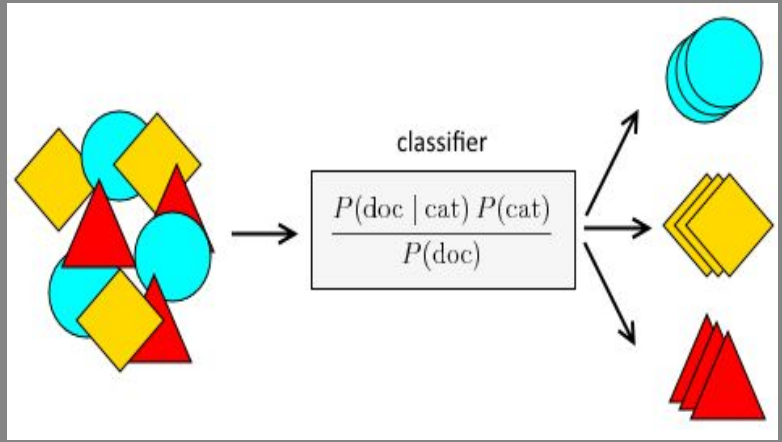
Chair Of Digital Libraries and
Web Information Systems

UNIVERSITÄT PASSAU

Twitter Text Classification



October 11, 2016

Dilip Madhu Kumar, Revathi Sadanand, Salman Khatri

Overview

1. Introduction
2. Implementation
3. Future Work
4. Conclusion

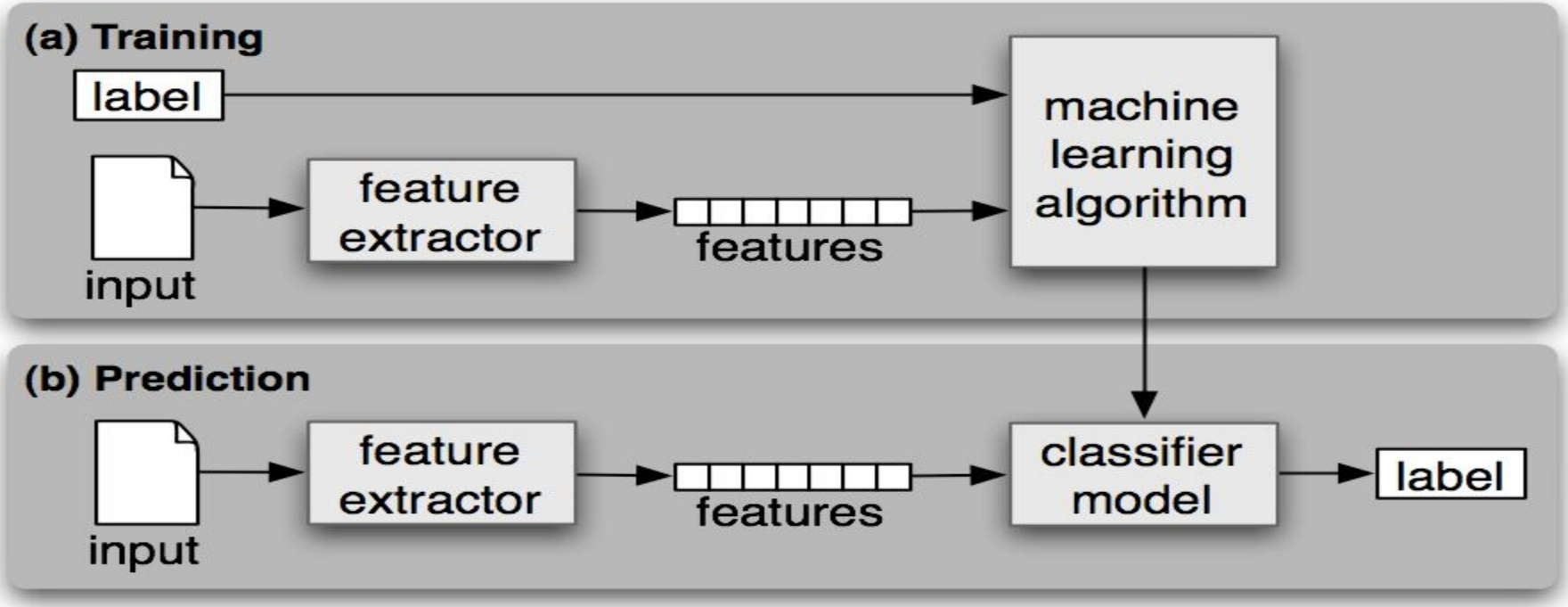
Introduction

- Text Classification
- Supervised Learning

Text Classification

The task of choosing the correct class label for a given input.

