ETH zürich

Project 2: Natural Language Processing for Global Health

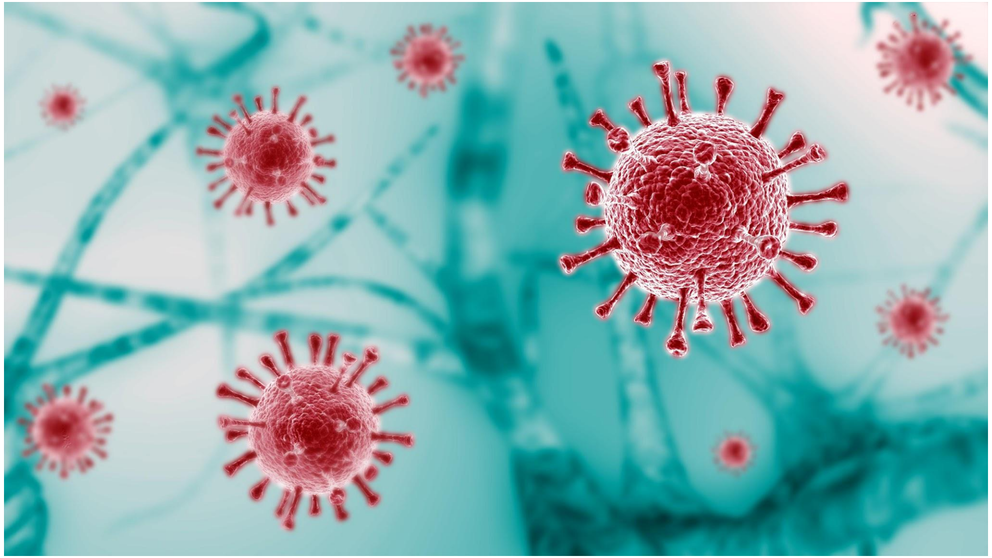
Alizée Pace, Alice Bizeul

04.04.2023

Sentiment Analysis at the rescue of Global Health

Sentiment analysis is a subfield of NLP that involves using ML techniques to automatically **identify and extract subjective information from text**, such as opinions, emotions, attitudes, and sentiments.

With the rise of social media platforms like Twitter, Facebook, and Instagram, **people are sharing their opinions and experiences online, in real-time more than ever before.**



Sentiment analysis therefore becomes a great tool for :

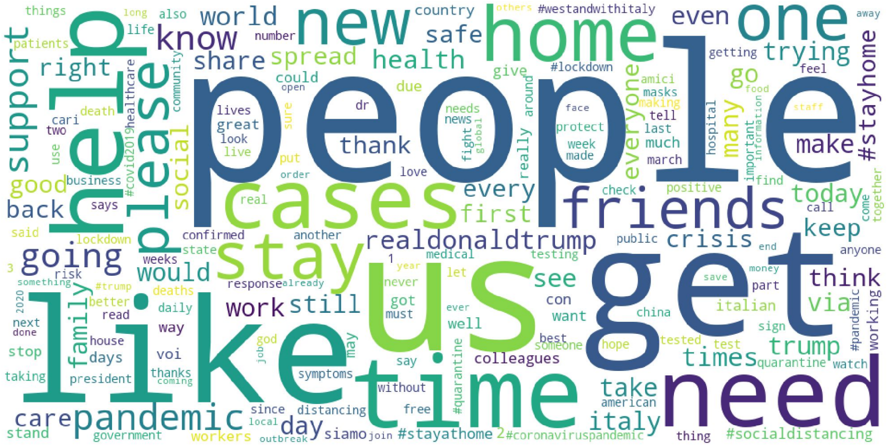
- Commercial purposes i.e., customer opinion mining, brand reputation tracking
- **Global opinion tracking i.e., political analysis, crisis management - e.g., Covid-19 pandemic**
- …

Project Goals

This project consists in **performing a sentiment analysis of Covid-related tweets** and deploying your model to **answer research questions related to the Covid-19 pandemic**:

1. Design a **pre-processing** pipeline for tweets
2. Explore diverse experimental settings for sentiment analysis:
   a. **Lexicon** based methods
   b. **Word embeddings** based methods
   c. **Transformer** based methods
3. **Evaluate** sentiment analysis performance and selecting the best approach
4. Define **relevant research questions and provide insights** using your chosen approach

Bonus: go beyond sentiment analysis and explore **emotion analysis or topic modelling**. Use your findings for research purposes

