Natural Language Processing
01. Introduction

Juan José Alegría

March 11, 2025

Disclaimer

- These slides are **heavily** based on Felipe Bravo's work
- At the same time, a significant part of the content presented in his slides took inspiration from other resources such as textbooks and publications.
  - The neural network part of the course is based on Yoav Goldberg's book [Goldberg, 2017]
  - Non-neural network topics, such as probabilistic language models, are taken from Michael Collins' Columbia course [Collins, 2013].
  - Also from the draft of the third edition of Dan Jurafsky and James H. Martin's book [Jurafsky and Martin, 2023].
  - In addition, some slides have been adapted from online tutorials and other courses, such as Christopher Manning's Stanford course[1].

[1] ttp://web.stanford.edu/class/cs224n/

Table of contents

What is Natural Language Processing?

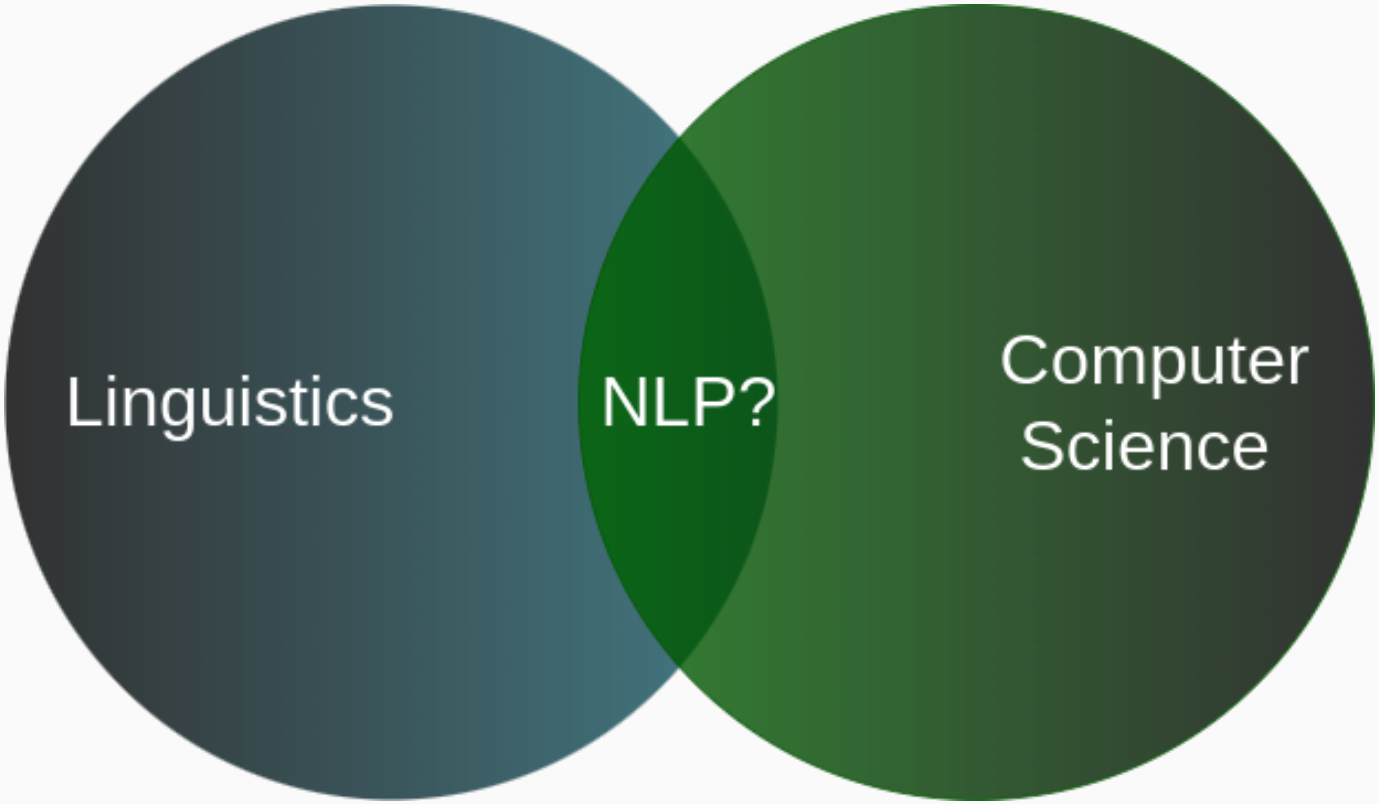
What is Natural Language Processing?

"Natural language processing (NLP) is the field of designing methods and algorithms that take as input or produce as output unstructured, natural language data."
[Goldberg, 2017]

Why do we need/want Natural Language Processing?

- The amount of digitized textual data generated daily is huge (e.g., the Web, social media, medical records, digitized books) $\rightarrow$ there is information there that we may want to retrieve!
- We would like to do many things that involve text $\rightarrow$ translation, summarization, question answering, etc.
- Tasks 👀

