PAC Candidate Issues Analytics

Shiyi Shen
Graduate School of Arts and Science
Brandeis University
Waltham, 02453
shiyis@cs.brandeis.edu

Abstract

This project adopts Snorkel, a weak-supervised toolkit and Google T5, which is good at different language modeling tasks and tries to collect and understand the data pulled from the FEC database. The system was able to successfully extract the keywords using Snorkel but with a rather low coverage rate due to the condition of the data. With the data scraped, the paraphrase model was able to summarize and recover certain missing data due to the constraint of the scraper.

Keywords: transfer learning; super PACs; snorkel;weak supervised learning; snorkel

1. Introduction

The super PACs (the political action committee) determine their financial spending either based on how the legislator is going to spend the money once they are in office or on how it can influence the campaign results. Not much, however, is said about the relations between the finances and the issues that the candidate presents on their personal profile[1]. This project, therefore, seeks to explore such relations. The system currently consists of two parts: the first part is the keyword extraction with the weak-supervised learning model, Snorkel (Ratner et al). The second part adopts one of the language modeling system, more specifically the text-to-text transformer developed by Google AI, the "T5" model. This project also tries to explore the merits of current state-of-the-art technology its breakthrough accomplishments on unsupervised data and the data rich era where raw unprocessed data could be pulled from the Internet for further research.

2. System Design

2.1 Data collection

After a simple crawl with selenium and Google search, albeit small, out of the 587 candidates 200 of them their personal/campaign website was able to elicit meaningful data. More specifically, the issue page on their personal site was able to be found, crawled. The data, with simple clean, was injected into csv, which results in 3717 data points with average of 87 characters per text.

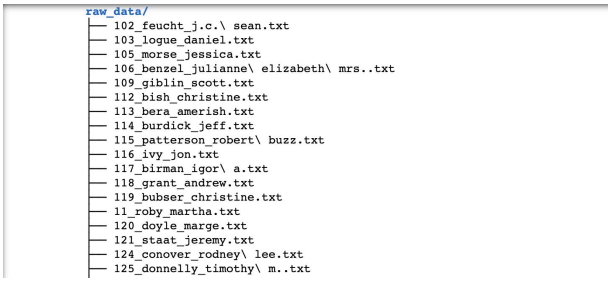


Figure 2.1: depicts the basic structure of the crawled data. Each number represents the order in which each political candidate in 2021 is scraped.

2.1 *Snorkel*

Create labels or topics from raw data weak supervised labeling model allows the generation of large amounts of data without the the burden of hand-labeling[2].

This model helps alleviate the bottleneck of current machines learning tasks that are manifested in the following ways: 1) not having enough labeled data; 2) not having enough field expertise in labeling data; 3) not having sufficient time to label data.

```
topic 1:      social_policy
keyword 1:    ['welfare', 'woman', 'women', 'lgbtq', 'gun control', 'abortion', 'planned parenthood',
'race', 'ethnicity', 'religion', 'gay marriage', 'adoption', 'women in combat', 'gender',
'reproduction', 'death penalty', 'seniors', 'prochoice', 'prolife', 'birth control', 'minority']
not_match 1:  ['continue reading', 'follow us on', 'follow us in', 'follow on', 'on social media',
'donate now']
required 1:   ['social policy']
label 1:      0
topic 2:      occupation
keyword 2:    ['basic income', 'equal pay', 'taxes', 'labor', 'sick leave', 'pay', 'payment',
'overtime', 'tariffs', 'pension', 'banks', 'banking', 'jobs', 'career', 'gnd', 'minimum wage',
'poverty', 'workers', 'farmers']
not_match 2:  ['continue reading', 'follow us on', 'follow us in', 'follow on', 'on social media',
'donate now']
required 2:   ['minimum wage', 'pay scale', 'payroll']
label 2:      1
topic 3:      economy
keyword 3:    ['economy', 'economic growth', 'federal funding', 'businesses', 'small businesses',
'spend', 'spending', 'booming', 'stagnant', 'money', 'funding', 'fund', 'finance', 'market', 'value',
'trade', 'stocks', 'real estate', 'assets']
not_match 3:  ['continue reading', 'follow us on', 'follow us in', 'follow on', 'on social media',
'donate now']
required 3:   ['economic growth']
```

Figure 2.2: shows a set of pre-defined topics that will be fed into the heuristic labeling function(lfs) of a snorkel model.

