Social Computing (CS 6454/CS 4803)

Computational Methods for Social Computing I

Mohit Chandra

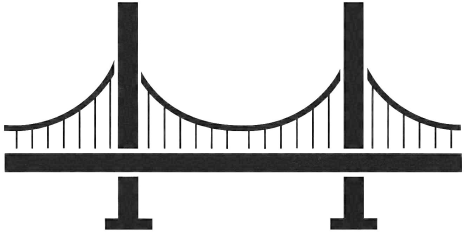
School of Interactive Computing
Georgia Tech

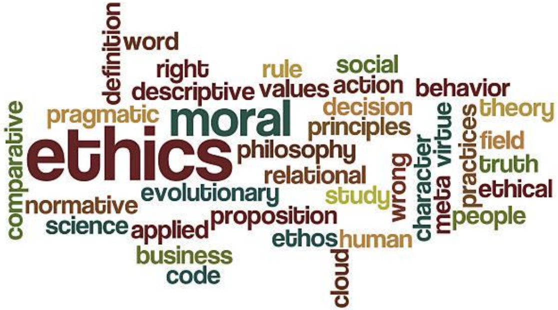
Outline

- Introduction and Computational Framework
- Ethical Data Collection
- Data Analysis and Understanding
- Basics of Natural Language Processing for Social Computing

