some of the important linguistic features. Hence, we shifted our approach to first validate this claim that ”BERT in its word embeddings does not contain enough information to get humor and sarcasm” and if the claim is true then add these features additionally and evaluate the results again. If the results are better then we can state our claim with even more confidence. Since this work was exciting and very little research is currently done in this area we skipped the part of doing applause detection on the transcript of audio and video data and focused on understanding BERT embeddings more.

3 Related work

In the context of detecting applause in campaign speeches (Gillick and Bamman, 2018) proved experimentally that lexical features carry the most information but a variety of features are predictive (i.e. they include prosody, long-term contextual dependencies) and are theoretically motivated, designed to capture rhetorical techniques. (Guerini et al., 2013) and (Liu et al., 2017) approach the task of applause prediction by looking at lexical features of individual sentences, that immediately precede audience applause. Models used by both the papers use a similar concept as proposed by (Danescu-Niculescu-Mizil et al., 2012), which formulates a dataset for binary classification. The used dataset contained sentences that were tagged for applause coupled with another sentence that did not lead to applause.

(Guerini et al., 2015) examines a set of features designed to capture aspects of euphony, or “the inherent pleasantness of the sounds of words” that might make an utterance memorable or persuasive such as rhyme, alliteration, homogeneity, and plosives. Using the CORPS dataset (Guerini et al., 2013), which consists of the text of several thousand political speeches dating from 1917 to 2011, the authors define persuasive sentences, which are ones that are preceded by annotations of either applause or laughter.

Another paper in the domain of predicting audience reaction but not directly related to our approach is by (Navarretta, 2017) that uses simple spoken sequences, speech pauses and co-speech gestures in an annotated video-and audio-recorded speech by Barack Obama at the Annual White House Correspondents’ Association Dinner to predict the audience response. It states that information about spoken sequences, pauses, and co-speech gestures can be used to predict the immediate response of the audience.

When we started exploring different models we came across (Zhang et al., 2016) which demonstrated that for the use case of sarcasm detection instead of defining features manually if we use deep neural network and let that figure out the features for itself then it provides better results. Hence, we took state-of-the-art model (Hazarika et al., 2018) on sarcasm detection on Reddit dataset (Khodak et al., 2017). Similarly, for humor detection we found (Chen and Soo, 2018) where deep learning facilitated improved performance in the humor detection task. Since that paper was mostly focused on getting a humor model to work in multiple languages rather than just focusing on English, we found out this model (de Oliveira and Rodrigo, 2017) to be more relevant to our use case.

4 Dataset

In the course of our complete project, we used three different datasets to validate our various analyses. Firstly for applause detection, the CORPS (Corpus of tagged Political Speeches) dataset, a political speech corpus developed by (Guerini et al., 2013) and has been annotated for audience reactions such as Applause or Laughter. The data is freely available for research purposes and can be obtained from (Guerini et al., 2013) in HTML format. Speeches, Author id and other attributes are parsed using HTML parsing techniques. Refer Table 1 for high level corpus statistics.

Example :

when john kerry suggests a global test, he goes right back to his beginnings in politics, when he said as he ran for congress the first time, he would only deploy troops under the authority of the united nations {**BOOING**}

during the 1980s, he opposed ronald reagan’s major defense initiatives that brought victory in the cold war.... {**NEUTRAL**}

you occasionally hear some bold talk from him, but it cannot disguise a 30-year record of coming down on the wrong side of virtually every major defense issue. {**APPLAUSE**}

For humor detection experiments, we used Google pre trained word2vec vectors, Yelp reviews dataset, funny one liner, proverbs, reuters and funny wikipedia headlines.

For sarcasm detection experiments, Reddit comments dataset (Khodak et al., 2017) was used.

4.1 Data preprocessing

A speech might contain multiple emotion tags such as Laughter; Applause, Standing Ovation etc. Since the distributions of the tags is very sparse, we formulated the complete problem as a binary classification problem. Tags except booing and neutral were mapped to class 1 and booing and neutral as to class 0 2.

