



# Classification of Poverty Condition Using Natural Language Processing

Guberney Muñetón-Santa<sup>1,2</sup> · Daniel Escobar-Grisales<sup>1</sup> ·  
Felipe Orlando López-Pabón<sup>1</sup> · Paula Andrea Pérez-Toro<sup>1,3</sup> ·  
Juan Rafael Orozco-Arroyave<sup>1,3</sup>

Accepted: 18 January 2022 / Published online: 8 February 2022  
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

## Abstract

This work introduces a methodology to classify between poor and extremely poor people through Natural Language Processing. The approach serves as a baseline to understand and classify poverty through the people's discourses using machine learning algorithms. Based on classical and modern word vector representations we propose two strategies for document level representations: (1) document-level features based on the concatenation of descriptive statistics and (2) Gaussian mixture models. Three classification methods are systematically evaluated: Support Vector Machines, Random Forest, and Extreme Gradient Boosting. The fourth best experiments yielded around 55% of accuracy, while the embeddings based on GloVe word vectors yielded a sensitivity of 79.6% which could be of great interest for the public policy makers to accurately find people who need to be prioritized in social programs.

**Keywords** Poverty · Natural language processing · Text classification · Word embedding · Document-level embedding · Machine learning

## 1 Introduction

Poverty is the oldest research topic in Economics, and it is the core of almost every public policy Ravallion (2015). The United Nations in the 2015 assembly defined 17 goals to be fulfilled by 2030; the first one consists in ending poverty in all of its forms worldwide. Specifically, the goal 1.2 aims to reduce at least in one half the proportion of men, women, and children at all ages living in poverty conditions. The idea of understanding the poverty conditions in all of its dimensions assumes that poverty should be characterized as multiple deprivations of human beings. The traditional way to assess poverty is exclusively based

---

✉ Guberney Muñetón-Santa  
guberney.muneton@udea.edu.co

<sup>1</sup> GITA Lab. Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia

<sup>2</sup> Instituto de Estudios Regionales, Universidad de Antioquia, Medellín, Colombia

<sup>3</sup> Pattern Recognition Lab., Friedrich Alexander-Universität, Erlangen-Nürnberg, Erlangen, Germany

on the monetary domain, but this approach has changed in the last years. The multidimensional nature of poverty is accepted by academic and public policy makers Sen (1999), Alkire et al. (2015), Naraya et al. (2000), Nolan and Whelan (2011), PNUD (sps49), World Bank (sps72). However, the question about which dimensions are the most informative to describe poverty may vary depending on the territory, and despite its high importance, there is no consensus in this concept Villatoro and Santos (2019). Besides, how to measure the dimensions “chosen” by governments to help their people is also a matter of debate. Further, the measurement itself is a research problem because multiple indexes have to be considered. In this regard, different poverty measures come from different evaluations that consider different human realities Laderchi et al. (2003). Notwithstanding the lack of accurate and specific measures to evaluate poverty, it is excellent news to make progress in uncovering missing dimensions that affect people the most and to show that traditional approaches have left aside relevant aspects regarding the poverty condition Alkire (2007), Biggeri and Santi (2012).

Giving voice to the marginalized people democratizes the creation of social knowledge about poverty. It improves the knowledge to make better informed public policy decisions. Participatory studies and on-field research evidence poverty as a multi-dimensional problem Narayan et al. (1999). Listening to poor people is one of the main methods to understand their reasoning to make decisions Banerjee et al. (2011), Nussbaum (2001). Nevertheless, the perspective of privileging the voices of the poor people entails a methodological challenge in obtaining, processing, and analyzing text data.

Although social media provides “easy accessible” data that would help in understanding the experience of poor people Caplan et al. (2017), a significant limitation with such data is that we are unsure about the real social conditions of those who post. Most of the available data to train language models come from the interaction between Internet users, mainly young people from rich countries. Existing language models reproduce particular points of view mainly valid in the rich parts of the world due to possible bias in the Internet users and the selected texts Bender et al. (2021), Salvatore et al. (2020), Prabhakaran et al. (2019). A similar phenomenon occurs when the source of information to understand poverty is the mass media Chiquito et al. (2019). This is mainly because the person who delivers the discourse or the idea is not poor, so the language is adapted and the ideas are refined to make them appropriate for the target market of a given news agency. Only a few times it is directly the poor person who gives the discourse. Thus, marginalized people are under-represented in the databases used to train language models Bender et al. (2021). Although the social media data have advantages to complement socioeconomic indicators, it requires special attention to resolve critical issues about its quality and representativeness of the target population Salvatore et al. (2020).

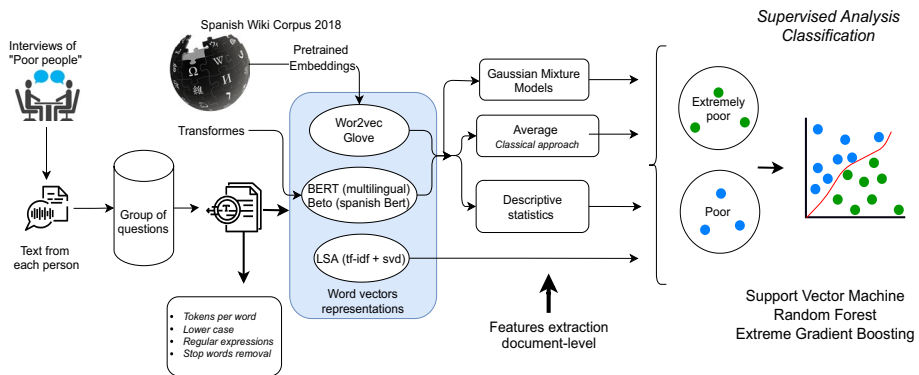
In the few cases where data are available, there is a specific problem due to their high dimensionality and lack of structure Caplan et al. (2017). Word vector representations are prominent methods in social science that have helped in understanding different cultural and linguistic aspects of the discourse Kozłowski et al. (2019). Computer science, Natural Language Processing (NLP), and advanced computational power, enabled the automatic analysis of texts and have generated a real revolution in the social sciences research Evans and Aceves (2016). With the development of different NLP algorithms, complex human characteristics such as feelings, opinions, social status, roles, arguments, or meanings have been successfully modeled considering unstructured data Aggarwal and Zhai (2012), Evans and Aceves (2016), Jo (2018). The most promising approaches to model text data in social sciences applications are those based on word embeddings. For instance Kozłowski et al. (2019) illustrates how to use word embeddings in sociological analysis to

model cultural aspects of humans. The same perspective also applies to evaluate psychological aspects Boyd and Schwartz (2020). Machine learning approaches to predict poverty indicators with unconventional data are mainly based on satellite images and meta data of phone-calls (e.g., localization, duration of the call, and others) recorded for further analyses combined with survey data Blumenstock et al. (2015), Jean et al. (2016), Steele et al. (2017), Engstrom et al. (2017), Gebru et al. (2017), Pokhriyal and Jacques (2017), Pandey et al. (2018), Ledesma et al. (2020), Pokhriyal et al. (2020), Lee and Braithwaite (2020), Ayush et al. (2020). Information extracted from social media have also been used to predict poverty Salvatore et al. (2020); Pulse (2014) but, as mentioned above, these approaches have different limitations. With the aim to make progress on automated studies based on NLP methods, we performed a systematic review of the literature and only found the work of Sheehan et al. (2019), which methods are based on NLP to predict social indicators. The authors used the wikipedia geolocated text data to create the models. Besides the scarcity of the studies and the natural difficulty to access the data, the works based on unconventional data mainly rely on external sources (satellite images, wikipedia texts, localization from the mobile network, etc), so specific phenomena related with the feeling of people about poverty are not appropriately captured or has not included as an important variable to tackle this problem. Conversely, NLP methods can be used to extract specific information about what the people think of poverty, what poverty is for them, which implications does it have in their lives, etc. The main advantage of such an approach is that specific feelings, typically hidden in other variables, could be uncovered and grouped into abstract concepts spontaneously expressed by the people through their own language Oved et al. (2020). Although all these are abstract concepts and their description may vary among people and cultures, we believe that text analysis is the best approach to develop methodologies where the opinion of poor people is taken into account as the main source of information.

