Topic Modeling and Sentiment Analysis of Svevo's Letters Corpus

Luca Crozzoli[1]

[1] problem statement, solution design, solution development, data gathering, writing

Course of AA 2019-2020 - Ingegneria Elettronica e Informatica [IN20]

1 Problem statement

From a corpus of 894 letters sent by and received to Italo Svevo[1], we will perform topic modeling and sentiment analysis. We will find k different topics by means of LDA[2] considering the 30 most relevant words for each one. By associating each letter with the most relevant topic, we will find who are the people which each topic is most associated with and how the interest on different topics evolve over the time. Then we will perform sentiment analysis by means of NRC Emotion Lexicon[3]. From the corpus we will study the distribution of positive, negative sentiments and emotions: Anger, Anticipation, Disgust ,Fear ,Joy ,Sadness ,Surprise ,Trust. We will find how these are related with people and topics and how they evolve over the time.

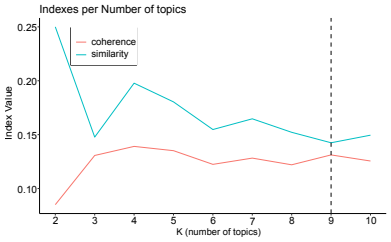
2 Data

The corpus contains information about letters, most relevantly: the date, the corpus section name (marks letters exchanged with the same pair of people) ,the sender and recipient names, the languages used in the letter, the main language and the text of the letter. After analyzing the corpus, we found that 826 (92.4%) of the letters have Italian as the main language, while the other languages are 1.1% English, 3.1% German, and 3.4% French. Given the high proportion of letters with Italian as the primary language, we ignored all the others. Furthermore, 31 people out of 41 have exchanged fewer than five letters with Svevo, so we classified those 31 people as "Others." We also found that the majority of correspondences are from Svevo's wife, Livia Venziani (612 letters) and Eugenio Montale (62 letters), suggesting a possible high bias in the topics of wife devotion and literature.

3 Performance Indexes

Due to the unsupervised nature of techniques used, is not easy to evaluate models. Particularly in LDA we needed to tune the hyper parameter k, the number of topics. To determine the proper value of k we varied k in [2,..,10] and calculated two indexes: 1) Topic Coherency[4]: We computed the arithmetic mean of coherences calculated on each of the k topics; 2)Jaccard Similarity[5]: We computed the arithmetic mean of the Jaccard similarities calculated among the k topics, considering the 30 most relevant words for each topic.

Figure 1:



4 Topic Modeling

The solution obtained is based on the LDA which took as input our DocumentTermMatrix (DTM)[6] derived from a pre-processed corpus. To obtain the pre-processed corpus first we removed numbers and punctuation, converted words to lower case, removed stop words in Italian, French, English, German and also words like ettore, schmitz, italo, svevo, commonly used by Italo Svevo to sign letters. Then, by using POS tagging[7] , we created a lemma dictionary to lemmatize[8] all Italian words. We also removed all the pronouns, adpositions, determiners and verbs that appears more than 1000 times. From the pre-processed corpus we created different DTMs, respectively for unigrams, bigrams and trigrams[9] [10]. For each of the three, we removed sparse terms[11], with a degree of sparsity of 0.99%, and terms that appeared less than three times in the corpus. Then we combined the three matrices to obtain one DTM composed by unigrams, bigrams and trigrams. We used this DTM as inputs for the LDA, executed many times by varying k in [2,..,10]. We chose the ideal k value considering the point that minimizes the differences between Jaccard Similarity and Coherence(lower Similarity and higher Coherence). From **Figure 1** we can see that the model that minimizes the difference is the one with k = 9. We chose k=9, ran LDA and plotted word clouds **Figure 2**. Then we classified each letter of the corpus considering the presence of the most relevant topic in each one. We

Figure 2:

