PolityCheck: A LSTM Approach to Detecting Political Ideology

Vishakha Dudani
Department of Computer Science
Stony Brook University (SUNY)
112078436

Abhishek Reddy
Department of Computer Science
Stony Brook University (SUNY)
112052058

Hamza Mahmood
Department of Technology and Society
Stony Brook University (SUNY)
111869947

Abstract

Political ideology detection is an exciting field of study in the domain of Natural Language Processing. Especially when considering the scenario where two popular presidential candidates or multiple political figures are embroiled in a debate, we observe a wide range of perspectives and understand the stance over prevalent issues faced on a government or civil level. This paper intends to analyze and measure how liberal and conservative a candidate is by extracting key ideological sentiments from congressional debates transcripts, available textual sources and tweets from the most recent GOP elections.

1 Introduction

We are trying to classify sentences said or written by famous politicians into two categories: Liberal or Conservative. The motive behind choosing this domain is that in modern era no candidate follows the partys beliefs blindly; for example, conservatives support equal work opportunities for women. So in todays world no speech given by politicians are overly extreme to the point they are classified as Liberal or Conservative, it is more of an amalgamation of beliefs, our set task is to define scores of the sentences which will tell us how much that sentence belong to particular class.

The general approach to tackle this problem has revolved around using bag-of-words classifiers and features which have often failed to capture the essential semantic information of data. Therefore, detecting ideological bias is a difficult task, even for humans. There is a need for understanding politics and current affairs at a deeper level. Additionally, linguists have tried to manually annotate on subtle phrases in language involved in order to detect agreement and disagreement. Then, based on these findings, implement various features to be captured automatically via supervised classification (Tonelli, 2016). Neural network models using target-word representations have also been constructed to predict stance of political candidates when delivering on key policy reforms (Chen et al, 2017).

There are some ideas and words on which these parties draw a long border between them. Words like tax, Tariff, Economy, marijuana, 'death-tax'. Each party uses a fixed set of words and idea in each election cycle. Our project focuses on training the LSTMs using labeled sentences and different word categories to get better accuracy compared with Bag of Words (BoW) and vanilla RNN models. Using the model we add features accordingly that include more terminologies specific to political reforms and policies.

The application has practical uses in the domains of tech policy as we aim to understand the trends showcased by candidates of how they vary language and speech. To summarize our work, we did the following:

1. We used a neural network approach for detecting political ideology.
2. Implemented an LSTM based model to
3. Our model performs better than the bag-of-words model with a improved accuracy of 10.4%.
4. Analysis shows that our model works best when sentences are longer and have more target/context words for more accurate ideology prediction.
5. We investigated that a bi-directional LSTM or attention based model are potential av-

enues which can improve our accuracy further.

2 Task Definition

We are evaluating on multiple data sources taken from congressional records available online, transcript versions of historic presidential debates and DW-NOMINATE scores (Lewis et al. 2017). Given the longevity of data, we add both breadth and depth into understanding the ideological context of presidents. However, with the assumption that language is subject to change and that the context of debate is relative to the existing members of the congress in a given point in time, we bifurcate our model to include year to provide a more focused analysis.

We have also scraped twitter data to test the hypothesis on contemporary political figures in the most recent GOP elections, to understand the way they shift and alter their point of view during the span of the electoral campaign.

Latest research in political ideology detection have used recursive neural network, LSTMs (Iyyer et al, 2014), multi-view document attention models (Skeina et al. 2018) and Hidden Markov based models to map Ideology trees by capturing cue words that signify the entropy over the next ideology state (Sim et al, 2013).

2.1 Dataset Details

Data was collected from two sources, first from the presidential Candidate debates from this link. The data consists of transcripts of every session of Congress in both the House and the Senate. The site provided PDF-format files of week/month long periods of the record, so we randomly downloaded sections from each year and took random pages from those sections.

The second source of data was GOP.gov which provided PDFs for each day, so we randomly downloaded days throughout the year and took random pages from those days. We then used pdfminer, a python tool to convert all pdfs into transcripts.

