Mining Social Data

Ifeoma Adaji

Learning objectives

- Define social data mining
- Identify data sources
- Explain how to extract data for analysis
- Identify and explain possible algorithms for analysis
- Evaluate results

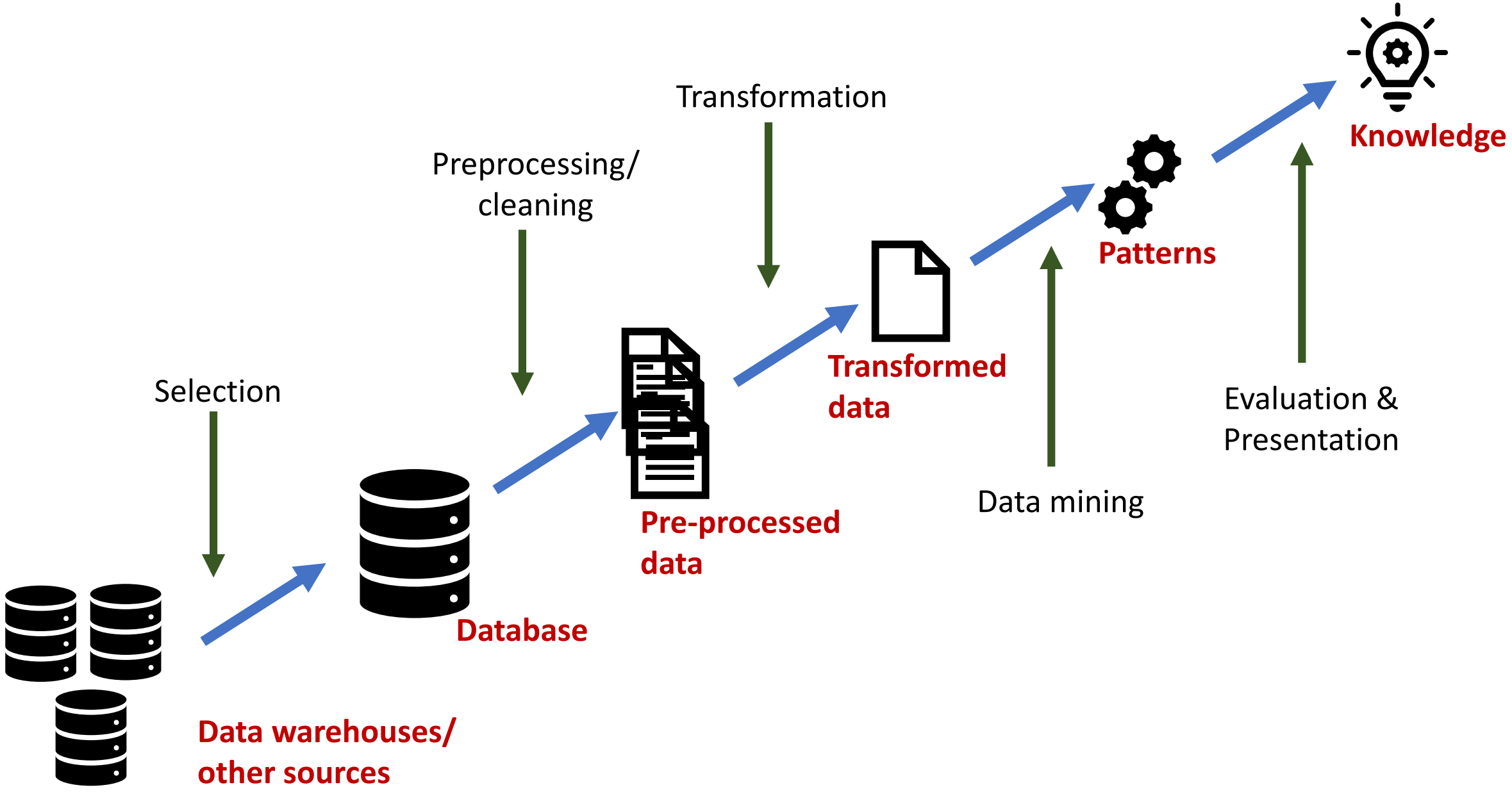
Why data mining

- We live in a world where vast amounts of data are collected daily on social platforms
  - Think of the amount of data generated on social media: Facebook, Instagram, TikTok, Twitter, Netflix
  - Think of data generated on e-commerce platforms: Amazon, Walmart, Ebay
- Data is generated frequently on multiple platforms
- Data is noisy
- Tools are needed to uncover valuable information from these data, to transform such to organized knowledge

