NLP: Natural Language Propaganda

Andrew Fogarty — W266 — Professor Cer

Abstract

Who are the targets of insurgent propaganda? I investigate the ability to classify the targets (e.g, the U.S. or Kabul) of insurgent propaganda messages using a novel corpus containing over 11,000 Taliban statements from 2014 to 2020. In experiments with Convolutional Neural Network (CNN) and transformer architectures, I demonstrate that the audiences of insurgent messages are best captured by transformers, likely owing to its encoder-decoder architecture. This paper's contribution is twofold: First, it offers a new and novel data set with utility in classification and summarization tasks for machine learning. Second, it suggests that since the audience of messaging can be reliably identified, new opportunities are afforded to analysts to look closer at the contrasts in language to better understand the targets of information.

1 Introduction

Who are the targets of insurgent propaganda messaging? Are insurgent messages uniform across warring parties or are they conditional on a specific audience? Political propaganda, existing accounts assume, are a means to indoctrinate the masses with pro-regime (or organizational) values and attitudes (Huang 2015, Garth and O'Donnell 2018). This uniform view seemingly reduces the need to study the targets of propaganda messaging. I challenge this view by showing that the content of propaganda messages are conditional on a specific audience and that those audiences can be reliably identified. To test this claim, I conduct experiments with varied neural network architecture to evaluate its ability to achieve high precision[1] and recall[2] (e.g., $F1$) in a novel corpus of roughly 11,000 Taliban propaganda statements from 2014 to 2020.

Two main findings emerge. First, there is clear evidence that the content and thus the language used in propaganda is differential – meaning it changes depending on the audience. This means that we can use machine learning algorithms to exploit this variation in order to label data by target audience. Consequently, analysts should not only consider the content of the text they are analyzing (e.g., is it positive or negative) but also the target. Second, since 2017, the Taliban have increasingly released propaganda targeting the U.S. (see Figure 1). This is likely in part due the U.S.' increased willingness to negotiate with the Taliban and to withdraw from Afghanistan.

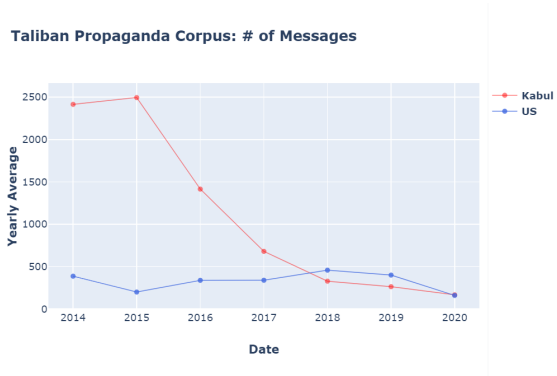

Figure 1: Corpus Overview

2 Literature Review

My research touches two broad tasks found in the literature: classifying sentiment and propaganda. Sentiment is comprised of two factors: (1) a judgment (e.g., is it positive or negative), and (2) a target (Liu 2012, 12). Predictive subtasks are then either based on: (a) predicting judgment or (b) predicting judgment given a target (Nakov et.

[1] Precision measures the extent to which the classifier produces false positives.

[2] Recall measures the extent to which the classifier produces false negatives.

al 2019, Rosenthal et. al 2017). These individual and collective tasks are extensively published on under the *SemEval* moniker. Far less commonly, researchers have focused on predicting the *target* of information (Bastos and Farkas 2019), which is the focus of my research, as a 2015 article posits that it is the first to do so using Twitter data (Lo et. al 2015).

In Lo et. al, the target audience is estimated by combining unsupervised (e.g., Latent Dirichlet Allocation) and supervised techniques (support vector machines) – drawing on Twitter followers and topic modeling. Bastos and Farkas derived a target audience typology inductively and consequentially labeled tweets (but did not use machine learning to analyze the text). By contrast, I use rule-based approaches to code my dependent variable, *target audience*, and use the text in each document as inputs for classification.

In terms of propaganda, researchers (Vlad et. al 2019, Yoosuf and Yang 2019) use BERT and BERT-ensemble models to detect certain techniques (e.g., appeal to authority, red-herring) or falsified information (e.g., fake news). While important in its own right, this research places the *content* of the text at the center of the analysis rather than *target* of the text. Taken together, this brief summary illustrates the need to broaden the scope of our research to consider not just whether we can predict sentiment, but the intended audience of that sentiment as well.