The data was cleaned up from JSON format to CSVs, then splitting the text into speakers and speech text. In the debates, the entries which were not in the format that would support this preprocessing method, were simply discarded. Extremely Long sentences were split if it exceeded the length of 20 words. Once we collected the training data, we needed labels to determine how liberal or conservative each sentence is. We combined our data from the Congressional Record with DW-NOMINATE scores from voteview.com. This score uses the voting records from members of congress to score their political ideology on a scale from -1 to 1. This scale is normalized across each member of Congress to every other member of their Congress based on the majority of recorded votes. A score of -1 denotes the member who is the exceedingly liberal while a score of +1 means that member is exceedingly conservative. The data collected had speech and debates from 1960 to 2016.

After training the model, we scraped tweets under the Twitter handles of Donald Trump and Hillary Clinton to test the model. We extract the keywords and perform lemmatization as part of the pre-processing stage. The tweets are then padded to match the sentence length needed by the model.

2.2 Baseline Model(s)

For our baseline we utilized a Bag-of-words model that takes important keywords based on crucial reforms such as taxes, drugs and gun violence. All the sentences in the corpus were converted into their bag-of words representation to indicate word counts. The overall vocabulary of the corpus is over 200,000 words but most words are not frequent. We eliminated all stop-words and selected the top 8000 words for the modeling. These bag of words were weighted using Term Frequency Inverse Document Frequency (TFIDF).

We took the frequency of the words and employed trigram probability to determine the ideology of candidates. The mean square error of our baseline model comes out to be 0.158. The next step is to extend the existing model by reconstructing it as a LSTM, which we will see that it substantially improves our ideology score accuracy.

2.3 The Issues

The baseline model has a number of limitations. Firstly, since we are using trigram maximum likelihood estimation (MLE), this model lacks the syntactic awareness about the text in general, as there are a limited number of political figures who are strictly neutral. Our sentences can be longer and N-gram estimation has the disadvantage of not dealing with long distance dependencies so the overall score will be flawed. Many trigrams in the corpus might be infrequent so the model will not able to make a better generalization on a candidate's ideology. For that, we need a sizable sample in order to get a good estimate. To cater to the zero probability problem, we can perform some type of smoothing but in the long term that can still prove ineffective as the model still will be unable to capture the subtle languages in the text. Therefore, we propose a LSTM model that is able to cover both long and short term dependencies.

3 Our Approach

We treat the ideological detection problem as a supervised learning task where we attempt to predict the rhetoric associated with each sentence spoken by a speaker or political figure. Using an LSTM works well when solving a problem specific to language modeling, transliteration or dependency recognition for part of speech tagging. Our model allows us to capture syntactic and semantic composition of the sentences with the ability to track subtle usage and changes of grammar and words. With respect to our selected data, the LSTM-based model will prove to be highly effective in understanding complex, verbose text where the ordering of words is of significant importance in our evaluation.

3.1 Idea 1: Recursive Neural Network

The paper which we are following for this project[1], introduced how we can solve the problem using Recursive Neural Network. The paper addresses the problem by detecting phrase level ideology scores and propagating the individual phrase scores to the sentence level.

3.2 Idea 2: Dependency Graphs

Second idea was to use dependency graphs to represent opinion words and context words and giving more weights to opinion words. Apart from the dependency graphs of the sentences we could have used a relational graph database which tells world knowledge to the model. [5]

3.3 Idea 3: Long Short Term Memory

The final idea which we locked in was to solve this problem by training a supervised regressor, we choose an LSTM model as it has the capability to sustain relevant information and forget the information which is not necessary as opposed to the traditional RNN based neural models.

3.4 Implementation Details

The input text is split into sentences and each word in a sentence is encoded into a unique integer representing the words, which is then fed into the LSTM. We now have to take care of the input vector, each sentence is now padded to form a fixed vector of size varying form 50, 100, 200. Each word in the sentence is then converted into a dense vector of size 100 and 200 (We can choose these parameters). This is fed to a sequential LSTM which is a type of recurrent neural network. An RNN is useful because it can be trained to understand sequential information. At any point in a sequence they can remember all the information they have seen so far, as they keep state of all prior data. LSTM as opposed to RNNs have the functionality that allows them to retain specific information and to forget the ones which are not so important for the task under consideration.

This is often useful in text analysis, as sentences have a structure where the order of words is important. A problem with a generic RNN is the vanishing gradient problem, where the network has trouble remembering long-term dependencies in the data. If two words are too further apart, then the network will not recognize their relationship. This is where an LSTM network has the upper advantage compared to an RNN.