Statistics	Count
Total number of speeches	3618
Total number of speakers	197
Total number of words	7,901,893
Total number of tags	66,082

Table 1: Corpus Main Statistics

We also experimented with various data preprocessing methods:
Strategy 1: Following the data preprocessing done in (Song et al., 2018), stop words and punctuation marks are removed and a context window of preceding 30 words was used to map the speech snippet to the tag. Remaining sections were again broken into group of 30 words and mapped to tag 'neutral'
Strategy 2: Since BERT has inbuilt data cleaning modules only context window was changed to capture better understanding while breaking up the speech on tags. A contextual window of preceding three sentences is used and remaining section is mapped to tag 'neutral'
Strategy 3: In order to further balance the class imbalance, a contextual window of preceding three sentences is used for both the classes and if the number of sentences for neutral class record is less than 3, then the record is dropped. Refer to 3 for final class distribution.

Tag	Count
Neutral	51,245
Applause	46,294
Laughter	14,050
other	3599

Table 2: Raw Speech Tags Statistics

Statistics	Count
Number of class 0 speech fragments	63,398
Number of class 1 speech fragments	51,991

Table 3: Class distribution after Strategy 3

5 Baselines

Based on our literature survey, we compared two widely used models for the purpose of applause response prediction and used them to set our project baseline. The results obtained are mentioned in Table 4.

Model	Precision	Recall	F1
SVM	59%	59%	59%
C-LSTM-CNN	63%	70%	66%

Table 4: Baseline Evaluation

5.1 PoS With SVM

As a first step to build our baseline, and based on the suggestions from the CORPS research (Guerini et al., 2013), we built our classifier using naive parts-of-speech tagging with support vector machines. The processed sentences from the dataset were tokenized using NLTK Word Tokenizer, and the tokens were tagged using NLTK Pos-tagger. The tagged tokens were then passed to a SVM classifier, to predict whether a given sentence would get a applause or not. We were able to achieve an F1 score of 59%, which shows that the SVM might be biased in classifying a given sentence, to a specific category. Since, the results obtained were better than the baseline paper, we did not perform hyperparameter tuning, but experimented with other state-of-the-art methods to improve the results.

5.2 C-LSTM-CNN

As an improvement to the vanilla SVM based classifier, we implemented a Context-LSTM-CNN (Song et al., 2018), which is based on the computationally efficient FOFE(Fixed Size Originally Forgetting) method (Zhang et al., 2015) and an architecture that combined an LSTM and CNN for the focus sentence. An LSTM layer was used to encode the focus sentences, the generated embeddigns were then passed to convolutional layers with small size kernels and max-pooling, and local features were extracted. Parallely, the full left and right contexts were encoded using FOFE, which was then passed through separate fully connected layer, concatenated with the previous generated word embeddings. The concatenated outputs were then used as inputs to the final sigmoid output layer. The model was trained using Adamax optimizer and it was able to achieve an accuracy of 62.23% on the test set.

The Pos with SVM model suggested by (Guerini et al., 2013) was taken as the baseline, and further enhancements were performed to improve the baseline accuracy.

6 Approach

Within the baseline models, we tried to find the caveats on the types of questions it was wrongly classifying. Initially, we looked at the confusion matrix and found it highly skewed providing a relatively high F1 score (Milestone 1). On further normalization of our data and evaluating the results (Confusion matrices indicated in Milestone 2), we realized that there a good number of humorous and sarcastic sentences that were not classified with applause. To be able to obtain these results, we would need to use a contextual model that can help us detect humor or sarcasm. On looking at the different linguistic features we decided to target our approach towards humor and sarcasm. The entire approach was divided into a four step process.

- Firstly, we decided to see how BERT performs as in when fine-tuned over the downstream task of applause detection.
- Secondly, we need to find the co-relation between applause and humour as well as applause and sarcasm to predict how our model could improve using these features.
- Thirdly, we need to perform experiments to understand whether our existing baseline model of BERT, already determines these values.
- Lastly, we needed to append the features to our existing baseline to observe how advantageous it was to explicitly include these additional features.

