Text Mining in Social Media

Text, Web and Social Media Analytics Lab

Prof. Dr. Diana Hristova

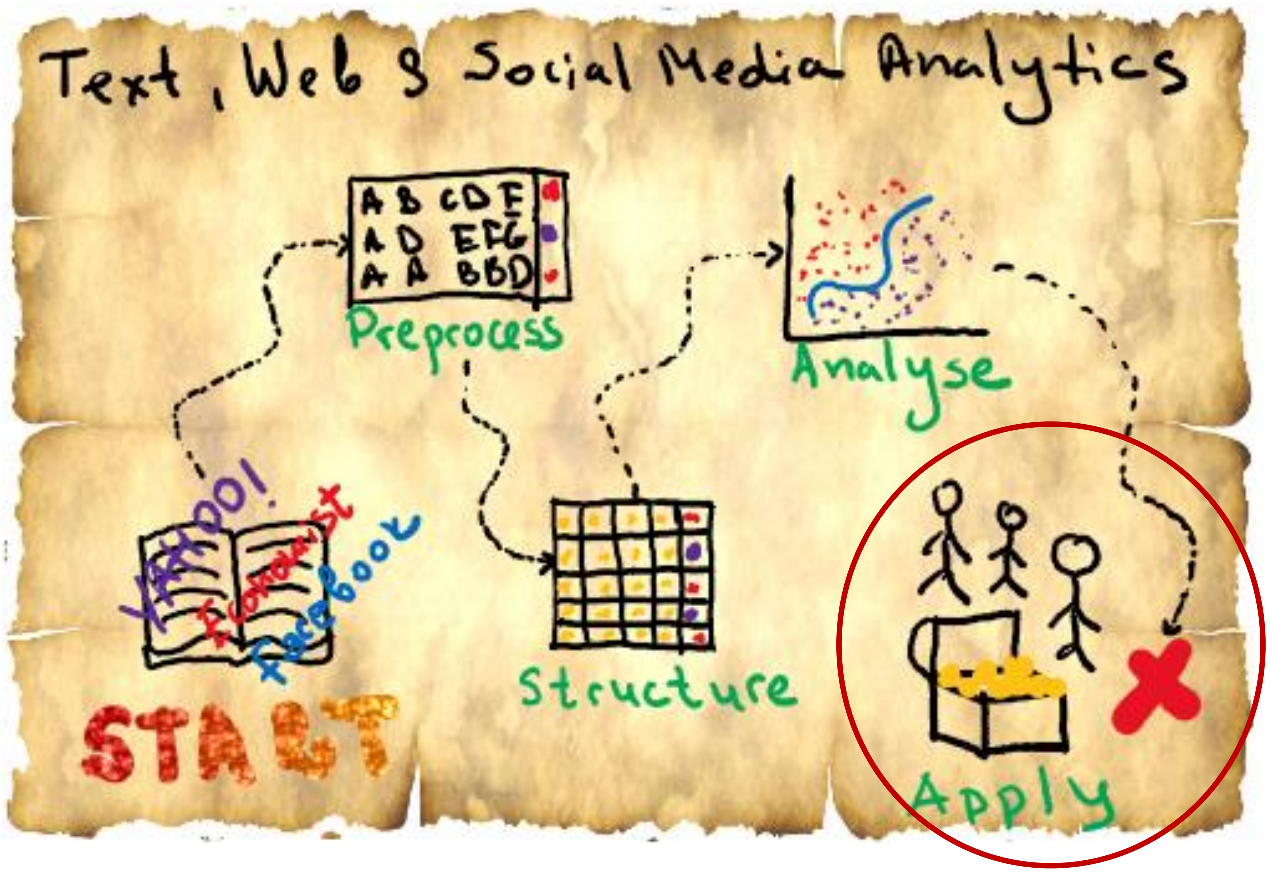
Exercise 6: Clustering

Can one group present please?

Summary

What did we learn last week?

Social Media Analytics: Treasury map



Course structure

Date	Lecture	Exercise
12.04.2021	Introduction	Technical Installation
19.04.2021	Text Preprocessing	Projects kick-off
26.04.2021	Text Representation	Preprocessing Newsgroups
03.05.2021	Text Representation (2)	Text Representation Newsgroups
10.05.2021	Text Classification	Text Representation Newsgroups (2)
17.05.2021	Text Clustering/Capgemni	Newsgroups Topic Classification
31.05.2021	Text Mining in Social Media	Newsgroups Topic Clustering
07.06.2021	Mining Social Graphs	Sentiment Analysis and Time Series in Twitter
14.06.2021	Projects Status Update	Projects Status Update
21.06.2021	Web Analytics	Mining Social Graphs in Twitter
28.06.2021	Mock Exam	Web Analytics in E-commerce
05.07.2021	Final Presentation	Final Presentation
19.07.2021	Submit Code & Written report	
t.b.a.	Exam	

What will we learn today?

At the end of this lecture, you will:

1. Know what is social media and why social media analytics matters.
2. Be aware of the main social media sites and possibilities to access their data.
3. Understand the special nature of text analytics and its applications for social media data.

What is social media?

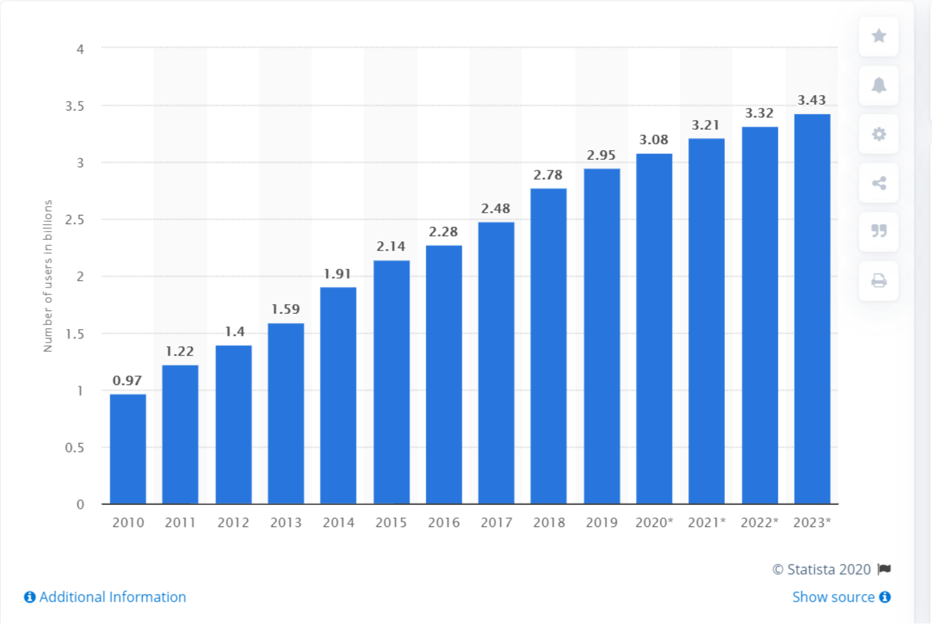
Social Media: Introduction

What is social media?

- **Web-based** applications that allow users to **share different types of content** such as comments, personal information, photos, videos, external content.
- They "…facilitate the development of **social networks** online by connecting a profile with those of other individuals and/or groups." (Obar, J,. 2015, p. 2).