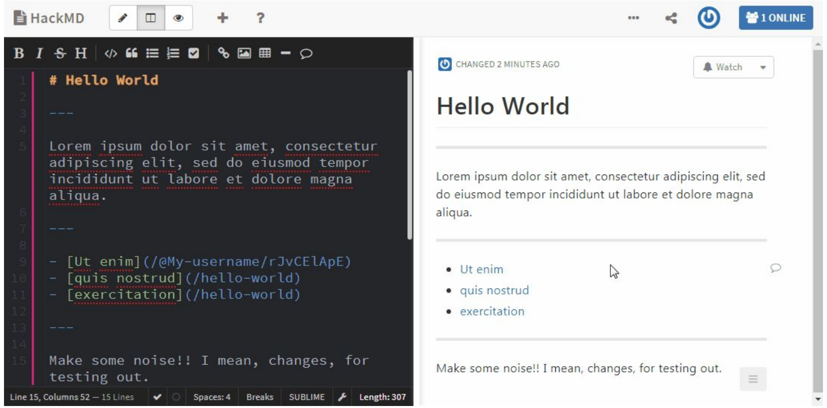


Word cloud extracted from covid-19 related tweets[0]

Wicke, Philipp, and Marianna M. Bolognesi. "Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter." *PloS one* 15.9 (2020): e0240010.

Project Guidelines - Part 1

- Details regarding project steps and datasets are available on the course’s Moodle page
- Deliverables are a markdown **(recommended)**, latex, word **or** notebook **report**:
  - **Follow the structure detailed on Moodle**
  - Questions will cover explanations of methods used, analysis & visualisations of results for each step, code snippets used to perform the tasks
  - **Please be concise** (getting full points does not require writing an essay …)
  - Code snippets should be code blocks/functions that allowed you to perform the sub-task (we want to have an overview of how you technically approach each task, code will not be ran)
  - **No rules regarding ChatGPT except a clear reference when used (same rule applies for all tools and references used to conduct this project) !**



```
# Hello World

---

Lorem ipsum dolor sit amet, consectetur
adipiscing elit, sed do eiusmod tempor
incididunt ut labore et dolore magna
aliqua.

---

- [Ut enim](/@My-username/rJvCElApE)
- [quis nostrud](/hello-world)
- [exercitation](/hello-world)

---

Make some noise!! I mean, changes, for
testing out.
```

Line 15, Columns 52 — 15 Lines | Spaces: 4 | Breaks | SUBLIME | Length: 307

CHANGED 2 MINUTES AGO

Hello World

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

- Ut enim
- quis nostrud
- exercitation

Make some noise!! I mean, changes, for testing out.

Hackmd.io

Project Guidelines - Part 2

- **Deadline for handing-in your report on Moodle:** May, 16th 2023
- The project is worth 70 points. The completion of each question grants you a fixed number of points. For additional bonus points, you are required to extend your work to a additional task (topic modelling or emotion analysis) and complete all associated requirements.
- **For most tasks there is no right or wrong answer. The goal of this project is for you to reason about the task at hand, gain experience with NLP and motivate each choice appropriately.**
- The team with the best report will be inquired to present their project for a non-cumulative bonus of 0.25 to the final grade.

Part 0: Understanding TweetsCOV19

TweetsCOV19 is a semantically annotated corpus of Tweets about the COVID-19 pandemic. It aims at capturing online discourse about various aspects of the pandemic and its societal impact.

A subset of the TweetsCOV19 dataset consisting of xxx tweets in total is made available on Moodle. It reflects the societal discourse about COVID-19 on Twitter in the period of October 2019 until May 2020.

To extract the dataset, a seed list of 268 COVID-19-related keywords was defined. Tweets in TweetsCOV19 contain at least one keyword from the set of seed terms, are written in English and published throughout the aforementioned time period.

Meta information entails a positive (1-5) and negative score (-5–1) for each tweet reflecting the polarity of the tweet, author location, timestamp among others.

