Divisive Language and Propaganda Detection using Multi-head Attention Transformers with Deep Learning BERT-based Language Models for Binary Classification

Norman John Mapes Jr., Anna White, Radhika Medury, Sumeet Dua
Louisiana Tech Office of Research and Partnerships

NLP4IF 2019 Shared Task - Team: ltuorp, Rank SLC #1

Abstract

On the NLP4IF 2019 sentence level propaganda classification task, we used a BERT language model that was pre-trained on Wikipedia and BookCorpus as team ltuorp ranking #1 of 26. It uses deep learning in the form of an attention transformer. We substituted the final layer of the neural network to a linear real valued output neuron from a layer of softmaxes. The backpropagation trained the entire neural network and not just the last layer. Training took 3 epochs and on our computation resources this took approximately one day. The pre-trained model consisted of uncased words and there were 12-layers, 768-hidden neurons with 12-heads for a total of 110 million parameters. The articles used in the training data promote divisive language similar to state-actor-funded influence operations on social media. Twitter shows state-sponsored examples designed to maximize division occurring across political lines, ranging from "Obama calls me a clinger, Hillary calls me deplorable, … and Trump calls me an American" oriented to the political right, to Russian propaganda featuring "Black Lives Matter" material with suggestions of institutional racism in US police forces oriented to the political left. We hope that raising awareness through our work will reduce the polarizing dialogue for the betterment of nations.

1 Introduction and Related Works

A question can be posed "What is an influence operation also known as?" Our system was trained to answer these questions but in the form of a cloze comprehension test "_____ is an influence operation." Likewise, Wikipedia and BookCorpus were used to develop an unsupervised language model built from the cloze questions by deleting 10% of the words from the corpora. Then the model was fed forward and a softmax output selected the most appropriate word, if this word was correct no training was done, if it was incorrect then the error was backpropagated through the network from the last layer's neurons to the first layer's word embeddings that were the inputs. Because an attention-based transformer can discern the difference between a river "bank" and a deposit "bank" depending on the context of the words, these word embeddings are considered dynamic. This contrasts with static word embeddings that were popularized by Mikolov et al. 2013, where bank has the same embedding regardless of context. Our model looks both to the left in the sentence and to the right and encodes the position of a word using a sinusoidal addition to the embeddings giving it awareness of the order of words. The model we based our approach on is called BERT by Google Research (Devlin et al. 2018). We independently discovered the value of using BERT like in D. Giovanni, 2019. BERT has undergone many changes to become RoBERTa (Liu et al. 2019) from Facebook. BERT and its related works have remained close to state of the art on tasks such as SQuAD (Rajpurkar et al. 2016). Although these results are less than a year old and nearly perform question answering better than humans, the superhuman level has been achieved recently in a very rapidly moving field. But it cannot be said this was unexpected given the results that IBM had when it bested the two

Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 103–106

strongest Jeopardy Champions (Markoff, 2011) for a million-dollar prize nearly 8 years ago.

2 Methodology

Our approach was based upon a very recent state-of-the-art release by Google Research (Github, 2019), we worked in the Python programming language to preprocess the data, set parameters, train, validate and predict propaganda. To accelerate the pace of our feedback loop (data to predictions to metric of success) we used a train/test split of 80/20 on the first 10% of the training data. We trained for optimal F1 score and noted Matthew's Correlation Coefficient, and ROC AUC for additional tuning. These values were optimized using a manual grid search for F1 score while monitoring the other metrics. If one of the monitored metrics performed particularly poorly, then we chose a model with more competitive values for all the metrics. We began with a robust model of TF-IDF and Random Forest to establish a baseline around which we can experiment with several other models. In the end we found the unsupervised language model BERT to be most effective after supervised re-training.

