Government plans in the 2016 and 2021 Peruvian presidential elections: A natural language processing analysis of the health chapters

Rodrigo M. Carrillo-Larco1,2, Manuel Castillo-Cara3, Jesús Lovón-Melgarejo4

1Department of Epidemiology and Biostatistics, Imperial College London, London, SW7 2AZ, UK
2CRONICAS Centre of Excellence in Chronic Diseases, Universidad Peruana Cayetano Heredia, Lima, Peru
3Universidad de Lima, Lima, Peru
4Paul Sabatier University, Toulouse, France

Abstract
Background: While clinical medicine has exploded, electronic health records for Natural Language Processing (NLP) analyses, public health, and health policy research have not yet adopted these algorithms. We aimed to dissect the health chapters of the government plans of the 2016 and 2021 Peruvian presidential elections, and to compare different NLP algorithms.

Methods: From the government plans (18 in 2016; 19 in 2021) we extracted each sentence from the health chapters. We used five NLP algorithms to extract keywords and phrases from each plan: Term Frequency–Inverse Document Frequency (TF-IDF), Latent Dirichlet Allocation (LDA), TextRank, Keywords Bidirectional Encoder Representations from Transformers (KeyBERT), and Rapid Automatic Keywords Extraction (Rake).

Results: In 2016 we analysed 630 sentences, whereas in 2021 there were 1,685 sentences. The TF-IDF algorithm showed that in 2016, 26 terms appeared with a frequency of 0.08 or greater, while in 2021 27 terms met this criterion. The LDA algorithm defined two groups. The first included terms related to things the population would receive (e.g., 'insurance'), while the second included terms about the health system (e.g., 'capacity'). In 2021, most of the government plans belonged to the second group. The TextRank analysis provided keywords showing that 'universal health coverage' appeared frequently in 2016, while in 2021 keywords about the COVID-19
pandemic were often found. The KeyBERT algorithm provided keywords based on the context of the text. These keywords identified some underlying characteristics of the political party (e.g., political spectrum such as left-wing). The Rake algorithm delivered phrases, in which we found ‘universal health coverage’ in 2016 and 2021.

**Conclusion:** The NLP analysis could be used to inform on the underlying priorities in each government plan. NLP analysis could also be included in research of health policies and politics during general elections and provide informative summaries for the general population.

**Keywords**
Public health, health policy, Natural Language Processing, Latin America and the Caribbean, Peru, COVID-19

This article is included in the [Coronavirus (COVID-19) collection](https://wellcomecollection.org/collections/coronavirus).

**Corresponding author:** Rodrigo M. Carrillo-Larco (rcarrill@ic.ac.uk)
**Author roles:** Carrillo-Larco RM: Conceptualization, Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing; Castillo-Cara M: Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; Lovón-Melgarejo J: Conceptualization, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by the Wellcome Trust International Training Fellowship[214185/Z/18/Z to RMC-L]. The funder had no role in this work and decision to submit this work for publication. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Copyright:** © 2022 Carrillo-Larco RM et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Carrillo-Larco RM, Castillo-Cara M and Lovón-Melgarejo J. Government plans in the 2016 and 2021 Peruvian presidential elections: A natural language processing analysis of the health chapters [version 5; peer review: 1 approved, 4 approved with reservations] Wellcome Open Research 2022, 6:177 https://doi.org/10.12688/wellcomeopenres.16867.5

**First published:** 08 Jul 2021, 6:177 https://doi.org/10.12688/wellcomeopenres.16867.1
Introduction

Over the last few years, researchers working in clinical medicine have adopted artificial intelligence and deep learning techniques such as Natural Language Processing (NLP). Due to this, electronic health records have become unique data sources because they contain free text annotations that can inform NLP models for a variety of tasks (e.g., risk prediction). On the other hand, public health research appears not to have benefited from NLP algorithms, despite the fact that this research field also has large data sources of text, such as reports and policy briefs. In this line, health policy research could also use NLP algorithms to scrutinize laws and other documents, such as health-related government plans and proposals during presidential election periods. This could help identify patterns within and between plans, to study underlying topics and key words or ideas, and to assess coherency within the text. Furthermore, following the global call to have evidence-based health policies, NLP algorithms could also help to study government plans in the context of the scientific literature (e.g., SciBERT). Overall, NLP could offer novel and insightful ways to dissect health-related government plans, so that practitioners, researchers, and public health experts have further arguments to imagine the future scenario of the health sector in their country. Similarly, the general population could benefit from this NLP analysis to compare across political parties and make a much more informed vote. Our work contributes to the NLP health literature by leveraging on standard and well-known algorithms to analyse health policy/politics documents. Because most previous NLP health research focused on electronic health records (i.e., clinical oriented) or social media (i.e., population health oriented), our work expanded the use of NLP in health research to new documents (policy/politics documents) with potential public health implications. Furthermore, another contribution of our work also lies in the fact that we used a sort of text documents that, arguably, may be available to a much broader audience. Even though electronic health records and social media are valuable sources of information for NLP analyses, electronic health records may be seldom available in low- and middle-income countries where people could also engage less in social media. Nevertheless, in most countries during election period there may be some sort of documents summarizing the proposals by the candidates who may share these as much as possible for electors to familiarize themselves with the candidates.

We aimed to describe the health chapters of the government plans in the ongoing (April 2021) and last (2016) presidential elections in Peru. In addition, we compared and discussed different NLP algorithms. This work sought to introduce NLP into the health policy research agenda, while providing a new way of seeing, reading, and understanding the health chapters of government proposals during general elections.

Methods

Study design

This is an analysis of government plans from the 2016 and 2021 Peruvian presidential elections. For this analysis we used NLP algorithms: Term Frequency – Inverse Document Frequency (TF-IDF), Latent Dirichlet Allocation (LDA), TextRank, Keywords Bidirectional Encoder Representations from Transformers (KeyBERT) and Rapid Automatic Keywords Extraction (Rake).

In this analysis, we focused on the government plans regardless of the political party. That is, we do not make specific references or conclusions about the political parties; however, the names of the parties included in the study are presented in the tables and figures. We do not make specific arguments about the political parties to avoid introducing political bias or preferences. We aimed to make this a data-drive analysis to dissect the content of the government plans, and not a quality assessment of the government plans (or the political parties behind it).

Scenario

According to the World Bank, Peru is an upper-middle income country in South America with a total population of 32.5 million people. Peru is a unitary presidential republic, in which the president is elected in general elections every five years; the most recent presidential elections took place in 2016, and the next one will take place in April 2021. The campaign for the forthcoming presidential elections started in December 2020 with over 15 candidates (i.e., political parties). Each candidate must publish a government plan.

Data sources

In this study, we analysed the government plans of the 2016 and 2021 presidential elections in Peru. These plans are open access and in the public domain, so that all citizens can read these and become informed of their proposals. For this work, which is health-oriented, we only selected the health chapters.

For the 2016 presidential elections, there were 19 government plans presented, although we analysed only 18 in this study as one was only available as a scanned image and could not be analysed as text. For the 2021 presidential elections, there were 20 government plans and we analysed just 19 as one was only available as a scanned image and could not be analysed as text. As we aimed to analyse the government plans, and not the political party, we analysed all government plans regardless of whether the party or the candidate was disqualified during the campaign period.

Within the health chapter of each government plan, we copied and pasted onto a spreadsheet each sentence in a row; in other words, for each government plan, we had as many rows as sentences. Furthermore, the dataset we generated had three columns: the name of the political party, a sentence indicator...
(e.g., political_party_A_sentence_1), and the sentences we extracted. All sentences were copied as they were in the original government plan. The government plans were translated into English using Google translate.

Natural Language Processing Analysis

Data preparation. NLP is a branch of artificial intelligence which aims to process and understand the human language. Not only does NLP need to understand the text (words), but it also needs to make sense of the context to provide accurate meaning to the text. To begin with NLP, we need to put a text into a two-dimensional matrix, In the model Bag of Words, each text or document is represented by the group of words in it.

As a previous step to the construction of the data matrix, a vocabulary is elaborated with the union of all the terms that appear in a document. From this, a matrix is built in which each row represents a document, and each column (each characteristic) corresponds to a term.

For this analysis we conducted a simple pre-processing, which included deleting HTML labels and non alphabetic characters for all datasets. Moreover, an advanced preprocessing of the datasets depends on the algorithm used. Therefore, in the description of the algorithms, the data pre-processing that has been used will be described for each algorithm.

Algorithms

We used five NLP algorithms: TF-IDF, LDA, TextRank, KeyBERT, and Rake. We chose these algorithms because they are frequently used in the NLP literature, they are available through most libraries in Python and R facilitating the replication and outreach of our work, and they satisfied our goals (i.e., to describe the health chapters in government plans by summarising words and terms).

TF-IDF and TextRank reveals keywords; TF-IDF does so based on the relative frequency of each term whereas TextRank selects keywords based on a trained model. Compared to TF-IDF and TextRank, KeyBERT as well as Rake reveal keywords or short phrases (e.g., 3-5 words) based on trained models. LDA, on the other hand, groups texts into unknown clusters and the user labels such clusters (i.e., as in unsupervised machine learning with K-means). Altogether, these algorithms provide framework to describe the health chapters in government plans going from unique terms (TF-IDF and TextRank), to short phrases (KeyBERT and Rake), and grouping into unknown clusters (LDA).