This paper aims to evaluate the suitability of different word embedding methods to classify between poor and extremely poor people considering texts resulting from 367 interviews given by people experiencing poverty. In this regard, the analysis consists of classical NLP methods such as Latent Semantic Analysis (LSA) Dumais (2004) and Term Frequency-Inverse Document Frequency (TF-IDF) Salton and Buckley (1988), and also novel methods based word embeddings including word2Vec Mikolov et al. (2013), Global Vectors for Word Representation (GloVe) Pennington et al. (2014), and Bidirectional Encoder Representations from Transformers (BERT) Devlin et al. (2018) together with its Spanish version called BETO Canete et al. (2020). The resulting word embeddings are further processed following two statistical approaches: (i) concatenation of several statistical descriptors and (ii) super-vectors created with Gaussian Mixture Models (GMM). The classification is performed with three classifiers which are among the most widely used in machine learning applications: Support Vectors Machine (SVM), Random Forest (RF), and eXtreme Gradient boosting (XGboost).

## 2 Contributions of this Study

Besides the evaluation of a comprehensive list of word embedding methods, this study explores the use of several statistical functionals additional to the classically used mean value to create the representations of each document. Furthermore, Gaussian mixture models (GMM) are used to group document level representations resulting from the GloVe, Word2Vec, BERT, and BetO word embeddings. Finally, considering the specificity of the



**Fig. 1** Summary of the proposed methodology

topic addressed in the corpus and the variety of methods included in the experiments, we think that this work is a baseline for further studies that intend to model poverty using NLP.

### 3 Materials and Methods

The general methodology proposed in this study is summarized in Fig. 1. The process starts with in-person interviews performed by experts in social sciences who visited poor families in different areas of Medellín, Colombia. These interviews were recorded using personal devices and later transliterated by the same social science professionals. A total of five open themes about poverty are included in the questionnaire administered to each person. These themes are included with the aim to motivate the participants to spontaneously talk about poverty. The resulting texts are pre-processed and different word embedding methods are applied to create different representations of each document. The list of embeddings include Word2Vec, GloVe, BERT, and BETO. Additionally, TF-IDF and LSA are also used to create a more comprehensive list of representations. Finally, the automatic discrimination between poor and extremely poor people is performed with three different classifiers: SVM, RF and XGBoost. Further details of this methodology are presented in the following subsections.

#### 3.1 Materials

#### 3.2 Participants and Data Collection

##### 3.2.1 Inclusion Criteria

We conducted the study in the city of Medellín, Colombia, with families that are part of the social program to fight against extreme poverty, namely “*Medellín Solidaria: Familias Medellín*”. Families were selected using three main inclusion criteria: (i) those with a low score in the *Sisben* index, which is a public policy tool to target vulnerable and poor people in Colombia such that allows to conduct social policies; (ii) families that have been in the

program for at least three consecutive years; and (iii) families that have experienced stumbling in their process to overcome the poverty.

### 3.2.2 Data Collection Process

The questionnaire was administered by a group of 16 social science professionals who are in charge of monitoring the poverty condition of the families as a part of their job within the social program. They visited the families twice per month, so they knew the family's history. These professionals were trained with the theoretical and technical background necessary to apply the questionnaires. The representative person of each family was interviewed as the person who most of the times is in charge of making the household decisions. From a total of 431 families, a subset of 367 completed the questionnaire. Data collection was conducted with semi-structured interviews that targeted four main themes: General definition of poverty, deprivations, causes, and opportunities. The themes were addressed with auxiliary questions to guide the conversation. As a result, we concatenated all of the answers of the four themes to obtain one text per person.

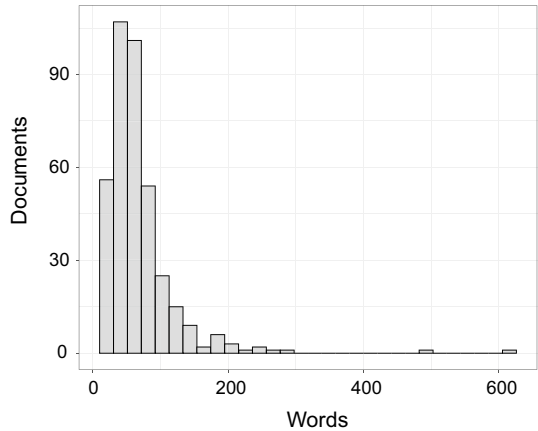
The interviews aimed to characterize the participants in terms of the poverty dimensions considered as relevant for themselves. Other concepts included are: possible reasons they attribute to their condition, difficulties they have gone through, how they usually obtain their resources, which are their desires or hopes, and which are their preferences when making decisions.

### 3.2.3 Labeling Process

Each person is labeled as poor or extremely poor in the corpus according to the national government social target tool called *Sisben* in its version III – System for the identification of potential beneficiaries of the Colombian social system in its third version –. Note that the National Planning Department in Colombia has started to change this tool to its fourth version, which follows a different methodology to label the people; however, the updated data were not available by the time of writing this paper. *Sisben* is an instrument of social policy for targeting poor people in all public policies of Colombia Departamento Nacional de Planeación (sps18). The instrument has questions about household services, quality of housing, health, education, and income. The estimated score is used to classify the population on a scale from 0 to 100. The scores allow to rank the population according to their poverty and vulnerability conditions. Lower scores represent poor or vulnerable while high scores indicate that the household is not vulnerable, so it is not eligible to receive help from the government. In this sense, the score is the main input to define the admission of a given person or family to different social programs. Notice that *sisben's* scores are dynamic, which means that one family could have been labeled as extremely poor at the moment of the interview, but had increased the score recently due to the application of social policies in the city. What is important to highlight is that, even though the score could have been increased, it does not necessarily mean that all social conditions of the family have improved. This is part of the limitations of the methodology used by government agencies to assign/measure the scores.

Besides the labels of poor and extremely poor, the process included also the tagging of the sentences and ideas given by the participants according to four concepts that were intended to be included in the questionnaire, namely: general concept of poverty, deprivations, causes, and opportunities. The interviewers carefully read each of the answers

**Fig. 2** Distribution before pre-processing



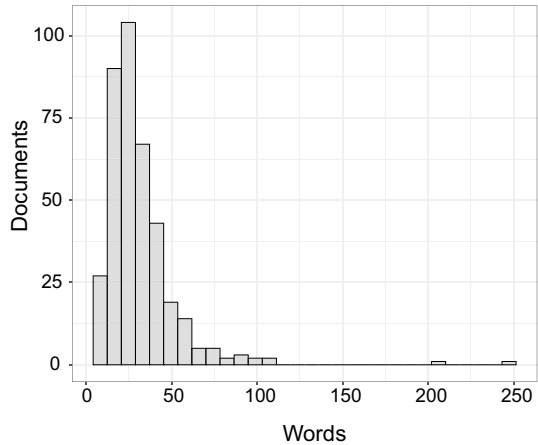
given by each participant and tagged the concepts according to the capability approach theory introduced in Sen (1985, 2009). Although this information is not considered in this study because we are currently focused on setting the baseline to automatically classify between poor and extremely poor people, we think that this tags and the corpus itself constitute a very important input to the research community that is working on modeling different aspects of computational social sciences. For instance, the evaluation of different topic-modeling methods could be one of the first steps in this direction.