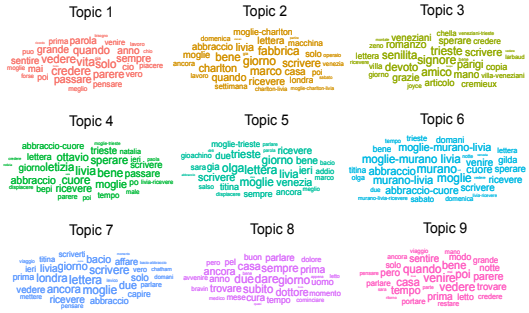
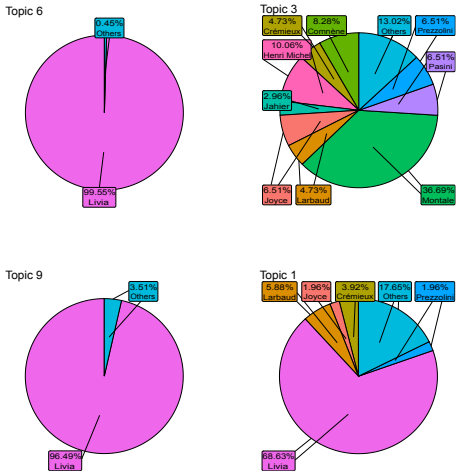


Figure 3:



sorted topics according to their probability within the entire corpus[12] and found that only 6,3,9, and 1 were the most important. For this reason we limited our analysis on these ones. To see the association between people and topics we considered individuals' names for those who had more than five letters exchanged, while we marked the rest of the people as "Others". We grouped letters sharing the same topic, and for each group, we calculated the percentage of letters addressed to recipients **Figure 3**. With the same method we plotted the proportion of letters per topic over time. We grouped letters by year (in not all the years there was an exchange of letters) and we calculated the percentage of letters associated with all 9 different topics (the most important 6,3,9,1, are outlined in red) **Figure 4**.

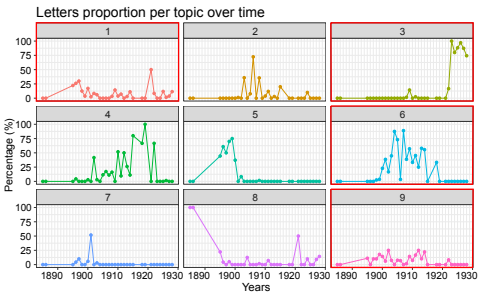
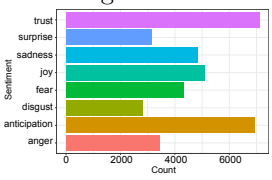


Figure 4:

5 Sentiment Analysis

Figure 5:



We used the pre-processed corpus and, by means of NRC Emotion Lexicon, we calculated the presence of the eight emotions (described in section 1) for each letter. Then for each emotion we summed all the eight values obtained from each letter, to characterize the presence of the emotions in the whole corpus **Figure 5**. Next we grouped letters sharing the same topic considering the most relevant ones 6,3,9,1, and we calculated the percentage of the presence of the emotions for each of them **Figure 6**. Then we repeated the same procedure but this time we grouped letters sharing the same correspondence to see how emotions are related with each person, considering only the ones whit whom Svevo had more contact **Figure 7**. Furthermore, with the same approach, we plotted sentiment and emotions proportion over the time, each one with the correspondent counterpart. We calculated the percentage of the presence of the emotions and sentiments for each year, in which there was an exchange of letters **Figure 8**.

