Introduction to Machine Learning project: Analysis of Svevo's letters corpus

Davide Basso[1] and Francesco Tomba[2]

[1] Sentiment analysis, solution design, solution development, writing
[2] Topic modelling, solution design, solution development, writing

Course of AA 2020-2021 - Master Degree in Data Science and Scientific Computing

1 Problem statement

In this project we had to perform topic modeling and sentiment analysis on a collection of ∼900 letters sent and received by the Italian writer Italo Svevo during his lifetime. The aim is to discover which are the main topics and sentiments that characterize the corpus and also to understand how these evolve over the years and by considering different senders and recipients.

2 Data description and exploration

Letters are written in various languages and dialects. Almost 90% of them (826 letters in total) is written in Italian, so for our analysis we decided to consider only this subset since for any other language the number of letters is not enough to perform a statistically meaningful analysis.

Svevo had most of the conversations with 9 people. So only people who had more than 10 conversations with Svevo were taken into account and the others (63 letters) were tagged as 'OTHERS'. Since almost 75% of the conversations were kept with his wife, we expect that the dataset will present some imbalance towards topics and sentiments related to her. In particular we expect that themes as "family" or "travels" as well as feelings like "joy" or "trust" will cover the vast majority of the letters.

3 Proposed solution

The input of the model will be the corpus of the Svevo letters for both the analysis and the output we want to obtain is a list of topics discussed within the corpus. The solution we propose is based on applying Latent Dirichlet Analysis

(LDA). Evaluating the performance of the model is difficult because there is no prior classification of these letters to rely on. We decided to compare the frequency of some topic with respect to biographical events of Svevo's life (such as travels and publications). Another way we used was to choose a "coherency metric" from the ones found in literature, in particular we used the "U-Mass" metric [6]. The score obtained is a negative number associated with each topic, the closer to 0 the better the model has performed in extracting the topic.

For what concerns sentiment analysis, we decided to extend our research also in finding emotions present in letters. To do so we used "NRC Word-Emotion Association Lexicon, aka NRC Emotion Lexicon, aka EmoLex"[3]. In this way we were able to have measurements in terms of scoring points regarding whether a sentiment was more present or not in a letter. The higher the value, the higher the presence in the text of words related to a certain sentiment/emotion.

4 Experimental evaluation

4.1 Experimental procedure

Topic modeling

The solution proposed was implemented in python, using the spaCy library [1] for preprocessing the text and then the gensim library [5] to import and train the LDA model. From the text were eliminated punctuation and stopwords. We removed also some other custom words (such as "mano" and "signore") and some expressions of Trieste's dialect such as ("xe" and "el"). The library also performs POS tagging and we decided for that purpose to keep only nouns, proper nouns, and verbs. We found in literature [2] that pruning the dictionary from the less and more frequent terms helps the model in learning. Due to the fact that the dataset is composed of texts of different length we tought it would be useful to test the effectiveness of a tf-idf approach.

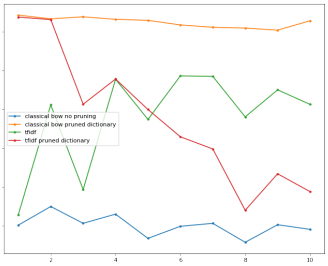

Figure 1: Plot of the coherency scores as a function of the number of topics of the LDA models trained with different dictionary approaches. The dictionary was pruned by removing words that do no appear in less than 5 documents and in no more than the 50% of the corpus

LDA needs also the number of topics (n) to be "a priori" fixed, so we had to find an optimal n. The approach was to increase the number of topics and calculate the coherency score, as we can see from results in figure 1. We chose to use the model trained on 4 topics using bag-of-word on a pruned dictionary for the following analysis. The aim was also to avoid to choose too few or too much topics. We saw in general that models performed better for n = 4.

Sentiment analysis

EmoLex is implemented in R package *syuzhet*. We opted for this method even if some researches has shown that for negatives categorization this package isn't an optimal choice[4] (but still is the most complete tool that also enables support for Italian language). Pre-processing steps involved case lowering the entire corpus, tokenization and use of POS tagging in order to maintain nouns, verbs, adjectives and punctuation. In particular the latter one was not removed because, even if we are considering a corpus of early 900's letters, quotation marks, exclamation marks or other punctuation characters could be interesting from sentiment analysis point of view. All these steps were performed using *spaCy* library[1].

In this way, at each pre-processed letter it corresponds a vector containing values that belong to one emotion or sentiment. Since letters length could vary, with the aim to normalize our results we opted for dividing each sentiment and emotion score present in the letter respectively by the sum of the sentiment and emotion ones.

