In Search for Patents: Modelling Technology Innovation

Bisevac Nezla
International Burch University, Sarajevo,
Bosnia and Herzegovina, Bsc

***Abstract:-* One of the world's largest repositories of technological innovation is patent data. To extract conclusions on how a particular choice of technology relies on scientific knowledge and what leads to a successful deployment of a patented technology, we apply data science techniques and propose a model for automating patent search and patent analytics. A multi-method approach including keyword extraction and k-means clustering is applied to a dataset of air pollution technology patents. The results provide insights into the drivers of technology innovation in reducing air pollution and the role of patenting in promoting innovation in this field. Keywords extraction of the patent title showed that: (1) there are ways to observe is the patent is successful when it comes to reducing air pollution, (2) the number of keywords combined does not lead to results improvement due to the presence of false patents in the dataset, and (3) using k-means clustering as a natural language processing technique for categorizing patents based on their feature can be a powerful tool in patent analytics, allowing us to quickly and efficiently analyze patent datasets. The study also points out to the problem of collecting patents: problem of datasets and the amount of false patents they can contain. The paper concludes that the combination of keywords extraction and k-means clustering is an effective research methodology to explore of nature of the patents in the imported dataset and to evaluate the relationship between technology innovation and patenting.**

Keywords: ***Air pollution; patent data; keyword extraction; k-means; patent analytics; data science;***

I. INTRODUCTION

Air pollution is a critical global issue affecting human health and the environment, caused by industrialization and urbanization that increases pollutants in the air. It contributes to 7 million premature deaths annually, with over a million in China, and decreases productivity, increases health problems, and weakens national security.

As technological patents are a powerful indicator of innovation, but also scientific knowledge, modelling a heterogeneous nature of patents to gain insights into the evolution of particular technology so that policymakers can better allocate R\&D budgets and adopt technological solutions to societal problems, such as air pollution, is a research problem worth exploring.

We use patent analytics to analyse patents related to air pollution, identify trends in innovation, and better understand the problem. We also examine the commercial and societal factors driving the filing of patents related to air pollution, and explore the limitations and challenges of these solutions in different parts of the world.

Ultimately, the goal of this research is to shed light on the relationship between technological innovation and scientific knowledge in the field of air pollution, and to assess the effectiveness of current solutions in reducing air pollution and improving air quality globally.

Overall, the aim of this study is to contribute to a better understanding of the challenges and opportunities in the global fight against air pollution, and to provide insights into how technology and scientific knowledge can be leveraged to reduce the impact of air pollution on human health and the environment.

To perform this task, we will use patent databases and scientific articles cited within the patent applications and approvals. To model this relationship between science and technology, we have chosen technologies that address the air pollution problem, and we then track its patent data, explore and apply patent analytics methods on them. Using the patent data on new technologies addressing air pollution problem, we believe, can enable more efficient and evidence-based policy making and lead to better measures of progress on the sustainable development goals (SDGs), focused on air pollution but it is also possible to generalise this methodology further and address other societal issues as well.

II. LITERATURE REVIEW

Air pollution as a field hasn't been a dominant research topic in the field of patent analytics. For those that have addressed the problem in the scope of patent analytics, various methodologies have been applied to analyse patent data for technology management, where the majority used text-mining techniques that can be narrowed to keywords extraction from patent documents [2], visualisation method using different tools, use of software tools for patent analytics.

[1] is based on establishing that technological interdependencies can help predict future innovation dynamics. The empirical observation can be explained with a simple model of network-dependent knowledge creation. Also, the model is validated by making out-of-sample predictions of growth rates, conditional and unconditional of neighbouring growth rates. Neighbourhood patenting growth rate is computed for each technology as the ANNG (average nearest neighbour growth rate).

The potential gap in this paper, or something that was left behind is the data preprocessing. The patents were not analysed before the network creation step. Even though data preprocessing is not crucial in this analysis, omitting data preprocessing can lead to lower performance results.

A method that was proposed in [3] was that prediction of technological success can be done by using neural network models. The approach is implemented using two Neural Networks models for accuracy comparison: a Wide and Deep Neural Network (WDNN) and a Recurrent Neural Network (RNN). As it was stated in the paper, predicting technology's success is a complex task in terms of prediction accuracy and data availability.

