Opinion Analysis on Business
Selection Prediction of NYT's Editors

Aria Chen

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 Business to get clues to what New York Times considers worth promoting. The main purpose is to figure out the sentiment orientation of the extracted topics from New York Times so that it can be a regressor for future investment. The project extracts the explicit aspects of the comments and detects the sentiment orientation on different aspects to predict the business opinions of editors by classifying whether a comment will get picked by editors.

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

1 Introduction

With the development of Nature Language Processing techniques, it is possible to explore the sentiment in the market and use it for investment prediction. In this project, I explore the sentiment of New York Times on Business by analyzing its editors' selection on different topics extracted from the comments. The hypothesis is that the editors' selection on comments reflects their opinions under the topic of *Business*, so the editors select the comments which are aligned to their opinions and therefore can reflect the attitude of New York Times and the way they influence the Public and investment market.

The methodology of the project is mainly aspect-based analysis and tries to use the sentiment orientation of comments on different aspects to perform the binary classification. The results prove that the hypothesis works.

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,

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

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 *Business*.

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. **Fig. 1** gives an overview of the analysis design.

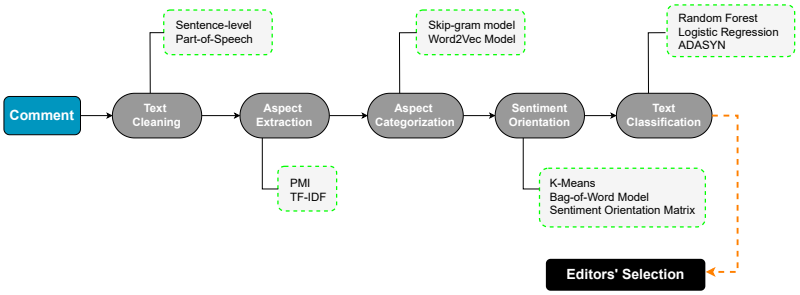

Fig. 1: Project Design

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 *Business* topics, people are discussing different sub-topics, which are called as *aspects* in this work, such as car. 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 March 2018 under the section name of *Business* from Kaggle[1]. The data contain 20,715 comments. The features of data used for analysis are the followings:

- `newDesk` describes the topic of the articles commented, only *Business* 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 selected, **1** for selected

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

3.2 Results and Evaluation

After implementing all the procedures introduced above, I successfully get 9 aspect categories and their corresponding aspect expressions, as listed in **Table 1**. **Fig.2** shows the sentiment orientation distribution of data on different aspects, the distribution of the target variable.

Aspect Category	Aspect Expression
trump	potus, dt, djt, donald, rhetoric, trumps, obama, resignation, thug, dennison, president, david, donald trump, presidency, traitor, administration, bolton, barack
facebook	fb, platform, snapchat, linkedin, yahoo, instagram, gmail, whatsapp, app, apps
world	mission, civilization, jeopardize, wallow, the world, mud, planet, unleash
medium	social medium, shine, web, commentary, addict, habit, twitter, disseminate
trade	trading, nafta, currency, wto, reign, resolution, procedure, negotiation, wwii, trade war, ww, afghanistan, loser, prepare, missile, iran, loom, peace, war
tariff	tarriffs, duty, restriction, davidson, retaliation, propose, sanction
car	vehicle, road, autopilot, driver, motor, mode, roadway, uber
pay	earn, sue, span, jack, sacrifice, owe, dividend, payroll, bonus
country	nation, britain, africa, europe, italy, quota, alliance, pacific

Table 1: Aspects of Business

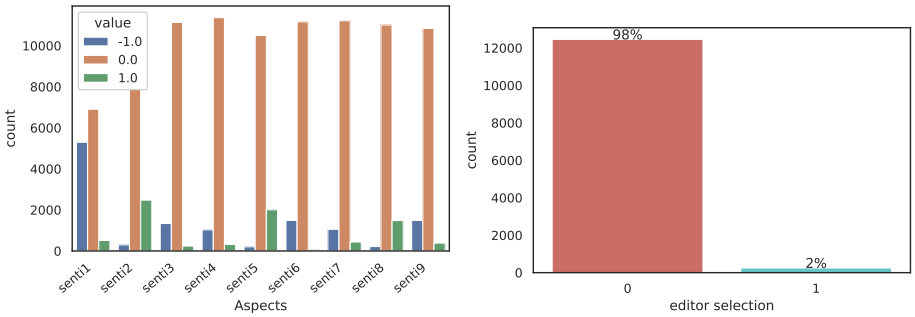

Fig. 2: Distribution of Input variables and Target variable

The comments with no opinion sentence are abandon, **26,717** are used for classification task. And **12,710** of them are **not** selected by editors, while **254** are

