Analisi di *Sentiment Penn TreeBank*

A.A. 2023/2024

Rosso Carlo

Contents

1 Introduction

The analysis of the corpora is divided in three phases:

1. summary of the information provided in the paper `https://aclanthology.org/D13-1170/`, which introduce the dataset for the first time;
2. analysis of the dataset, in order to understand the structure of the data and the possible applications. Particularly, I've never done such an analysis, so I asked chatGPT to help me in this task;
3. integration of the analysis after a discussion with the professor who follows me in the project.

2 Summary from the paper

Further progress towards understanding compositionality in tasks such as sentiment detection requires richer supervised training and evaluation resources and more powerful models of computation. Particularly, in this note, we will discuss about the former. The authors introduce a Sentiment Treebank, which includes fine grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences. The *Sentiment Penn Treebank* is the first corpus with fully labeled parse trees that allows for a complete analysis of the compositional effects of sentiment in language.
The corpus is based on the dataset introduced by Pang and Lee (2005) and consists of 11,855 sentences extracted from movie reviews. It was parsed with the Stanford parser and includes a total of 215,154 unique phrases from those parse trees, each annotated by 3 human judges.

2.1 Origin

It is considered the corpus of movie review excerpts from the `rottentomatoes.com` website originally collected and published by Pang and Lee (2005). The original dataset includes 10.662 sentences, half of which were considered positive and the other half negative.
The Stanford Parser (...) in used to parses all 10.662 sentences. In approximately 1.100 cases it splits the snippet into multiple sentences. It was used Amazon Mechanical Turk to label the resulting 215.154 phrases.
Starting at length 20, the majority are full sentences. One of the findings from labeling sentences based on *reader's perception* is that many of them could be considered neutral. Furthermore, the longer the sentence, the stronger the sentiment.

3 My analysis of the dataset

According to ChatGPT at the start it is important to describe the dataset in general, while most of the following information is already given in the previous section, I will repeat those:

- **Origin of the dataset**: `rottentomatoes.com`, then elaborated by Pang and Lee, the Stanford Parser, Amazon Mechanical Turk. Finally three judges were asked to decide on the final sentiment of the reviews. Note that the dataset is made of movie reviews;
- **Number of patterns**: 215.154 phrases from 10.662 sentences, but they are treated independently;
- **Number of classes**: 5: negative, somewhat negative, neutral, somewhat positive, positive. Actually the value given to each sentence ranges in 25 classes, but in the paper they are grouped in 5, since the most extreme values are not very common. I suppose I will use only 5;
- **Language**: English;

3.1 Labels distribution

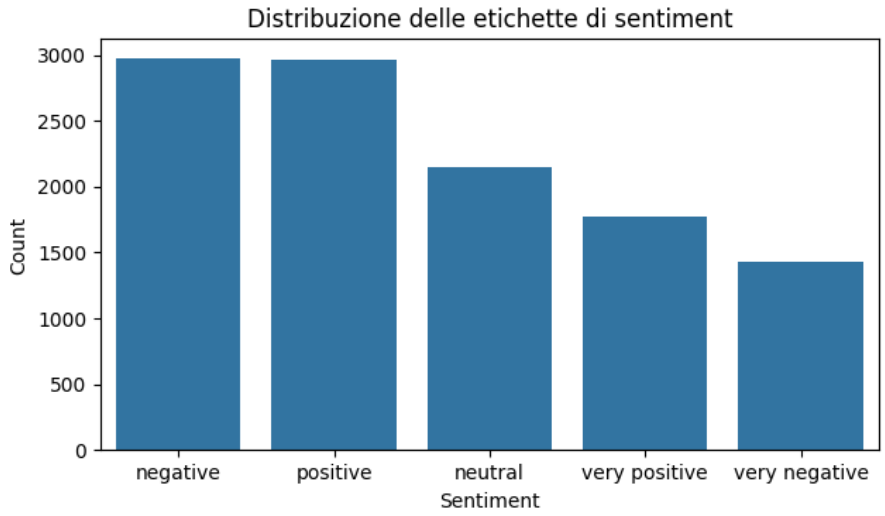

Figure 1: Labels distribution

Back to the actual numbers, we got the following distribution:

- **Negative**: 1432;
- **Somewhat negative**: 2971;
- **Neutral**: 2144;
- **Somewhat positive**: 2966;
- **Positive**: 1773;

