WELCOME TO DATA SCIENCE, ANALYTIC AND ACADEMIC RESEARCH

DR. ANDY ASARE

2025

UNIVERSITY OF CALGARY

UNIVERSITY OF CALGARY

Natural Language Processing (NLP)

UNIVERSITY OF CALGARY

Learning Objectives

1. Define key NLP concepts such as tokenization, stemming, lemmatization, and vectorization.
2. Apply NLP techniques such as sentiment analysis and topic modeling (e.g., LDA) to analyze textual data from academic papers, news, or social media.
3. Assess the effectiveness of NLP models based on performance metrics such as accuracy, precision, and recall.

Natural Language Processing (NLP)

NLP Defined

Natural Language Processing (NLP) is a field of artificial intelligence that focuses on the interaction between computers and human (natural) languages.

It enables machines to understand, interpret, generate, and respond to text or speech in a way that is meaningful and useful.

Text Analytics and Text Mining

Text Analytics is a broader concept that includes information retrieval and data mining such as searching and identifying relevant documents for a given set of key terms.

Text mining is the semiautomated process of extracting useful information and knowledge from large amounts of unstructured textual data sources.

The goal of text miming is to transform text into data for analysis, through the application of natural language processing (NLP) and analytical methods.

Importance of Text Mining and Analytics

1. Identify relevant information from large text datasets.
2. Categorize text into predefined groups (e.g., spam vs. non-spam).
3. Determine the sentiment expressed in text (positive, negative, neutral).
4. Identify main topics within a collection of texts.
5. Create concise summaries of large documents.
6. Identify and classify entities (e.g., names of people, places, organizations).

Topics and Sources of Text

Topics

People, events, products, services, politics, entertainment, sports, etc..

Sources: Blogs, microblogs, forums, reviews, comments

Text Analytics Platforms

Text Analytics Process

Text Analytics Process

Keywords or search words: Words used to search for information

Information extraction: Identifying key phrases and relationships within text

Topic tracking: Identifying information of interest from multiple sources such as social media, news articles, radio etc.

Text Mining Data Cleaning Techniques

- **Web scraping:** Defines the extraction of data from a website. The data collected is then exported into a format that is more useful for further analysis

- **Corpus** is a large and structured set of texts (now usually stored and processed electronically) prepared for the purpose of conducting knowledge discovery.

- **Terms** A term is a single word or multiword phrase extracted directly from the corpus of a specific domain by means of NLP methods.

- **Stemming** is the process of reducing inflected words (changing/deriving) affixes such as (-ed,-ize, -s,-de,mis) to their stem (or base or root)

- **Stop words** (or noise words) are words that are filtered out prior to or after processing natural language data (i.e., of Stopwords are words in NLP are, "is", "an", "the", etc.)

Text Mining Data Cleaning Techniques

- **Synonyms** are syntactically different words (i.e., spelled differently) with identical or at least similar meanings (e.g., movie, film, and motion picture).
- **Polysemes** also called homonyms, are syntactically identical words (i.e., spelled the same) with different meanings (e.g., bow can mean “to bend forward,” “the front of the ship,” “the weapon that shoots arrows,” or “a kind of tied ribbon”).
- **Tokenization** is splitting a body of text into smaller units, such as individual words or terms. Each of these smaller units are called tokens.
- **Lemmatization** extracts the root of the word, e.g ‘dancing', dance’. Different from stemming, lemmatization understands the context and provides the root words rather than simply removing the suffix or prefix of the word.

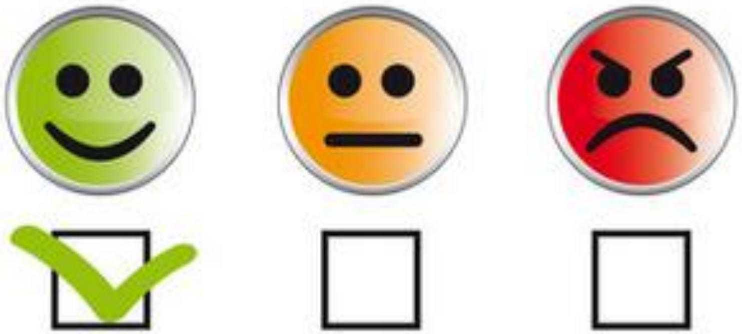
- **Remove duplicates and empty columns**

UNIVERSITY OF CALGARY

Text Mining Analysis Techniques

Sentiment Analysis is the process of analyzing whether a piece of writing such as customer and public comments and reviews are positive, negative or neutral using Natural Language Processing (NLP) methods and algorithms .

Conducting Sentiment analysis enables data analysts within large enterprises to analyze public **opinion**, **emotions, polarity** and **subjectivity** of customers perception about brands, services, and product.

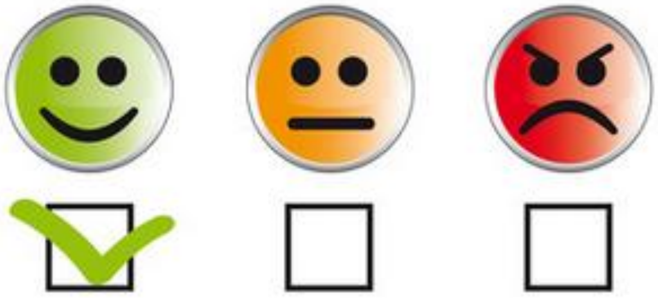


Describe the processes in sentiment analysis

Rule-based systems use a set of human-crafted rules to store, sort and manipulate data to identify subjectivity, polarity, or the subject of fact and opinion.