Introduction and Computational Framework

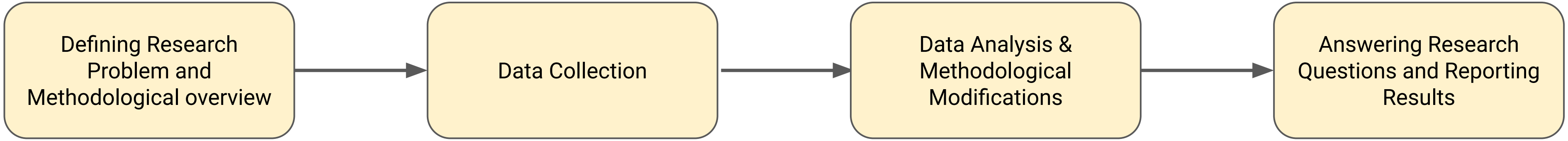
Introduction





Slide Credits: Diyi Yang

Computational Framework



Defining Research Problem and Methodological overview

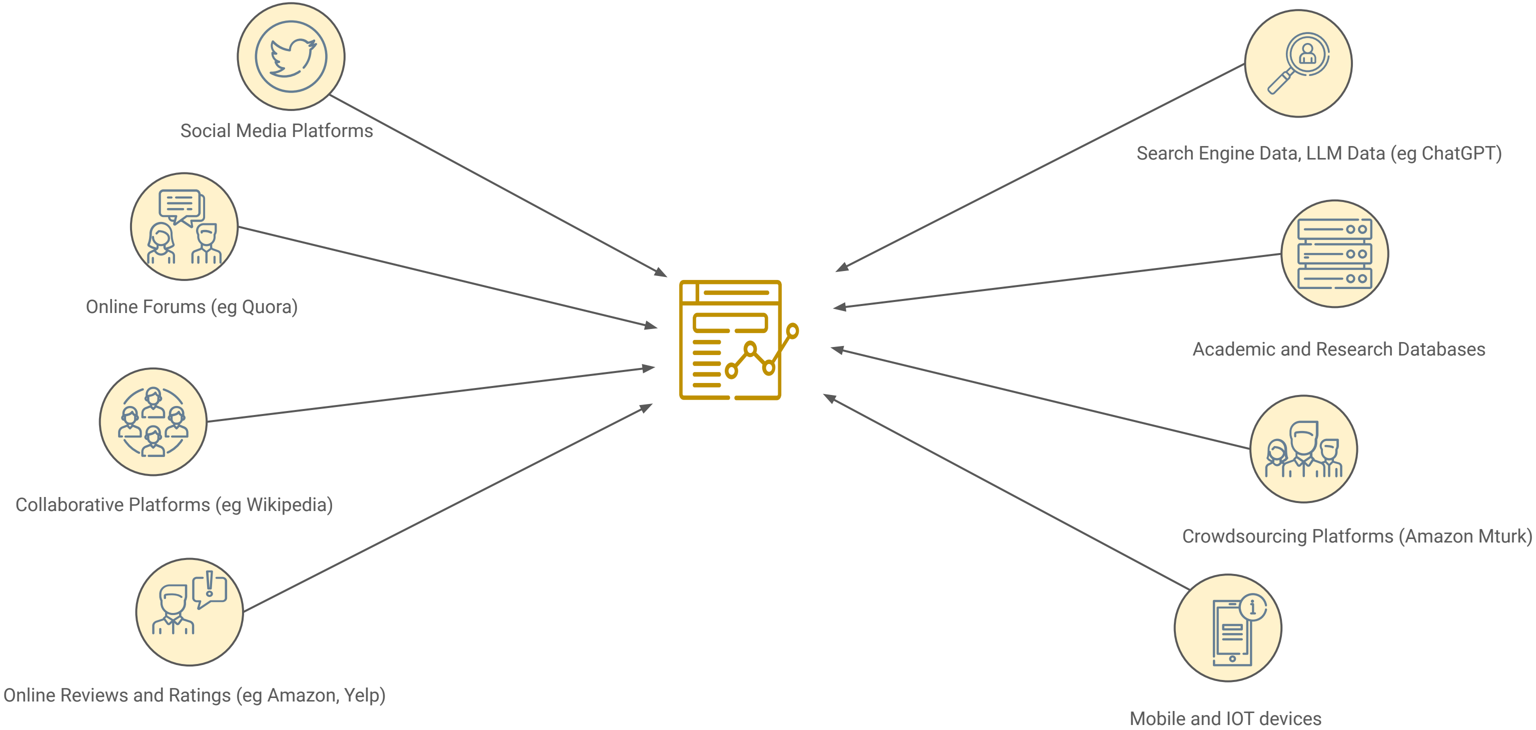
Data Collection

Data Analysis & Methodological Modifications

Answering Research Questions and Reporting Results

Ethical Data Collection

Data Sources


Social Media Platforms
Online Forums (eg Quora)
Collaborative Platforms (eg Wikipedia)
Online Reviews and Ratings (eg Amazon, Yelp)
Search Engine Data, LLM Data (eg ChatGPT)
Academic and Research Databases
Crowdsourcing Platforms (Amazon Mturk)
Mobile and IOT devices

Approaches for Data Collection

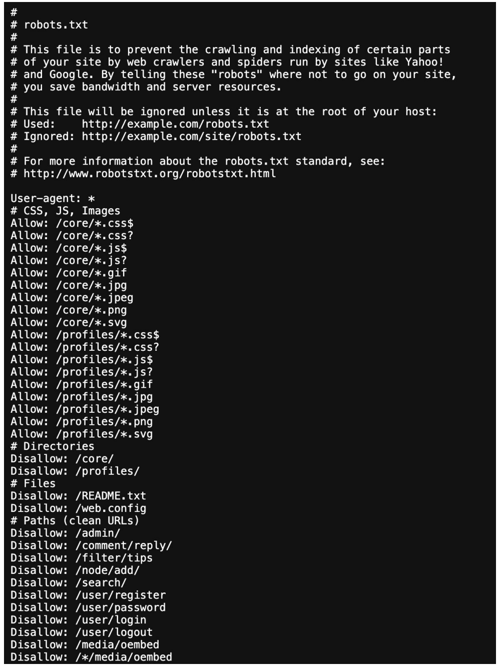
- API from the platforms (Reddit API, GPT-3.5 etc)
  - Ethical way for collecting data.
  - Often paid and rate-limited.
  - Dependent upon the providers for the service and metadata.

- Web-Scraping
  - Ethical (if following robots.txt and Terms of Service agreement (ToS))
  - No-payment required and little to no rate-limit.
  - Metadata often not available.