6.1 Infrastructure and Setup

The project was done on Google Cloud Platform using both Google Colab and Virtual Machine. Code was written over Tensorflow and developed by adapting for GPU and various copies of Colab notebooks were run in parallel to perform various experiments as described later in results. Most of the training was performed on a GCP VM Instance running n1-standard-8 (8 vCPUs, 30 GB memory) with an additional GPU of NVIDIA Tesla V100. The instance was running a deep-learning image adapted for tensorflow provided by GCP along with a 500GB persistent disk space to keep hold of the datasets. The following libraries were explicitly used to help in using BERT:

1. The BERT Uncased Base model provided at `https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1`.
2. BERT tensorflow pip package (`https://pypi.org/project/bert-tensorflow/`) that provides some helper functions to work along with the model.

6.2 Existing Work

To understand the impact of humour and sarcasm on applause detection, it was important to use one of the best models to help determine them. After performing analysis, it was decided to use CASCADE (Hazarika et al., 2018) for detecting sarcasm while using the Humor Detection in Yelp reviews (de Oliveira and Rodrigo, 2017), for humour detection. The source code was provided by the CASCADE team `https://github.com/SenticNet/CASCADE` and the pre-trained model for humour detection was provided at `https://github.com/cschen13/yelp-humor-detection`. Also, to help us get started with using BERT, we were able to reuse some work from Google Research [1].

6.2.1 Sarcasm Model

The model uses SARC dataset (Khodak et al., 2017) to obtain 77% accuracy. The following steps are performed:

1. As different people have different idiolects and authorship styles, we obtain stylometric information for each of the speakers by

[1] `https://colab.research.google.com/github/google-research/bert/blob/master/predicting_movie_reviews_with_bert_on_tf_hub.ipynb`

Total Applause	With Sarcasm and Humor	Only Sarcasm	Only Humor	Rest
63398	9423	5350	30833	17792

Table 5: Feature Dependency

creating *ParagraphVectors* (Le and Mikolov, 2014)

2. As a next step, using the pre-trained FastText embedding the users personality was scored across the five personality traits using the model from (Celli et al., 2013)

3. The models were fused using generalized canonical co-relation analysis (GCCA) by adapting the code provided by (Benton et al., 2016) and the result was a user embedding

4. Additionally, as sarcasm can depend on context, a similar approach of stylometric features were obtained on political subreddit

5. Lastly, these user embeddings and discource embedding along with the sarcasm annotated reddit data (Khodak et al., 2017) was passed through a CNN network to train a model which was used to annotate and generate the embeddings for our corpus.

6.2.2 Humour Model

To generate humour scores on our dataset, following steps are performed:

1. Pre-trained Word2vec model on google vectors is incrementally trained on Yelp reviews and other public available datasets. Using this word2vec model, embeddings for our vocabulary are extracted. If a word is not present in the word2vec model vocabulary then a random vector of length 100 with values in [-.25, .25] range is used. .25 is the variance of pre trained vectors. To calculate the embedding for a speech snippet, average of its all constituent word embeddings is used

2. Finally, the generated embeddings were used to train a CNN classifier based on the model proposed by (Kim, 2014).

6.3 The Approach

6.3.1 Raw BERT on Applause Detection

To evaluate and see how a contextual embedding of a advanced model like BERT, the BERT model provided by Google Research was fine-tuned on the downstream task of applause detection on the same CORPS dataset. It had a single fully-connected layer at the end to perform classification. The results obtained were very positive as the accuracy was 77.2%. This showed a clear advantage of using transfer learning from an advanced contextual model which can be fine-tuned for a downstream task.

6.3.2 Co-Relation Identification

To validate the fundamental assumption that sarcasm and humour will improve the accuracy, we first annotated our primary dataset CORPS (Guerini et al., 2013) using the humor and sarcasm model. The results obtained are shown in Table 5. We can a relatively strong co-relation between applause and humor, while a relatively small yet importance relationship between applause and sarcasm.

6.3.3 Experiments on Baseline

As the next step, we needed to run experiments on the baseline algorithm of using BERT for sarcasm and humor detection. We use a simple classification on BERT using our existing datasets for sarcasm and humor to observe the results. The flow is shown in Figure 1.

1. Trying to identify sarcasm, we see a 55.1% accuracy which is a similar probability to tossing a coin. Hence, it might be interesting to append these features to the our existing model to aid applause detection.