TF-IDF: This is the most basic NLP algorithm. Each position \(d, t\) in the matrix has a value for the metric Term Frequency which is denoted as \(f_{d,t}\) and reflects the number of times the term \(t\) appears in the document \(d\). There are variants to the Term Frequency metric, so that the matrix can include:

- \(f_{d,t} \in [0, 1]\) - if the term appears or not in the document.
- \(\log (1 + f_{d,t})\) - logarithmic scale.
- \(t f_{d,t} \max, t f_\max\) - scale in relation to the most frequently used term in the document.

There are terms which appear several times in a document, but which not carry much information; these could include 'the', 'a', 'an' and 'of', among others. The metric Inverse Document Frequency penalises the terms in a magnitude equal to the frequency in which they appear in the text. The inverse frequency of a term is computed as:

\[
id f_t = \log \left( \frac{\# \text{documents}}{\# \text{documents that include word } t} \right)
\]

Because both metrics (Term Frequency and Inverse Document Frequency) provide information to the NLP algorithm to dissect the text, it is common practice to combine them as the TF-IDF metric, which can be expressed as:

\[
t f = id f_{d,t} = t f_{d,t} \times id f_t
\]

For TF-IDF, we used the dataset partitioned by phrases to analyse globally because a wide corpus of words is needed, so here the frequency of words in general is analysed (for all political parties).

Hence, we used the TfidfVectorizer class from the Scikit-Learn library with the parameter stop_words='english'; this parameter reads all the text from each government plan and removes non-relevant terms from a text such as 'the' and 'as', among others. Moreover, lemmatisation was used with the class WordNetLemmatizer from nltk library. The following filler words were excluded: execute, force and change; although these are not traditional filler words, in the political context they are filler words.

LDA: LDA builds a generative model from a set of observations that belong to unobserved groups. Basically, it considers each group as a probability distribution over the characteristics, and each observation as generated from a mixture of the groups' probability distributions.

This model is frequently used to categorise a document into subjects or topics. For each topic, this model estimates the probability of the terms, while for the document, this model estimates the relevance of each topic.

Therefore we used LDA to define and describe, through the most relevant terms, the categories in which each data entry can be assigned to. We trained the LDA model with a dataset in which each sentence was an instance, and the categorisation was then conducted for the government plan.

The input for training the LDA algorithm was all the government plans (separately for each year), having removed the stop words and lemmatized the terms. Then, with the cleaned (lemmatized and without stop words) text for each political party, the model assigns each party to a group defined during the LDA training. We used LatentDirichletAllocation class from the Scikit-Learn library with the parameter n_components=2; this parameter reads the number of topics to be created. For this work, we defined two groups (i.e., we specified that the LDA algorithm should create two groups following an unsupervised approach from the data). Two is
the absolute minimum number of groups and we chose such option this being an exploratory work compounded with other NLP algorithms, and we did not focus solely on the LDA results. For a discussion on this limitation, please refer to the Strengths and Limitations section. Finally, for the data pre-processing, we used stop_words and WordNetLemmatizer classes from nltk library.

**TextRank:** This algorithm is a graph-based algorithm and it is based on PageRank developed by Google, where PageRank is used to define the position of websites in the search engine through the extraction of key terms.

TextRank informs about the semantic sequence used in the text with unique keywords. We used the library Gensim.

For this analysis, we used the full government plans to extract the key words in each plan. Following the recommendations in the documentation for TextRank, we did not pre-process the data and thus used the original text as it was.

**KeyBERT:** KeyBERT is an adapted keyword extraction tool based on BERT. In the last couple of years, algorithms based on BERT are pushing the boundaries of the state of the art technology in social and technical multiple aspects of the natural language processing tasks. BERT is a pre-trained model on the Book-Corpus and English Wikipedia datasets. It uses a multi-head attention mechanism, so it can create better representations for entities considering the context, and leveraging the classic word-based approach.

For this analysis we used the full government plans to extract the keywords in each one. Also following the documentation for the KeyBERT algorithm, the texts were not pre-processed and used as they were. A limitation of this algorithm is that it retrieves a maximum of six keywords; though this did now affect our work because we sought five keywords as it was the case for other algorithms.

**Rake:** Rake is used to extract keywords and key phrases. To use the Rake algorithm, we followed these steps:

1. In the text we identified stop words and punctuation.
2. We removed stop words and punctuation, and generated a list of the phrases that were separated by them.
3. We calculated the number of times each word appeared in all the phrases (i.e., frequency of a given word).
4. For two given different words in the text, we estimated how many times they were together in the same phrase (i.e., a metric of co-occurrence).
5. For every given word, we calculated a score: frequency (step 3) divided by the co-occurrence (step 4).
6. A score for a complete given phrase was computed as the sum of the scores (step 5) of each word in such phrase.

For this analysis we used the full government plans to extract the phrases in each one. Also following the documentation for the Rake algorithm, the texts were not pre-processed and used as they were. We used the class Rake from the rake_nltk library; for further details including the equation for the Rake method please refer to reference 15.

**Ethics.** The underlying data for this study is accessible in the public domain and did not include the names of individuals, but the name of the political parties which created each government plan; this information is also in the public domain. Therefore, we considered this work of minimal risk and did not seek approval by an ethics committee or institutional review board.

**Results**

**Data characteristics**

The original dataset with the 2016 government plans had 559 rows (i.e., sentences) in total, which represented 18 government plans; the shortest government plan contributed with four sentences, while the longest contributed with 96 sentences (See Table 1). The original dataset with the 2021 government plans had 1,586 rows (i.e., sentences) in total, which represented 19 government plans; the shortest government plan had 10 sentences, while the longest had 215 sentences (See Table 1).

**TF-IDF**

The TF-IDF analysis showed differences between the plans in 2016 and 2021. When we set a threshold of 0.08 to select terms (i.e., terms that represented 8 per cent or more of all the words in the text), 26 terms met this criterion in 2016 and 27 terms met this criterion in 2021. These terms are shown in Figure 1 for the years 2016 and 2021.

In 2016, across all government plans, the term 'care' had a TF-IDF frequency close to 0.40, while the terms 'system', 'service', 'level', 'public', 'population' and 'national' had a frequency above 0.15 (see Figure 1(a)). In 2021, across all government plans, the term 'care' had a TD-IDF frequency close to 0.35; the terms 'system', 'public' and 'level' had a frequency above 0.20 (Figure 1(b)).

**LDA**

The LDA algorithm was trained in a dataset with as many rows as sentences per government plan, to define two groups of ten words each (see Table 2); we did not define more groups, or more terms per group, because the dataset was small. For the prediction phase, we used the full text of each government plan (i.e., a dataset with as many rows as sentences per government plans), and analysed which group each government plan would belong to each group. We did this analysis for the years 2016 and 2021 separately.

Overall, in both 2016 and 2021, Group 0 appeared to cluster terms signaling things the population would receive (see Table 2). For example in 2016 and in 2021, Group 0 included 'service', 'access' and 'insurance'. Conversely, in both 2016 and 2021, Group 1 appeared to cluster terms that are related with...
Table 1. Characteristics (number of sentences) of the original 2016 and 2021 datasets. Symbol — means that the political party did not stand for election in that year.

| Political Party          | Dataset 2016 | Dataset 2021 |
|--------------------------|--------------|--------------|
| Acción Popular           | 8            | 18           |
| Alianza para el Progreso | 6            | 95           |
| Alianza Popular          | 33           | —            |
| APRA                     | —            | 84           |
| Avanza Pais              | —            | 10           |
| Democracia Directa       | 4            | 40           |
| Frente Amplio            | 67           | 83           |
| Frente Esperanza         | 22           | 35           |
| Fuerza Popular           | 25           | 122          |
| Juntos por el Peru       | —            | 124          |
| Orden                    | 74           | —            |
| Partido Humanista Peruano| 7            | —            |
| Partido Morado           | —            | 165          |
| Partido Nacionalista Peruano | 42    | 98           |
| Partido Popular Cristiano| —           | 138          |
| Perú Libertario          | 15           | —            |
| Peru Libre               | —            | 55           |
| Perú Nación              | 6            | —            |
| Perú Patria Segura       | 96           | 102          |
| Perú Posible             | 23           | —            |
| Peruanos por el Cambio   | 94           | —            |
| Podemos                  | —            | 32           |
| Progresando Perú         | 6            | —            |
| Renovación Popular       | —            | 29           |
| Siempre Unidos           | 20           | —            |
| Solidaridad Nacional – UPP| 11       | —            |
| Somos Perú               | —            | 215          |
| Unión por el Perú        | —            | 44           |
| Victoria Nacional        | —            | 97           |

TextRank

The TextRank algorithm shows the keywords in the text. The number of keywords depends on the size and coherence of the text; that is, longer texts and those with more complexity would have more keywords than small texts with poor context. For an informative representation, we chose the top six keywords per government plan. Keyword could have between one and three words.

In both 2016 (see Figure 4, Figure 5 and Figure 6) and 2021 (see Figure 7, Figure 8 and Figure 9), the term 'health' was the most frequent keyword. In 2016, terms regarding 'universal health coverage' were also frequent; in 2021 however, terms about 'universal health coverage' were not present. In 2016, the terms were more general, and appeared to focus on improving the health system with words like 'hospital', 'population', 'public' or 'region'.

