A Software Pipeline Based on Sentiment Analysis to Analyze Narrative Medicine Texts

Ileana Scarpino, Chiara Zucco, Mario Cannataro

Data Analytics Research Center, Department of Medical and Surgical Sciences,
University "Magna Graecia" of Catanzaro, Italy
ileana.scarpino@studenti.unicz.it, chiara.zucco@unicz.it,
cannataro@unicz.it

Abstract. By using social media people can exchange sentiments and emotions, allowing to understand public opinion on specific issues. Sentiment Analysis (SA) is a novel text-mining (TM) and natural language processing (NLP) methodology to extract sentiment, opinions and emotions from written texts, usually provided through social media or questionnaires. Sharing medical and clinical experiences of patients through social media, is the target of the so-called Narrative Medicine (NM). Here we report some research experiences in applying SA techniques to analyze NM texts. A problem to be faced in NM is the automatic analysis of a potentially large set of documents. Application of SA is useful for having immediate analysis and extracting information from medical literature quickly. Here we present a software pipeline based on SA and TM which allows to effectively analyze NM texts. First experimental results allow to discover topics related to diseases.

Keywords: Sentiment Analysis · Text Mining · Topic Modeling · Narrative Medicine.

1 Introduction

Scientific research has shown that to make a targeted and personalized diagnostic process it is necessary to pay attention not only to the patient's illness, but also to his/her psychological and emotional state. The narration of patients and caregivers is an essential element of contemporary medicine and it is based on the active participation of the subjects involved. Through their stories, people become protagonists of their own healing process [1]. A medicine practiced with narrative competence, i.e. Narrative Medicine (NM), may be better able to recognize patients and diseases, empathize with colleagues, accompany patients and their families through the vicissitudes of the disease [2–5]. In particular, NM has many advantages: it improves clinical practice, allows a more in-depth diagnosis, promotes adherence to therapy, helps and consolidates choices, fosters relationships between patient, family and healthcare staff, improves the quality of service and the therapy strategy, verifies and allows feedback on the functionality of the therapy, promotes the formation of communities that help the patient on a social and psychological level.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_45

Although NM is a practice born in the late 1960s [6], it has been scarcely treated from a methodological point of view [7]. Only the oncological branch is projected towards the analysis of patients' narratives [8]. The methodologies that have characterized the analysis of written and oral narrative material can be grouped into three main strands: Thematic Analysis allows to count the frequency of the words and themes proposed by the patient [9]; Linguistic Analysis allows to differentiate the narratives by gender complexity [10]; Content analysis implements various procedures for quantitative survey of the narrative structure and its qualitative content [11, 12].

Consequently to the social media growth, people exchange opinions, feelings and emotions and they also share personal experiences related to their diseases, by giving birth to online blogs and forums for sharing their experience, as for instance the Italian blog "Viverla Tutta" [13]. Analyzing the information present on online health forums and communities may help patients by improving the diagnostic process on the basis of similar experiences.

It is known how clinical narratives contain a moderate amount of sentiment terms that reflect the objectivity and preciseness of the clinical writing style [14]. The interest towards the extraction of opinions and emotions from textual online resourced has led to the development of Sentiment Analysis (SA), that combines Text Mining (TM) and Natural Language Processing (NLP) and whose aim is the extraction of subjective information from texts, focusing not only on the topic, but also on the opinions expressed in the texts [15]. SA tools lead to a polarity analysis that can be positive, negative or neutral [16]. Text Mining techniques are widely used to perform biomedical knowledge extraction [17] to obtain relevant information from vast online databases of health science literature or patients' electronic health records.

Previous studies have shown how sentiment analysis applied to clinical documents has the potential for assisting patients with information for self assessing treatments, providing health professionals with more insights into patients' health conditions, or even managing relations between patients and doctors [18].

In this paper we present an application of several NLP and TM methods to examine various aspects of the coexistence between patients with their illness. The main contribution is the proposal of a semi-automatic software pipeline for narrative medicine, a domain in which automatic approaches for the analysis are still poor. The rest of the paper is organized as follows. Section 2 presents the proposed software pipeline. Section 3 presents the final results showing both clinical and predominantly emotional aspects. Finally Section 4 presents the conclusions and future work.

