Policy Violation Detection through Customer Reviews

Ghodratollah Aalipour
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14623

ABSTRACT

Natural language processing, or NLP as a branch of artificial intelligence (AI), has found many important applications in human daily life. In particular, it has improved ways that computers and humans interact. In this report, we investigate some connections between NLP and business by analyzing a dataset of 1525 negative reviews given by customers of low-rated McDonald's fast-food restaurants in the United States. The reviews are labeled by one or several labels from eight categories. We build a model to classify each review to the most appropriate category. At the end, we evaluate our model performance by considering its accuracy and confusion matrix.

Introduction

Human language, an old and rich form of communication carries a lot of information that can be investigated in several aspects. So extracting, modeling and exploiting this great amount of information has made NLP a vital bridge between human linguistic information and digital data. But there are subtleties in this path. NLP is challenging because of the ambiguity, and the changing shape of the data. Ambiguity is a consequence of the several meaning a word can have and so it is difficult to identity the targeted meaning without having enough knowledge. The changing shape is due to several factors including new technologies. The new tools and technologies bring new types of structures and shortened phrases that become popular among users. For instance, with the rise of Twitter, new words and phrases started appearing in our formal languages. So we have to update our NLP models over time.

Under the light of recent advances in computers' power of computing, having a good data for feeding a machine learning model plays a critical role in every industry. Modern types of datasets don't limit to tabular datasets of observation with explicitly determined features. Instead, we usually have unstructured data awaiting for preprocessing and feature selection. Among these types of data, the linguistic data forms a significant part and carries a remarkable amount of information. So taking insight out of these written texts would be very beneficial to understating of what people in the society talk about. The society may have various groups of people such as customers of a company, twitter users, etc. The first step toward analyzing text data would be feature generation from the text. Always having a good set of features is one of the most important machine learning challenges.

There are some basic and intuitive approaches for feature generation. Presence of some words such as nice, happy, fun, etc in a positive/negative classification problem would be the baseline for the feature generation. But a good set of features does not limit to these features only. We still can find more complicated features such as POS tagging, relation identification, etc. Also statistical features such as frequencies and likelihood can be applied to add to the early set of features. The author of this report found feature extraction as one of the most important parts of NLP.

Contribution of NLP to Business

NLP is used to let computers understand human language, extract information, answer questions or complete other tasks. This interaction between computers and human beings is categorized based on the task that needs to be accomplished. NLP has been found very useful in business for interaction with customers in a large scale. In order to better understand customer preferences, many companies now analyze customer call recordings and reviews. In the rest of the paper, we see how NLP can contribute to business. We first start with some contributions of NLP to business (see also [2]).

- Getting feedback from customer through their opinions and reviews
- Detect customers who like/hate company's products and tune the recommendation systems accordingly
- Improve products receiving most complaints from customers
- Detect unhappy customers and help them be happy
- Building strong relationships with your customers, vendors and suppliers
- Increase sales buy tuning the products to the customer preferences

There are also several startups which have brought these plans to applications. In [3] and [4], there are extended lists and descriptions of some of these companies and what they do. We briefly review them here:

Expect Labs: NLP in Voice Recognition. This company was founded to enable other vendors create intelligent voice-driven interfaces for any app or device. The company is building cutting-edge NLP technologies to understand speech and also answer questions.

SwiftKey: NLP in Text Prediction. This is an innovative startup for text prediction technology and aims to significantly improve the accuracy, speed and fluency. NLP techniques are exploited to predict favorite words, phrases and emojis.

NetBase: NLP in Social Media Analysis. This startup employs social web data to apply sentiment analysis and detect the topic by reading and understanding millions of social media postings every day.

FiscalNote: NLP in Predicting Government Legislation. This company creates technologies for analyzing political and legal information by employing NLP models. FiscalNote's analytics platform aims to improve the way policy and legislative experts work by providing them with the data they need for critical decisions on market-moving issues. Their first product, named Prophecy, monitors and forecasts the outcome of state and federal legislation. The company claims that Prophecy has over 94% accuracy on forecasting policy outcomes.