We built an LSTM Neural Network using the Keras open-source library. Keras gave us the leverage of creating a network layer by layer. The first layer is comprised of the embedding layer. Instead of using a sparse representation of the words in our sentences, words are represented by dense vectors, where each vector represents the projection of a word into a continuous vector space. The position of a word within the vector space is based on the words that most often surround the word when it is used. The following layer is an LSTM layer, the main part of this net-

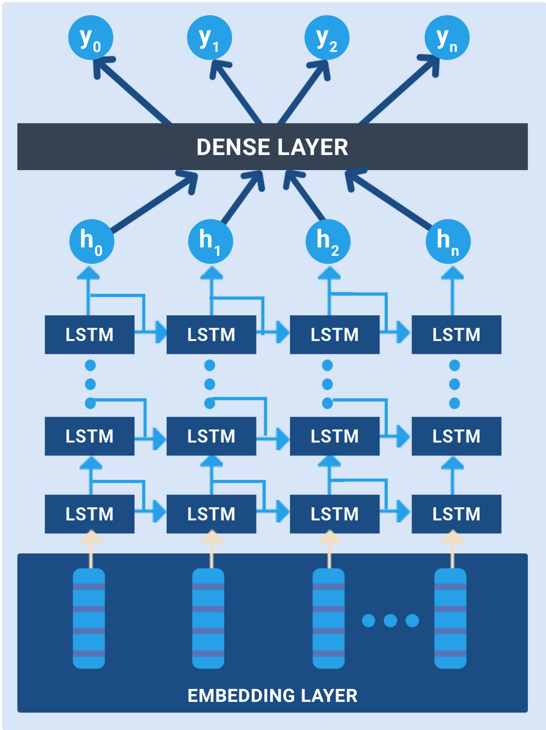

Figure 1: Architecture of our LSTM-based model.

work, followed by a dense layer, which provides the prediction.

4 Evaluation

In order to validate our results as sound and justifiable, our classification of presidential candidates that we input should accurately tag them as liberal or conservative based on our training data. We use the root mean squared (MSE) values as our metric on the calculated results from our training model. For each liberal and conservative candidate scored between a scale from -1 to 1, we compare it with the political ideology score obtained from voteview.com (Lewis et al. 2017). The score vector is represented by a scale between -1 to +1.

4.1 Evaluation Measures

We wanted to test our model on a more contemporary data set; hence, we decided to take Twitter as our source of tweets of the most recent candidates of the presidential elections.

We used our model to extract the most republican and democratic-centric tweets of Hillary Clinton and Donald trump which we can visualize in figures 3, 4, 5 and 6. The underlying logic for analyzing both spectrum for the candidates is that at times in order to huddle potential voters into their party's side, they write phrases and tweets as part of their campaign strategy.

Model	Mean Square Error (MSE)
Baseline	0.1701
LSTM (Political Terms)	0.1584
LSTM (All)	**0.1523**

Figure 2: Numerical result comparison

4.2 Optimal Hyperparameters

For our best model, we trained our data using the following hyper parameters:

1. LSTM Layer Size = 50
2. Embedding Length = 50
3. Max Sentence Length = 100
4. Epoch = 2

Our best hyper parameters given above were reached from the following strategy: Since the quality labelled data available to us was limited; hence, we choose to do a K-fold of 3 folds to train our LSTM model, and at each fold we ran the model by varying the parameters items given below.

- LSTM Layer size with values 50, 100 and 200.
- Embedding Vector Length of sizes 50, 100 and 200.
- Maximum Sentence Length of sizes 100, 200 and 300.

For each fold we train the model, then repeat the process over all possible combinations of the hyper parameters chosen above and we calculated the average mean squared error for each fold and then based on these results which is shown in a table included in the results section, we arrive at the best possible solution with the given hyper parameters, shown above.

4.3 Results

The table below illustrates the results obtained by our model on 3 fold training of all the combination of the hyperparameters mentioned in the section above. We have achieved a maximum accuracy of 60% for best hyper-parameter which are, LSTM

Layer Size = 50, Embedding Length = 50 and Max Sentence Length = 100. Also the average test MSE is around 1.56. The average is over the each fold of the training data.

