Public Opinion Analysis on Politics

Prediction on Selection by NYT's Editors

Junqian Zhang - 944556

[1] University of Milan
[2] Data Science and Economics

Abstract. The Public comment the articles of New York Times to present their opinions, but not all the comments will be selected by the editors. This project analyzes the public comments on Politics to get clues to what New York Times considers worth promoting by aspect-based method and classification.

Keywords: Opinion Analysis · Aspect Extraction · Sentiment Orientation · Text Classification.

1 Introduction

The editors of New York Times will not pick all the comments from the public. The aim of this project is to predict whether a comment can get the editors' selection under the topic of *Politics*. This work wants to explore the relation between the Public's opinion and editors' sentiment. Opinion analysis is an active research field, and thanks to great many researchers' work, plenty of methods for opinion analysis have been proposed. The methodology of the work is mainly generated from *Sentiment Analysis and Opinion Mining* (Bing Liu, 2012)[1]. It is aspect-based analysis and tries to use the sentiment orientation of comments on different aspects to perform the binary classification.

2 Problem Statement and Methodology

The main challenge in this work is to find the relation between the sentiment of one group - the Public and the sentiment of the other group - the editors. The sentiment of the Public is represented by the sentiment orientation matrix, and the sentiment of the editors are represented by whether they pick a comment. The analysis regards the frequent nouns or noun phrases as aspects, which are all explicit aspects and detects the sentiment orientation of comments on these aspects to build sentiment orientation matrix.

The *opinion* of the Public discussed in this project is a quadruple,

[1] Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012

$$(e_i, a_{ij}, s_{ijk}, h_k)$$

where e_i is the topic of the comment, a_{ij} is an aspect of e_i, s_{ijk} is the sentiment orientation on aspect a_{ij} of topic e_i, h_k is the opinion holder. In this paper, the topic e_i is fixed at *Politics*.

This work tries to solve the problem by performing classification using the sentiment orientation matrix. The whole procedure is implemented as the following five steps.

2.1 Text Cleaning

The dataset used for analysis is a handful of comments. For the following steps, all the comments are cleaned based on sentence-level. All the sentences are tokenized and words in the sentences are executed lemmatization to get uni-grams by **Part-of-Speech** Tagging. However, at the same time, stop words are not abandoned.

2.2 Aspect Extraction

Under the *Politics* topics, people are discussing different sub-topics, which are called as *aspects* in this work, such as tax. This project focus only on explicit aspects based on frequency which are nouns and noun phrases appearing in the comment body. In this case, all the noun phrases are assumed to be composed of two words, i.e., all the noun phrase are bi-grams.

Based on the uni-grams, bi-grams are built at sentence-level to find candidate noun phrases. This project assumes that the candidate noun phrases should be in the pattern of two nouns, or stop word plus noun or adjective plus noun. So, to efficiently get the bi-grams, bi-grams without nouns inside are abandoned. To obtain the valid noun phrases and rank them, I calculate Pointwise Mutual Information (**PMI**) scores for all the bi-grams. PMI values take into account of the correlation between the two words inside the noun phrase, avoiding the case where the noun phrases in fact are the aspects of the aspects. 50 bi-grams with highest PMI scores are selected but the meaningless ones among them are abandoned.

Except these selected noun phrases, all the other bi-grams are splitted into uni-grams again and corpus is recleaned the corpus by removing the stop words. With comments as documents, I calculate the term frequency–inverse document frequency (**TF-IDF**) for all the nouns and candidate noun phrases. 20 features with highest TF-IDF values are selected as aspects.

2.3 Aspect Categorization

The aspects found from last step are the explicit aspects for this analysis, but these aspects have other expressions because different people may have different describing habits. So the target of this step is to group aspect expressions into *aspect categories*.

Assuming that aspects expressions who belong to the same category, have the same context, it is necessary to compare the context of words and phrases. A **Word2Vec model** is trained to represent all the uni-grams and bi-grams as vectors, and thus it is possible to compare the context between each other. These vectors can be seen as a description of the context of each element in the vocabulary. **Skip-gram model** is used to get the input for the word embedding representations. Through skip-gram, all the sentences in all the comments are organized into sequences. The length of context window is 3 and each gram concerns the 2 grams surrounded as neighbours.