- Data-Donation
  - Ethical (if approved by IRB)
  - Payments and incentives are often required.
  - Metadata often not available.

Web-Scraping Ethics

- Python libraries such as BeautifulSoup and Selenium (more features)
- Respect for Website Resources
  - Respecting robots.txt and making requests at a reasonable rate.
- Privacy Concerns
  - Address the ethical implications of scraping personal data. Is that data public or private (behind a login-wall)?
- Data Use and Sharing
  - Ethical considerations in the use and distribution of scraped data.
  - Avoiding data misuse and adhering to fair use principles. (https://www.go-fair.org/fair-principles/)



```
#
# robots.txt
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html

User-agent: *
# CSS, JS, Images
Allow: /core/*.css$
Allow: /core/*.css?
Allow: /core/*.js$
Allow: /core/*.js?
Allow: /core/*.gif
Allow: /core/*.jpg
Allow: /core/*.jpeg
Allow: /core/*.png
Allow: /core/*.svg
Allow: /profiles/*.css$
Allow: /profiles/*.css?
Allow: /profiles/*.js$
Allow: /profiles/*.js?
Allow: /profiles/*.gif
Allow: /profiles/*.jpg
Allow: /profiles/*.jpeg
Allow: /profiles/*.png
Allow: /profiles/*.svg
# Directories
Disallow: /core/
Disallow: /profiles/
# Files
Disallow: /README.txt
Disallow: /web.config
# Paths (clean URLs)
Disallow: /admin/
Disallow: /comment/reply/
Disallow: /filter/tips
Disallow: /node/add/
Disallow: /search/
Disallow: /user/register
Disallow: /user/password
Disallow: /user/login
Disallow: /user/logout
Disallow: /media/oembed
Disallow: /*/media/oembed
```

IRB

- Purpose of the IRB
  - To protect the rights and welfare of human research subjects.
  - Ensure ethical standards are met in accordance with federal regulations and institutional policies.

- Key Functions
  - Review and approve research proposals involving human subjects.
  - Monitor research to ensure ongoing compliance.
  - Address concerns or complaints related to research ethics.

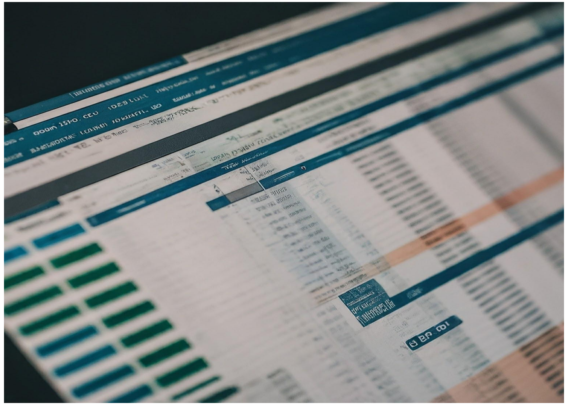
- When to involve IRB?
  - When working with data collected through interaction with human subjects.
  - Eg: Interviews, Data Donation, Crowdworker annotations.

Data Analysis and Understanding

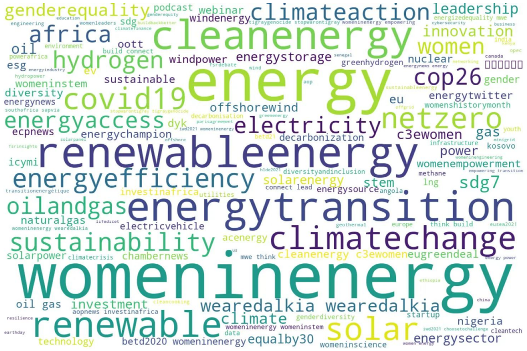
Why look into the data?