### 3.2.4 Questionnaire

The four themes and their corresponding auxiliary questions included in the semi-structured interviews are:

- What is poverty for you? How would you describe it? Do you consider yourself poor? Why? [General definition of poverty]
- What things do you feel you are missing? What things would you like to enjoy but they are impossible to get now? [Deprivations]
- For what reason (or reasons) do you think you are living in these conditions? [Causes]
- What do you think would be necessary to overcome this current living condition? [Opportunities]

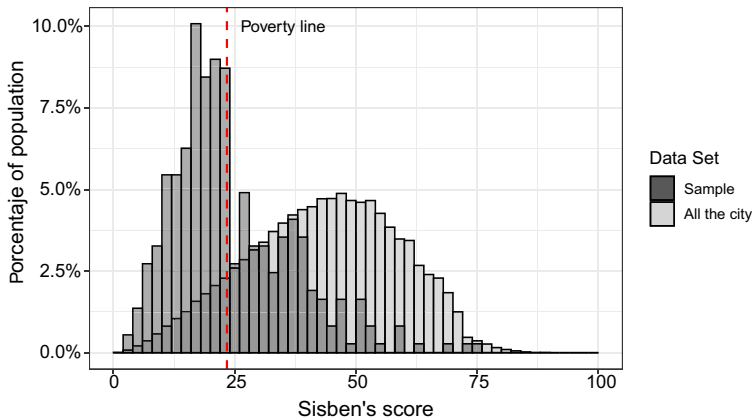
After collecting all answers and structuring the corpus for this study we found that the median value of the number of words per person was 55, the minimum 14, and the maximum 609; those values were calculated by counting each word said by the interviewee and without running any pre-processing; the vocabulary size was 25,013. After pre-processing, the median of the number of words turned into 27, the minimum 8, and the maximum 247; and the vocabulary size decreased to 11,572. Figures 2 and 3 show that the distribution of the words per person does not change with the pre-processing. The word count excludes the interviewer's words.

**Fig. 3** Distribution after preprocessing

The following are examples of answers provided by different interviewed persons, we include them with the aim to provide the reader with a more complete picture of the concepts we are modeling to discriminate between poor and extremely poor people.

- (i) “Que uno como ser humano no tenga con que hacer una agua de panela [General definition of poverty]. De todo, una casa donde vivir, uno consigue trabajo y alimentos, pero sin casa es muy difícil. Una vivienda digna [Deprivations]. Por no saber pensar o por falta de oportunidades [Opportunities]. Un buen empleo para mi hijo y una vivienda [Opportunities].” *English translation:* “*The fact that one as a human being does not have enough to make an “Agua de panela”<sup>1</sup> [General definition of poverty]. Everything, a house to live in, one gets a job and food, but without a house it is very difficult. A decent house [Deprivations]. Due to lack of knowledge or lack of opportunities [Opportunities]. A good job for my son and a house [Opportunities].*”
- (ii) “No contar con recursos tanto materiales como espirituales [General definition of poverty]. Me falta tener a mis hijos más cerca de mí. Disfrutar de mis hijos, mis nietos y de buena salud, viajar y conocer [Deprivations]. Lo que llamamos circunstancias de la vida que son atribuidas a las malas decisiones, no aprovechar las oportunidades que nos da la vida a diario [Opportunities]. Tener a mis hijos más cerca, es difícil, pero si se me puede mejorar la salud y aprovechar más las oportunidades [Opportunities].” *English translation:* “*Not having both material and spiritual resources [General definition of poverty]. I am missing to have my children closer to me. To enjoy my children, my grandchildren and a good health, to travel and to see places [Deprivations]. What we call life circumstances that are attributed to bad decisions, not taking advantage of the opportunities that life gives us every day [Opportunities]. Having my children closer, it is difficult, but if I can improve my health and take advantage of more opportunities [Opportunities].*”
- (iii) “La pobreza es cuando la persona vive en la calle, no tiene donde dormir, ni que comer o también cuando uno tiene donde vivir pero no tiene con qué comer. Si consigue para una cosa no tiene para la otra. Si conseguimos para el arriendo solamente

<sup>1</sup> Agua de panela is a very traditional beverage made out of sugarcane.



**Fig. 4** *Sisben's* score distribution for the total population of Medellín and for the sample included in this study

conseguimos para una comida. Nos ha tocado pedir [General definition of poverty]. Nos hace falta una casa propia y empleo [Deprivations]. Porque mi mamá es muy malgeniada, no sabe decir las cosas. Cuando vivíamos en Puerto Berrío vivíamos muy bien, y a mi mamá la privaron de la libertad durante 8 años. Yo y mi cuñado estuvimos al frente del cuidado de la familia. No tuvimos un papá ni una mamá que nos dijera cómo hacer las cosas. Todos hacían lo que les diera la gana [Causes]. El empleo para tener una vida mejor y tener buen diálogo [Deprivations].” *English translation:* “Poverty is when a person lives on the street, has nowhere to sleep, nothing to eat, or when one has a place to live but nothing to eat. If you get enough for one thing you don’t have enough for the other. If we get enough to pay the rent, we only get enough for one meal. We have had to beg for [General definition of poverty]. We need our own house and a job [Deprivations]. Because my mother is always in a bad-mood, she doesn’t know how to say things. When we lived in Puerto Berrío we lived very well and my mother was deprived of her freedom for 8 years. Me and my brother-in-law were in charge of taking care of the family. We didn’t have a father or a mother telling us how to do things. Everyone did whatever they wanted [Causes]. Employment to have a better life and to have a good dialogue [Opportunities].”

### 3.3 Descriptive Analysis

Figure 4 shows the distribution of the interviewed population according to their *Sisben's* score (dark-gray bars) in comparison with the distribution of the whole population of Medellín (light-gray bars). We used the value of 23.4 points as a cut-off score to define the threshold between poor and extremely poor people. Such a score was defined by the Department of Social Prosperity of the Colombian government according to the Resolution # 481 of 2014, issued by the National Agency for Overcoming Extreme Poverty (ANSPE in Spanish). Hence, people with a score smaller than 23.4 points are considered as extremely poor, and people with a score above are considered as poor in our corpus. Note that the government does not label the entire population of the city with the *Sisben* score because this tool is available on-demand by each individual. By 2017 in Medellín there were 1,943,631 people registered in the *Sisben*. From the figure it can also be observed that



**Table 1** Man Whitney test comparing poor and extremely poor populations in years of education and age

Variable	Measures	Poor	Extremely poor	W-value	<i>p</i> -value
Years of education	<i>N</i>	151	151	13554	0.023
	Mean	7.48	7.03		
	Median	9	6		
	<i>r</i>	.002	– .03		
Age	<i>N</i>	151	151	18106	0.042
	Mean	40.73	42.33		
	Median	39	44.5		
	<i>r</i>	.095	0.16		

*N* number of samples, *r*: Pearson's correlation between the Sisben score and the given variable.

There were six missing values for years of education and age

our data are left-tail distributed, which means that the interviews were mainly administered to people with small *Sisben* scores. The 'outliers' observed in the right part of the figure, i.e., with high scores, are families that participated in the poverty alleviation program years ago and met the inclusion criteria for this study. It is worth-noting to stress the fact that our corpus has 56.1% people below the poverty line threshold compared to 11.3% in the distribution of the Medellín city.

To complement the descriptive distribution of database, we calculated independent tests using the target variable against years of education, age, and place of residence. The variable sex was not tested because 93.2% of the interviewed people were women, head of the household, which is common among poor families. Pearson's Chi-square tests were performed for nominal and categorical variables. We tested the normality with the Shapiro-test, which suggests strong evidence of non-normality for the years of education and age variables. The Mann Whitney test was used for numerical variables. We selected a significance level of 0.05 to evaluate the tests. The results obtained in the Mann Whitney test are provided in Table 1.

The Mann Whitney test indicates that there is weak evidence for assuming any difference between poor and extremely poor samples based on age ( $p$ -value = 0.042) and years of education ( $p$ -value = 0.023) (See Table 1). The  $p$ -values obtained are less than the threshold value for  $p$  at 5%, but they are higher than 1%. It means that error rate type I is less than 0.023 for years of education and 0.042 to age. There is no strong evidence for an existing bias. Also, it is important to highlight that one of the main difficulties of the studies about evaluating poverty by directly interviewing the people, is to create an homogeneous sample without any bias. Finally, no correlation is observed between years of education and the *Sisben*'s score in none of the two groups.

Table 2 shows the results obtained in the Chi-squared tests where the variables residence and years of education are compared between poor and extremely poor people. Note that when grouping the years of education the conclusion is different than in the previous case, the poor and extremely poor classes are independent of years of education ( $p$ -value = 0.382). The group of extremely poor people has a mean of seven years of education while the poor group has a mean of 7.5 years. In terms of the residence, there is no evidence to differentiate poor and extremely poor people ( $\chi^2$  with  $p$ -value = 0.064). It is important to highlight that all residence places considered in this database correspond to geographic areas in the city where vulnerable people are located.

