HOW TO UNDERSTAND MILLIONS OF WORDS IN A MINUTE

SENTIMENT ANALYSIS

"The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral."

Oxford Dictionaries

ABOUT

Sentiment Analysis allows estimation of proportions for categories in a target population without classifying each individual document.

Key advantage is its flexibility:
1. Repetition both over and in real-time
2. No need to encode every word manually
3. Results adapt to pattern changes.

PROCEDURE

The procedure starts with the construction of the training set. With the results from the training set (T), the key information is used to begin the sentiment analysis process (S).

1. T → 2. S

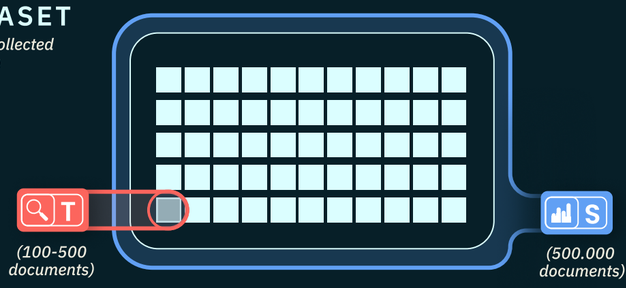
DATA EXTRACTION

Sentiment Analysis begins with gathering relevant text data of the topic. Some possible sources are:

- online articles
- blogs
- social media
- customer reviews

DATASET

all collected data



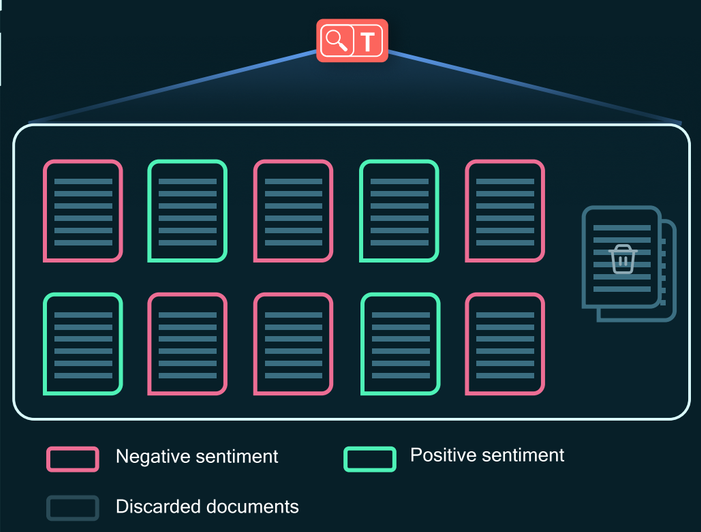
THE CONSTRUCTION OF THE TRAINING SET

START
THE EXAMPLE

1 CATEGORY DEFINITION

Label each text document into categories.
Documents must be divided into so that:
- No document belongs to several categories
- No uncategorized documents

Ex. label the texts "positive" or "negative". ' Off-topic documents, different languages, and spam are removed.
This is done manually by data scientists.



2 TEXT PREPROCESSING

Transform collected texts into data variables to be computed by simplifying text into a short list of meaningful unigrams. This is done automatically through software.

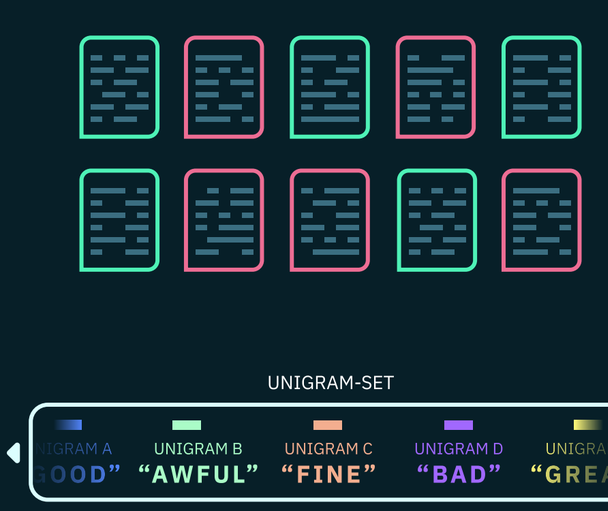
3 UNIGRAM PREPROCESSING

Lower-case, and lack of punctuation.
Documents are converted into unigrams

Unigrams: A one word sequence
Ex. Running or runner are reduced to RUN
Unigrams deletion: if appears in fewer than 1% and more than 99% of all documents

4 UNIGRAM COUNTING

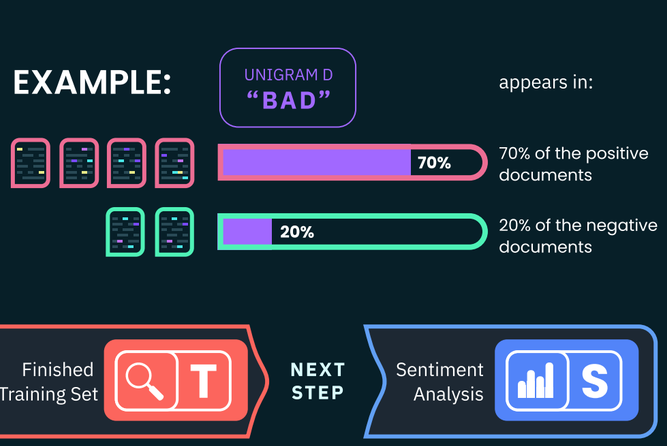
A set of polarized unigrams *(ex: good VS bad)* are chosen to regulate positive and negative sentiment ratio of the training set.
Unigrams are counted on the entire training set and both categories.