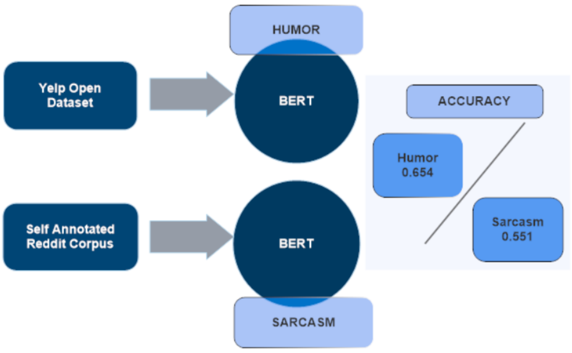
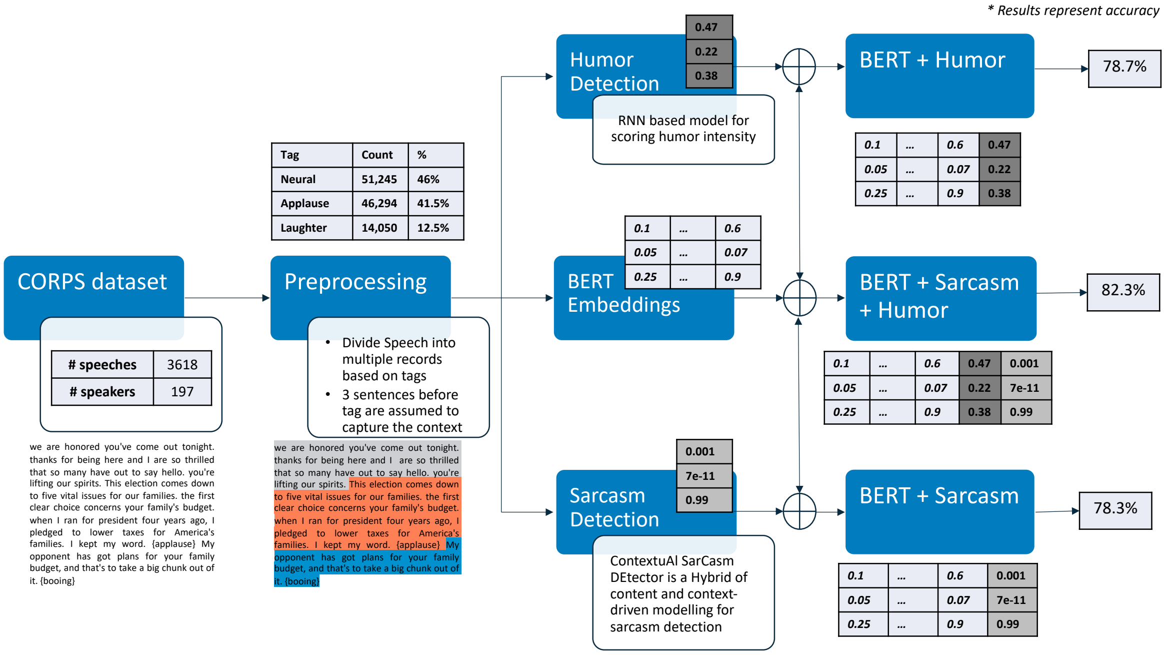


Figure 1: Linguistic Features experiment on BERT


* Results represent accuracy
CORPS dataset
Preprocessing
BERT Embeddings
Humor Detection
RNN based model for scoring humor intensity
Sarcasm Detection
ContextuAl SarCasm DEtector is a Hybrid of content and context-driven modelling for sarcasm detection
BERT + Humor
78.7%
BERT + Sarcasm + Humor
82.3%
BERT + Sarcasm
78.3%