Natural Language Processing



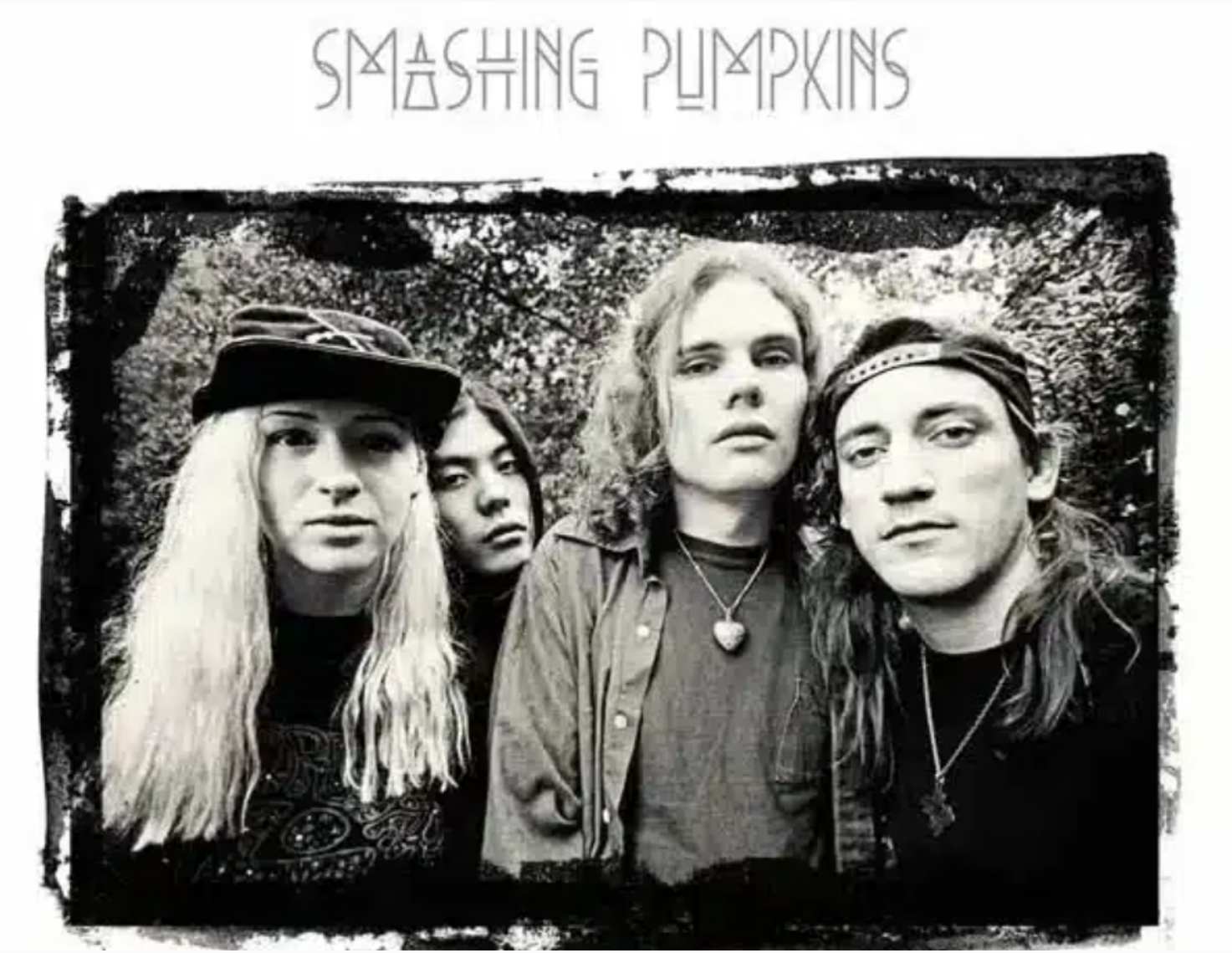
`Important:` There are two related fields at the intersection of linguistics and computer science: Natural Language Processing and Computational Linguistics

Natural Language Processing and Computational Linguistics

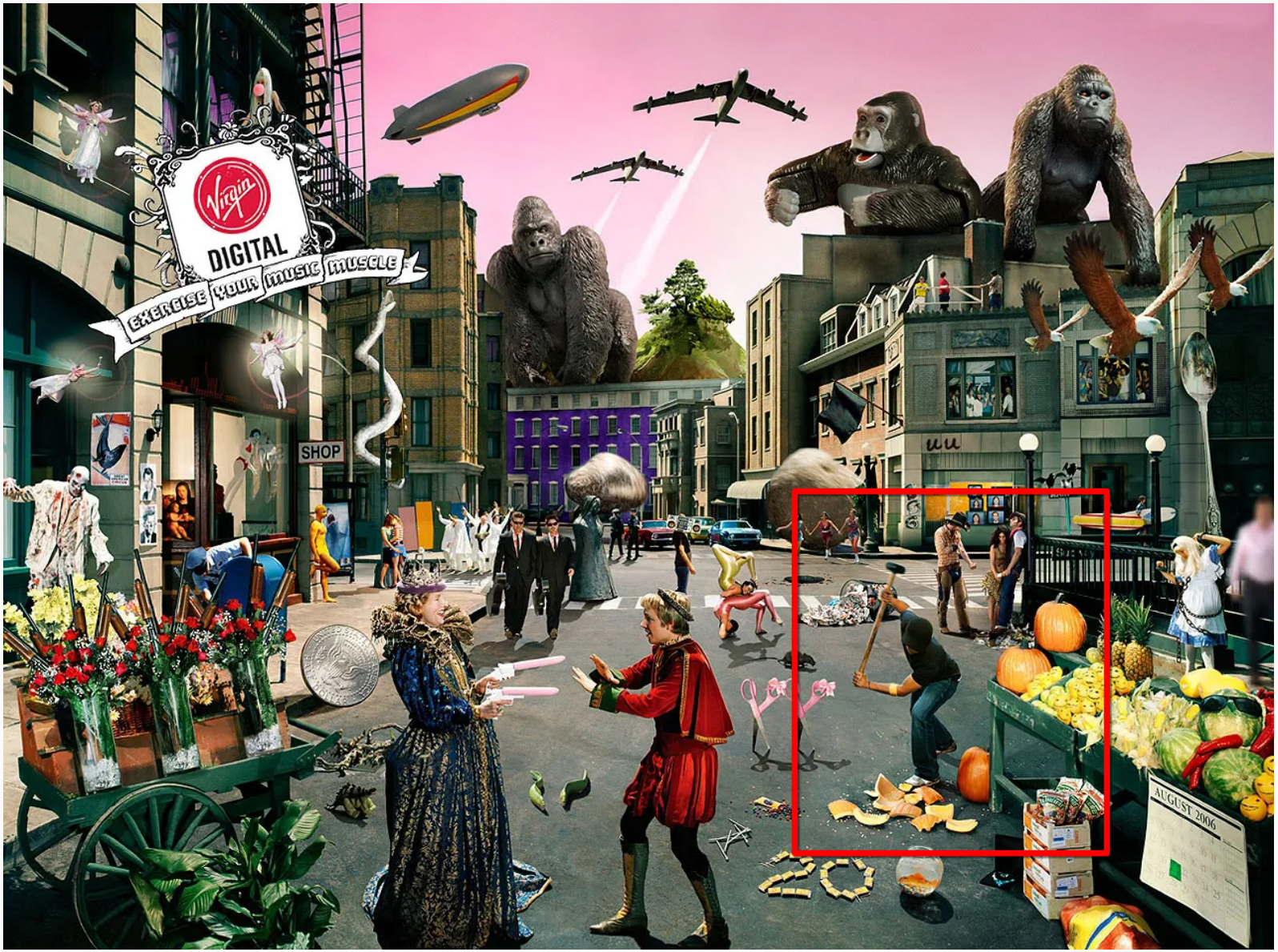
- Natural language processing (NLP) develops methods for solving practical problems involving language [Johnson, 2014].
  - Automatic speech recognition.
  - Machine translation.
  - Information retrieval from documents.
- Computational linguistics (CL) studies the computational processes underlying (human) language.
  - How do we understand language?
  - How do we produce language?
  - How do we learn language?
- Similar methods and models are used in NLP and CL, but with a different focus. In CL, language is the object of study.