In 2021, we found words relevant in the context of the COVID-19 pandemic, like 'national emergency'. We also found words that reflected the characteristics of the health system at the beginning of the COVID-19 pandemic, for example 'low investment' and 'lacking'.

Bear in mind that although TextRank provides the most frequent keywords, these words alone do not provide insights about the context of the text. Therefore, it becomes relevant to study the context of the text, so we also used the KeyBERT (group of terms) and Rake (to study phrases) algorithms.

For the year 2016, we observed that the government plans likely to belong to Group 0, were much less likely to belong to Group 1 (see Figure 2(a), Figure 2(b)). There were four government plans with a probability between 0.40 and 0.60 of belonging to either Group 0 or Group 1. There was strong evidence suggesting that seven (out of eighteen) government plans would belong, almost exclusively, to Group 1 (see Figure 2(a), Figure 2(b)).

The distinction in favour of Group 0 was less clear when analysing the 2021 government plans (see Figure 3(a), Figure 3(b)), when we observed one government plan very likely to belong to Group 0 (in 2016 there were four). Conversely, there were nine government plans with high probability of belonging to Group 1, yet very low probability of belonging to Group 0.

Figure 2 shows the LDA results for the government plans in 2016 and Figure 3 for the plans in 2021. While these figures show that political parties are more inclined to either Group 0 or Group 1, a direct comparison between 2016 and 2021 is troublesome because: i) the same exact political parties did not participate in the two elections; ii) the composition of the political parties in terms of roles, interests and priorities may have changed between 2016 and 2021; and iii) amidst the COVID-19 pandemic, the focus of the health-related government plans would be different between 2016 and 2021. Perhaps in response to the COVID-19 pandemic, the difference observed between Figure 2B and Figure 3B could signal that the 2021 plans focused more on terms related to the structure of the health system (i.e., Group 1 appeared to cluster terms that are related with the structure of the health system).
Figure 1. Frequent terms in the government plans (2016 and 2021) as per the TF-IDF algorithm.
### Table 2. Keywords (probability for each term in the group) classified to each group by LDA for each Government plan year.

| Government plan year | Group 0                  | Group 1                  |
|----------------------|--------------------------|--------------------------|
| 2016                 | system: 0.0181           | care: 0.0242             |
|                      | public: 0.0122           | level: 0.0145            |
|                      | insurance: 0.0121        | national: 0.0127         |
|                      | population: 0.0114       | service: 0.0094          |
|                      | service: 0.0108          | hospital: 0.0082         |
|                      | universal: 0.0082        | quality: 0.0073          |
|                      | year: 0.0076             | disease: 0.0069          |
|                      | private: 0.0075          | population: 0.0065       |
|                      | access: 0.0067           | program: 0.0062          |
|                      | resource: 0.0061         | system: 0.0062           |
| 2021                 | population: 0.0162       | care: 0.0207             |
|                      | public: 0.0103           | level: 0.0160            |
|                      | system: 0.0087           | system: 0.0156           |
|                      | insurance: 0.0079        | management: 0.0106       |
|                      | access: 0.0079           | national: 0.0105         |
|                      | social: 0.0076           | service: 0.0102          |
|                      | financing: 0.0072        | public: 0.0091           |
|                      | budget: 0.0060           | hospital: 0.0082         |
|                      | medical: 0.0058          | capacity: 0.0070         |
|                      | country: 0.0054          | regional: 0.0061         |

KeyBERT

TextRank\(^6\) and KeyBERT\(^7\) provides keywords (between one and three words), but the latter uses a variation of the BERT algorithm to chose keywords based on the context of the text. Moreover, KeyBERT shows the main keywords with their score. For this algorithm, the five main keywords have been extracted using the same government plans, i.e., 2016 (see Figure 10, Figure 11 and Figure 12) and 2021 (see Figure 13, Figure 14 and Figure 15).

Consistent with what we observed using the TextRank algorithm, the 2016 (see Figure 10, Figure 11 and Figure 12) keywords appeared to be more general. The terms talked about, for example, 'health solidarity', 'health crisis', and 'health inefficient'. However, the KeyBERT algorithm did find other keywords which revealed more concise concepts: 'campaign vaccination', 'equipped hospital' and 'health education'. For one government plan (see Figure 12(c)), all keywords were about diseases: 'dyslipidaemias diabetes', 'diabetes obesity', 'mortality morbidity', 'anemia vulnerable', and 'malaria dengue'. Furthermore, the keywords retrieved with the KeyBERT algorithm also revealed underlying characteristics of the political party. For example (see Figure 11(e)), the keywords of a left-wing political party were 'Peruvian richest', 'subsidise poorest' and 'discriminatory medical'.

As also pointed out with the TextRank algorithm, in 2021 (see Figure 13, Figure 14 and Figure 15) the keywords appeared to be more concrete (e.g., 'health networks', 'Peruvians insured' and 'hospitals needed') and also covered topics related with the COVID-19 pandemic (e.g., 'health pandemic', 'management pandemic' and 'eaths pandemic'). In one government plan (see Figure 14(b)), four out of five keywords were about vaccination: 'vaccines reforming', 'make vaccines', 'vaccines Peru', and 'Peru vaccination'. Similarly, other government plan had four keywords about diseases: 'malnutrition tuberculosis', 'childhood malnutrition', 'anemia chronic' and 'morbidity mortality'.

Rake

For visualisation purposes, we chose the top five key phrases per government plan in 2016 (see Table 3) and 2021 (see Table 4).

Therefore, in 2016, a topic that was recurrent in four government plans was 'universal health coverage'. When the phrases addressed specific diseases/conditions, these were malnutrition, mental health, communicable diseases, and sexually transmitted diseases (see Table 3).

On the other hand, in 2021, 'universal health coverage' was also a frequent topic found in phrases of four government plans. Mental health also appeared often (in three government plans). There were also phrases related to COVID-19, and the fact that the health system in Peru is fragmented (see Table 4).

Discussion

Main findings

We used standard and well-known NLP algorithms to study health chapters of the government plans of the political parties in the 2016 (18 plans) and 2021 (19 plans) general elections in Peru. We did not aim to develop new NLP algorithms nor to modify the available algorithms, rather, we aimed to apply available NLP algorithms in a novel topic: health policy/politics.

The TF-IDF algorithm revealed a similar number of terms between 2016 and 2021. In both years, the term 'health' had the highest frequency. In both years, all the other terms were not specific to a disease, a program or policy, or a new proposal. An analysis of frequency of terms may not be very informative, thus an in-depth analysis with other algorithms may be needed.

The LDA analysis defined two groups: one gathering words signalling things the population would receive (e.g., 'insurance'), and the other with terms about the health system (e.g., 'capacity'). The LDA analysis also assigned government plans to either group. This suggested whether a government plan would focus on delivering goods or services to the population, whereas others would focus on improving the health system.