3 Corpus Construction

The novel propaganda corpus was built using a query database which contained translated copies of every Taliban propaganda message since at least 2010. Roughly 100 messages were combined into a single text file and then downloaded, yielding 131 text files and 11,553 messages. Given these files, I devised algorithms that searched and extracted the following information: (1) the body of the text, (2) the original source language (e.g., Pashto, Dari), (3) the message body, (4) the message title, (5) the message author, and (6) the message summary.[3] The dependent variable, *target*, or the intended audience of the message was coded in two ways: (1) if the summary contained selective key words (e.g., U.S., Kabul, puppets, invaders, Trump), or (2) in the absence of the former, word counts were generated and then compared with a specified list drawn from domain knowledge used to label the message. For instance, if *America* or *puppet* were the most commonly used words in two messages, then the messages were coded as *U.S.* and *Kabul* respectively. From there, a significant amount of data cleaning was then done to prepare the corpus for analysis.

4 Data Overview

The resulting corpus is highly imbalanced[4] and is comprised of varied length messages as measured by word count (see table below). The imbalance, instead of being a feature of collection bias or error, is likely due to the fact that the Taliban's ultimate competition is with the Afghan government and thus it makes substantive sense for a majority of messages to be targeted against Kabul. To simplify the analysis, only messages targeting Kabul or the U.S. were retained for the classification task.

Table 1: Describing the Data

Targets	N	Avg. Length	Max Length
Kabul	7763	188	5472
U.S.	2282	584	5112
Other	1508	207	1427

5 Model Selection

To test whether or not insurgent messages are conditional on specific audiences, I selected four competing models: (1) CNN (Kim, 2014), (2) Hierarchical Attention Network (HAN) (Yang et. al 2016), (3) BERT$_{base,uncased}$ (Devlin et. al 2018), and (4) T5$_{small}$ (Raffel et. al 2019). The CNN was chosen because it is a robust baseline, the HAN was chosen because its architecture is optimized to classify documents, BERT was chosen because it is the canonical transformer, and T5 was chosen because it represents the state-of-the-art with its encoder-decoder architecture.[5]

[3] The code for data extraction, cleaning, and modeling are located here: https://github.com/afogarty85/nlp_w266

[4] Hence the logic to choose $F1$ instead of accuracy as an evaluation metric.

[5] The transformers were built using *huggingface* while the CNN and HAN models were prepared by the University of Waterloo at: https://github.com/castorini/hedwig/tree/master/models.

6 Experimental Setup

In all experiments, I trained the models for five epochs with mixed precision in PyTorch and split the data as follows: 80% of the data for training, 10% for validation, and 10% for testing. The test coefficients were derived from the model checkpoint with the lowest validation loss. Hyper parameter tuning was conducted on the validation set with *optuna*, leveraging its asynchronous successive halving pruner (Li et. al 2018). To manage the data's imbalance, I used a weighted random sampler (leveraging replacement) to inject mostly balanced batches of data into the model during training.

For the CNN and HAN models, 200-dimensional GloVe embeddings were used (Pennington et. al 2014). Words that did not occur more than twice as well thousands of transliterated words that referenced locations, like *nahr-e-saraj* were removed as it is unclear, *a priori*, that their inclusion would be helpful. Two new vectors were created to account for padding and the remaining unknown words. Each unknown word shared a randomly generated vector based on GloVe's variance.

For the transformer models, I used the list of transliterated words that could not be matched with GloVe and removed them from the corpus. While transformers are incredibly powerful, it is unlikely that they are pre-trained on text that includes Afghan villages and locations and therefore such terms are only likely to complicate its learning.

7 Results

The experiments show relatively close agreement across the four models, with BERT, the central model of interest, barely and unexpectedly outperforming the CNN.

Table 2: Results

Model	Test Loss	Test F1
Zero Rule	NA	0.67
BERT	0.315	0.87
CNN	0.265	0.867
HAN	0.365	0.834
T5	0.16	0.825

8 Analysis

Along the way in this solo project several mistakes were made such as: (1) using 1 output neuron for binary classification, (2) using weighted random samplers on the test set, and (3) not pre-processing my corpus to be in concordance with the pre-trained embeddings. By correcting these issues, among other tweaks, I found notable evaluation metric and substantive improvements in my results (e.g., precision was not always 1.0). Perhaps unsurprisingly, I found the learning rate to be the most important hyper parameter as the coefficients are determinative in predicting each class.

In general, continual drops in the validation loss were very difficult to achieve across all models. One then might suspect the following about the data: its insufficient and/or noisy. One method I used to evaluate both suspicions was to sample my data by fractions of its total and plot the $F1$ metric results (see image below). Since the graph shows that the $F1$ metrics are roughly similar, 0.85 - 0.88, while the best performing model was produced with only 60% of the total data, this means that noise is complicating the model's learning. This makes intuitive sense because a majority of my data involves the Taliban declaring some short-term battle victory (e.g., they blew up a vehicle) while the rest are longer diatribes about issues like U.S. invasion, the peace talks, and the incompatibility of democracy and Islam. While indeed more data would help settle this issue, it can also be solved by a more careful consideration of what comprises the input data.