Language

Language can be difficult



Language can be difficult



5 Challenging Properties of Human Language

1. **Ambiguity**: Human language is highly ambiguous.
    - For example: *I ate pizza with friends* vs. *I ate pizza with olives* vs. *I ate pizza with a fork*.
    - All these sentences have similar grammatical properties but differ radically in what the prepositional phrase modifies: (1) the pronoun “I”, (2) the noun “pizza”, (3) the verb “eat”.
2. **Dynamism**: Language is ever changing and evolving (e.g, Hashtags in Twitter).

5 Challenging Properties of Human Language

3. **Discreteness**: we cannot infer the relation between two words from the letters they are made of (e.g., hamburger and pizza).
4. **Compositionality**: the meaning of a sentence goes beyond the individual meaning of their words.
5. **Sparseness**: The way in which words (discrete symbols) can be combined to form meanings is practically infinite.

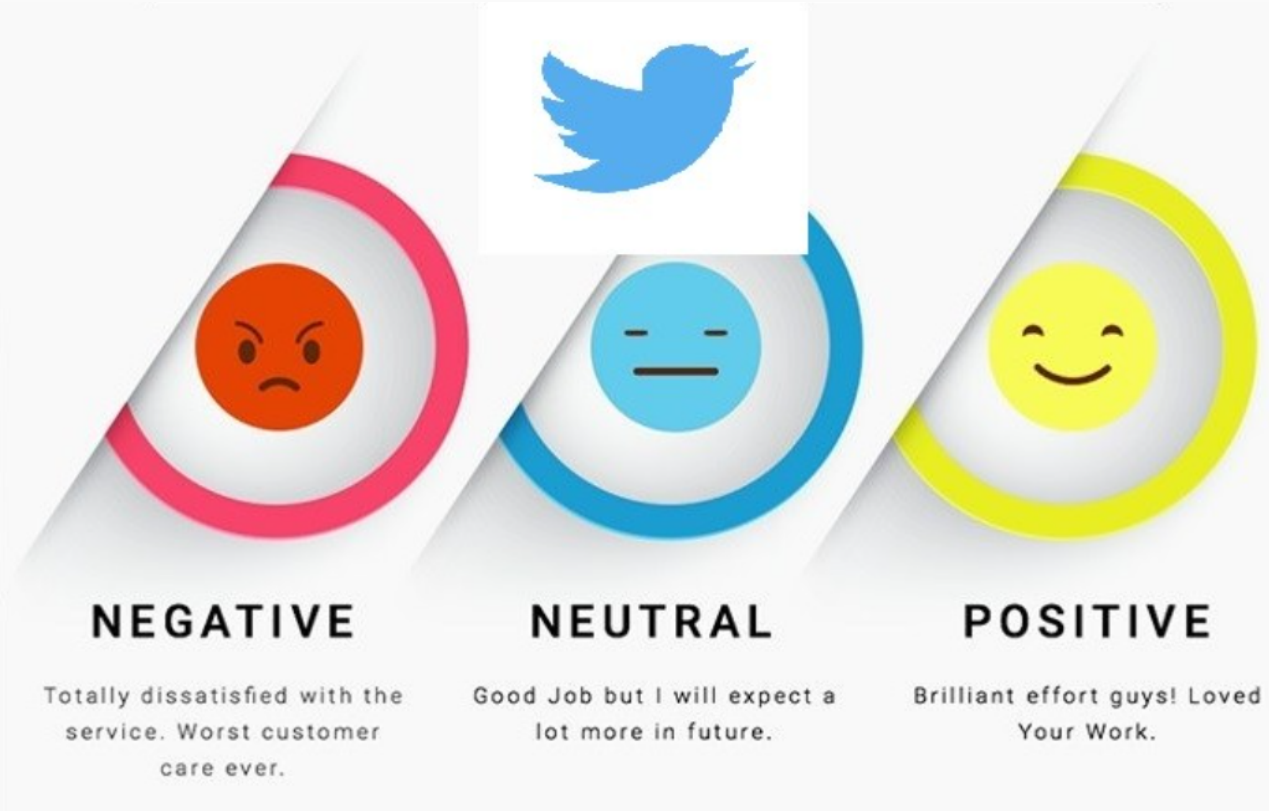
NLP and Machine Learning

