Machine Learning and Data Mining Project: Analysis of Svevo's Letter Corpus

Gabriele Sarti

Academic Year 2018-2019

1 Problem statement

Natural language processing has been a very fertile field for innovation in recent years. While most of the research work around language focuses on modern applications such as social networks, past literary works are seldom analyzed despite having a great potential in giving additional clues about authors and their productions.

In this work, I apply some natural language processing techniques on the multilingual epistolary corpus of Italo Svevo, one of the great Italian novelists of the twentieth century and a pioneer of the psychological novel in Italy, in order to gain insights about topics and emotions contained in his letters. While previous works[6] focused on visualizing the relations between individuals in the corpus, this proposal is peculiar since it aims to extract new information from the existing corpus, highlighting relations between topics, individuals and emotions and exploring how those relations evolve through time. An exhaustive overview of the project, with code included, is available on GitHub[12].

2 Data and assessment

The Svevo letter corpus dataset was created in 2017 by C. Fenu from an original corpus compiled in 1966[6]. It contains a total of 894 letters in more than four languages, including dialects. In addition to letter texts, the dataset contains information about dates in which letters were sent, the names of senders and receivers, their locations and the languages used throughout the letter, for a total of 12 variables for each observation.
The main challenges of analyzing the Svevo letter corpus are the sparse presence of multiple languages and the implicit unbalancedness of the corpus, since 826 out of 894 letters are written in Italian and 639 between them are sent by or addressed to Svevo's wife. These aspects were crucial for many design choices I had to take in order to make the project viable with limited time and resources.

3 Preprocessing

For the topic modeling part I decided to consider exclusively the Italian letter corpus since the French, German and English letters were not enough to provide meaningful results and were hardly translatable. The preprocessing pipeline for this section was structured as follows: firstly, I tokenized the texts and converted all tokens to lowercase; secondly, I removed punctuation, stop-words and non-alpha words from the tokens; thirdly, I used part-of-speech tagging[1] to keep only nouns and verb and finally, I performed the lemmatization of the remaining tokens. All those steps were taken in order to maximize the amount of information about topics while minimizing the size of the dictionary created from those tokens. A crucial step was to perform an additional filtering of the dictionary, removing all words recurring in less than 5 letters or more than the 5% of the total corpus to remove outliers, especially greetings and expressions without any real value for this analysis.

For the sentiment analysis part, since I was able to exploit multilingual lexicons, the original corpus was considered in its totality, without applying any kind of preprocessing.

4 Proposed solution

To perform the topic modeling, I opted for a probabilistic approach in the form of a latent Dirichlet allocation[2][3] model since it is one of the most effective approaches for this task. The model used was the one contained in the gensim library[10], with the peculiar characteristic of being trained by passing through the whole Italian corpus 200 times in order to make the topics division more precise.

4.1 Performance indexes for topic modeling

In order to choose the appropriate number of topics for my model, I trained models on a range between 2 and 40 topics and computed the silhouette index[11] and the extrinsic UCI coherence score[9] for each model. By doing so, I obtained the results shown in the Figures 1 and 2, which led me to reduce my research scope between two and six topics, where the values of those indices are acceptable.

4.2 Sentiment analysis

I decided to use the NRC Word-Emotion Association Lexicon[8] implemented in the Syuzhet Package[7] because it is able to encompass the four main languages included in the corpus, extracting sentiment scores for eight base emotions using more than 14'000 lemmas associated with semantic areas. With this approach, each letter in the corpus has n points for each emotion, with n being the number of words associated to that emotion that are present in the letter. I converted

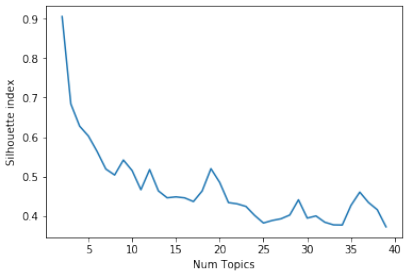

Figure 1: Average silhouette score

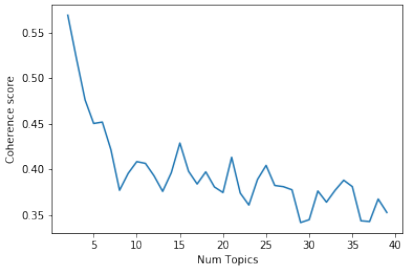

Figure 2: Extrinsic UCI coherence score

these scores to percentage scores by dividing them for the total sentiment score of their respective letters, in order to avoid unbalances between long and short letters. A dataset with all the percentage sentiment score for each letter was generated and then joined to the original dataset to extrapolate and visualize our findings.

