Topic Modeling and Sentiment Analysis of Svevo's Letters

Andres Bermeo Marinelli[1] and Samuele Coletti[2]

[1]Topic Modeling, problem statement, solution design, solution development, writing
[2]Sentiment Analysis, problem statement, solution design, solution development, writing

Course of AA 2020-2021 - [1]DSSC/[2]SSA

1 Problem statement

In this paper we will perform topic modeling and sentiment analysis on a corpus of 894 letters written/received by italian writer Italo Svevo[1]. We will extract the K main topics that characterize the corpus by means of LDA[2] along with the 10 most relevant words for each topic. By associating each document with the most relevant topic, we will study to whom each topic is mostly associated with and how topic interest varies over time.

As for sentiment analysis, by means of NRC Emotion Lexicon[3], we will study the incidence of positive and negative sentiments and eight emotions such as "Anger","Anticipation", "Disgust", "Fear", "Joy", "Sadness", "Surprise", and "Trust". We will also study how these change over time and to whom they are mostly associated with.

2 Data description

The input corpus contains information about each letter, most importantly: the date, the location and name of the sender/recipient, the languages present, the main language, and the letter itself.

We immediately realize that 826 letters (92.3%) have Italian as the main language while the remaining are spread amongst French (3.3%), German (3.1%), and English(1.1%). Due to their relatively small presence, letters whose main language wasn't italian were eliminated.

Furthermore, only 9 people had at least 10 correspondences with Svevo while all others had values below 5. This latter group was classified as "Others" for simplicity. The majority of letters (612) were exchanged with his wife Livia Veneziani, while the second most frequent correspondent, Eugenio Montale, only had 62 exchanges. This alerts us to the fact that there may be a large bias towards topics of family and daily activities.

3 Assessment and performance indexes

Being unsupervised techniques, it is hard to asses our models.

In regard to parameter tuning, for topic modeling, LDA requires as input the number of topics `K`. To choose this value, we varied `K` in $[1, ..., 10]$ and used three indexes: 1) Topic Coherency[4], 2) Mean "Internal" Jaccard Similarity[5] among the `K` topics by using the 10 most relevant words of each topic, 3) A measure of topic overlap for models with consecutive `K`. In other words, on average, how similar are the topics of a model with `K` topics to those of a model with `K` + 1 topics (always using Jaccard similarity).

4 Topic Modeling

Our solution is based on the LDA model applied to our corpus after applying some preprocessing steps. First, we converted all words to lower case and removed punctuation and stop words acquired from external sources[6]. We extended the list of stop words to include stop words in English, French, and German as well as typical words in triestino such as "xe", "ghe", "mulo". Next, we built bigram[7] and trigram[7] models, and then used lemmatization[8] in order to reduce each word to its lemma. Finally, by means of POS tagging[9], we kept only proper nouns, nouns, and verbs.

We built a dictionary with the preprocessed corpus and applied LDA with `K`= $[1, ..., 10]$ using both a BOW[10] and a TD-IDF[10] approach. We then pruned[11] the dictionary to contain only words that appeared more than 10 times overall and less than 50% in the whole corpus. We applied LDA using the pruned dictionary with both a BOW and a TF-IDF approach.

The ideal number of topics was chosen as the point that maximizes the distance between topic coherency and internal similarity while also checking that the topic overlap is low.

We can see in Figure (1) that the model that maximizes this distance is the TF-IDF approach on a pruned dictionary with $\mathtt{K}_{ideal} = 9$. However, when we analyzed the relevance of each topic in each document we concluded that only four -topic 0, 5, 7, and 8- were the most important. All remaining topics ranked third place or lower in relevance for each document and for this reason we limit our further analysis to these ones.

Next, we associated each letter to the most relevant topic and found that the majority of letters were associated to topic

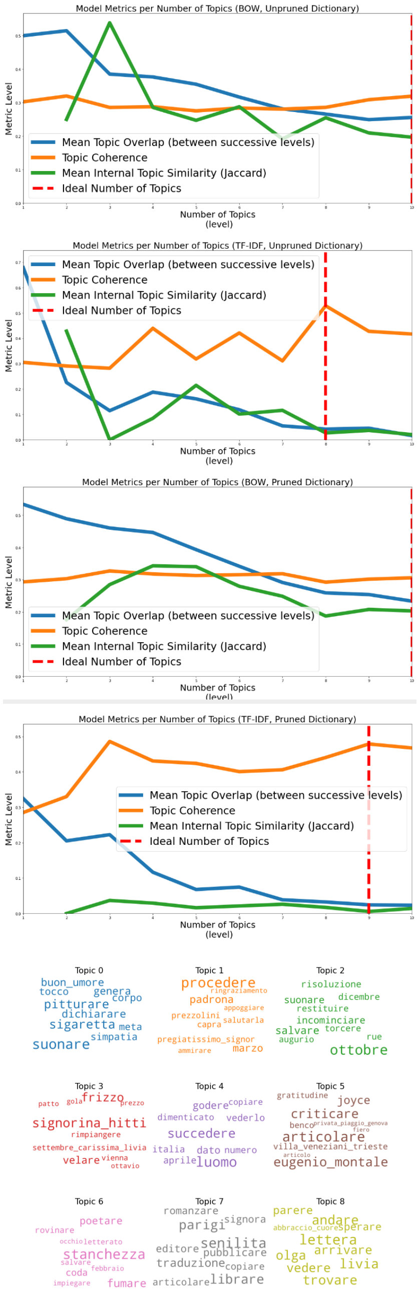

Figure 1:

8: family. This is to be expected given that 90% of letters are exchanged with his wife. To fix this issue, we associated to each letter both the most relevant and second most relevant topics.