Twiggle: NLP in ecommerce. This company develops ecommerce search engines with a higher degree of accuracy and performance than current available options. The system analyzes normal sentences for key phrases in product descriptions. The article available in [4] claims that as reported by the Wall Street Journal, Twiggle's query language tool has the biggest appeal for Alibaba.

Analysis of McDonald's Customer Reviews

In this project, we will process customer's negative review of low rated McDonald restaurants in United States. This dataset is available in [1] and consists of 1525 negative reviews of McDonald's fast-food restaurants in the United States. Based on each review, a label is assigned determining which business policy of company are violated. The reviews are labeled by one or a combination of the following categories:

- BadFood
- Cost
- Filthy
- MissingFood
- OrderProblem
- RudeService
- ScaryMcDs
- SlowService

Most of reviews are at least a paragraph long. Some of the reviews are presented below:

- *"Terrible customer service. I came in at 9:30pm and stood in front of the register and no one bothered to say anything or help me for 5 minutes. There was no one else waiting for their food inside either, just outside at the window. I left and went to Chickfila next door and was greeted before I was all the way inside. This McDonalds is also dirty, the floor was covered with dropped food. Obviously filled with surly and unhappy workers."*
- *"I'm not crazy about this McDonald's. This is primarily because they are so slow. My gosh what exactly is the hold up? It's FAST food people. Also, this morning, I guess the worker thought his mic was off, but it wasn't. I now know that he is trying to get as many hours as possible because he needs money BAD. Spread the word. Anyway, this location is on a little access road and you have to go back the way you came because there is no exit from it at the other end. It would have helped if there was one. So, in the end I think I'll avoid this location and find another. This should be easy as there is no shortage of Mickey D's in this piece."*

Objective: Design a model to take a (negative) review and determine which policy is violated. This is a linguistic classification problem.

Data Exploratory and Preprocessing

In building machine learning models, understanding the data, cleaning, processing and removing inconsistency are the main and primary steps. The data engineer should understand the data from different aspects such as what is the outcome, how the model works and what label would have the most misclassification. As mentioned earlier and also visible from the figure, some reviews are assigned with a combination of labels rather than one single label. After looking at the labels, we realize that there are 145 different labels in this dataset that are obtained by permutations of the original eight labels above. In the following figure, we have listed the top 15 frequent labels.

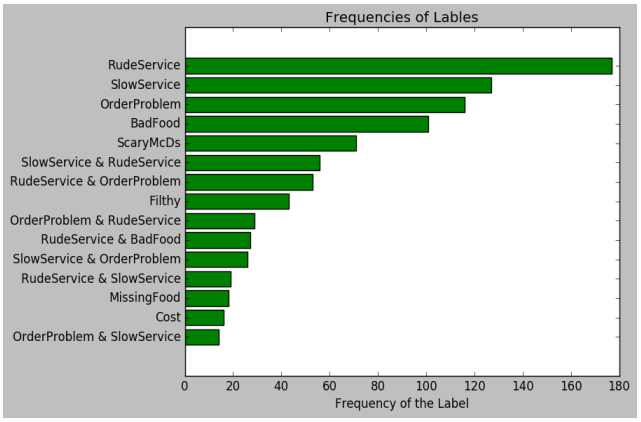

Figure 1: Frequencies of Raw Labels

Observation Labeled by NA/NAN

Besides the following issues, there are other data instances that are labeled by either na or nan. For instance, the following reviews are examples of reviews which are labeled "na" or "nan". After reading the reviews, some of the reviews assigned by "na" are actually positive, some meaningless and some are negative but the category is not determined. The followings are some of such reviews:

> *"I see im not the only one giving 1 star only because there is not a 25 star thats all i need to say"*
>
> *"i left the hilton late last night and i was really thirsty this was an easy inout option for me from sahara i did the drive through and was on my way with a diet coke in about 2 minutesthis is one of the remaining early style mcdonalds outlets and theres a classic chevrolet parallel to the drive through where you can take photoslooking through the windows i could see the interior is also the classic design ill have to go back and sit inside do i see a shamrock shake in my near future"*
>
> *"regular mcdonalds close to the highway which can be good and bad "*
>
> *"my order was fresh and prepared correctly they were friendly and moved the drivethru line quickly during the lunch rush very good"*
>
> *"nice mcdonalds inside but they have a remote that locks the bathroom ghetto much"*

These ambiguities motivated us to remove instances with missing values from our dataset, losing over 300 reviews.

