Social Media Analytics - CS-EJ5621

Lecture 5

Aalto University
School of Science

FITech
NETWORK UNIVERSITY

Aqdas Malik (aqdas.a.malik@aalto.fi)

09.10.2020

Course practicalities

- **Missing Quizzes and Case project proposals**
- **Quiz 4 due today (2359)**
- **Guest lecturer**


social
media
Brand
sources
help
infinite
art
direct
reviews
including
blogs
transform
tweets
individuals
Internet
one
electronic
also
monitoring
sites
altered
neologism
broadcast
Answers
almost
inputting
sharing
using
different
television
bookmarking
accessible
print
devices
use
instant
skills
describe
microblogs
knowledge
interaction
Customer
thread
U.S
news
small
radios
even
six
marketing
search
made
communications
references
discuss
types
big
free
adopting
may
reach
many
people
network
content
information
See
Tagged
technologies
Music
since
now
multiple
December
applications
consumers
projects
participants
optimization
total
video
technological
networking
Examples
production
community
syndication
Distinction
edit
training
globally
Community
newspapers
example
properties
industrial
September
become
article
anyone
economy
statistics
book
Online
virtual
user-generated
brands
platforms
refer
internet
compared
generally
communities
enable
include
like
physical
publish
producers
page
worlds
time
typically
means
aggregation
global
However
consumer-generated
issues
government
mobile
Business
forms
used
tools
build
others
million

Bikesh Raj Upreti

MSc. (IS); Ph.D. “Application of text mining methods”

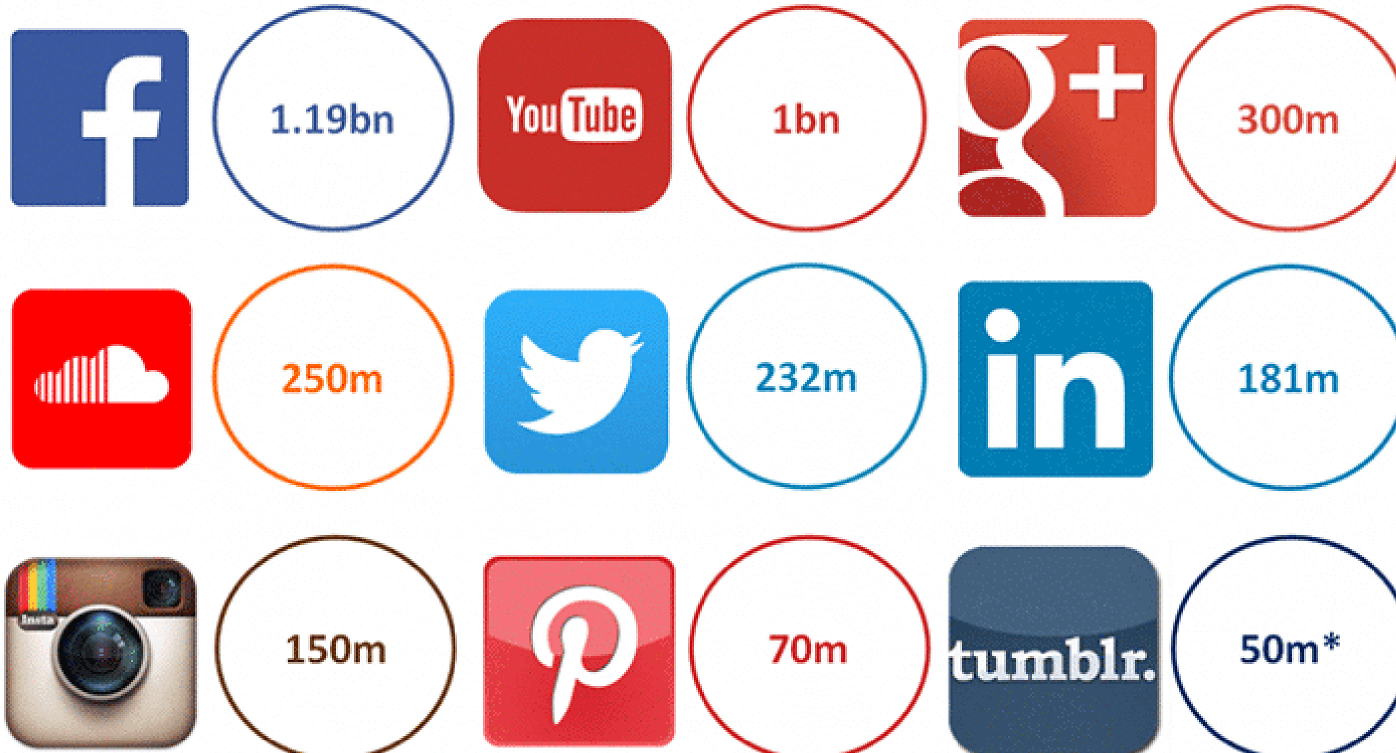
Research Area: Application of machine learning, Text mining, Statistical analysis in business domain

