Machine Learning and Data Mining project: Analysis of Svevo's letters corpus

Valeria Insogna[1], Roberta Pascale[1], and Cecilia Zagni[1]

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

Course of AA 21/22 - Introduction to Machine Learning

1 Problem statement

This project aims to analyze the epistolary of the italian writer Italo Svevo, in order to extract information regarding the **main topics** and the **predominant sentiments** in the letters. Text mining techniques, in particular topic modeling (unsupervised learning) and sentiment analysis (supervised learning), were applied to identify, respectively, topics and sentiments. Further analysis on the text processing results were carried out to highlight links existing among topics, persons and sentiments, and to appreciate their evolution in time.

2 Data overview

The dataset under analysis is public and it consists in a corpus of 894 letters written or received by Italo Svevo in the years going from 1885 to 1928. The corpus is multilingual (Tab. 1), even at single letter level. The correspondents exchanging the letters are around 50. Basic statistics on the dataset (Tab. 1) show that Italian is the predominant language in the corpus and that most of the letters are addressed to Livia Veneziani, Svevo's wife. The last two aspects highlight the dataset unbalance, together with the wide range of letters length which varies between 2000 and 30 words, with an average of about 300. Beside the letters content, the recipients and languages, the dataset includes for each letter information about the date and the location of sender and receiver, for a total of 12 independent variables. There are no missing data.

3 Data preprocessing

Preprocessing techniques have been applied in order to structure the data in a suitable format to conduct patterns analysis, and hence, topic modeling. First,

Italian	826	Livia Veneziani	639
French	30	Eugenio Montale	62
German	28	Marieanne Comnène	30
English	10	James Joyce	19
		Others	144

Table 1: Corpus main languages and recipients

the corpus was restricted to letters written mainly in Italian, being the other languages underrepresented (less than 8%) and thus, unlikely to provide relevant information. Punctuation and non-alphanumeric words were removed. Later, tokenization, POS-tagging and lemmatization (via UDPipe R package [9]) were applied to the corpus. After these steps, lowercase transformation of the text was carried out. The Lexicon was reduced to only nouns, verbs and proper nouns and italian stop words were removed (including custom recurring words like months/days of the week). Lastly, the Lexicon was filtered to keep lemmas occurring in more than 0.5% and less than 5% of the corpus letters. The last two steps were fundamental in the definition of a dictionary clear of most of the outliers, greeting forms and recurring recipients names, which could have influenced negatively the topic modeling.

Except for the restriction to only italian letters, no preprocessing step was performed for sentiment analysis. This choice was made to compare the results from the two text mining techniques on the same corpus of letters.

4 Proposed solution

The estimation of topics in the corpus was performed adopting one of the most effective topic modeling techniques, Latent Dirichlet Allocation (LDA) [1]. The model was trained using up to 300 collapsed Gibbs sampling [4] from the textmineR package [6] in order to achieve maximum log-likelihood stability. Hyperparameters tuning against model coherence score maximization was performed to select the optimal combination of LDA input parameters (number of topics, α and β). Coherence score was chosen as index of model goodness being a measure of topics quality that showed to be highly correlated with human interpretability, often considered the gold standard for topic quality [7]. After experimental attempts, a baseline for models to be trained was set up letting the number of topics range between 3 and 10, and varying the symmetric parameters for both priors. The choice of an asymmetric prior for the document-topic distribution was discarded because of the lack of previous knowledge of the corpus content and therefore, of the possibility to make an informed choice for the values of the α. Hyperparameters tuning (Fig. 1) showed that coherence tends to decrease increasing the number of topics.

For the sentiment analysis, because of the lack of learning data to build a Svevo vocabulary, a general pre-trained model was selected: NRC Word-

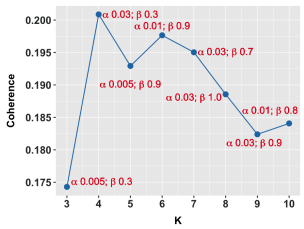


Figure 1: Coherence score varying k number of topics. Each point is the model with the highest coherence score resulting from the hyperparameters tuning for a given k.

Emotion Association Lexicon [8] implemented in Syuzhet R package [5]. This tool extracts sentiment scores for eight emotions relying on a crowd-sourced lexicon of around 14000 lemmas associated with semantic areas. Applying it, the count of occurrences of words associated to each sentiment per letter was obtained. To avoid biased scores due to the varying letter lengths, the counts were converted in relative frequencies dividing each sentiment observation by the sum of sentiments occurrences in the letter.

