Social Media Analysis

Ashish Kumar Jha

Agenda

Article critiques

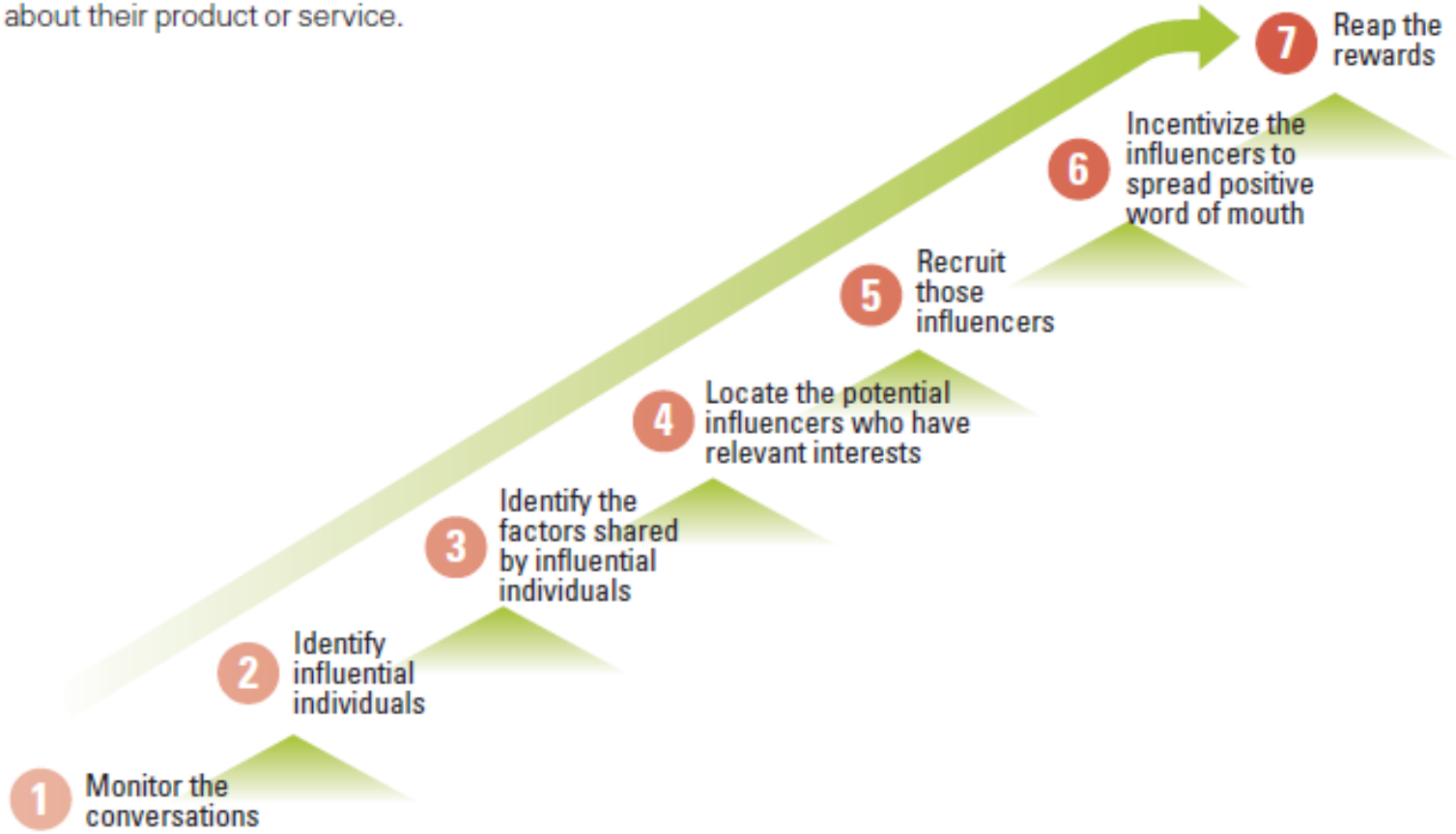
Understanding the textual data

Cleaning and handling text data

Introduction to Linguistics

SEVEN STEPSTO SOCIAL MEDIA SUCCESS

Our research suggests that by using this seven-step framework to identify and recruit individuals who are influential on social media, businesses can promote social media word of mouth about their product or service.