Since all the words and phrases are in a numeric form, the cosine similarity distance between them can be calculated to see the difference of their contexts. With the word embedding model, for each aspect category extracted in the last step, I take the top 10 words or phrases that are closest to it, which are regarded as different aspect expressions. But some expression may be not nouns or noun phrases, so these are removed from the category.

2.4 Sentiment Orientation

This paper uses the *sentiment orientation matrix* as a tool to represent the opinions from the Public. Rows of sentiment orientation matrix represent comments, and columns represent aspects. The values of matrix are among **-1**, **0**, and **1**, representing *positive*, *neutral* and *negative* respectively. Since there are no labels on polarity, unsupervised learning method is adopted.

The sentences, who contains at least one aspect, are called *opinion sentences* here. For each comment, if it does not contain any opinion sentence on one aspect, then **0** is assigned to it on that aspect, regarding as *neutral*. For each aspect, all the comments with opinion sentences on this aspect are collected and these comments abandon the sentences which do not contain this aspect. Since there is no sentiment orientation label and comments show only two different sentiment orientations, *positive* and *negative*, **K-Means** is implemented to cluster the comments into 2 groups.

The comments from **Bag-of-Word** model is used to represent the comments by **TF-IDF** approach. The two comments which are closest to centroid of each cluster, are called *center comments*. The two center comments represent two sentiment polarities of the aspect. I check the text of center comments, and label them *positive* and *negative* manually. Then the comments in the same cluster are labeled after the center comment. In this way, **-1** is assigned to comments which are labeled with negative on one aspect, and otherwise **1** is assigned to comments which are labeled with positive.

2.5 Text Classification

The final step is to build the relation between the sentiment orientation matrix and editor's selection by classification. This step uses **Random Forest** and **Logistic Regression** as tool.

3 Experiment Results

3.1 Data Description

The data analyzed in the project are the comments made on the articles published in New York Times in April 2017 under the section name of *Politics* from Kaggle[1]. The data contain 38,381 comments from 51 articles. The features of data used for analysis are the followings:

- `sectionName` describes the topic of the articles commented, only *Politics* are selectd in this analysis
- `commentBody` stores the contents of the comment
- `editorsSelection` describes whether a comment is picked by editors, and is the target variable, and **0** for not picked, **1** for picked

3.2 Results and Evaluation

After implementing all the procedures introduced above, I successfully get 14 aspect categories and their corresponding aspects expressions, as listed in **Table 1**. **Figure 1** shows the sentiment orientation distribution of data on different aspects, the distribution of target variable.

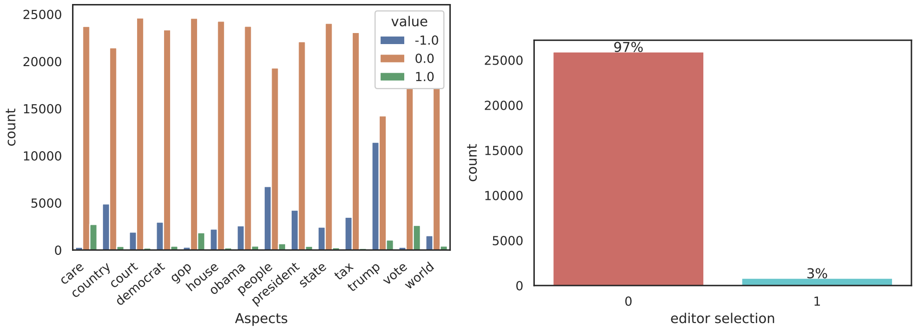

Fig. 1. Distribution of Input variables and Target variable

[1] Data source: https://www.kaggle.com/aashita/nyt-comments

Table 1. Aspect category and its aspect expressions (Ordered by first letter)

Aspect Category	Aspect expression
care	care, healthcare, health care, health
country	country, nation, usa, america, planet, citizenry, united state
court	court, scotus, supreme court, supreme, bench, sc, illegitimate, overturn, scalia
democrat	democrat, dems, democrats', dem, legislator, runoff
gop	gop, the gop, republicans, rep, caucus, faction, reckon, spine
house	house, occupant, visitor, wh, whitewater, white house, chairman, invitation, jason
obama	obama, gall, obamas, president obama, predecessor, bush, stonewalling, successor
people	people, folk, citizen, constituent, illegals, americans, person, woman, benefactor
president	president, potus, duly, pres, jimmy, presidency, sociopath, occupant, businessman
state	state, governor, assistant, county, resident, district, georgia, alabama, indiana, catholic
tax	tax, amt, monies, income tax, earnings, tax rate, adjust, loophole
trump	trump, djt, donald, donnie, trumps, dt, imbecile, apologist, sycophant
vote	vote, runoff, stein, voting, ballot, disenfranchise, reelect, slim, turnout, repub
world	world, region, the world, eu, sordid, th, isolate, beacon

