Scientific texts analysis and classification by using machine learning methods

Nicolas LAFITTE

September 10, 2019

Abstract

The project is part of the Artificial Intelligence (AI) and Machine Learning (ML) training course held at Sorbonne University (Paris, France) during 2018-2019 academic year. As final project, a work on the analysis and classification of scientific texts have been chosen.
The initial motivations were to discover and learn mathematical algorithms and methods used in AI and ML in order to evaluate any application for the Fluigent European project HoliFAB. It aims at adapting an existing pilot line for the production of microfluidic instruments, and develops hardware and software strategies for the optimization of the production and system. Some mathematical developments at the beginning of the project led to the implementation of regression functions that auto-place and auto-wire a system layout.

However the motivation for the current project raises from the frustration as a researcher to not be able to read and study all papers relevant of a field. Making a bibliography on a specific topic is common and easy to perform thanks to different database and search engine available on the internet. However, when it comes to study a broad scientific field for a global understanding or new research investigation or market analysis, the task often lacks an intensive study of the domain.

Text mining, which is the task of extracting meaningful information from text, is developed here on a database of scientific papers. More precisely Latent Dirichlet Algorithms (LDA) are studied in order statistically categorize text and extract topics. The tool helps to find and name topics and applied on different database of different years check for evolution in the scientific research and next innovation that will probably lead the market.

Keywords: mining, classification, clustering, information extraction, topic extraction, information retrieval, Latent Dirichlet Algorithm (LDA)

Contents

Chapter 1

Introduction

Disclaimer: This report is widely inspired by the review of Allahyari et al., *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques* [Allahyari et al., 2017].

In a very recent paper [Tshitoyan et al., 2019], researchers at Lawrence Berkeley National Laboratory have developed an artificial intelligence (AI) that paves the way to predict discoveries in science. The scientists gathered 3.3 million articles on materials science from 1,000 different journals published between 1922 and 2018, and trained the renowned Word2vec algorithm in order to build statistical connections between words that are in the same context (cf. words embedding). On the one hand, the program was able to classify well known thermoelectric materials explicitly mentioned in the scientific abstracts alongside the word “thermoelectric” or associated words like ‘ZT’, ‘zT’, ‘seebeck’, ‘thermoelectric’, ‘thermoelectrics’, ‘thermoelectrical’, ‘thermoelectricity’, ‘thermoelectrically’ or ‘thermopower’. However by mathematical projection of all the materials, it is also indicating a relationship that is not explicitly written in the text Figure 1.1. Particularly Figure 1.1-c demonstrates how words that are chemical formula, ie. totally new character strings, are associated through concept words expressing applications (electronic, optoelectronic, photovoltaic), physical parameters (bandgap, heusler compound) or others known thermoelectric material (PbTe, Cu_2Te, Cu_5Te_3).
On the other hand, articles from 2000 to 2018 were removed and 18 new predictive models were trained in order to predict discoveries the following years and evaluate prediction abilities (Figures 1.2 and 1.3). Each of these models ranked materials according to their similarity to the word “thermoelectric” (or “ferroelectric”, “photovoltaic”, “topological insulator”), and took the top 50 that were not studied as thermoelectrics as of that year. It turns out, many of these materials were subsequently reported as thermoelectrics in future years.

This work is perfect example of the target of the current work, ie. study future research and markets on a specific field that is here microfluidics.