Kumar & Mirchandani, 2012

Major steps

Identify influencer

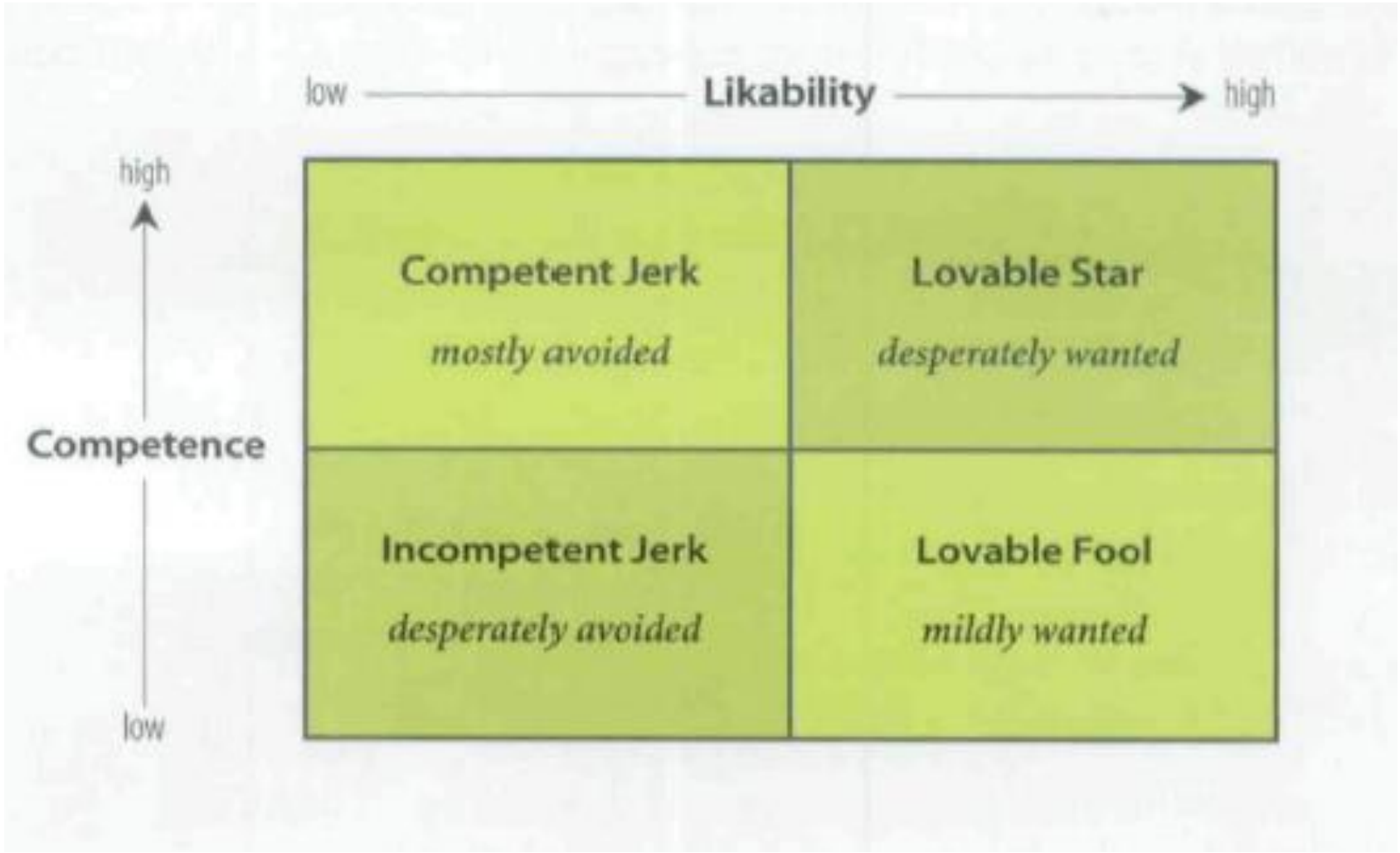
- Number of times messages were forwarded
- Number of connections jumped
- Number of comments and replies

Identify Ideal influencer

- Activeness
- Clout
- Talkativeness- retweet
- Likeminded friends/followers

Metrics

- CIE
  - How much influence an influencer has on their followers
- Stickiness Index
  - How much a person discusses the topic of interest or related topics