First a set of keywords are pre-defined for extracting the topics. There are three categories for the extraction and each of the labeling functions uses a regex to identify the matched, unmatched, and required keywords, as shown in the Figure 2.2

2.2 *Google T5*

Then the system goes on exploring the Google T5[1] model for further paraphrasing and interpretation of the crawled data. Transfer learning has long been appraised of having the merits of varied language modeling and low-level to high-level machine learning tasks. T5 on top of that is an explorative project alongside other DNNs that have populated the arena of machine learning as a result of the recent booming interests in AI.

Filling the gap of patchy data and interpretability
As due to the condition of how the data was collected and the quantity, there's a certain gap that could be filled between the raw data, processing the data, and the interpretation and training of the model. Under such premise, the T5 model was introduced in hope to fill the gap between retrieving less that satisfactory quality data and still able to train a machine learning model.

Initially a pre-trained T5 model was used for the current tasks. However, due to computational cost, I was not able to achieve a somewhat satisfactory result. In turn, I tried to utilize the multiprocesses package to accelerate the classifying speed. Yet, the model kept hanging at the last two rounds of the training loop. To quickly move forward with the project, I bootstrapped a model with both the training and classifying cycle. Below are demonstrations of the sample input and output of the final paraphrased results.

3. Output Demo

Below are demonstrations of how at each step the data is processed, collected, trained and put together.

3.1 *Data Collection*

	CAND_ID	CAND_NAM	CAND_PTY_	CAND_OFFI	CAND_OFFI	ELECT_TYPE	STATE_FULL	CAND_PTY_	campaign_site
1	H6AK00045	YOUNG, DO	REP	AK	0	House	Alaska	Republican	http://alaskansfordonyoung.com
2	H8AK01031	NELSON, TI	REP	AK	0	House	Alaska	Republican	https://johnnelsonforalaskans.com/
3	H8AK00140	GALVIN, AL'	IND	AK	0	House	Alaska	Independen	https://www.alyse4alaska.com/
4	H0AL01097	AVERHART,	DEM	AL	1	House	Alabama	Democrat	https://jamesaverhart.com/
5	H0AL01105	GARDNER, I	DEM	AL	1	House	Alabama	Democrat	https://kianigardner.com/
6	H0AL01139	COLLINS, FF	DEM	AL	1	House	Alabama	Democrat	https://rickcollinscampaign.com/
8	H0AL01063	LAMBERT, [	REP	AL	1	House	Alabama	Republican	https://www.weslambertforcongres
9	H0AL01071	PRINGLE, CI	REP	AL	1	House	Alabama	Republican	https://pringleforcongress.com/
10	H0AL01089	HIGHTOWE	REP	AL	1	House	Alabama	Republican	http://www.billhightower.com
11	H0AL01121	CASTORANI	REP	AL	1	House	Alabama	Republican	https://www.castoraniforcongress.c
13	H0AL02202	HARVEY-HA	DEM	AL	2	House	Alabama	Democrat	http://harveyhall2congress.com/
15	H0AL02087	ROBY, MAR	REP	AL	2	House	Alabama	Republican	https://martharoby.com
17	H0AL02145	COLEMAN, .	REP	AL	2	House	Alabama	Republican	https://jeffcolemanal.com/
18	H0AL02152	KING, TROY	REP	AL	2	House	Alabama	Republican	http://www.troyking4congress.com/

Figure 3.1: candidates personal web search results

First the candidate's personal web page links are pulled from the Google search page with a simple crawl through the *googlesearch* package in Python.

