# Social Media

Uijeong “UJ” Hwang

# What Qualifies as “Social Media”?

- Social media is digital technology that allows the sharing of ideas, information, events, and experiences through text and visuals within virtual networks and communities.

# What Qualifies as “Social Media”?

- Social media is digital technology that allows the sharing of ideas, information, events, and experiences through text and visuals within virtual networks and communities.
- It typically features user-generated content that seeks to engage a wider audience through likes, shares, comments, and discussions.

# What Qualifies as “Social Media”?

- Social media is digital technology that allows the sharing of ideas, information, events, and experiences through text and visuals within virtual networks and communities.
- It typically features user-generated content that seeks to engage a wider audience through likes, shares, comments, and discussions.
- Types of social media:
  - Social Networks
  - Media Sharing Networks
  - Discussion Forums/Communities
  - Microblogging
  - Messaging
  - Professional/Business

# Most Widely Used Social Media Platforms

Ranking of social media platforms by monthly active users (as of early 2024):

1. **Facebook**: ~3 billion
2. **YouTube**: ~2.5 billion
3. **WhatsApp**: ~2.4 billion
4. **Instagram**: ~2.3 billion
5. **WeChat**: ~1.3 billion
6. **TikTok**: ~1.2 billion
7. **LinkedIn**: ~930 million
8. **Snapchat**: ~750 million
9. **Reddit**: ~550 million
10. **Twitter/X**: ~500 million

# Types of Social Media Data

- **Textual Data**
  - User posts, comments, and replies as sources of sentiment and discussion.
- **Interaction Data**
  - Likes, upvotes, replies, and shares as indicators of engagement and popularity.
- **User Metadata**
  - Location (when available), post timing, and user demographics.

# Is Social Media Useful?

Five Characteristics of Big Data

- Volume: The amount of data being generated
- Velocity: The speed at which data is generated
- Variety: The different types of data being generated (structured, semi-structured, or unstructured)
- Veracity: The trustworthiness of the data
- Value: The value of the data

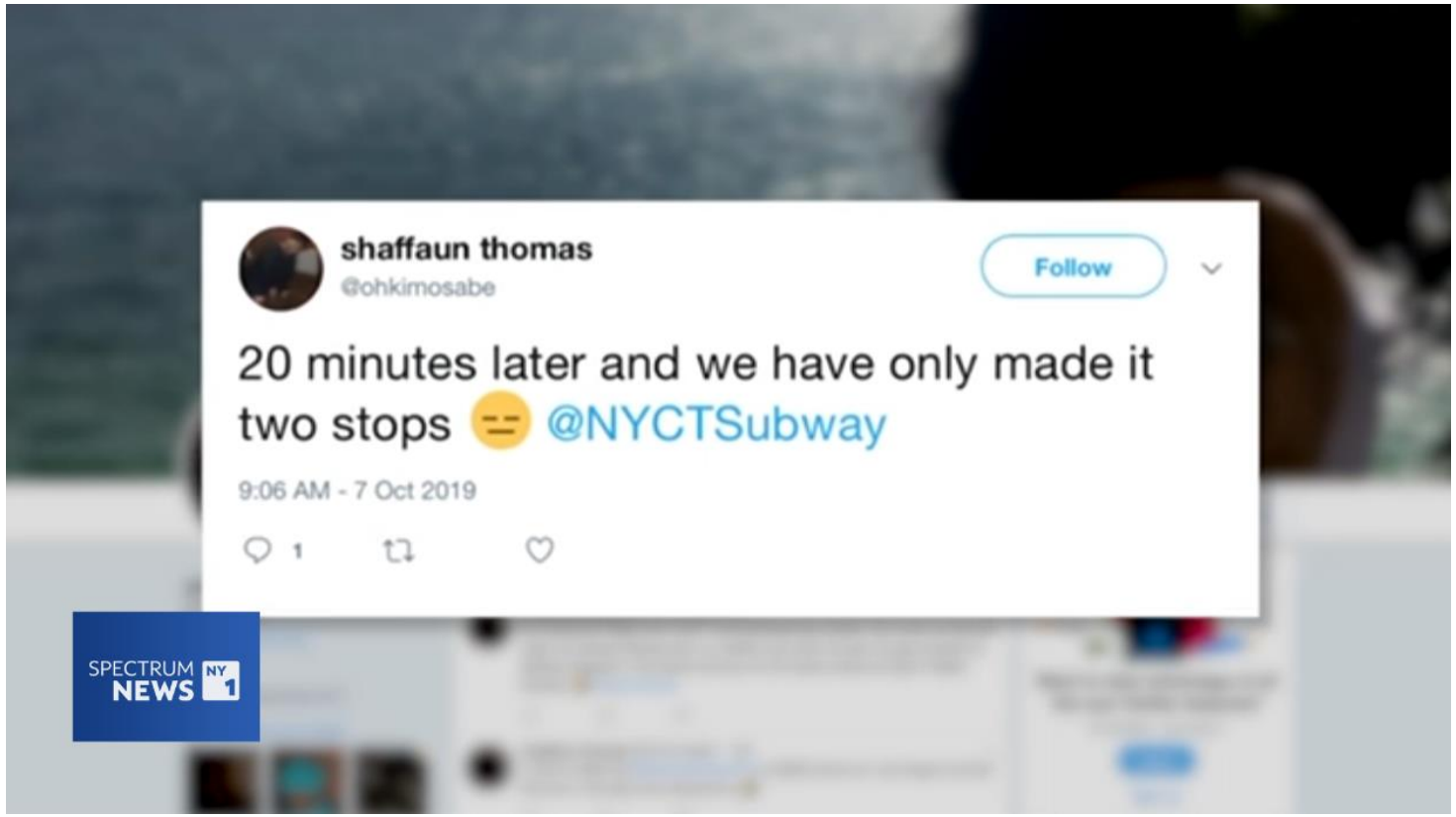
# Is Social Media Useful?