5 Sentiment Analysis

We used the same preprocessing steps as above but we excluded bigrams and trigrams. Then, by means of the NRC Emotion Lexicon we associated to each word, for each of the eight emotions, a value between 0 and 1 which indicated if that word was associated with that particular emotion or not. We then counted the number of words relative to each emotion for each letter and added these values to characterize the overall presence of each.

In order to study the relation between emotions and people we considered only the individuals with whom Svevo had most contact: Livia Veneziani, Eugenio Montale, James Joyce, and Paul Henri Michel. For each individual, we added the number of words relative to each emotion and found the relative frequency.

Analogously, for each year we counted the words relative to each emotion and found the relative frequency by considering only the words used in the corpus for that specific year. This enabled us to characterize the presence of each emotion over time and see how these changed. To highlight interactions, we plotted each positive emotion with its negative counterpart: joy-sadness, trust-fear, surprise-anger, anticipation-disgust. This process was repeated for positive and negative sentiments to gauge which one was more prevalent over time.

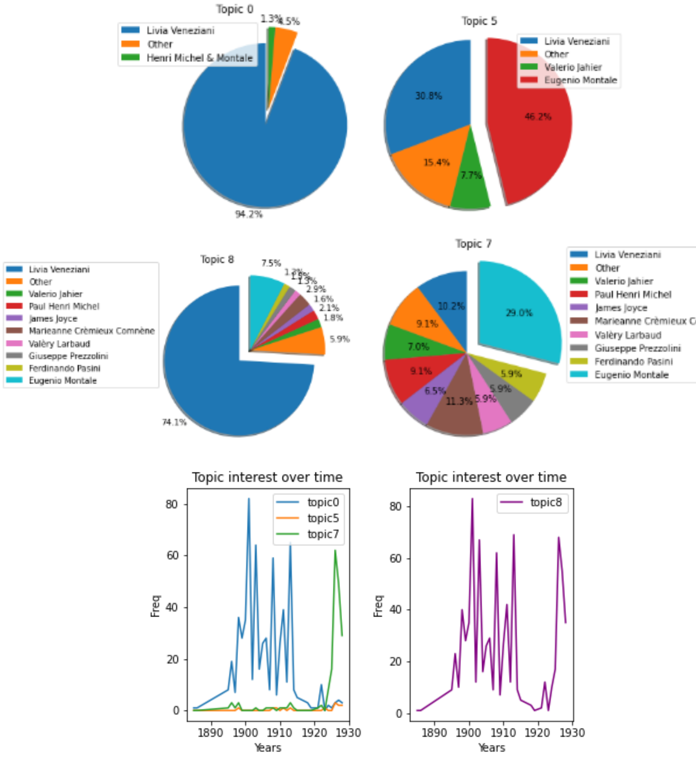

Figure 2:

6 Results and Discussion

For topic modeling, we can associate a theme to each topic. For topic 0: daily activities/passions, topic 1: nature/nature walks, topic 2: festivities, topic 3: personal problems, topic 4: success from his book, topic 5: gratitude towards his friends, topic 6: tiredness from writing, topic 7: literature, topic 8: family.

As mentioned above, the most important topics were topic 0, 5, 7, and 8 which tells us that Svevo mostly talked about family, literature, his friends and his passions. He talks about his activities mainly with his wife, who is also mostly associated to the topic of family. This is to be expected, since, as mentioned previously, 90% of the examined letters are exchanged with her. However, it is interesting to see that he also talks about family with fellow writers and friends and not just about literature. As expected, he talks about his gratitude to his friends mostly with Montale who aided in making his book popular. He talks about literature with more or less everyone which indicates his great passion for his profession. Finally, we notice that interest in literature and gratitude towards friends spikes around 1925, approximately when his book became popular. We notice that the behaviour of interest for topic 8 is basically the union of topic 0 and 7. This can be explained by remembering that for the majority of letters, LDA showed topic 8 as the most relevant which led us to associate to each document also topics ranked second place in relevance in order to extract some useful knowledge.

Sentiments are mainly positive during his entire lifetime with the presence of only two significant changes in tendency during the year of the death of his mother (1895) and in 1918 (a year of turbulence for Trieste due to its annexation to Italy). In 1923 his last book ”La coscienza di Zeno” comes out and, although not immediately popular, signals the beginning of a period in which positive feelings prevail. In fact, thanks to the help of James Joyce and Eugenio Montale, his last book starts to attain success in Italy and in Europe, particularly in France. We notice that positive emotions such as surprise, anticipation, and trust dominate over their negative counterparts throughout his life while sadness dominates over joy except for the period where his book became popular. We see that for all four of his main correspondents the main emotions are trust, joy and anticipation.