The TextRank revealed some key words. Interestingly, the keyword phrase 'universal health coverage' were frequent in 2016, but this did not appear in 2021, when keywords about the COVID-19 pandemic (e.g., 'national emergency') and the limited capacities of the health system (e.g., 'low investment') were
Figure 2. Membership of the 2016 government plans to each group classified by the LDA algorithm.
Figure 3. Membership of the 2021 government plans to each group classified by the LDA algorithm.
Figure 4. TextRank results, keywords and scores, for government plans in 2016 (political parties one through six).
Figure 5. TextRank results, keywords and scores, for government plans in 2016 (political parties seven through twelve).
Figure 6. TextRank results, keywords and scores, for government plans in 2016 (political parties thirteen through eighteen).
Figure 7. TextRank results, keywords and scores, for government plans in 2021 (political parties one through six).
Figure 8. TextRank results, keywords and scores, for government plans in 2021 (political parties seven through twelve).
Figure 9. TextRank results, keywords and scores, for government plans in 2021 (political parties thirteen through nineteen).
Figure 10. KeyBERT results, keywords, and scores, for government plans in 2016 (political parties one through six).
Figure 11. KeyBERT results, keywords, and scores, for government plans in 2016 (political parties seven through twelve).
Figure 12. KeyBERT results, keywords, and scores, for government plans in 2016 (political parties thirteen through eighteen).
Figure 13. KeyBERT results, keywords, and scores, for government plans in 2021 (political parties one through six).
Figure 14. KeyBERT results, keywords, and scores, for government plans in 2021 (political parties seven through twelve).
Figure 15. KeyBERT results, keywords, and scores, for government plans in 2021 (political parties thirteen through nineteen).
| Political Party | Keywords Phrases |
|----------------|------------------|
| Acción Popular | 'national agreement child malnutrition', 'health plan agreed upon', '24 hours per day', 'drinking water service coverage', 'universal service coverage' |
| Alianza para el progreso | 'apothecary modules installed inside warehouses', 'multiple minsa health service providers', 'health service providers', 'quality generic medicines', 'promoting universal insurance' |
| Alianza popular | 'drastically reduce chronic child malnutrition', 'remuneration reform', 'generating mechanisms', 'single health information management system', 'expand universal health insurance throughout' |
| Democracia directa | 'culturally appropriate health service delivery system', 'integrated health care based', 'health care model centered', 'public health system', 'traditionally excluded groups' |
| Frente amplio | 'private health system meets required standards regulated', 'mental health centers creation', 'health system deficient mental health system', 'health needs territorialize health care', 'national health information system modernize' |
| Frente esperanza | 'must reach 2 beds per thousand inhabitants', '1000 USD per year per inhabitant', 'subsidising brand name drugs', 'also faces serious problems' |
| Fuerza popular | 'respecting compulsory licensing rules promote access', 'accompanying multidisciplinary community health teams', 'seguro integral de salud' (national health system), 'ensuring better working conditions', 'sustainable development goal 2' |
| Orden | 'mobile health units', 'health must assume full responsibility', 'establishing free mobile health units', '50 million soles', 'rational comprehensive national health plan' |
| Partido humanista peruano | 'preventive health services full coverage', 'healthy person without access', 'free trade agreement signed', 'single patient without care', 'life total health plan' |
| Political Party                  | Keywords Phrases                                                                 |
|--------------------------------|---------------------------------------------------------------------------------|
| Partido nacionalista peruano    | ‘one thousand seven hundred million soles’, ‘deadlines progressive decentralization sustainability surveillance’, ‘approximately four hundred million soles’, ‘universal insurance revolve around coverage’, ‘total peruvian population target 100’ |
| Perú libertario                 | ‘achieving total free public health’, ‘17 billion dollars annually’, ‘abolish discriminatory medical care’, ‘entire educational revolution must’, ‘state must set fees’ |
| Perú nación                     | ‘developing properly equipped hospital centers’, ‘develop public policies aimed’, ‘health services currently provided’, ‘health care bonds’, ‘health care personnel’ |
| Perú patria segura              | ‘health programs communicable disease programs fully strengthened’, ‘economic status marginal populations virtually unattended’, ‘high quality management system inefficient remuneration’, ‘centers managing excellent health services number’, ‘std prevention programs well strengthened’ |
| Perú posible                    | ‘intangible health solidarity fund’, ‘elderly extend mental health care’, ‘mass communication strategies throughout’, ‘more health plan peru’, ‘provide free primary care’ |
| Peruanos por el cambio           | ‘generating three intervention programs per year’, ‘human resources directorate, regional governments’, ‘moving towards universal health coverage’, ‘least 4 annual programs focused’, ‘regional governments strategic guideline’ |
| Progresando Perú                | ‘contemporary scientific medical knowledge’, ‘reforming health care services’, ‘localised health care’, ‘intercultural health program’, ‘high mortality rate’ |
| Siempre unidos                  | ‘achieving one hospital per 500 thousand inhabitants’, ‘hospitalisation significantly improve emergency services encourage’, ‘state must provide budgetary support’, ‘reduce patient waiting time’, ‘current 3rd level hospitals’ |
| Solidaridad nacional -UPP       | ‘25 per 1000 live births’, ‘online id card key’, ‘offer family planning services’, ‘promote healthy living habits’, ‘higher infant mortality rates’ |
Table 4. List of top five Rake key phrases for government plans in 2021.

| Political Party          | Keywords Phrases                                                                 |
|--------------------------|----------------------------------------------------------------------------------|
| Acción Popular           | 'equipped primary care medical posts', 'technologically innovate office services', 'basic health plan appropriate', 'allow solving health problems', 'reduce health care gaps' |
| Alianza para el progreso | 'tier health insurance fund administrative institution', 'several parallel systems creates many problems', 'high cost fund strategic', 'universal insurance system strategic', 'health care centers' |
| APRA                     | 'health professionals per inhabitant according', 'integrated medical information management system', 'articulated territorial assistance network', 'universal digital medical record system', 'articulated territorial assistance networks' |
| Avanza país              | 'latin american average levels', 'health services continues', 'efficient health strategies', 'public health services', 'promote citizen participation' |
| Democracia directa      | 'properly equipped isolation areas within', 'budget program "0131 control"', 'promotes healthy lifestyle habits', 'health sector receives approximately', 'mental health problems attended' |
| Frente amplio           | 'private health system meets required standards', 'comprehensive care obligations using political', 'implementing integrated health networks open', 'national health care plan approved', 'comprehensive health career policy throughout' |
| Frente esperanza        | 'prepare adequately trained health professionals', 'also faces serious problems', 'clearly identify pandemic diseases', 'must reach 2 beds', 'health personnel widely trained' |
| Fuerza popular           | 'subsidized insurance regime must guarantee equal benefits', 'longer build temporary medical care centers', '3 thousand new detected per day', 'approximately one million rapid serological tests', 'advance towards universal social security' |
| Juntos por el Perú       | 'facilitates severe administrative sanctions regulatory framework', 'public system implemented covid sequelae treatment program', 'single public health system implemented national central', 'intersectoral action regional governments assume responsibility', 'operation public pharmacy chains operating' |
| Political Party                     | Keywords Phrases                                                                                                                                                                                                 |
|------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Partido morado                     | ‘reinforce said “integration”’, ‘rapid test occurs around 20 days’, ‘integrity: public networks must talk’, ‘mainly use rapid tests’, ‘regional health directorates must give way’ |
| Partido nacionalista peruano       | ‘200 properly equipped provincial epidemiology offices’, ‘32 certified biosafety level iii laboratories’, ‘people receiving mental health services establish’, ‘private sector must work together’, ‘high complexity national service works’ |
| Partido popular cristiano          | ‘gender approach national teaching articulation system service created resolution act’, ‘competency strengthening programs implemented people development plans approved multidisciplinary health teams’, ‘cost disease fund high cost disease fund created nationally approved’, ‘proven health spending efficiency monitoring system resolution act definition’, ‘adequate working conditions approved health career document resolutive act’ |
| Perú libre                         | ‘would surely end discriminatory medical care’, ‘expanding universal public health coverage’, ‘163 million usd per year’, ‘peruvian legislation currently contemplates’, ‘regional medical resident program’ |
| Perú patria segura                 | ‘std prevention programs well strengthened reducing cases’, ‘goal communicable disease programs fully strengthened’, ‘economic condition marginalised populations practically lacking’, ‘achieve strengthen communicable disease programs’, ‘marginalised populations practically lacking’ |
| Podemos                            | ‘four unrelated subsystems coexist — confusingly —’, ‘health science popular soup kitchens repowering’, ‘minsa promote comprehensive health insurance transfer’, ‘precarious since 379 health establishments’, ‘1000 health establishments nationwide creation’ |
| Renovación Popular                 | ‘national continuous professional qualification module’, ‘national health promotion plan’, ‘health sector create centers’, ‘health sector centers’, ‘inadequate public management’ |
| Somos Perú                         | ‘standardised electronic medical record’, ‘healthy country whose paradigms’, ‘community mental health centers throughout’, ‘health financing must include health promotion’, ‘implement community mental health centers’ |
| Unión por el Perú                  | ‘acute respiratory infections ari would’, ‘200 million new soles’, ‘eliminate chronic child malnutrition’, ‘2020 covid 19 pandemic’, ‘single integrated health system’ |
| Victoria nacional                  | ‘electronic medical records implemented 100%’, ‘yet achieved universal health insurance’, ‘electronic medical records’, ‘effective universal health insurance system’, ‘130 deaths per million’ |
frequent. This provides preliminary evidence on how the main focus of the government plans shifted between 2016 and 2021.

The KeyBERT analysis revealed keywords, but these were learnt based on the context of the text. Interestingly, these keywords provided more insight about the underlying profile of the political party. This suggests that it would be possible to identify characteristics of the writer (in this case of the political party) based on the text and its content.

The Rake analysis provided key phrases, instead of keywords, and the results provided more insight about the documents. For example, this algorithm revealed that phrases ‘universal health coverage’ appeared both in 2016 and 2021. Similarly, phrases with “mental health” were found in both years. Phrases about COVID-19 were only found in 2021.

Interpretation and potential explanations
The first finding was that the dataset with the 2021 government plans was larger than the 2016 dataset. Similarly, the TF-IDF analysis also showed that the number of terms with a frequency of ≥0.10 was 4.8-fold higher in 2021 than in 2016. This could imply that health has gained more relevance over the last five years, particularly since 2020 when the COVID-19 pandemic revealed the limitations and deficiencies of the health systems in Peru.

The TF-IDF analysis showed that most words in 2016 were also found in 2021. This may suggest that the same subjects in 2016 are still valid issues or concerns in 2021. Two words in 2016 that were not repeated in 2021 were ‘private’ and ‘universal’. Probably, issues regarding access to private medical care, or lack of universal health coverage, were less concerning in 2021. On the other hand, words that were in 2021 but not in 2016 included ‘COVID’ and ‘capacity’. The former is self-explanatory, because ‘COVID’ was not a concern in 2016 at all. The term ‘capacity’ could represent the lack of resources, hospitalization beds (both overall and in intensive care units) and oxygen that Peru suffered during the 2021 COVID-19 waves. This subject —lack of capacity— may have driven the health policy agenda, thus gaining relevance in the 2021 TF-IDF analysis.

The LDA analysis provided two groups, and also showed to which group each government plan is most similar. Our hypothesis is that Group 0 included terms or things the population would receive or directly interact with, and Group 1 included terms or things that the healthcare provider would deliver. In 2016, the LDA analysis suggested that most of the government plans would belong to Group 1; this could imply that the 2016 government plans focused on how to improve the services they provide or deliver. In 2021, there were even more government plans highly likely to belong to Group 1, probably because the COVID-19 pandemic made them realise the urgent need for structural changes to improve the health system.