Part 1: Designing a pre-processing pipeline

Working with real-world data requires some pre-processing or data cleaning as well as some data exploration:

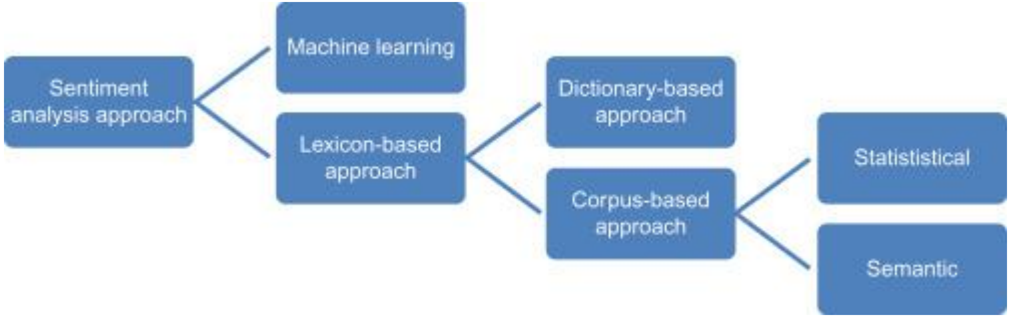
- **Data pre-processing:** familiarize yourself with the data, establish a tweet pre-processing pipeline (lemmatization, tokenization, emoji-mentions-url removal, …), define crucial steps, motivate and explain each one
- **Data exploration:** reason on the analysis and visualisations that would bring valuable insights to you for the remaining of this project. Visualise the distribution of most common uni- and bi-grams as well as the distribution of sentiment scores.
- **Metric choice:** select and motivate the metric you will use to evaluate your sentiment analysis
- **Train-test split:** perform a train-validation-test split of your data, motivate your approach.

Remark: all performance measurements should be done on your test set.

Part 2: Exploring lexicon based approaches

One of the most straightforward approach to sentiment analysis are lexicon matching and rule-based approaches:

- **Dictionary based approach (e.g., VADER):** a dictionary of lexicons (list of adjectives and adverbs with semantic orientation annotation) can be created manually as well as automatically generated. WorldNet or any other kind of online thesaurus can be used to discover the synonyms and antonyms to expand that dictionary. The final orientation of a piece of text is the result of the aggregation of all lexicons found in the text.



Part 3: Exploring word embeddings

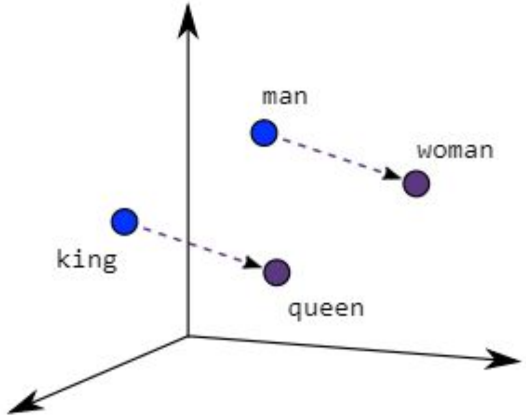
In NLP, **a word embedding is a representation of a word.** Typically, the representation is a real-valued vector that encodes the meaning of the word in such a way that words that are closer in the vector space are expected to be similar in meaning:

Rule-based approaches:

- Bag of Words (BOW)
- TF-IDF

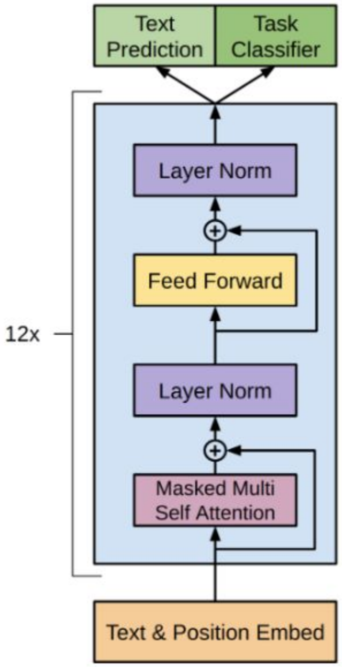
Unsupervised learning approaches:

