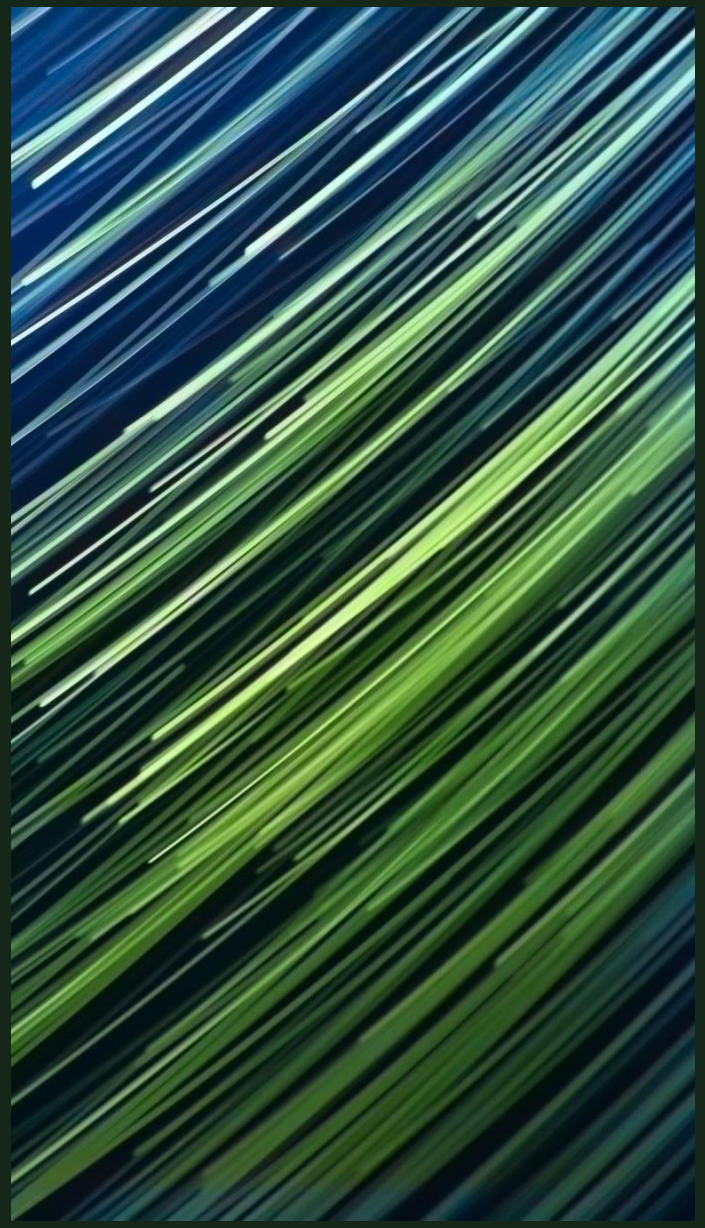


Binary Classification in the analysis of sentiments

Cristian KEVORCHIAN, Ph.D.

Faculty of mathematics and Computer Science

University of Bucharest

Text Mining

- Text mining is the process of extracting (non-trivial) information from a variety of textual forms with a high degree of accuracy.
- It entails the automatic detection and retrieval of previously unknown data from a variety of textual forms.
- Text mining is similar to data mining, however it works with text structures rather than data structure.
- The arrangement and structuring of data such that it can be subjected to both qualitative and quantitative analysis is one of the initial steps in text mining operations.

Sentiment Analysis

- Text mining has a subdomain called sentiment analysis.
- In order to identify, extract, and quantify emotional states and in this context of subjective information, the sentiment analysis is based on natural language processing, text analysis, computational linguistics, and neuroscience..
- Sentiment analysis is frequently used to assess textual structures such as reviews and survey replies, as well as a variety of online text messages, healthcare resources, and a variety of marketing communications...

Categorical Data

- Categorical data is a term used to describe variables that have a range of possible values.
- The number of available values is restricted to a small number of options.
- I/O variables must be numeric in ML models. In order for the model to evaluate the data, it must be categorical.
- A categorical variable (also known as a nominal variable) is a variable that consists of a finite number of discrete values with no link between them.