As pointed out before with the LDA algorithm, the TextRank analysis also showed that keywords in 2016 were more general than in 2021, when we also found keywords addressing the COVID-19 pandemic. Interestingly, in 2021 the TextRank analysis also revealed a critique view of the health system at the beginning of the COVID-19 pandemic. This suggests that some government plans were not only about proposing or offering new programs or interventions, but they also evaluated the health system beforehand, ideally to inform their proposals, making these as specific as possible to solve the main problems or limitations of the health system.

Even though TextRank and KeyBERT would deliver keywords, the keywords obtained with KeyBERT gave more context about the government plan or the political party because this algorithm learnt from the context in which the words were embedded. Government plans in which most of the keywords obtained with KeyBERT were about diseases, could suggest that they will focus on these illness, their risk factors and consequences; in other words, the proposals could be disease-specific, perhaps aiming to provide diagnosis and treatment. Similarly, government plans in which many of the keywords were about vaccination -in 2021 presumably regarding COVID-19-, could suggest that their priority would be to get vaccines, and deliver these to the best of their ability; it could also suggest that they will focus —mostly— on the pandemic, while other problems would be addressed in parallel or with less priority. Finally, in accordance with the fact that the KeyBERT algorithm finds keywords based on the context of the text, this algorithm gave insights about the underlying characteristics of the political party. This algorithm could be used to dissect the profile of the political party and how they may address a given issue, above and beyond the frequency of single words and learning from the context. This could be useful to understand the vision of the document and underlying priorities.

The Rake analysis provided key phrases that the unsupervised model could compose from the original texts. Both in 2016 and 2021, a recurrent topic was universal health coverage, signalling that this was a hot topic which, apparently, has not been solved since 2016. In this line, the 2021 phrases addressed the fragmentation of the healthcare system in Peru, where several systems coexist (e.g., private, public, social security, and military forces). This fragmentation has challenged, and sometimes limited, the response to the COVID-19 pandemic. Finally, mental health was addressed both in 2016 and 2021. This implies that the concern about the mental health of the population has risen since 2016. Also, this suggests that the improvement achieved so far still needs further work to secure optimal mental health for the population. Communicable diseases were also mentioned in both 2016 and 2021; however, non-communicable diseases were absent from these phrases, despite the large burden they impose on the population and health systems.

Finally, the contribution of this work lies in the fact that we expanded the data sources frequently used in NLP health research, which has mostly leveraged on texts from electronic health records and social media. Unfortunately, electronic health records may not be available in many low- and middle-income countries, where people may not engage as much in some social media outlets as their peers in high-income countries. Thus, our work may democratize the use of NLP in health research by leveraging on a new data source (e.g., materials...
distributed during national elections) that, arguably, is accessible to a much broader population. Future work, moving from electronic health records and social media, could focus on texts from newspapers and open-access documents published by authorities such as reports from government officers, summaries from government meetings, and summaries of debates at the congress; again, all these sources may be available to a large population above and beyond electronic health records and social media.

Strengths and limitations
We used novel techniques to study the health chapters of the government plans of the 2016 and 2021 Peruvian presidential elections. In so doing, we provided novel insights about these documents to better inform practitioners, researchers, public health experts, and the general population in Peru. This exercise can be replicated in other countries hosting elections this year or soon. We also hope to have sparked interest in NLP techniques to strengthen public health and health policy research, which have lots of data sources which can benefit from NLP analysis.

There are, however, limitations we must acknowledge. First, we focused on the health chapter/section of each government plan which surely contained much of the information for the health sector. Nevertheless, we cannot completely rule out that other chapters could have included additional information about their health-related plans; for example, the chapter/section about economics could have addressed the budget for the health sector or plans of investments. We argued that the most relevant health information must have been included in the health chapters/sections herein analysed, so that arguments outside these chapters would not substantially change our findings or conclusions. Second, we used the government plans as they were, and translated them into English. To the best of our knowledge, the most used NLP algorithms are only available for texts in English. We acknowledge that some words/terms may not have had the ideal translation from Spanish to English through Google Translator; however, this potential limitation affecting single words may have not biased the overall findings and conclusions. We hope that this work sparks interest in the Spanish-speaking scientific community to use NLP into clinical medicine and public health research, while also developing NLP algorithms in Spanish. In addition to more algorithms in Spanish, the scientific community should also secure datasets in Spanish which currently lack to train available and new NLP models. Where possible, these datasets of text in Spanish (and other languages) should be tagged and untagged, as well as annotated or annotated. Third, the analyses and interpretations were mostly data-driven, and should be interpreted in that context. This work was not designed to be a comprehensive political (though NLP is also useful for studying political speech)\textsuperscript{34}, anthropological, or linguistic scrutiny of the 2016 and 2021 government plans in Peru, rather the application of novel artificial intelligence techniques to further expand the understanding of the health chapters in these government plans. Fourth, we did not compare the government plan of the same political party in 2016 and 2021. A prospective easement of the political parties was beyond the scope of this work, and such exercise would not be possible for all parties. Fifth, the text of the health chapters varied in length. The LDA analysis was trained in a dataset with all the government plans together. Government plans with longer text, hence with more words and information, could have driven the algorithm results; conversely, shorter government plans could have provided less information to the LDA algorithm. For transparency, we reported the length of each government plan and the results should be interpreted considering this potential limitation. Sixth, the LDA algorithm was applied with the absolute minimum number of groups (i.e., to identify two clusters). We could have tried to find 5, 10, 20 or more groups; however, dividing the texts in many thin groups would have not been ideal. First, we only worked with the health chapters which were, on average, one or two pages long; in other words, these chapter did not have much content to be dissected into many clusters. Had we analysed the full plans (not only the health chapters), finding more groups would have helped to separate all sections. Second, the government plans were concise straight-to-the-point documents (sometime bullet points); that is, they did not cover several topics nor in detail. This made it difficult to dissect the texts into many thin groups.

Conclusion
This NLP analysis of the health chapters of government plans for the general elections in Peru in 2016 and 2021, showed that NLP are a useful tool to dissect these documents in terms of keywords and phrases based on frequency and context. The NLP analysis could inform about the underlying priorities or main subjects addressed by each government plan, while also revealing the profile of the political parties. NLP analysis could also be included in the research of health policies and politics during general elections, and could provide informative summaries for the general population.

Data availability
Underlying data
Figshare: Data - Government plans in the 2016 and 2021 Peruvian presidential elections: A natural language processing analysis of the health chapters. https://doi.org/10.6084/m9.figshare.14466699.v2.

The project contains the following underlying data:

- planesSalud 2016.csv (dataset where each row is a sentence of the health chapter in each government plan. For example, if the government plan ‘ABC’ has a health chapter with 15 sentences, there will be 15 rows (each containing a sentence))

- planesSalud 2021.csv (dataset where each row is a sentence of the health chapter in each government plan. For example, if the government plan ‘ABC’ has a health chapter with 15 sentences, there will be 15 rows (each containing a sentence))

- planesSalud Comp 2016.csv (dataset where each row is a health chapter in each government plan. There
will be as many rows as government plans, each row containing the health chapter in full)

• planesSalud Comp 2021.csv (dataset where each row is a health chapter in each government plan. There will be as many rows as government plans, each row containing the health chapter in full)

• Jupyter Notebook with analysis code in Python

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Contributions
RMC-L conceived the idea with support from JLM and MC-C. RMC-L built the datasets. JLM and MC-C conducted the analyses. All authors wrote de manuscript. All authors approved the submitted version.