3.2 *Topic Labeling*

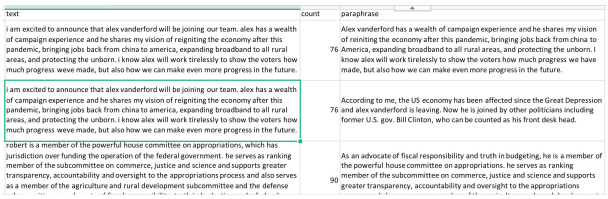
	name	text	count
0	ADERHOLT	i am excited to announce that alex vanderford will be joining our team. alex has a wealth of campaign experience and he shares my vision of reigniting the economy after this pandemic, bringing jobs back from china to america, expanding broadband to all rural areas, and protecting the unborn. i know alex will work tirelessly to show the voters how much progress weve made, but also how we can make even more progress in the future.	76
1	ADERHOLT	robert is a member of the powerful house committee on appropriations, which has jurisdiction over funding the operation of the federal government. he serves as ranking member of the subcommittee on commerce, justice and science and supports greater transparency, accountability and oversight to the appropriations process and also serves as a member of the agriculture and rural development subcommittee and the defense subcommittee. an advocate of fiscal responsibility, truth in budgeting and a federal government that operates within its means, he tries to bring commonsense solutions to the appropriations committee.	90
2	ADERHOLT	born on july 22, 1965, and raised in alabama, robert and his wife, caroline, have their residence in haleyville along with their daughter, mary elliott, and their son, robert hayes. when congress is in session, his family joins him in washington, d.c. area.	43
3	ADERHOLT	i take pride in the fact that president trump received a higher percentage of the vote in the 4th congressional district than any congressional district in the nation. during the past four years, i have formed a personal relationship with president trump, and we have fought side by side for rural america.	52

Figure 3.2: text to be fed into the *Snorkel* labeling functions for keyword matching

After a simple crawl using *Selenium,* each candidate with a personal web page is able to produce a certain amount of texts, which then gets separated into smaller chunk of an average length of 87 characters per text. It results in 3717 data points; although not a lot, it was really boiled down to data that the quality would not really get in the way of training the model (for the ones included, there's not much missing data, wrong punctuation, misspelling and each text was able to capture the essence of the message through human evaluation).

3.3 *Paraphrasing*

[1] https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html



text	count	paraphrase
i am excited to announce that alex vanderford will be joining our team. alex has a wealth of campaign experience and he shares my vision of reigniting the economy after this pandemic, bringing jobs back from china to america, expanding broadband to all rural areas, and protecting the unborn. i know alex will work tirelessly to show the voters how much progress weve made, but also how we can make even more progress in the future.	76	Alex vanderford has a wealth of campaign experience and he shares my vision of reinitting the economy after this pandemic, bringing jobs back from china to America, expanding broadband to all rural areas, and protecting the unborn. I know alex will work tirelessly to show the voters how much progress we have made, but also how we can make even more progress in the future.
i am excited to announce that alex vanderford will be joining our team. alex has a wealth of campaign experience and he shares my vision of reigniting the economy after this pandemic, bringing jobs back from china to america, expanding broadband to all rural areas, and protecting the unborn. i know alex will work tirelessly to show the voters how much progress weve made, but also how we can make even more progress in the future.	76	According to me, the US economy has been affected since the Great Depression and alex vanderford is leaving. Now he is joined by other politicians including former U.S. gov. Bill Clinton, who can be counted as his front desk head.
robert is a member of the powerful house committee on appropriations, which has jurisdiction over funding the operation of the federal government. he serves as ranking member of the subcommittee on commerce, justice and science and supports greater transparency, accountability and oversight to the appropriations process and also serves as a member of the agriculture and rural development subcommittee and the defense	90	As an advocate of fiscal responsibility and truth in budgeting, he is a member of the powerful house committee on appropriations. he serves as ranking member of the subcommittee on commerce, justice and science and supports greater transparency, accountability and oversight to the appropriations

Figure 3.3: paraphrase of the original text using T5

Due to the lack of large amounts of data, I experimented to explode/expand the data by trying current state-of-the-art model

3.4 *Demo*

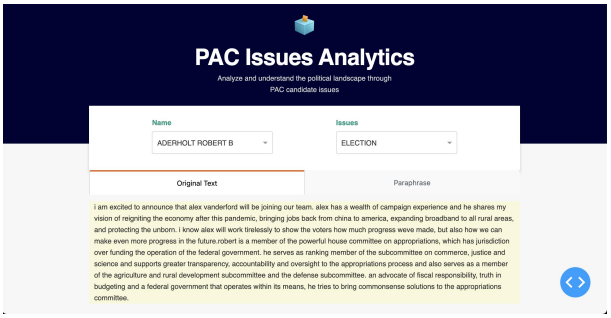


Figure 3.4: sample inputs and outputs are put together using the web services provided by *Dash*

Here everything described in the previous section is put together using the web services provided by *Dash,* which is a python quick web bootstrapping tool that's convenient to use to put together a simple web frame.