Natural Language Processing and Machine Learning

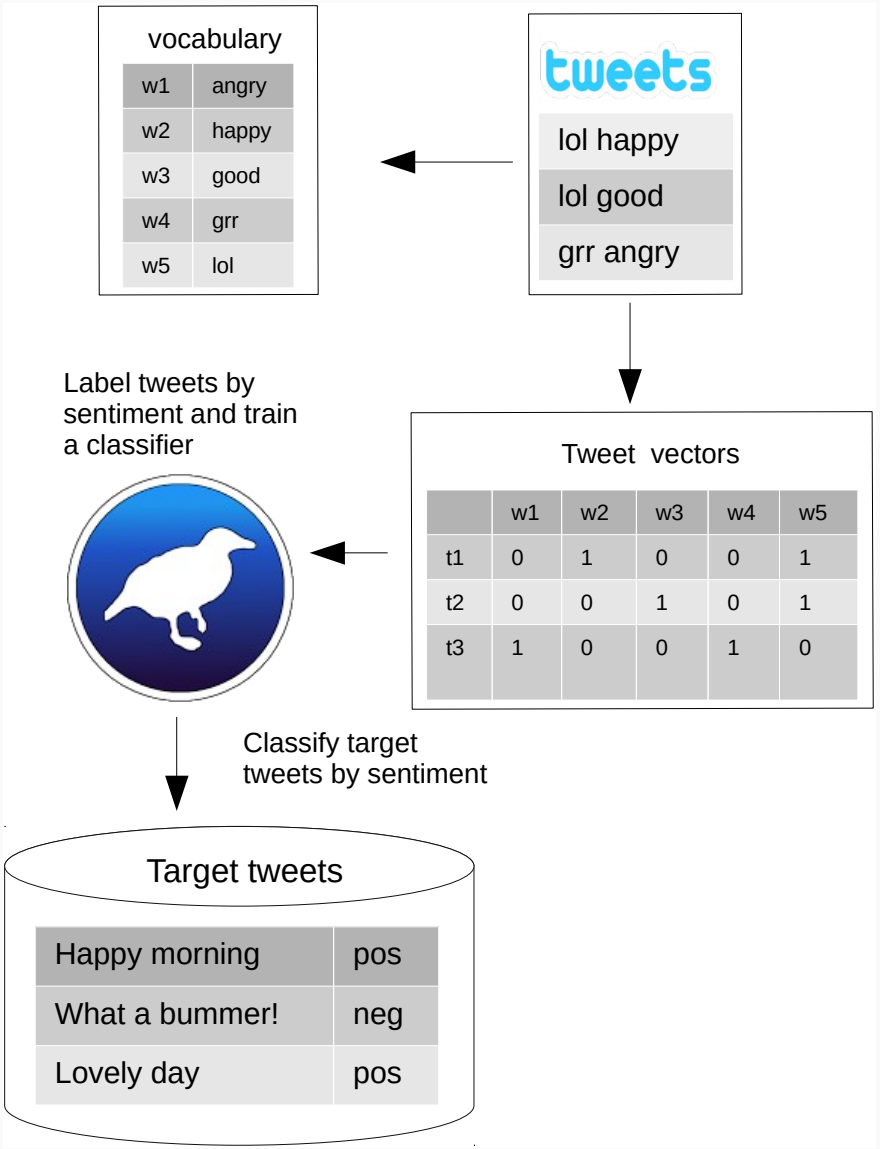
- Although we humans are great users of language, we are also very poor at formally understanding and describing the rules that govern language.
- Understanding and producing language using computers is highly challenging.
- The best known set of methods for dealing with language data is based on supervised machine learning.
- Supervised machine learning: attempt to infer usage patterns and regularities from a set of pre-annotated input and output pairs (a.k.a training dataset).

Example: Sentiment classification

- Objective: Classify tweets as positive or negative
- We already have a **corpus** of labeled tweets (positive/negative)
- How can we train a classifier?

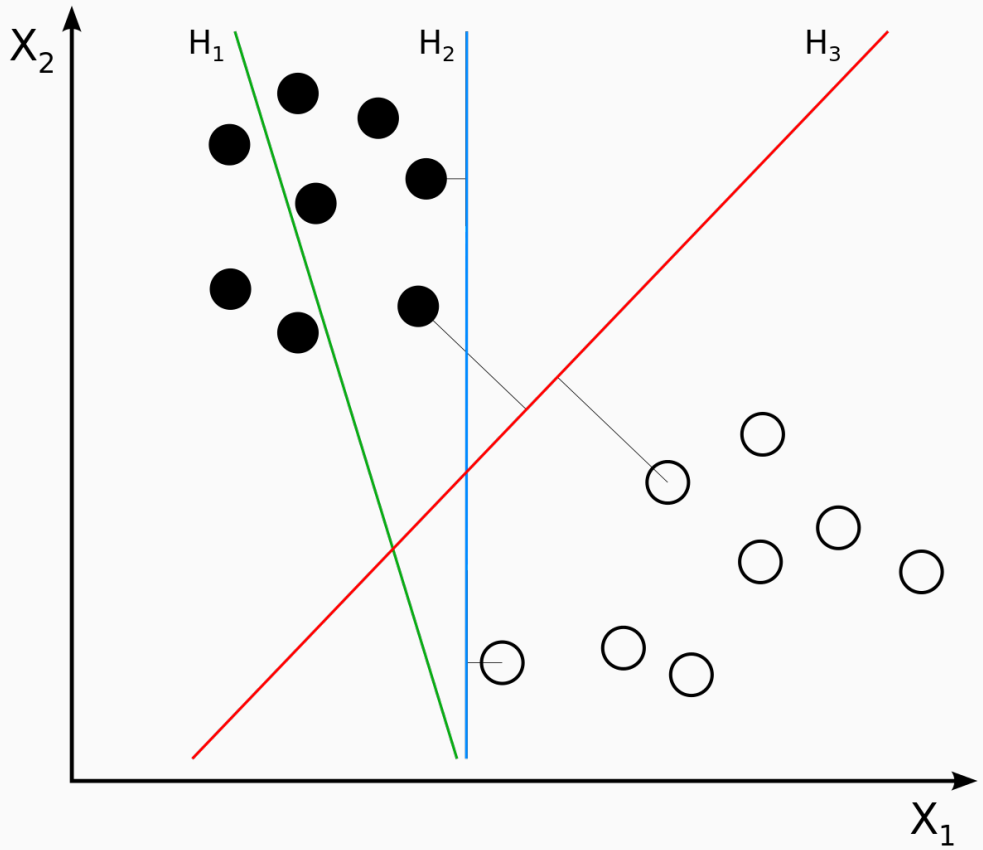


Input: Bag of Words



Supervised Learning: Support Vector Machines (SVMs)

- Idea: Find a hyperplane that separates the classes with the maximum margin (largest separation).