**Table 2** Chi-squared tests for comparing poor and extremely poor populations in residence and intervals of years of education

Variable	N and factors	Poor	Extremely poor	$\chi^2$	p-value
Residence	N	151	151	8.88	0.064
	Aranjuez	17	24		
	Buenos Aires	18	34		
	Candelaria	56	48		
	Castilla	11	8		
	San Javier	49	37		
Years of education	N	151	151	1.92	0.382
	0–5 years	60	71		
	6–11 years	82	70		
	Over 11 years	9	10		

N number of samples. Factors in the case of Residence correspond to names of areas in the city of Medellín. There were six missing values for years of education and age

### 3.4 Pre-processing

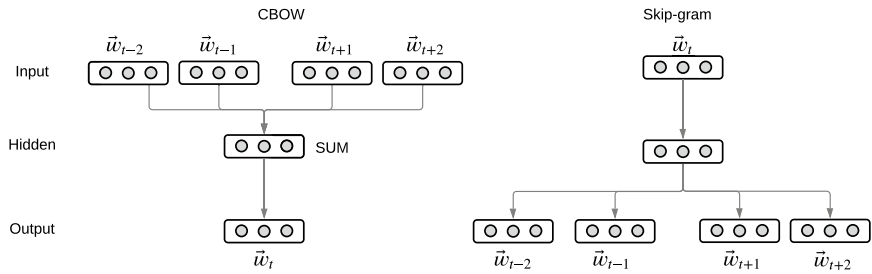
One document with the combination of the answers given per person in the interviews is created. Typical steps in pre-processing text data include tokenization, lemmatization, dependency parsing and parts of speech tagging. We addressed these four steps using the UDPipe framework which is available as a CRAN package of the R programming language Wijffels (2019). A total of 308 stop words (meaningless words) and also numerical digits, punctuation marks, and words with three or more repeated letters were removed from the documents. All this process and estimates were performed in Spanish language.

### 3.5 Methods

#### 3.5.1 Word Representation Learning and Word Embeddings

Word representation learning is an unsupervised process where patterns resulting from text data are represented in vector spaces. Texts are composed with combinations of words which can be represented by word embeddings. The idea behind word embeddings is to create a vector space where the text holds its semantic properties Pilehvar and Camacho-Collados (2020). It is possible to infer the semantic relationship of the words in a given text according to the distributional hypothesis: *A word is characterized by the company it keeps* Harris (1954). Vector space models are related to the distributional hypothesis. Turney and Pantel (2010) presents a comprehensive survey with applications of earlier word vector space models. There are two main approaches: static and contextual representations. The first one depends on identifying the meaning of a particular word, while the second one adds the meaning of the words surrounding the target word. Contextual representation resolves natural language ambiguity when words or phrases have multiple meanings or interpretations.

Static methods are based on word counting and co-occurrence term matrices, which is not the case when modern deep learning -based methods that use contextual information to train representations are applied. The most popular counting-based approach is TF-IDF.



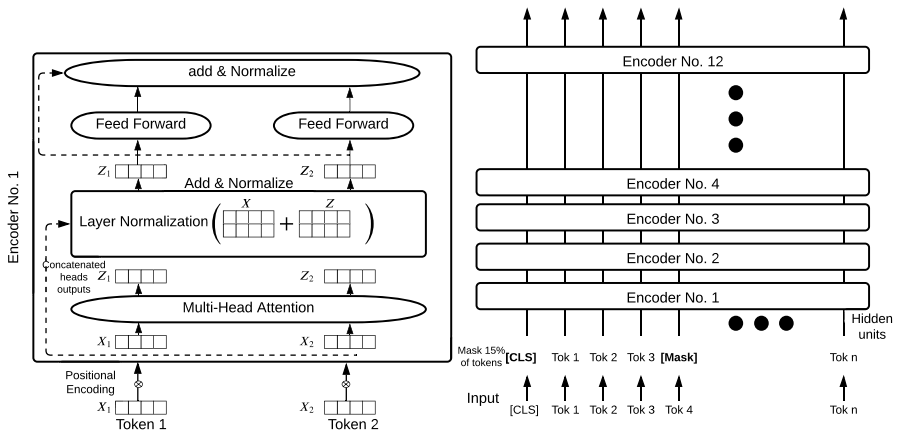
**Fig. 5** CBOW and Skip-gram strategies of Word2Vec Mikolov et al. (2013). Figure adapted From Pilehvar and Camacho-Collados (2020)

It is a weighting approach where high values are assigned to terms that are common in a given corresponding document but not as common in other documents of the corpus Salton and Buckley (1988).

The LSA is another widely used counting-based method where vectors of text data are processed by applying Singular Value Decomposition (SVD) upon the term frequency matrix to inferring relations of usage of words in passages Landauer et al. (1998). Although this is the most common approach, we decided to do the SVD upon the TF-IDF matrix because it produces better representations of the documents and reduces the noise Salton and Buckley (1988). This method provides a text representation in terms of the topics or latent features. According to the literature, one of its benefits is the reduction of the dimensionality of the original term frequency matrix and the elimination of noise. Additionally, although the sequential order of the words is not appropriately modeled by LSA, the method allows getting relevant semantic aspects Landauer et al. (1998); Stein et al. (2019); Turney and Pantel (2010).

Another word representation is Word2Vec, which was first introduced in Mikolov et al. (2013). It is a special type of distributed word representation that popularized the use of neural networks as an alternative to create word representation spaces. Word2Vec embeddings can be generated following two different strategies: Continuous Bag-Of-Words (CBOW) and Skip-Gram models. In the second strategy a binary classifier is trained and its weights are the embeddings that represent the semantic meaning. Figure 5 presents the general architecture of the Word2Vec approaches. The Skip-gram model predicts the words in the surrounding context given the target word, while the CBOW predicts the target word given the context. The input to word2Vec embeddings are one-hot vectors, so the co-occurrence and statistical information are missing. The embedding term refers to the word vector representation obtained from the application of a fully connected neural network that predicts the next word based on the prior ones.

Another model to obtain word embeddings is GloVe Pennington et al. (2014). This is a distributed word representation that uses word co-occurrence information and applies dimensionality reduction to predict the weights. GloVe is considered as a predictive model which is based on a log-bilinear regression applied over the co-occurrence matrix and the local context window Pennington et al. (2014). To extract GloVe representations it is necessary to define the word co-occurrence matrix  $\mathbf{X}$ , where each element  $X_{ij}$  represents the number of times the  $j$ -th word occurs in the context of the  $i$ -th word. The pairs  $X_{ij}$  are constrained according to Equation 1, where  $\mathbf{w}_j$  is the vector of the objective word, and  $\mathbf{w}_i$  is vector of the context word; also  $b_j$  and  $b_i$  are biases for the objective and context words, respectively. The optimization problem is based on the cost function illustrated in



**Fig. 6** Architecture of BERT Devlin et al. (2018). The right-hand side illustrates the general architecture and the left-hand side is the zoom-in of one encoder. Figure adapted from Alammar (2020)

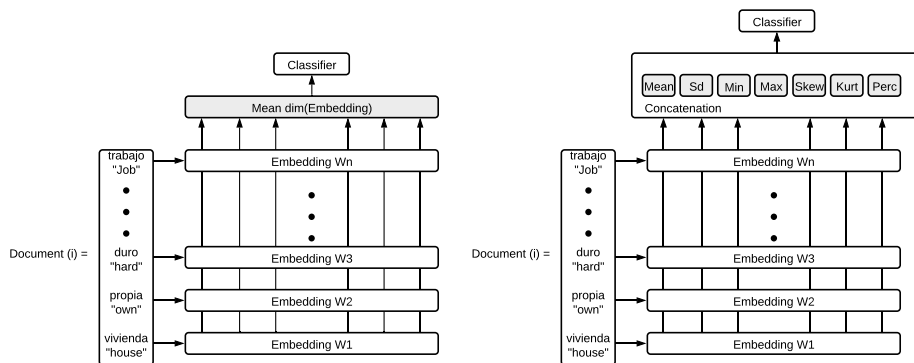
Equation 2, which is a least squared problem where  $V$  is the size of the vocabulary. The weighting function  $f(X_{ij})$  in Equation 3 is introduced to avoid focusing on common word pairs. The term  $X_{max}$  is the maximum value of co-occurrence count fixed as 100 cutoff point argument by the authors for all their experiments Pennington et al. (2014).

