Machine Learning and Data Mining project: Corpus Analysis

Baurice Nafack

Course of AY 2021-2022 - Introduction to Machine Learning and Evolutionary Robotics, Prof Eric Medvet

1 Problem statement

Italo Svevo, a pioneer of the psychological novel in Italy and one of the greatest Italian novelists, wrote and received letters in multiple languages during the twentieth century. The letters were recorded and stored in a database. Thus, the purpose of this project is to analyze those epistolary corpora to gain insight into topics and sentiments expressed (positive or negative) in his letters. This approach intends to extract information from the corpora by looking at the relationships between subjects, people, and emotions (fear, joy, sadness, anger), as well as how those interactions change over time.

2 Algorithms use to Approch the problem

The number of topics discussed in letters will be determine by using a Correspondance Analysis (CA). A visualization technique for identifying and displaying the relationship between categories. For the topic modeling, Latent Dirichlet allocation (LDA) method will be apply for fitting topic model. This is a stochastic algorithm that could have different results depending on where the algorithm starts, so we will need to specify a seed to ensure reproducibility of results.

To find the emotion and sentiment express in each letter, the FEEL-IT (Emotion and Sentiment Classification for Italian Language) model building in python release in 2021 [1] will be used.

3 Data description

The Svevo letter corpus dataset contains a total of 894 letters written by Italo Svevo including 826 letters written in Italian, 28 letters written in German, 30 in French, and 10 in English all written between 1885 and 1928. Data includes information about: the name of the corpus section, the index of the letter in the section, the date of the letter, the year of the letter, the sender of the letter, the sender's location, the recipient's location, languages used in the letter, the main language, and the text of the letter. There are 12 variables in total. Letters were mostly written in Italian with 816 sending by Ettore Schmitz, 30 by Eugenio Montale, 15 by Marieanne Crémieux Comnène, 11 by James Joyce, 8 by Benjamin Crémieux, 8 by Valerio Jahier, 5 by Valéry Larbaud, and 1 by Benjamin Larbaud to Svevo and his wife. We noticed an unbalanced distribution of data.

4 Experimental Procedure

For the experimental procedure, other corpora not writing in Italian were not considered since they did not contain enough letters to add meaningful information.

4.1 Data cleaning and Preprocessing

The following step were consider : remove all punctuation and numbers, lowercase the letters, delete the unnecessary white spaces and useless words (u, essere, p, dopo etc), remove italian stopwords and proper nouns, and the letters where then converted to DocumentTermMatrix which automaticaly apply tokenization.

4.2 Results and discussion

The Correspondence Analysis identifies the top five groups of topics that have been discussed in time, left figure 1. This number was used to fit the LDA to identify the top words discussed in each topic, right figure 1 with β (tf-idf) the probability of that term being generated from that topic. The topic modeling process has identified groupings of terms that we can understand as human readers and assigned to a group of discussion. The following topic was finally identified: Work (topic 1), Severo Book(topic 2), Daily Life (topic 3), Opinion (topic 4), and Travel (topic 5). By exploring the wordcloud figure(2), the following grouping of words and interpretation was found *viaggo and londra* refer to travel, *affari and compagnia* refer to Work, *pareva, credo, and trovo* refer to the opinion, *bella,oggi, and adeso* refer to daily life and *scrivere and lettere* refer to a book.

The study of the topic discussion proportions over year provides a distant view of the topic in the data. I aggregate mean topic proportions per year of all letters. These aggregated topic proportions can then be visualized in figure (3). It shows that topics around opinion, daily life, and travel dominate the first decades. In the second decade, the discussion arount work was the most important topic. In the last decade, the discussion around Severo's book became the most important topic. This can be interprete by saying books become more valuable with time. Based on the proportion of topics discussed by sender, it becomes evident that the majority of discussion around Severo's book were related to Eugenio Montale, James Joyce, Valerio Jahier, and Valery Larbaud, while Ettore Schmitz's approximaly discuss all topics with same proportion.