Five Characteristics of Big Data

- Volume: The amount of data being generated 👍
- Velocity: The speed at which data is generated 👍
- Variety: The different types of data being generated (structured, semi-structured, or unstructured) 👍
- Veracity: The trustworthiness of the data 👎
- Value: The value of the data 👍

# Leveraging Social Media: Urban mobility

- The NYC MTA uses social media to track real-time feedback from subway riders regarding delays, overcrowding, and service interruptions.




https://ny1.com/nyc/manhattan/news/2019/10/11/a-look-inside-the-mta-s-social-media-operations

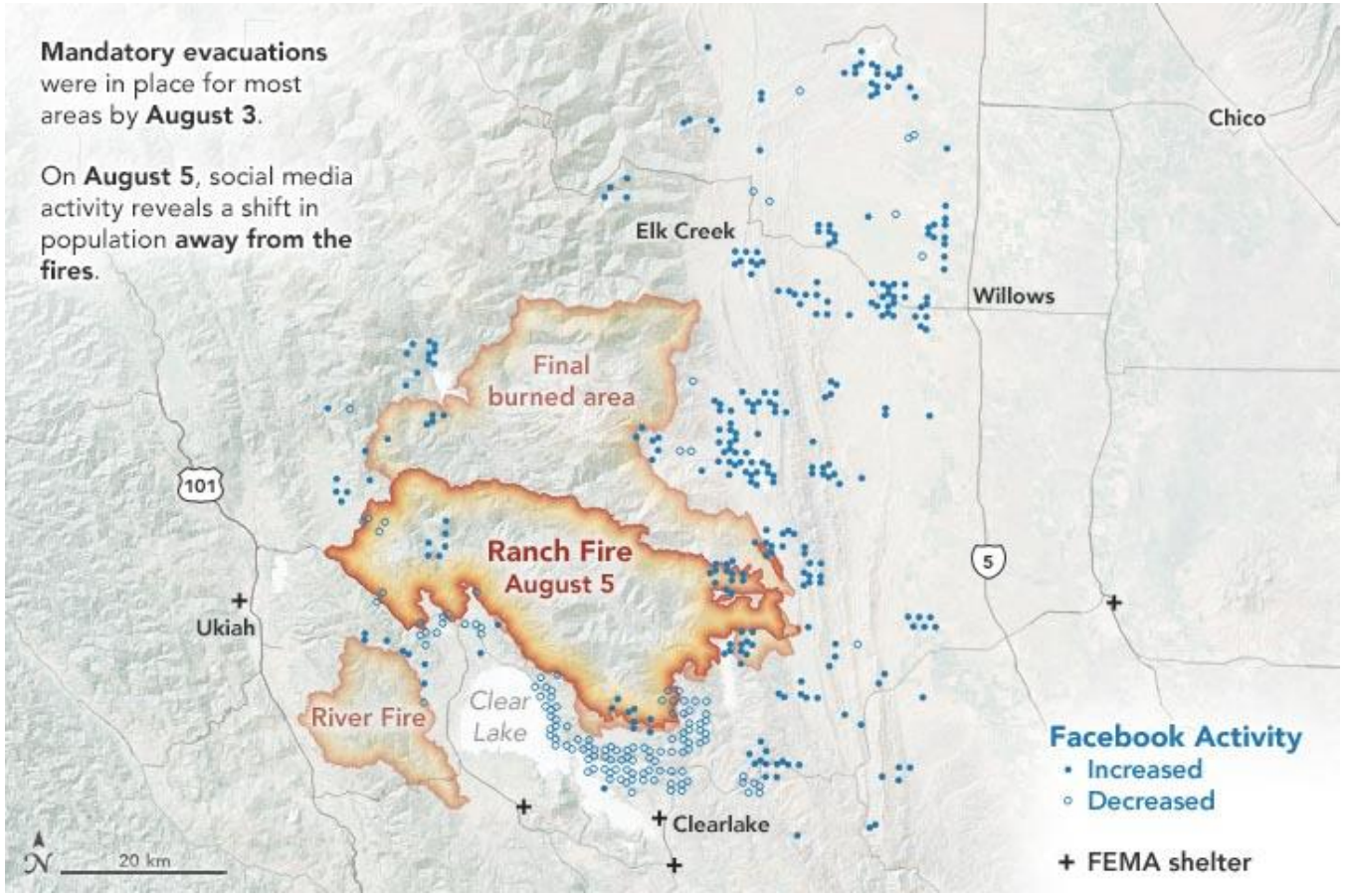
# Leveraging Social Media: Disaster response

- During Hurricane Harvey in 2017, emergency response teams used social media to identify areas needing urgent help.


https://www.gbtribune.com/news/trending-topics/help-victims-of-hurricane-harvey-by-using-social-media-and-texting/

# Leveraging Social Media: Disaster response

- During the California wildfires, social media data was used to track the spread of fires and coordinate evacuations.




https://wildfiretoday.com/2019/04/01/social-media-can-help-identify-evacuation-patterns/