$$\mathbf{w}_i^T \mathbf{w}_j + b_i + b_j = \log(X_{ij}) \quad (1)$$

$$J = \sum_{i=1}^V \sum_{j=1}^V f(X_{ij}) (\mathbf{w}_i^T \mathbf{w}_j + b_i + b_j - \log X_{ij})^2 \quad (2)$$

$$f(X_{ij}) = \begin{cases} \left(\frac{X_{ij}}{X_{max}}\right)^\alpha & \text{if } X_{ij} < X_{max} \\ 1 & \text{otherwise.} \end{cases} \quad (3)$$

Both GloVe and Word2Vec are used in many text classification tasks Mitra and Jenamani (2020), Stein et al. (2019), Jang et al. (2019), Rezaeinia et al. (2019), Abdillahi et al. (2020), Yu et al. (2017). We encourage the readers to see the works of Li et al. (2020) and Minaee et al. (2020) for a more detailed reference of different NLP applications in text classification tasks. After the introduction of the GloVe method, the research community wanted to continue making progress in creating word representation methods where the meaning of the words surrounding the target one were modeled. One of the most popular and successful approaches towards this direction is BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. (2018), and its improvements known as ALBERT, DistilBERT, and RoBERTa. These are transformers-based methods that create the word vectors automatically and the resulting representations are contextualized, unlike the static embeddings obtained with GloVe and Word2Vec Rogers et al. (2020). Transformers-based models have achieved great success in different tasks by using the so called attention mechanism methods Devlin et al. (2018). The architecture of the neural network implemented in BERT is a stack of transformer encoder layers, as shown in Fig. 6. Inside each layer, the model has two main components: the attention mechanism with multiples heads and the



**Fig. 7** Two approaches for document level representation. The left-hand side illustrates the typical approach based on mean values and the right-hand side shows the list of the statistical functionals proposed in the SFDL model to create the document-level representation

fully connected neural network Rogers et al. (2020). Such an attention mechanism allows to learn how to put more attention on the specific inputs to create appropriately weighted representations. Transformers-based methods have been widely used since 2018 thanks to their excellent performance in different text classification tasks Rogers et al. (2020). Besides BERT, in this work we also used BETO Canete et al. (2020), which is the Spanish version of BERT, i.e., BETO is trained with Spanish data.

### 3.6 Document Level Representation

Word embeddings do not provide representation for sentences or documents, so different statistical functionals have to be estimated to create models at a sentence level. In this work we compute seven statistics to model all words in each document, such that there will be a document level representation. The typical approach consists in estimating only mean values over the embeddings extracted from each word in a given document. Besides, with the aim to create a more comprehensive representation per document, we propose to estimate Statistical Features at a Document-Level (SFDL) from the word embeddings. The estimated statistics are used as the input to the classifier, as shown in Fig. 7. The left side of the graph illustrates the typical way to calculate the document-level representation using the mean value of each word embedding Mitra and Jenamani (2020); Kenter et al. (2016). The right side of the graph shows the way that we propose for concatenating seven statistical functionals from the word embeddings. The mean value, standard deviation, minimum, maximum, skewness, kurtosis, and the hundred percentiles of the distribution are included in the SFDL model.

### 3.7 Feature Representation Through Gaussian Mixture Models (GMM)

GMM belong to the family of probabilistic methods where the probability density function is represented as a weighted sum of Gaussian distributions. The method represents a multi-dimensional distribution as a sum of uni-dimensional distributions Reynolds (2009). A GMM distribution is represented as in equation 4.

$$p(\mathbf{x} \mid w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \sum_{i=1}^M w_i g(\mathbf{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (4)$$

Where  $\mathbf{x}$  is a feature vector,  $w_i$  is the weight for the  $i$ -th Gaussian such that  $\sum_i w_i = 1$ , and  $M$  is the number of Gaussian components in the model. Each  $d$ -dimensional Gaussian component is represented as in equation 5, where  $\boldsymbol{\mu}_i$  is the mean vector and  $\boldsymbol{\Sigma}_i$  the covariance matrix.

$$g(\mathbf{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{d/2} \boldsymbol{\Sigma}_i^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right) \quad (5)$$

The expectation-maximization (EM) algorithm is used to fit GMM to the data in the training set. The EM is an iterative process to estimate the parameters  $w_i$ ,  $\boldsymbol{\mu}_i$ , and  $\boldsymbol{\Sigma}_i$  such that maximize the likelihood of the GMM with the data Reynolds (2009). The GMM likelihood follows the Equation 6, and the estimated model has greater GMM likelihood such that:

$$p(\mathbf{X} \mid \bar{w}, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}) \geq p(\mathbf{X} \mid w, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

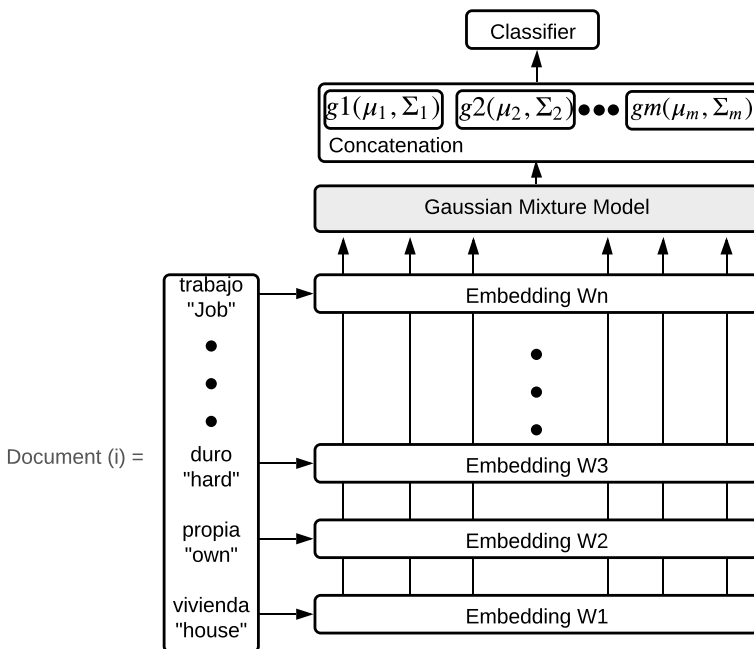
Where parameters with bar are part of the new model in each iteration until a given convergence threshold is reached Reynolds (2009).

$$p(\mathbf{X} \mid w, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{t=1}^T p(\mathbf{x}_t \mid w, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (6)$$

In our case, we used four and eight components per document, i.e.,  $d = 4$  and  $d = 8$ , to avoid losing information due to the minimum amount of words per document. The maximum components are given by the minimum amount of words per document. The document level representation with GMMs includes number of components, mean vector, and the covariance matrix  $\boldsymbol{\Sigma}_i$  which in this case is assumed as diagonal. Figure 8 shows the process to create the document level representation using GMM. This approach has been used in other applications like identity verification with speech Reynolds et al. (2000), keystroke dynamics Escobar-Grisales et al. (2020), to assess Parkinson's disease over time Arias-Vergara et al. (2016), and also to model adversarial attacks on GMM vectors Li et al. (2020). In the literature review, we did not find this method applied to text representation with word embeddings.

## 4 Experiments

The texts considered in this study are modeled with two different count methods: TF-IDF and LSA. Also, different word embedding methods are used to create document-level representations, including Word2Vec, GloVe, BERT, and Beto. We computed the average/mean value of the word embeddings to form a static vector representations. Also, the concatenation of several statistics are used. Additionally, a GMM based approach was implemented by using the dynamics of the word embedding representations. The dimensionality of the matrix for these experiments was 302 documents  $\times$  1900 features for GloVe and Word2Vec; and 302  $\times$  4708 for BERT and Beto. Two different versions of the GMM are evaluated, one with 4 Gaussians and the other one with 8.