The limitation here is that it was almost impossible to access the accurate Big Data, because the data that is available to a wider audience is limited to a small dataset that was extracted only from one source. Since the data is our most important resource when building the model, using small datasets can lead to weaker performance of the model. Having more data means better chances for our model to score better results.

Another limitation here is that patent data is categorised by technology based on keywords matching only the title of the patent application. Using only patent titles as relevant, without checking for a description can lead to excluding important records - which can, again, result in low performances.

III. METHODOLOGY

We take the multi-method approach to research design as depicted in Fig.1, and use a variety of techniques and tools. Raw data will be imported into the appropriate tool and processed. Then, a processed dataset will be used in order to create visualisation in Tableau, create keywords using Python pandas, visualise keywords and their frequency, and finally use them to perform *k-means clustering*.

The main input to our model is the dataset we create by using filters in the advanced search in Google Patents[1] search engine.

Based on the chosen filters, the search string *(air pollution) country:US before:priority:20220101 after:priority:20000101 status:GRANT language:ENGLISH type:PATENT litigation:NO* resulted in 135828 data points in the dataset.

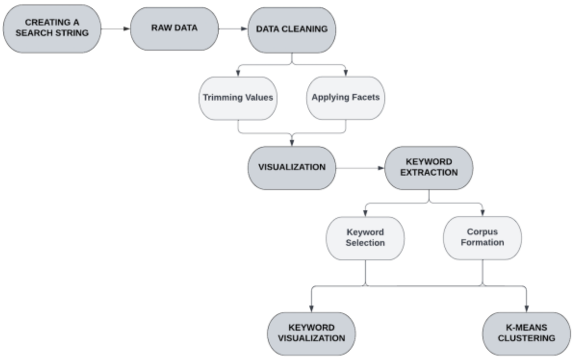
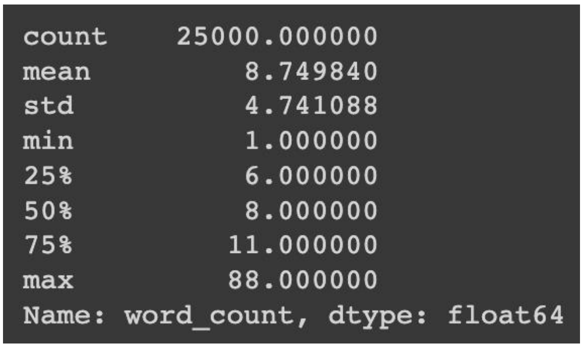


Figure 1. Overall research process.

A. The Dataset Statistics



```
count    25000.000000
mean         8.749840
std          4.741088
min          1.000000
25%          6.000000
50%          8.000000
75%         11.000000
max         88.000000
Name: word_count, dtype: float64
```

Figure 2. Descriptive statistics after cleaning the dataset.

Descriptive statistics can be used to summarise and describe the main features of a dataset, such as:

- Central tendency (mean, median, mode);
- Spread (range, variance, standard deviation);
- Shape of the distribution (skewness, kurtosis);
- Identifying outliers;

As seen in Fig.2, with these metrics, we can get an overall understanding of the dataset and detect any patterns, distributions, or anomalies in the data.

Average word count is about 9 words per title. The word count ranges from a minimum of 1 to a maximum of 88. The word count is important to give us an indication of the size of the dataset that we are handling as well as the variation in word counts across the rows.

B. Text pre-processing

Text pre-processing involves two main tasks: noise removal, which removes unnecessary information, and normalisation, which standardises the text. Noise removal eliminates any data elements that are not crucial for the main text analysis task. Normalisation reduces multiple variations of a word to a common base form, which can be achieved through stemming or lemmatisation techniques. See the image below

[1] Google's own patent search tool which draws on patents from more than 100 of the biggest patent offices from around the world, enabling the search through more than 120 million documents. https://patents.google.com/

for a visual representation of the text preprocessing workflow.

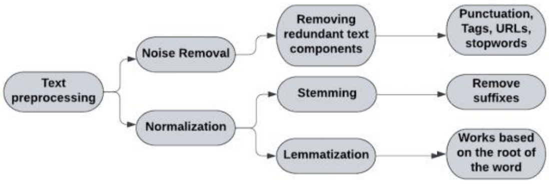

Text preprocessing
Noise Removal
Normalization
Removing redundant text components
Stemming
Lemmatization
Punctuation, Tags, URLs, stopwords
Remove suffixes
Works based on the root of the word

Figure 3. Text preprocessing procedure.

