Machine Learning for Text Analysis and Classification

Sean P. Rogers
NULab for Digital Humanities and
Computational Social Science
PyData VT, July 2024

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

About the Speaker

Assistant Director
NULab for Digital Humanities and Computational Social Science
Northeastern University
Email: se.rogers@northeastern.edu

MassMutual Center of Excellence PhD Fellow/Gund Graduate Fellow
Vermont Complex Systems Center
University of Vermont
Email: sean.rogers@uvm.edu

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Workshop Agenda

- Objectives and Goals
- Introduce Text Analysis Concepts
- Introduce Machine Learning Concepts
- Discuss Textual Feature Engineering and Model Selection
- Ethics Discussion
- Demo (follow along if you want)

Feel free to ask questions at any point during the presentation!

Workshop Objectives

- Understand key concepts about Machine Learning
- Understand and key concepts of text analysis
- Be able to communicate discuss different ways machine learning can enable textual analysis
- How to train and test a model

Feel free to ask questions at any point during the presentation!

Text Analysis

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

What is Text Analysis?

- Broadly, text analysis is the art and science of extracting information from a body of text
- Data sources can include text from social media, digitized novels, newspapers, reports, plans, or any kind of textual information source one could think of
- Used by researchers and practitioners to gain insights into mood and behavior, and uncover relevant textual themes or dimensions
- Examples of Text Analysis Tasks: Named Entity Recognition, Classification

Feel free to ask questions at any point during the presentation!

Why Text Analysis?

- Allows researchers to examine themes in textual data
- Used by researchers and practitioners to gain insights into mood and behavior, and uncover relevant textual themes or dimensions
- Can be applied to large datasets

Feel free to ask questions at any point during the presentation!

Text Analysis: Key Terms and Concepts

Corpus - From the latin word for body - the body of text that is being used for a project

Normalization - Converting text to a uniform format with respect to case and punctuation

Stemming - The process of moving prefixes and suffixes from a word to return it to its base form, may not produce real words

Lemmatization - Reduces a word to its base form by considering semantic context

Example: "The quick brown fox be jump over the lazy dog."

Stemming

"foxes" becomes "fox"

"jumping" becomes "jump"

"dogs" becomes "dog

Lemmatization

"foxes" becomes "fox"

"jumping" becomes "jump"

"are" becomes "be"

"dogs" remains "dogs"

Feel free to ask questions at any point during the presentation!

Text Analysis: Key Terms and Concepts Continued

N-Gram - A series if N-items from a given example of text, for example:

1-gram (unigram): "The", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog".
2-gram (bigram): "The quick", "quick brown", "brown fox", "fox jumps", "jumps over", "over the",

Vectorization - The Process of creating numerical representation of tokens
Term /Frequency Inverse Document Frequency (TF-IDF) is one such approach

Feel free to ask questions at any point during the presentation!

Machine Learning

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

What is Machine Learning?

- An ensemble of tools for developing algorithms and statistical models that enable computers to perform tasks without being explicitly programmed. Machine learning systems learn from data, identifying patterns and making predictions or decisions based on the information they gather.
- Machine learning is a subset of artificial intelligence, and is currently the most effective medium for AI

Feel free to ask questions at any point during the presentation!

Why Machine Learning?

- Reduction of Human Labor and time to result
- Increased accuracy and performance in some instances
- Reduced bias (we hope) if done correctly
- Extensive scientific provenance and support

Feel free to ask questions at any point during the presentation!

Key Concepts and Terms continued

Algorithm - a defined set of rules and processes to enable data processing: Examples of ML Algorithms we will cover are Random Forest Classifier, Linear SVM, and Naive Bayes
Model - Data+Algorithm: A model in machine learning refers to the computational system or process that can make predictions or decisions based on input data.
Training - The process of giving an algorithm example data to learn from to create a model
Testing - The process of assessing the predictions made by the model against known data
Validation - The process of using data unseen by the model during the training process to assess the performance of the model

Feel free to ask questions at any point during the presentation!

Key Concepts and Terms continued

Pipeline - A pipeline in machine learning is a sequence of data processing steps, where each step involves a transformation or a model

Features - Features, also known as variables or attributes, are the measurable properties or characteristics of the input data that are used by a machine learning model to make predictions or classifications.

One-hot-encoding - A feature encoding method where a row that meets a condition is give a value of one, and a row that does not meet the condition is a zero. This helps the model identify patterns.

Cross-validation - is a technique used in machine learning to assess how the results of a statistical analysis generalize to an independent dataset by dividing the data into multiple subsets, training the model on some subsets, and testing it on others.

Feel free to ask questions at any point during the presentation!

Key Concepts and Terms Continued

Accuracy - a metric that measures the correctness of a model by calculating the ratio of correctly predicted instances to the total instances.

Precision - a metric that quantifies the accuracy of positive predictions made by a model. It is calculated as the ratio of true positives to the sum of true positives and false positives.

Recall - a metric that measures the ability of a model to correctly identify all relevant instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives.

F1 Score - a metric that combines precision and recall into a single value. It is the harmonic mean of precision and recall and provides a balanced measure of a model's performance, particularly in situations with imbalanced classes.

Feel free to ask questions at any point during the presentation!

Key Concepts and Terms Continued

Supervised Learning - Learning process is guided by labeled data curated to generalize to external data

Semi Supervised Learning - Learning process is guided by labeled data which is then used to pseudo-label data that augments the training set

Unsupervised Learning - Operate without labeled data, use clustering and other techniques to detect underlying patterns in the data

Feel free to ask questions at any point during the presentation!

Key Concepts and Terms Continued

Cross Validation - A process of partitioning data in to different segments of training, test, and validation sets, to test the generalizability and validity of the model across different subsets of data

Classification Report - A means of evaluating the performance of a classification model by displaying the precision, recall, and F1 scores across classification classes.

Confusion Matrix - A confusion matrix is a visual representation of model performance with respect to true positives, false positives, true negatives, and false negatives across different classification classes

Feel free to ask questions at any point during the presentation!

ML Algorithms - Random Forest Classifier

- *Ensemble of Trees: Random Forest combines many decision trees.*
- *Random Features: Each tree considers random features for splitting.*
- *Bagging: Trees are trained on random subsets of data.*
- *Voting: Each tree votes for the class, and the most votes win.*
- *Reduced Overfitting: Random Forest reduces overfitting by combining predictions.*

Feel free to ask questions at any point during the presentation!

ML Algorithms - Linear Support Vector Machine

- Straight Line Separation: Linear SVM finds the best straight line to separate classes.
- It maximizes the space between the line and the closest points from each class.
- SVM minimizes classification errors while maximizing this space.
- Mainly used for binary classification, but can handle multi-class tasks too.

Feel free to ask questions at any point during the presentation!

Model Evaluation

- Accuracy Trap
- Balance between precision and recall
- Using F1 Score
- Cross Validation

Feel free to ask questions at any point during the presentation!

Thank you!

If you have any questions, contact us at:

Sean P. Rogers
Assistant Director, NULab for Digital
Humanities and Computational Social Science
Northeastern University
se.rogers@northeastern.edu

Slides, handouts, and data will be available post conference

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!