selected. Due to the fact of *unbiased* 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 **12,456** not selected and **12,534** selected.

The classification is implemented by Random Forest and Logistic Regression on category features. The evaluation strategy to evaluate the performance of the model is 10-fold **Cross Validation**. The classification performance of two classifier models are presented in **Table 2** & **Table 3**.

We can easily find that random forest has the best accuracy performance, and better results on precision and recall, while logistic regression's performance is not good, only a bit better than purely random classifier.

4 Concluding Remarks

From the result of random forest, *Accuracy*, *Precision*, *Recall* and *F1-Score* are not that low, and from the confusion matrix, we can easily find that most of the errors come from *False Negative*. Low *False Positive* tells that the topics extracted are meaningful since random forest can work by means of the topics as features. Although the result of the classification does not reach perfect performance, it is enough to prove that we can extract the sentiment of New York Times from their selection on Public comments. For the next step, we need to solve the problem of high *False Negative* and hence increase the overall high performance.

Table 2: Comparison of the results

No.	Description	CV Score	Test Score
Model a	Random Forest	0.70	0.69
Model b	Logistic Regression	0.58	0.56

Table 3: Classification Report of models

Model	Label	Precision	Recall	F1-score
Model a	0	0.65	0.82	0.72
	1	0.76	0.56	0.64
Model b	0	0.57	0.51	0.54
	1	0.56	0.61	0.59

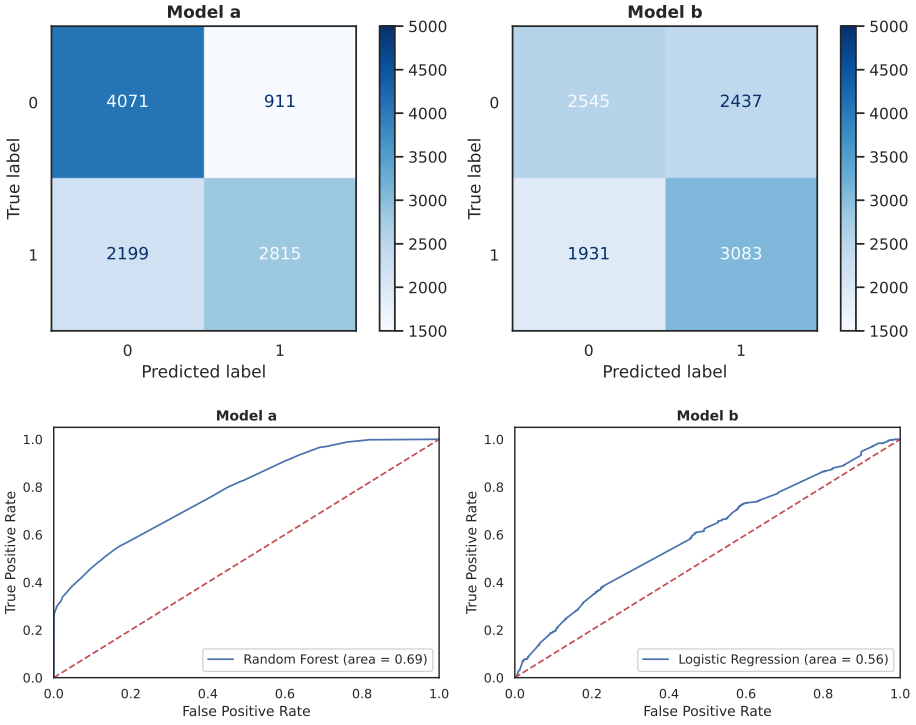

Fig. 3: Performance Summary