5 Evaluation procedure

For the topic modeling part, I decided not to go with the choice of two topics, which was the most precise division according to indices, since I wanted to discover more niche themes, and decided to follow an empirical procedure instead. I evaluated each model with six or less topics by the pertinence and interpretability of its keywords, and randomly sampled five texts presenting high scores for different topics of each model in order to assess the modeling accuracy. The final choice derived from this judgments was to use five topics for the modeling task. We finally evaluated the five-topics model by comparing the distribution in time of the topics with a timeline containing all meaningful events in the life of Svevo[5]. This procedure confirmed the pertinence of the design choices taken, since it produces an evident relation between topics in the letters and events in author's life.

For the sentiment analysis, my evaluation approach was to inspect random samples of five letters having high percentages scores of a specific sentiment to validate the accuracy of the lexicon, especially for languages other than English. I found that the scores were decent given the context, but many ironical texts were misinterpreted. Finally, I assessed the overall validity of the sentiment analysis by grouping letters by authors, year of sending and prevalent topic, averaging the sentiment score for those groups and comparing the results obtained this way with those presented by the Svevo Museum[4] and the Svevo timeline[5]. The sentiment obtained in consistent with previous finding and with author's life in general.

6 Results and discussion

I named the five topics extracted by our LDA model as "family", "work", "travel", "health" and "literature" using their characteristic words to extrapolate a possible name. The "family" topic, using words as "cuore, Livia, Letizia, Olga, abbracciare" is the largest one, spanning through Svevo entire life, and it is mostly related to his wife Livia, other members of the family and close friends. The "work" topic, using words as "fabbricare, operai, lavorare" focuses between 1898 and 1901, the period in which Svevo quits his job at a bank and starts working for his father-in-law, and is mostly related with his wife. The "travel" topic, using words as "viaggiare, Londra, Trieste, addio, ritornare" is also mostly related with Svevo wife and spans the period between 1900 and 1908 in which he had to travel a lot for his work. The "health" topic, using words as "dottore, dolore, curare, febbre" characterizes the years between 1885 and 1897, in which the father, the mother and a brother of the author die for various illnesses and is mostly related to his wife and to his brothers Ottavio and Elio. It is also the topic with the lowest positive score and highest negative score. Finally, the "literature" topic, using words as "senilità, Joyce, Zeno, romanzare, scrivere" is the most positive one, characterizing the golden years after the publication of the novel "La coscienza di Zeno" in 1923, in which his novels become internationally acclaimed and his set of interlocutors broadens. It is mostly related to the authors Eugenio Montale and James Joyce, and to critics Larbaud and Crémieux.

For the sentiment relations with individuals and years, Figures 3 and 4 a part of my findings. Most negative period coincides with the death of Svevo's relatives, while his years of fame are marked by a high spike in positivity. Also, it is interesting to note how the letters between Svevo and other authors such as Joyce and Montale are generally more positive than those between Svevo and his relatives. The letters with his brother Ottavio are particularly negative since they were sent shortly after the death of their mother. More complete visualizations for topics and sentiment are available on GitHub[12].

Figure 3: Sentiment evolution by year

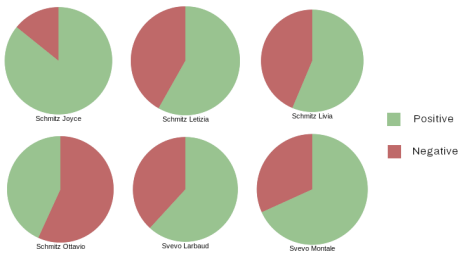

Figure 4: Sentiment by individual

References

[1] Explosion AI. spaCy POS tagging. https://spacy.io/usage/linguistic-features#section-pos-tagging, 2018. [Online; accessed 21 January 2019].

[2] Andrew Y.; Jordan Michael I Blei, David M.; Ng. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

[3] David M. Blei. Probabilistic Topic Models. Surveying a suite of algorithms that offer a solution to managing large document archives. *Magazine Communications of the ACM*, 2012.

[4] Museo Sveviano di Trieste. Museo Sveviano Digital Archive. http://www.museosveviano.it/ar/progetto/archivio-digitale/. [Online; accessed 21 January 2019].

[5] Museo Sveviano di Trieste. Svevo Timeline. http://www.museosveviano.it/ar/italo-svevo-la-vita/. [Online; accessed 21 January 2019].

[6] C. Fenu. Sentiment Analysis d'autore: l'epistolario di Italo Svevo. In *Proceedings of 2017 AIUCD 6th Conference on "The educational impacts of DSE"*, pages 149–155, 2017.

[7] Matthew L. Jockers. Syuzhet Package. https://github.com/mjockers/syuzhet, 2015.

[8] Saif M. Mohammad and Peter D. Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACLHLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, 2010.

[9] Quentin Pleplé. Topic Coherence to Evaluate Topic Models. http://qpleple.com/topic-coherence-to-evaluate-topic-models/, 2013. [Online; accessed 5 January 2019].

[10] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010. http://is.muni.cz/publication/884893/en.

[11] Peter J. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*, pages 53–65, 1987.

[12] Gabriele Sarti. Svevo letters analysis. https://github.com/gsarti/svevo-letters-analysis, 2019.