Data mining

- Knowledge mining from data (social data)
- Process of discovering interesting patterns and knowledge from large amounts of data
- Extracting novel and actionable patterns from data
  - Meaningful, readily apparent, not easily obtainable
  - Not just about retrieving information from a database
  - Involves methods of Machine learning, Information Retrieval, Statistical Analysis
- Step in the process of knowledge discovery

Steps in the process of knowledge discovery



Data selection

- Various sources of social data
  - Data explorer e.g. Stack Overflow
  - APIs e.g. Reddit, Twitter
  - Web crawling e.eg Beautiful soup (Python)
  - Other data repositories e.g. https://www.kaggle.com/, https://snap.stanford.edu/

Pre-processing and cleaning

- To remove noise and inconsistent data
- Everything you do before data mining
- Anonymization - removing identifying information
- Normalization - weighted all features equally
- Feature Selection – including only necessary features
- Noise/Spam removal

Data mining algorithms

- Supervised learning
  - The use of labeled datasets to train algorithms that classify data or predict outcomes accurately
  - Requires prior knowledge of the data class labels
  - Often used in classification (predicting label) & regression (predicting quantity) problems
- Unsupervised learning
  - Uses unlabeled data to discover patterns that help solve clustering or association problems
  - Particularly useful when one is unsure of common properties within a data set.

Supervised learning

- Classification
  - Uses an algorithm to accurately assign test data into specific categories.
  - It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined.
  - Start with a set of labelled training data
  - Build a model based on the training data
  - Use model to categorize new data

Classification example

Steps in knowledge discovery

ID	Verified account	# Days old	# Followers	# Following	Frequency of posts	Profile picture	Troll?
1	No	1,883	200k	4k	200	1	No
2	Yes	201	65k	8K	18	1	No
3	No	89	1	2k	344	0	Yes
4	No	11	181	9k	566	0	Yes
5	No	33	456	2k	12	0	Yes
6	No	62	655	9k	96	0	Yes
7	Yes	978	121k	1k	15	1	No
8	No	34	30	3k	240	1	Yes
9	No	131	12	3k	103	0	Yes
10	Yes	222	11k	1k	34	0	No

Features

- Verified account
- # days old
- # followers
- # following
- Frequency of posts
- Profile picture

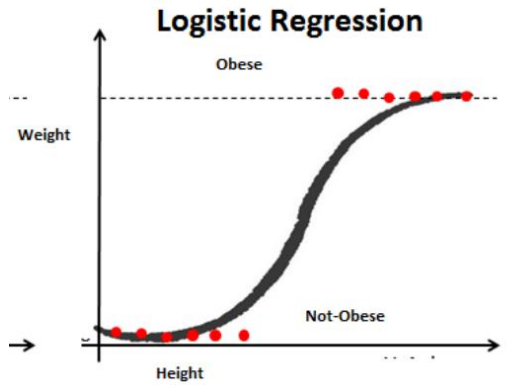
Label

- Troll (Yes/No)

Classification algorithm

- Logistic regression
- Random forests
- Naive Bayes

Classification algorithms



- Logistic regression
  - Used when dependent variable (what one is predicting) is categorical; true/false, yes/no
  - Used to model the probability of a class/event happening
  - Uses a logistic function to model the relationship between independent variables and dependent variables
  - Should not be used when the number of observations is less than number of features
- Naive Bayes
  - Works on Bayes theorem of probability to predict the class of unknown data sets
  - Assumes independence among predictors – main problem

Formula

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B	= events
$P(A\|B)$	= probability of A given B is true
$P(B\|A)$	= probability of B given A is true
$P(A), P(B)$	= the independent probabilities of A and B

Img source: https://www.analyticsvidhya.com/blog/2020/12/beginners-take-how-logistic-regression-is-related-to-linear-regression/

What is the probability that a user will like the post of a friend?