Stop words are commonly used words in a language, such as prepositions and pronouns, that are removed from text to focus on the more relevant words in context. A list of common words will be added to the stopwords list during text preprocessing to improve the dataset for ML. The dataset has 25k rows with 16k different words, and common/uncommon words will be determined by personal judgement to avoid incorrect grouping. Common words unlikely to relate to air pollution will be added to the stopwords list, while uncommon words related to air pollution will form a separate list.

To get a better overview of pre-processed dataset and visualise the generated text corpus, we will generate a word cloud.

To prepare text for the NLP, it must be converted to a format that can be understood by the algorithm. This involves tokenization and vectorization, with the bag of words model used for text preparation. The CountVectorizer will be used to tokenize the text and establish a vocabulary of known words. This involves creating a CountVectorizer object and using the fit_transform method to construct the vocabulary.

Figure 4. Wordcloud.

C. K-means clustering

We will use k-means clustering on three versions of corpus:

- Uni keywords, first version - consists of one-word keywords. This dataset is the largest and therefore will have the biggest number of clusters. Also, chances of errors in clustering are the highest here
- Bi keywords, second version - consists of two-word keywords and is 5 times shorter than the first dataset, therefore, smaller chances of errors
- Three keywords, last version - consists of three-word keywords, has the same size as the second version, but returns the most accurate results.

Clustering can produce poor results if the initial starting point is bad. The elbow method is a way to determine the best number of clusters for the K-means algorithm, which is consistent and independent of the data. The resulting number of clusters will be the default for the clustering process. The K-means inertia graph can be used to gain a better understanding of this process.

Creation of datasets will be done using the functions created for the purpose of visualisation of preprocessing results: *get_top_n_words, get_top_n2_words* and *get_top_n3_words*.Keywords will be first generated using the get_top_n_words function and then, to be able to apply k means, they will be vectorized. The wider range for determining the number of clusters we choose, the higher possibility is that the elbow method will return a linear function and smaller number of clusters, or even, a 0, which means no clusters at all.

D. Elbow Method

Elbow method will be performed for the uni and bi keywords dataset. Tri keywords dataset will be omitted since it's the same size as bi keywords dataset and therefore, results won't show significant difference. For the first dataset elbow method will be run for the range from 1-8 since this is the largest of three datasets, for the second dataset range will be from 1 to 5 and for the third dataset from 1 to 4. Results can be seen in the pictures below. Ideal number of clusters for the first dataset is 4 and 3 for the second.

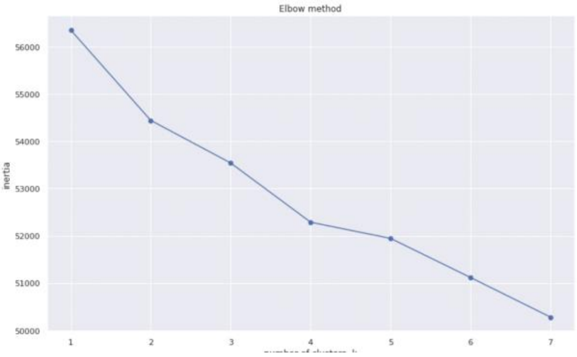


Figure 5. Elbow method for uni keywords dataset

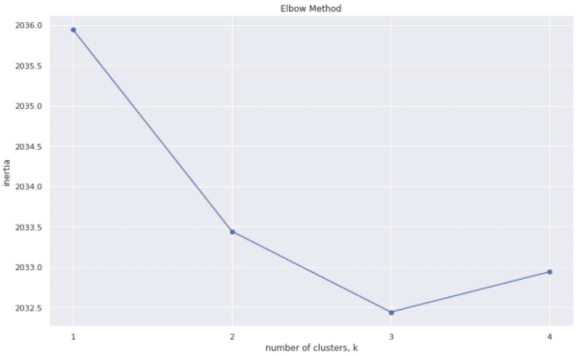


Figure 6. Elbow method for bi keywords dataset

IV. RESULTS AND DISCUSSION

Three step procedure will be followed to see the performance of k-means for the first two datasets:

1. Number of patents containing the chosen word will be evaluated
2. After determining the purpose of the patent, it will be split in one of three categories
    a. Patent is directly related to reducing the air pollution - labelled as useful
    b. Patent is not directly related to reducing the air pollution - labelled as helpful
    c. Patent is not related to air pollution in any ways - labelled as purposeless

For the first cluster in the uni keywords dataset, words multi with frequency of 504 and coating with frequency of 280 are the most dominant. Multi doesn't tell us much but, coating is usually applied to the surface of an object. Its purpose can be functional and decorative and can be applied as liquids, gases or solids.