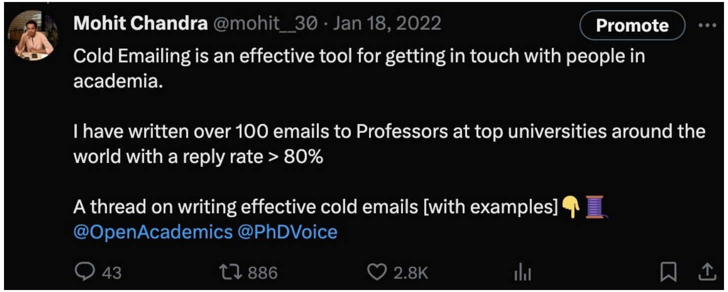
- Different platforms offer different affordances and data/metadata can vary based on how people use different features of the platform.
- Reveal interesting patterns
  - Understanding human behavior , topics and contexts.
  - Detecting data outliers.
  - Nuances of data such as text length, unique vocabulary.
- Deciding further steps.
  - Decision for using specific tools/algorithms/models.
  - Modifications in the methodology for answering RQs.

pandas



Why look into the data?

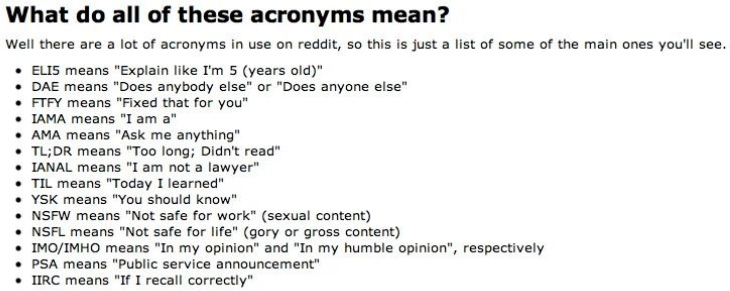



Mohit Chandra @mohit__30 · Jan 18, 2022

Cold Emailing is an effective tool for getting in touch with people in academia.

I have written over 100 emails to Professors at top universities around the world with a reply rate > 80%

A thread on writing effective cold emails [with examples]👇🧵
@OpenAcademics @PhDVoice


What do all of these acronyms mean?

Well there are a lot of acronyms in use on reddit, so this is just a list of some of the main ones you'll see.

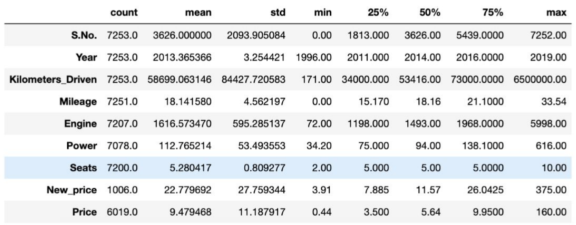
- ELI5 means "Explain like I'm 5 (years old)"
- DAE means "Does anybody else" or "Does anyone else"
- FTFY means "Fixed that for you"
- IAMA means "I am a"
- AMA means "Ask me anything"
- TL;DR means "Too long; Didn't read"
- IANAL means "I am not a lawyer"
- TIL means "Today I learned"
- YSK means "You should know"
- NSFW means "Not safe for work" (sexual content)
- NSFL means "Not safe for life" (gory or gross content)
- IMO/IMHO means "In my opinion" and "In my humble opinion", respectively
- PSA means "Public service announcement"
- IIRC means "If I recall correctly"

Image Credits:
https://www.equality-energytransitions.org/main-hashtags-in-women-in-energy-twitter-network/
https://twitter.com/mohit__30/status/1483395288266321923
https://www.reddit.com/r/coolguides/comments/jxoaeo/commonly_used_acronyms_around_reddit_guide_for/

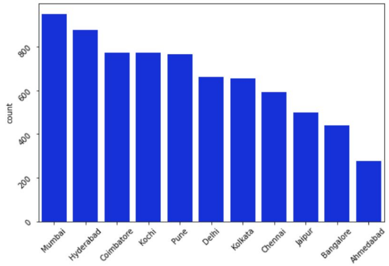
Exploratory Data Analysis

- EDA is a crucial step after the data collection phase.
    - Understanding Data Structure
    - Identifying Patterns
    - Spotting Anomalies

