Sentiment analysis on Twitter data and predicting the election

author: Liang Chen

1 Motivation

With the development of social media, applications like Twitter and Facebook become an important part of our social life. Especially after Donald Trump was elected to be the president of the United State in 2016, Twitter is now a place to share people's political opinions, not only to just share their daily life. Traditional election polls use an old fashioned way to collect data and always have deviations on their samples, which cause inaccurate predictions especially since 2016 . Because of the Anonymity of Twitter, people might be more likely to express their real opinion on social media and the convenience of tweets also give everyone an equal chance to speak, so it is very interesting for us to know if tweets about the election can be used to predict the result. In this project, we will train a model for tweet sentiment analysis, this model should identify a given tweet is positive or negative, word2vec might be applied to convert tweet to matrices for machine learning, and Neural Network or other approaches may be also used to train this model.

2 Data Set

The dataset used to train the sentiment analysis model is provided by a Kaggle competition called Tweet Sentiment Extraction. The dataset consists of two files, a training dataset, and a test dataset. The training dataset is 27481x4, contains text ID, the text of a tweet, selected text which includes a word or phrase drawn from the tweet that encapsulates the provided sentiment, and a sentiment label.
Dataset Link: `https://www.kaggle.com/c/tweet-sentiment-extraction/overview/description`
We will perform a sentiment analysis on US Election 2020 tweets dataset using the model obtained from the above dataset. The dataset consists of 2 files, one for each candidate: tweets for Donald Trump and tweets for Joe Biden.
Dataset Link: `https://www.kaggle.com/manchunhui/us-election-2020-tweets`

3 Related Work

Numbers of techniques were used in sentiment analysis, including naive bayes, logistic regression, random forests, and support vector machines. In our work, in addition to these methods, we will also apply a LSTM sentiment analysis.
Sentiment Analysis Using Simplified Long Short-term Memory Recurrent Neural Networks
https://arxiv.org/abs/2005.03993
There are some similar work, Cram et al. [2017] use tweets to analyze the general election of UK, and Yaqub et al. [2020] use tweets to do sentiment analysis based on tweets' location.

4 Problem Formulation

This project aims to analyze sentiment on tweets data. Specifically, given a new piece of tweets, we would like to predict the sentiment of the tweets as either "positive" or "neutral" or "negative". Formally, we can define our problem as:

Given a set of training data, $\{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)\}$, where $\mathbf{x_i}$ represents the word vector for a single piece of tweet data, y represents the true label indicating if this piece of data has positive, neutral, or negative tendency, i.e., $y_i \in \{$"positive", "neutral","negative"$\}$. This problem is essentially a supervised machine learning problem.

More specifically, we use the pre-trained word2vec to transfer our tweets dataset to custom word matrix, and use it as input into a classification model which yields multi-class outputs. Different classification models are compared, e.g., random forest, gradient boosting, SVM, neural nets, LSTM, etc. We then apply the customized search function to tune hyperparameters of models in an automatical way. After we find the best classification model, we apply this model to another newly-released unlabeled dataset: US Election 2020 tweets dataset, in order to label the tweets.

Besides machine learning algorithms applied for this problem, data preprocessing and exploratory data analysis are also performed in order to get better insight of the dataset.

5 Methods

Listed methods may be applied to this project

- Naive Bayes(Baseline)
  Naive Bayes is the most general and traditional algorithm for classification
- Neural Network:
  Neural network takes training samples as input data and accordingly outputs the predicted labels. For multi-class classification, softmax function is used to output the probability for each class.
- LSTM:
  Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. It can be used for classification problem.
- Random Forest:
  We use random forest as an ensemble method here for classification by establishing a large number of decision trees and computing majority voting of the individual trees as the final output class.
- SVM
  An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.
- CNN
  In deep learning, convolutional neural network (aka CNN) is a class of deep neural networks. It can extract features from data. CNN has wide applications in image classification, medical image analysis, natural language processing, financial time series, etc.

6 Evaluation

- Random Forest:
  In random forest, we minimize the entropy:$-p_i \log_2 p_i$, where $i \in \{\text{“positive”}, \text{“neutral”}, \text{“negative”}\}$.
- Neural Network and LSTM:
  In the Vanilla Neural Net and LSTM, we use cross-entropy as our loss function:

$$\text{CrossEntropy Loss } = -\left(y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right) \log\left(1 - \hat{y}_i\right)\right)$$

- SVM:
  In SVM, we minimize the hinge loss.

$$\begin{aligned} \min_{\mathbf{w}, b, \xi \geq 0} \quad & \frac{1}{2}\|\mathbf{w}\|_p^p + C\|\xi\|_q^q \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n \end{aligned}$$

We aim to apply cross validation technique to tune hyperparameters of the models, as well as to use accuracy, F1 score, ROC curve to evaluate the models.

References

Laura Cram, Clare Llewellyn, Robin L. Hill, and Walid Magdy. UK general election 2017: a twitter analysis. *CoRR*, abs/1706.02271, 2017. URL `http://arxiv.org/abs/1706.02271`.

Ussama Yaqub, Nitesh Sharma, Rachit Pabreja, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. Location-based sentiment analyses and visualization of twitter election data. *Digit. Gov.: Res. Pract.*, 1 (2), April 2020. ISSN 2691-199X. doi: 10.1145/3339909. URL `https://doi.org/10.1145/3339909`.