Casciaro & Lobo, 2005

Manufacture likeability

Promote familiarity

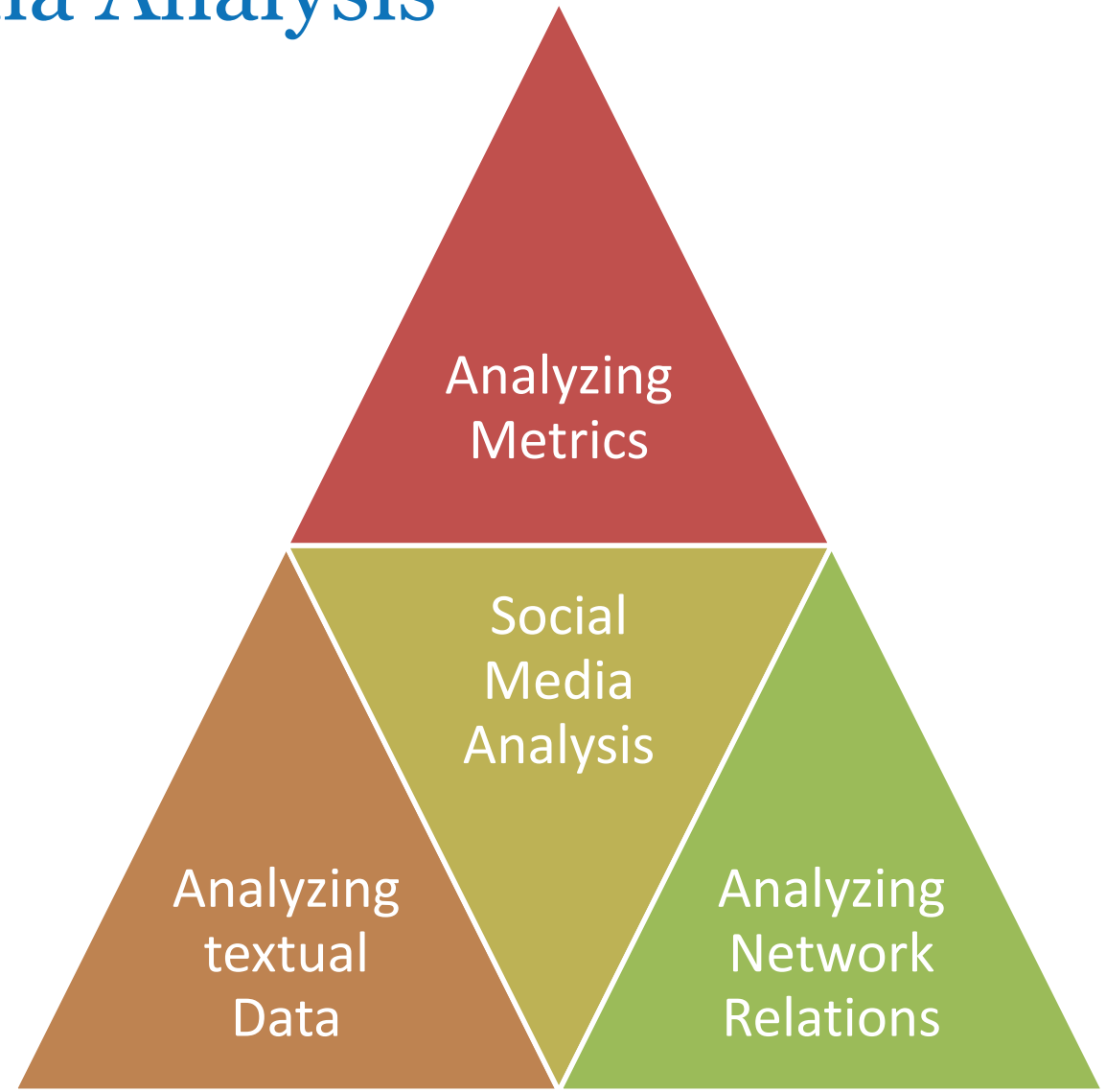
Redefine similarity

Foster Bonding

What to do about the Jerk

Implications for social media

Social Media Analysis



Text Analysis

Text data fundamentals

- Largest kind of dataset in modern world
- Provides deepest insight
- Is considered as “Unstructured data”