- Important Steps
    - Data type and statistics analysis (analyzing value ranges and type of each variable)
    - Missing value identification
    - Univariate Distribution Analysis for each variable.
    - Multivariate Distribution Analysis.
    - Topic Modeling.



	count	mean	std	min	25%	50%	75%	max
S.No.	7253.0	3626.000000	2093.905084	0.00	1813.000	3626.00	5439.0000	7252.00
Year	7253.0	2013.365366	3.254421	1996.00	2011.000	2014.00	2016.0000	2019.00
Kilometers_Driven	7253.0	58699.063146	84427.720583	171.00	34000.000	53416.00	73000.0000	6500000.00
Mileage	7251.0	18.141580	4.562197	0.00	15.170	18.16	21.1000	33.54
Engine	7207.0	1616.573470	595.285137	72.00	1198.000	1493.00	1968.0000	5998.00
Power	7078.0	112.765214	53.493553	34.20	75.000	94.00	138.1000	616.00
Seats	7200.0	5.280417	0.809277	2.00	5.000	5.00	5.0000	10.00
New_price	1006.0	22.779692	27.759344	3.91	7.885	11.57	26.0425	375.00
Price	6019.0	9.479468	11.187917	0.44	3.500	5.64	9.9500	160.00



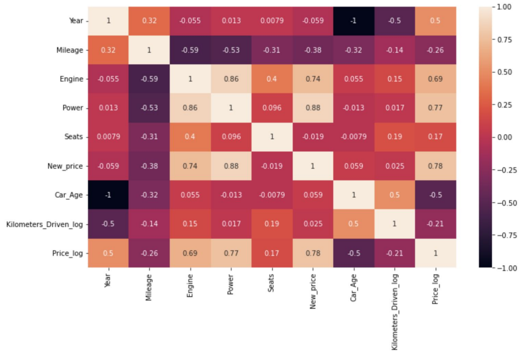


Image Credits:
https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/

Basics of Natural Language Processing for Social Computing

Text Pre-processing Steps

Raw text obtained from various online platforms requires processing for it to be useful for analysis.

The steps can vary based on the chosen platform and the presented problem.

Pre-processing steps:

- Tokenization
- Stop word and punctuation removal
- Lemmatization (Optional)
- Lowercasing
- Emoticon, Hashtags, URL removal (Optional)
- Spelling Correction (Optional)

Using Text/Visual Data

Processing text/visual data requires converting the human knowledge into machine-understandable language

Create vectors/matrices/embeddings from raw data representing information in numerical formats.

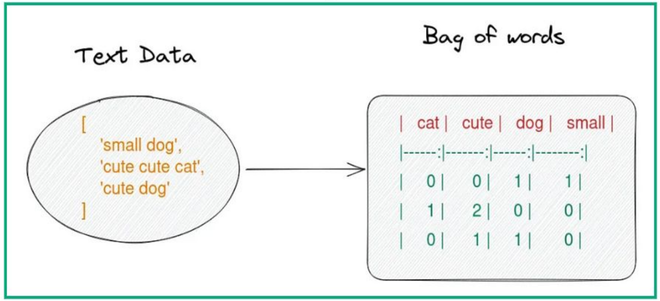
Various Techniques:

- Bag of Words
- TF-IDF Vectors
- Word embeddings (Word2Vec)
- Sentence level contextual embeddings (BERT)

Bag of Words / TF-IDF Vectors

Bag of Words aims at creating a count vector with the tokens occurring in the text that are part of the known vocabulary

Suffers from issues such as large size, no contextual knowledge, no importance for rare/useful words.



[
'small dog',
'cute cute cat',
'cute dog'
]

cat	cute	dog	small
0	0	1	1
1	2	0	0
0	1	1	0

TF-IDF rescales the frequency of words by how often they appear in all documents.