3.2 Texts length distribution

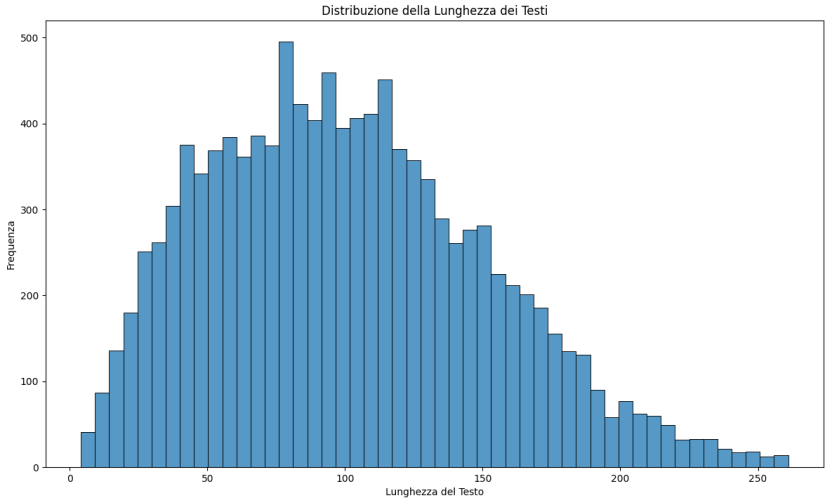

Figure 2: Texts length distribution

Where the mean of the texts length is 101 and the median is 97.

3.3 Most frequent words

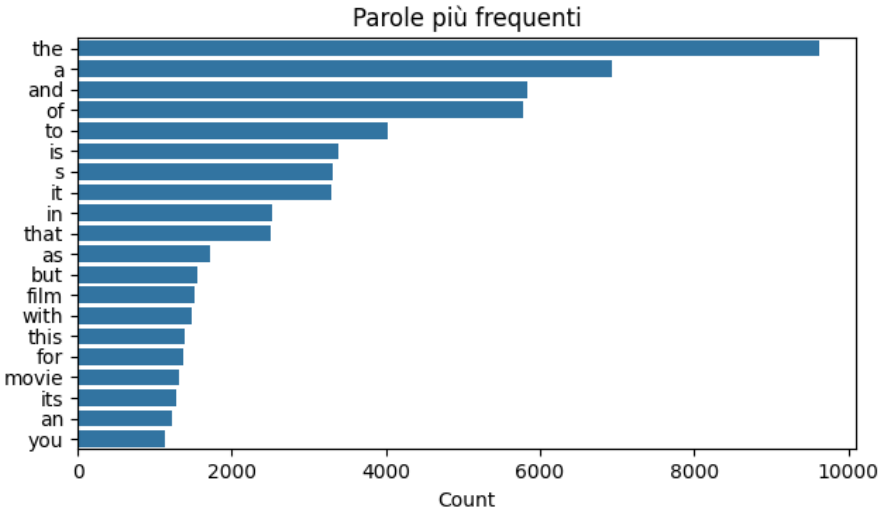

Figure 3: Most frequent words

Finally the most frequent words are:

- **the**: 9615;
- **a**: 6934;

- **and**: 5841;

- **of**: 5787;

- **to**: 4033;

3.4 Sentiment visualization

The sentiment visualization is done through word clouds, where the size of the word is proportional to the frequency of the word in the dataset, within the label. The following figures show the word clouds for each label.