Important R packages to deal with

- Tidyverse – dplyr, tidyr
- Stringr
- Tm
- NLP
- Topicmodels
- Text2Vec

Text data organization

Letter – “r”

Word – “red”

Sentence – “red fox is sleeping”

R/text analysis packages use the following technical conventions

- Character
- String
- Token

Data can be stored as

- Document/ CSVs
- Corpus
- DTM

Character strings

```
x = c('... Of Your Fake Dimension', 'Ephemeron', 'Dryswch', 'Isotasy', 'Memory')
```

Raw text cannot be analyzed quantitively

Need to transform the text to quantitative equivalents depending on the query

Basic Transformations

Paste

Substr

nchar

Grep

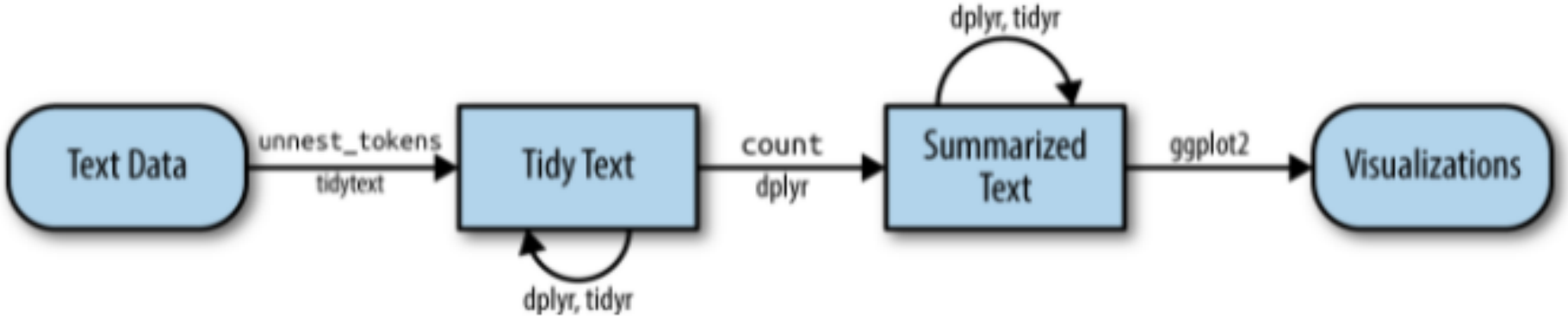
- **Regular expression (regex)**
- **One of the most powerful textual commands**
- **Will take a long time to get used to**
  - ^r.*fox$
  - **^** : starts with, so ^r means starts with r
    - **.** : any character
    - * : match the preceding zero or more times
    - **fox**: match ‘fox’
    - **$** : ends with preceding
- DPLYR can replace some functions

Text Organization

Generally stored as raw text, imported as such

Convert to tibble to make more sense

Best to convert it to tokens for ease of manipulation



Now you can start to do operations

- Count number of word per sentence
- Frequency analysis etc.

table(token1$line)

Tf-IDF

One of the most-important ways to manage and analyze documents

Term-frequency- Inverse document frequency

- how frequently a word occurs in a document

$$idf(\text{term}) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right)$$

- statistic **tf-idf** is intended to measure how important a word is to a document in a collection (or corpus) of documents

Most words occur very infrequently in any language and follows an exponentially declining distribution

- Zipf's law states that the frequency that a word appears is inversely proportional to its rank.