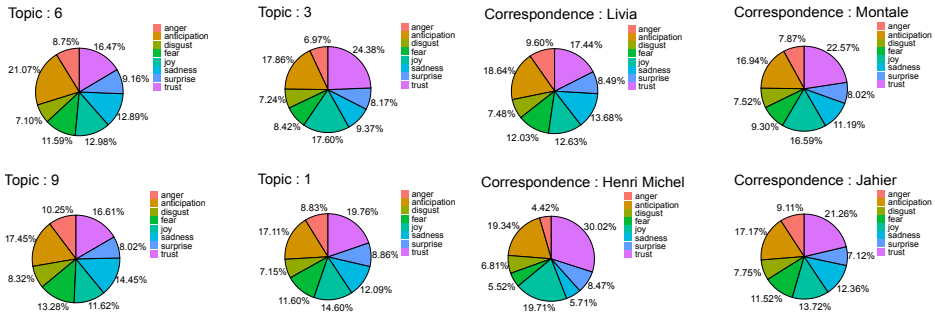


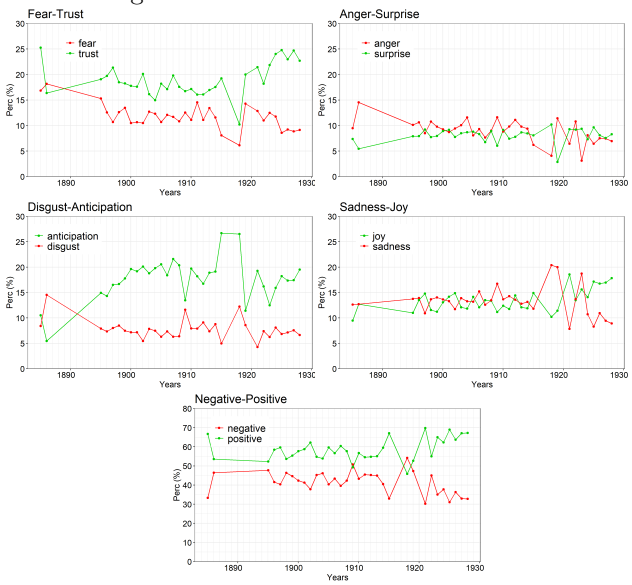
Figure 6:

Figure 7:

6 Conclusions

For each topic we assigned a label. Topic 1: personal thoughts; Topic 2: work; Topic 3: literature; Topic 5: family affair; Topic 6: wife devotion; Topics 4 and 7: family; Topic 8: health; Topic 9: travel. The most relevant topics were 6,3,9, and 1, which demonstrates that Svevo wrote mostly about his devotion to his wife and literature. From **Figure 4** we see that the interest in topic 6 starts after 1896, when Svevo married Livia Veneziani, while interest in literature peaks in 1923, when Svevo released his book "La coscienza di Zeno", and in 1925/26, when Eugenio Montale and James Joyce suggested the book to French Critics thus increasing its popularity. Health topic interest peaks around 1886 and 1895, when Svevo's brother and mother died due to their aggravated health conditions. Family topic covers the most of Svevo's lifetime, furthermore, in family affair topic, Svevo mentioned his future in-law Olga and Gioachino Veneziani, before the marriage with Livia. The writer worked and travelled for the Veneziani family's industry from 1899, period in which the interest in the work and travel topics increased. **Figure 8** tells us positive sentiments are dominant, apart in 1918 when Svevo experienced a period of tension, due to the annexation of his hometown Trieste to the kingdom of Italy. For each pair of emotions there is one more prevalent of its counterpart. In the graphs where negative emotions are dominant (anger, sadness) we can see there are some phases in which the positive counterpart prevails: particularly in 1925 when "La coscienza di Zeno" became popular. In **Figure 6 - 7** we see that for each topic and correspondence the main emotions are trust and anticipation.

Figure 8:



References

[1] "Italo Svevo" , Wikipedia, Wikimedia Foundation, 3 Gen 2022, https://it.wikipedia.org/wiki/Italo_Svevo

[2] Arga Adyatama, Joe Nathan Christian ,"Topic Modelling with LDA", 17 Apr. 2020, https://rpubs.com/Argaadya/topic_lda

[3] Saif Mohammad, "NRC Word-Emotion Association Lexicon", https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[4] Thomas W. Jones, "Topic Modeling", 27 June 2021, https://cran.r-project.org/web/packages/textmineR/vignettes/c_topic_modeling.html#:~:text=Probabilistic-,coherence,-measures%20how%20associated

[5] finnstats in R bloggers,"How to Calculate Jaccard Similarity in R", 13 Nov. 2021, https://www.r-bloggers.com/2021/11/how-to-calculate-jaccard-similarity-in-r/

[6] "Document-term-matrix",Wikipedia, Wikimedia Foundation,4 September 2021 https://en.wikipedia.org/wiki/Document-term_matrix

[7] Jan Wijffels, "UDPipe Natural Language Processing - Basic Analytical Use Cases", 2 Dec. 2021, https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-usecase-postagging-lemmatisation.html

[8] Stanford NLP Group Cambridge University Press, "Stemming and lemmatization",2008 https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[9] Kavita Ganesan, "What are N-Grams?", https://kavita-ganesan.com/what-are-n-grams/

[10] Mythreyi S., "Data Science Specialization Capstone: Milestone Report", http://rstudio-pubs-static.s3.amazonaws.com/162548_3d7b082333834b4bb0817c58b8f14d9b.html#:~:text=N%2Dgrams%20Identification

[11] "How does the removeSparseTerms in R work?", https://stackoverflow.com/questions/28763389/how-does-the-removesparseterms-in-r-work#:~:text=sparse%20argument%20to-,removeSparseTerms(),-%2C%20sparsity%20refers%20to

[12] Martin Schweinberger, "Topic Modeling with R", 2021-11-24, https://slcladal.github.io/topicmodels.html#approach-1