We will now discuss the parameters that we experimented with in our final model and chose according to performance on the validation set. The BERT parameter of sentence length was set to the first 50 words. If a sentence was longer than 50 words, then the 51st and beyond were discarded. Our batch size during training was 32 and 500 during prediction. Gradually increasing the training batch size usually improves performance. However, we were running at maximum memory on our computational resources and were unable to increase batch sizes. Our learning rate began at 1e-5 and gradually increased according to the default warm-up schedule.

Attention is defined as:

$$Attention(Q, K, V) = Softmax(\ QK^T)V \quad (1)$$

$$Softmax(X) = \frac{\exp(X)}{\sum \exp(X)} \quad (2)$$

Where softmax takes a vector X and Q, K, V are all embeddings of dimensionality 768. For a more detailed low-level understanding of attention see Vaswani et al. 2017. Because each time the neural network is initialized a new random number is used for the embeddings it is useful to ensemble attention neural networks for multi-head results. Each head gives a generally unique interpretation of the sentences. In our case we used 12 attention heads and 12 transformer blocks. Attention gives a particularly interesting result, as it selects for words which have an additional significance when used together, effectively capturing the interaction and sending this signal through to the next layer. This interaction along with the position encoding give the transformer the ability to consider context. For more discussion of transformers see Devlin et al. 2018. The dataset used is described in D. Giovanni 2019.

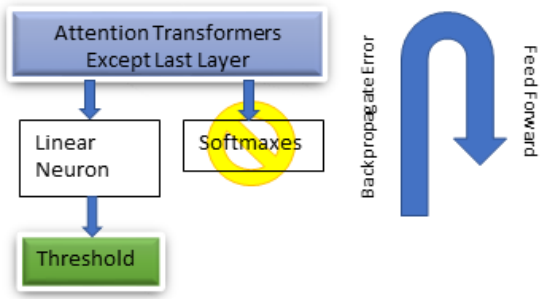

Figure 1 BERT-based attention transformer model with softmax layer substituted for a real valued neuron.

3 Results

On the development set we obtained two scores: one that was our internal 80/20 split on the 10% of the training data and the second that was based on the full set submissions to the webserver as team ltuorp. We selected the model parameters that were best for both. We found a threshold of 0.3 to classify propaganda was most effective for higher F1 scores. The threshold was selected using a manual grid search. By using a threshold, we formulated the problem as a regression problem. During training 0 was non-propaganda and 1 was propaganda. Then predictions were taken on the validation data and run through the regression model. If the predicted value was less than 0.3 it was classified as non-propaganda if it was equal to or greater than 0.3 it was classified as propaganda. We believe by having multiple datasets we were able to develop a better model. These datasets are both the language model that encompasses all of Wikipedia and BookCorpus and the partitioned training data. Had time allowed we would have used yet another frame of

reference on the development set by performing 10 fold cross validation or leave one out validation.

A thought-provoking finding is that even though there are 18 categories of propaganda we were able to perform binary classification with a precision of 60.1, a recall of 66.5 and an F1 of 63.2 indicating that most of the propaganda follows a repeatable pattern in language and does not require human level intelligence or the need to recognize complex patterns to discern whether or not a sentence is propaganda. The baseline is 43.7, 38.8 and 49.4 respectively for comparison. The remaining 36.8 of F1 however would require a more complex model to classify. Because most propaganda follows a pattern it is possible to objectively and automatedly evaluate a publisher. For instance, news network X was found to have Y% more biased news than news network Z. Governments, critical readers, fact checking organizations, policy advisors, news companies, social media and internet companies can all make informed judgments based on the results of using these models.

4 Discussion and Future Work

The impact of our results cannot be overstated. Peer and near-peer competitors to the USA and allies spend money to influence US elections to a favorable outcome for the rival at the expense of US voters who potentially fail to secure a superior candidate. When analyzing home-grown propaganda, it is eerily similar, to the point of being indistinguishable from the foreign influence operations' divisive language that was found on social media such as Twitter and Facebook ads such as those in Figure 1. (Persily, 2017 and Twitter Data Release, 2019 and House Intelligence Committee 2017).