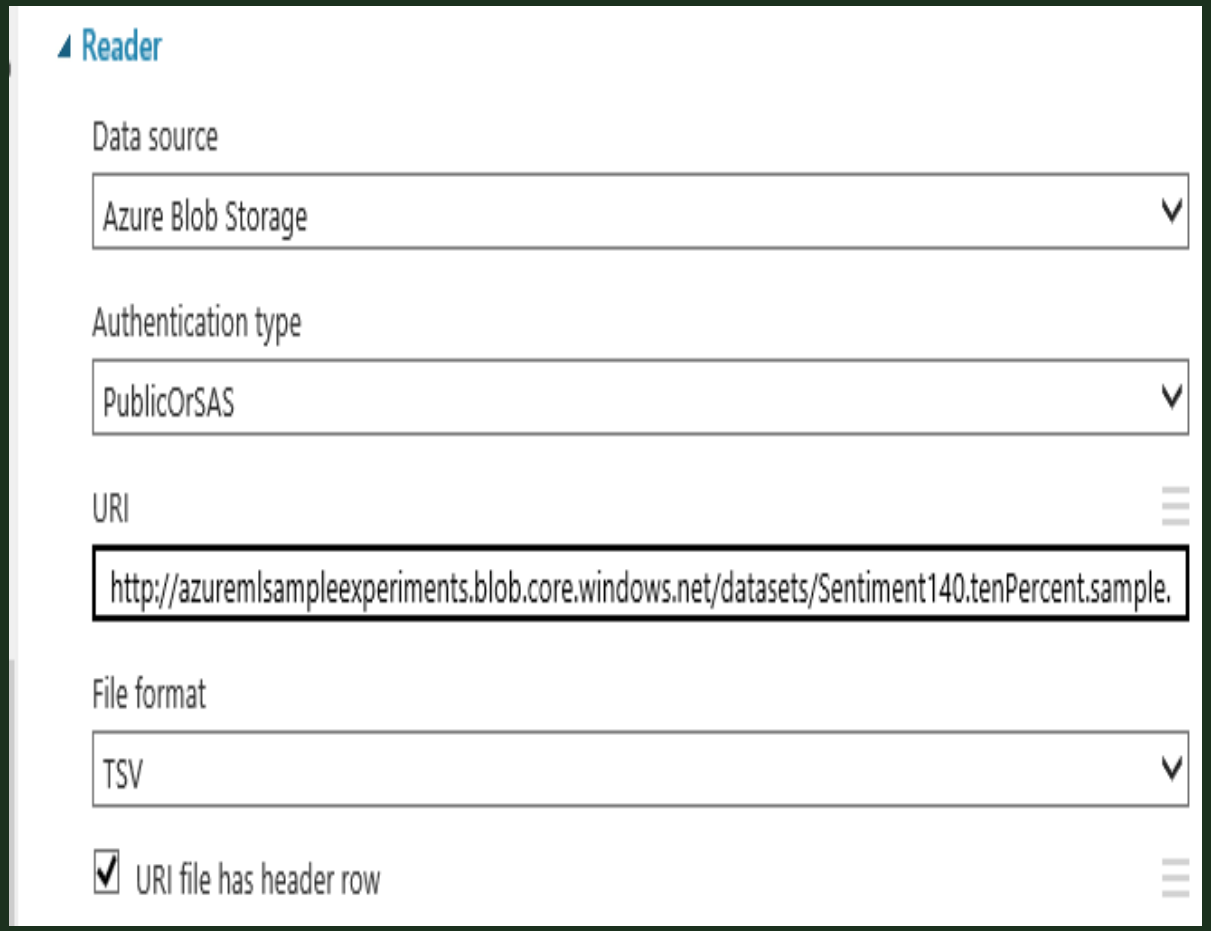
Ordinal Encoding and One-Hot Encoding

- **Ordinal Encoding** – A variable is made up of a finite number of discrete values that are arranged in a ranking order. Ex. Red – 1, Green – 2, Blue - 3
- b) One-Hot Encoding – Each state is encoded using a distinct bit of state in one-hot encoding. One-hot is named after the fact that only one bit is "hot" or TRUE at any given moment. For instance, a three-state one-hot encoded FSM would have state encodings of red –(1,0,0), green – (0,1,0) blue-(0,0,1)
- c) Dummy variable – A dummy variable is a variable with values of 0 and 1, with the values indicating whether something is present or not.where Red – (1,0) Green – (0,1) Blue –(0,0). C categorii necesita C-1 variabile dummy.