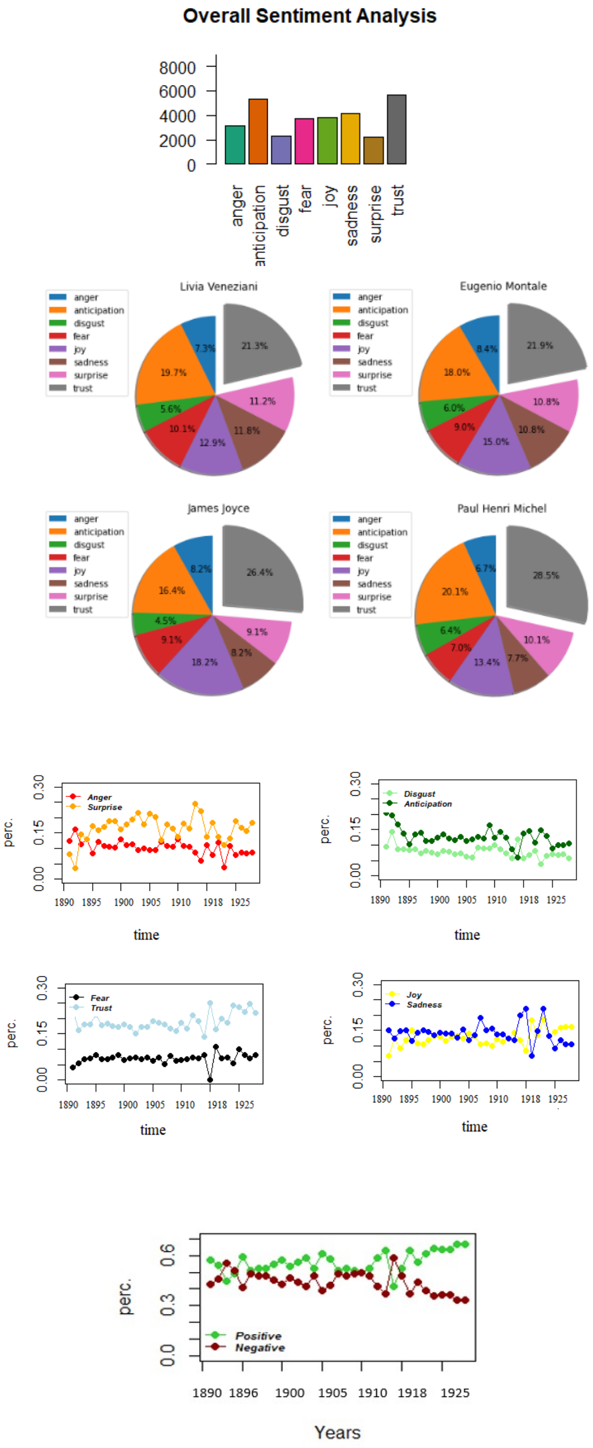

Figure 3:

References

[1] Raffini, Daniel. "Italo Svevo: Vita e Opere." Studenti.it, www.studenti.it/italo-svevo-vita-opere.html.

[2] Doll, Tyler. "LDA Topic Modeling." Medium, Towards Data Science, 11 Mar. 2019, towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd.

[3] Saif, Mohammad. NRC Emotion Lexicon, saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

[4] Kapadia, Shashank. "Evaluate Topic Models: Latent Dirichlet Allocation (LDA)." Medium, Towards Data Science, 29 Dec. 2020, towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0.

[5] "Jaccard Index." Wikipedia, Wikimedia Foundation, 30 Jan. 2021, en.wikipedia.org/wiki/Jaccard_index.

[6] Singh, Shubham. "How To Remove Stopwords In Python: Stemming and Lemmatization." Analytics Vidhya, 23 Dec. 2020, www.analyticsvidhya.com/blog/2019/08/how-to-remove-stopwords-text-normalization-nltk-spacy-gensim-python/.

[7] Ganesan, Kavita. "What Are N-Grams?" Kavita Ganesan, Ph.D, 30 July 2020, kavita-ganesan.com/what-are-n-grams/#.YCrP_Ooo8p8.

[8] "Chapter 2, Section 2: Stemming and Lemmatization." Introduction to Information Retrieval, by Christopher D. Manning et al., Cambridge University Press, 2018. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[9] "POS Tags and Part-of-Speech Tagging." Sketch Engine, 25 Sept. 2020, www.sketchengine.eu/blog/pos-tags/.

[10] Huilgol, Purva. "BoW Model and TF-IDF For Creating Feature From Text." Analytics Vidhya,

23 Dec. 2020, www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/.

[11] Maier, Daniel & Niekler, Andreas & Wiedemann, Gregor & Stoltenberg, Daniela. (2019). How document sampling and vocabulary pruning affect the results of topic models. 10.31219/osf.io/2rh6g.

[12] Prabhakaran, Selva. "Topic Modeling Visualization - How to Present Results of LDA Model?" ML+, 17 Sept. 2020, www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/.