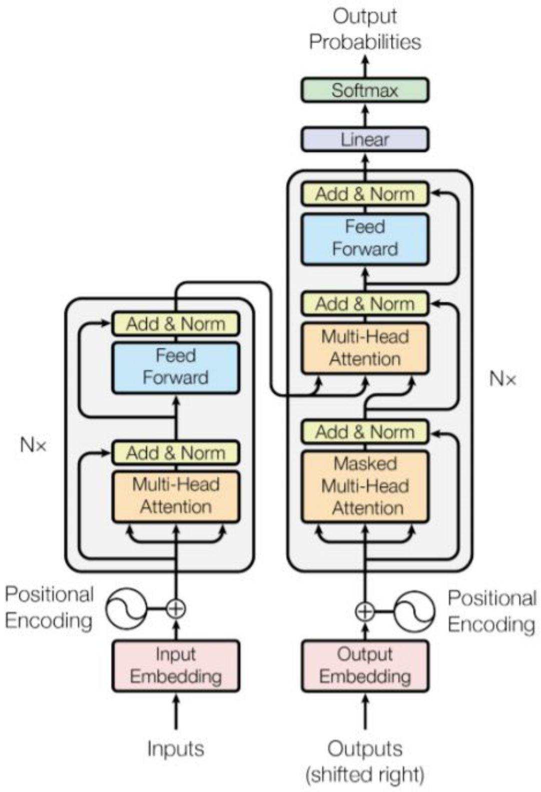
- Word2Vec
- GloVe
- FastTex



Using word embeddings as representations of each word, create and motivate a design rule for tweet representations, compare semantic knowledge of each approach (i.e., PCA, UMAP, t-SNE) and train a classifier for sentiment analysis

Part 4: Large language models & Transformer architectures





- **Main architecture details:**
  - Input embedding
  - Multi-head self-attention
  - Feed-forward network
  - Normalization and residual connections

- **Self-supervised** language models, with different pre-training objectives
  - BERT: Masked Language Modelling and Next Sentence Prediction
  - GPT?

- Then **fine-tune** the models to your task of interest. How can you best do this?

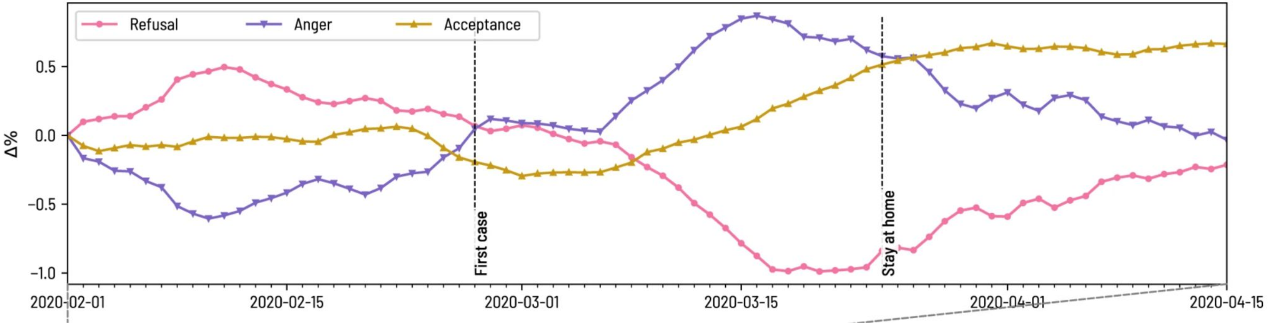
- How do the LLM language representations compare to embedding approaches investigated up to now?

Compute resources: we recommend **Google colab** with GPU.

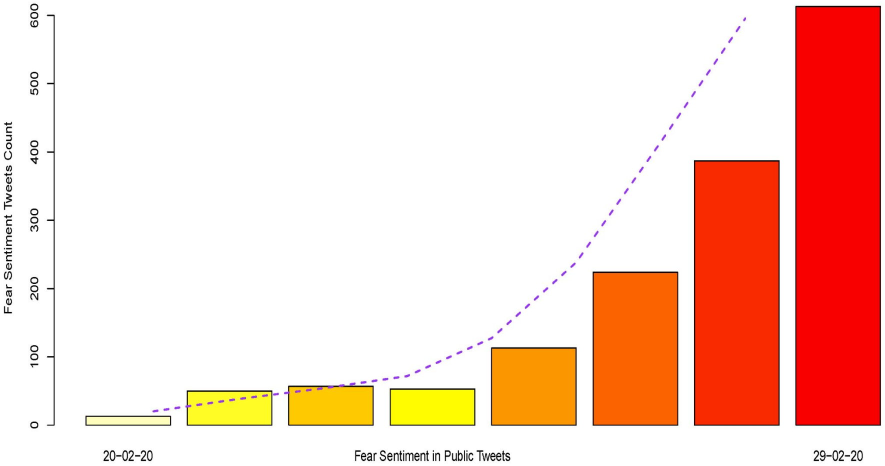
Part 5: Insights for Global Health

Examples from the literature

Did the sentiment of COVID-related tweets change over time during the pandemic?