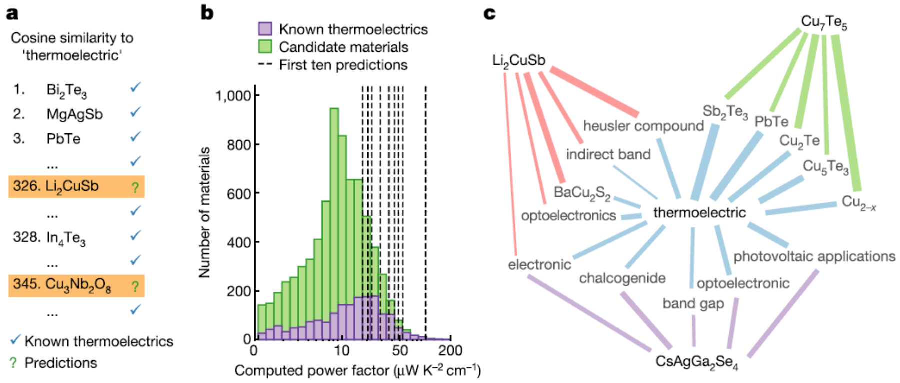

Figure 1.1: Prediction of new thermoelectric materials based on 3.3 million scientific articles from 1,000 different journals published between 1922 and 2018 [Tshitoyan et al., 2019].
a. Ranking table of thermoelectric materials. Materials with a check symbol are found in the context with thermoelectric terms (ie. 'ZT', 'zT', 'seebeck', 'thermoelectric', 'thermoelectrics', 'thermoelectrical', 'thermoelectricity', 'thermoelectrically' or 'thermopower'). However materials with a interrogation point are not explicitly studied as thermoelectric but are "mathematically" close and potential predictions that can be tested in the future.
b. Distributions of the power factors (computed with specific chemistry calculations) for 1,820 known thermoelectrics in the literature (purple) and 7,663 candidate materials not yet studied as thermoelectric (green). Dashed lines show the 10 first predictions of table **a**: Li_2CuSb . . .
c. Graph showing how the context words of materials predicted to be thermoelectrics connect to the word thermoelectric. The materials are the first (Li_2CuSb), third ($CsAgGa_2Se_4$) and fourth (Cu_7Te_5) predictions of table **a**. Examination of the context words demonstrates that the algorithm is making associations on the basis of crystal structure, co-mentions with other materials for the same application, between different applications and key phrases that describe the material's known properties.

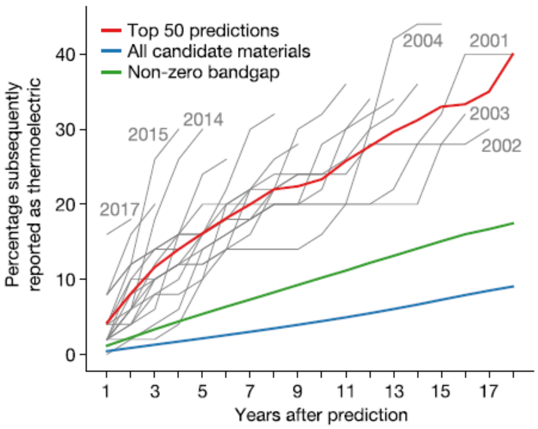

Figure 1.2: Predictions of thermoelectric materials based on previously published papers [Tshitoyan et al., 2019].

For example, predictions for 2001 are performed using abstracts from 2000 and earlier, and the grey lines plot the cumulative percentage of predicted materials subsequently reported as thermoelectrics in the years following their predictions. The results are averaged (red curve) and are 8 times more likely to have been studied as thermoelectrics within the next five years as compared to a randomly chosen unstudied material from our corpus at that time (blue curve) or three times more likely than a random material with a non-zero band gap (green curve).

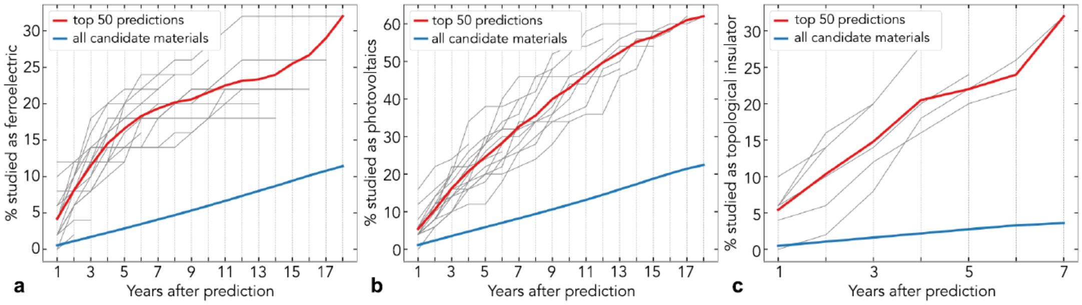

Figure 1.3: Same prediction methods as Figure 1.2, but with words related to ferroelectric (graph **a**), photovoltaic (graph **b**) and topological insulator (graph **c**) (Supplementary material of [Tshitoyan et al., 2019]).

Chapter 2

Implementation

The main source of machine-interpretable data for the materials research community has come from structured property databases. Beyond property values, publications contain valuable knowledge regarding the connections and relationships between data items as interpreted by the authors. To improve the identification and use of this knowledge, several studies have focused on the retrieval of information from scientific literature using supervised natural language processing, which requires large hand-labelled datasets for training.

2.1 Pipeline

2.2 Data extraction

2.2.1 Sources

2.2.2 Converting pdf to dataframe

2.3 Text transformation

The texts need some transformation to be relevant for algorithms or models applied afterwards. Typically models sensitive to term appearance frequency to determine importance of the words will be biased by recurrent term like in English: *I, and, or . . .* that do not carry much meaning.
A transformer is an abstraction that includes feature transformers and learned models: ie. A transformer implements a method, which converts one dataframe into another, generally by appending a new column.

2.3.1 Tokenization

Tokenization is the process of taking the text (such a sentence) and breaking it into individual terms (usually words).

2.3.2 Stop words

Stop words process takes as input a sequence of strings (e.g. the output of the tokenization) and drops all the stop words.
Stop words are words which should be excluded because the words appears frequently and carry as much meaning.

2.3.3 Stemming

2.4 Algorithm

2.4.1 Latent Dirichlet Algorithm

2.4.2 Others

Chapter 3

Results

Chapter 4

Conclusion and perspectives

Chapter 5

Annexes

Bibliography

[Allahyari et al., 2017] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.

[Tshitoyan et al., 2019] Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.