- TF of a term or word is the number of times the term appears in a document
- IDF of a term reflects the proportion of documents in the corpus that contain the term.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

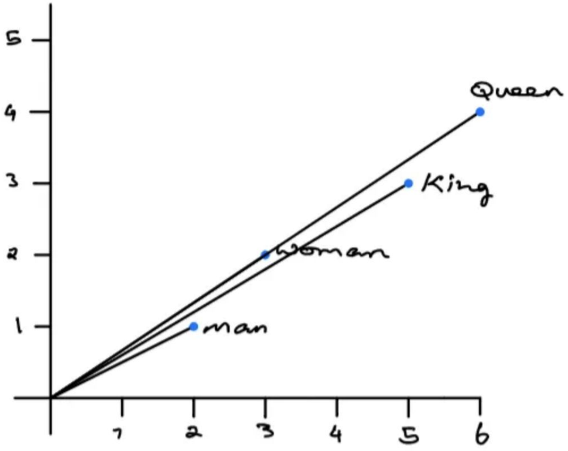
Credits:
https://ayselaydin.medium.com/4-bag-of-words-model-in-nlp-434cb38cdd1b
http://filotechnologia.blogspot.com/2014/01/a-simple-java-class-for-tfidf-scoring.html

Word Embeddings (Word2Vec)

Word embeddings are a technique where individual words are transformed into a numerical representation of the word (a vector)

The embeddings have fixed size and can capture the semantics of the text

Given a large enough dataset, Word2Vec can make strong estimates about a word’s meaning based on its occurrences in the text.

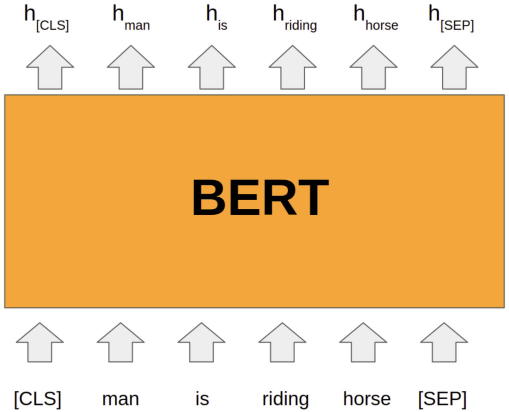


Credits:
https://towardsdatascience.com/word2vec-explained-49c52b4ccb71

Contextual Embeddings (BERT and Beyond)

Transformer models have revolutionised NLP and currently common to be used for creating embeddings

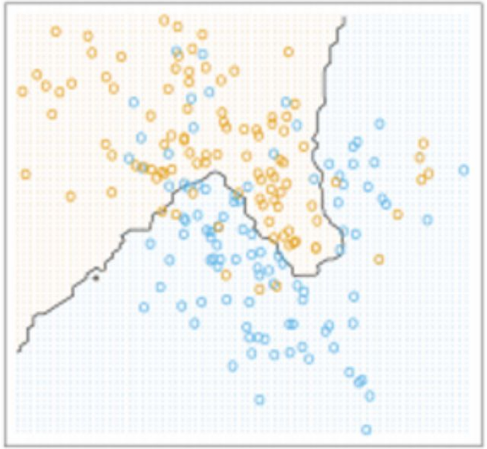
Bidirectional Encoder Representations from Transformers (BERT) embeddings create a fixed length vector capturing the sentence level contextual semantics.



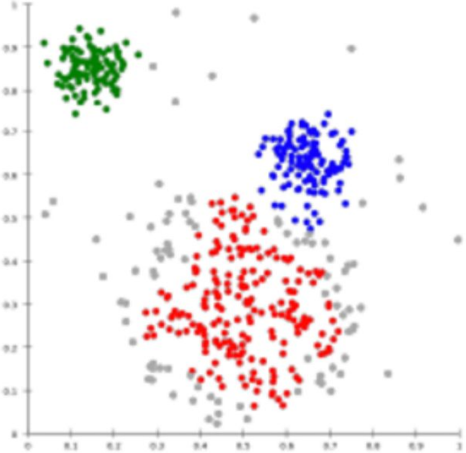
LLM based embeddings such as OpenAI embeddings have become popular recently

