Machine Learning project: Analysis of Svevo's letters corpus

https://github.com/Francesc0rtu/Svevo-corpus-analysis

Erika Lena[1], Bernardo Manfriani[1], and Francesco Ortu [1]

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

Course of AA 2021/2022 - Introduction to Machine Learning

1 Problem statement

The current work aims to investigate a corpus of letters of the Italian writer Italo Svevo. The letters are written in Italian, English, French and German, between the years 1895 and 1928. The following analysis has been carried out through the use of different text mining tools, such as Topic Modeling and Sentiment Analysis. In order to give a formal description of the problem, we need to define some basic definitions which will be broadly used: a *word* is an item of the dictionary $V = \{w_1 \dots w_n\}$, a *document* D is a collection of words, a *corpus* C is a collection of documents and a *topic* t is a subset of words. A *sentiment* is a categorical variable, initially simply defined as $s \in S = \{positive, negative, neutral\}$ and then extended to $S = \{joy, fear, anger, sadness\}$.
In the following table, we summarized different problems which have been addressed:

P	**input**	**output**
main topics	C	T
topic-person	$(T, p) \subset T \times P$	$\{\alpha_p(t) \in [0,1] : t \in T\}$ s.t. $\sum_{t \in T} \alpha_p(t)$
topic-time	$(T, \lambda) \subset T \times \Lambda$	$\{\beta_\lambda(t) \in [0,1] : t \in T\}$ s.t. $\sum_{t \in T} \beta_\lambda(t)$
sentiment-person	$(S, p) \subset S \times P$	$\{\gamma_p(s) \in [0,1] : s \in S\}$ s.t. $\sum_{s \in S} \gamma_p(s)$
sentiment-time	$(S, \lambda) \subset S \times \Lambda$	$\{\phi_\lambda(s) \in [0,1] : s \in S\}$ s.t. $\sum_{s \in S} \phi_\lambda(s)$
sentiment-topics	$(S, t) \subset S \times T$	$\{\psi_t(s) \in [0,1] : s \in S\}$ s.t. $\sum_{s \in S} \psi_t(s)$

where C is the corpus of documents, T and S are the set of topics and sentiments respectively, P is the set of persons.
The values returned as output are defined as follows:

- $\alpha_p(t)$ is the percentage of topic t in the correspondence with p
- $\beta_\lambda(t)$ is the percentage of topic t in the letters exchanged in the year λ
- $\gamma_p(s)$ is the percentage of sentiment s in the correspondence with p
- $\phi_\lambda(s)$ is the percentage of sentiment s in the letters exchanged in the year λ
- $\psi_t(s)$ is the percentage of sentiment s in the topic t

2 Assessment and performance indexes

The approach used to address each task is unsupervised, for this reason it is particularly difficult to establish performance metrics and indexes. If for sentiment analysis we could rely on predefined lists of opinion words and pre-trained tools, for topic modeling we needed a way to evaluate the text categorization obtained through the application of LDA[1]. A fundamental parameter to be tuned is the number of topics k; to properly assign it, we made use of three performance indexes:

- *perplexity*: a statistical measure of how well a probability model predicts a sample.
- *coherence*: measures the degree of semantic similarity between high scoring words in the topic.
- *silhouette score*: measures how similar a object is to its own cluster compared to other clusters.

The perplexity score measures the goodness of the fit. However, it is shown that this index often do not reflect human judgment[2]. For this reason we considered also the coherence index. In order to have a measure for cluster validation, we used the silhouette score, which is highly useful since it considers both cluster cohesion and separation [3]. The silhouette score is calculated between topics using the Hellinger distance. Through the use of these three scores, it was possible to find a suitable value for k, the number of topics.

3 Proposed solution

A first important step to address any possible solution is the pre-processing of the data. We start by removing punctuation, stop-words and upper-cases, then we apply lemmatization; all these steps are performed with respect to the main language used in each of the letters. To address the topic modeling tasks we start by using LDA and we find a suitable number of topics k, as discussed in paragraph 2. A visual inspection of the topics is carried out to check for how much the selected topics are human understandable. To inspect how the topics are structured, we compute the distances between pairs of topics using Hellinger distance and then, we use hierarchical clustering with average linkage to group them. Clustering allows to group nearest and overlapping topics, a key point though is in deciding the number of clusters to be used. We do not have a way to assess this parameter, except visual inspection. Main topics can then be chosen by ordering clusters with respect to their prevalence in the whole corpus.

Once we have chosen the topics, we group documents by sender-receiver pairs to analyze the persons each topic is most associated to and we use document-topic probabilities, summed up with respect to each person. We made the same thing grouping by year to analyze topics trend over time. To handle sentiments in each letter we start by using a predefined list of words, labeled as *positive, negative* or *neutral* with a corresponding level of polarity [4]. To each document

is assigned a percentage for each of these three sentiments by considering how each of its words is labeled. We repeated the same steps performed for the topics, grouping documents by person and by year and summing up the values associated to each sentiment. To further analyze the sentiments expressed in each letter, we also use a pre-trained model[5], which takes in input a sentence and returns a categorical value among $\{joy, fear, sadness, anger\}$, to classify the emotions expressed in the corpus.