Handling Multi-labels for a Single Review

As we pointed out earlier, some reviews are assigned with more than one label. That's because of several violated policies mentioned in the review. Here are some of such reviews:

Review	Label
"im not a huge mcds lover but ive been to better ones this is by far the worst one ive ever been too its filthy inside and if you get drive through they completely screw up your order every time the staff is terribly unfriendly and nobody seems to care"	"RudeService & OrderProblem & Filthy'
"this has to be one of the worst and slowest mcdonalds franchises there is cant figure out why my egg mcmuffin is always on a stale untoasted english muffin bought a chocolate shake today and threw it away"	"BadFood & SlowService"
"super slow service foods terrible like its been sitting and then reheated everything is out napkins iced tea"	"SlowService & MissingFood & BadFood"
"25 minutes in drive through line gunshots from the apartments behind worst mcdonalds ever"	"SlowService & ScaryMcDs

So the question is which one is the main one. As there is not any weight for labels, it is not easy to detect which one is the major violation. After a careful look at reviews and the labels, one possible approach would be assigning the first available label among a sequence of labels given for the review. But by taking this approach, there is inconsistency in the target value as "RudeService" will become the most frequent violation:

Label	**Frequency**
BadFood	255
Cost	28
Filthy	72
MissingFood	23
OrderProblem	192
RudeService	328
ScaryMcDs	99
SlowService	284

Continuing with this set of labels will force us to handle an imbalanced dataset. So we modify our approach above by reducing the number of classes and also taking the hierarchical connection between the labels in such a way that we get a uniform distribution among the labels. We merge some classes into other classes as follows: As "Filthy" has the lowest frequency, we change a multi-label into "Filthy" if the multi-label contains "Filthy". Otherwise, we merge "Cost" and "MissingFood" to "OrderProblem" and "ScarMCDs" to "Filthy" as shown by the following figure.

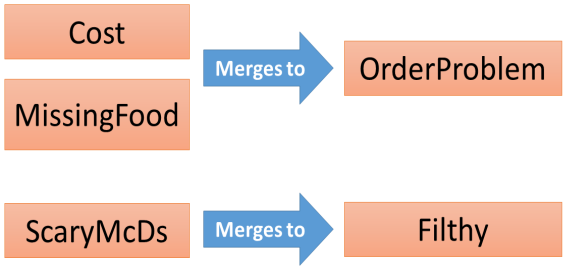

Figure 2: Merging Smaller Classes to a Large Class

Example:. In the table below, we have included some examples for this transformation.

Original Label	**Projected To**
RudeService & OrderProblem & Filthy	Filthy
SlowService & OrderProblem	SlowService
BadFood & RudeService & SlowService	BadFood
SlowService & MissingFood & BadFood	SlowService
na & ScaryMcDs	Filth

As a result of the transformation above, we get the following distribution for the new five labels. The transformation has moved the distribution closer to the uniform distribution.

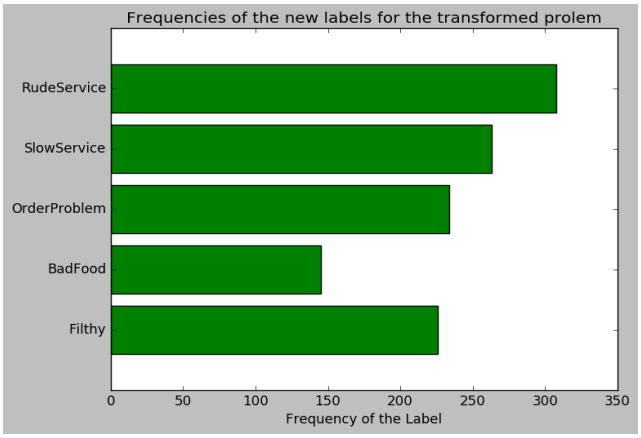

Figure 3: Distribution of New Labels

Another possible approach is to use these to generate additional instances. If we do not change the reviews and just copy it per given label, we would increase the size of the dataset but the trade-off is that the system will confuse about the correct labeling. So following this approach would require modifying the review each time we make a copy of that. The following may be proposed for modifications:

