Dan Burke

Dburke01@syr.edu

Homework 4

MB Naïve Bayes

05/22/2023

|  |  |
| --- | --- |
| Homework 5 – Option 2 | |
| **Introduction** | Sentiment analysis has a multitude of economic applications with text data both on the side of the consumer and that of the producer. This report attempts to address one application of sentiment analysis and lie detection within the scope of restaurant reviews.  This report intends to apply the Naïve Bayes approach to restaurant review text data sourced from course materials. This data possesses lie, sentiment, review and features.  Naïve Bayes is a probabilistic machine learning process utilizing the Bayes’ theorem. This process assumes that all features (columns) of the data set are conditionally independent. Within the scope of this report the features refer to induvial words; all individual words contained within the dataset. |
| **Analysis** | data preparation The crypto currency article has been sourced from NewsAPI.org. The data is fetched via an API (Application Programming Interface) call (http request), whose response is formatted list of news articles in the form of JSON (JavaScript Object Notation).  The API HTTP request, JSON format response required this effort to both create custom methods to retrieve and format the data. This was combined into one topic agnostic method, which accepts the starting date, sorting method, topic to query, API key and Boolean “save (or export if in virtualized environment) to file” parameters. This custom method retrieves, formats and leverages regular expressions to remove unwanted/irrelevant text from the raw data.  This custom data retrieval and data formatting method was the best approach as this effort required multiple topics to query from NewsAPI.org. Encapsulating this functionality into a custom method allows for reusability and reproducible results (in the form of formatting) when retrieving data from this source. This method also labels the data on a per topic basis (using the term to query parameter), as article data fetched from NewsAPI.org does not possess an explicitly defined label (nor should it as it is only a data source not a tool for conducting analysis).    Figure Data Retrieval and Preparation Custom Function    Figure Method Calls, Exporting to CSV File    Figure Data Preparation Method Resulting Dataframe  **Data Types**  The Multinominal (MB) Naïve Bayes approach requires all data to be numerical, this requires two steps. First to check the cleaned data frames column types as well as count vectorizing (establishing word occurrence counts).  Two columns required a change to a categorical data type; “lie” and “sentiment”. Once these data types have been properly configured and verified the CountVectorizer method of the SKLearn Python library was utilized.    Figure Mutation to Categorical Data Type  **Label Extraction**  MB Naïve Bayes models only allows for one “label” or “classifier”. This required the effort to create two data frame (though not necessary labels could be extracted and applied to models separately while using the same text data). Both “lie” and “sentiment” were stored within a “labels” variable.  With these separate data frames (lie and sentiment) it was now necessary to count vectorize the text data, in order to format it properly for generation of a MB Naïve Bayes model.  Once the text data intended for use with a model has been count vectorized it was then necessary to create the “test, train, split” of the data. This is separating the data into a training (model generation) and test (prediction) segments. A 80% training and 20% testing split was utilized.    Figure Test Train Split - Both Lie and Sentiment  **Fitting the Model**  Once the test, train, split has been created it was then required to fit the data to the model. X referring to the textual data, y referring to the label (classifier).  **Sentiment**    **Fitting the Model**  **Lie Detection**    . |
| **results** | technical results **Sentiment Analysis**    **Lie Detection**    **Confusion Matrices**  **Sentiment**    **Lie Detection** |
| **conclusions** | The Multinomial Naïve Bayes approached produced results of acceptable accuracy ~78.9% within sentiment analysis and undesirable accuracy ~57.8% for lie detection.  Though the Sentiment analysis model is, both the sentiment and lie detection models could not be integrated into a production environment due to accuracy.  Besides expanding the size of the dataset; a modification to the stop word list as well as adjustments of the test, train, split distribution may increase results. The use of an additional analysis method; VADER (Valence Aware Dictionary for Sentiment Reasoning) and the use of sentiment lexicons may give more insight to  . |