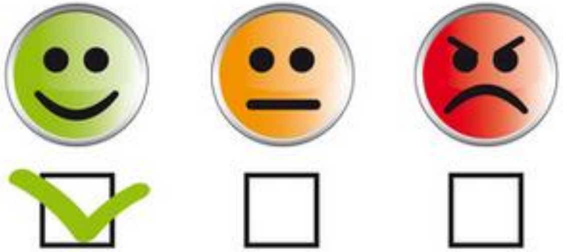
Automatic systems use machine learning techniques to categorize data into positive, negative, or neutral.

Hybrid systems- combines both rule-based and automatic approaches to convert data into positive, negative, or neutral.



Challenges of sentiment analysis

- **Language oriented Issues:** Due to the characteristics of different languages such as Chinese, Arabic, and Hindi conducting sentiment analysis on non-English language becomes a challenge.
- **Spam and Fake Opinions-** The anonymity of web users allows opinion spammers to post fake positive and negative opinions or reviews to promote their product or to discredit their competitors.
- **Irony and sarcasm-** people or customers sometimes express their negative sentiments using positive words
- **Word ambiguity:** Sentiments can sometimes be unclear, indefinite, or vague



Model Evaluation

UNIVERSITY OF CALGARY

Predictive Models for Sentiment Analysis

Predictive Model	Interpretation
Logistic Regression (LR)	A model used for binary or multiclass classification that estimates the probability that a given input belongs to a particular class using a sigmoid function
K-Nearest Neighbors (KNN)	A simple algorithm that classifies a data point based on how its neighbors are classified. It looks at the ‘k’ closest training examples and assigns the most common class among them.
Naïve Bayes (NB)	Estimates the probability of a class given the words in a message. Probabilistic assumptions perform effectively, likely due to clear, well-defined sentimental cues.
Support Vector Classifier (SVC)	Finds the best hyperplane to separate data points from different classes
Random Forest Classifier (RFC)	Builds many decision trees and averages their predictions for better accuracy and generalization.
Decision Tree Classifier (DTC)	A flowchart-like model where data is split into branches based on decision rules (e.g., “Is word X present?”).
Gradient Boosting Classifier (GBC)	Builds a strong predictive model by combining multiple weak learners (usually decision trees) in a sequential manner. Each new model is trained to correct the errors made by the previous ones.

Model Prediction Process

- **Evaluation: Evaluate the model performance using** a separate dataset (validation or test set) that it has not used during the training phase. Use performance metrics such as accuracy, precision, recall, and F1 score to assess how well the model generalizes to new, unseen data.

- **Prediction:** Once the model is trained and evaluated, it can be used to predict the class or category of new, unlabeled data.

Predictive Models for Sentiment Analysis-1

For Classification Models (like sentiment analysis)

- **Accuracy** – Proportion of correct predictions.
- **Precision** – How many predicted positives are truly positive.
- **Recall (Sensitivity)** – How many actual positives were correctly identified.
- **F1 Score** – Harmonic mean of precision and recall; balances both.
- **AUC-ROC (Area Under Curve - Receiver Operating Characteristic)** – Measures ability to distinguish between classes.
- **Confusion Matrix** – Table showing true positives, false positives, true negatives, and false negatives.

Predicting the score of a model (Model Evaluation)-1

SCENARIO

Suppose you're building a model to predict whether a country will increase imports next month (Yes = import will increase, No = import will not increase).

Situation	What Happens	Example in Import Prediction
True Positive (TP)	Model predicts **70 increase**, and imports **actually increase**.	True Positives (TP) = 70
False Positive (FP)	Model predicts **5 increase**, but imports **do not increase**.	False Positives (FP) = 5
True Negative (TN)	Model predicts **20 no increase**, and imports **actually don't increase**.	True Negatives (TN) = 20
False Negative (FN)	Model predicts **5 no increase but** imports **actually increase**.	False Negatives (FN) = 5

Predictive Models for Sentiment Analysis-2

For Regression Models

- **Mean Absolute Error (MAE)** – Average of absolute differences between predicted and actual values.
- **Mean Squared Error (MSE)** – Average of squared differences; penalizes large errors.
- **Root Mean Squared Error (RMSE)** – Square root of MSE; interpretable in same units as target variable.
- **R-squared (R^2)** – Proportion of variance explained by the model.

Predictive Models for Sentiment Analysis-3

For Clustering Models (e.g., k-means)

- **Silhouette Score** – Measures how similar a point is to its own cluster vs. others.
- **Davies-Bouldin Index** – Lower is better; measures intra-cluster similarity.
- **Calinski-Harabasz Index** – Higher is better; ratio of between-cluster dispersion to within-cluster dispersion.

QUESTIONS?

ACTIVITIES

Why Don’t We Scrap Some Data ?

UNIVERSITY OF CALGARY