# speeches	3618
# speakers	197

Tag	Count	%
Neural	51,245	46%
Applause	46,294	41.5%
Laughter	14,050	12.5%

- Divide Speech into multiple records based on tags
- 3 sentences before tag are assumed to capture the context

we are honored you've come out tonight. thanks for being here and I are so thrilled that so many have out to say hello. you're lifting our spirits. This election comes down to five vital issues for our families. the first clear choice concerns your family's budget. when I ran for president four years ago, I pledged to lower taxes for America's families. I kept my word. {applause} My opponent has got plans for your family budget, and that's to take a big chunk out of it. {booing}

we are honored you've come out tonight. thanks for being here and I are so thrilled that so many have out to say hello. you're lifting our spirits. This election comes down to five vital issues for our families. the first clear choice concerns your family's budget. when I ran for president four years ago, I pledged to lower taxes for America's families. I kept my word. {applause} My opponent has got plans for your family budget, and that's to take a big chunk out of it. {booing}

0.47	
0.22	
0.38	

0.1	...	0.6
0.05	...	0.07
0.25	...	0.9

0.001	
7e-11	
0.99	

0.1	...	0.6	0.47
0.05	...	0.07	0.22
0.25	...	0.9	0.38

0.1	...	0.6	0.47	0.001
0.05	...	0.07	0.22	7e-11
0.25	...	0.9	0.38	0.99

0.1	...	0.6	0.001
0.05	...	0.07	7e-11
0.25	...	0.9	0.99

Figure 2: Architecture Diagram

2. While running it for humor, we see a slightly better accuracy of 65.4% accuracy. As it's not exceptional, adding this feature additionally should improve the accuracy of our existing applause model.

6.3.4 Feature augmentation on baseline

Based on our findings, we decided to augment the features obtained from BERT with the sarcasm and humour annotations. We had three different options how we wanted to append these new features:

1. Append the class with to the embedding obtained from BERT. That is, if the sentence contains humour or not and the same with sarcasm.

2. Append the probability of sarcasm and humour to the embedding. Obtained by a simple softmax over the scores.

3. We also experimented adding only humour, only sarcasm along with adding both of them.

We performed all the three different variations and ran experiments across all the different variations to obtain results. The architecture can be seen in Figure-2.

6.4 Results & Analysis

The results were quite positive as we noticed a 5% improvement in F1 score after adding the two features as can be seen in Figure-3. Also, in Table-6 and Table-7 we can see all the populated results of the different experiments. The following points can be inferred:

1. Using context embedding such as available from BERT were very helpful in raising the baseline as compared to the existing CNN and SVM approaches that had been tried before.

2. These results help us understand that using BERT alone might not suffice to understand the complex linguistic features. If any of these features can aid the detection of the task at end, it might be wise to detect them via an alternate model.

3. The results also show that it was the best to append the scores instead of the class or the probabilities. This is interesting as it shows that the level of sarcasm across different statements also help the model classify correctly.

4. As expected, we also notice a much better performance while fine-tuning BERT towards a downstream task.

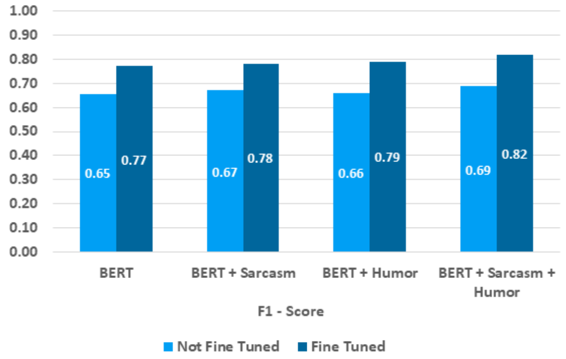


Figure 3: Comparison of Models

6.5 Code Correlation

1. Pre-processing: /Preprocessing/data_preprocessing.py contains all the pre-processing steps that were done to modify the provided dataset for our task.

2. Baselines: SVM and CNN can be seen in their respective folders. Only the code is attached as the checkpoint files were quite large.

3. Sarcasm: The work can be seen in the /sarcasm folder where the code was modified to perform all the different steps on CORPS dataset. The data was transformed to obtain the user embedding using /sarcasm/cascade/users/custom_data_to_csv.py and /sarcasm/cascade/users/custom/train_personality.py while the final training was /sarcasm/cascade/src/getSarcasm.py and /sarcasm/cascade/src/train_cascade.py

4. Humor: Most of the work can be seen in the file under /humor/humor.py which uses the pre-trained model to help annotate the data.