4. Evaluation

4.1 *Snorkel Eval*

```
Start topic training
                          j Polarity  Coverage  Overlaps  Conflicts
keyword_social_policy     0      [0]  0.008881  0.000807   0.000807
keyword_occupation        1      [1]  0.018568  0.003229   0.003229
keyword_economy           2      [2]  0.020721  0.004575   0.004575
keyword_election          3      [3]  0.038751  0.004844   0.004844
keyword_education         4      [4]  0.023950  0.001884   0.001884
keyword_healthcare        5      [5]  0.016685  0.004306   0.004306
keyword_security_policy   6      [6]  0.007804  0.002960   0.002960
keyword_domestic_policy   7      [7]  0.010495  0.002691   0.002691
keyword_foreign_policy    8      [8]  0.008073  0.001884   0.001884
keyword_immigration       9      [9]  0.010495  0.004037   0.004037
keyword_environment      10     [10]  0.006459  0.000269   0.000269
keyword_science          11     [11]  0.000807  0.000000   0.000000
keyword_infra_transpo    12     [12]  0.010495  0.001346   0.001346
keyword_public_health    13     [13]  0.002422  0.000538   0.000538
```

Figure 4.1: here is a simple run of the evaluation on the keywords defined

Within the system, there are metrics to evaluate the pre-defined keywords to gauge whether if a keyword is ill-defined. Because there's no ground truth to determine the quality of the labels. The metrics that *Snorkel* had was for facilitating the system to do an estimate on how well all the labels did through measuring the *Coverage, Polarity, Conflicts,* and *Overlaps* of each of labels defined. Here the evaluation shows an overall lower coverage rate. This might be due to the fact the quantity of the data is not significant enough.

4.2 *Google T5 Eval*

Due to some technical difficulties, the evaluation on the T5 model was not complete. However, I did go through some numbers of text paraphrases with human evaluation. I came to understand that the way that T5 classification works is through beam search, so I included in the hyperparameter the Top-10 search results if there's any, and I came to realize that although the model was able to capture the gist of some of the texts, it has a tendency to change the formality and the overall tone of the original data as shown below,

womens reproductive rights are under constant attack and were still a long way away from realizing full pay equity for women. in congress, ive been standing up to the trump administrations efforts to defund planned parenthood and restrict access to abortions, contraceptives and family planning services.	46	In congress, ive stood up to the trump administrations efforts to defund planned parenthood and reduce access to abortions, contraceptives and family planning services. I don't have right to abortion; I just care for my own life.

Figure 4.2: shows the change of tone in the paraphrase text

It shows the limitation of the pre-trained model, as it's stated the document that it's trained on previous unlabeled data with zero supervision, albeit that the model was able to capture the essence of the issue here. However, the part that captured the message in the original text was also very much present in the original text, and the model only added some extra noise to accomplish the paraphrasing -- something that could use a little more investigation in the future.

5. Conclusion

This paper introduced a heuristic approach for labeling and categorizing political/rhetoric text messages in order to understand the political scene and how it can correlate with other fields *Snorkel* and *Google T5* text-to-text attention based transformer model was introduced for further experimentation. The limitation of the project is its direction, scope, and evaluation. From diving into the project, however, I was able to understand the basic downstream tasks of building and putting together even a simple system and the difficulties. In the future, the paper will try to experiment more on pulling other data of the political candidates (their facebook, Twitter accounts) to draw a better

connection. By the end, the system is expected to output the connection of their financial reports and their actual political stances, in hope that people could be rid of the hassle to go through the FEC database to have to learn about the politicians and super PACs that they care about, which could be puzzling at times.

References

[1]Magee, C., 2000. Why Do Political Action Committees Give Money to Candidates? Campaign Contributions, Policy Choices, and Election Outcomes. *SSRN Electronic Journal,*.

[2]Ratner, A., Bach, S., Ehrenberg, H., Fries, J., Wu, S. and Ré, C., 2017. Snorkel. *Proceedings of the VLDB Endowment*, 11(3), pp.269-282.