Embedding Length	LSTM Layer Size	Max Sentence Length	Avg. Test Accuracy	Avg. Test MSE
50	200	200	0.591782	0.15682
50	50	300	0.594184	0.157012
50	100	200	0.581832	0.158559
50	100	100	0.593755	0.157604
50	200	300	0.595471	0.156557
50	50	200	0.602505	0.158099
50	50	100	0.593927	0.156637
50	100	300	0.59041	0.156395
50	200	100	0.591954	0.155704

Figure 3: Result with various hyperparameters

We wanted to highlight the top 10 tweets of both Hillary Clinton and Donald Trump in which they deliver remarks that are either conservative or liberal in their nature of wording. In figure 5, 6, 7 and 8 we identify the dates where we saw Trump and Clinton posting tweets reflecting the bipartisan nature of the election campaign. The main reason is to maximize their votes in different regions. An interesting insight would be to recognize the geographical location of the mentioned tweets for more elaboration on their election campaign. Our understanding is that certain target words manipulate the scale from left to right and a slight tweak on it can make some level of difference.

actual	predicted	difference	Sentence Text
0.782	-1.224	-0.442	This Congress ought to do its job and act today on a real continuing resolution to keep the full Government...
0.863	-1.206	-0.344	Four years ago, this body amended the Food Stamp Act in order to expand and im prove the food stamp program...
0.863	-1.155	-0.292	While there are a number of legislative proposals to address this issue, the consensus is clear: we need a fail-safe...
0.739	-1.152	-0.413	It is estimated that we have lost over 1 million jobs in the auto and related industries since the mid-1970's because...
0.691	-1.142	-0.451	Since 1992, the Great Lakes Shipwreck Historical Society has operated the Great Lakes Shipwreck Museum to ...
0.541	-1.135	-0.593	In the coming decades, some scholars believe, food scarcity will be the normal con dition of life on earth-and not...
0.739	-1.126	-0.386	But let there be no change by usurpation; for though this is one instance of good, it is the customary weapon by...
0.919	-1.115	-0.196	Speaker, I rise in support of the gentleman's amendment...
0.919	-1.098	-0.179	This year's infrastructure grade from the American Society of Civil Engineers is a ''D+.'' Sound infrastructure...

Figure 4: Prediction of Worst Sentences

Moreover, we obtained the result of scoring the tweets of Hillary Clinton and Donald Trump during GOP campaign, by our trained model. As we can see from figure 9, 10, they stick to a centrist ideology to attract voters, while Donald Trump's graphs has peeks on the conservative side of the spectrum, whereas we can also observe Clinton's tweets peaking to the liberal end of the spectrum.

4.4 Analysis

Testing our hypothesis of whether the candidate is leaning towards the left or the right, we captured tweets of the most recent GOP election candidates, then performed multi-variate statistical analysis on

Date	Tweet Content
07-16-2015	"@ChanRog I liter give fist pump I see @realDonaldTrump tell thing way & call credibl
08-10-2015	".@BrandenRoderick I pleas see wonder statement made media.I'm surpris special person
08-21-2015	"@MattyJack33: @TomLlamasABC So refresh someon call left winger pose journalists. Keep Trump!
10-04-2015	"@AndreaTantaros: DonaldTrump right. And sadli Christian treat better Hussein; Gaddafi ISIS. Clinton may go jail...
11-27-2015	"@VaughnVhalen: Donald Trump Rise 38% Nationally; Ted Cruz Edg 2nd 12%, Great twitter quot Donald Trump: ""Never apolog fake...
02-27-2016	"I self-fund campaign therefor control lobbyist special interest like lightweight Rubio Ted Cruz!
03-20-2016	"@gamzorz: @megynkelli Dont worri Trump They lose thousand viewer money. Roger Ail tell Megyn stop
06-06-2016	"I get bad mark certain pundit I small campaign staff. But small good flexibl save money number one!
07-01-2016	"Yet anoth terrorist attack today Israel -- father shot Palestinian terrorist kill while:https://t.co/Cv1HzKVbiT
10-16-2016	"Hillary' staff thought email scandal might blow over. Who would trust peopl nation security? https://t.co/EvBCQoZRG2

