Final Project Dataset - IMDB Dataset of 50K Movie Reviews

- This is a dataset for **binary** sentiment classification containing substantially more data than previous benchmark datasets. 25,000 highly polar movie reviews for training and 25,000 for testing.
- Classification task.
- Challenge: (1) Transforming Text into Numerical Data; (2) High-dimensional Feature

Example

Positive or negative review?

I saw this movie when I was about 12 when it came out. I recall the scariest scene was the big bird eating men dangling helplessly from parachutes right out of the air. The horror. The horror.As a young kid going to these cheesy B films on Saturday afternoons, I still was tired of the formula for these monster type movies that usually included the hero, a beautiful woman who might be the daughter of a professor and a happy resolution when the monster died in the end. I didn't care much for the romantic angle as a 12 year old and the predictable plots. I love them now for the unintentional humor. But, about a year or so later, I saw Psycho when it came out and I loved that the star, Janet Leigh, was bumped off early in the film. I sat up and took notice at that point. Since screenwriters are making up the story, make it up to be as scary as possible and not from a well-worn formula. There are no rules. **Negative**

not so obvious review...

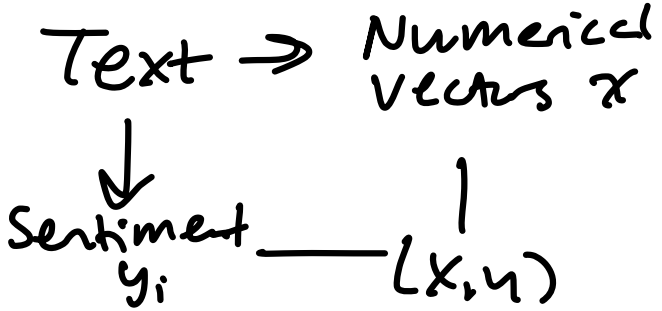
true answer

- **Example**: "I stopped in because I was hungry for some snacks. Browsed the store since I had some spare time and found it to be **clean** and **well stocked**. Wide isles, **good** selection of bakery stuff, flowers and all your usual groceries. I like that they had some locally made snacks too like California Kettle Corn and Taco Works (out of SLO). Found the semi-unusual chips I wanted that other stores don't always carry, grabbed a Diet Coke by the checkout and I was **happy**. **Convenient** parking lot and the staff was **friendly** and **helpful**. Thanks Albertsons!"

- True Rating: 4
- Averaged rating of this user: 3.79
- Averaged rating of this restaurant: 3.
- **Question**: What are useful words for you to predict the rating of this review?

Challenges in dealing with text data

- Numericalization is necessary
- Data size can be **big** for both dataset and dimension of features.
- Eliminate unnecessary things in textual documents:
  - Personal Pronoun: He, She, I, We... REMOVE
  - Determiners – Determiners tend to mark nouns where a determiner usually will be followed by a noun examples: the, a, an, another
  - Coordinating conjunctions: for, an, nor, but, or, yet, so
  - Prepositions – in, under, towards, before
- Transform **adverb** to **adjective**: Happily to Happy.



Numericalization

- Create a sentiment dictionary:

$$D = \{w_1, \ldots, w_d\},$$

where w_i are words. For example,

$$D = \{\textit{Friendly}, \textit{Happy}, \textit{Sad}, \textit{good}, \ldots\}$$

- **Bag-of-words**: For any piece of text $\boldsymbol{t}$, we can transform this text to a d-dimensional vector