References

1. Kersloot MG, van Putten FJP, Abu-Hanna A, et al.: Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. J Biomed Semantics. 2020; 11(1): 14.
  Published Abstract | Publisher Full Text | Free Full Text
2. Melia JA, Basta MN, Toyoda Y, et al.: Natural language processing in surgery: A systematic review and meta-analysis. Ann Surg. 2021; 273(3): 500–508.
  Published Abstract | Publisher Full Text
3. Mahmoudi E, Kamdar N, Kim N, et al.: Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. BMJ. 2020; 369: m5958.
  Published Abstract | Publisher Full Text | Free Full Text
4. O’Connor B, Stewart BM, Smith NA: A systematic review of natural language processing in radiology: a systematic review. Radiology. 2016; 279(2): 329–343.
  Published Abstract | Publisher Full Text | Free Full Text
5. Mahmoudi E, Kamdar N, Kim N, et al.: Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc. 2019; 26(4): 364–379.
  Published Abstract | Publisher Full Text | Free Full Text
6. Dreisbach C, Koleck TA, Bourne PE, et al.: A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform. 2019; 125: 37–46.
  Published Abstract | Publisher Full Text | Free Full Text
7. Sheikhalishahi S, Miotto R, Dudley JT, et al.: Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform. 2019; 7(2): e12239.
  Published Abstract | Publisher Full Text | Free Full Text
8. O’Connor B, Stewart BM; Smith NA: Learning to extract international relations from political context. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013; 1094–1104.
  Reference Source
9. Leeson W, Resnick A, Alexander D, et al.: Natural Language Processing (NLP) in Qualitative Public Health Research: A Proof of Concept Study. Int J Qual Methods. 2019; 18.
  Published Full Text
10. Alag S: Analysis of covid-19 clinical trials: A data-driven, ontology-based, and natural language processing approach. PLoS One. 2020; 15(6): e0239694.
  Published Abstract | Publisher Full Text | Free Full Text
11. Ambalavanan AK, Devakonda MV: Using the contextual language model BERT for multi-criteria classification of scientific articles. J Biomed Semantics. 2020; 112: 103578.
  Published Abstract | Publisher Full Text
12. Beltagy I, Lo K, Cohan A: Scibert: Pretrained language model for scientific text. Association for Computational Linguistics. EMNLP, 2019; 3615–3620.
  Publisher Full Text
13. Sharma R, Li Y: Self-supervised contextual keyword and keyphrase retrieval with self-labelling. 2019.
  Publisher Full Text
14. Campos R, Mangaravite V, Pasquali A, et al.: Yake! keyword extraction from single documents using multiple local features. Inform Sciences. 2020; 509: 257–289.
  Publisher Full Text
15. Rose S, Engel D, Cranmer N, et al.: Automatic keyword extraction from individual documents. Text Mining: Applications and Theory. 2019; 1: 1–20.
  Published Full Text
16. Robertson S: Understanding inverse document frequency: on theoretical arguments for idf. J Doc. 2004; 60(5): 503–520.
  Publisher Full Text
17. Blei DM, Ng AV, Jordan MI: Latent dirichlet allocation. J Mach Learn Res. 2003; 3: 993–1022.
  Reference Source
18. Mihalcea R, Tarau P: Textrank: Bringing order into text. 2004; 404-411.
  Reference Source
19. Devlin J, Chang MW, Lee K, et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805. 2018.
  Reference Source
20. Berry MW, Kogan J: Text mining: Applications and Theory. West Sussex, PO1985Q, UK: John Wiley & Sons, 2010.
  Reference Source
21. The World Bank: Data. 2021; [Online; accessed 28-march-2021].
  Reference Source
22. EduGestores: Elecciones 2016 – Planes y propuestas de los candidatos. 2021; [Online; accessed 28-march-2021].
  Reference Source
23. El Peruano: Elecciones 2021: mira aquí los planes de gobierno de los candidatos a la presidencia. 2021; [Online; accessed 28-march-2021].
  Reference Source
24. Yoav G, Graeme H: Neural network methods in natural language processing. Morgan & Claypool: San Rafael, 5R, USA, 2017; 104-113.
  Publisher Full Text
25. Scikit-Learn: TfidfVectorizer. 2021; [Online; accessed 28-march-2021].
  Reference Source
26. Bird S, Klein E, Loper E: Natural language processing with Python: analyzing text with the natural language toolkit. "O’Reilly Media, Inc.", 2009.
  Reference Source
27. Scikit-Learn: Latent Dirichlet Allocation. 2021; [Online; accessed 28-march-2021].
  Reference Source
28. Rehurek R, Sojka P: Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC2010 Work shop on New Challenges for NLP Frameworks. Valletta, Malta, ELRA, 2010; 45-50.
  Reference Source
29. Zhang T, Kishore V, Wu F, et al.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv: 1904.09675. 2019.
  Publisher Full Text
30. Smit A, Jain S, Rajpurkar P, et al.: Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. arXiv preprint arXiv: 2004.09167. 2020.
  Publisher Full Text
31. Keyword Extraction with BERT: KeyBERT. 2021; [Online; accessed 28-march-2021].
  Reference Source
32. Alcalde-Rabanal JE, Lazo-González O, Nigenda G: [The health system of peru]. Salud Publica Mex. 2011; 53 Suppl 2: s43-54.
  PubMed Abstract
33. GBD 2019 Diseases and Injuries Collaborators: Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020; 396(10258): 1204-1222.
  PubMed Abstract | Publisher Full Text | Free Full Text
34. Katre PD: NLP Based Text Analytics and Visualization of Political Speeches. International Journal of Recent Technology and Engineering. 2019; 8(3): 8574–8579.
  Publisher Full Text
Open Peer Review

Current Peer Review Status: ✔ ? ? ? ?

Version 5

Reviewer Report 28 January 2025

https://doi.org/10.21956/wellcomeopenres.20483.r115207

© 2025 Khatter K. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Kiran Khatter  
BML Munjal University, Gurgaon, India

This paper presents a study on the application of Natural Language Processing (NLP) algorithms to analyze the health chapters of government plans from the 2016 and 2021 Peruvian presidential elections. The study addresses a gap in the use of NLP for health policy analysis, moving beyond traditional applications like electronic health records or social media analysis. This expands the scope of NLP applications in public health. The study compares different NLP algorithms to extract and analyze keywords and phrases, aiming to highlight underlying themes and priorities in health policy during these election cycles.

- The discussion on the practical applications of the findings is brief. More concrete examples of how the insights could inform health policy, public health strategies, or electoral decision-making would strengthen the paper's impact.
- The use of Google Translate for data translation could introduce biases. This limitation should be addressed with more detail in the paper, including potential effects on findings and how these were mitigated.
- The paper lacks sufficient detail on the preprocessing steps, feature extraction, and validation for the datasets used. Clear descriptions of these processes would enhance reproducibility.
- The rationale for selecting the specific NLP algorithms is underexplored. Including a justification for their suitability over alternative approaches and a discussion of their comparative effectiveness would add depth.
- The paper acknowledges limitations but could provide a more detailed exploration, including the impact of dataset size variability, algorithmic constraints (e.g., LDA's fixed groupings), and generalizability to other countries or contexts.
- The qualitative implications of findings, such as insights into political party priorities or public health challenges, could be expanded. Incorporating more context-based analysis would make the results more impactful.
- The writing could be improved for clarity and conciseness. Some sections are overly technical and may not be easily accessible to non-specialist readers. Defining key terms and providing examples would improve comprehension.
Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Image Processing, NLP, Blockchain, Fuzzy/Neutrosophic sets

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 10 January 2025

https://doi.org/10.21956/wellcomeopenres.20483.r116415

© 2025 Arisal A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Andria Arisal
National Research and Innovation Agency, Jakarta, Indonesia

This manuscript effectively demonstrates how NLP techniques can be applied to extract important healthcare-related topics in political documents, aiding the public in comparing priority topics across political parties. The analysis is comprehensive and well-executed.

However, the reviewer suggests the following improvements to enhance the clarity, methodological rigor, and impact of the manuscript:
1. Move the discussion of the manuscript’s contribution from the Discussion section to the Introduction to better frame the study’s significance upfront.
2. Clearly explain the differences between TextRank and KeyBERT in the methods section and
provide an argument for using one over the other. Or, if the authors want to use both, include a head-to-head comparison (e.g., in a single side-by-side figure) to better contextualize their results. Avoid spreading their results across separate sets of figures (e.g., Figures 4-9 for TextRank and Figures 10-15 for KeyBERT).

3. Remove “health*” as a keyword from TextRank and KeyBERT analyses, as it is a ubiquitous term in the dataset and does not provide additional insight into distinguishing topics.

4. The current number of bar charts is overwhelming and diverts the reader’s attention from the core discussion. Consider more integrative and comparative visualization techniques, such as:
   - Word clustering with intersections among political parties to highlight shared priorities.
   - LDAvis for visualizing group classifications in LDA, offering a more intuitive understanding of topic distributions.
   - Combine Figure 1 into a grouped bar chart to display the 2016 and 2021 data series together.
   - Present a grouped bar chart showing terms with the highest aggregate TextRank or KeyBERT scores across political parties to emphasize shared interests.

6. The authors chose two clusters for the LDA analysis but did not justify this decision. Including coherence scores or perplexity metrics to determine the optimal number of clusters would improve methodological rigor.

If possible the authors could also:

1. Add statistical tests to validate observed trends and differences between 2016 and 2021, such as tests for significant differences in keyword frequencies or topic distributions across years.
2. The imbalance in dataset size between 2016 and 2021 likely influences results. Normalize keyword frequencies and clustering probabilities to account for differences in the volume of data.
3. The discussion could benefit from more actionable insights. For example, explain how these findings could influence public health policy or decision-making in Peru and other similar contexts.
4. Address ethical considerations around using government plans for NLP analysis, such as potential implications of misinterpretation or misrepresentation of political priorities.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.
Reviewer Expertise: Data Analytics, Natural Language Processing, High Performance Computing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 07 January 2025

https://doi.org/10.21956/wellcomeopenres.20483.r113436

© 2025 Raza S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Shaina Raza
Vector Institute, Toronto, Ontario, Canada

The paper presents an important problem, however, I would like authors to mention where is there contributions, is it methods or problem they solved?
I see methods are quite simplistic? is it planned? Can they include some analysis through named entity recognition as in recent clinal NLP and NER methods, or they can use some LLMs (free ones) to get some insights to make it better.
It depends what authors are trying to convey, their computational method or solving health NLP.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: NLP, LLMs
I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

淑女Ayoub Bagheri
Utrecht University, Utrecht, The Netherlands

Hereby I am submitting my reviews on the paper entitled “Government plans in the 2016 and 2021 Peruvian presidential elections: A natural language processing analysis of the health chapters.” The paper is well written and while the topic is of much interest, there are some serious concerns regarding the methods, analysis, and experimental design of the paper. Below are comments in detail:

○ Please provide a link to your toolkit as it is mentioned that it is available to use for other researchers.

○ Could you elaborate on the contributions of this work?

○ I would like to know why the authors choose these methods and how do they compare them with each other.