2 Narrative Medicine Analysis Pipeline

The proposed analysis pipeline includes two main stages, NLP methods for the preprocessing of the input NM texts, and TM methods to analyze them (see Figure 1). The analysis pipeline is implemented in *Python*.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_45

The experimentation of the pipeline has been performed by using textual sources extracted from the "Viverla tutta" data source, a blog dedicated to NM. A total of 12 texts, each one written by a patient, related to three diseases (fibromyalgia, cancer and diabetes) were selected. Then 4 patient testimonials were selected for each kind of disease, as shown in Table 1.

Table 1. Data files from the Italian blog "Viverla tutta" [13] are grouped by the common disease of patients.

Disease	File name	File dimension (Bytes)
Fibromyalgia	Patient_1.txt	2,138
	Patient_2.txt	2,962
	Patient_3.txt	3,892
	Patient_4.txt	2,111
Tumor	Patient_5.txt	1,177
	Patient_6.txt	2,710
	Patient_7.txt	2,705
	Patient_8.txt	2,327
Diabetes	Patient_9.txt	1,178
	Patient_10.txt	892
	Patient_11.txt	1,513
	Patient_12.txt	6,260

The main steps of the pipeline are described below. After loading and reading text files, we proceed with data preparation, which is particularly important in text mining, because it works on textual data that must be made suitable to the subsequent steps. Pre-processing consists of several stages, i.e. lowercasing, punctuation and special characters removal, tokenization, stemming, stopwords removal and Part-of-Speech tagging. In particular, the *Nltk* [1], the *SpaCy* [2] and the *gensim* [3] *Python* library for the Italian language have been used for text preprocessing.

Term Frequency - Inverse Document Frequency (TF-IDF) statistically quantifies how important a term is within a document in relation to other similar documents. It replaces number of occurrences with a weighted frequency value. TF-IDF algorithm associates greater importance to less frequent but more relevant terms

Topic modeling is a technique used in NLP. It is a statistical model that searches for topics from a corpus of texts. The topics are not known a priori but they are identified by the algorithm based on word frequency in the documents. In particular, Latent Dirichlet Allocation (LDA) is an unsupervised learning al-

[1] https://www.nltk.org/
[2] https://spacy.io
[3] https://radimrehurek.com/gensim/index.html

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_45

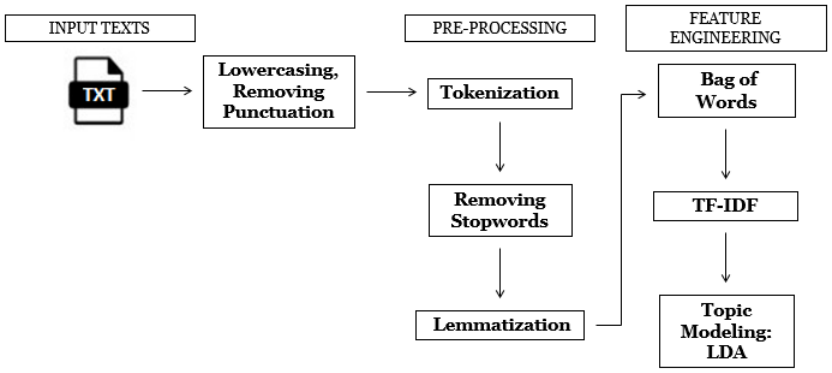

Fig. 1. The proposed pipeline for the analysis of NM texts.

Table 2. TF-IDF of terms in the analyzed dataset. Terms carrying more information are underlined. Underlined values have more weight in the document of belonging than all the others texts. Empty cells represent terms missing in some documents but present in others.

Disease	Terms	Patient 1	Patient 2	Patient 3	Patient 4
Fibromyalgia	Illness	0.1155	0.2652	0.1269	
	Syndrome			0.25377	
	Fibromyalgia	0.1155		0.29607	
	Rheumatoid arthritis		0.1326		
	Inflammation			0.0845	
	Pain	0.05774		0.25377	0.22829
	Sensibility	0.05774		0.08459	0.05707
	Life	0.11547	0.13258	0.0423	0.22829
		Patient 5	**Patient 6**	**Patient 7**	**Patient 8**
Tumor	Tumor	0.0811	0.0971	0.1918	0.0508
	Carcinoma	0.0811	0.0486	0.0479	0.0508
	Chemotherapy		0.0971	0.0959	0.1017
	Radiotherapy	0.0811		0.4795	
	Surgery		0.0486	0.0959	0.1525
	Relapses		0.0486		
	Defeat			0.0959	
	Life	0.0811	0.1943	0.2877	0.3558
		Patient 9	**Patient 10**	**Patient 11**	**Patient 12**
Diabetes	Illness			0.207	0.3105
	Diabetes		0.2572		0.3103
	Hyperglycemia		0.1715		0.0282
	Life		0.0857	0.138	0.2823
	Win	0.0814	0.0857		0.0282