Number of social network users worldwide from 2010 to 2023

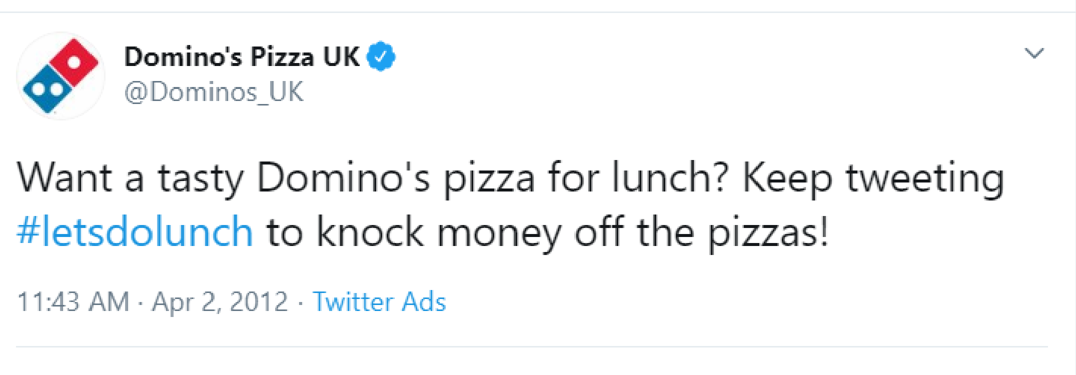
(in billions)



Source: https://www.statista.com/

Obar, Jonathan A., and Steven S. Wildman. "Social media definition and the governance challenge-an introduction to the special issue." *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy* 39, no. 9 (2015): 745-750.

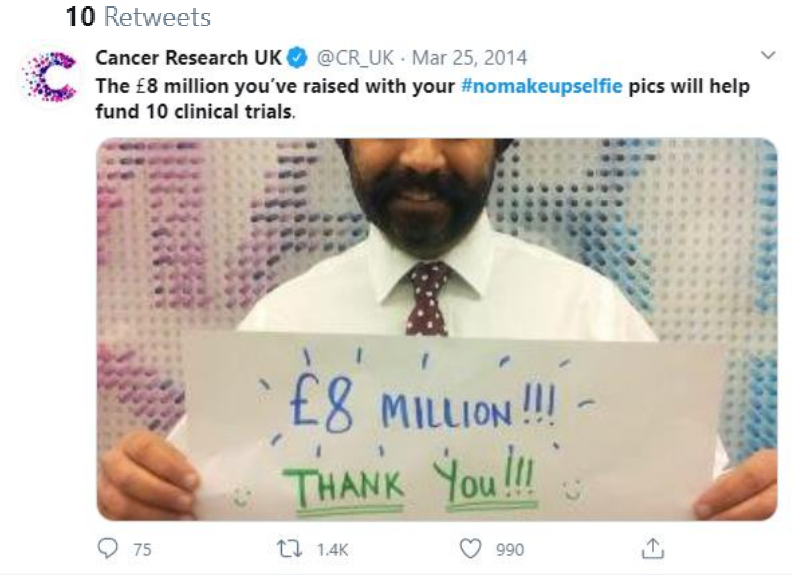
Social Media: Success stories



Domino's Pizza UK
@Dominos_UK

Want a tasty Domino's pizza for lunch? Keep tweeting #letsdolunch to knock money off the pizzas!

11:43 AM · Apr 2, 2012 · Twitter Ads

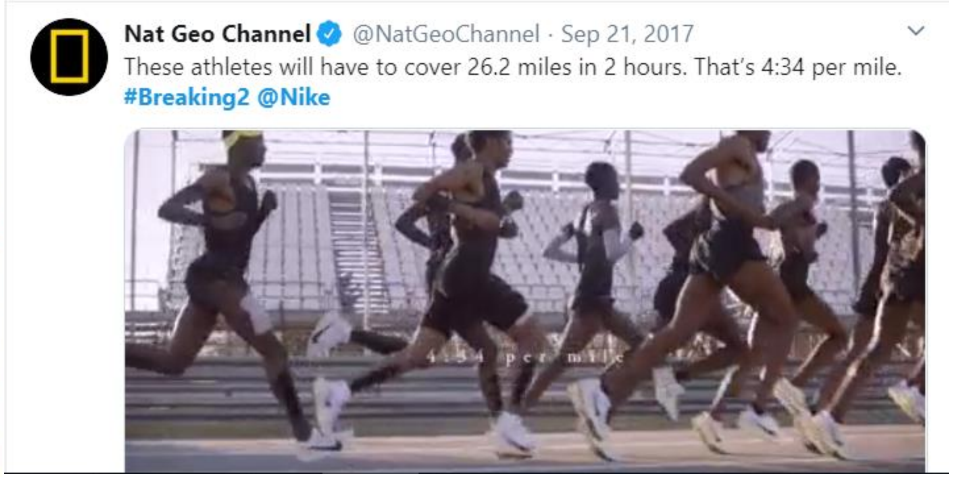


10 Retweets

Cancer Research UK @CR_UK · Mar 25, 2014

The £8 million you've raised with your #nomakeupselfie pics will help fund 10 clinical trials.

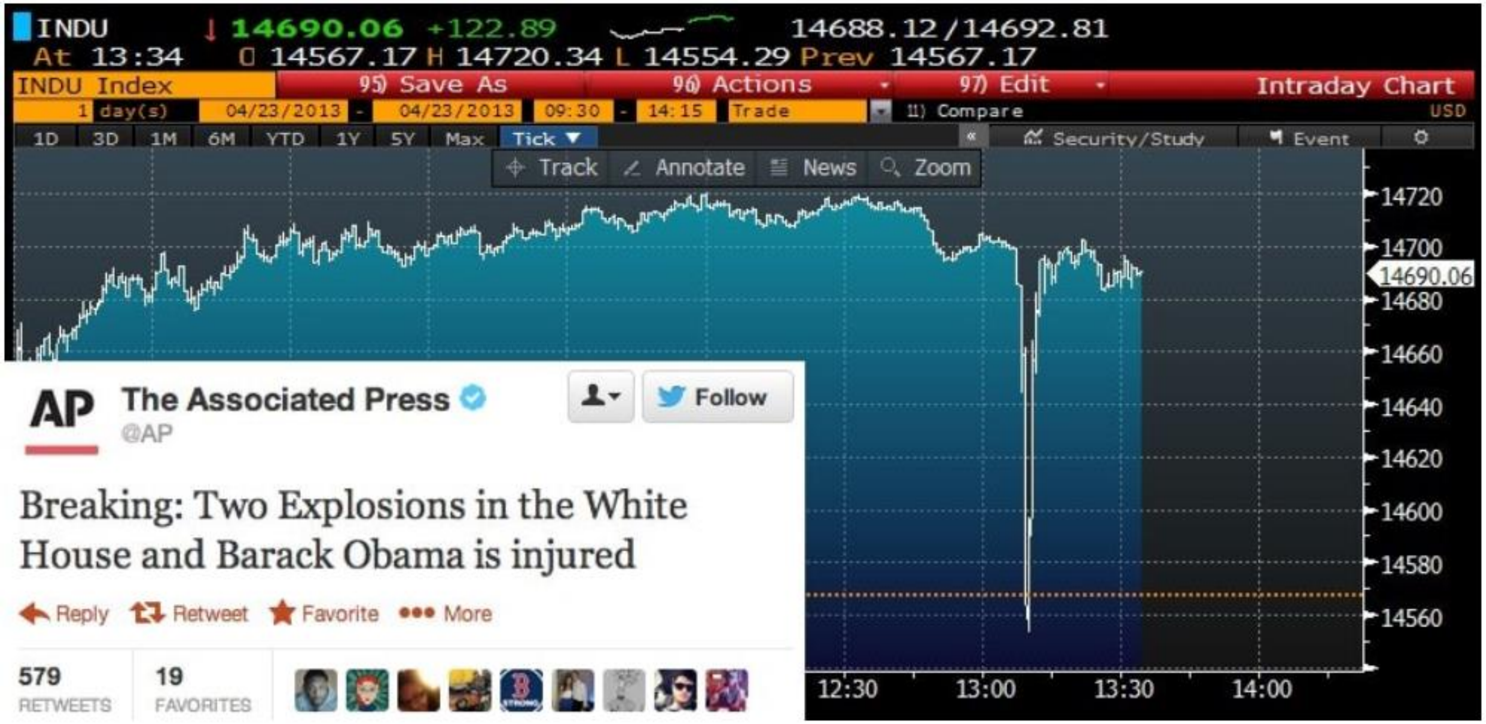
75 1.4K 990



Nat Geo Channel @NatGeoChannel · Sep 21, 2017

These athletes will have to cover 26.2 miles in 2 hours. That's 4:34 per mile. #Breaking2 @Nike

Social Media: Success stories (?)