5. BERT: All the major experiments and work was done in BERT were modifications done to the file. Comments are written to explain the changes in code that were required to perform the experiments.

7 Error analysis

Indepth error analysis was carried out to validate the hypothesis and experiments carried out. Weherin both positive - negative examples and their humor sarcasm scores were considered to validate our analysis. Further, examples neither classified correctly by any of the models and data annotation specific issues were also analyzed. A brief summary of the same is presented in the following sections.

7.1 Baseline Error Analysis - SVM

Following are major shortcomings of the POS-tagger based SVM baseline :

1. **Correctly classified Example - POS Tags** : As the model was based on rudimentary features contextual learnings present in speech was not at all capture, Majority of the predictions done by SVM were based on the positive or negative nature of words.

   For example, SVM was able to correctly identify the tag for the following statement :

   ”for here we are inundated by sights and sounds that make it impossible for us to doubt that ideals can, indeed, be made into realities thank you, mr president thank you”

   The positive characteristics from the words thank you and emphasis from words like ’inundated’ was correctly classified as a Applause worthy response by SVM.

2. **Incorrect Examples - Contextual Learning** : More complex examples having contextual learning were not at all captured correctly by SVM. For example, the following sentence though having an positive tone due to appreciation coming from the prior part was wrongly classified.

   ”i want to thank all the other statehouse officials i know the speaker of the house is with us, speaker tom craddick and nadine”

3. Except for these all other major shortcomings from BERT described in the following section were also very clearly present in SVM. Thus, having a very low prediction accuracy.

7.2 Error Analysis - BERT

BERT could very easily capture context and solve major issues lacking in SVM. Thus, improving the performance of the model. Following are major issues encountered by BERT:

1. **Incorrect Examples - Linguistic Features** : Even though BERT could solve the incorrect

	Only Sarcasm	**Only Humor**	**Sarcasm and Humor**
Class	0.795	0.800	0.81
Probability	0.792	0.801	0.809

Table 6: Results as F1 scores when fine-tuned

	Only Sarcasm	**Only Humor**	**Sarcasm and Humor**
Class	0.65	0.651	0.0.66
Probability	0.62	0.66	0.661

Table 7: Results as F1 scores when not fine-tuned

example mentioned in 7.1 if failed at the task of capturing Linguistic Features arising due to sarcasm and humor.

For example, the following speech having humor probability of **0.9074** could not identify Applause correctly.

" but you know, what was different back in the '60s was that the passions welled up as they always do, and yet there seems to have been a generation of folks in the leadership who just looked at that, and said, "ok, we give up we surrender!" why?! to this day, i look back on that period, and i still can't quite figure out why you had a generation of people who, when the traditional values and mores and structures were challenged, didn't just look the kids in the eye and say, ***"look, you're the kids, we're the grownups now, you do what we tell you ""***

Similarly, examples having predominant high humor (**0.88**) and sarcasm probabilities (**0.84**) also faced a similar issue.

"the surplus argument particularly intrigues me because, well, i don't know–it could be that i'm just too simple-minded for politics, as reagan used to say but i look at that discussion, and i wonder if we ran out and a bought a car, and a few weeks later we got a letter from the dealer, and the dealer told us, "we've just been over the books and we've found that you've overpaid for the car and we're going to have a meeting tomorrow of the salesmen and directors to decide how much of the money that we're going to send back to you"

2. Except for these the BERT based model still had a few other errors outlined in the following section which even our Linguistic features based model couldn't catch.

7.3 Error Analysis - BERT + Linguistic Features

Addition of these probabilities obtained from the humor and sarcasm model substantially helped the model improve the linguistic feature capturing characteristics. Though there are still a gamut of issues yet to be resolved which are listed below with relevant examples :

1. **Pragmatic Inferences** : Unlike human understanding BERT based model is not able to capture Applause based on common sense. For example, in the following example unlike a human who would understand the placement of word "people" when "not government or hard earned money" is mentioned, BERT was unable to encompass such a pragmatic meaning. Thus, failing to predict the Applause.