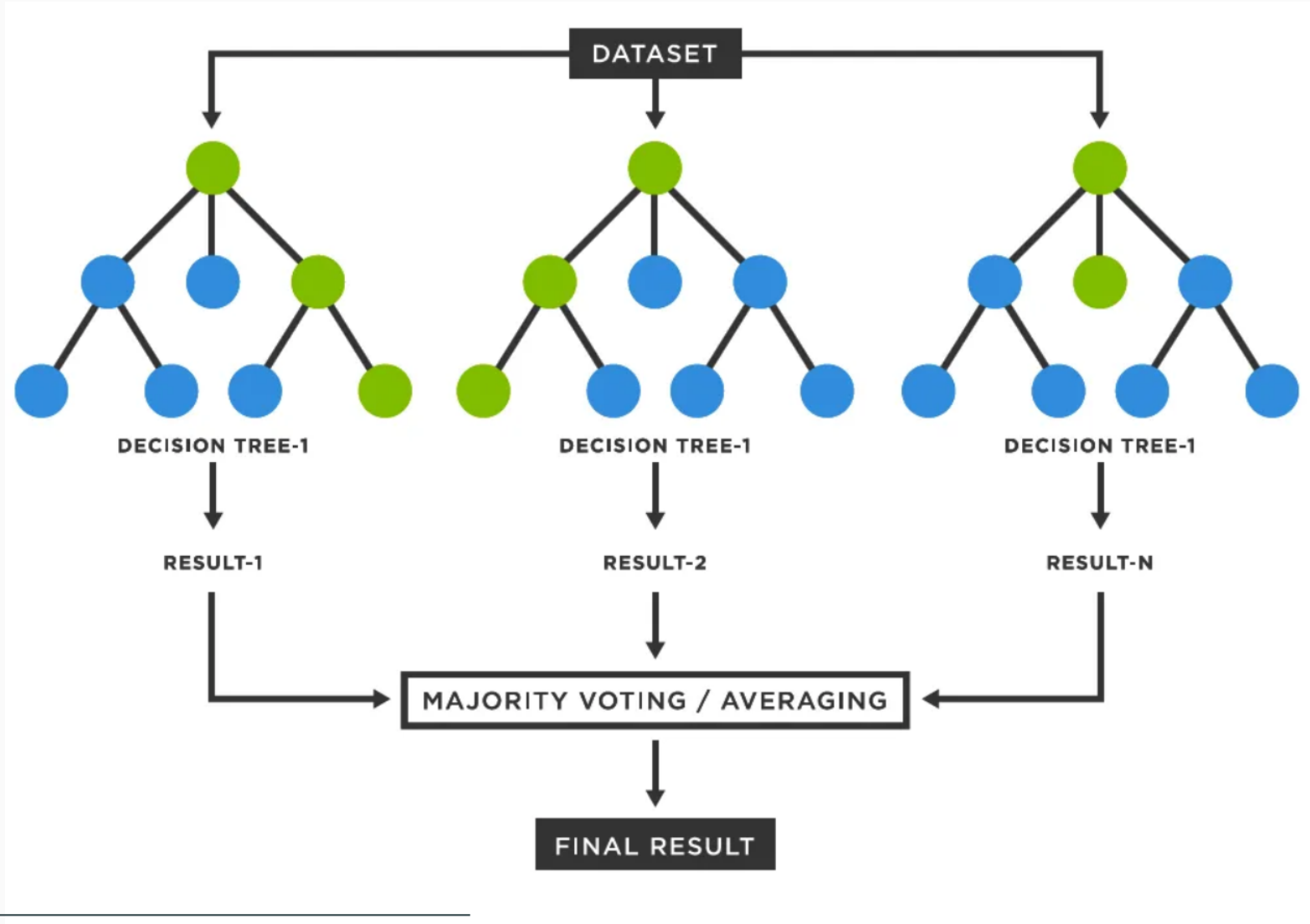


- H_3 separates the classes with the maximum margin.

[2] Image source: *Wikipedia* - Support vector machine

Supervised Learning: Random Forest

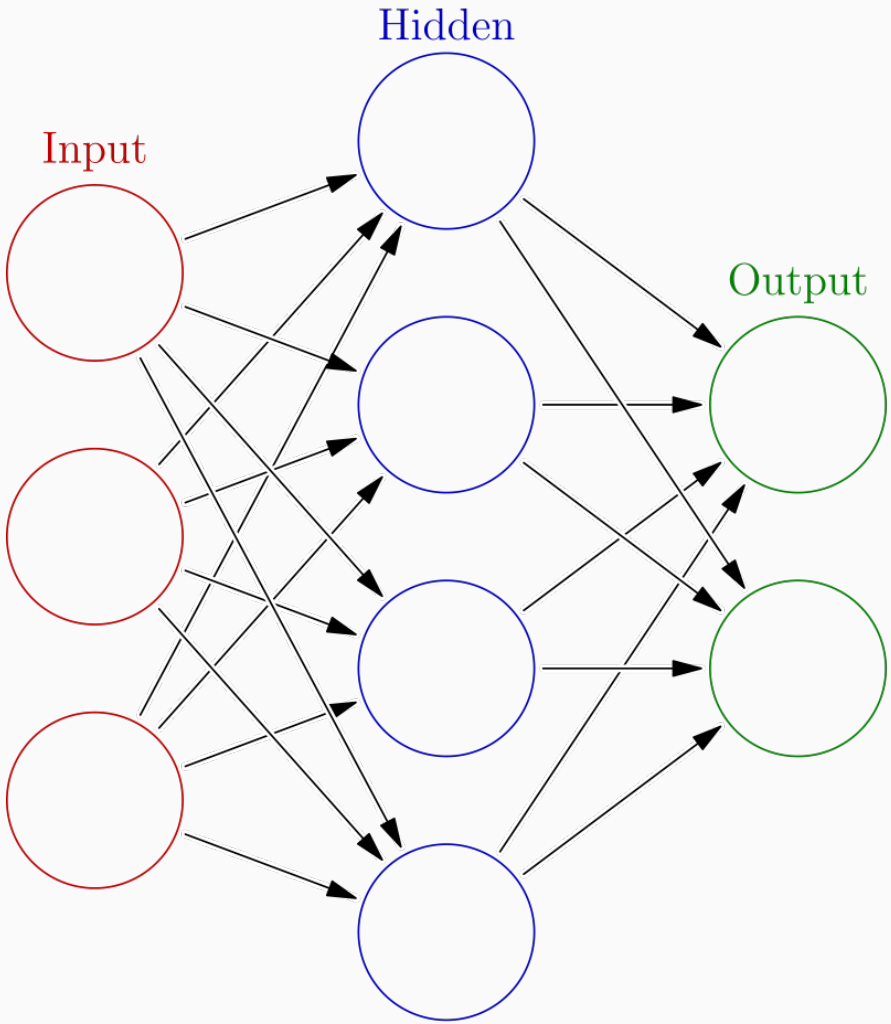
- Idea: Train several (almost) independent trees and then perform majority voting



[3] Image source: *Medium* - Random Forests

Supervised Learning: (shallow) Neural Network

- Idea: Train a neural network (a.k.a. multilayer perception), using the bag of words as input



[4] Image source: *Wikipedia* - Neural network (machine learning)

Supervised Learning: Can we do better?