- Deletion of some sentences
- Using word net to replace some words and phrases with their synonym

Feature Extraction

After the previous data processing operations, we have a clean and consistency linguistic data. The next step toward building a model is feature extraction.

Some linguistically motivated features. Presence of some emotional words like tasty, nice, friendly, garbage, disappointing, dirty can give us some sense about the general sense of the review. The dependency relation between the concepts in the text can also be useful in this regard. For example the presence of some words like "noisy", "crowdy", "downtown" may direct us to conclude that the customer is complaining about the location. Comprehending the text is also another approach. Detection of the set of features that almost determine which policy is the subject of the review and making connection between the phrases would have a great impact on designing a classifier.

Lemmatizing, POS tagging, and n-grams. The primary methods for feature selection is making bag of words. For each category of interesting words, we found the top frequent ones and put them in our bag. After lemmatizing words, we filter over the noun, verbs, adjective, and adverbs to find the top frequent ones. By selecting a specific number of each, we build a binary dataset representing the existence of each of these words. After some initial machine learning models, we realize that applying bigrams and trigrams helps build a better performance model. So we also include bigram and trigram. That which portion of them we take remains as a problem for cross-validation dataset. To illustrate the features for our model, we include the top 20 bigrams and also the top 10 trigrams in the tables below:

Bigram	Freq	Bigram	Freq
('in', 'the')	431	('my', 'order')	212
('to', 'the')	344	('to', 'get')	207
('of', 'the')	308	('and', 'i')	205
('and', 'the')	248	('to', 'be')	198
('it', 'was')	236	('this', 'mcdonalds')	197
('i', 'was')	236	('on', 'the')	195
('drive', 'thru')	228	('for', 'a')	186
('the', 'drive')	227	('when', 'i')	184
('i', 'have')	225	('for', 'the')	165
('at', 'the')	220	('this', 'location')	164

Trigram	Freq
('the', 'drive', 'thru')	150
('in', 'front', 'of')	59
('one', 'of', 'the')	59
('the', 'drive', 'through')	57
('i', 'had', 'to')	48
('in', 'the', 'drive')	48
('to', 'this', 'mcdonalds')	43
('up', 'to', 'the')	42
('this', 'mcdonalds', 'is')	42
('to', 'the', 'window')	41

More Features by POS Tagging

As we can see from the tables above, the top bigrams and trigrams are somewhat unrelated to the reviews. Actually the later bigrams and trigrams are more informative than the top 20 ones. This motivates us to extract more features. A careful look at the labels such as "Filthy", "RudeService" encourage us to consider POS tags such as nouns, adjectives and adverbs. This is a part of our earlier discussion that some features are linguistically motivative. We use the pos_tag built–in function of nltk to extract these POSs.

POS = "NN"	Freq	POS = "VB"	Freq
'order'	731	'be'	4998
'food'	657	'have'	1465
'time'	448	'get'	973
'drive'	413	'go'	820
'service'	404	'do'	517
'place'	351	'take'	388
'dont'	334	'say'	358
'people'	289	'give'	344
'location'	279	'come'	305

We also find the top frequent adjective and adverbs. Each of these four POS has a long list of tokens. If we combine them and join the resulting list to the list of bigrams and trigrams, then we have a single but large list of features.

Hyper-parameter Tuning. Just like to the n-grams case, it remains as an interesting hyper–parameter adjustment to select which portion of each POS list. For our model, we set these parameter as $(100, 100, 130, 110, 100, 150)$, for bigrams, trigrams, nouns, verbs, adjectives and adverbs respectively. This combination forms 690 features. Using these features, we generate a binary dataset representing the presence of each in the sample review.

