DEPARTMENT OF MATHEMATICS AND GEOSCIENCE, UNIVERSITY OF TRIESTE

Topic Modeling and Sentiment Analysis of Italo Svevo Epistolary Corpus

Gabriele Sarti, with Eric Medvet (Machine Learning Lab UniTS) and Cristina Fenu (Hortis Library of Trieste)

Aim

In this work, I apply natural language processing techniques on the multilingual epistolary corpus of Italo Svevo, one of the great Italian novelists of the 20th century, in order to gain insights about topics and sentiment expressed in his letters. More specifically, this proposal aims to extract new information from the available corpus, highlighting relations between topics, individuals and emotions and exploring how those connections evolve through time.

Dataset and Challenges

The Svevo letter corpus dataset was compiled in 2017 by C. Fenu for her submission at the 6th AIUCD conference. It contains a total of **894 letters** written by Italo Svevo in more than **four languages** (mainly Italian, French, English, German and Trieste's dialect) between 1885 and 1928.

In addition to letter bodies, the dataset contains dates of sending, names of addressees, locations of sending and reception, sub-corpus belonging and languages used for a total of 12 variables for each observation.

The main challenges faced throughout the analysis were the **sparse multilinguality** and the **implicit unbalancedness** of the corpus: 826 out of 894 letters are mainly in Italian and 639 of them are addressed to Svevo's wife.

Italian → 826 letters **German** → 28 letters
French → 30 letters **English** → 10 letters

A. Main languages used inside the corpus

Livia Veneziani → 639 letters
Eugenio Montale → 62 letters
Marianne Comnène → 30 letters
Benjamin Crémieux → 19 letters
James Joyce → 19 letters
Others → 89 letters

B. Most frequent addressees inside the corpus

Preprocessing

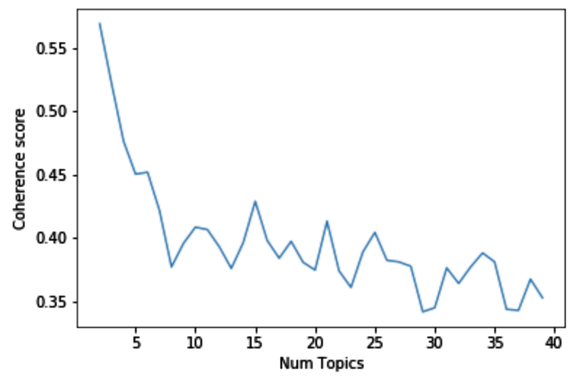
The following preprocessing steps were taken for the topic modeling part:

1. **Only the Italian corpus was considered** since the other corpora didn't contain enough letters to add meaningful information.
2. Letters were **tokenized** and converted to **lowercase**.
3. **Punctuation**, italian **stop-words** and **non-alpha** words were removed.
4. POS-tagging was used to keep **only nouns and verbs**.
5. Remaining tokens were **lemmatized** and added to the dictionary.
6. Dictionary was filtered removing all tokens recurring in **less than 5 letters and in more than 5%** of the total corpus to remove greetings and too common/rare expressions.

All preprocessing steps were performed using *spaCy* lexicons. No preprocessing steps were taken for the sentiment analysis part.

Proposed Approach

For the topic modeling part, after trying advanced approaches such as lda2vec and two-steps LDA I settled for *gensim*'s **Latent Dirichlet Allocation** with dynamically-computed asymmetric priors. The number of passes through the whole corpus in order to train the LDA model was set to 200 to ensure consistency in the results. The **silhouette index** and the **extrinsic UCI coherence** scores were used as indicators to choose an appropriate number of topics for the LDA model.



C. Coherence score between 2 and 40 LDA topics.

For sentiment analysis I used the **NRC Word-Emotion Association Lexicon (EmoLex)** implemented in the R package *syuzhet*. The approach was fitting since EmoLex contains sentiment scores for eight base emotions associated with more than 14 000 lemmas in more than 100 languages. Scores for each sentiment of a letter were normalized in order to avoid unbalances between letters of different lengths.