Breaking: Two Explosions in the White House and Barack Obama is injured

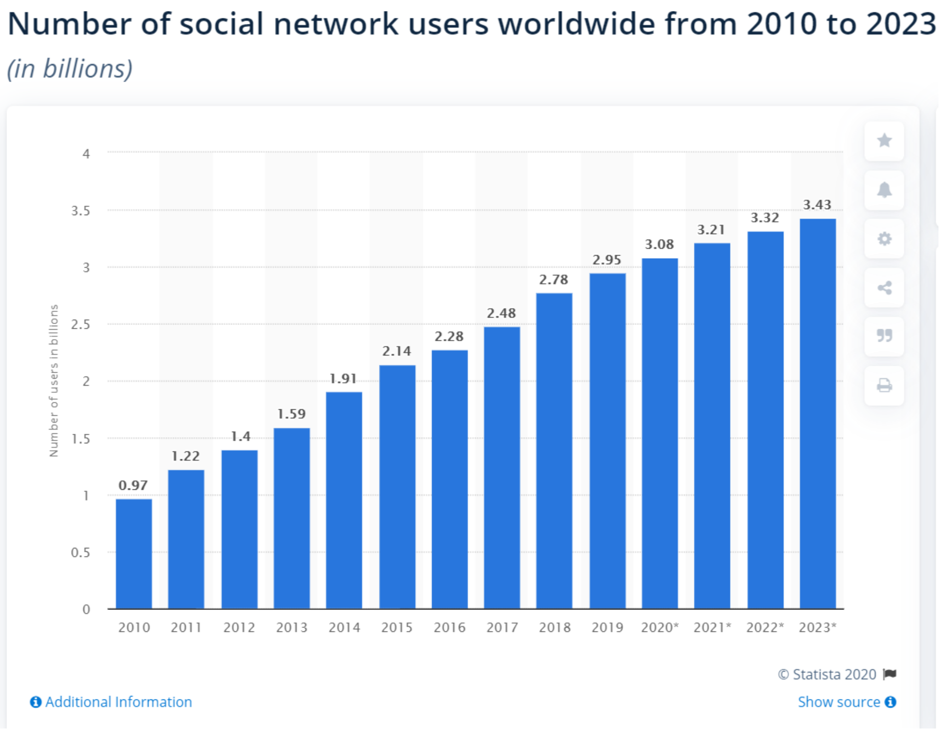
This chart shows the Dow Jones Industrial Average during Tuesday afternoon's drop, caused by a fake A.P. tweet, inset at left.

https://www.washingtonpost.com/news/worldviews/wp/2013/04/23/syrian-hackers-claim-ap-hack-that-tipped-stock-market-by-136-billion-is-it-terrorism/

Social Media: Introduction

What is social media?

- **Web-based** applications that allow users to **share different types of content** such as comments, personal information, photos, videos, external content.
- They "…facilitate the development of **social networks** online by connecting a profile with those of other individuals and/or groups." (Obar, J,. 2015, p. 2).
- **Social media analytics:** gathering and analysing the **data** generated by social media.



Source: https://www.statista.com/

Obar, Jonathan A., and Steven S. Wildman. "Social media definition and the governance challenge-an introduction to the special issue." *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy* 39, no. 9 (2015): 745-750.

Social Media Analytics: Why it matters?

Why does social media analytics matter?

Social Media Analytics: Why it matters?

- Social Media is so important that companies can‘t ignore it.
- **But:** a bad social media strategy can have devastating consequences

➔ Develop a strategy based on the data by…

…understanding the status quo (as-is):

- Who are your customers (e.g. age)?
- How do they feel about your company/product?
- How did this develop over time (e.g. new products)?
- How are your competitors doing in social media?

…and improving the future (to-be):

- How to improve users‘ perception of your company/product?
- How to increase your customer base?

Social Media: Main Sites

Which social media site is the most popular one worldwide (number of active users)?

a: Facebook

b: Twitter

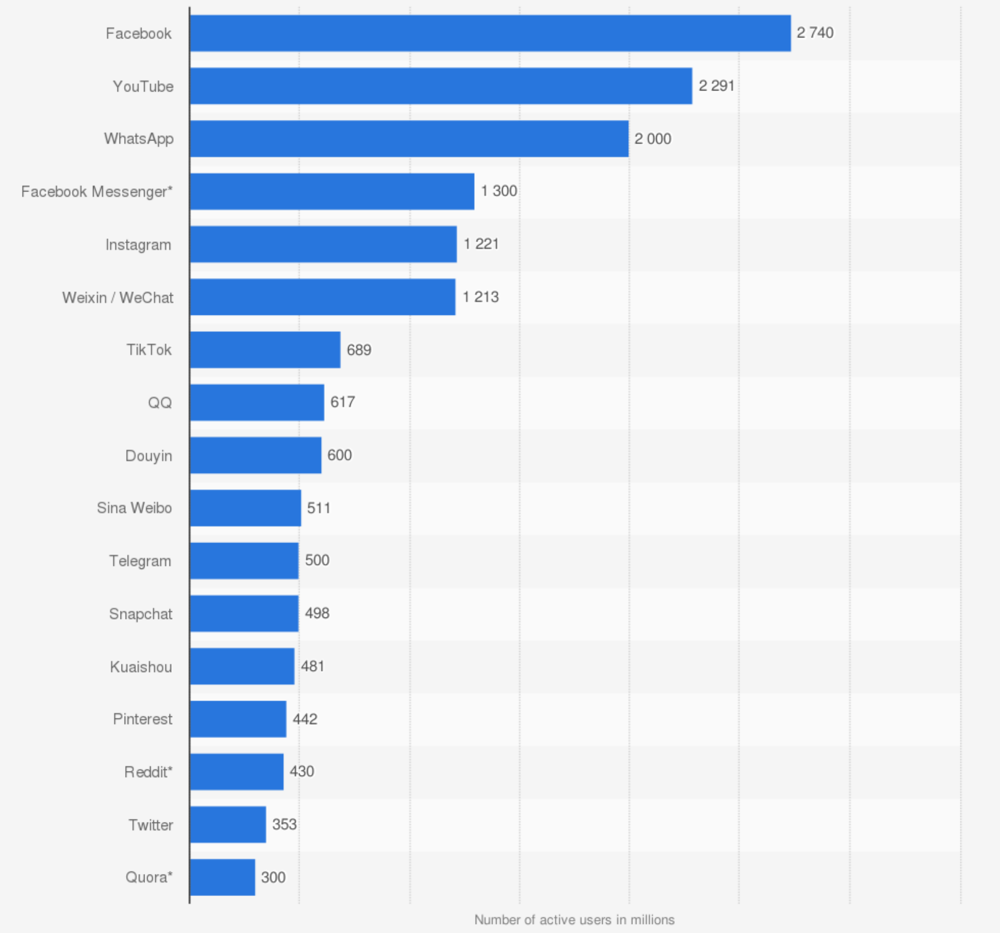
c: YouTube

d: WeChat

e: LinkedIn

Social Media: Main Sites (2)

Most popular social networks worldwide as of January 2021, ranked by number of active users (in millions)