○ Is it an improvement if you are using a lemmatisation method in TF-IDF and LDA?

○ What is the reasoning when you cluster data into only two groups in LDA? It would be nice if you look at 5, 10, 20, and 50 as well.

○ Could you explain in the paper what is the input to the LDA model?

○ What are the limitations of using KeyBERT? Do you have to use fixed vectors?

○ Could you provide the equation for the rake method and describe what was the library you used for this?

○ If you look at the frequencies, words similar to 'health' should appear in most of the documents, therefore with TF-IDF they should gain a very low weight. How can you respond to this argument? Isn’t ‘health’ a stopword in this analysis?
In Table 2, would you provide the probabilities for each term in groups as well?

How do you compare Fig. 2b and Fig. 3b?

The keywords for Fig. 4c are not clear.

From page 8 until 26 you just see figures and tables with no text explaining them.

The first point in the discussion is about applying novel techniques in this study. I must say that I do not agree with the authors as I do not see any novel method in the paper. The paper does not propose any novelty (that, of course, is not the focus of the study and the journal), while except KeyBERT (2020), the other methods are quite used in many studies: TF-IDF (1972), LDA (2003), Rake (2010), textrank (2004).

Simply because you are looking at the health chapters, TF-IDF should remove the term ‘health’ from the analysis, unless you say that many of the documents do not contain the word health.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Natural Language Processing, Data Science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
Rodrigo M Carrillo-Larco

We are thankful for the insightful comments by the reviewer. We appreciate the time he invested in reading our work. Please see our point-by-point answers below. We hope these satisfy the reviewer's expectations.

C1. Please provide a link to your toolkit as it is mentioned that it is available to use for other researchers.
R1. We have uploaded the Jupyter Notebooks for these analyses to the Figshare repository reported in the Data Availability statement, which was updated accordingly. The sentence in the introduction about the ‘toolkit’ was removed.

C2. Could you elaborate on the contributions of this work?
R2. Yes, the contribution of this work has been further elaborated in the Introduction and Discussion sections. The new lines read:

[introduction] Our work contributes to the NLP health literature by leveraging on standard and well-known algorithms to analyse health policy/politics documents. Because most previous NLP health research focused on electronic health records (i.e., clinical oriented) or social media (i.e., population health oriented), our work expanded the use of NLP in health research to new documents (policy/politics documents) with potential public health implications. Furthermore, another contribution of our work also lies in the fact that we used a sort of text documents that, arguably, may be available to a much broader audience. Even though electronic health records and social media are valuable sources of information for NLP analyses, electronic health records may be seldom available in low- and middle-income countries where people could also engage less in social media. Nevertheless, in most countries during election period there may be some sort of documents summarizing the proposals by the candidates who may share these as much as possible for electors to familiarize themselves with the candidates.

[discussion] Finally, the contribution of this work lies in the fact that we expanded the data sources frequently used in NLP health research, which has mostly leveraged on texts from electronic health records and social media. Unfortunately, electronic health records may not be available in many low- and middle-income countries, where people may not engage as much in some social media outlets as their peers in high-income countries. Thus, our work may democratize the use of NLP in health research by leveraging on a new data source (e.g., materials distributed during national elections) that, arguably, is accessible to a much broader population. Future work, moving from electronic health records and social media, could focus on texts from newspapers and open-access documents published by authorities such as reports from government officers, summaries from government meetings, and summaries of debates at the congress; again, all these sources may be available to a large population above and beyond electronic health records and social media.

C3. I would like to know why the authors choose these methods and how do they compare them with each other.
R3. We have provided this information in the Methods section, right before the description of the algorithms. The new lines read:
We used five NLP algorithms: TF-IDF, LDA, TextRank, KeyBERT, and Rake. We chose these algorithms because they are frequently used in the NLP literature, they are available through most libraries in Python and R facilitating the replication and outreach of our work, and they satisfied our goals (i.e., to describe the health chapters in government plans by summarising words and terms).

TF-IDF and TextRank reveals keywords; TF-IDF does so based on the relative frequency of each term whereas TextRank selects keywords based on a trained model. Compared to TF-IDF and TextRank, KeyBERT as well as Rake reveal keywords or short phrases (e.g., 3-5 words) based on trained models. LDA, on the other hand, groups texts into unknown clusters and the user labels such clusters (i.e., as in unsupervised machine learning with K-means). Altogether, these algorithms provide a framework to describe the health chapters in government plans going from unique terms (TF-IDF and TextRank), to short phrases (KeyBERT and Rake), and grouping into unknown clusters (LDA).

C4. Is it an improvement if you are using a lemmatisation method in TF-IDF and LDA?
R4. We followed the standard recommendations for TF-IDF and LDA as shown in references 25, 26, and 27 in the manuscript. For both algorithms, lemmatisation was suggested before implementing them.

C5. What is the reasoning when you cluster data into only two groups in LDA? It would be nice if you look at 5, 10, 20, and 50 as well.
R5. We have justified our choice of two (02) groups in LDA. The new lines read (new text underlined):

For this work, we defined two groups (i.e., we specified that the LDA algorithm should create two groups following an unsupervised approach from the data). Two is the absolute minimum number of groups and we chose such option this being an exploratory work compounded with other NLP algorithms, and we did not focus solely on the LDA results. For a discussion on this limitation, please refer to the Strengths and Limitations section.

We also discussed this decision, specifically in Strengths and Limitations section:

Sixth, the LDA algorithm was applied with the absolute minimum number of groups (i.e., to identify two clusters). We could have tried to find 5, 10, 20 or more groups; however, dividing the texts in many thin groups would have not been ideal. First, we only worked with the health chapters which were, on average, one or two pages long; in other words, these chapters did not have much content to be dissected into many clusters. Had we analysed the full plans (not only the health chapters), finding more groups would have helped to separate all sections. Second, the government plans were concise straight-to-the-point documents (sometime bullet points); that is, they did not cover several topics nor in detail. This made it difficult to dissect the texts into many thin groups.

C6. Could you explain in the paper what is the input to the LDA model?
R6. We added these lines:

The input for training the LDA algorithm was all the government plans (separately for each year),
having removed the stop words and lemmatized the terms. Then, with the cleaned (lemmatized and without stop words) text for each political party, the model assigns each party to a group defined during the LDA training.

C7. What are the limitations of using KeyBERT? Do you have to use fixed vectors?  
R7. The question was not very clear to us. We apologise unreservedly and hope this answer addresses the issue. We added these lines in the Methods section where we described the KeyBERT analysis:

A limitation of this algorithm is that it retrieves a maximum of six keywords; though this did not affect our work because we sought five keywords as it was the case for other algorithms.

C8. Could you provide the equation for the rake method and describe what was the library you used for this?  
R8. We added these lines:

We used the class Rake from the rake_nltk library; for further details including the equation for the Rake method please refer to reference 15.

C9. If you look at the frequencies, words similar to ‘health’ should appear in most of the documents, therefore with TF-IDF they should gain a very low weight. How can you respond to this argument? Isn’t ‘health’ a stopword in this analysis?  
R9. We have included the term ‘health’ as a stop word (i.e., excluded the term ‘health’ from the TF-IDF analysis). The TF-IDF figures and results have been updated accordingly. Notably, the main results did not change.

Following this recommendation, and because TF-IDF has the same pre-processing as the LDA algorithm, the term ‘health’ was also excluded from the LDA analysis. We updated the results and table accordingly.

The term ‘health’ was not removed from the other analyses because they use the semantic context of the text and removing a word would go against this recommendation. Moreover, removing the word ‘health’ could change some relevant key words or short phrases such as ‘public health’, ‘universal health’, or ‘health insurance’.

C10. In Table 2, would you provide the probabilities for each term in groups as well?  
R10. Table 2 now includes the probabilities for each of the ten terms. Please, keep in mind that the terms have changed as well (not the same as in the original submission), because the LDA analysis was re-ran having removed the word ‘health’ (i.e., ‘health’ is now a stop word).

C11. How do you compare Fig. 2b and Fig. 3b?  
R11. We have included these lines in the Results section, at the end of the LDA sub-heading.

Figure 2 shows the LDA results for the government plans in 2016 and Figure 3 for the plans in 2021. While these figures show that political parties are more inclined to either Group 0 or Group 1, a direct comparison between 2016 and 2021 is troublesome because: i) the same exact political
parties did not participate in the two elections; ii) the composition of the political parties in terms of roles, interests and priorities may have changed between 2016 and 2021; and iii) amidst the COVID-19 pandemic, the focus of the health-related government plans would be different between 2016 and 2021. Perhaps in response to the COVID-19 pandemic, the difference observed between Figure 2B and Figure 3B could signal that the 2021 plans focused more on terms related to the structure of the health system (i.e., Group 1 appeared to cluster terms that are related with the structure of the health system).

C12. The keywords for Fig. 4c are not clear.
R12. Unfortunately, we could not customize the very first label for Figure 4C. The very first label was a three-word term which font/size we could not customize without compromising the other labels. Even though we tried different methods to change that label only, we could not find an optimal solution. We apologise for the inconvenience. Happily, the message and interpretation of Figure 4C was not altered.

C13. From page 8 until 26 you just see figures and tables with no text explaining them.
R13. The text explaining or describing the figures and tables are just before the pages containing only figures and table. This was the layout decided by the journal's typesetters for the PDF file. However, the HTML version of the manuscript has the figures and tables embedded in the text. We apologise for the inconvenience and we will work with the editors to improve the layout of the PDF version.