   "we understand, as well, whose money we spend in washington, d c the money we spend in washington is not the government's money, it is the hard earned money"

2. **Numerical References** : Quite a few speeches recieved Applause due to utilization of statistics or comparative analogies. BERT was clearly not able to capture the same. For example, this has been clearly depicted in the following example :

   " recently, there was a little town their traffic signs were only 5 feet high, and they decided to raise them, for better visibility for the motorists, to raise them to 7 feet above

the ground and the federal government came in and said they had a program that would do that for them they lowered the pavement 2 feet "
As the example had a clearly low humor and sarcasm score as well the Applause can majorly be attributed to the numeric comparison.

3. **Additional Linguistic Features** Additional linguistic features likes promises, dramatic assertions made in political speeches were majorly missed. For example, the following example very clearly denotes a dramatic promise carried in the speech which would have lead to an Applause by the ehanced as well as baseline BERT model was unable to capture the same.

   "but we need to find the courage and the creativity to solve the problem we're not like some of those countries who give you your wages for a year and a half and all of your benefits if you lose the job in america people need to work and you just think about it, about half our problems would go away overnight if everybody in this country who wanted to work had a job"

4. **Data specific issues** : On the flip side our model did help us highlight a few data processing or annotation improvements which are required for this specific use case. For example, the following example should clearly have had a very high Applause score but due to their presence in a larger part of the speech the audience may have refrained from applauding. Thus, giving a 'NEGATIVE' tag to the data. But the model was very well able to capture the same and provided us a scope to think about ways to incorporate human tendencies into the modeling framework.

   "this saturday will be an historic milestone for all of afghanistan, especially for the more than 4 million women who will be heading to the polls because we acted, the people of iraq are now free from the tyranny of saddam hussein"

Majority of the issues noticed in the above section are areas of improvement for BERT and are currently under research. A further analyses of the literature to understand the errors better have lead use to explore these mentioned in (Wallace et al., 2019),(Kovaleva et al., 2019),(Belinkov et al., 2017), (Zhang et al., 2019) which form as major areas of future work.

8 Contributions of group members

Each project member was well involved and contributed significantly to the project. An overall break-down of the work done by each member is mentioned below:

- Literature Survey: The team collectively researched on existing problems which were similar to the one we were trying to solve. Ankita worked on understanding our baseline paper (Guerini et al., 2013), and researched on the ways to improve the results. Apurva and Parag explored the other model architectures that could be relevant to our problem statement and tried to analyze state-of-the-art models. Kartik and Chirag worked on the understanding the Humor and Sarcasm models and research on ways, these models could add value to our implementation.

- Data Preprocessing: Apurva worked on processing the data, so that we could use it for our experiments.

- Experiments: Ankita implemented the Pos with SVM model, referenced from our baseline paper, and performed the experiment wherein the class scores from Sarcasm model were added to BERT embeddings(with/without fine-tuning) to predict applause. Apurva performed the experiment wherein the probabilities obtained from the Sarcasm model were appended to BERT(with/without fine-tuning). Parag implemented the BERT Model which was used as the base, on top of which all enhancements were made. Parag performed the experiment wherein the probabilities obtained from the Humor Model were appended to BERT (with/without fine-tuning). Kartik implemented the C-LSTM-CNN model catering to our dataset, implemented the Humor Model, and performed the experiment wherein the class scores from Humor model were added to BERT embeddings(with/without fine-tuning). Chirag set up the Google Cloud Platform for the project use, and trained the Sarcasm model on it. He performed the experiments wherein the class scores and probabilities obtained from the Humor and Sarcasm model, were appended to BERT embeddings(with/without fine-tuning).

- Results/Error Analysis: The team collectively analyzed the results obtained and explored various research papers, to perform error analysis. Ankita and Apurva worked on gathering examples and performing error analysis. Parag, Kartik and Chirag developed inferences from the results obtained, and visually analyzed the results.

- Report Writing: The team divided the sections of the reports equally for each of the milestones.

9 Conclusion

In summary, the results pertaining to the improvement of BERT using linguistic features and the research carried out for understanding the future work specifically turned out to be surprisingly. The error analysis precisely mentioning the stark improvement in the results using linguistic features has not only been the major takeaway from the project but also an interesting area of futher exploration for the team.