Linguistics

Analysis of language is theoretical linguistics

- Development of universal grammar at core of NLP
- Major areas

–Phonetics-study of speech sounds

–Phonology –Study of sounds for meaning like stress, tones

–Morphology-Study of internal structure of words

–Syntax-study of language structure

–Semantics-Study of intension i.e. intrinsic meaning of words/phrases

Linguistics

Text can be divided into

- Lexical analysis
- Semantic analysis

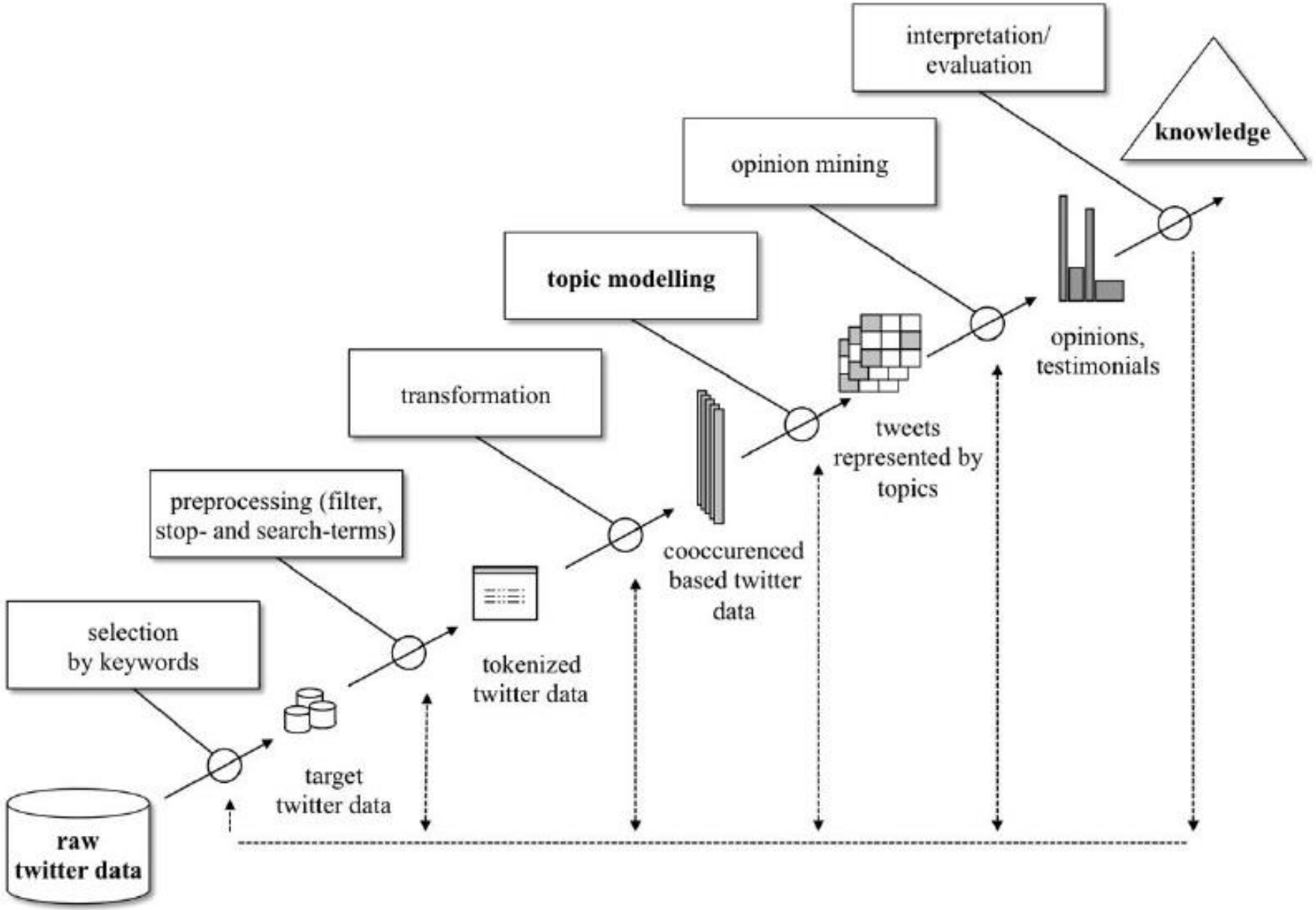
Lexicon

- Vocabulary of the language
- Lexical analysis will focus on the words and their formation, frequency of usage

Semantics

- branch of linguistics studying the meaning of words
- Semantic analysis describes the process of understanding natural language–the way that humans communicate–based on meaning and context.
- It focusses on analyzing why the word is used and what are the meaning behind the usage

Customer opinion mining



Sentiment Analysis

Sentiment analysis/Emotion mining is a part of broader field of Natural Language Processing

Sentiment is a simplification of the thoughts of the holder based on written or spoken content.