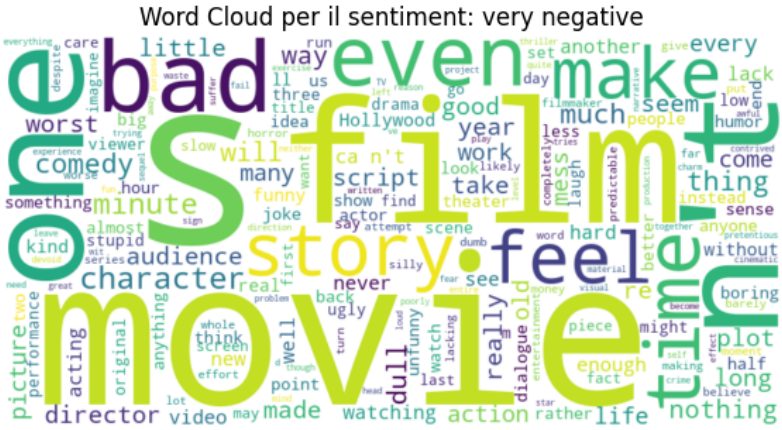

Figure 4: Negative word cloud

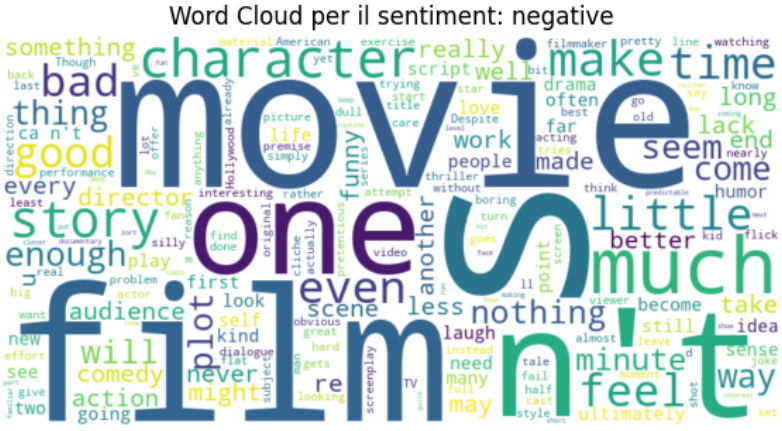

Figure 5: Somewhat negative word cloud

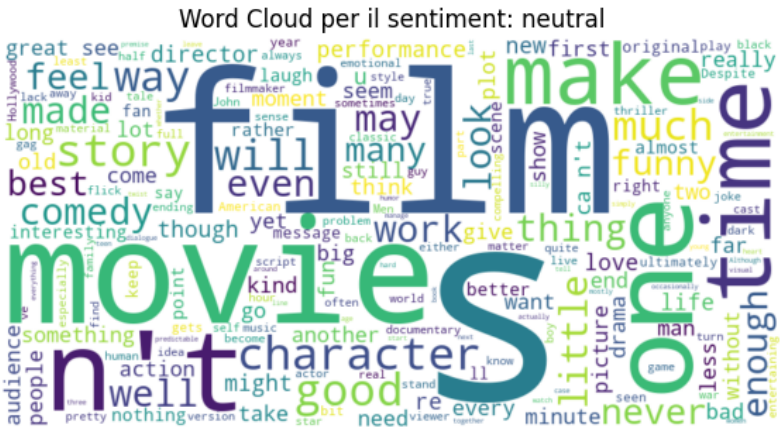

Figure 6: Neutral word cloud

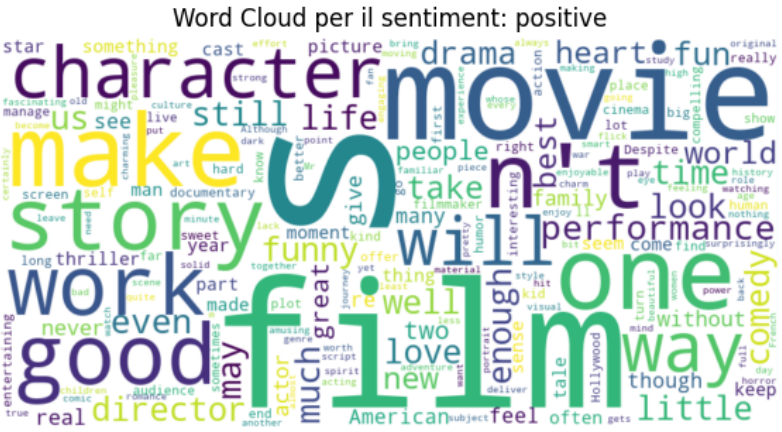

Figure 7: Somewhat positive word cloud

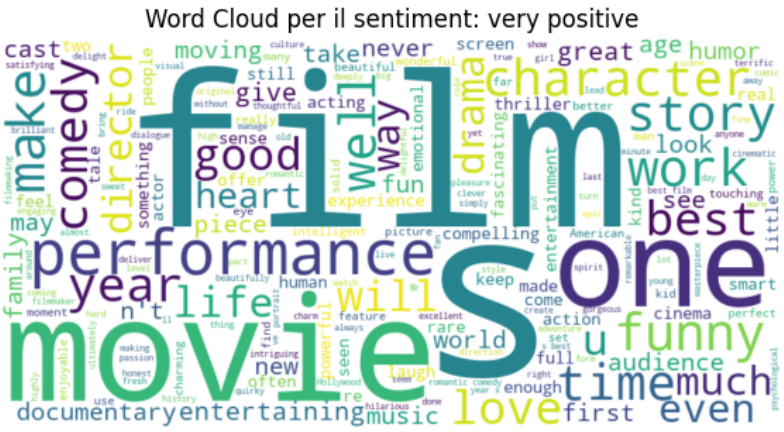

Figure 8: Positive word cloud

4 Integrations

- **Vocabulary dimension**: 17.178 words;
- **Total number of words**: 193.932 words;