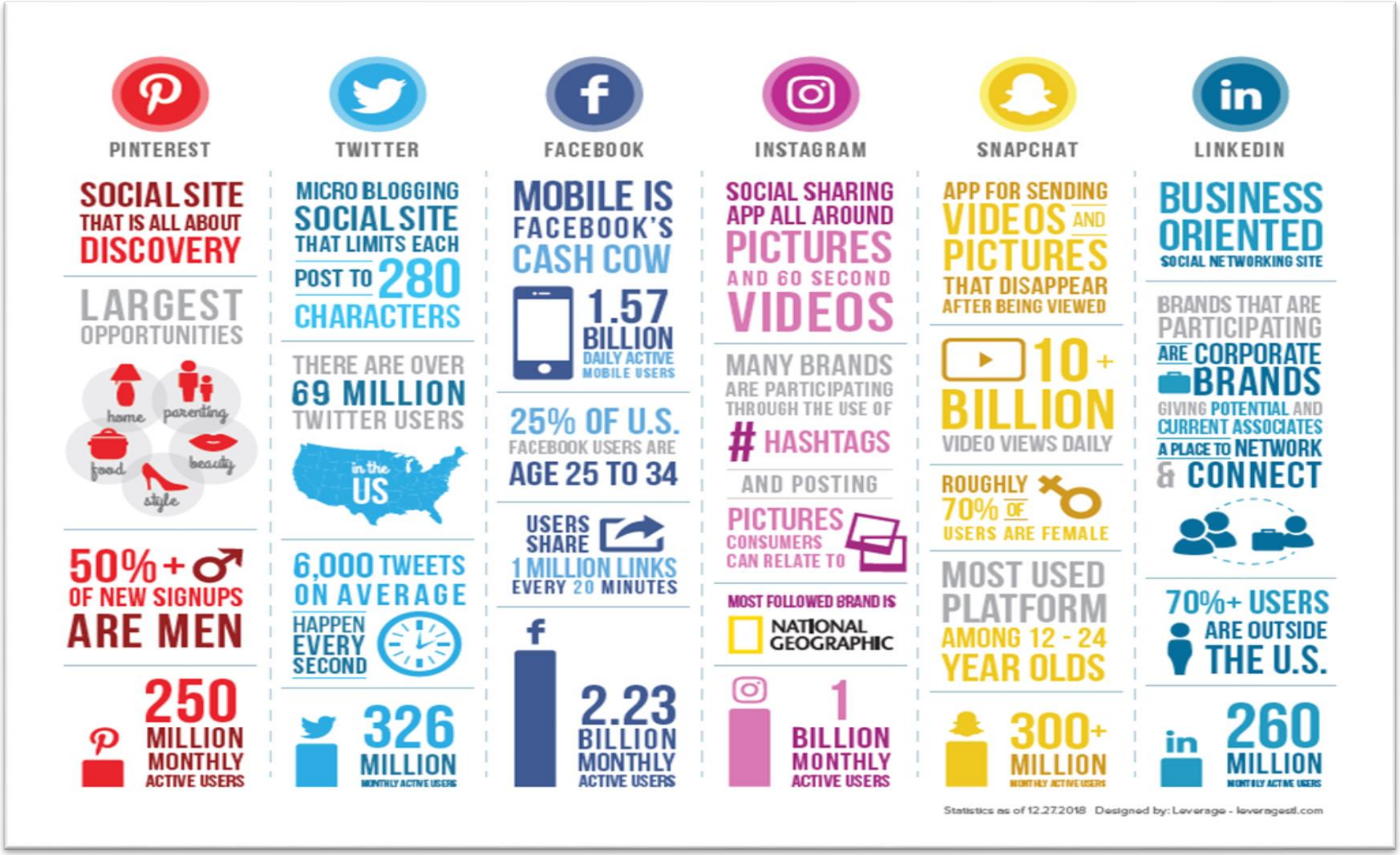
Sources
We Are Social; Various sources (Company data); Hootsuite; DataReportal
© Statista 2021

Additional Information:
Worldwide; Various sources (Company data); DataReportal; January, 2021; social networks and messenger/chat app/voip not include Douyin

https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Social Media Analytics: Main Sites

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law



PINTEREST

SOCIAL SITE THAT IS ALL ABOUT DISCOVERY

LARGEST OPPORTUNITIES

50%+ ♂ OF NEW SIGNUPS ARE MEN

250 MILLION MONTHLY ACTIVE USERS

TWITTER

MICRO BLOGGING SOCIAL SITE THAT LIMITS EACH POST TO 280 CHARACTERS

THERE ARE OVER 69 MILLION TWITTER USERS in the US

6,000 TWEETS ON AVERAGE HAPPEN EVERY SECOND

326 MILLION MONTHLY ACTIVE USERS

FACEBOOK

MOBILE IS FACEBOOK'S CASH COW

1.57 BILLION DAILY ACTIVE MOBILE USERS

25% OF U.S. FACEBOOK USERS ARE AGE 25 TO 34

USERS SHARE 1 MILLION LINKS EVERY 20 MINUTES

2.23 BILLION MONTHLY ACTIVE USERS

INSTAGRAM

SOCIAL SHARING APP ALL AROUND PICTURES AND 60 SECOND VIDEOS

MANY BRANDS ARE PARTICIPATING THROUGH THE USE OF # HASHTAGS AND POSTING PICTURES CONSUMERS CAN RELATE TO

MOST FOLLOWED BRAND IS NATIONAL GEOGRAPHIC

1 BILLION MONTHLY ACTIVE USERS

SNAPCHAT

APP FOR SENDING VIDEOS AND PICTURES THAT DISAPPEAR AFTER BEING VIEWED

10+ BILLION VIDEO VIEWS DAILY

ROUGHLY 70% OF USERS ARE FEMALE

MOST USED PLATFORM AMONG 12 - 24 YEAR OLDS

300+ MILLION MONTHLY ACTIVE USERS

LINKEDIN

BUSINESS ORIENTED SOCIAL NETWORKING SITE

BRANDS THAT ARE PARTICIPATING ARE CORPORATE BRANDS GIVING POTENTIAL AND CURRENT ASSOCIATES A PLACE TO NETWORK & CONNECT

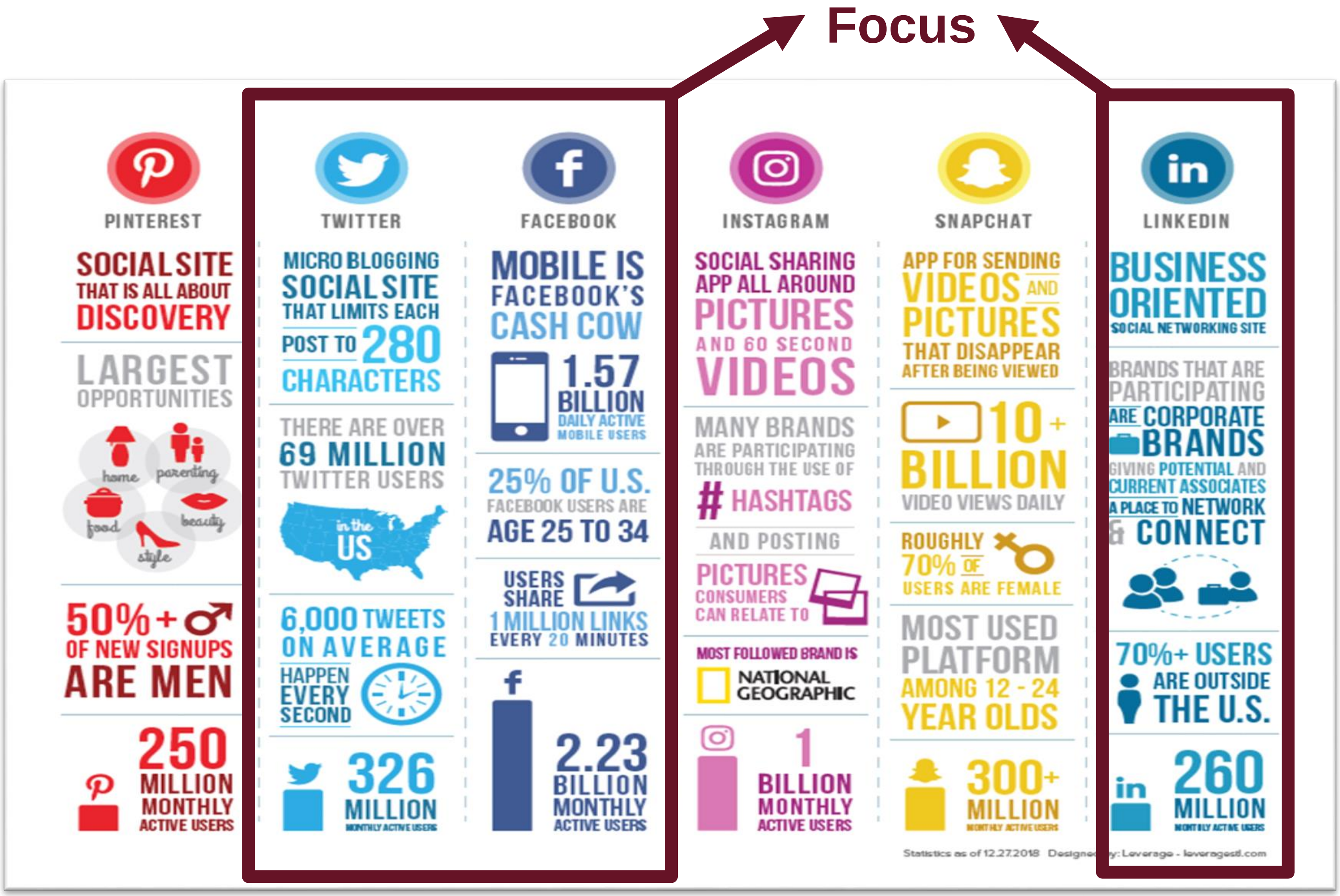
70%+ USERS ARE OUTSIDE THE U.S.