$$\boldsymbol{t} \to B(\boldsymbol{t}) = (I(w_i \in \boldsymbol{t}))_{i=1}^{d}$$

- **Example**: $\boldsymbol{t}$ = Today is a good day. Then

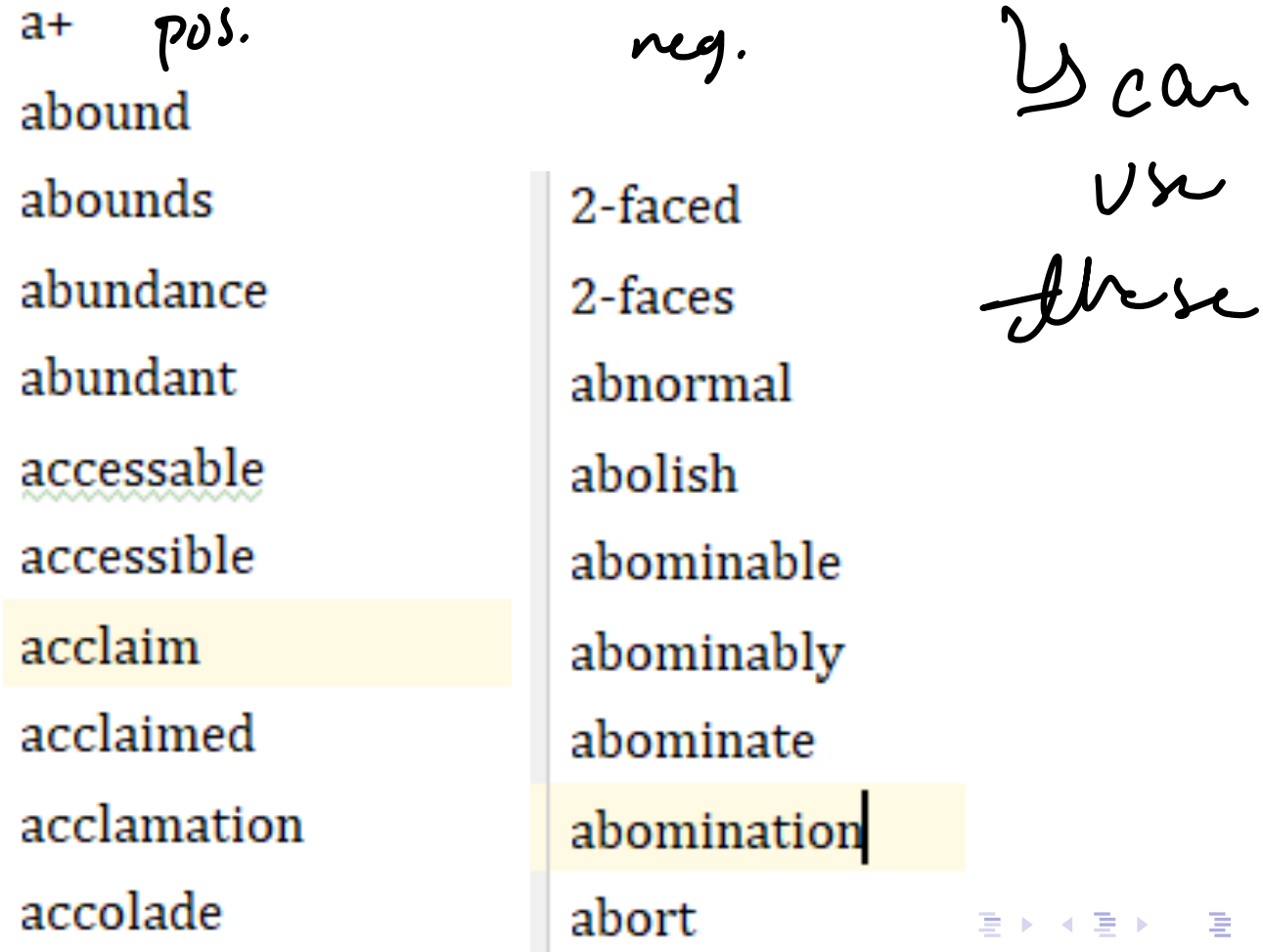
$$B(\boldsymbol{t}) = (0, 0, 0, 1, \ldots, 0)$$

the non-highlighted words don't appear in the dictionary

The fourth element is 1, indicating the existence of the word "good".

Opinion Lexicon

- Pre-defined opinion lexicon as your dictionary:
- Available:
  https://www.kaggle.com/datasets/nltkdata/opinion-lexicon



Term Frequency and Inverse document Frequency (TF-IDF)

if word is unimportant, can remove it

- Term Frequency (tf): Term Frequency: It is a measure of how frequently a term (usually a word) appears in a textual document. For example: The TF of word **Good** in a text document "Today is a good day" is 1/5. **Implication**: Large TF means this term is an important feature.
- Inverse Document Frequency (idf): For a term t, its inverse document frequency is calculated as

$$IDF(t) = \log\left(\frac{\text{Number of documents}}{\text{Number of documents with term t}}\right)$$

If ratio close to 1 → not important

Implication: Large idf means this term is an important feature.

- TF-IDF = TF x IDF: Use TF-IDF to choose words into dictionary

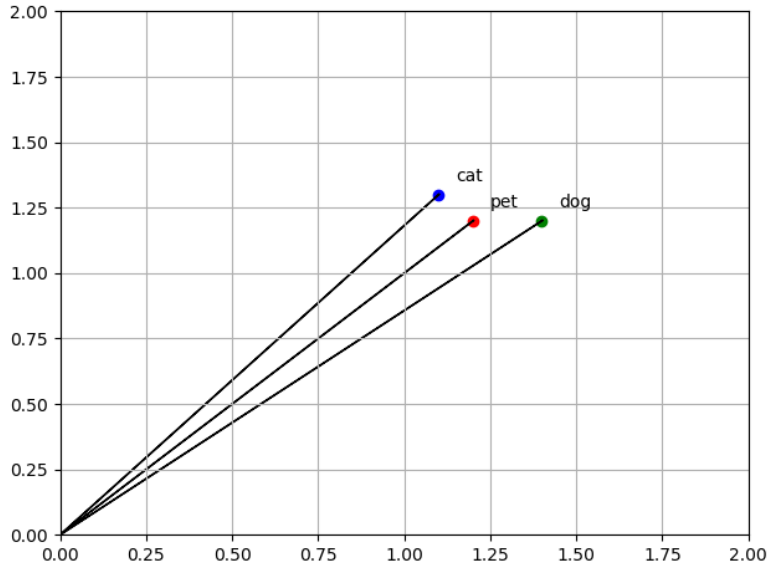
Word Embeddings

- **Objective**: represent a word as a numerical vector.
- **Typical Methods**: Word2Vec
- Intuition is that words appearing in the same contexts share semantic meaning

"*Joe has a* **pet dog** , and Curry has a **pet cat.** "

context window

context window



If two words are close in meaning, they should be close in embedding

"Shallow" neural network