- We can use several ML models, some that we don't even mentioned (KNN, naive bayes, etc)
- But can we improve the input?
- Maybe we can new features, not just counting words.
- This are called *hand-crafted* features
- Ideas?

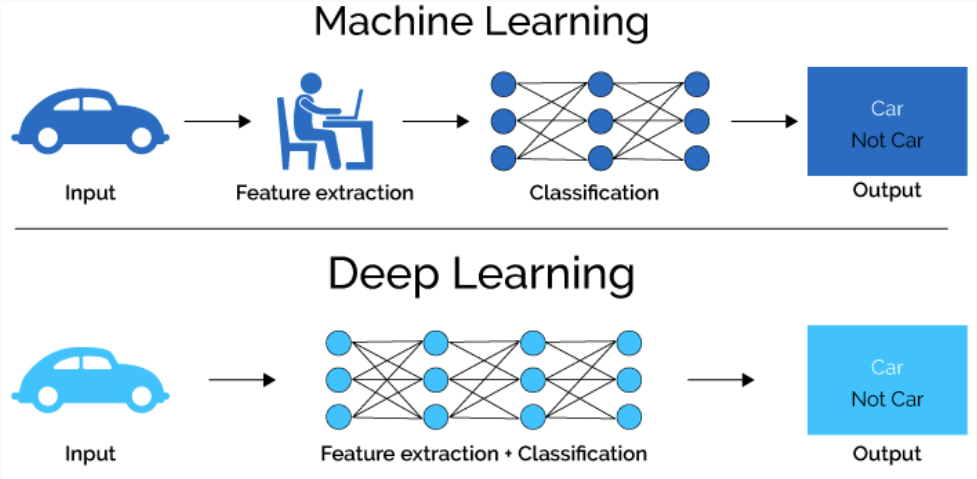
Hand-crafted features

- Some hand-crafted features devised by the team that achieved the best performance [Mohammad et al., 2013] in the “Sentiment Analysis in Twitter task” in the Semantic Evaluation Workshop (2013) [Nakov et al., 2013]:
  1. Word *n*-grams.
  2. Character *n*-grams.
  3. Part-of-speech tags.
  4. The number of elongated words (words with one character repeated more than two times).
  5. The number of words with all characters in uppercase.
  6. The presence of positive or negative emoticons.
  7. The number of individual negations.
  8. The number of contiguous sequences of dots, question marks and exclamation marks.
  9. Features derived from polarity lexicons [Mohammad et al., 2013].
  10. Etc.

Hand-crafted features

- Designing hand-crafted features is hard. Feature engineering requires domain knowledge and is usually a bottleneck in the process.
- It would be nice if the models themselves were able to find the best way to represent the input 👀 👀 👀

Deep Learning



- In Deep Learning, we give the input to the model with “little” preprocessing: the model itself learns how to best represent the input.
- Deep Learning yields state-of-the-art results in most NLP tasks.
- Large amounts of training data and faster multicore GPU machines are key in the success of deep learning.
- **Neural networks** and **word embeddings** play a key role in modern NLP models.