Supervised Learning



Implementation

Data Preparation and Cleaning

- The Twitter Political Corpus contains tweets that have been hand labeled for their topics, specifically, discussing politics or not discussing politics.

Twitter Political Corpus

- Then the Tweets were cleaned by removing the HTML content, non-letters, stop-words and converting it into lowercase, finally splitting into individual words

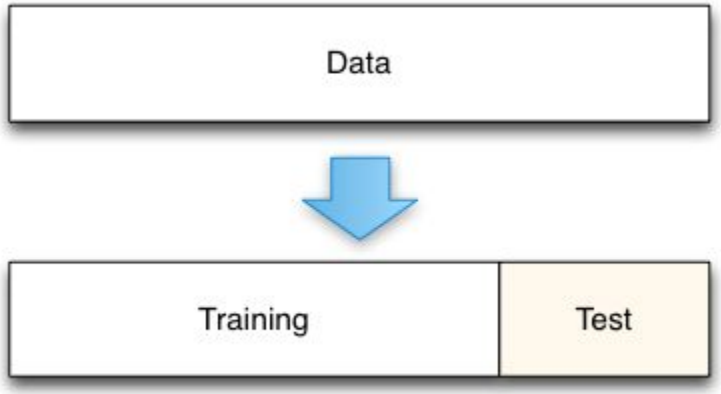
Data Pre-processing

- Count-Vectorizer

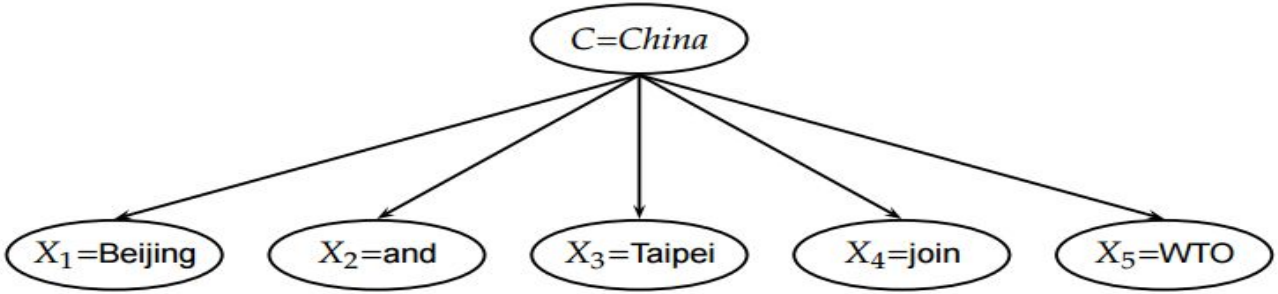
- TF-IDF

Train Test Split

Splitting the available data into random training and testing subsets, drastically reduces the number of samples which can be used for learning the model.

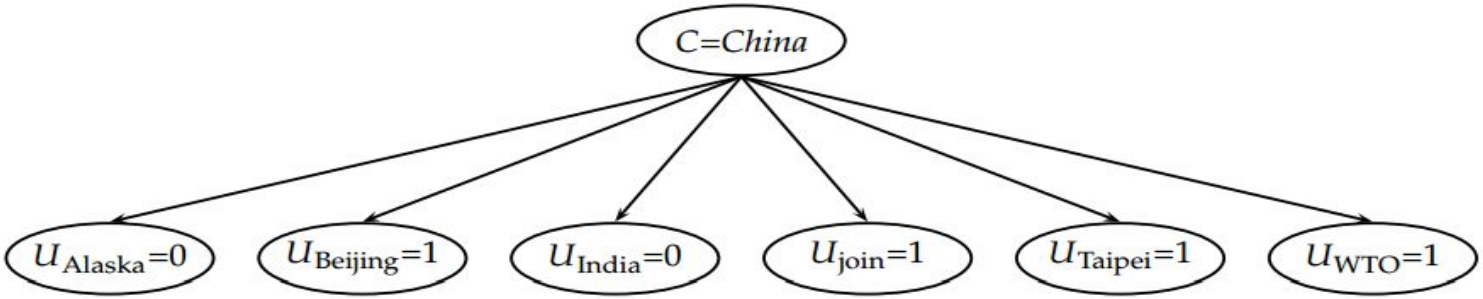


Classifier: Multinomial Naive Bayes



- Estimates the Conditional probability of a particular term.
- Consider number of occurrences of a term.