# Leveraging Social Media: Public sentiment

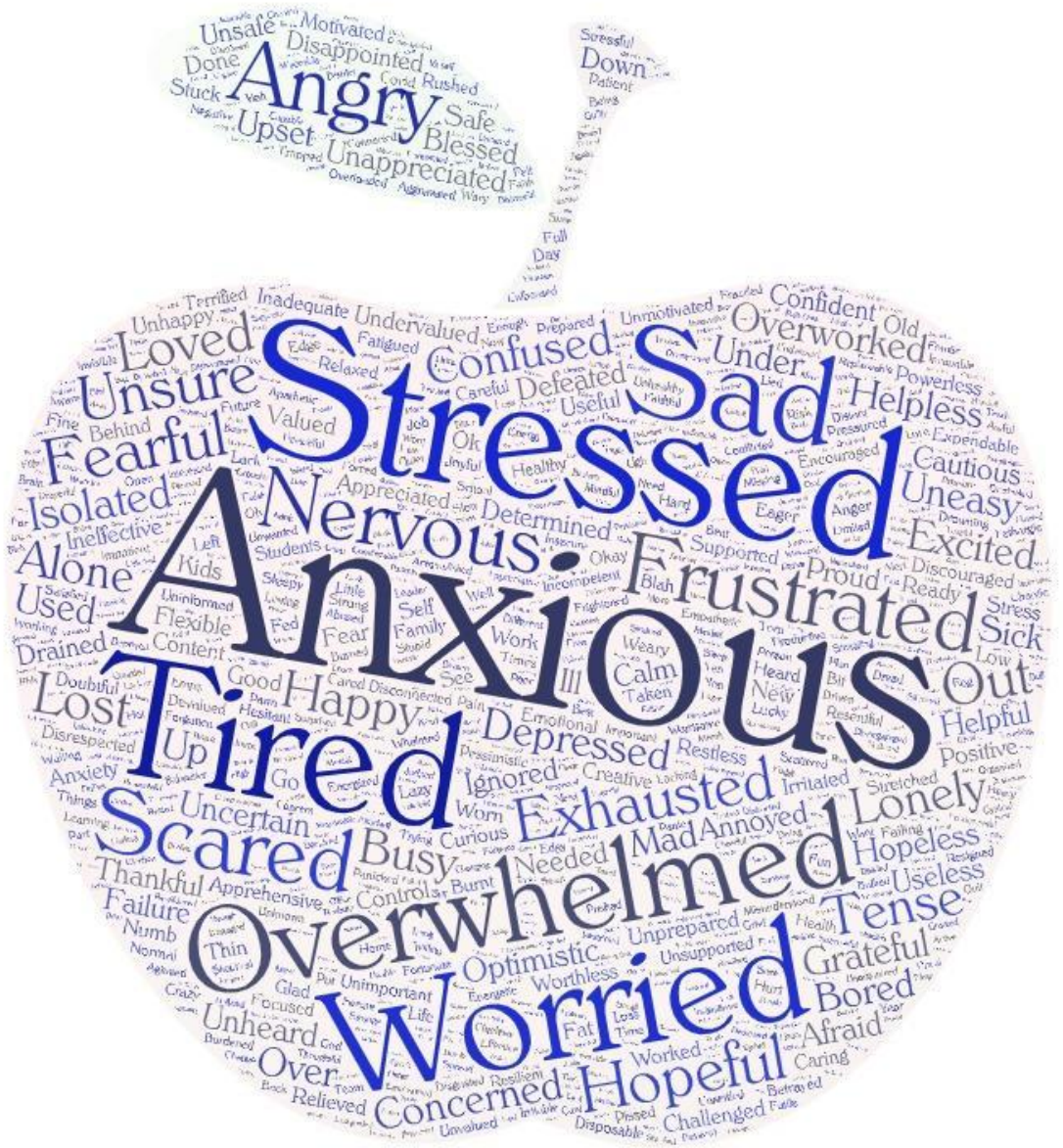
- The city of Seattle monitored social media to understand public sentiment regarding newly installed bike lanes.



https://www.seattlebikeblog.com/2018/03/23/anti-bike-lane-group-tweets-that-single-moms-dont-bike-it-backfires-spectacularly/

# Leveraging Social Media: Public sentiment

- During the COVID-19 pandemic, social media data was used to monitor public sentiment about lockdowns, masks, and vaccines. Platforms like Twitter, Facebook, and YouTube helped authorities understand misinformation trends and sentiment, tailoring public health messaging accordingly.