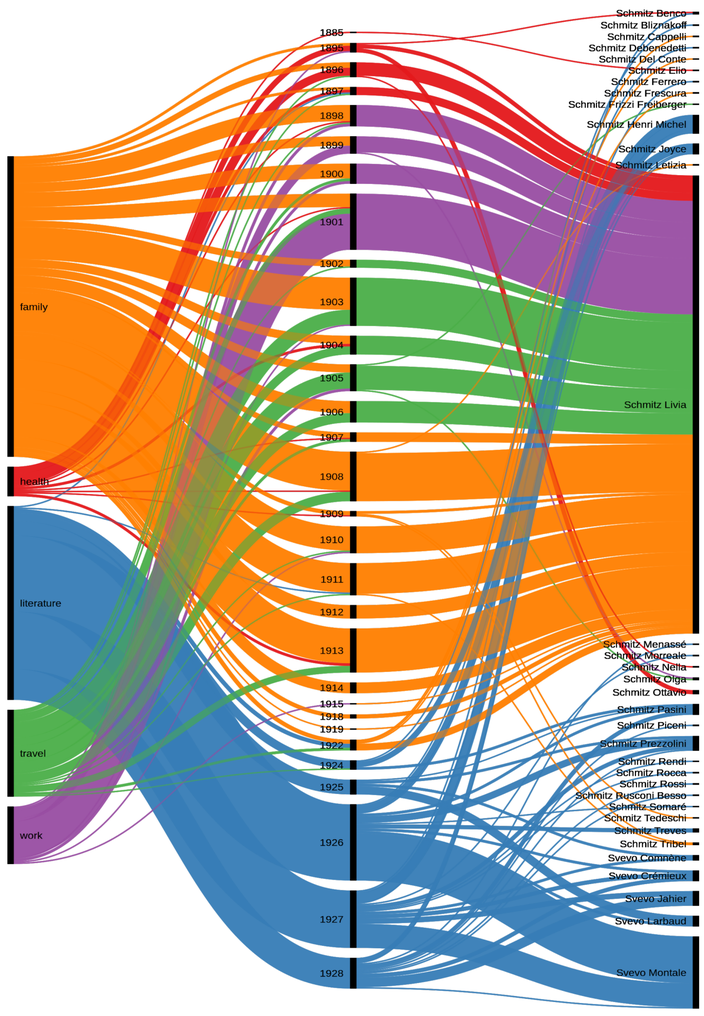
Evaluation

After shrinking the model research scope between two and six topics, each LDA model was evaluated by the pertinence and interpretability of its keywords, randomly sampling five texts presenting high scores for different topics of each model to assess accuracy. After the evaluation I finally opted for the **five-topic LDA model,** which was subsequently used to compare topics distribution over time with Svevo's life timeline.

In order to validate sentiment distribution over the corpus I inspected random samples of letters having high sentiment scores to assess the validity of the lexicon, especially for non-English lemmas. I found that scores were decent given the context but many letters containing **irony**, a staple in Svevo literary style, were misinterpreted as positive. Also, many terms commonly associated to a negative sentiment (e.g. "cigarette") are positive in the Svevian universe, requiring a **fine-tuned lexicon based on the author** which is currently under construction. Sentiment results were also validated against Svevo's life timeline.