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_45

gorithm that allows to identify a certain number of topics, called "latent topics" in a corpus of documents. LDA generates a set of words called "keywords", to which a weight is associated. A higher weight discriminates the most relevant topics, whose number needs to be set before training. In the present analysis the number of topics has been set equal to the number of documents to avoid algorithm generating repetitions or similar topics. Following the training of LDA, perplexity and coherence score have been calculated as evaluation metrics to assess LDA model's clarity. Perplexity is a well known intrinsic evaluation metric. Topic Coherence measures, score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference [19, 20].

3 Results

Two types of studies were conducted: first patient testimonials released from the blog "Viverla Tutta" were analyzed together by relating terms specific to each disease so that it was possible to test that the focus was specific for each disease. Subsequently, texts were analyzed one-to-one to verify if some words were more present in one document than other.

Table 2 summarizes the TF-IDF results. The underlined values refer to terms that are the most relevant in comparison with all the documents analyzed, and therefore they carry more information. The empty cells instead represent the missing terms in the documents and therefore words that are not written by a patient in a given document.

Analyzing Table 2 by disease it can be seen that each patient, despite being affected by the same pathology, lives illness in a personal way by adopting different terms for its definition. Some fibromyalgia patients talk about "illness", as for instance Patient 2. Patient 3 talks about "syndrome", terms that is absent in other documents. He/she describes physical symptoms as "numbness", "tingling", "photophobia", "intolerance", "sleep", rather than focusing on the consequences that disease had in his/her life. The word "pain" predominates in documents 3 and 4 while the terms present in all documents are those related to the patient's experience such as "living", "life", "past".

Every single patient of the considered subset "Tumor" defines his/her disease as "tumor" and talks about of "carcinoma". The terms present in all the sources are "live" and "life", showing the desire of dealing with the disease and conquer it. In addition, there is the possibility that some subjects have undergone a "surgery"; the word "relapse" occurred only in a document means that at least one patient has developed relapses.

With respect to diabetes patients, the term "diabetes" has an important frequency value in document 4 and it compares in document 2 in which "drink" and "water" are found, recalling that among the symptoms of diabetes there is "polydipsia". Other characteristic features of this disease are "blood", "sugar", "glucometer", "insulin".

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_45

LDA similarity was applied to compare how similar documents are in topics from their topic distributions. It has been allocated how much each document belongs to 1st, 2nd, 3rd or 4th topic, through numerical value indicating strength with which the text is related to a given topic. We were expecting that the main focus was the disease told by patients, but amount of verbs is also relevant. Observing carefully the LDA results, it is possible to find two types of results, in some topics a more clinical aspect and in others a more emotional one is found. Perplexity and coherence scores did not give surprising results, as shown in Table 3. By decreasing the number of topics analyzed, the probability that a document could be associated with multiple topics increases.

Table 3. Perplexity and Coherence score: intrinsic evaluation metrics of the model. The log perplexity results indicate poor human interpretable topics. Coherence scores represent a probability that of course vary in a $[0, 1]$ range. Therefore the overall results are not satisfying.

Disease	Perplexity	Coherence score
Fibromyalgia	-7.556	0.394
Tumor	-7.503	0.304
Diabetes	-7.580	0.367

4 Conclusions

A first contribution of the paper is the proposal of a software pipeline, based on different *Python* libraries, for the analysis of NM texts using NLP and TM techniques. The pipeline has been applied to analyze some NM texts provided by several patients affected by three pathologies: fibromyalgia, cancer and diabetes. Currently, the pipeline implementation is not publicly available, but it will be released to be public as a future development of the study. The presented information extraction process allowed to analyse various aspects of the coexistence of patients with their illness. The TF-IDF analysis showed that in patients with the same pathological condition, the narration of their experience was focused on discordant aspects that depend on personality traits which leads them to live their experience in a different way. On the other hand, there are also similar aspects between various patients though they are suffering from different pathologies. The identified relevant words vary from clinical to emotional topic.