Connection	Like post
Friend	No
Family	Yes
Celebrity	Yes
Friend	Yes
Friend	Yes
Family	Yes
Celebrity	No
Celebrity	No
Friend	Yes
Celebrity	Yes
Friend	No
Family	Yes
Family	Yes
Celebrity	No
Friend	???

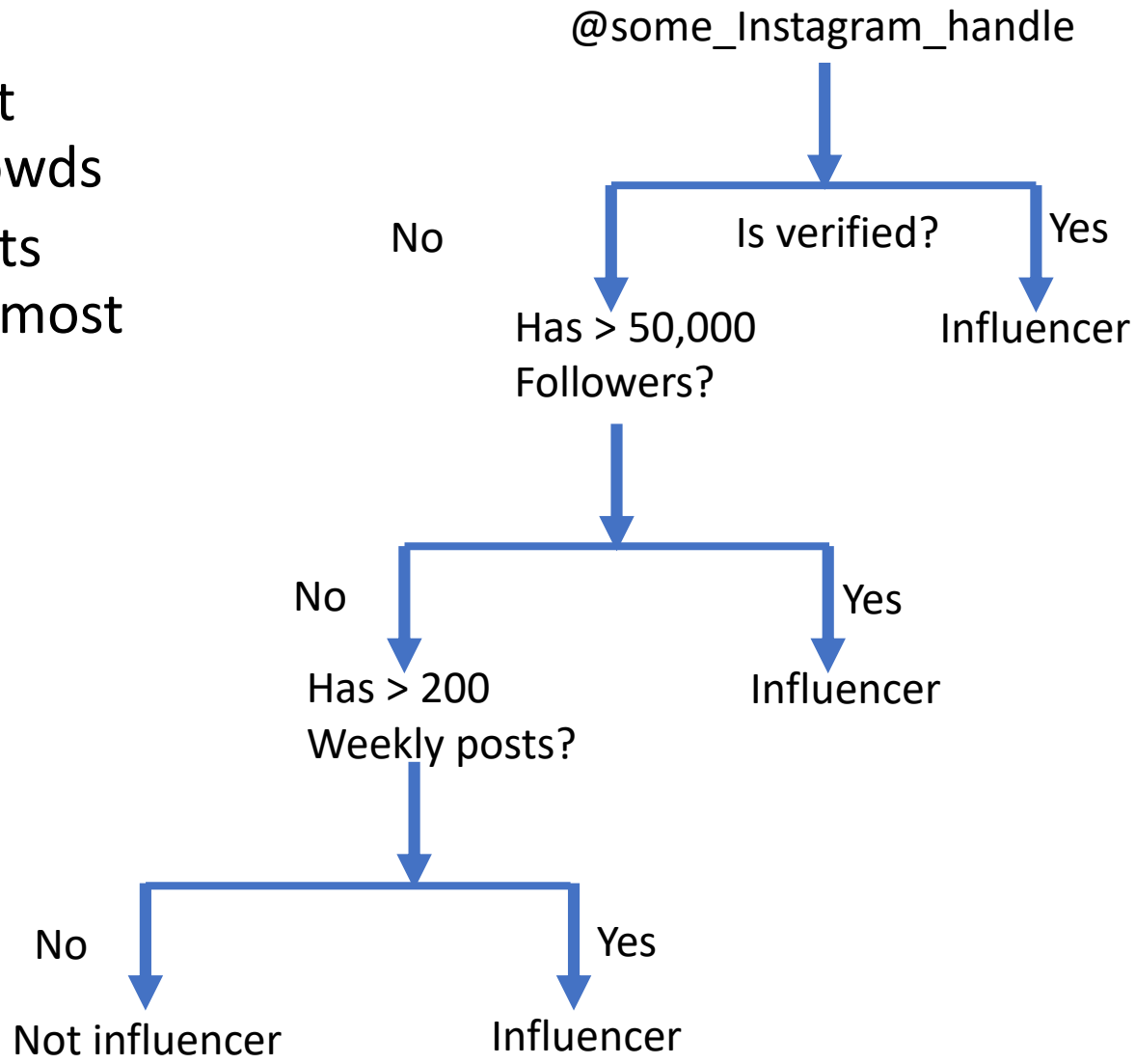
Frequency Table		
Connection	**No**	**Yes**
Family		4
Celebrity	3	2
Friend	2	3
Grand Total	5	9