Classifiers

In building our models, we use three classifiers from scikitlearn and nltk: NaiveBayes, DecisionTree, and SVM. Gen-

erally the decision trees do not perform as well as the other two classifiers. So we only focus on these two. Since each time we have shuffle the data to split it to training and test sets, we get different results for accuracy. For ten split of the data into two training and test sets, we have the following accuracies

Model	Accuracy for 10 Random Testset	Median
Baseline	0.26 (Upon Fair Split)	26%
Naive Bayes	0.48, 0.47, 0.43, 0.48, 0.47, 0.43, 0.48, 0.51, 0.44, 0.53	47%
SVM	0.49, 0.52, 0.48, 0.58, 0.54, 0.55, 0.46, 0.55, 0.52, 0.51	52%
Keras NN	0.45, 0.43, 0.43, 0.48, 0.49, 0.49, 0.46, 0.44, 0.45, 0.43	44%

The last classifier is deep network made by Keras. Notice that the network is an ordinary multi-layer perceptron.

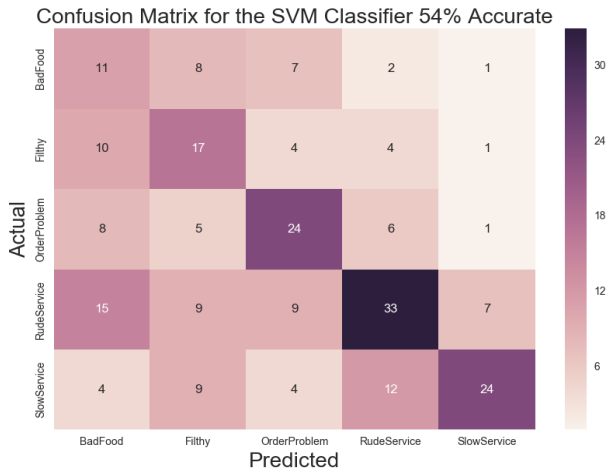

Figure 4: Confusion Matrix for SVM

Conclusion

In this project, we worked on a challenging dataset of negative reviews of McDonald restaurants. The data sets had several samples with a missing value for the label. Moreover, the reviews were labeled by a combination of eight available labels. To build an initial system, we first transformed the labels to convert the multi-label classification into a multi-class classification problem. We conducted this task by merging smaller classes into the larges ones resulting fewer number of labels. Unfortunately, we may miss some information in this transformation. Moreover, this transformation will confuse the system later as the corresponding review is also eligible for other labels. So it remains to determine a good transformation. The other approach would be preserving the multi-label classification problem but instead build a classifier to classify multi-labels.

Another challenging aspect of this dataset is its fairly small size. The author has been thinking to augmenting the datasets by applying some data-augmentation methods such as autoencoders. Generally, other architectures of neural networks such as LSTM or GRU are interesting to apply for this problem as the dataset is a sequential data and these types of networks are good candidates for the sequential data.

Word embedding seems to be one of the promising feature model. We apply an ordinary neural network through Keras and our features generated. But it would be interesting to feed word vectors to the Keras and see the performance.

How McDonald Restaurants Can Benefit?

We built a model to assign a label to each review. This model can be employed by the restaurant to improve the service and business in several ways. For instance, the model can be employed to filter over the customers and identify those who have had bas interaction with the employees. The business can also receive feedback on the product through this reviews. For example if many customers have been complaining about the quality of the food, then the business should either stop producing the product or improve its quality.

1. REFERENCES

[1] Data for Everyone, available at https://www.crowdflower.com/data-for-everyone/.

[2] Marco Lagi, Natural Language Processing – Business Applicationshttps, www.techemergence.com/natural-language-processing-business-applications, January 23, 2017.

[3] Gelareh Taghizadeh, Top 5 Companies Innovating With Natural Language Processing, http://blog.ventureradar.com/2015/09/22/top-5-companies-innovating-with-natural-language-processing/, September 22, 2015.

[4] 10 Emerging Startups In Natural Language Processing (NLP), https://www.fungglobalretailtech.com/research/10-emerging-startups-natural-language-processing-nlp.