260 MILLION MONTHLY ACTIVE USERS

Statistics as of 12.27.2018 Designed by: Leverage - leveragestl.com

Source: https://www.leveragestl.com/social-media-infographic

Social Media Analytics: Main Sites



Focus

PINTEREST

SOCIAL SITE THAT IS ALL ABOUT DISCOVERY

LARGEST OPPORTUNITIES

home
parenting
food
beauty
style

50%+ ♂ OF NEW SIGNUPS ARE MEN

250 MILLION MONTHLY ACTIVE USERS

TWITTER

MICRO BLOGGING SOCIAL SITE THAT LIMITS EACH POST TO 280 CHARACTERS

THERE ARE OVER 69 MILLION TWITTER USERS in the US

6,000 TWEETS ON AVERAGE HAPPEN EVERY SECOND

326 MILLION MONTHLY ACTIVE USERS

FACEBOOK

MOBILE IS FACEBOOK'S CASH COW

1.57 BILLION DAILY ACTIVE MOBILE USERS

25% OF U.S. FACEBOOK USERS ARE AGE 25 TO 34

USERS SHARE 1 MILLION LINKS EVERY 20 MINUTES

2.23 BILLION MONTHLY ACTIVE USERS

INSTAGRAM

SOCIAL SHARING APP ALL AROUND PICTURES AND 60 SECOND VIDEOS

MANY BRANDS ARE PARTICIPATING THROUGH THE USE OF # HASHTAGS AND POSTING PICTURES CONSUMERS CAN RELATE TO

MOST FOLLOWED BRAND IS NATIONAL GEOGRAPHIC

1 BILLION MONTHLY ACTIVE USERS

SNAPCHAT

APP FOR SENDING VIDEOS AND PICTURES THAT DISAPPEAR AFTER BEING VIEWED

10+ BILLION VIDEO VIEWS DAILY

ROUGHLY 70% OF USERS ARE FEMALE

MOST USED PLATFORM AMONG 12 - 24 YEAR OLDS

300+ MILLION MONTHLY ACTIVE USERS

LINKEDIN

BUSINESS ORIENTED SOCIAL NETWORKING SITE

BRANDS THAT ARE PARTICIPATING ARE CORPORATE BRANDS GIVING POTENTIAL AND CURRENT ASSOCIATES A PLACE TO NETWORK & CONNECT

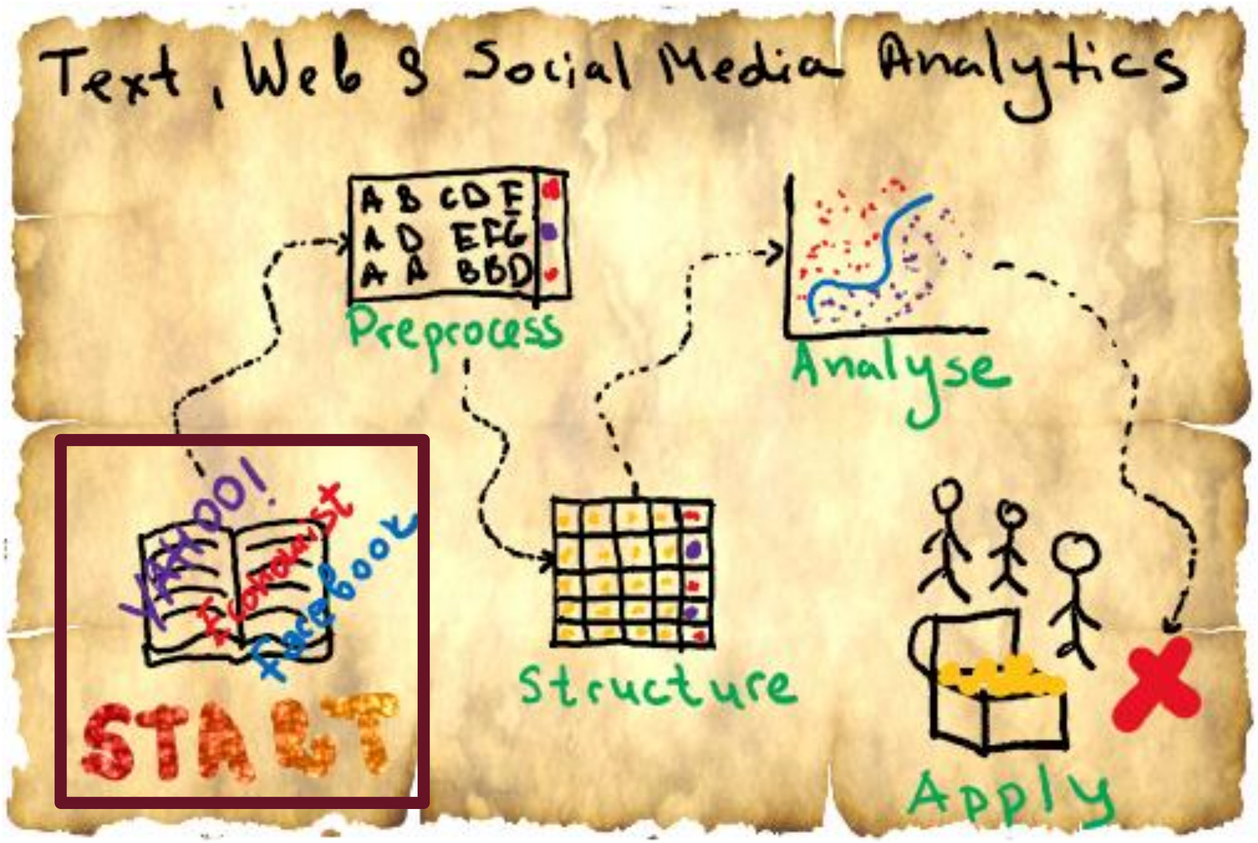
70%+ USERS ARE OUTSIDE THE U.S.