P(Yes | Friend) = P(Friend | Yes) * P(Yes) / P (Friend)
P (Friend |Yes) = 3/9 = 0.33, P(Friend) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64
P (Yes | Friend) = 0.33 * 0.64 / 0.36 = 0.60

Likelihood table				
Connection	**No**	**Yes**		
Family		4	=4/14	0.29
Celebrity	3	2	=5/14	0.36
Friend	2	3	=5/14	0.36
All	**5**	**9**		
	=5/14	=9/14		
	0.36	0.64		

Classification algorithms

- Random forests
  - Consists of many decision trees
  - Prediction of the group is more accurate that prediction of individual tree – wisdom of crowds
  - Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes the model's prediction
  - Low correlation between models is the key.
  - The trees protect each other from their individual errors
- Other classification algorithms include
  - Support vector machines
  - Neural networks



Unsupervised learning - Clustering

- Used for exploratory data mining
- Determines which samples are similar to each other based on the similarity of the data elements
- The algorithm does not require labels showing how the samples should be group together
- Commonly used in recommendation systems
- Examples
  - k-means clustering

K-means clustering

- Aims to partition n observations into k clusters in which each observation/data point belongs to the cluster with the nearest mean (cluster centers or cluster centroid)
- Initialize K random centroids.
  - Pick K random data points and make those your starting points or
  - Pick K random values for each variable.
- For every data point, look at which centroid is nearest to it.
  - Using some sort of measurement like Euclidean or Cosine distance (commonly used in text mining)
- Assign the data point to the nearest centroid.
- For every centroid, move the centroid to the average of the points assigned to that centroid.
- Repeat the last three steps until the centroid assignment no longer changes.
  - The algorithm is said to have “converged” once there are no more changes.

Evaluation

Accuracy

$$Accuracy = \frac{number\ of\ correctly\ classified\ test\ sample}{total\ number\ of\ test\ sample}$$

Precision

$$Precision = \frac{number\ of\ relevant\ results\ that\ were\ ranked\ or\ retrieved}{number\ of\ retrieved\ results}$$

Recall

$$Recall = \frac{number\ of\ relevant\ results\ that\ were\ ranked\ or\ retrieved}{number\ of\ relevant\ results}$$

F1 Score

$$score = 2 * \frac{precision * recall}{precision + recall}$$

Applications of social media data mining

- Advertising
  - Determine what messages are most effective among certain demographic groups
  - Determine the best time of the day to run a specific ad on a specific digital platform
- Influencer marketing
  - Helpful in identifying influencers/users with strong follower base and engagement rates
- Predictive analysis
  - Predicting churn/customer attrition, misinformation spread, etc.
- Recommender systems
- Trend analysis
- Event detection
- Spam detection

Problems

- Privacy
- Ethical issues
- Bias
- Data ownership
- Identity theft – fraud

Summary

- Valuable information is hidden in vast amounts of social media data
- Mining this data can discover actionable knowledge that is otherwise difficult to find
- Mining can be done using various methods including clustering and classification

References

- Mining the Social Web, 3rd Edition By Matthew A. Russell, Mikhail Klassen
- Data Mining: Concepts and Techniques, 3rd Edition By Jiawei Han, Jian Pei, Micheline Kamber