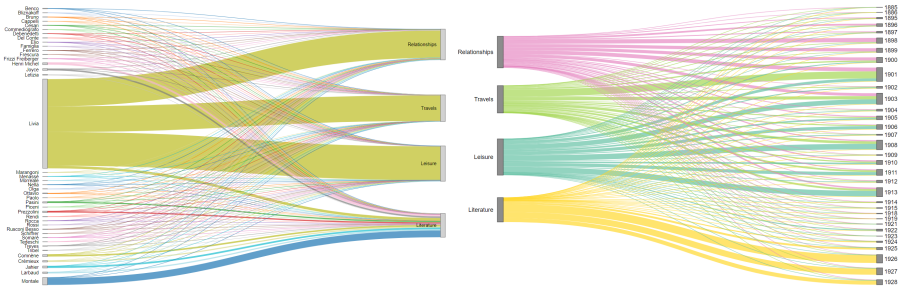
5 Experimental assessment

To select a model for topic estimation, visual inspection and interpretation of the topics keywords of the models with the highest coherence score was performed. The 4-topic model was chosen and the topics were labelled as: **Literature** (main driving words "traduzione, critico, letteratura, pubblicazione"), **Travels** ("inghilterra, marsiglia, battello, stazione"), **Leisure** ("comperare, bere, vestito") and **Relationships** ("gelosia, amore, figlia, sacrificio"). To assess the pertinence of the model, the distribution of topics in time (Fig.2b) with all the meaningful events in the author's life [3] were compared.

For the sentiment analysis, because of the adoption of an off-the-shelf pretrained tool, classical effectiveness indexes for classification problems could not be used. The goodness of the solution was evaluated via human judgment, analysing the results obtained for the letters associated with higher sentiment scores for each emotion. The accuracy of the lexicon was good in more than 80% of the times (which is in agreement with the baseline assumed in similar studies [10] [2]). However, it was noticed that some words were misclassified as negative because of the usage of irony (for example when Svevo refers to his wife with the term "capra") or because of the different stigma that words had in Svevo's epoch.

6 Results and discussion

Analyzing the main topics through the years (Fig.2b), is evident that "Literature" was dominant in the years of Svevo's sucess, after the publication of "La coscienza di Zeno" in 1926. The "Travels" topic is, instead, more related to Svevo's intense years of business trips from 1901, after joining his father-in-law's activity. The "Relationships" topic is mostly related to his wife Livia (Fig.2a),



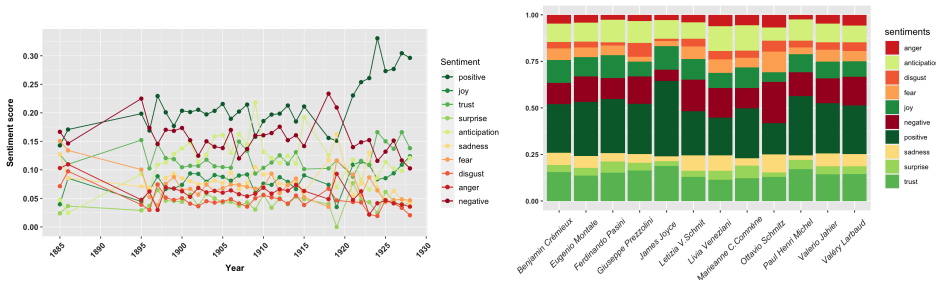
(a) Person-topic association

(b) Year-topic association

Figure 2: Topic modeling highlights

whilst the "Leisure" topic was strongly present in the years of travels before the author committed himself completely to writing and before his literary success. It stands out that Livia is the person to which all topics are more connected with, except for the "Literature", which is more linked with other famous writers such as Montale and Joyce. The associations are sound, both respect to the recipients professions both respect to the percentage of letters exchanged with Livia, which is the highest.

The sentiments evolution over the years (Fig.3a) shows that the peak of positivity coincides with the last years of Svevo's life, characterized by his work success. On the other hand, the most negative sentiments coincide with the death of his relatives, in particular his mother in 1985. The analysis of the combination between sentiment scores and topics confirms this evaluation, since "Literature" is the most positive topic, while "Relationships" is the most negative one. Also the association between the emotions and Svevo's correspondents (Fig.3b) confirms this pattern: the letters addressed to other authors like Joyce and Paul H. Micheal are generally more positive than those addressed to his family, in particular to his brother Ottavio.



(a) Sentiments scores evolution in time

(b) Sentiment scores of top 12 recipients

Figure 3: Sentiment analysis highlights

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] Eric Chen, Zhiyun Lu, Hao Xu, Liangliang Cao, Yu Zhang, and James Fan. A large scale speech sentiment corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6549–6555, Marseille, France, May 2020. European Language Resources Association.

[3] Vita di Italo Svevo. Museo sveviano di trieste. https://www.museosveviano.it/italo-svevo/.

[4] Chase Geigle. Inference methods for latent dirichlet allocation. Technical report, University of Illinois at Urbana-Champaign, 2017.

[5] Matthew Jockers. Package 'syuzhet'. *URL: https://cran. r-project. org/web/packages/syuzhet*, 2017.

[6] T Jones. textminer: Functions for text mining and topic modeling. https://CRAN.R-project.org/package=textmineR, 2021.

[7] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[8] Saif Mohammad and Peter Turney. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA, June 2010. Association for Computational Linguistics.

[9] Jan Wijffels. Udpipe natural language processing - text annotation. https://cran.r-project.org/web/packages/udpipe/index.html, 2021.

[10] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.