Image Credits:
https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-bert/

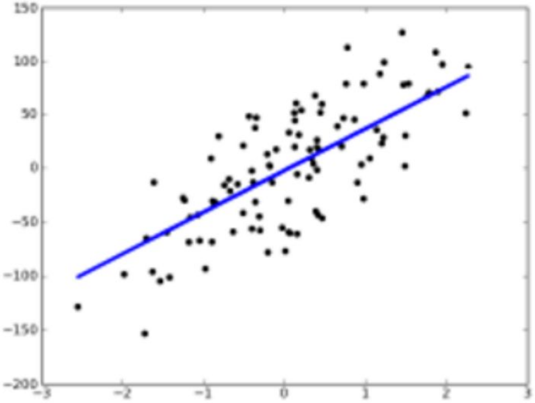
Common Methods for Computational Analysis



Classification



Clustering



Regression

Slide Credits: Diyi Yang

Use of Classification or Regression

Two major uses of supervised classification/regression

- Prediction:
    - Train a model on a sample of data (x,y) to predict for some new data x'
- Interpretation or Explanation
    - Train a model on a sample of data (x,y) to to understand the relationship between x and y

Slide Credits: David Bamman

Text Classification Applications (Prediction)

Task Name	**Task Description**	**Output Variable**
Sentiment Analysis	Identifying the sentiment of the author from the text	Binary: Positive,Negative Multiclass: Angry, Sad, Laugh, Envy etc.
Hate Speech Detection	Identifying the presence/absence of hate-speech.	Binary: Presence/Absence Multiclass: Type of hate
Language Identification	Identifying the language present in the text.	Binary or multiclass
Fake News Detection	Identifying the presence of misinformation/fake news in the text.	Generally Binary Sometimes multiclass
Author Identification	Determining the author of a text based on writing style	Generally multiclass
Health Monitoring	Detecting mentions of illnesses, symptoms, or health-related concerns	Generally multiclass Sometimes binary

Regression Applications

Task Name	**Task Description**
Trend Forecasting	Modeling and predicting the trajectory of trends based on historical data
Health Outcome Prediction	Estimating health outcomes from social media data, such as predicting the spread of diseases based on user posts.
Sentiment Intensity Prediction	Quantifying the degree of sentiment (how positive or negative) expressed in text.
Predicting User Engagement	Estimating the metrics that a post might receive, based on content and contextual features

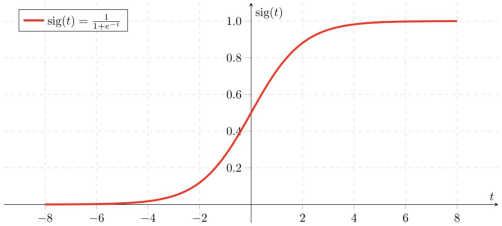
Machine Learning Models

Social Computing employ a wide variety of machine learning models from traditional models to newer deep learning models.

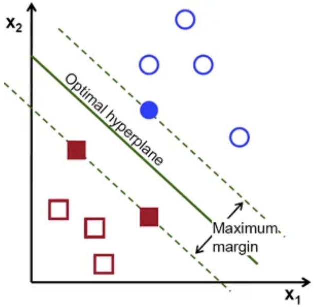
While traditional models provide the ease in training, require less data and are more explainable, they do not perform on the level of newer deep learning based models

Deep Learning models require large amount of data, greater computing resources and are less explainable. However, they perform better than traditional models on most tasks.

Traditional Machine Learning Models



Logistic Regression

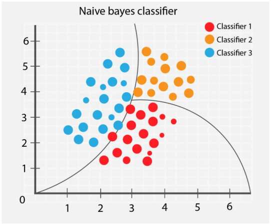


Support Vector Machine

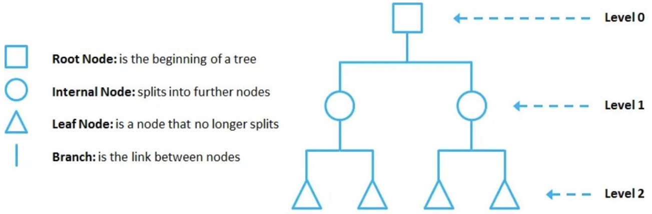
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$



Naive Bayes Classifier



Root Node: is the beginning of a tree

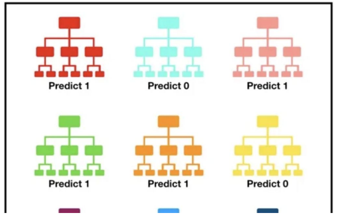
Internal Node: splits into further nodes

Leaf Node: is a node that no longer splits

Branch: is the link between nodes

In this example, a DT of 2 levels.