C14. The first point in the discussion is about applying novel techniques in this study. I must say that I do not agree with the authors as I do not see any novel method in the paper. The paper does not propose any novelty (that, of course, is not the focus of the study and the journal), while except KeyBERT (2020), the other methods are quite used in many studies: TF-IDF (1972), LDA (2003), Rake (2010), textrank (2004).
R14. Indeed. We appreciate the discrepancy and the opportunity to clarify. We modified the very first sentence in the Discussion (new text underlined):

_We used standard and well-known NLP algorithms to study health chapters of the government plans of the political parties in the 2016 (18 plans) and 2021 (19 plans) general elections in Peru. We did not aim to develop new NLP algorithms nor to modify the available algorithms, rather, we aimed to apply available NLP algorithms in a novel topic: health policy/politics._

C15. Simply because you are looking at the health chapters, TF-IDF should remove the term 'health' from the analysis, unless you say that many of the documents do not contain the word health.
R15. The word ‘health’ was removed from the TF-IDF and LDA analyses (i.e., it was set as a stop word). The results, figures, and tables for the TF-IDF and LDA analyses were updated accordingly.

**Competing Interests:** None
Thank you for the improvements. The article now is fit to be indexed.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

**Competing Interests:** I taught a course on parallel algorithms at the same department and same University as the authors from around 2010 to 2015. Mr. Lovon-Melgarejo was an undergraduate student and Mr. Castillo-Cara was a member of the teaching staff.

**Reviewer Expertise:** AI, machine learning, data mining, text mining

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
In the last review, I asked the authors to exclude some words from the TF-IDF analysis, and update figure 1. The authors did it. But the authors are using the previous analysis (analysis in the previous version without words excluded) in the paragraph starting with "The TF-IDF analysis suggested different underlying approaches between the 2016 and 2021 government plans. Our hypothesis is that the 2016 plans were more general, because frequent terms were 'agreed', 'agreement', 'political' and 'plan'; overall, these terms would suggest a non-specific approach. "...

Please review all the text (not only that paragraph) in order to assure that you are not presenting the results of the old analysis.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: I taught a course on parallel algorithms at the same department and same University as the authors from around 2010 to 2015. Mr. Lovon-Melgarejo was an undergraduate student and Mr. Castillo-Cara was a member of the teaching staff.

Reviewer Expertise: AI, machine learning, data mining, text mining

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have
significant reservations, as outlined above.

Author Response 13 Jan 2022

Rodrigo M Carrillo-Larco

Thank you very much for pointing this out. Two paragraphs in the discussion have been modified accordingly.

**Competing Interests:** No competing interests were disclosed.

---

**Version 2**

Reviewer Report 07 December 2021

https://doi.org/10.21956/wellcomeopenres.19245.r47104

© 2021 Rodriguez Rafael G. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Glen Dario Rodriguez Rafael

Unidad de Postgrado de la Facultad de Ingeniería de Informática y Sistemas, Universidad Nacional Mayor de San Marcos, Lima, Peru

I thank the author for their efforts in improving this paper. I have only one observation standing: Regarding Q2 of first review, "change"; and "execute" and "force" have not been excluded in Figures 1(b) and 1(a), so I am not sure you excluded them for both years. I mentioned 3 words as examples of filler words, but there are more words without intrinsic “policy” meaning: based, force, agreed, execute, addressed, exempting, promote. A filler word would be any common adjective or verb (based, agreed, addressed are adjective, execute, promote, exempt-ing are verbs) that can be used for many purposes; and some nouns (force).

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Yes

**Competing Interests:** I taught a course on parallel algorithms at the same department and same University as the authors from around 2010 to 2015. Mr. Lovon-Melgarejo was an undergraduate student and Mr. Castillo-Cara was a member of the teaching staff.

**Reviewer Expertise:** AI, machine learning, data mining, text mining

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

**Author Response 09 Dec 2021**

Rodrigo M Carrillo-Larco

We appreciate the positive feedback by the reviewer. We apologise for the confusion when uploading and publishing the updated figures. We have excluded the following words: based, force, agreed, execute, addressed, exempting, promote lower, bio, isc, expected and trend. The text and figures have been updated accordingly.

**Competing Interests:** None.

---

**Version 1**

Reviewer Report 11 November 2021

https://doi.org/10.21956/wellcomeopenres.18605.r46096

© 2021 Rodriguez Rafael G. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Glen Dario Rodriguez Rafael

Unidad de Postgrado de la Facultad de Ingeniería de Informática y Sistemas, Universidad Nacional Mayor de San Marcos, Lima, Peru

I think this article is an interesting exploration on the topic of policy analysis using NLP. However, there are some issues that need to be addressed:
Regarding the statistical analysis, the text of the health chapters has a large variance in length and, for many political parties, is quite small. There is no discussion of how this fact could affect the conclusions drawn using NLP.

TF-IDF as used here is of scarce value, because "filler words" were not excluded (example: execute, force, change). These are no traditional filler words, but in the context of political discourse, they are filler words.

LDA analysis is interesting and worthy.

Discussion of KeyBERT analysis is too brief. Please discuss further.

The problem with NLP for Spanish text is not the lack of algorithms, it is the lack of datasets (text corpus, both POS tagged and untagged, annotated, etc.)

A couple of references that would be useful to add:
Leeson, W., Resnick, A., Alexander, D., & Rovers, J. (2019). Natural Language Processing (NLP) in qualitative public health research: a proof of concept study. International Journal of Qualitative Methods, 18.
Katre, P. D. NLP Based Text Analytics and Visualization of Political Speeches. International Journal of Recent Technology and Engineering 8(3):8574-8579

References
1. Leeson W, Resnick A, Alexander D, Rovers J: Natural Language Processing (NLP) in Qualitative Public Health Research: A Proof of Concept Study. International Journal of Qualitative Methods. 2019; 18. Publisher Full Text
2. Katre PD: NLP Based Text Analytics and Visualization ofPolitical Speeches. International Journal of Recent Technology and Engineering. 2019; 8 (3). Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature? Yes

Is the study design appropriate and is the work technically sound? Yes

Are sufficient details of methods and analysis provided to allow replication by others? Yes

If applicable, is the statistical analysis and its interpretation appropriate? Partly

Are all the source data underlying the results available to ensure full reproducibility? Yes

Are the conclusions drawn adequately supported by the results? Partly
**Competing Interests:** I taught a course on parallel algorithms at the same department and same University as the authors from around 2010 to 2015. Mr. Lovon-Melgarejo was an undergraduate student and Mr. Castillo-Cara was a member of the teaching staff.

**Reviewer Expertise:** AI, machine learning, data mining, text mining

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

**Author Response 13 Nov 2021**

Rodrigo M Carrillo-Larco

Q1. Regarding the statistical analysis, the text of the health chapters has a large variance in length and, for many political parties, is quite small. There is no discussion of how this fact could affect the conclusions drawn using NLP.

R1. This has been discussed in the limitations section: “Fifth, the text of the health chapters varied in length. The LDA analysis was trained in a dataset with all the government plans together. Government plans with longer text, hence with more words and information, could have driven the algorithm results; conversely, short government plans could have provided less information to the LDA algorithm. For transparency, we reported the length of each government plan and the results should be interpreted considering this potential limitation.”

Q2. TF-IDF as used here is of scarce value, because “filler words” were not excluded (example: execute, force, change). These are no traditional filler words, but in the context of political discourse, they are filler words.

R2. These “filler words” (execute, force and change) were excluded from the analysis. The methods section has been updated accordingly: “The following filler words were excluded: execute, force and change; although these are not traditional filler words, in the political context they are filler words.” Also, the TF-IDF results were updated and so were the figures (figure 1).

Q3. LDA analysis is interesting and worthy.

R3. We appreciate the positive feedback.

Q4. Discussion of KeyBERT analysis is too brief. Please discuss further.

R4. In the discussion, in the sub-section ‘Interpretation and potential explanations’, the second-to-last paragraph has been modified to discuss further.

Q5. The problem with NLP for Spanish text is not the lack of algorithms, it is the lack of datasets (text corpus, both POS tagged and untagged, annotated, etc.)

R5. This has been included in the limitations section: “In addition to more algorithms in
Spanish, the scientific community should also secure datasets in Spanish which currently lack to train available and new NLP models. Where possible, these datasets of text in Spanish (and other languages) should be tagged and untagged, as well as annotated or annotated.

Q6. A couple of references that would be useful to add:
Leeson, W., Resnick, A., Alexander, D., & Rovers, J. (2019). Natural Language Processing (NLP) in qualitative public health research: a proof of concept study. International Journal of Qualitative Methods, 18.1
Katre, P. D. NLP Based Text Analytics and Visualization of Political Speeches. International Journal of Recent Technology and Engineering 8(3):8574-8579

R6. These two references were included. The first reference was included at the end of this sentence, next to a reference we already had: “On the other hand, public health research appears not to have benefited from NLP algorithms, despite the fact that this research field also has large data sources of text, such as reports and policy briefs.” The second reference was included at the end of this new sentence in the limitations section: “This work was not designed to be a comprehensive political (though NLP is also useful for studying political speech) ...

Competing Interests: None.