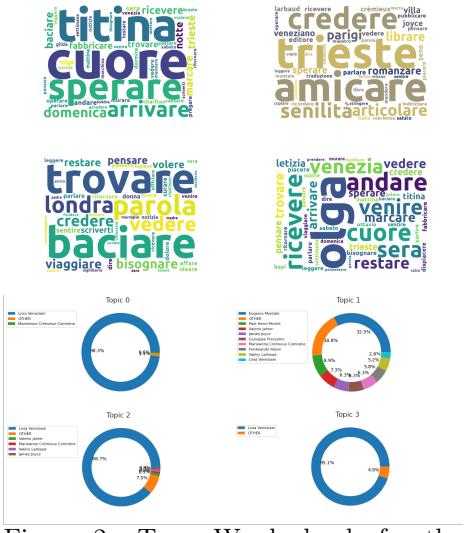

Figure 2: Top: Word clouds for the topic extraction in the letters, topics are numbered by rows [0, 1], [2, 3] Bottom: Donut plots of the topic's frequencies

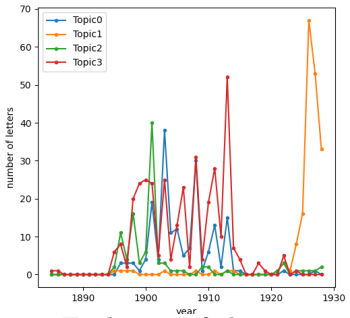

Figure 3: Evolution of the presence of the topics in the letters over the years

4.2 Analysis of the results

The output of the LDA model is a list of topics, at each one of them corresponds a list of the most frequent words and their respective frequencies. Figure 2 shows 4 word clouds representing the output of the model. From those result we noticed that: i) topics 0, 2, 3 contained words related respectively to "Love", "Travels" and "Family" and that those topics are present in the vast majority of the letters written to Livia Veneziani ii) Topic 1 in which the word "senilità" stands out instead is present in the vast majority of the letters related to the personalities of the literary landscape of those years.

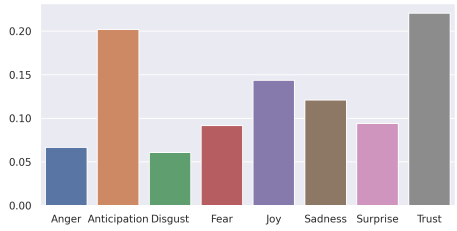

Figure 4: Overall emotions percentages

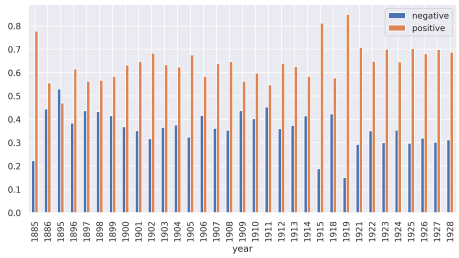

Figure 5: Sentiments percentages over years

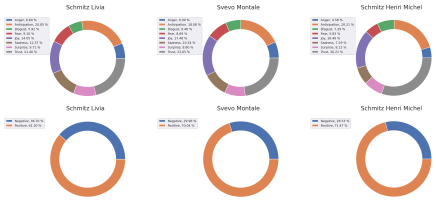

Figure 6: Sentiments and emotions for the top 3 writers/senders

To further analyze the results we plotted the evolution of the topics covered over time (figure 3), and we noticed that Topic 0, 2 and 3 were more present in the periods Svevo was away from Trieste while Topic 1 was more significant namely in 2 periods: late years of 1800 and 1925-27, coinciding with publication of the most important Svevo's works.

Results of our procedure for sentiment analysis are shown in plots 4, 5, 6; we computed overall percentage of sentiments and emotions and then by grouping them by year and different interlocutors.

In order to evaluate these results we compared some of the estimated sentiment and emotion normalized scores with the actual text of the corresponding letter. It turned out that results were pretty good, however in general negative emotions and sentiments seemed to be a little underestimated. It's interesting to observe how positive and negative sentiment percentages vary during years accordingly to specific events of Svevo's lifetime. In 1895 author's mother dies and we can notice that this is the only year where negative is the prevalent sentiment. On the other hand, 1919, year in which Svevo starts writing his masterpiece "La coscienza di Zeno", states the highest value for positive sentiment. Even maximum values for anticipation and trust, along with lowest ones for anger and fear, coincide with this specific year.

4.3 Final remarks and conclusions

The topic extraction model was able to find 4 different topics that are covered in the letters. As we previously said data we had were very imbalanced and so the analysis may be tuned further by considering only the letters sent to the wife and the letters sent to other people. Nevertheless, coherence between both topic and sentiment analysis with actual author's life might be a confirm of the goodness of the model proposed as solution to this problem.

References

[1] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

[2] Daniel Maier, Andreas Niekler, Gregor Wiedemann, and Daniela Stoltenberg. How document sampling and vocabulary pruning affect the results of topic models. *Computational Communication Research*, 2(2):139–152, 2020.

[3] Saif Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *CoRR*, abs/1308.6297, 2013.

[4] Maurizio Naldi. A review of sentiment computation methods with r packages, 2019.

[5] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[6] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA, 2015. Association for Computing Machinery.