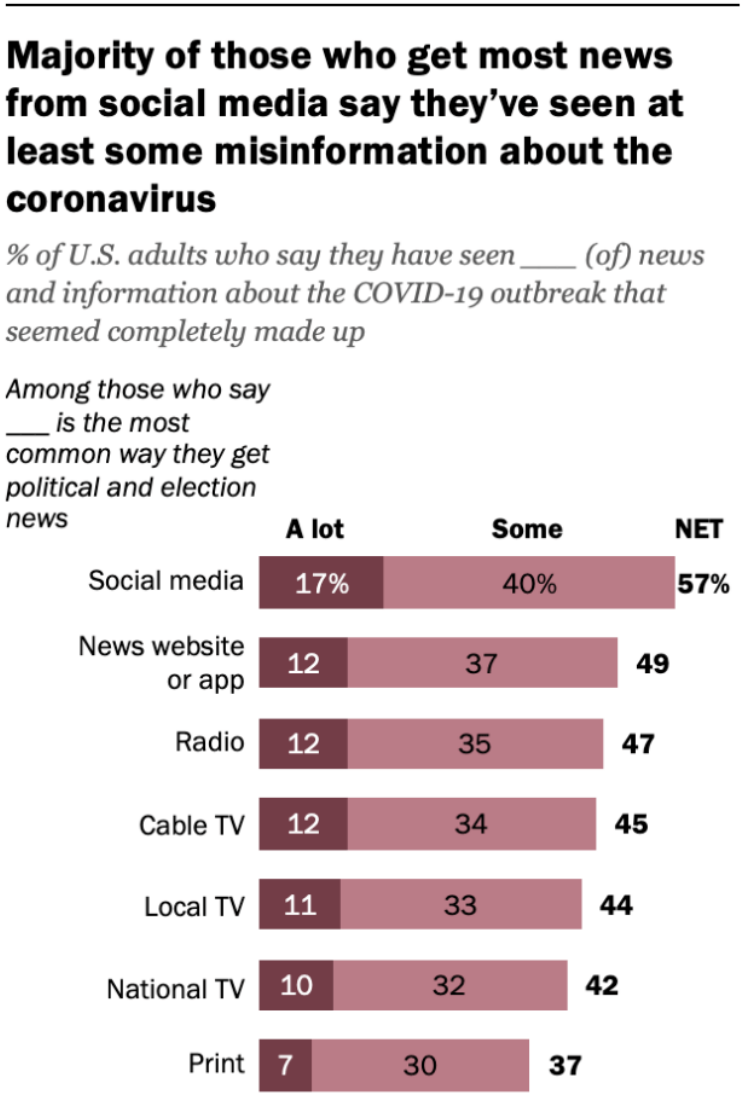



https://www.ccts.uky.edu/news/new-research-illuminates-school-staff-experiences-through-covid-19

# Challenges in Using Social Media Data

- **Data Quality and Bias**
  - Misinformation; Noises; Demographic biases

*"On social media platforms, users may create personal, unfiltered posts that are free from the responsibility of fact-checking. There are also numerous bots and trolls on these apps which exist solely to generate and spread misinformation surrounding health topics. Due to these two factors, the wave of misinformation on social media is unlike any that health departments may have encountered in the days before social media."* (Charlotte Ciampa, 2022)




https://www.naccho.org/blog/articles/social-media-during-a-public-health-crisis

https://www.pewresearch.org/journalism/2020/03/25/americans-who-primarily-get-news-through-social-media-are-least-likely-to-follow-covid-19-coverage-most-likely-to-report-seeing-made-up-news/

# Challenges in Using Social Media Data

- **Data Quality and Bias**
    - Misinformation; Noises; Demographic biases
- **Privacy and Ethics**
    - Importance of anonymizing data and handling ethical considerations

# Challenges in Using Social Media Data

- **Data Quality and Bias**
    - Misinformation; Noises; Demographic biases
- **Privacy and Ethics**
    - Importance of anonymizing data and handling ethical considerations
- **Technical Challenges**
    - Working with unstructured data and handling large datasets in real-time