5 RESULT OVERVIEW

This determines:
- Overall proportion of positive and negative documents in the trainig set sample.
- Occurrence and proportion of each unigram in each category.

This is a key information to perform the sentiment analysis



SENTIMENT ANALISYS

1 TARGET DATASET ANALYSIS

The algorithm analyzes the target dataset and checks all of the documents containing all the unigrams.

2 ESTABLISHING ASSUMPTIONS

- Sum of all positive and negative documents equals total documents in the dataset
- Sum of positive and negative documents with unigram equals all documents with unigram in dataset

3 TAKING VALUES OF THE TRAINING SET

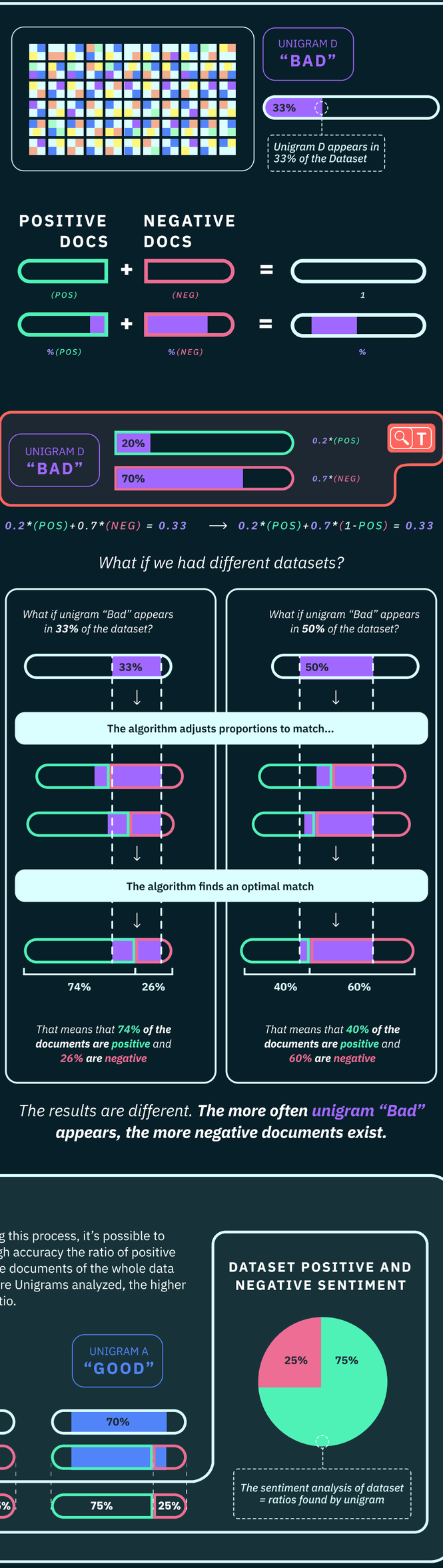
4 ADJUSTING PROPORTIONS

The algorithm adjusts the proportions of the categories until the number of the documents that contain a unigram will match the dataset, while the sum of the proportions stays the same.

Meaning:

Percent of documents with the unigram in each category is fixed

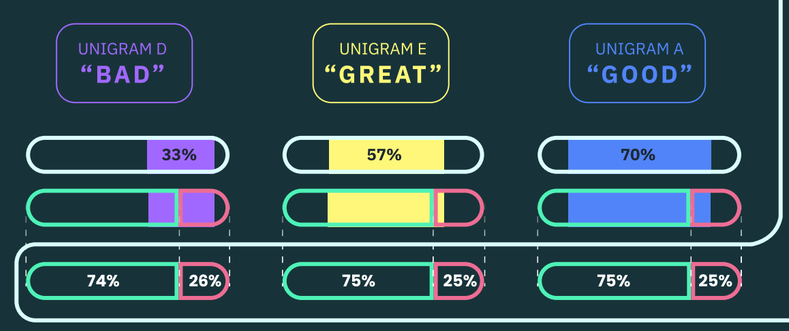
Proportion of categories is dependent on the presence of the unigram in the dataset



The results are different. ***The more often unigram "Bad" appears, the more negative documents exist.***

RESULTS

By repeating this process, it's possible to find with high accuracy the ratio of positive and negative documents of the whole data set. The more Unigrams analyzed, the higher accuracy ratio.



VISUAL EXPLANATIONS OF STATISTICAL MODELS

Sentiment Analysis

AUTHORS

Anel Alzhanova
Soraya Astaghforellahi
Maria Camila Coelho
Beatrice Foresti
Yaqing Luan
Nelly Saad
Severin Alois Schwailghofer

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Andrea Benedetti
Tommaso Elli
Beatrice Gobbo

Final Synthesis Design Studio
Sect. C3

LM in Communication Design
A.A. 2021/2022

DEN-SITY GN+

POLITECNICO MILANO 1863