For vast majority of applications, we use written words

Sentiment Analysis levels

Document level

Sentence level

Word/Phrase level

Broad Approaches

Lexicon Based

- Count no of words of positive and negative intent
- Subtract the count of words from the 2 sets and find the overall polarity
- Polarity = Positive polarity – Negative polarity
- Popular software – SentiStrength (Free to use)

Tf-idf

- Term frequency- inverse document frequency

Broad Approaches

Algorithm Based

- Model learns to associate a particular input (i.e. a text) to the corresponding output (tag) based on the test samples used for training
- In the prediction process (b), the feature extractor is used to transform unseen text inputs into feature vectors. These feature vectors are then fed into the model, which generates predicted tags

Approaches

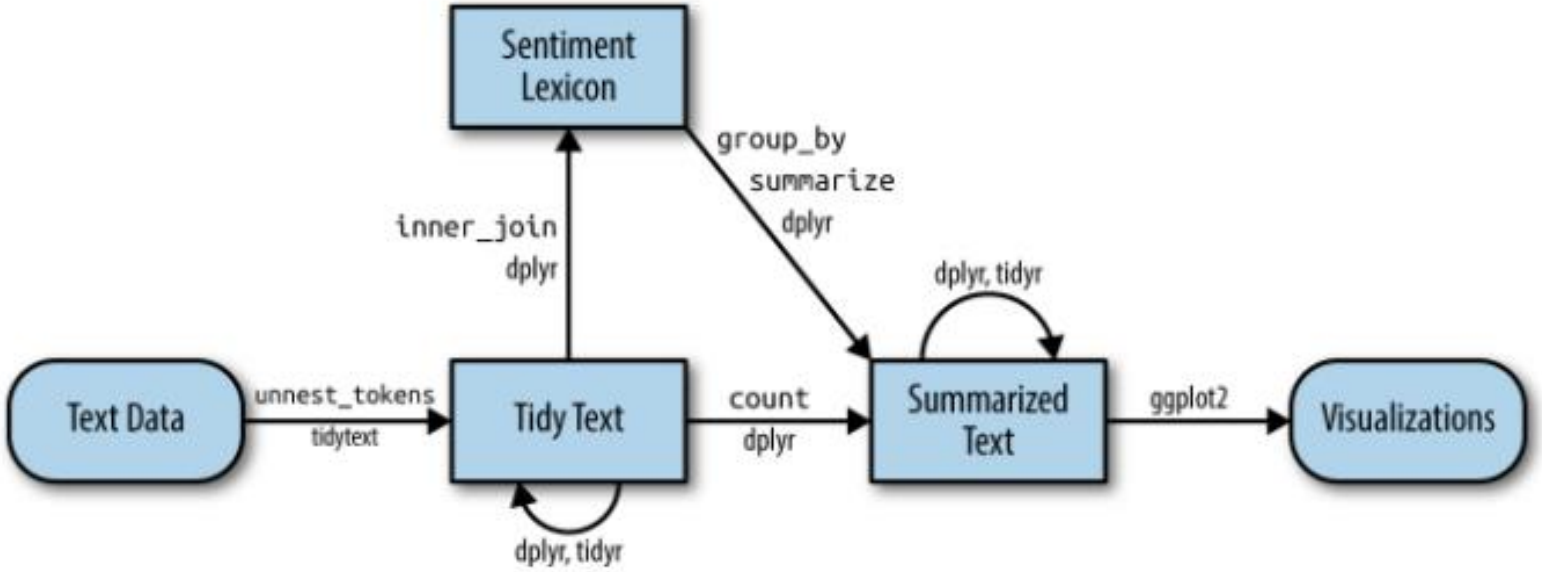
- Regression
- Naïve Bayes
- SVM
- Deep Learning

Feature Extraction

- Parsing, stemming, tokenization
- Transform text into numerical representation
- Each component of the vector represents the frequency of a word or expression in a predefined dictionary
- Bag-of-words approach based on either individual words or n-grams approach
- New approach includes creation on word vectors (word2vec)

Sentiment Analysis- R

- Consider the text as a combination of its individual words and the sentiment content of the whole text as the sum of the sentiment content of the individual words



Sentiment Analysis R

- **Lexical analysis – Datasets – (dictionary based methods)**

– AFINN from Finn Årup Nielsen

  - assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment

– bing from Bing Liu and collaborators

  - categorizes words in a binary fashion into positive and negative categories

– nrc from Saif Mohammad and Peter Turney

  - categorizes words in a binary fashion (“yes”/“no”) into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust

- **Based on unigrams (single words)**
- **Only for English**