# Social Media Analysis

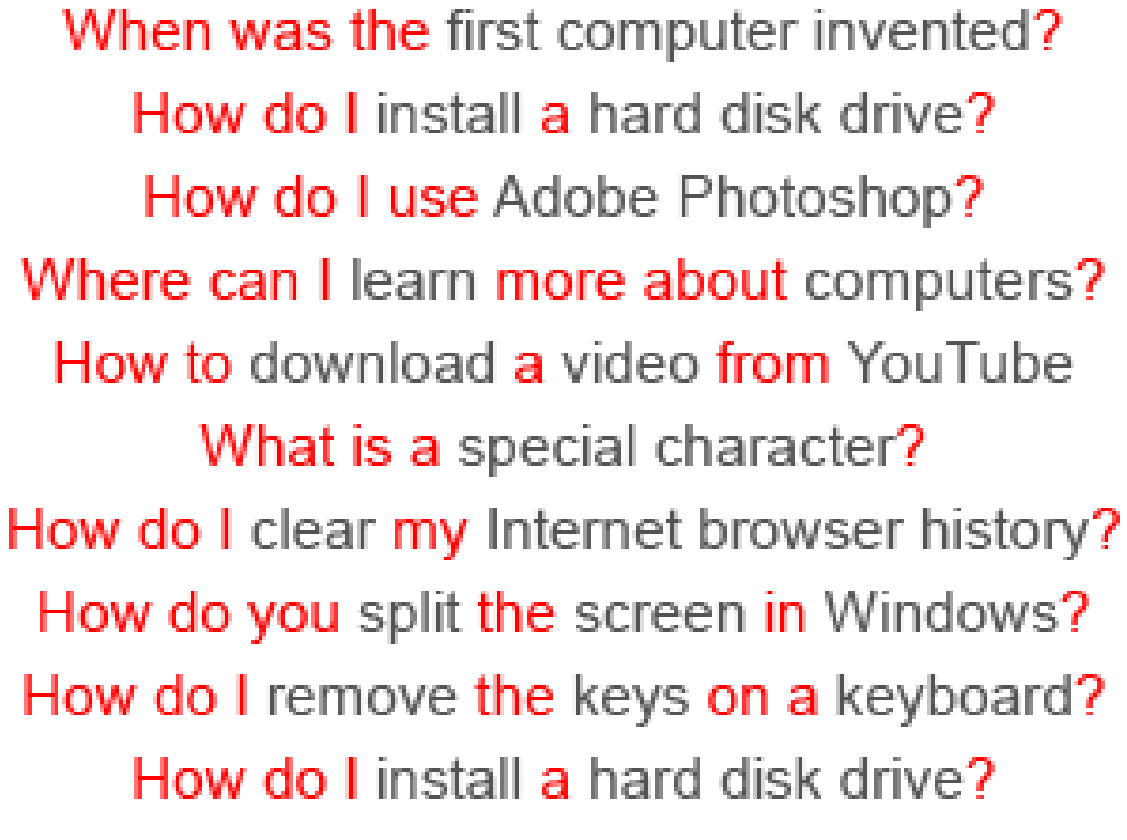
- **Data Collection**
  - API-based data extraction
  - Web scraping
- **Text Preprocessing/Normalization**
  - Tokenization; stopword removal; stemming and lemmatization
- **Sentiment Analysis**
  - Rule-based models; Deep learning models
- **Topic Modeling**
  - E.g., Latent Dirichlet Allocation (LDA)

# Tokenization

- Tokenization is a way of separating a piece of text into smaller units called tokens.
- It is a fundamental step in Natural Language Processing (NLP) since tokens are the building blocks of Natural Language.
- Three types of tokenization:
  - Word tokenization: Machine Learning → “Machine” / “Learning”
  - Character tokenization: Machine learning → “M” / “a” / “c” / “h” / “i” / “n” /…
  - Subwords tokenization: Machine Learning → “Machine” / “Learn” / “ing”

