Smart City Survey Annotations and Experiments

Introduction and Dataset

This project is about the annotations and model building for the smart city survey in San Antonio. The dataset is sample comments provided by the City of San Antonio from their smart city survey. Inside the dataset, the text of one comment is on each line along with the comment the question is associated with.

For this project, I first do annotations for the comments. I need to conceptualize and design a classification task for the comments. And some topic classes are recommended: Transportation systems, vehicles, and autonomy. Data governance and city data platforms and dashboards. Wireless communications and broadband applications. Public safety and security, and mission critical communications. Public utilities for energy, water, and waste management. I use total 11 topic classes for the annotation. In addition, I also need to give each comment either sentiment or opinion/suggestion. My study annotates whether each comment is specifying something negative, neutral, or positive with respect to San Antonio.

After received the comments annotations from our 'friend' group, I also get an agreement score between these two annotations. However, the score is low. Then I do the gold label carefully to ensure that each comment is labeled correctly.

Feature Engineering

Feature engineering is one of important steps in machine learning, which refers to a process of creating features with domain knowledge of data to make the machine learning algorithms work well. The more accurate information you provide, the better machine learning algorithms will be able to interpret the information. If we focus on the data in the beginning, we will get better results than focusing only on models. Feature engineering can help provide better data which in turn helps the model interpret it well and produce accurate results.

NLP is one branch of AI which studies how people interact with machines. To grasp the meaning of natural language, we have to understand how sentences are formed, how we express our ideas and thoughts with different words, signs, and other special characters. Moreover, we need to understand the real context of the sentences. The model will be able to interpret the sentences better if we can use these contexts as features and feed them to the models. There are some common features we can then extract from the sentences. For instance, the number of words, the number of capital words, the number of unique words, the number of punctuations, the number of stop words, the number of hashtags, average sentence length, and so on. We can define and extract the features according to our real dataset.

In my study, I use Count Vectorizer to extract features from the sample comments. It is a great tool provided by scikit-learn library in Python. We can use it to transform a given text into vector based on the number of times that each word appears in the entire text. This tool is really helpful when we have multiple such texts, through which we can convert each word in each text into vectors for further text analysis. Count Vectorizer builds a matrix in which every unique word is represented by a column of the matrix, and the value of each cell is the count of word in that particular text sample.

Model Building

logistic regression

For the experiment models, I first apply logistic regression. Regression or logistic regression is one of the most popular models used as algorithms for problems with binary classification with just 2 values, such as '0 or 1' (Kalaiselvi et al., 2014). In fact, logistic regression has been used in many fields, including the biological sciences. When the research goal is to category the data items into groups, the logistic regression is one of the good choices (Zhu et al., 2019). As in my study, my goal is to classify the topic and sentiment of the comment. We can apply logistic regression to find the best fit which is reasonable to describe the relationship between outcome variable and predictor variables.

Consider the following set (X, Y), where X is a matrix of values with i examples that represent input parameters and j features. Y is a vector with i examples which are the sample outcomes based on X. and the goal to train LRML is to predict which class the values belong to (Patil et al., 2021).

Dummy Classifier

This is a classifier model that doesn't find patterns in the data when it makes predictions. In essence, the default model exams what label is appears the most frequently in the training dataset and then makes predictions based on that label. This model is completely independent of the training data as the trends in the training dataset are ignored. It adopts one of the strategies to predict class labels. For instance, most frequent—the classifier always predict the most frequent class label in the training dataset. In our study, we use this strategy. In a word, dummy classifier is based on the idea that any analytic strategy for classification problems ought to be superior to a guessing approach.

Experiment Results Analysis

In my study, I mainly apply two models-logistic regression and dummy classifier. After I train the separate models, I get the precision, recall and F1 score for both of these two models. And the final results are as follows:

Experiment Results for ‘Topics’

	Logistic Regression	Dummy Classifier
Precision	0.4205	0.0155
Recall	0.3612	0.0909
F1	0.3566	0.0264

	Logistic Regression	Dummy Classifier
Precision	0.2300	0.2300
Recall	0.3333	0.3333
F1	0.2742	0.2722

Experiment Results for ‘Sentiment’

After comparing these results, we can see that the logistic regression model behaves better in modeling ‘topics’. But the score for modeling ‘sentiment’ in these two models has no big differences.

It is helpful to model against the benchmark while building machine learning models. There can be a point that we don’t have access to an obvious model. In this situation, a dummy model is a good choice because it applies a heuristic on the input data.

References

Kalaiselvi, C., & Nasira, G. M. (2014). A New Approach for Diagnosis of Diabetes and Prediction of Cancer Using ANFIS. *2014 World Congress on Computing and Communication Technologies*, 188–190. https://doi.org/10.1109/WCCCT.2014.66

Patil, V., & Ingle, D. R. (2021). Comparative Analysis of Different ML Classification Algorithms with Diabetes Prediction through Pima Indian Diabetics Dataset. *2021 International Conference on Intelligent Technologies (CONIT)*, 1–9. https://doi.org/10.1109/CONIT51480.2021.9498361

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*, 100179. https://doi.org/10.1016/j.imu.2019.100179