4 Experimental evaluation

4.1 Data

The dataset used is made of 894 letters, written to and by Italo Svevo between the years 1895 and 1928. They are written in four different languages divided as follows with respect to the main language used: 826 use Italian, 30 French, 28 German, 10 English. Letters are exchanged with 45 different persons, the correspondence is particularly intense (>15 letters) with the following ones: 639 to Livia, 62 to Montale, 30 to Comnène, 19 to Crémieux, 19 to Joyce and 17 Henri Michel. From this preliminary analysis we can already see how the dataset is highly unbalanced towards Italian letters class, which represents the 92.39% of the whole corpus. And the same happens with regard to the persons to which the letters are addressed: the correspondence with Livia is the 71.47% of the total.

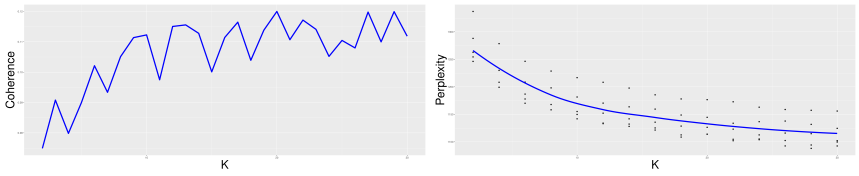


Figure 1: Plots of performance indexes: high coherence is better as well as low perplexity. The value of the silhouette score is almost constant for all $k > 1$.

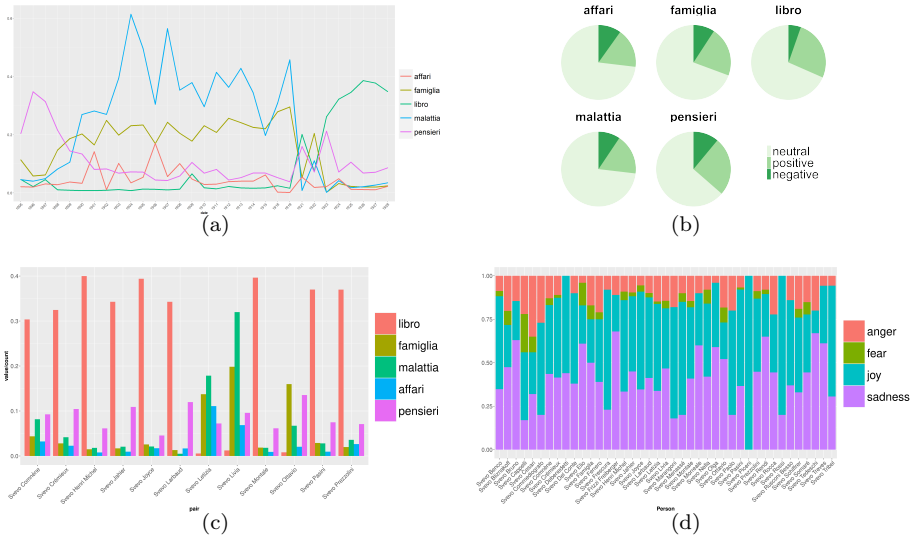
4.2 Procedure

Since the text mining methods we used are language dependent and the prevalent language in the corpus is the Italian, we decided to use only those letters which have Italian as main language. We believe this choice to do not have a negative impact, since the Italian corpus is about 92% of the total. We applied LDA[6] to our pre-processed corpus, and we compared previously discussed indexes to find a proper number of topics. This analysis seemed to suggest a value for k near 10[Fig. 1]; however, using 10 or an higher value for k resulted in some overlapping in the main words associated to different topics. By clustering them, we manually detected that a suitable number of clusters could be approximately 5, in order to achieve a better human understandability. A further check on the selected topics was done manually and the main ones were labeled: *"libro",*

"famiglia", "malattia", "affari", "pensieri". Other considerations on the reliability of selected topics were done by observing how they relates with persons and observing their trend over time along with relevant events in Svevo's life; same reasoning was applied also for sentiments evaluation.

4.3 Results and discussion

We obtained five main topics, the prevalent one is *"famiglia"* which appears to be discussed mainly with Livia, Letizia and Ottavio, three members of Svevo's family. Moreover, this result is consistent with the unbalanceness detected in the corpus and highly number of letters sent to and received from Livia. We can see how the topics concerning Svevo writing work are mainly discussed with contemporary authors, such as Montale, Joyce and Prezzolini. Even the trend of topics over time reveals something interesting, in particular with respect to what concerns the topic *"libro"*, which is highly present since the late 1910s, when he began to write his masterpiece *La coscienza di Zeno*. A similar analysis for the relationships between sentiments and persons can be inferred from charts, as well as their trend over time and their correlations with main topics. Even if results seem to be reasonable, there is much room for improvement. What has been done was a preliminary investigation of main concepts contained in the corpus. Further analysis could involve a more fine-grained selection of the documents, considering also other languages even if under-represented and a way to handle unbalanceness.



(a) (b) (c) (d)

Figure 2: (a) Topics trend over time, (b) percentage of sentiments assign to each topic, (c) topics associated to each sender-receiver pair, (d) emotions associated to each sender-receiver pair.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[3] Mika V. Mantyla, Maelick Claes, and Umar Farooq. Measuring lda topic stability from clusters of replicated runs. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '18, New York, NY, USA, 2018. Association for Computing Machinery.

[4] Ruben Izquierdo. Public-sentiment-lexicons. https://github.com/opener-project/public-sentiment-lexicons, 2014.

[5] 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.

[6] Bettina Grün and Kurt Hornik. topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 2011.