Feature Heshing

- Modulul Hashing de caracteristici în Azure Machine Learning Studio (clasic), este destinat transformarii unei secvente de text(în eng.) într-un set de caracteristici reprezentate prin numere întregi dupa care pe baza familiei de caracteristici hash prin intermediul unui algoritm de învățare automată se poate antrena un model de text mining.
- This approach can be used to encode variable-length text as equal-length vectors of numerical features. By replacing the string comparison with the hash value comparison, feature hash decreases the size of the data and makes it easier to locate the weights associated with the features..

Experiment Data Source



Data Volume:

- **160000 records**
- **Data Collections 10%**

Each instance in the initial data set has 6 fields:

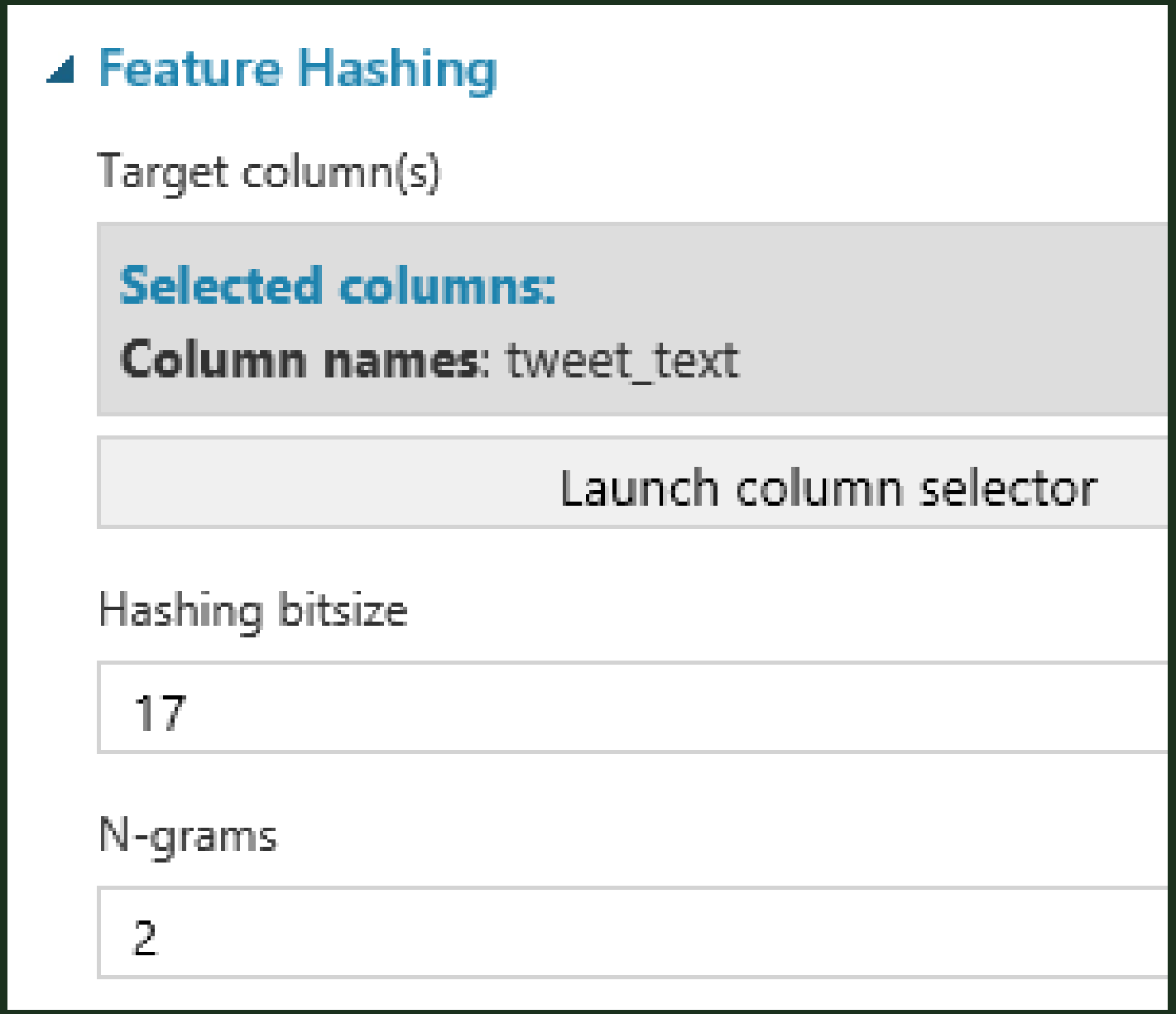
- sentiment_label - the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
- tweet_id - the id of the tweet
- time_stamp - the date of the tweet
- user_id - the user who posted the tweet
- tweet_text - the text of the tweet