**Fig. 8** Gaussian Mixture Model approach for document level representation

Before to train the classifiers, the dataset was randomly balanced by class. From a total of 367 interviews, 151 are poor and 216 are extremely poor. A total of 151 samples from the second group were randomly chosen to create a balanced corpus with 151 samples per class (a total of 302 documents). The negative class is tagged as extremely poor in the database, while the positive class is poor people labeled as one. For all experiments the 10-fold cross-validation strategy is followed. The automatic classification between poor and extremely poor participants is performed with three classification methods Li et al. (2020): SVM Joachims (1998), RF Breiman (2001), and xGBoost Chen and Guestrin (2016). Optimal parameters for the classifiers are found following a grid search strategy.

#### 4.1 Experiment One: Counting Methods

We calculated the TF-IDF and removed the words that did not appear in the 99% of the documents, which results in a dimensionality reduction from  $367 \times 2442$  (documents  $\times$  words) to  $338 \times 489$ . Besides the classical TF-IDF, to obtain another document matrix representation, we applied Singular Value Decomposition (SVD) to the TF-IDF matrix to get the Latent Semantic Analysis representation. The first 225 eigen-vectors are considered to extract 95% of explained variance. Notice that this procedure has an additional reduction in dimensionality, which on the one hand may result in a more compact representation, but on the other hand, may result in a loss of information.

**Table 3** Results with counting methods. Four performance metrics with mean value and standard deviation

Method	Accuracy	Specificity	Sensitivity	F1-Score
TF-IDF - Unigram - SVM	0.492 $\pm$ 0.03	0.562 $\pm$ 0.13	0.423 $\pm$ 0.14	0.443 $\pm$ 0.09
TF-IDF - Unigram - SVD - SVM	0.474 $\pm$ 0.02	0.48 $\pm$ 0.10	0.464 $\pm$ 0.08	0.466 $\pm$ 0.04
TF-IDF - Unigram - RF	0.485 $\pm$ 0.01	0.505 $\pm$ 0.02	0.466 $\pm$ 0.20	0.475 $\pm$ 0.02
TF-IDF - Unigram - SVD - RF	0.494 $\pm$ 0.02	0.501 $\pm$ 0.03	0.487 $\pm$ 0.03	0.490 $\pm$ 0.03
TF-IDF - Unigram - XGB	0.496 $\pm$ 0.01	0.495 $\pm$ 0.02	0.498 $\pm$ 0.02	0.497 $\pm$ 0.02
TF-IDF - Unigram - SVD - XGB	<b>0.552 <math>\pm</math> 0.03</b>	0.558 $\pm$ 0.38	0.546 $\pm$ 0.04	0.549 $\pm$ 0.03

The highest accuracy is highlighted in bold

## 4.2 Experiment Two: Word Embeddings methods

Four different word embedding methods are used to create vector representations for each word of the documents. Word2Vec and GloVe embeddings Pennington et al. (2014) are trained with the Spanish Wikipedia 2018 Corpus which contains about 709 million words and it was downloaded on February 2020<sup>2</sup>. For the case of Word2Vec, it was trained using the Skip-Gram method with a window size of 8 words and a minimum count of 5 occurrences per word Mikolov et al. (2013). 300-dimensional embeddings are created on each case, as suggested in Pennington et al. (2014). Both models, GloVe and Word2Vec, are trained in Python using the gensim module Řehůřek and Sojka (2010) implemented in a cluster computer with 96 cores and 256GB of RAM memory. For BERT and BETO we used a pre-trained model (multilingual base case) Devlin et al. (2018); Canete et al. (2020). Both are 768-dimensional representations (Hidden units) in their pre-trained version. Models resulting from the use of only mean values are considered as baseline for comparison purposes. Models with more statistics estimated at a document-level (SFDL) are also considered.

## 4.3 Experiment Three: Gaussian Mixture Model

We trained a GMM considering the embeddings resulting from word2Vec, GloVe, BERT, and Beto. Given the limited number of samples per class, weights of the GMM are excluded from the optimization procedure, hence only the mean vector and the diagonal of the covariance matrix are considered 8. Models with more than 8 Gaussians are not considered here because it resulted in losing about 15% of the documents.

## 5 Results and Discussion

The best accuracy is 55.2%, and it is obtained with the LSA model (SVD on the TF-IDF matrix) and xGBoost as classifier (See Table 3). The second best accuracy is 54.6% with GloVe and only mean values as document level representation. The classification is also

<sup>2</sup> “eswiki-latest-pages-articles.xml.bz2”



**Table 4** Results with word embeddings. Includes the comparison between word embeddings with mean values only and the SFDL document-level representation. Four performance metrics with mean value and standard deviation

Method	Accuracy	Specificity	Sensitivity	F1-Score
GloVe - Mean - SVM	0.496 $\pm$ 0.02	0.195 $\pm$ 0.07	<b>0.796 <math>\pm</math> 0.07</b>	0.611 $\pm$ 0.02
GloVe - SFDL - SVM	0.525 $\pm$ 0.03	0.498 $\pm$ 0.04	0.552 $\pm$ 0.05	0.537 $\pm$ 0.03
GloVe - Mean - XGB	<b>0.546 <math>\pm</math> 0.01</b>	0.554 $\pm$ 0.03	0.538 $\pm$ 0.01	0.542 $\pm$ 0.01
GloVe - SFDL - XGB	0.506 $\pm$ 0.02	0.501 $\pm$ 0.03	0.512 $\pm$ 0.03	0.510 $\pm$ 0.02
GloVe - Mean - RF	<b>0.545 <math>\pm</math> 0.02</b>	0.550 $\pm$ 0.02	0.541 $\pm$ 0.03	0.543 $\pm$ 0.02
GloVe - SFDL - RF	0.499 $\pm$ 0.02	0.524 $\pm$ 0.03	0.474 $\pm$ 0.03	0.486 $\pm$ 0.02
Wor2Vec - Mean - SVM	0.503 $\pm$ 0.02	0.320 $\pm$ 0.09	<b>0.686 <math>\pm</math> 0.07</b>	0.578 $\pm$ 0.03
Wor2Vec - SFDL - SVM	0.510 $\pm$ 0.02	0.539 $\pm$ 0.03	0.481 $\pm$ 0.02	0.495 $\pm$ 0.02
Wor2Vec - Mean - XGB	0.507 $\pm$ 0.03	0.504 $\pm$ 0.02	0.512 $\pm$ 0.04	0.509 $\pm$ 0.03
Wor2Vec - SFDL - XGB	0.501 $\pm$ 0.01	0.503 $\pm$ 0.01	0.498 $\pm$ 0.02	0.499 $\pm$ 0.01
Wor2Vec - Mean - RF	0.503 $\pm$ 0.02	0.491 $\pm$ 0.03	0.516 $\pm$ 0.02	0.510 $\pm$ 0.02
Wor2Vec - SFDL - RF	0.497 $\pm$ 0.02	0.511 $\pm$ 0.03	0.482 $\pm$ 0.03	0.489 $\pm$ 0.02
BERT - Mean - SVM	0.505 $\pm$ 0.02	<b>0.706 <math>\pm</math> 0.06</b>	0.304 $\pm$ 0.07	0.377 $\pm$ 0.05
BERT - SFDL - SVM	0.486 $\pm$ 0.01	0.490 $\pm$ 0.05	0.482 $\pm$ 0.06	0.482 $\pm$ 0.03
BERT - Mean - RF	0.485 $\pm$ 0.02	0.497 $\pm$ 0.03	0.473 $\pm$ 0.03	0.479 $\pm$ 0.02
BERT - SFDL - RF	0.477 $\pm$ 0.03	0.480 $\pm$ 0.03	0.474 $\pm$ 0.04	0.475 $\pm$ 0.03
BERT - Mean - XGB	0.484 $\pm$ 0.02	0.475 $\pm$ 0.02	0.493 $\pm$ 0.03	0.488 $\pm$ 0.03
BERT - SFDL - XGB	0.435 $\pm$ 0.02	0.429 $\pm$ 0.02	0.441 $\pm$ 0.02	0.438 $\pm$ 0.02
Beto - Mean - SVM	0.478 $\pm$ 0.04	<b>0.610 <math>\pm</math> 0.08</b>	0.347 $\pm$ 0.09	0.393 $\pm$ 0.08
Beto - SFDL - SVM	0.500 $\pm$ 0.02	0.492 $\pm$ 0.03	0.507 $\pm$ 0.03	0.503 $\pm$ 0.02
Beto - Mean - RF	0.462 $\pm$ 0.02	0.484 $\pm$ 0.03	0.441 $\pm$ 0.02	0.451 $\pm$ 0.02
Beto - SFDL - RF	0.461 $\pm$ 0.02	0.466 $\pm$ 0.03	0.456 $\pm$ 0.02	0.458 $\pm$ 0.02
Beto - Mean - XGB	0.476 $\pm$ 0.02	0.469 $\pm$ 0.03	0.484 $\pm$ 0.04	0.480 $\pm$ 0.03
Beto - SFDL - XGB	0.452 $\pm$ 0.02	0.458 $\pm$ 0.02	0.445 $\pm$ 0.03	0.448 $\pm$ 0.02