Experience in collecting and analyzing user-generated content from social media platform and discussion forums

Agenda

- Introduction
- Social media data
- Text pre-processing
- Text analytics methods
- Examples of social media text pre-processing

Active Monthly Users of the 'Big 9'

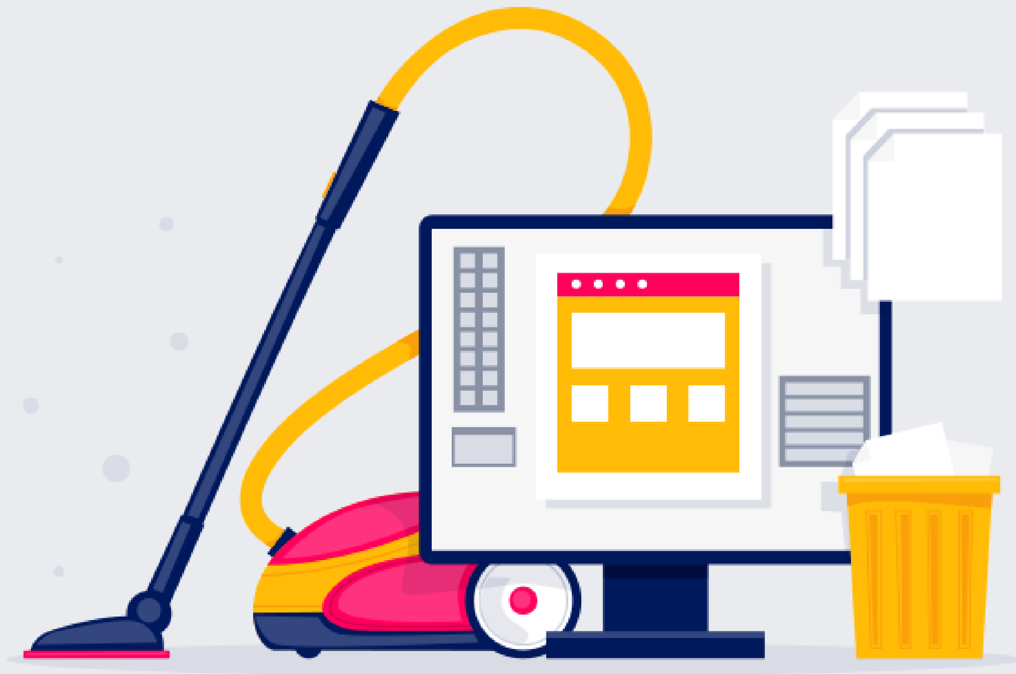


Social media analytics

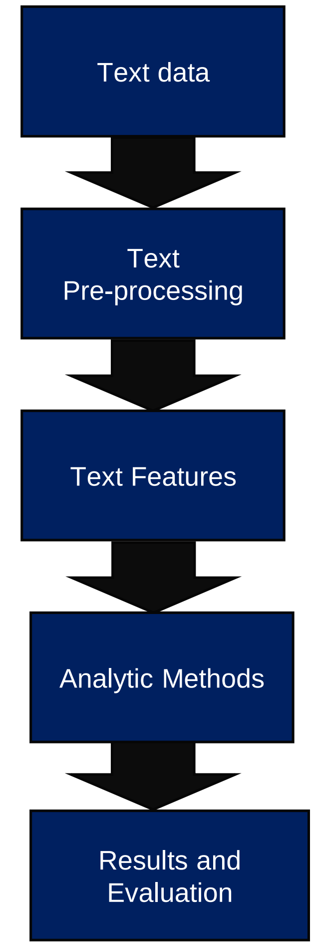
- Digital user generated content contains vast amount of information
- Different from experimental setup: Analysts and researcher are observer of phenomenon
- Compared to survey method provides more robust approach to data collection
- Interest from various domains e.g. business, politics, social and behavioral science
- User level data, time series data, and other metadata
- Text is among the dominating form of data

Social media text

- User generated content
- Among the free form of text
- Lacks structure and correctness
- Noisy: mixture of different language, spelling errors, URL links, tags and words out of dictionary
- Very challenging in cleaning



Source: https://iterable.com/blog/growth-marketing-platform-migration-guide-part-2-cleaning-data/