Will use only two fields sentiment-label and tweet_text

Data Normalisation and Metadata

- Before unstructured text, such as a tweet, can be evaluated, it usually needs to be reprocessed.
- Will be used a R script to remove punctuation marks, special character and digits, and then performed case normalization.
- After cleaning the text, it changed the metadata of the text column using the Metadata Editor module, as shown below.
  - The text column has been designated as a non-categorical column.
  - The text column was likewise designated as a non-feature.The reason is that we want the learner process to ignore the source text and not use it as a feature when training the model

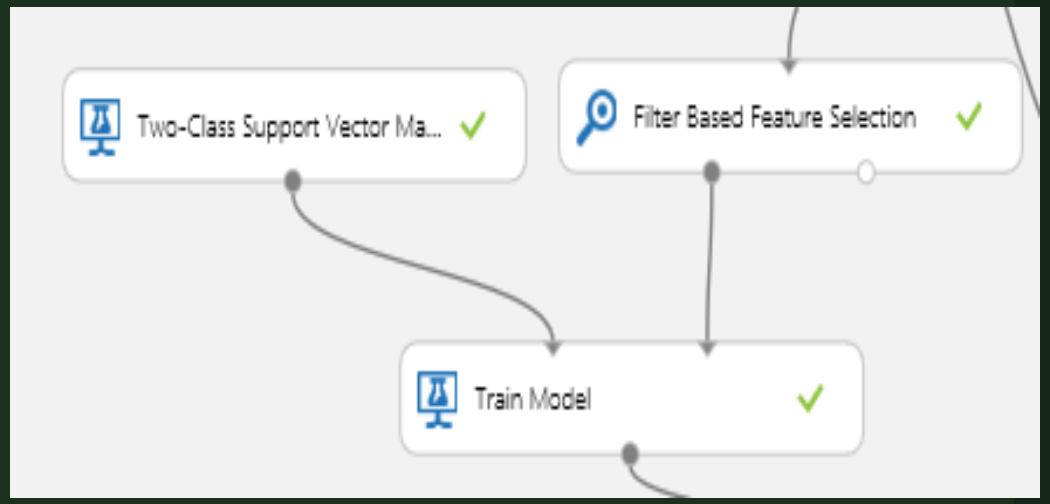
Hashing-ul Caracteristicilor

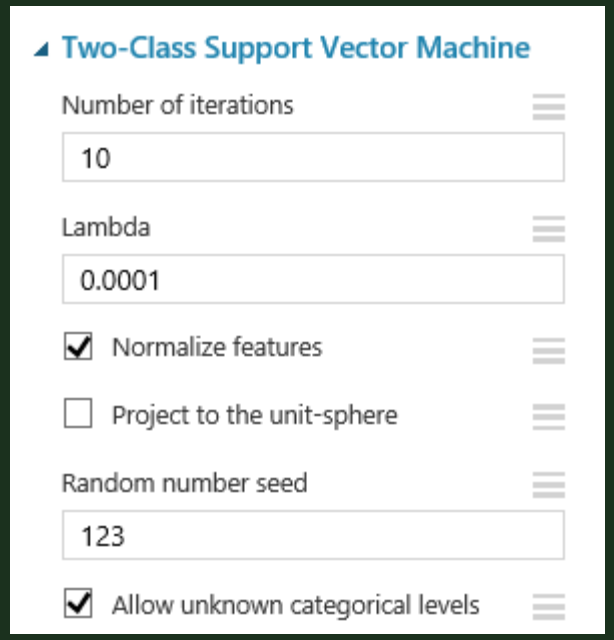


Feature Selections

- The classification complexity of a linear model is linear with respect to the number of features..
- **A text classification model can have too many features for a satisfactory solution, even with feature hashing.**
- **It selected a compact feature subset from the exhaustive array of extracted hashing features using the Filter Based Feature Selection module.**
- **The goal is to lower computing complexity while maintaining classification accuracy.**
- **Chi-squared score function to rank the hashing features in descending order, and returned the top 20,000 most relevant features with respect to the sentiment label, out of the 2^17 extracted features.**
- **Out of the 2^17 retrieved features, the Chi-squared score function was used to rank the hashing features in descending order, and the top 20,000 most important features with respect to the sentiment label were returned.**