Classifiers: Bernoulli Naive Bayes



- Generates boolean indicators(0 and 1).
- Includes non-occurrence terms

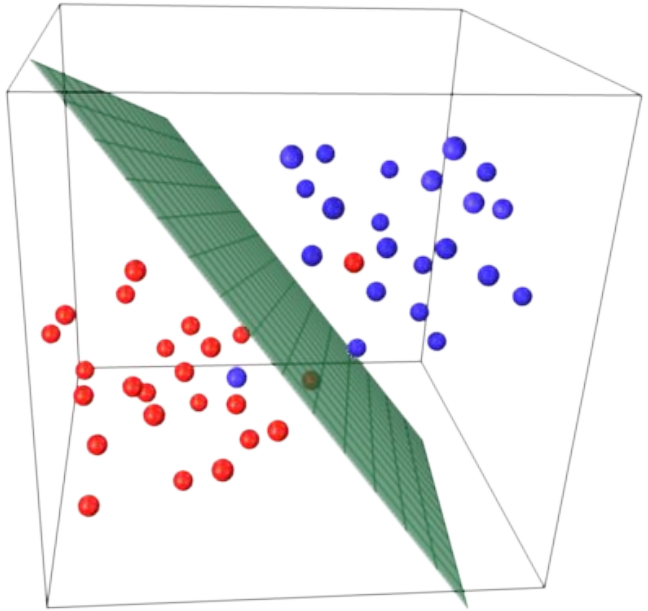
Classifiers: Bernoulli Naive Bayes

- Generates boolean indicators (0 and 1).

- Includes non-occurrence terms

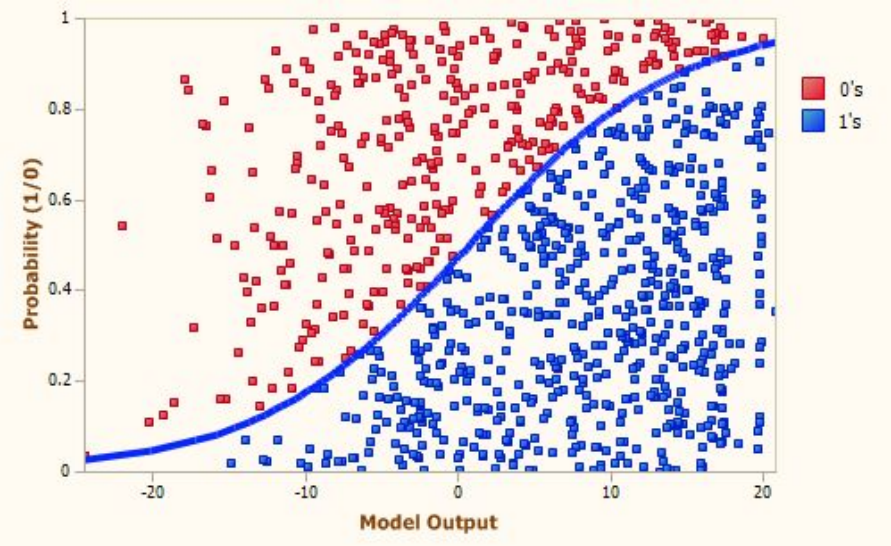
Classifiers: SVM

- Uses hyperplane for segregation
- Different “Kernel” functions depending on separating of the data points