9.1 Difficulties Faced

Like majority NLP tasks the major difficult faced was in creating the data. As political speeches tend to be large bodies of information exact annotation at a sentence-level though expected to be a straight-forward task required quite a few iterations. After multiple attempts between the milestones we could finally reach an approach solving majority of the issues we noticed. A few of them still remain which have been covere in error analysis section.

Another major challenging task was the identification of the right process to introduce the linguistic features into the BERT model. BERT being a highly parametrized network required a careful incorporation of not too granular but at the sametime impact worthy features. Thus, experiments with prediction labels and prediction probabilities were carried out.

Finally, the problem had large degrees of freedom due to the niche explored dataset, recent state-of-art model and very poor baselines. This ensured large amount of deliberation and increased number of failed experiments before reaching on satisfactory results for improvement. Thus, leaving the future work with atleast 1-2 pointers under each section.

9.2 Future Work

1. **Enhance Linguistic Features**
   Identify features beyond humor and sarcasm influencing applause. As per error analysis other linguistic features like assertiveness, promises could form as good indicators of applause in political speeches. This could help us target the remaining 28% of error analysis from the error analysis. Also, exploring other ways of combining these features into the BERT model. For example, using embeddings, prediction scores or incorporating a multi-task learning framework could be explored.

2. **Model Enhancement**
   Enhancing the model using audio visual data such as physical gestures from videos or audience response to improve model performance. Such experiments have been carried out on political speeches in the past (Navarretta, 2017) and thus can be used to improve the performance of our model.

3. **Analyses of BERT** :
   Our BERT based model still lacks a large number of characterictics and the recent research in this field has made us realize further indepth analyses which could be carried out for improving the model performance and at the sametime reducing the computation time.

4. **Extend Scope of Application**
   Extend the approach to domains beyond political speeches like books.

5. **Audience Profiling**
   We could use demographic data of audience as well as speakers to improve the performance of model. As explained in error analysis few reactions are also very audience specific.

References

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017). What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*.

Benton, A., Arora, R., and Dredze, M. (2016). Learning multiview embeddings of twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Berlin, Germany. Association for Computational Linguistics.

Celli, F., Pianesi, F., Stillwell, D., and Kosinski, M. (2013). Workshop on computational personality recognition: Shared task.

Chen, P.-Y. and Soo, V.-W. (2018). Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Danescu-Niculescu-Mizil, C., Cheng, J., Kleinberg, J., and Lee, L. (2012). You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 892–901. Association for Computational Linguistics.

de Oliveira, L. and Rodrigo, A. L. (2017). Humor detection in yelp reviews.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/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)*, pages 92–102.

Guerini, M., Giampiccolo, D., Moretti, G., Sprugnoli, R., and Strapparava, C. (2013). The new release of corps: A corpus of political speeches annotated with audience reactions. in multimodal communication in political speech. In *International Workshop on Political Speech*, pages 86–98.

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

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., and Mihalcea, R. (2018). Cascade: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848. Association for Computational Linguistics.

Khodak, M., Saunshi, N., and Vodrahalli, K. (2017). A large self-annotated corpus for sarcasm. *CoRR*, abs/1704.05579.

Kim, Y. (2014). Convolutional neural networks for sentence classification.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. *CoRR*, abs/1405.4053.

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 *Eleventh International AAAI Conference on Web and Social Media*.

Navarretta, C. (2017). Prediction of audience response from spoken sequences, speech pauses and co-speech gestures in humorous discourse by barack obama. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000327–000332. IEEE.

Song, X., Petrak, J., and Roberts, A. (2018). A deep neural network sentence level classification method with context information.

Wallace, E., Wang, Y., Li, S., Singh, S., and Gardner, M. (2019). Do nlp models know numbers? probing numeracy in embeddings.

Zhang, M., Zhang, Y., and Fu, G. (2016). Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2449–2460, Osaka, Japan. The COLING 2016 Organizing Committee.

Zhang, S., Jiang, H., Xu, M., Hou, J., and Dai, L. (2015). The fixed-size ordinally-forgetting encoding method for neural network language models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 495–500, Beijing, China. Association for Computational Linguistics.

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., and Zhou, X. (2019). Semantics-aware bert for language understanding.