The highest values of accuracy, sensitivity and specificity are highlighted in bold

with XGBoost. A similar result is obtained with GloVe, and the RF classification model (see Table 4). With the aim to validate whether a simpler classification method could result in a similar result, the Logistic Regression algorithm was implemented. The obtained accuracy was 51.0% when using the two best feature sets (LSA, and Glove - Mean), which supports the fact that the addressed problem is complex and perhaps needs more elaborated methods to find better results. The experiment using Word2Vec embeddings and GMM as document level representation yields an accuracy of 52% with an RF classifier (See Table 5). It can be observed that none of the models with more statistics (i.e., SFDL) exhibited better results than those obtained with the classical models based only on mean values of the vector representations.

Additionally, no pattern is observed regarding the ability of different word representations to classify poor vs. extremely poor subjects. All of the methods show similar performances which makes it evident the high complexity of the problem addressed in this work. In general terms, the experiments show that different feature extraction or document level representation methods do not influence the classification performance of poor and extremely poor people; the classical methods of document level representation

**Table 5** Results with Gaussian Mixture Models. Four performance metrics with mean value and standard deviation

Method	Accuracy	Specificity	Sensitivity	F1-Score
GloVe - GMM4 - SVM	0.517 ± 0.02	0.561 ± 0.13	0.474 ± 0.15	0.480 ± 0.10
GloVe - GMM8 - SVM	0.456 ± 0.02	0.303 ± 0.04	<b>0.610 ± 0.07</b>	0.526 ± 0.03
GloVe - GMM4 - XGB	0.487 ± 0.03	0.474 ± 0.03	0.499 ± 0.04	0.492 ± 0.03
GloVe - GMM8 - XGB	0.502 ± 0.02	0.491 ± 0.02	0.513 ± 0.03	0.507 ± 0.02
GloVe - GMM4 - RF	0.490 ± 0.02	0.485 ± 0.05	0.496 ± 0.04	0.492 ± 0.02
GloVe - GMM8 - RF	0.470 ± 0.021	0.495 ± 0.032	0.445 ± 0.033	0.456 ± 0.033
Wor2Vec - GMM4 - SVM	0.485 ± 0.01	0.453 ± 0.03	0.513 ± 0.02	0.499 ± 0.01
Wor2Vec - GMM8 - SVM	0.464 ± 0.02	0.482 ± 0.02	0.447 ± 0.03	0.454 ± 0.02
Wor2Vec - GMM4 - XGB	0.527 ± 0.02	0.519 ± 0.04	0.535 ± 0.03	0.53 ± 0.02
Wor2Vec - GMM8 - XGB	0.452 ± 0.02	0.452 ± 0.02	0.452 ± 0.03	0.452 ± 0.02
Wor2Vec - GMM4 - RF	<b>0.520 ± 0.02</b>	0.509 ± 0.03	0.534 ± 0.02	0.526 ± 0.02
Wor2Vec - GMM8 - RF	0.465 ± 0.02	0.470 ± 0.03	0.461 ± 0.02	0.463 ± 0.02
Beto - GMM4 - SVM	0.498 ± 0.02	0.554 ± 0.06	0.441 ± 0.05	0.466 ± 0.03
Beto - GMM8 - SVM	0.463 ± 0.03	<b>0.696 ± 0.06</b>	0.230 ± 0.04	0.298 ± 0.04
BERT - GMM4 - SVM	0.493 ± 0.02	0.493 ± 0.07	0.492 ± 0.07	0.490 ± 0.04
BERT - GMM8 - SVM	0.471 ± 0.02	0.453 ± 0.06	0.489 ± 0.06	0.479 ± 0.03
BERT - GMM4 - RF	0.476 ± 0.02	0.467 ± 0.02	0.484 ± 0.02	0.480 ± 0.02
BERT - GMM8 - RF	0.435 ± 0.02	0.454 ± 0.02	0.415 ± 0.05	0.422 ± 0.03
Beto - GMM4 - RF	0.481 ± 0.02	0.498 ± 0.02	0.464 ± 0.02	0.472 ± 0.01
Beto - GMM8 - RF	0.486 ± 0.01	0.494 ± 0.03	0.478 ± 0.02	0.481 ± 0.02
BERT - GMM4 - XGB	0.479 ± 0.02	0.462 ± 0.03	0.497 ± 0.02	0.488 ± 0.02
BERT - GMM8 - XGB	0.484 ± 0.02	0.464 ± 0.02	0.505 ± 0.02	0.495 ± 0.02
Beto - GMM4 - XGB	0.464 ± 0.02	0.472 ± 0.03	0.456 ± 0.02	0.459 ± 0.02
Beto - GMM8 - XGB	0.509 ± 0.02	0.510 ± 0.02	0.508 ± 0.03	0.508 ± 0.02

The highest values of accuracy, sensitivity and specificity are highlighted in bold

(TF-IDF) perform in a similar way to modern methods (Glove and Bert). Although the accuracies obtained along this study are not satisfactory, there are some results to highlight. For instance, several experiments show relatively high sensitivity and others showed high Specificity. As shown in Table 4, when using the GloVe model based on mean values and the SVM as classifier, the resulting sensitivity is 79.6%. Similarly, the BERT model with the SVM classifier yields a specificity of 70.6%. These two results indicate that it is possible to accurately detect poor people with the GloVe method, while the detection of extremely poor people is possible with the BERT model. This finding also suggests that one of the possible research topics in the near future is to evaluate different strategies to fuse information, e.g., early, late, or slow fusion, or even those based on modern recurrent neural networks like the gated multimodal units (GMU). The above results could be of interest for public policy makers with the aim to target social programs to extremely poor people.

It is worthwhile to note that GMM method is promising because it achieves similar results to those reported in the state-of-the-art, e.g., BERT and Beto. We believe that the GMM method does not perform well in this work due to the small size of the corpus. GMM

is typically used in topics like speech recognition or speaker verification, where thousands of data are available.

Finally, we want to stress the fact that the results observed in this study clearly highlight the complexity of the problems related to the modelling of social and economical conditions based on spontaneous discourses about their own experience. This observation about the high complexity of the topic was previously highlighted in Salganik et al. (2020).

## 6 Limitations of the Study

We believe that the main limitations of this study include: (1) the government's score (namely Sisben) is not designed to differentiate between poor and extremely poor people, therefore there is no a clear threshold to distinguish between these two groups; (2) the small size of the corpus considered here limits the chances of finding robust and optimal parameters for the representation models and the classifiers; (3) pre-trained vectors have difficulties with specific semantic fields, for instance those used in GloVe, Word2Vec, BERT, and Beto, use general databases with non an specific semantic domain Mitra and Jenamani (2020). Then, the resulting word vector representation does not belong to the same mental/lexical concept: poverty, in this case; and finally, (4) in general terms, some of the models used here have limitations of interpretation given the basis of deep neural networks to create the word vector representations Li et al. (2020).

## 7 Conclusions

In this paper, we present an approach to classify poor and extremely poor people through NLP. We trained several classifiers using people discourses about poverty in order to create embedding vectors that encode the concept of poverty. We use classical and modern word vector representations, and additionally propose two strategies for document-level features: concatenation of descriptive statistics, and Gaussian Mixture Models. We systematically tested the word and documents representation with three classifiers: SVM, RF, and XGBoost.