260 MILLION MONTHLY ACTIVE USERS

Statistics as of 12.27.2018 Designed by: Leverage - leveragestl.com

Source: https://www.leveragestl.com/social-media-infographic

Social Media Analytics: Treasury map

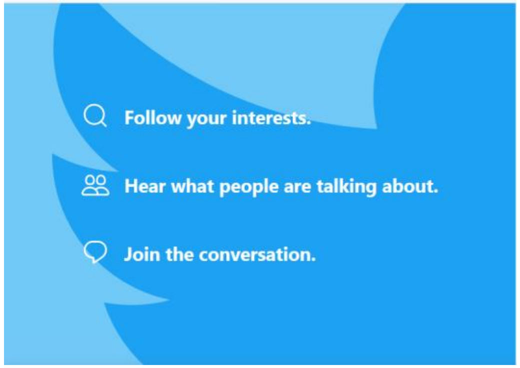


Social Media Analytics: Twitter

Main idea: ‚..real-time, highly social microblogging service…‘ (Russell and Klassen 2018, p. 9) for which you don‘t need to be connected with the user to consume content. Posts are called ‚tweets‘.

Available data (from https://developer.twitter.com/):

- ❑ **Search API:** retrieve historical tweets for the last 7 days for free, for older tweets you need a premium/enterprise account.
- ❑ **Streaming API:** real-time data based on keywords or as a random sample of real-time public tweets
- ❑ **Batch API:** “…full archive of public Twitter data.”, only Enterprise version
- ❑ **Ads API:** get data gathered by your Twitter Ad



Follow your interests.

Hear what people are talking about.

Join the conversation.

? What can Twitter data be used for?

Russell, Matthew A., and Mikhail Klassen. *Mining the social web: data mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and more*. O'Reilly Media, 2018.

Social Media Analytics: Twitter

Main idea: ‚..real-time, highly social microblogging service…‘ (Russell and Klassen 2018, p. 9) for which you don‘t need to be connected with the user to consume content. Posts are called ‚tweets‘.

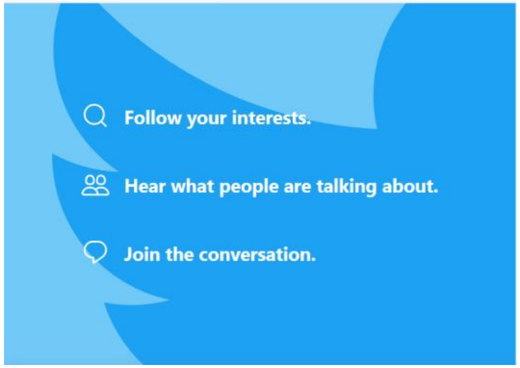
Available data (from https://developer.twitter.com/):

❑ **Search API:** retrieve historical tweets for the last 7 days for free, for older tweets you need a premium/enterprise account.

❑ **Streaming API:** real-time data based on keywords or as a random sample of real-time public tweets

❑ **Batch API:** "…full archive of public Twitter data.", only Enterprise version

❑ **Ads API:** get data gathered by your Twitter Ad



Follow your interests.

Hear what people are talking about.

Join the conversation.

Applications: celebrities' reach, sentiment for a company, product campaigns, trending topics, etc.,

Russell, Matthew A., and Mikhail Klassen. *Mining the social web: data mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and more*. O'Reilly Media, 2018.

Social Media Analytics: Facebook

Main idea: a social network in which users can conduct different activities between each other such as likes, comments, photos, tags. As opposed to Twitter, not all content is public and content visibility is guided by a friendship relationship.

Available data: you can get users' Facebook data only, if you register an application and users authorize it to access their data. The data is for free.

- ❑ **Facebook Graph API:** main access point to Facebook data such as users, posts, groups, events.
- ❑ **Facebook Marketing API:** a tool for automatically managing your marketing campaign on Facebook.

facebook
Facebook helps you connect and share with the people in your life.

? What can Facebook data be used for?

Social Media Analytics: Facebook

Main idea: a social network in which users can conduct different activities between each other such as likes, comments, photos, tags. As opposed to Twitter, not all content is public and content visibility is guided by a friendship relationship.

Available data: you can get users' Facebook data only, if you register an application and users authorize it to access their data. The data is for free.

- ❑ **Facebook Graph API:** main access point to Facebook data such as users, posts, groups, events.
- ❑ **Facebook Marketing API:** a tool for automatically managing your marketing campaign on Facebook.

Facebook helps you connect and share with the people in your life.

Applications: product campaigns, reach new customers, topic clusters, geo analytics, user similarity, etc.

Social Media Analytics: LinkedIn

LinkedIn

Willkommen in Ihrer beruflichen Community

Main idea: a professional social network where people provide their job history, professional interests and wishes and discuss and comment on professional topics.

Available data: due to the sensitive nature of the data, no graph data is available.

❑ **Profile API:** profile information for a given user, visible if public or the requestor is allowed to see

❑ **Connections API:** "..a list of 1st degree connections for a user who has granted access to his/her account." (linkedin.com)

❑ …..

? What can LinkedIn data be used for?

Social Media Analytics: LinkedIn

Linked in

Willkommen in Ihrer beruflichen Community

Main idea: a professional social network where people provide their job history, professional interests and wishes and discuss and comment on professional topics, can get a copy of your data.

Available data: due to the sensitive nature of the data, no graph data is available.

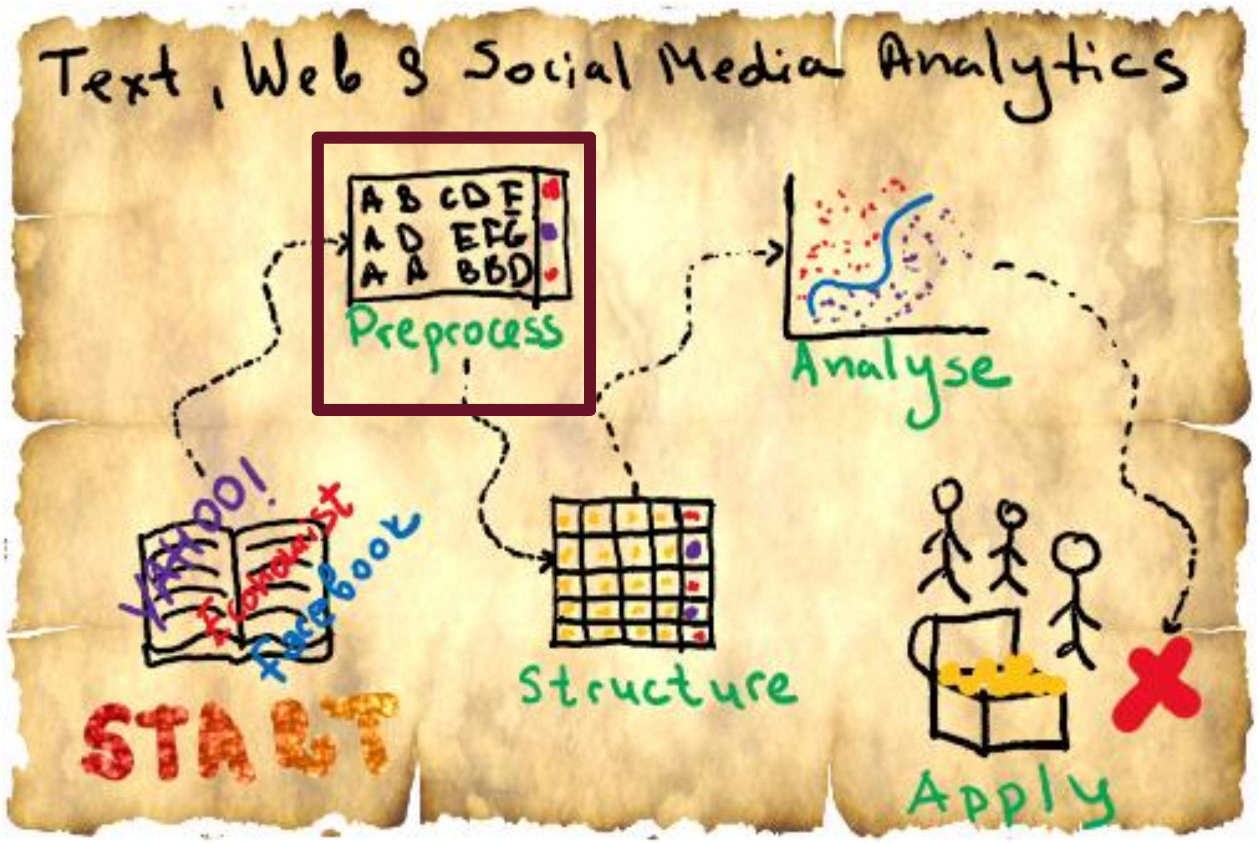
❑ **Profile API:** profile information for a given user, visible if public or the requestor is allowed to see

❑ **Connections API:** “..a list of 1st degree connections for a user who has granted access to his/her account.” (linkedin.com)

❑ …..