Figure 2 (Top Image) Russian propaganda using racially divisive content where 12,858 Rubles were spent. This is file P(1)0002156.pdf from the 2015-q2 archive in the citation above. Blue ovals have been placed to protect identities. 126 million Americans were exposed to organic content based on 3,393 Russian advertising campaigns. Any divisive topic was subject to use in these campaigns. (Bottom Image) Twitter based foreign information operations content.

In future works it would be significant to find divisive content such as those used in the Russian state-sponsored campaigns. It is often more subtle, image based, social media based and not found in traditional news sources. Also, it is usually disguised as counter-dialogue. However, this work and model gives a baseline upon which we can improve, using techniques such as the following.

We are very interested in the cloze question answering pre-training method that BERT uses. Perhaps in the future the model will be able to not penalize "good" answers. If there is a synonym that BERT predicts but it does not match the expected word, then it will train to reduce the probability of the acceptable but unexpected word occurring in that position.

Another future contribution will be the ability to reason using common sense. For example, in the Winograd Schema a question can be posed: "The city councilmen refused the demonstrators a permit because they [feared/advocated] violence." To answer the question the model must understand and have knowledge of the world and sentence structure to disambiguate the pronouns. It must also associate "councilmen refuse permit" as being incompatible with "councilmen fear". While "councilmen refuse permit" is compatible with "protesters who advocate violence". The best attempt only gets 70% accuracy on a default accuracy of 50%. (E. Davis 2019). This means

that future works will no doubt raise the level of performance on Winograd Schema, a measure of commonsense reasoning and therefore likely, also the sentence level propaganda detection task.

5 Conclusion

We demonstrated good performance on classifying propaganda by attaining first place of 26 on the SLC task. It is our hope that the model and methods described in this paper will be used to create a more informed public that is resistant to divisive messages masked as counter-dialogue. One could conjecture that the motivation of foreign information operations is to sew discord and to reduce unity of a society's populace. We remain politically neutral with a hope that divisive language is not used intentionally to polarize others and in cases of legitimate promotion of already divisive topics, that polarization can be functionally minimized as opposed to unintentionally creating further division of an audience while advancing politically charged causes such as healthcare or social security reform (Howard, 2018). It may not be apparent how this happens, but common devices identified in the FLC portion of this competition such as flag waving i.e. conflating the opposing viewpoint with being unpatriotic, etc. is one example of many possible. While some propaganda has an element of truth, it is up to the reader to discern that they are being targeted to promote the cause of an information operation that often has a conflicting motivation with the reader's.

Acknowledgments

We would like to acknowledge the Propaganda Analysis Project and Twitter for providing relevant and high-quality datasets. Without these datasets we would have no empirical conclusions or be able to further the discussion to what to do with the results obtained and future directions.

References

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Da San Martino, Giovanni, Y. Seunghak, B. Alberto, P. Rotislav, N. Preslav "Fine-Grained Analysis of Propaganda in News Articles" EMNLP-IJCNLP 2019

Liu, Yinhan, et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint arXiv:1907.11692 (2019).

Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." arXiv preprint arXiv:1606.05250 (2016).

Markoff, John. "Computer wins on 'jeopardy!': trivial, it's not." New York Times 16 (2011).

Github "TensorFlow code and pre-trained models for BERT" https://github.com/google-research/bert 2019

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Persily, Nathaniel. "The 2016 US Election: Can democracy survive the internet?." Journal of democracy 28.2 (2017): 63-76.

Twitter, "Election's Integrity Data Archive." https://about.twitter.com/en_us/values/elections-integrity.html#data (2019)

United State House Intelligence Committee "Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements" https://intelligence.house.gov/social-media-content/ (2017)

E. Davis, L. Morgenstern and C. Ortiz "The Winograd Schema Challenge" New York University https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html (2019)

Howard, Philip N., et al. "Social media, news and political information during the US election: Was polarizing content concentrated in swing states?." arXiv preprint arXiv:1802.03573 (2018).