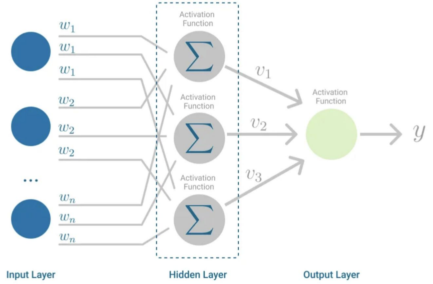
Decision Trees



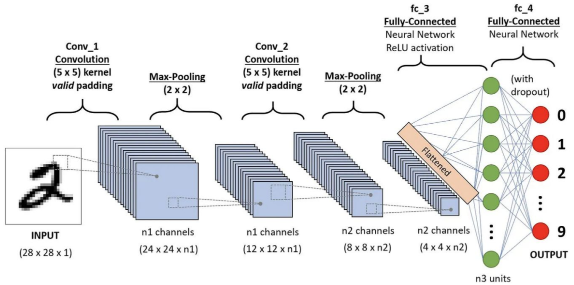
Random Forest Classifier

Credits:
https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
https://thatware.co/wp-content/uploads/2020/04/naive-bayes.png
https://towardsdatascience.com/understanding-random-forest-58381e0602d2

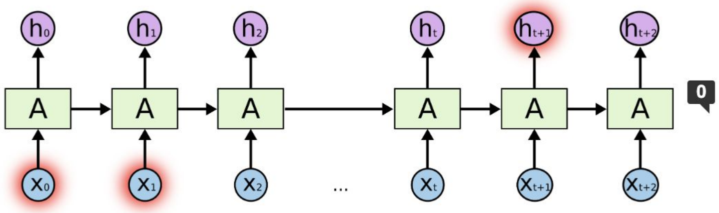
Deep Learning Models



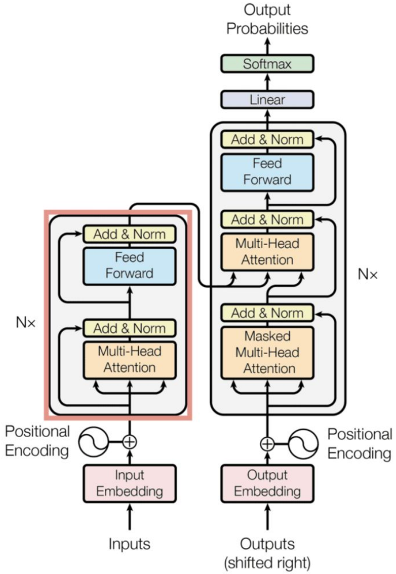
Multi-Layer Perceptron



Convolutional Neural Network



Long Short Term Memory (LSTM)



Transformer Models

Credits:
https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141
https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53
https://colah.github.io/posts/2015-08-Understanding-LSTMs/
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems,

Practical Approach Towards Machine Learning Models

- Understanding the data with its limitations and the size of the dataset helps in deciding the models that need to be applied
    - If data points are few in numbers? use traditional methods
    - Feature engineering might not be a feasible option. In those cases deep learning models might also be a good option.

- Always approach the problem from "ground up". Use simpler and more explainable models in the initial stages and then move towards complex deep learning models.

- Evaluation should involve
    - A cross-validation approach (with held out set in case of deep learning models)
    - Using tools for explainability and performing error analysis helps in understanding the pros and cons for the models
        - Use Libraries such as ELI5 and LIME