# Stop Words

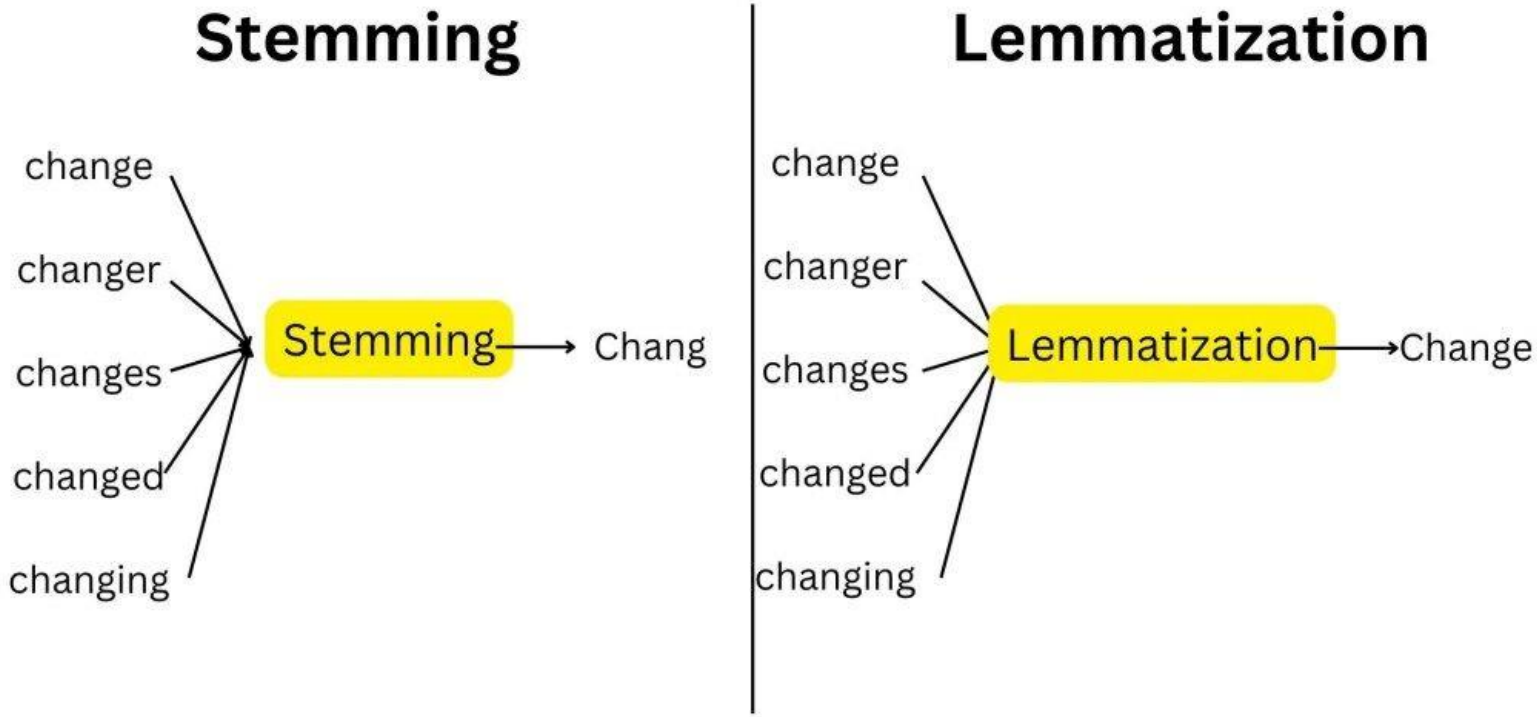
- Stop words are commonly used in Text Mining and NLP to eliminate words that are so widely used that they carry very little useful information.




https://www.computerhope.com/jargon/s/stopword.htm

# Stemming & Lemmatization

- Stemming is to remove prefixes and suffixes from a word to their root form.
- Lemmatization is to reduce a word to its base/root form, or ***lemma***, to make it easier to analyze.




https://seodepths.com/seo-research/nlp-seo-guide-use-cases/

# Sentiment Analysis: Rule-based model

- Rule-based models (such as dictionary-based models) use predefined lexicons or dictionaries of words that are assigned sentiment values (positive, negative, neutral).
- These models classify sentiment based on the words in the input text matching those in the dictionary.
- Limitations:
  - Lack of context understanding (e.g., “That’s sick”)
  - Over-simplification (e.g., “I liked the food” vs. “I kind of liked the food”)

# Sentiment Analysis: Deep learning model

- Deep learning models go beyond traditional rule-based methods by learning features and patterns directly from the data.
- ***Transformer*** architecture, for instance, use self-attention mechanisms to understand relationships between words in a sentence, which allows to capture complex linguistic structures, context, and semantic meaning.
- Like we did in the Computer Vision module, we will use a pretrained model to infer sentiment scores from the text data we get from Reddit.
- Let's try some state-of-the-art models (sign up required): https://huggingface.co/models?pipeline_tag=text-classification&library=transformers&language=en&sort=trending