Coating can contribute to air pollution as it produces hazardous air pollutants and heavy metals. In the dataset, there are 280 patents with coating in their title. Two of them, Insulator Coating For Reducing Power Line System Pollution Problems and Transparent Photocatalyst Coating, can help reduce air pollution and are therefore, both of these oatents can be labelled as useful.

However, there are many more patents that are not directly related to air pollution.

Patent for Non-yellowing Polyester Coating Composition could indirectly help reduce air pollution by potentially reducing waste and resources needed for re-coating or replacing the coated item. It doesn't directly affect air pollution but it can help in reducing it, and therefore it belongs in third category, helpful.

It is highly unlikely that Coating liquid, an image recording method and recording using same developed by Seiko Epson Corporation can address air pollution. The purpose of this patent is likely to improve the efficiency and quality of image recording processes, rather than reducing air pollution. This patent is in the third category, purposeless.

Most dominant keywords in the second cluster are system and device. System has the highest frequency so it is more likely to bring promising results compared to the other words.

Air Pollution Control System and Air Pollution Control Method developed by Mitsubishi, has its purpose in providing a system and method for controlling air pollution, which may involve capturing and treating pollutants in the air, such as particulates, chemicals, and gasses. Patent System and a Method for Assessing and Reducing Air Pollution by Regulating Airflow Ventilation developed by Urecsys, can measure levels of pollutants in the air and adjust the ventilation and airflow to reduce those levels. These systems could help to reduce the amount of pollutants that enter the environment and contribute to air pollution. Both of these patents can be labeled as useful.

Air-Conditioning System, developed by Uniflair, is an example of a patent that is not directly related to air pollution, but can contribute to it. This type of patent is in the grey zone, we are sure it's not in the first or last group, but it has a high possibility to be in the second group.

Patents from this cluster that are not related to air pollution at all are Automated Competitive Bidding System and Process Developed in 2000 by Tariq Khalidi, Universal Verification and Validation System and Method of Computer-aided Software Quality Assurance and Testing by Sofia Passova developed also in 2000 but granted in 2003 and Systems and Methods for Monitoring Travel Conditions developed by United Parcel Service of America, granted in 2009. All three are in the category of purposeless.

Second keyword of the bi-keywords dataset, Semiconductor device is an electronic component that relies on the electronic properties of a semiconductor material[2] and can cause underground air pollution and generate toxic waste[3]. *Micro silicon fuel cell, method of fabrication and self-powered semiconductor device integrating a micro fuel cell* is one of the very few patents in this group that can be labelled as useful.

This group mostly consists of patents that are not directly related to reducing air pollution, including *Method and apparatus for forming a thin semiconductor film, method and apparatus for producing a semiconductor device, and electro-optical apparatus* developed by Sony Corporation, *Method of creating a high performance organic semiconductor device* developed by Precision Dynamics Corporation.

Although the patents in the dataset are focused on solving air pollution, a patent from this cluster, *Apparatus for Production of High Purity Carbon Monoxide*, published in the year 2016 represents a rarity. As the name itself says, it is a device that produces CO with a high degree of purity (this type of CO is used in various industrial processes). Although it is about the production of less harmful CO, the problem is still the production itself. In other words, the production of CO often involves the use of fossil fuels, which release carbon dioxide (CO2) and other pollutants into the atmosphere.Therefore, while the apparatus for production of high purity CO itself may not contribute to air pollution, the use of CO in industrial processes and the production of CO through the use of fossil fuels can contribute to air pollution, instead of helping to suppress it.

V. CONCLUSION

It is difficult to determine which patent has contributed the most to reducing air pollution, as the impact of a patent on air pollution can depend on many factors, including the specific technology covered by the patent, how widely the technology is adopted and used, and the types and levels of pollutants being targeted.