NLP Tasks

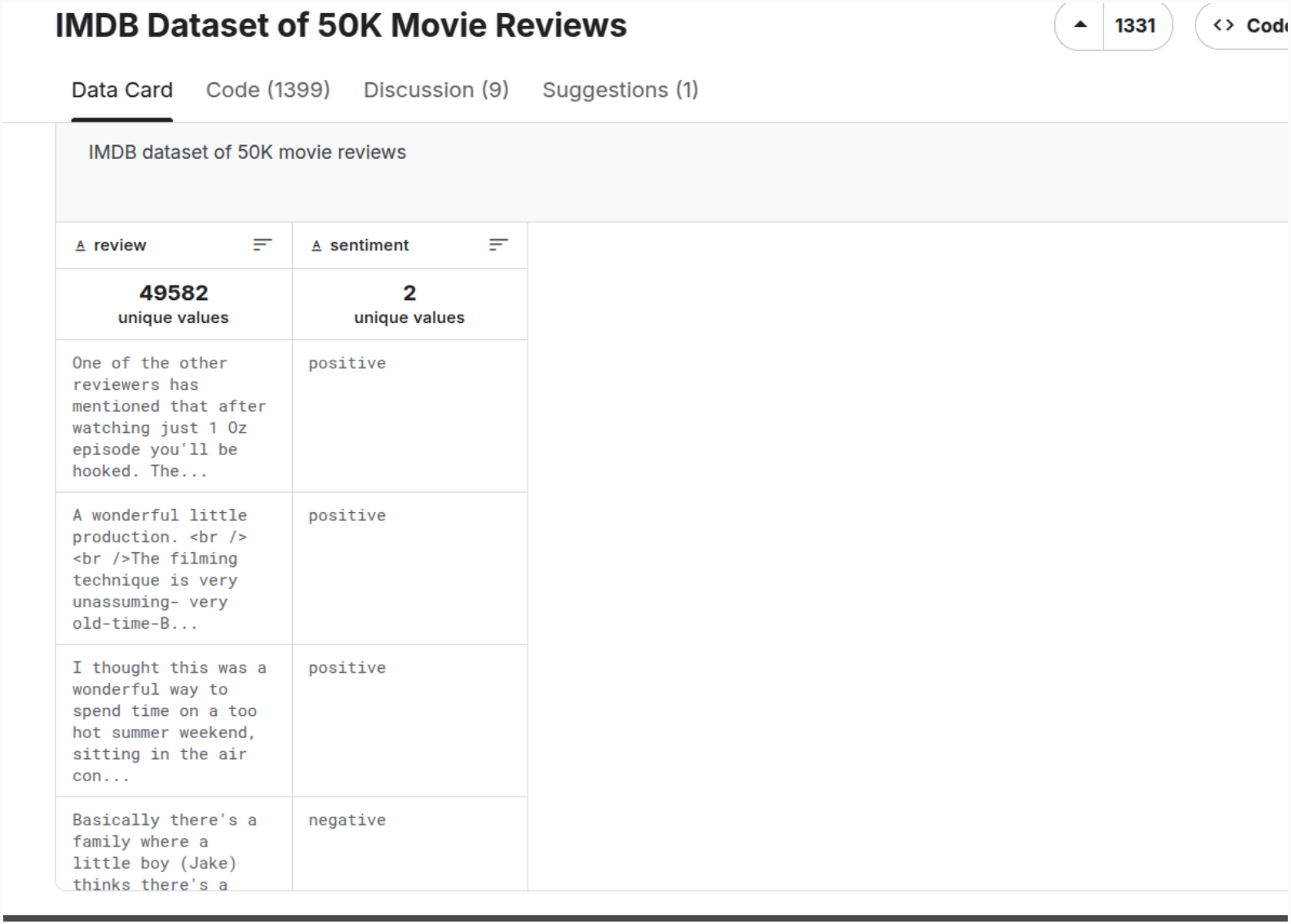
Tasks

- There are many possible problems that we would like to tackle using NLP methods; these are usually called ***tasks***.
- Some examples:
  - **Text classification**: given a document, classify it into one of several classes. E.g., classify a news article into *SPORTS*, *GOSSIP*, *POLITICS* o *ECONOMY*.
  - **Sentiment Analysis**: classify a document according to its polarity. It's a type of text classification but with some particular nuances (sarcasm, irony, negations, etc)
  - **Named Entity Recognition (NER)**: given a text, recognize and classify entities (people, places, organizations, etc).
  - **Parts-Of-Speech (POS) Tagging**: classify each word in the text into nouns, verbs, adjectives, etc.
  - **Machine translation**: translate a sentence from a source language to a target language.
  - **Question Answering**: given a paragraph p and a question q, find in the p the answer to q.
  - Etc.

Tasks and data labeling

- We are in the supervised learning framework, so to train a model for any task, we need labeled data.
- Clearly, the labeling for each task is different.

Data labeling for text classification



IMDB Dataset of 50K Movie Reviews

1331 Code

Data Card | Code (1399) | Discussion (9) | Suggestions (1)

IMDB dataset of 50K movie reviews

review	sentiment
49582 unique values	2 unique values
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive
A wonderful little production. The filming technique is very unassuming- very old-time-B...	positive
I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air con...	positive
Basically there's a family where a little boy (Jake) thinks there's a	negative

[5] Source: *Kaggle* - IMDB Dataset of 50K Movie Reviews

Data labeling for POS tagging and NER

Name_Entity_Recognition_data

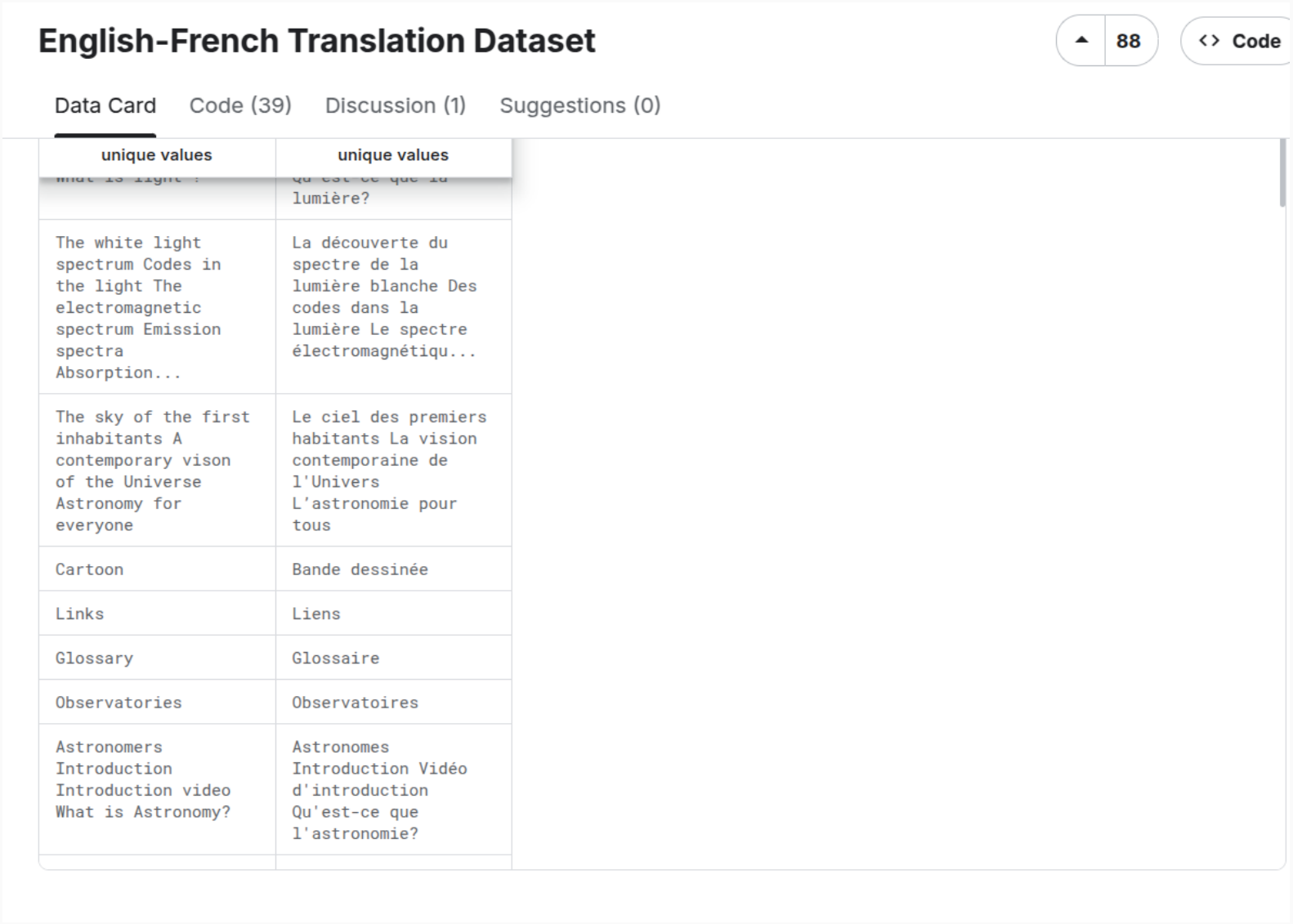
57 Code

Data Card | Code (0) | Discussion (0) | Suggestions (0)