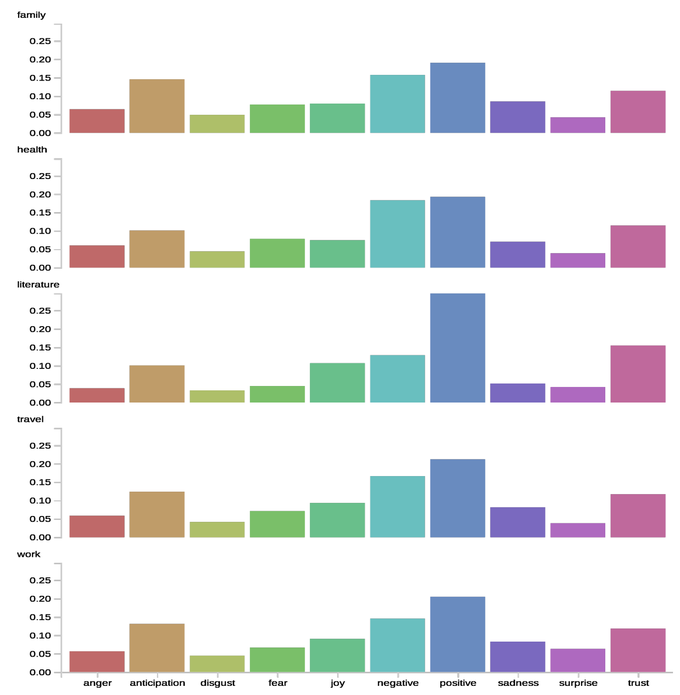
Results and Discussion

ICharacterizing words for each topic were used to determine an interpretable label for each of them. Final choices were ***family, work, travel, health*** and ***literature***.



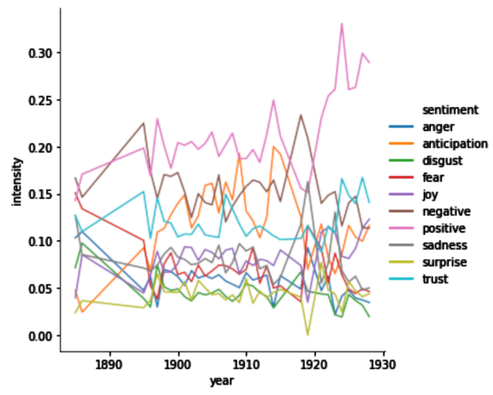
D. Topic relations with addressees over time.

We used an alluvional diagram to visualize the distribution of topics over time and their relation to addressees: in the left side of the diagram the choice of color represents different topics, while in the right side it highlights different phases of Svevo's life. It is interesting to note the correspondence between the former and the latter. We also analyzed the relation between prevalent topics and sentiment.



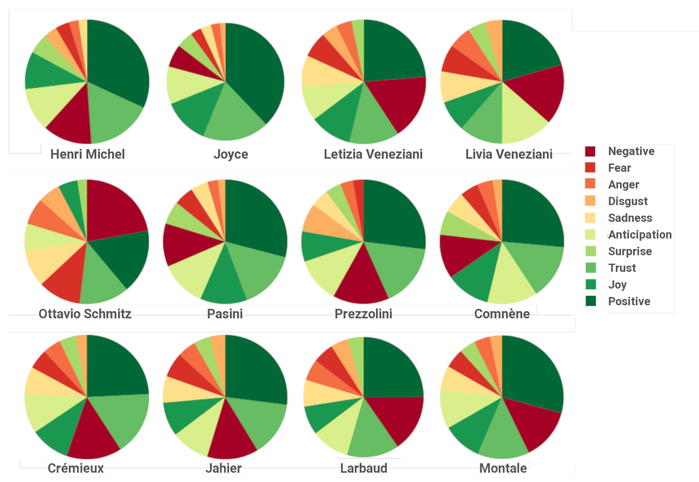
E. Topics in relation to sentiment scores

Concerning sentiment distribution, figure F shows how peaks of negative sentiment coincide with the death of Svevo's relatives, while his years of fame are marked by a high spike in positivity. These finding are consistent with our matching between sentiment scores and topics found in letters, since "literature" is the most positive topic, while "health" is the most negative one.



F. Sentiment scores progression over time.

Finally, it is interesting to note from figure G how the interactions between Svevo and other authors such as Joyce and Montale were generally more positive than those with Svevo's relatives. This can be partially explained by the unbalance in letter quantities for those categories and the periods in which those letters were written, which match with my aforementioned findings.



G. Sentiment scores in relation to addressees.

Additional Info and Sources

An in-depth overview of the project, complete with code, bibliography and additional visualizations is available on Github at the following link:

github.com/gsarti/svevo-letters-analysis