The comments with no opinion sentence are abandon, **26,717** are used for classification task. And **25,907** of them are **not** selected by editors, while **810** are selected. Due to the fact of *ubiased* labels, to prevent overfitting and bad performance on *True Positive*, this paper uses **Adaptive Synthetic Sampling** (ADASYN) to get a more balanced dataset. After adjustment, the data therefore is transformed into **25,907** not selected and **26,092** selected.

The classification is implemented by Random Forest and Logistic Regression on numeric input and categorical input (values of sentiment orientation transferred to dummy and 0 as baseline) respectively. The evaluation strategy to evaluate the performance of the model is 10-fold **Cross Validation**. The classification report of four classifier model are presented in **Table 2-7**.

We can easily find that random forest on numeric variable has the best accuracy performance, and good result on precision and recall, while logistic regression's performance is not good, only a bit better than purely random classifier. However, the classifiers with category input both have high true positive.

Table 2. Comparison of the results

No.	Description	CV Score	Test Score
Model a	Random Forest (Numeric)	0.750	0.747
Model b	Logistic Regression (Numeric)	0.562	0.561
Model c	Random Forest (Categorical)	0.676	0.672
Model d	Logistic Regression (Categorical)	0.592	0.600

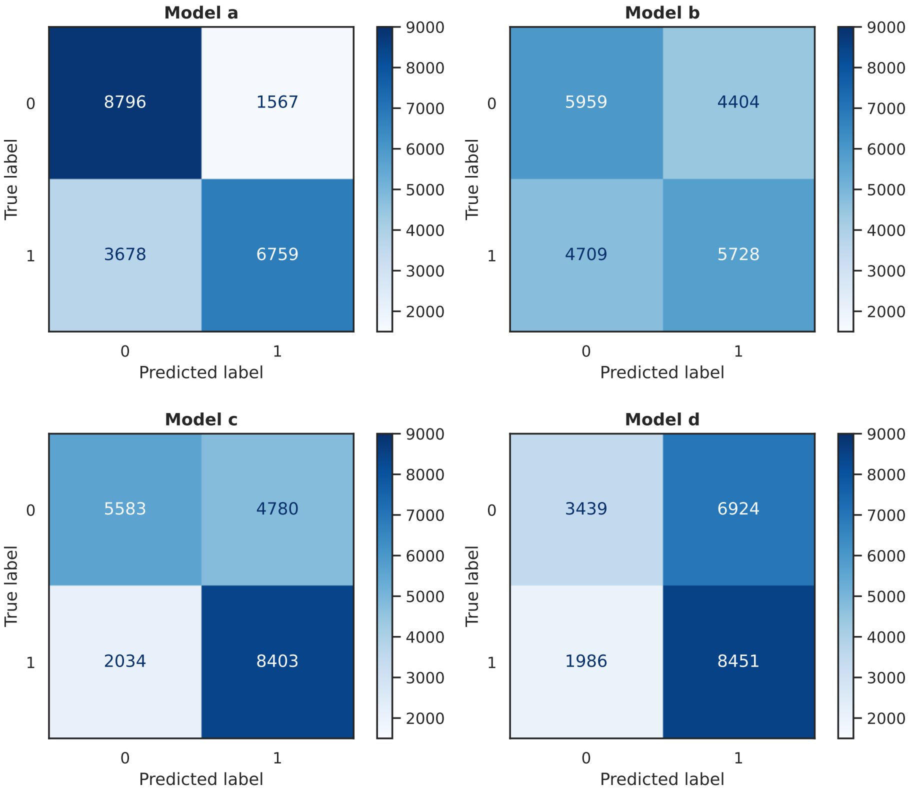

Fig. 2. Confusion Matrix

Table 3. Classification Report of Model a

Label	Precision	Recall	F1-score	Support
0	0.70	0.85	0.77	10363
1	0.81	0.64	0.72	10437