The analysis of emotion and sentiment using the FEEL-IT model helps to find the proportions of words related to them to have their trend profile over the year, see figure (4). It shows the expressions of anger, joy, sadness, and fear linearly increase in the first 2 decades, but in the last decade, the sadness emotion almost disappears. The negative sentiment is dominant in the first two decades, but in the last decade, the positive sentiment becomes dominant. Figure 5 show that those emotion and sentiment are mostly connected to Ettore Schmitz who express more sadness and negativity. the data are unbalanced, this affects the interpretation of the letter sentiment and emotion over the decade. The analysis of emotion per topic shows that the discussion around the Severo's book has a proportion of joy and an almost equal proportion of expression around other topics, it also has more positivity than the order topic of discussion figure 6

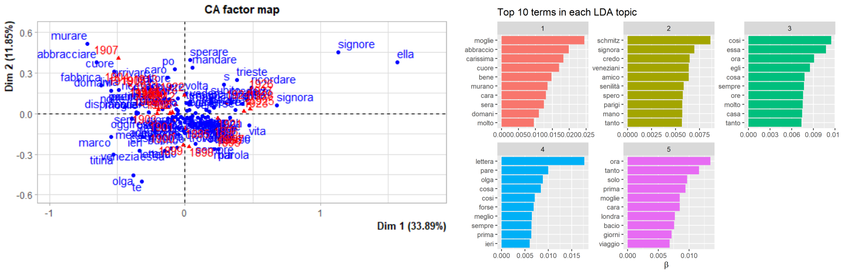

Figure 1: Correspondance Analysis result(left) and LDA result(right figure).

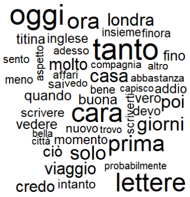

Figure 2: Word cloud.

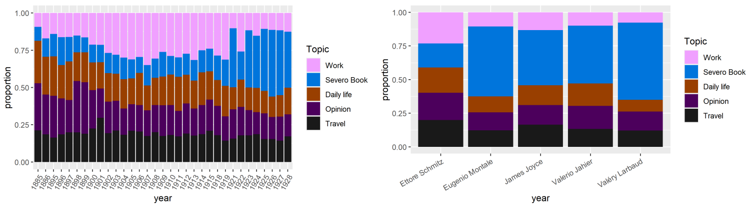

Figure 3: Topic proportion per yer and per sender

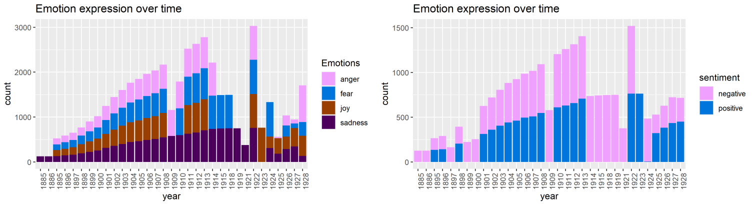

Figure 4: Emotion and sentiment express over year on different letters

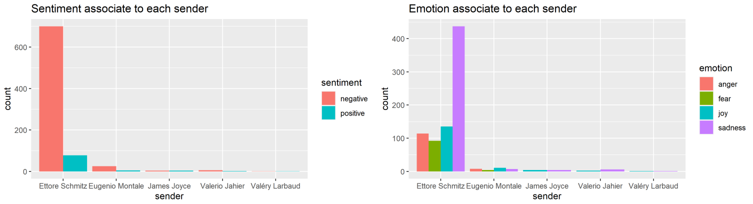

Figure 5: Emotion and sentiment express by each sender

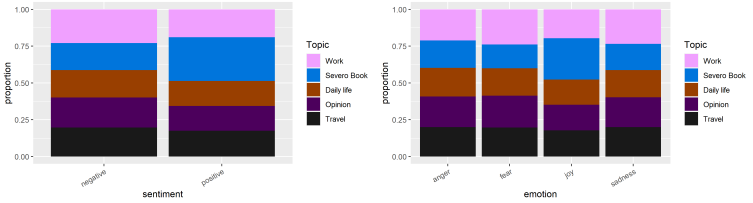

Figure 6: Emotion and sentiment per topic

References

[1] Federico Bianchi, Debora Nozza, and Dirk Hovy. "FEEL-IT: Emotion and Sentiment Classification for the Italian Language". In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2021.