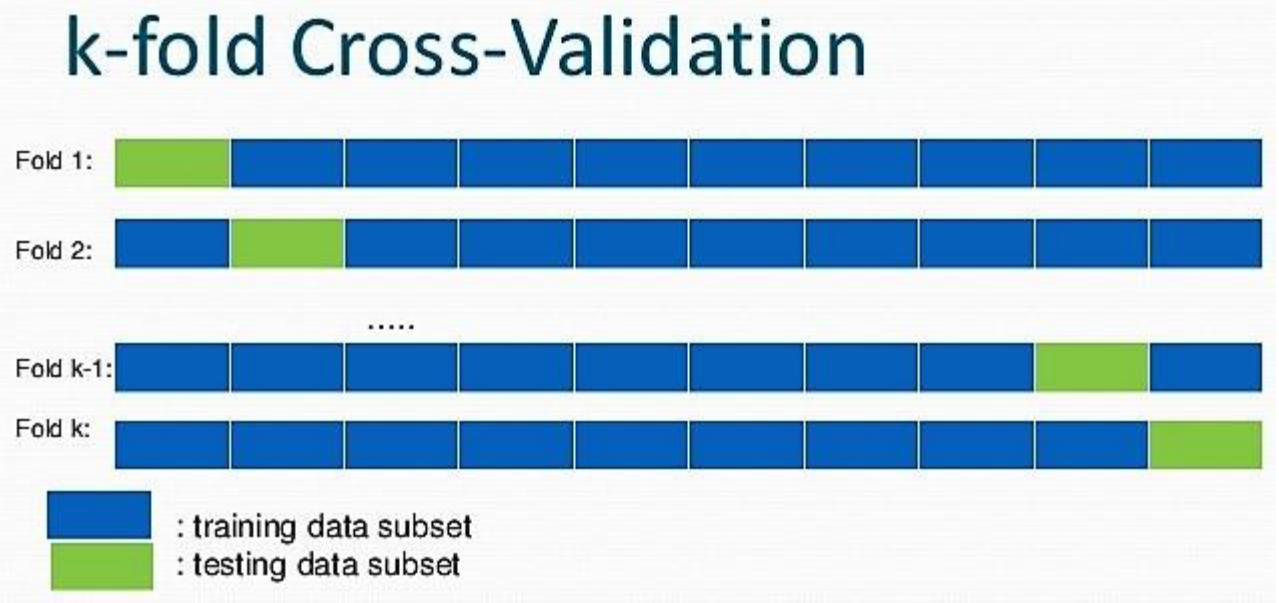


Classifier: Logistic Regression

- Desired outcome of the dependent variable can have two possible types
- Focusses on the conditional probability



Evaluation: 10-Fold Cross Validation



Evaluation: Confusion Matrix

- Describes the performance and accuracy
- Shows the number of correct and incorrect predictions made by the classifiers

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Evaluation: Classification Report

- Shows the main classification metrics
- Precision, Recall, F1-score against the Class labels

Future Work

Future Work: Sentiment Analysis

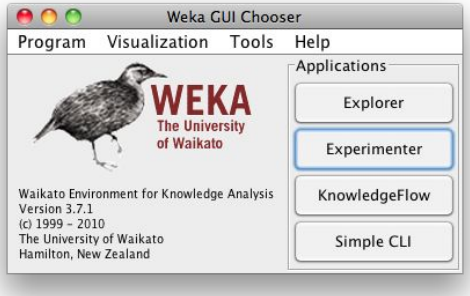
- It is a machine learning method to extract, identify, or otherwise characterize the sentiment content of a text unit and is also referred to as opinion mining.

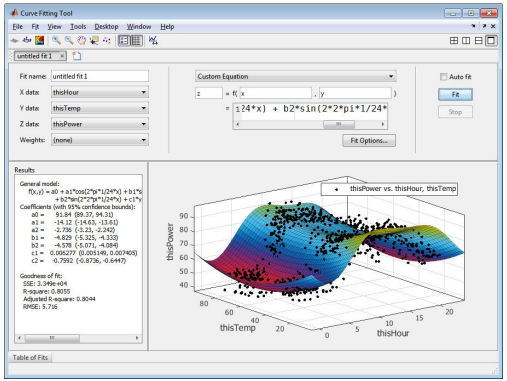
Future Work: Neural Network

- Convolution Neural Network for Text Classification.
- Keras deep learning learning libraries along with Theona / TensorFLow back-end.

Future Work: Tools

- **Tools which can be used**





Conclusion

- Comparing classifiers based on accuracy and time.
- Distinguish between political and nonpolitical tweets.
- Processing speed of the computer.
- Used Cross-validation for more accurate evaluation of the models
- Performance evaluation : concluded Logistic regression yields better accuracy.

Thank You for your Attention

Questions