Text pre-processing

- Objective is to clean the data
- Reduce noise, remove uninformative words, reduce variation
- An important step in text analytics
- Example steps:

  Tokenizing -> Lower case -> Remove numbers -> Remove URL -> Remove username -> Remove retweet header -> Lemmatize/Stemming ->Remove stopwords -> Remove punctuations

Text pre-processing

- Tokenization: Converting text into list of words (Separating words)
- Stopwords: Common words in language, usually do not add value in interpretation (pronouns, articles and auxiliary verbs)
- Removals: Number, URL links, username requires pattern matching
- Stemming: Heuristic approach of reducing words to word stem (basic form)
- Lemmatization: Use of morphological analysis to reduce words to dictionary form

Text feature representation

- Matrix form with document (tweets as row) and features as column
- Matrix cell value: scoring or count based on various methods
- Depends upon analysis method
- Can be binary, frequency counts, topics produced by topic model, or embedding vectors
- Some examples: Bag of words with frequency count, TF-IDF feature representation, Topic models, Vector embeddings

Text feature representation

- Bag of words: Word order not preserved
- Classical method: Word counts
- Example: Words as column, documents as row, word counts as entry
- Can also be binary counts

Documets /Terms	a	all	and	dog	drinks	friend	good	had	I	is	meal	nice	need	who
I had a good friend who had good dog	1	0	0	1	0	1	2	2	1	0	0	0	0	1
good friend and good dog is all I need	0	1	1	1	0	1	2	0	1	1	0	0	1	0
I had a nice meal and a nice drinks	2	0	1	0	1	0	0	1	1	0	1	2	0	0
dog is a nice friend	1	0	0	1	0	1	0	0	0	1	0	1	0	0

Term frequency – Inverse document frequency (TFIDF)

- Reflects word importance: Improvement over word counts
- TF = (Word frequency in the document) / (Total word counts in the document)
- IDF = log(Total number of documents / Number of documents with the word ”W” in it
- TF-IDF = TF * IDF

Documets /Terms	a	all	and	dog	drinks	friend	good	had	I	is	meal	nice	need	who	Total words
I had a good friend who had good dog	1	0	0	1	0	1	2	2	1	0	0	0	0	1	9
good friend and good dog is all I need	0	1	1	1	0	1	2	0	1	1	0	0	1	0	9
I had a nice meal and a nice drinks	2	0	1	0	1	0	0	1	1	0	1	2	0	0	9
dog is a nice friend	1	0	0	1	0	1	0	0	0	1	0	1	0	0	5
Number of documents with the term	3	1	2	3	1	3	2	2	3	2	1	2	1	1	
Inversce document Frequency	0,125	0,602	0,301	0,125	0,602	0,125	0,301	0,301	0,125	0,301	0,602	0,301	0,602	0,602	

Popular text analytics methods

Mention counts (Keyword counts) -> How to overcome variations

- One of the most popular method

Sentiment Analysis

- Using sentiment dictionary

Topic discovery

- Still challenging due to nature of text i.e. short and sparse (spread out)

Trend analysis

- Time-series analysis

Sentiment analysis

- **Two Approach:**
  - Dictionary based method:
    - Words are associated with sentiment scores
    - Prepared by the experts and tested
    - Example includes: Harvard Inquirer, SentiWordnet, LWIC, Vader
    - Sentiment for text is based on the word scores
  - Learning from meta-data:
    - Using machine learning to learn the sentiment rule from already classified data
    - Requires manual effort in classifying data
    - Easier to validate performance

Topic models

- **Popular suite of methods in text mining**
- **Latent Dirichlet Allocation (LDA)**
- **Assumptions:**
  - Text documents are observation. The words in the vocabulary are organized as a topic and documents are made from the words that are drawn from the topics
  - Collection of documents can be described in terms of topics that are hidden and common across the documents
  - So the model is formulated as: Given how words co-occur in documents we can infer topics

(http://www.cs.columbia.edu/~blei/topicmodeling_software.html)

Words of caution!

- Social media text are short and noisy
- Analytics is not equal to automation! manual validation of results are still important
- Several iteration of cleaning and pre-processing to improve the results
- Methods that rely on co-occurrence statistics (e.g. topic model) suffer from sparsity (spreading words)
- Methods that uses context window statistics (word embeddings) tends to perform better (can be useful in reducing variations of words)

Python demo (text pre-processing)

Stemming & Lemmatization - GUI interface

http://text-processing.com/demo/stem/

Next lecture – 16.10.2020

- Thematic analysis
- Sentiment analysis
- Social network analysis

Thank you