Table 4. Classification Report of Model b

Label	Precision	Recall	F1-score	Support
0	0.56	0.56	0.56	10363
1	0.56	0.56	0.56	10437

4 Concluding Remarks

From the result of random forest on numeric variable, *Accuracy*, *Precision*, *Recall* and *F1-Score* are not low for the classification task, all higher than 0.6, meaning that the sentiment orientation from the Public on different aspects has correlation with editor's selection decision and the aspects extracted do influence the selection by editors. From the confusion matrix, we can easily find that most of error come from *False Negative*. But the classifiers on categorical input have good performance on *False Negative*. This phenomenon means that part of sentiment polarities are wrong while the clustering has good performance.

To improve the result and find more clues on the relation, there is a lot to do for the future work. *Firstly*, for sentiment orientation judgement, the number of labels for sentiment orientation is small and thus it brings uncertain to the analysis. On the one hand, to be more precise on the polarity, other information retrieval tools can be exploited. In the comments under *Politics*, not many sentiment words, but great amount of Sarcasm and comparative words appear in the context. Other methods other than Bag-of-Word model should be used to compare the emotion of the context. On the other hand, *positive* and *negative* are not sufficient to describe people's opinion on an aspect of *Politics*. Multi-classification can be implemented. *Secondly*, this project does not concern any implicit aspects. Part of the comments are abandoned because they do not contain aspects extracted. Topic model or sentiment words for detecting implicit aspects can be implemented to make use of more comments. Thirdly, this work does not reveal what the editors sentiment orientation on different aspects is. It is necessary to explore this connection because it is the direct reason why they selected a comment. What's more, for better classification models for this problem, other classifiers or ensemble methods ought to be exploited.

Table 5. Classification Report of Model c

Label	Precision	Recall	F1-score	Support
0	0.75	0.51	0.61	10363
1	0.63	0.83	0.72	10437

Table 6. Classification Report of Model d

Label	Precision	Recall	F1-score	Support
0	0.63	0.47	0.54	10363
1	0.58	0.73	0.65	10437

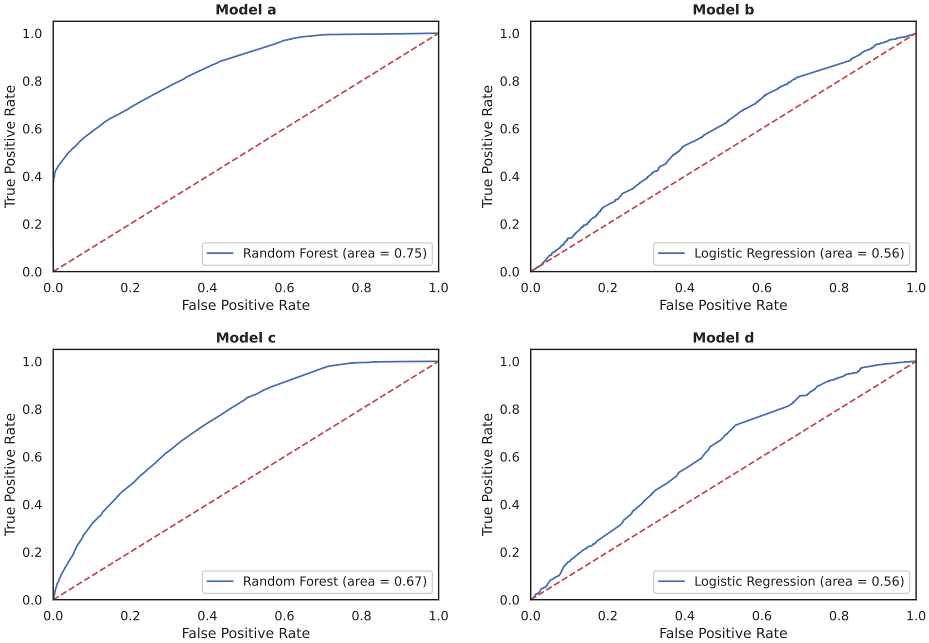

Fig. 3. Receiver Operating Characteristic

References

1. Mining Hu and Bing Liu, ”Mining and summarizing customer reviews”, *Proceedings of the 10th ACM SIGKDD International conference on knowledge discovery and data mining*, 2004
2. Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012