Train Model



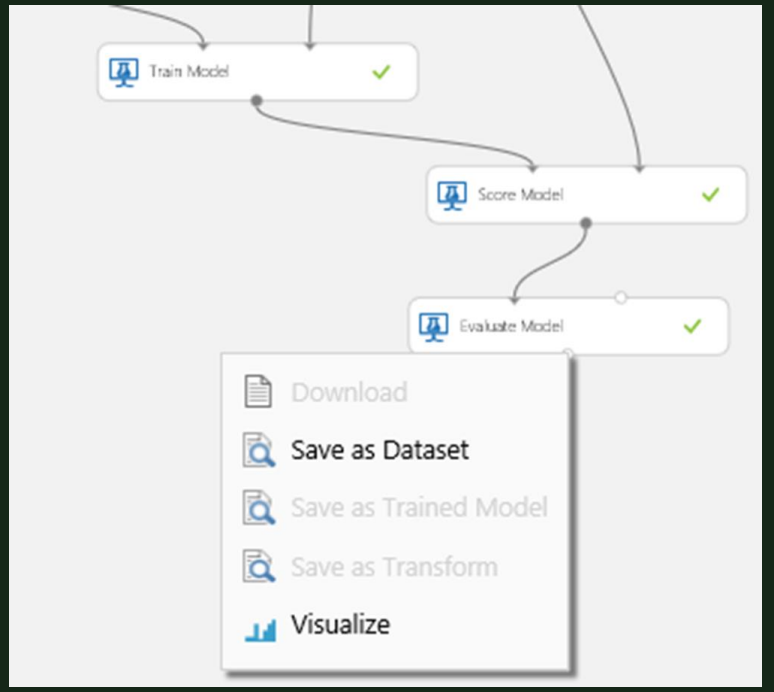


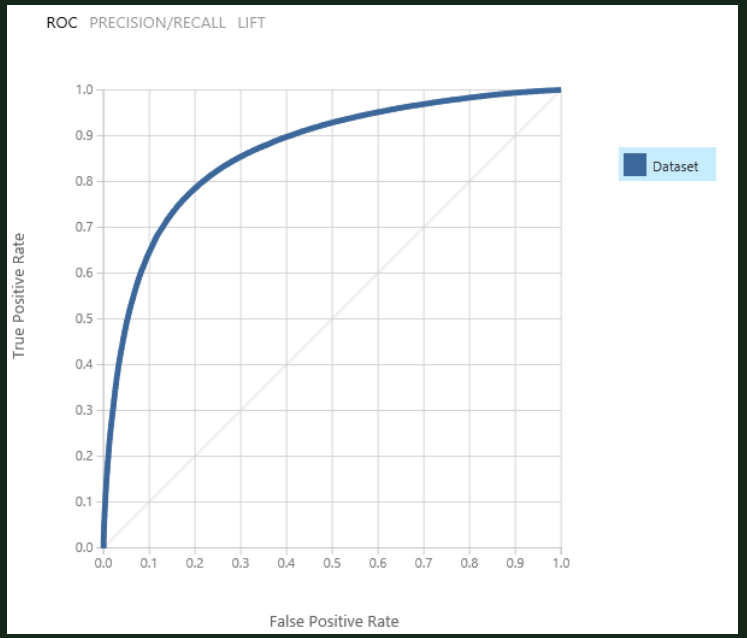
The text characteristics developed in the preceding steps (the training data) are connected to the *Train Model module to train the model. (we will use SVM).

- The output model and the test data set are connected to the Score Model module in order to score the tweets of the test set in order to evaluate the trained Support Vector Machine model's generalization capabilities on unseen data.
- As we illustrated below, connect the out predictions to the Evaluate Model module to acquire a number of performance evaluation metrics. Note that the results below are based on the model being trained on the entire Sentiment140 dataset. Please replace the 10% sample data related to the experiment with the entire data set to obtain the same results.

The Trained Model Evaluation

- To get the evaluation metrics (ROC, precision/recall, and lift) depicted in the following charts, use the **Evaluate Model** module.
- The metrics displayed here were obtained by training the algorithm on the entire Sentiment140 dataset. As a result, you should replace the 10% sample dataset with the entire data set to replicate these results.





Thank You,
for your attention!