Since the data on the results of the patent have not been

[2] https://en.wikipedia.org/wiki/Semiconductor_device

[3] https://www.ncbi.nlm.nih.gov › articles › PMC1566445

found, and, because we do not have statistical data to confirm it, we cannot determine with certainty whether the patent is successful or not - the patent analytics is still mostly based on the critical thinking of the individual. However, by using machine learning techniques, as well as text preprocessing, which help to remove patents that do not belong to the problem at all, the process of patent validation can be facilitated and made less time-consuming.

Nevertheless, we saw that, using the keywords extraction and combining those keywords, there are ways to see if the patent is useful or not. The approach we chose in identifying the most powerful keywords was to take the same number of patents from each category, inspect their description and by critical thinking and scientific knowledge identify the category they belong to (useful/helpful/purposeless). Keywords with the highest number of patents marked as useful are the most powerful keywords - meaning that patents containing these words are more rational to be the change in reducing air pollution than other keywords in the datasets. Therefore, patents that include keywords coating, device and vehicle are those that are most likely to actually reduce air pollution.

This means that patent inventions should be more oriented towards solving the problem of major air pollutant contributors, vehicles. Also, all of these words were a part of the uni-keywords dataset which tells us that the larger number of words combined don’t guarantee the success of the algorithm. These results are formed based on the dataset, which, unfortunately, contained a lot of falsey data There were not many patents related to stove design or other kitchen tools, and the reason for that is that indoor pollution is more related to third-world countries and the dataset is for the USA only.

Patent labels can also help us to determine if the invention was driven by the desire to solve a problem or by commercial factors. Many patents were in the field of air pollution, but had nothing in common with it, like the patent for bidding systems or patent about software tools. This pulls a new problem - if there are a couple of recognised patents that didn’t help air pollution at all, how many more are in the dataset that were foreseen? Are the patent datasets reliable enough?

Another important thing to note is that when checking the dataset, attention should be paid to the number of patents that have 90% similarity. Not taking care of this step can lead to wrong conclusions when visualising the data. In our case, there were numerous patents filed multiple times with only one word replaced in the title, mostly with its synonym. Furthermore, all the derived conclusions are based on the results of the dataset, which contains a lot of invalid patents.

We cannot say with certainty that the results would be significantly better if the filters currently offered by the google patents database were improved, but we can say that they would be significantly better if the process of patent approval and more detailed review of them were improved. Air pollution is a serious problem that is responsible for numerous health problems, even death, and the seriousness of the problems it brings with it should not be misused.

Another problem we encountered during the analysis is the discovery of a patent that, in a process aimed at reducing air pollution, actually contributes to air pollution itself. Can this patent introduce us to a new problem: are we actually taking one step back with the desire to take two steps forward?

It is imperative that measures are taken to improve the quality of the patent approval process and the examination of the patents. This can be achieved by having stricter guidelines and standards in place to ensure that only valid patents are approved. Furthermore, the development of more advanced technologies, such as machine learning algorithms, can help to identify and reject invalid patents. This will not only improve the accuracy of patent analysis, but it will also help to ensure that patents are not misused for personal or financial gain at the expense of public health and the environment.

Additionally, better resources should be made available to review and validate patents, so that invalid patents are detected before they are approved, and resources are not wasted on invalid patents. Another important thing is for authors to make sure that the patent they are developing should be oriented towards solving the problem or creating steps that can reduce the problem, and not the opposite. If the desire to solve the problem leads to the creation of a new problem, it is necessary to try to find elements that help to suppress the newly created problem.

By taking these steps, we can create a more efficient and trustworthy patent system that supports the development of innovative solutions to address the pressing environmental and health problems we face today.

Reducing air pollution is a complex challenge that requires a multi-faceted approach, involving not just technology but also changes in policy, regulations, and consumer behaviour. So, while a small number of useful patents for individual technologies can play a role in reducing air pollution, they are just one piece of a larger puzzle in addressing this global challenge.

REFERENCES

[1] Pichler, A., Lafond, F., & Farmer, J. D. (2020). Technological interdependencies predict innovation dynamics. Technological Forecasting and Social Change, 159, 120580.

[2] Noh, H., Jo, Y., & L, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. Expert Systems with Applications, 42(11), 4983–4993.

[3] Saade, M., Jneid, M., & Saleh, I. (2019, September). Predicting technology success based on patent data using a wide and deep neural network and a recurrent neural network. In 2019 9th International Conference on Information and Communication Technologies (ICICT) (pp. 1-6). IEEE.