From an LDA perspective, the unsatisfactory results should be improved in future works by considering a larger number of textual NM sources. Moreover, to overcome any limitations due to the use of articulated texts, the pre-processing may be improved, enhancing different machine learning techniques to improve performance.

The proposed pipeline can be tested and used in various application as for instance to help patients by improving the diagnostic treatment on the basis of

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_45

similar experiences. As stated before, the main limitation of the study is the small number of analyzed text. In future works, we plan to extend the dataset by adding more textual data and by also increasing the number of considered diseases. As a second future development, we are interested in evaluating whether a correlation between extracted topics and the disease severity may be assessed, and to also consider the sentiment related to NM texts.

References

1. Hurwitz, B.: Narrative [in] Medicine. In: Spinozzi, P., Hurwitz, B., Discourses in the Biosciences, Vol. 8 pp. 13–30. Vandenhoeck & Ruprecht Unipress, Gottingen (2011)
2. Evans, M.: Reflections on the humanities in medical education. Medical Education **36**(6), 508–13 (2002)
3. Charon, R.: Narrative Medicine: Honoring the Stories of Illness. Oxford University Press, (2006)
4. Zannini, L.: Medical humanities e medicina narrativa. Nuove prospettive nella formazione dei professionisti della cura. Raffaello Cortina Editore, Milano (2008)
5. Bernegger, G., Castiglioni, M., Garrino, L.: Un medico tra radure, tigri e jazz. A colloquio con Rita Charon. Rivista per le Medical Humanities **28**(8), 67–77 (2014)
6. Gottschalk, L., Gleser GC., : The measurement of psychological states through the content analysis of verbal behavior. University of California Press, Berkeley, (1969)
7. Overcash, JA.: Narrative research: a review of methodology and relevance to clinical practice. Critical Review Oncology/Hematology **48**(2), 179–184. USA (2003)
8. Jordens, CF., Little, M., Paul, K., Sayers, EJ.: Life Disruption and Generic Complexity: a Social Linguistic Analysis of Narratives of Cancer Illness. Social Science & Medicine **53**(9), 1227–1236 (2001)
9. Owen, WF.: Interpretative Themes in Relational Communication. Q J Speech **70**(3), 274–287 (1984)
10. Bakhtin, MM.: The Problem of Speech Genres. In Emerson C. & Holquist M., Speech Genres and Other Late Essays. University of Texas Press, Austin (1986)
11. Gottschalk, L., Gleser GC., : The measurement of psychological states through the content analysis of verbal behavior. University of California Press, Berkeley, (1969)
12. Weber, RP.: Basic Content Analysis. SAGE, Newbury Park (1990)
13. http://www.viverlatutta.it/. Last accessed 31 March 2021
14. Deng, Y., Stoehr, M.,& Denecke, K.,: Retrieving attitudes: Sentiment analysis from clinical narratives. In: MedIR@ SIGIR, pp. 12-15 (2014).
15. Vinodhini, G., Chandrasekaran, R. M.: Sentiment analysis and opinion mining: a survey. International Journal **2**(6), 282–292 (2012)
16. Zucco, C., Calabrese, B., Agapito, G., Guzzi, P. H., & Cannataro, M.: Sentiment analysis for mining texts and social networks data: Methods and tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(1), e1333 (2020).
17. Neustein, A., Sagar Imambi, S., Rodrigues, M., Teixeira, A., & Ferreira, L.: Text Mining of Web-Based Medical Content. De Gruyter, (2014)
18. Liu, S., Lee, I.,: Extracting features with medical sentiment lexicon and position encoding for drug reviews. Health information science and systems. **7**(1): 1-10, (2019).
19. Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022, (2003), doi:10.1162/jmlr.2003.3.4-5.993
20. Griffiths, T. L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101 Suppl 1(Suppl 1), 5228–5235, (2004) https://doi.org/10.1073/pnas.0307752101

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_45