Figure 5: Trump's top 10 most conservative tweets

Date	Tweet Content
07-07-2015	"@CoachJMan: If major bind togeth support Trump revers downward spiral US on. http://t.co/J1FEV8eHg1
08-06-2015	"@HowardJax50: @FrankLuntz NBC say Trump lead HS grad less...doesn't bother say Trump lead colleg grad
05-15-2016	"The fail @nytim wrote yet anoth hit piec me. All impress nice I treat women found nothing. A joke!
06-02-2016	"Same fail @nytim ""reporter"" wrote discredit women' stori last week wrote anoth terribl stori today- never learn!
06-25-2016	"Thought prayer everyon West Virginia- deal devast floods. #ImWithY
06-25-2016	"On 13th tee box @TrumpScotland grand daughter Kai! @DonaldJTrumpJr https://t.co/7ii4KUfsab
07-26-2016	"Dem want talk ISI b/c Hillary' foreign intervent unleash ISI & refuge plan make easier come here.
07-30-2016	"#CrookedHillari = Obama' third term would terribl news econom growth - seen below. https://t.co/y9WJoUaaql
08-22-2016	"@Jimbos2002: @Morning_Jo Video: Hillari refer black super predat need brought heel. https://t.co/pMIHWayMRw
10-10-2016	"In administr EVERi American treat equal protect equal honor equal #Debat #BigLeagueTruth

Figure 6: Trump's top 10 most Liberal tweets during the campaign

Date	Tweet Content
9/12/2016	The report Hillary' email left basi fact veer danger territory. https://t.co/ycp0KW7LZL
9/27/2016	Have small busi owner dad may get multi-million-dollar bailouts, teach va... https://t.co/3LV3pmjRBI
10/27/2016	"Cast vote ultim way go high go low." —@FLOTU https://t.co/tTgeqxNqYm
10/22/2016	"Whether agre disagre me, whether vote not, I believ disagre without disagreeable." —Hillari
9/15/2016	"Peopl accus kind things—you'v probabl seen that. But nobodi ever accus quitting." —Hillari
9/15/2016	"I'm go close campaign way I began career...focus opportun kid fair families." —Hillari
10/21/2016	"We need get better find way disagre matter policy, agre question decency." #AlSmithDinn
9/18/2016	"I'm go close campaign way I began career...focus opportun children fair families." —Hillari
11/1/2016	"The thought Donald Trump nuclear weapon scare death. It scare everyone." https://t.co/E77BgjQGj7
9/14/2016	2. How handl non-cancel contractu oblig parti whose interest conflict Unit States?

Figure 7: Clinton's top 10 most conservative tweets (max score +0.3)

Date	Tweet Content
8/19/2016	RT @MarlonDMarshall: Trump paint entir African American commun live poverti job show he' complet to...
8/25/2016	Trump famous post anti-Semit graphic first appear white supremacist website. https://t.co/JgH87OylQC
8/23/2016	The gender wage gap even wider women color. It' time ensur equal pay. #BlackWomensEqualPay
8/25/2016	State regul fine one Trump' casino repeatedli remov black dealer floor.
10/20/2016	RT @BernieSanders: Donald Trump' vision America: tax break billionair larg corporations, poverti wage work familie...
9/16/2016	Trump' birther stem innat belief led discrimin black tenant earli career. Can't undone.
9/16/2016	Trump spent year peddl racist conspiraci aim undermin first African American president. He can't take back.
10/17/2016	RT @HoustonChron: We found civic vital reprint recommend @HillaryClinton 4 languag wide spoken https://t...
11/8/2016	RT @TheDemocrats: Let us bend arc histori toward justic more. https://t.co/JjkvA5ZKTj
10/11/2016	The violenc transgend American face—particularli transgend women color—i rebuk us. We t...

Figure 8: Clinton's top 10 most democratic tweets

the tweets that spanned from the beginning till the end of the campaign. We observed interesting trends through visualizing a time-period chart.

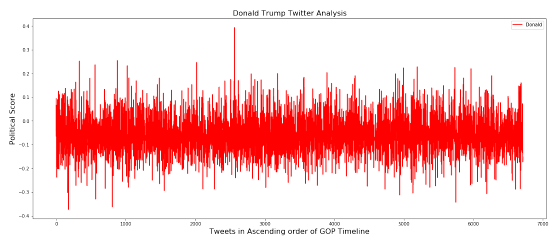

Figure 9: Analysis of Donald Trump's tweets

Closely looking at figure 9, we observe that Trump's tweets are more hinged towards the positive side of the scale with a number of peaks

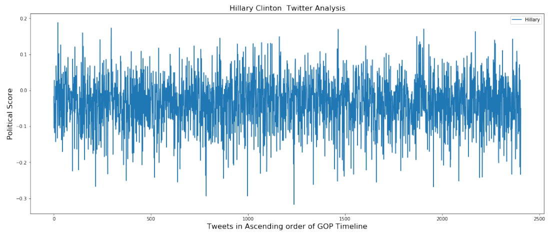

Figure 10: Analysis of Hilary Clinton's tweets

hitting a maximum of 0.4 which we essentially categorize as republican-centric. In the case of Hillary Clinton's tweets in figure 10, we see the peaks reaching more in the negative side of the spectrum, up to a minimum of -0.3; therefore, classifying her as a democratic candidate.