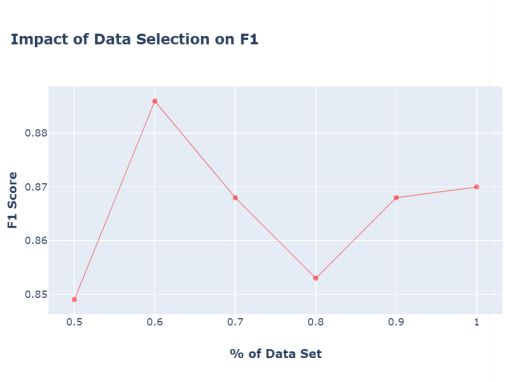

Figure 2: BERT Model

Lastly, it is worth exploring why the CNN, paired with GloVe embeddings, performs nearly as well as state-of-the-art transformer architecture.

While I do not have a definitive answer, it may be due to the following factors: (1) GloVe's vocabulary supports much of the language used in the text,[6] (2) localized features of high importance (e.g., the CNN's primary task) are sufficient to classify documents without relying on and interpreting the context, and (3) since CNNs can be tuned far more quickly than transformers, this yields a higher likelihood of learning ideal coefficients that generalize well.

9 Conclusion

The premise that propaganda is largely devised uniformly to influence as many people as possible remains a central tenant of our theories of propaganda. My evidence suggests a more complicated picture: propaganda messages are devised asymmetrically conditional on the audience which means that the language used to target one audience is different from the language used to target another audience. Given this, researchers should move beyond sentiment evaluations that merely judge an item by positive, negative, or neutral, but also the *target* of that sentiment. After all, my research shows that we can discern the target of the opinion from the text itself.

10 References

Bastos, Marco, and Johan Farkas. "Donald Trump Is My President!": The Internet Research Agency Propaganda Machine." Social Media+ Society 5, no. 3 (2019): 2056305119865466

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Huang, Haifeng. "Propaganda as signaling." Comparative Politics 47, no. 4 (2015): 419-444.

Jowett, Garth S., and Victoria O'Donnell. Propaganda & persuasion. Sage publications, 2018.

Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

[6] GloVe has vectors for Afghan provinces such as *Paktika*, for instance.

Li, Liam, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. "Massively parallel hyperparameter tuning." arXiv preprint arXiv:1810.05934 (2018).

Nakov, Preslav, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. "SemEval-2016 task 4: Sentiment analysis in Twitter." arXiv preprint arXiv:1912.01973 (2019).

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." arXiv preprint arXiv:1910.10683 (2019).

Rosenthal, Sara, Noura Farra, and Preslav Nakov. "SemEval-2017 task 4: Sentiment analysis in Twitter." In Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), pp. 502-518. 2017.

Vlad, George-Alexandru, Mircea-Adrian Tanase, Cristian Onose, and Dumitru-Clementin Cercel. "Sentence-Level Propaganda Detection in News Articles with Transfer Learning and BERT-BiLSTM-Capsule Model." In Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, pp. 148-154. 2019.

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. "Hierarchical attention networks for document classification." In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 1480-1489. 2016.

Yoosuf, Shehel, and Yin Yang. "Fine-grained propaganda detection with fine-tuned BERT." In Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom:

Censorship, Disinformation, and Propaganda, pp. 87-91. 2019.

11 Appendix

I envisioned this project to be an entire end-to-end natural language processing project, meaning I wanted to build my own corpus and analyze it using as many models as I could to maximize my exposure to the field. My data set offered additional flexibility in being able to use it for text generation as well as summarization tasks.

11.1 Summarization

While there are a number of different criteria by which we can judge a model's ability to summarize, by using ROUGE, we primarily care about recall which asks: how many n-grams in the summaries are in the body of the text? The results from T5 are listed in the table below:

Table 3: T5: Summarization Results

	ROUGE 1	ROUGE 2	ROUGE L
Precision	0.532	0.309	0.507
Recall	0.492	0.289	0.469
F1	0.506	0.296	0.483

11.2 Text Generation

I also used the DistilGPT-2 language model to generate text, which was exciting in its own right. I combined beam search with top_k and top_p to filter the vocabulary and produce the samples enumerated below.

1. The Afghan National Army reported that a mine explosion on the Kabul Ghouta district, at around 9:00am on Saturday night, killed at least 17 people and destroyed several properties. The blast took place as it was being...

2. The Afghan National Army reported a bomb blast that killed at least six Afghan soldiers in Kabul yesterday morning, but the exact number is not known. However, officials of the Taliban have denied the attack and said no such incident took place...