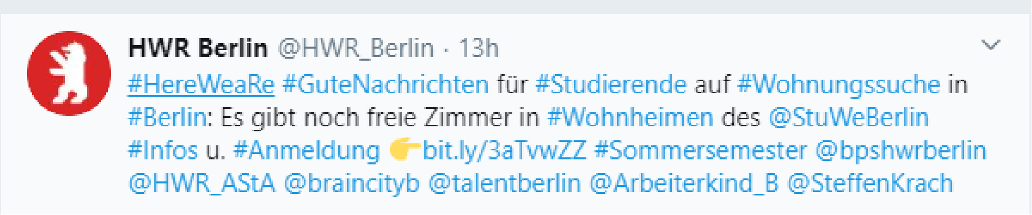
Applications: cluster users in professional groups and work experience, find the most trending professional topics

Social Media Analytics: Treasury map



Text Preprocessing with Twitter

Why do social media texts require additional preprocessing than other texts such as books?



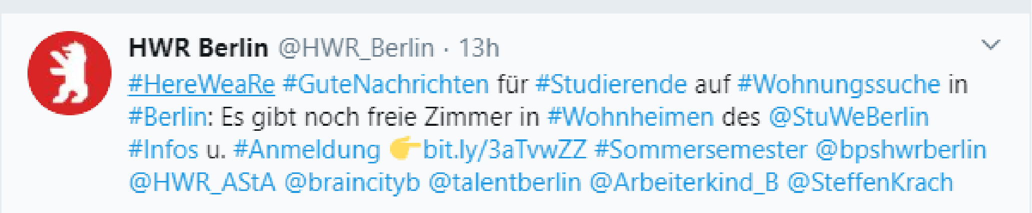
HWR Berlin @HWR_Berlin · 13h

#HereWeaRe #GuteNachrichten für #Studierende auf #Wohnungssuche in #Berlin: Es gibt noch freie Zimmer in #Wohnheimen des @StuWeBerlin #Infos u. #Anmeldung 👉bit.ly/3aTvwZZ #Sommersemester @bpshwrberlin @HWR_AStA @braincityb @talentberlin @Arbeiterkind_B @SteffenKrach

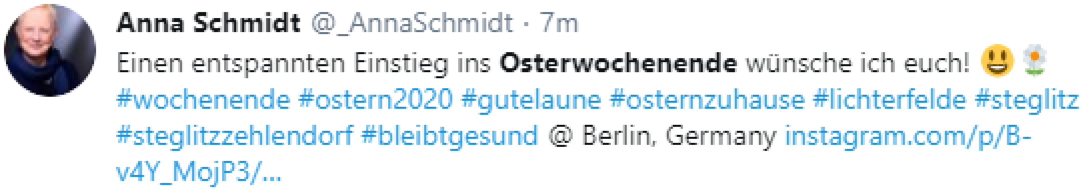
Text Preprocessing in Social Media Analytics

Social Media texts are different, because they contain non-language parts such as:

- URL: e.g. bit.ly/3aTvwZZ
- Hashtags: #StuWeBerlin
- Mentions: @HWR_AStA
- Reserved words: RT(=Retweet), FAV(=Favourite Message)
- Emojies: 😃 👋
- Smilies: ☺, ☹



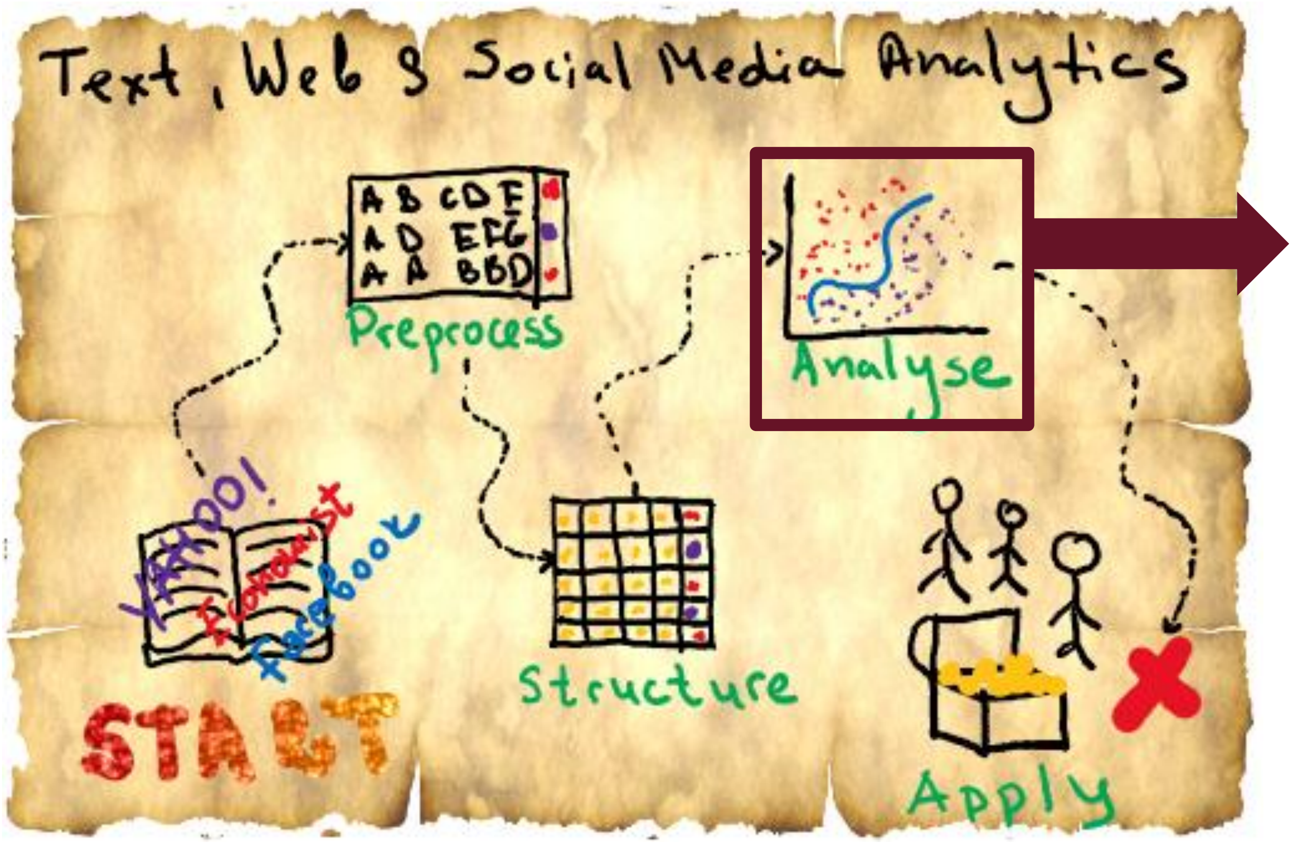
HWR Berlin @HWR_Berlin · 13h
#HereWeaRe #GuteNachrichten für #Studierende auf #Wohnungssuche in #Berlin: Es gibt noch freie Zimmer in #Wohnheimen des @StuWeBerlin #Infos u. #Anmeldung 👉bit.ly/3aTvwZZ #Sommersemester @bpshwrberlin @HWR_AStA @braincityb @talentberlin @Arbeiterkind_B @SteffenKrach



Anna Schmidt @_AnnaSchmidt · 7m
Einen entspannten Einstieg ins **Osterwochenende** wünsche ich euch! 😃🌼
#wochenende #ostern2020 #gutelaune #osternzuhause #lichterfelde #steglitz #steglitzzehlendorf #bleibtgesund @ Berlin, Germany instagram.com/p/B-v4Y_MojP3/...