COVID−19's Sentiment Curve: The Fear Pandemic

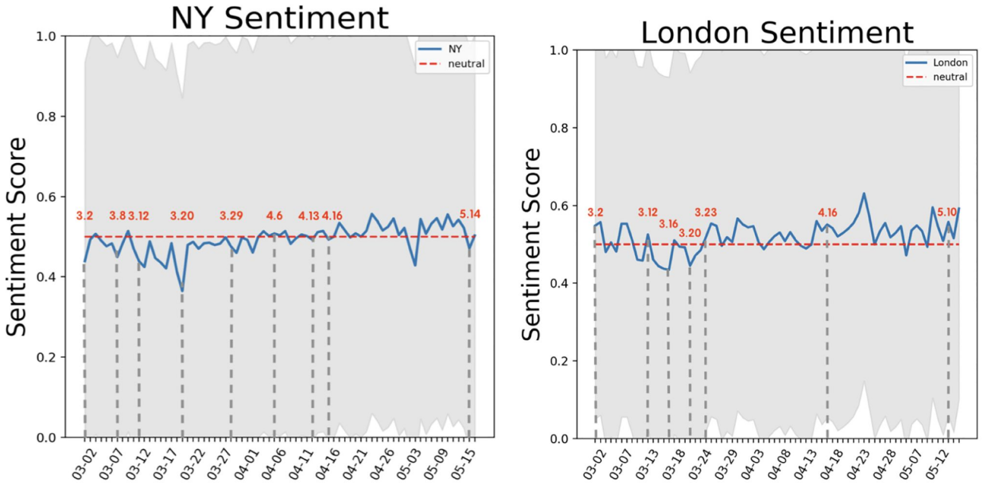


Sources: Aeillo et al., 2020. *Nature*.
Samuel et al., 2020. *Information*.

Part 5: Insights for Global Health

Examples from the literature

Did the sentiment of COVID-related tweets depend on the user country?

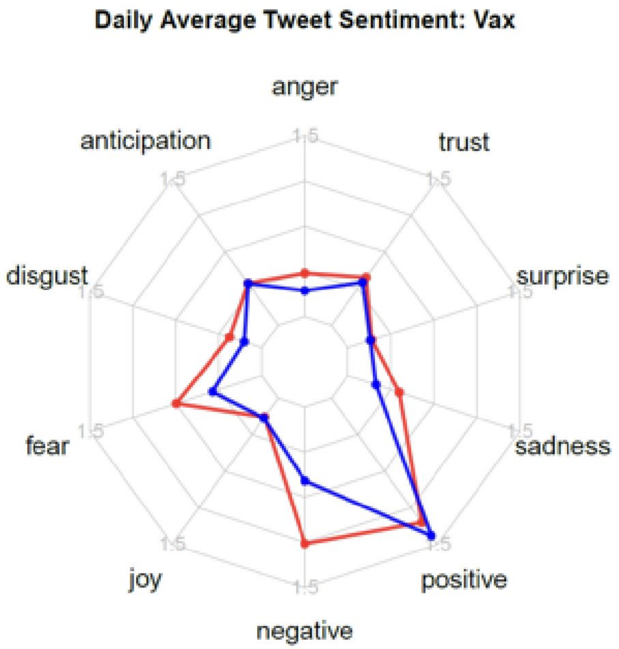


Source: Yao et al., 2021. *Cities*.

Part 5: Insights for Global Health

Examples from the literature

What was the main sentiment towards public health measures in tweets?



Source: Berts et al., 2021.
JMIR Public Health Surveill.

Part 5: Insights for Global Health

Analyse Tweet Sentiments as a function of **Time**, **Space** or **Topic** (choose one!).

For example:

- Did the sentiment of COVID-related tweets change over time during the pandemic?
- Did the sentiment of COVID-related tweets depend on the user country?
- What was the main sentiment towards public health measures in tweets?

- Choose and motivate your research question.
- Choose one of the approaches developed up to now to answer your question. Discuss pros/cons of your method.
- Analyse your results and compare to existing literature on the topic.

Bonus task: Emotion Analysis or Topic Modelling

Extend the binary sentiment analysis task to other common NLP tasks:

- a **more nuanced emotion analysis** (e.g. distinguish optimism/pessimism, thankfulness, anxiety, sadness, anger, surprise, jokes, etc).
- Alternative supervised task: we propose another related dataset with such labels for training. Deploy on our data and analyse.

or

- Identify the **most common topics** in the tweet corpus.
- We suggest using Latent Dirichlet Allocation, a common topic modelling approach (unsupervised).
- Assumes a collection of documents (tweets), each containing a mixture of topics. LDA identifies topics as sets of words which often co-occur within a document.
- Use third-party libraries!

What public health insights do you gain with this new task? How does this compare with any prior work in the literature?

Questions?

Also feel free to ask on the Moodle forum.