Sentence #	Sentence	POS	Tag
47959	47575	47214	33318
47959 unique values	**47575** unique values	**47214** unique values	**33318** unique values
Sentence: 1	Thousands of demonstrators have marched through London to protest the war in Iraq and demand the wit...	['NNS', 'IN', 'NNS', 'VBP', 'VBN', 'IN', 'NNP', 'TO', 'VB', 'DT', 'NN', 'IN', 'NNP', 'CC', 'VB', 'DT...	['O', 'O', 'O', 'O', 'O', 'O', 'B-geo', 'O', 'O', 'O', 'O', 'B-geo', 'O', 'O', 'O', 'O', '...
Sentence: 2	Families of soldiers killed in the conflict joined the protesters who carried banners with such slog...	['NNS', 'IN', 'NNS', 'VBN', 'IN', 'DT', 'NN', 'VBD', 'DT', 'NNS', 'WP', 'VBD', 'NNS', 'IN', 'JJ', 'N...	['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-per', ...
Sentence: 3	They marched from the Houses of Parliament to a rally in Hyde Park .	['PRP', 'VBD', 'IN', 'DT', 'NNS', 'IN', 'NN', 'TO', 'DT', 'NN', 'IN', 'NNP', 'NNP', '.']	['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-geo', 'I-geo', 'O']
Sentence: 4	Police put the number of marchers at 10,000 while	['NNS', 'VBD', 'DT', 'NN', 'IN', 'NNS', 'IN', 'CD', 'IN',	['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O',

[6] Source: *Kaggle* - Name_Entity_Recognition_data

Data labeling for machine translation



English-French Translation Dataset

88 Code

Data Card | Code (39) | Discussion (1) | Suggestions (0)

unique values	unique values
	lumière?
The white light spectrum Codes in the light The electromagnetic spectrum Emission spectra Absorption...	La découverte du spectre de la lumière blanche Des codes dans la lumière Le spectre électromagnétiqu...
The sky of the first inhabitants A contemporary vison of the Universe Astronomy for everyone	Le ciel des premiers habitants La vision contemporaine de l'Univers L'astronomie pour tous
Cartoon	Bande dessinée
Links	Liens
Glossary	Glossaire
Observatories	Observatoires
Astronomers Introduction Introduction video What is Astronomy?	Astronomes Introduction Vidéo d'introduction Qu'est-ce que l'astronomie?

[7] Source: *Kaggle* - English-French Translation Dataset

Brief history of NLP

NLP History

NLP progress can be divided into three main waves: 1) rationalism, 2) empiricism, and 3) deep learning [Deng and Liu, 2018].

- **(1950 - 1990) Rationalism:** approaches endeavored to design hand-crafted rules to incorporate knowledge and reasoning mechanisms into intelligent NLP systems (e.g, ELIZA for simulating a Rogerian psychotherapist, MARGIE for structuring real-world information into concept ontologies).
- **(1991 - 2009) Empiricism:** characterized by the exploitation of data corpora and of (shallow) machine learning and statistical models (e.g., Naive Bayes, HMMs, IBM translation models).
- **(2010 -) Deep Learning:** feature engineering (considered as a bottleneck) is replaced with representation learning and/or deep neural networks (e.g., DeepL Translator). A very influential paper in this revolution: [Collobert et al., 2011].

A fourth wave?

- Large Language Models (LLMs) like GPT, DeepSeek R1, Llama, Gemino, etc., are deep neural networks trained on large corpora (hundreds of billions of tokens) and a large parameter space (billions) to predict the next word from a fixed-size context.
- One of the most striking features of these models is their ability for few-shot, one-shot, and zero-shot learning, often referred to as “in-context learning”.
- This implies their capacity to acquire new tasks with minimal human-annotated data, simply by providing the appropriate instruction or prompt.
- Thus, despite being rooted in the deep learning paradigm, they introduce a disruptive approach to NLP.

Roadmap and evaluations

Roadmap

Unit 1: Foundations of NLP

1. Introduction to NLP
2. Vector Semantics
3. Fundamental questions about language
4. Probabilistic language models
5. Linear models

Roadmap

Unit 2: Neural networks

5. Neural Networks
6. Word Vectors
7. Recurrent Neural Networks
8. Sequence-to-sequence + Attention
9. Transformers + BERT

Roadmap

Unit 3: Large Language Models and applications

10. GPT + LLMs and emergent abilities
11. Retrieval Augmented Generation
12. Interpretability
13. Agents
14. Ethics

Evaluations

- NC: 2 tests (“controles”): after Unit 1 and Unit 2
- NP: 1 group presentation of a paper (selected by you from a set of options we’ll provide) the last week of the semester
- NT: 3 group homework assignment (“tareas grupales”)

Final grade: (NC + NP + NT) / 3

Other stuff

- **Course repository**: `https://github.com/dccuchile/CC6205`
- The schedule will be listed there!

References i


Collins, M. (2013).
Statistical nlp lecture notes.
Lecture Notes.


Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011).
Natural language processing (almost) from scratch.
Journal of machine learning research, 12(Aug):2493–2537.


Deng, L. and Liu, Y. (2018).
***Deep Learning in Natural Language Processing*.**
Springer.


Goldberg, Y. (2017).
Neural network methods for natural language processing.
Synthesis Lectures on Human Language Technologies, 10(1):1–309.

References ii


Johnson, M. (2014).
Introduction to computational linguistics and natural language processing (slides).
2014 Machine Learning Summer School.


Jurafsky, D. and Martin, J. H. (2023).
Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.
Prentice Hall, Upper Saddle River, NJ, USA, 3rd (draft) edition.


Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013).
Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets.
Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013).

References iii

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013).
Semeval-2013 task 2: Sentiment analysis in twitter.
In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.