Additional preprocessing to reduce noise.

Social Media Analytics: Treasury map

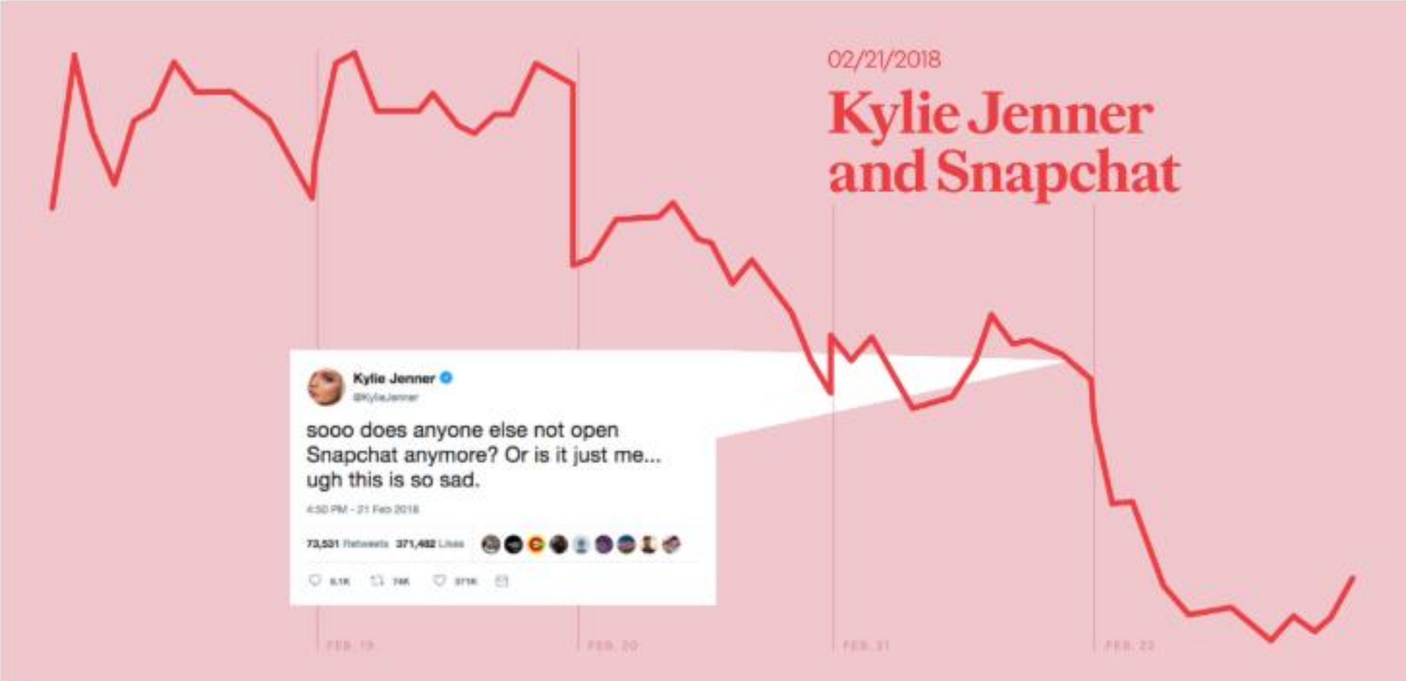


- Sentiment analysis
- Network analysis

Text Analysis in Social Media Analytics: Sentiment

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

- As opposed to other texts such as news, texts in social media are emotional and subjective.
- Due to the network structure, emotions amplify and can have serious consequences.



Kylie Jenner
@KylieJenner

sooo does anyone else not open Snapchat anymore? Or is it just me... ugh this is so sad.

4:50 PM - 21 Feb 2018

73,531 Retweets 371,482 Likes

6.1K 74K 371K

Source: https://www.ogilvy.com/feed/11-tweets-that-turned-the-stock-market-upside-down/

Text Analysis in Social Media Analytics: Sentiment (2)

- Sentiment Analysis aims at determining whether a piece of text is positively, neutrally or negatively mooded.

Example:

- „The movie was horrible, I definitely don‘t recommend it.“ ➔ negative
- „It was a fantastic movie with a very interesting plot.“ ➔ positive
- „I went to see the movie yesterday.“➔ neutral

▶ Calculate social media sentiment in Python

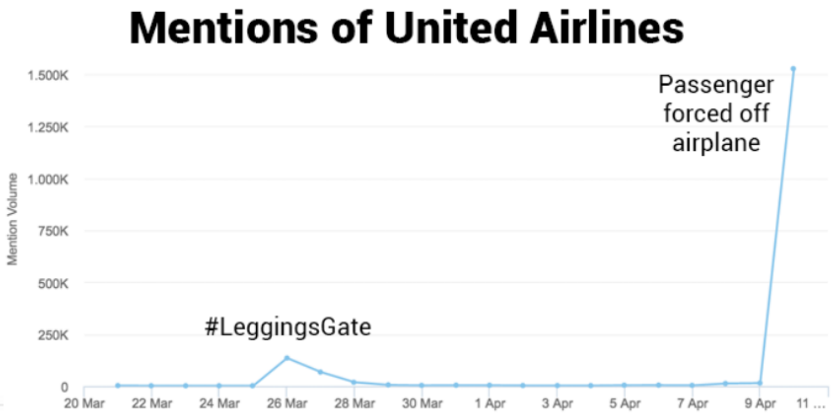
Comparison of Python Sentiment Packages

Package	Output	Method	Comment
VADER (NLTK)	• Negative • Neutral • Positive • Compound in $[-1,1]$, combination of positive and negative	Rule/lexicon-based, word based	Good for short texts, handles smileys
CoreNLP	0: very negative 1: negative 2: neutral 3: positive 4: very positive	Tree-based and neural network models	Uses a deep learning model that has to be loaded
TextBlob	• Polarity in $[-1,1]$ • Subjectivity in $[0,1]$	Pattern analyser and Naïve Bayes analyser (trained on movie reviews)	Based on the most commonly occurring negative and positive adjectives (Pattern)

Text Analysis in Social Media Analytics: Time Series Analysis

Why it matters?

- Examine effects of social media campaigns (e.g. #Breaking2)
- Identify trends in sentiment and react
- Examine trending topics over time



Twitter, Facebook & Instagram analysis via Brandwatch | 20 March - 10 April 2017

Jayse D. Anspach @JayseDavid · Apr 10, 2017

@United overbook #flight3411 and decided to force random passengers off the plane. Here's how they did it:

Summary and Outlook

Summary:

- Social media content quality can have serious positive and negative consequences on products and companies.
- Main social media sites such as Twitter, Facebook and LinkedIn offer free access to their data via APIs.
- Text preprocessing for social media content requires additional cleaning.
- Sentiment analysis in Python can be done with VADER, TextBlob or NLPCore.
- Time series analysis in social media is crucial for product campaigns and long-term sentiment analysis.

Outlook: Social media data has an underlying network structure that can be analysed to determine most influential users, topic clusters and information spread dynamics.

Questions?

Exercise 7

In a minute, six break-out rooms will be created. Choose the room that corresponds to your group in Moodle e.g. Room 1= Group 1. In your project group discuss and document the solution for Exercise 7 (in Moodle).