Few limitations of our model would be since the labelled data we have is based on the particular person's political score and not exactly on the context of the speech or text this might lead to incorrect labelling and which in-turn might lead to an imperfect model hence our accuracy of 60 percent, which is although quiet exceptional the incorrect values can be attributed to the claim in our model.

4.5 Code

We have shared the main runnable code file and the data related to it in the below mentioned drive link.

1. Drive link for code base

2. Our work is loosely based on code in this, repository .

3. We added a new file LSTMlearn.ipynb, which contain the logic to run the LSTM, and also we have added an application of the model which we have obtained, which to understand the way GOP candidates (Donald, Hillary) shift and alter their point of view during the span of the electoral campaign.

4. We required the below software and frameworks to complete our project.

   - Tensorflow v1.11
   - Keras v2.20
   - Python 2.7
   - Anaconda 2

5. **Infrastructure**: Since we are training a LSTM model which requires higher computational capabilities we have setup our working code in Google compute engine of the configurations shown below.

   - 12 virtual CPU's
   - 44 GB RAM.

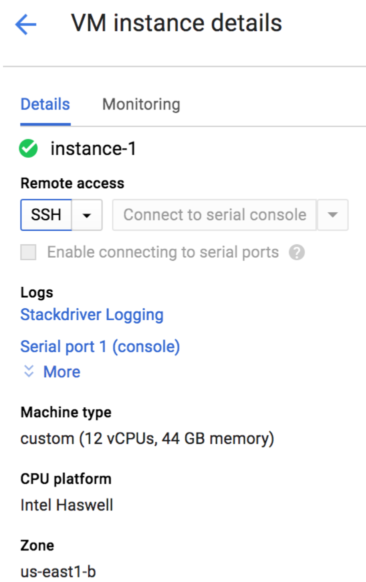

Figure 11: Configuration settings for training our model

5 Conclusions

There are lots of work done in the Political Ideology detection area, but none of them used LSTMs or any other Neural Models. In one of they papers [1], they do use Recursive Neural Network which inspired us to do this project. Political Ideology is generally assumed to be something in line with the party the politician is representing. But in today's world no speeches are made in extreme. That's why our model also predicted most of the sentences as central or neutral. Hence, we decided to analyze the political ideology by scoring each of the sentences said or wrote by the politicians. This gives a clearer picture in different time frames.

The speeches made by the politicians can be huge and their transcripts generally don't maintain the punctuation properly. So for long sentences having a neural model with memory cell will be boost to the performance. That was the main reason why we used LSTM for this project.

We trained the LSTM also with common political words, tax and drug related words so that it has some knowledge about the topics. The model performed better than just using vanilla LSTM.

6 Future Work

- Due to time restriction we randomly selected data from the sources, we can extract the complete data and can train the model with more sentences.
- We can try different flavors of LSTMs for better accuracy in our scores. One idea discussed in [2] is to use a convolutional, bi-directional LSTM hybrid which has the ability to capture the target context and semantics of text.
- In one of the previous works, the paper gives different weights to opinion words and context words differently, taking inspiration from this work we can implement an attention model along with LSTM which gives opinion words more attention.

7 References

[1] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber,Philip Resnik: Political Ideology Detection Using Recursive Neural Networks.

[2] Li, Xilian Chen, Wei Wang, Tengjiao Huang, Weijing. (2017). Target-Specific Convolutional Bi-directional LSTM Neural Network for Political Ideology Analysis.64-72. 10.1007/978-3-319-63564-4-5.

[3] Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, Noah A. Smith. Measuring Ideological Proportions in Political Speeches. ACL (2013).

[4] Stefano Menini and Sara Tonelli. 2016. Agreement and disagreement: Comparison of points of view in the political domain. In Proceedings of the 26th International Conference on Computational Linguistics, pages 24612470

[5] Wei Chen, Xiao Zhang, Tengjiao Wang, Bishan Yang, Yi Li1 2017. Opinion-aware Knowledge Graph for Political Ideology Detection