There are no strong differences between the word and document representation methods used in this study. Accuracies of around 55% were obtained with the TF-IDF method and with the GloVe word embedding approach. The SVM classifier seems to be the most robust among the three classification methods evaluated in the experiments. Some models exhibited relatively high sensitivity (79.6%), which could be of interest for public policy makers to target social programs.

This paper allows to open the discussion about poverty and NLP. Our research belongs to the emerging area about the integration of artificial intelligence methods with social problems. The insights for this work are the use text data collected from poor people to understand their poverty condition and their own concept of poverty. In this regard, we tested several state-of-the-art methods to create text representations. The most relevant outcome was the hypotheses for further analyses. Thus, one major challenge consists in developing the theoretical framework necessary to find appropriate semantic fields such that allow to create accurate models for the concept of poverty. The main rationale for this challenge is that the embeddings used here (and in most of the studies focused on language modeling) are trained with data of general propose, therefore the resulting embeddings

have an encoding bias Bender et al. (2021). Specific concepts like poverty are probably underrepresented.

Another hypothesis is related to the methods currently used by governments to label people as poor or extremely poor. Those labels seem not to appropriately represent real situations that make the people to feel poor and create their own discourse about poverty. We suggest revising the consensus about the meaning of poverty through participatory studies giving voice to the people. The motivation underlying the participatory approach is to draw attention to important aspects of how the people live their life and understand the idea of poverty. Seeing poverty in terms of relevant concerns of the people involved is in line with human experiences that really matter to tackle the phenomena in the context.

**Acknowledgements** This study was partially funded by CODI from the University of Antioquia, grant # PRG2020-34068

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Abdillah, J., Asror, I., Wibowo, Y. F. A., et al. (2020). Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 4(4), 723–729.
- Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Berlin: Springer Science & Business Media.
- Alammar, J. (2020). . The illustrated transformer. <http://jalammar.github.io/illustrated-transformer/>. Accessed: 2020-10-05
- Alkire, S. (2007). The missing dimensions of poverty data: Introduction to the special issue. *Oxford development studies*, 35(4), 347–359.
- Alkire, S., Roche, J. M., Ballon, P., Foster, J., Santos, M. E., & Seth, S. (2015). *Multidimensional poverty measurement and analysis*. USA: Oxford University Press.
- Arias-Vergara, T., Vásquez-Correa, J.C., Orozco-Arroyave, J.R., Vargas-Bonilla, J.F., Nöth, E. (2016). Parkinson's disease progression assessment from speech using gmm-ubm. In *Interspeech*, pp. 1933–1937
- Ayush, K., UzKent, B., Burke, M., Lobell, D., Ermon, S. (2020). Generating interpretable poverty maps using object detection in satellite images. arXiv preprint [arXiv:2002.01612](https://arxiv.org/abs/2002.01612)
- Banerjee, A.V., Banerjee, A., Duflo, E. (2011). . Poor economics: A radical rethinking of the way to fight global poverty. Public Affairs
- Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). . On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623
- Biggeri, M., & Santi, M. (2012). The missing dimensions of children's well-being and well-becoming in education systems: Capabilities and philosophy for children. *Journal of Human Development and Capabilities*, 13(3), 373–395. <https://doi.org/10.1080/19452829.2012.694858>
- Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone meta-data. *Science*, 350(6264), 1073–1076.
- Boyd, R.L., Schwartz, H.A. (2020). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology* p. 0261927X20967028
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Canete, J., Chaperon, G., Fuentes, R., Pérez, J. (2020). . Spanish pre-trained bert model and evaluation data. PML4DC at ICLR 2020
- Caplan, M. A., Purser, G., & Kindle, P. A. (2017). Personal accounts of poverty: A thematic analysis of social media. *Journal of Evidence-Informed Social Work*, 14(6), 433–456.
- Chen, T., Guestrin, C. (2016). . Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794

- Chiquito, A. B., Pinardi, L. C., & Llull, G. (2019). *La pobreza en la prensa*. Palabras claves en los diarios de Argentina, Brasil: Colombia y México. CLACSO.
- Departamento Nacional de Planeación: Actualización de los criterios para la determinación, identificación y selección de beneficiarios de programas sociales (2008). <https://colaboracion.dnp.gov.co/CDT/Conpes/Social/117.pdf>
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2018) . Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
- Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1), 188–230.
- Engstrom, R., Hersh, J., Newhouse, D. (2017) . Poverty from space: using high-resolution satellite imagery for estimating economic well-being. Working Paper 8284, The World Bank
- Escobar-Grisales, D., Vásquez-Correa, J., Vargas-Bonilla, J. F., Orozco-Arroyave, J. R., et al. (2020). Identity verification in virtual education using biometric analysis based on keystroke dynamics. *TecnoLógicas*, 23(47), 193–207.
- Evans, J. A., & Aceves, P. (2016). Machine translation: Mining text for social theory. *Annual Review of Sociology*, 42, 21–50.
- Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50), 13108–13113.
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2–3), 146–162.
- Jang, B., Kim, I., & Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *PloS One*, 14(8), e0220976.
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790–794.
- Jo, T. (2018). Text mining: Concepts, implementation, and big data challenge, vol. 45. Springer
- Joachims, T. (1998) . Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pp. 137–142. Springer
- Kenter, T., Borisov, A., de Rijke, M. (2016). Siamese CBOW: Optimizing word embeddings for sentence representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pp. 941–951. Association for Computational Linguistics, Berlin, Germany . <https://doi.org/10.18653/v1/P16-1089>. <https://www.aclweb.org/anthology/P16-1089>
- Kozłowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5), 905–949.
- Laderchi, C. R., Saith, R., & Stewart, F. (2003). Does it matter that we do not agree on the definition of poverty? A comparison of four approaches. *Oxford Development Studies*, 31(3), 243–274. <https://doi.org/10.1080/1360081032000111698>.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3), 259–284.
- Ledesma, C., Garonita, O.L., Flores, L.J., Tingzon, I., & Dalisay, D. (2020). Interpretable poverty mapping using social media data, satellite images, and geospatial information. arXiv preprint [arXiv:2011.13563](https://arxiv.org/abs/2011.13563)
- Lee, K., & Braithwaite, J. (2020). High-resolution poverty maps in sub-saharan africa. arXiv preprint [arXiv:2009.00544](https://arxiv.org/abs/2009.00544)
- Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P.S., & He, L. (2020) . A text classification survey: From shallow to deep learning. arXiv preprint [arXiv:2008.00364](https://arxiv.org/abs/2008.00364)
- Li, X., Zhong, J., Wu, X., Yu, J., Liu, X., & Meng, H. (2020) . Adversarial attacks on gmm i-vector based speaker verification systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6579–6583. IEEE
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 3111–3119.
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep learning based text classification: A comprehensive review. arXiv preprint [arXiv:2004.03705](https://arxiv.org/abs/2004.03705)
- Mitra, S., & Jenamani, M. (2020). Hybrid improved document-level embedding (hide). arXiv preprint [arXiv:2006.01203](https://arxiv.org/abs/2006.01203)
- Naraya, D., Patel, R., Schafft, K., Rademacher, A., & Koch-Schulte, S. (2000). *Can anyone hear us? The World Bank: Voices of the poor*.
- Narayan, D., Patel, R., Schafft, K., Rademacher, A., & Koch-Schulte, S. (1999). Can Anyone Hear Us? Voices From 47 Countries. Tech. rep., World Bank . <http://siteresources.worldbank.org/INTPOVERTY/Resources/335642-1124115102975/1555199-1124115187705/voll.pdf>

- Nolan, B., & Whelan, C. T. (2011). *Poverty and deprivation in Europe*. Oxford: Oxford University Press.
- Nussbaum, M.C. (2001). *Women and human development: The capabilities approach*, vol. 3. Cambridge University Press
- Oved, N., Feder, A., & Reichart, R. (2020). Predicting in-game actions from interviews of nba players. *Computational Linguistics*, 46(3), 667–712.
- Pandey, S., Agarwal, T., & Krishnan, N.C. (2018). Multi-task deep learning for predicting poverty from satellite images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 7793–7798 <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16441/16388>
- Pennington, J., Socher, R., & Manning, C.D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543
- Pilehvar, M. T., & Camacho-Collados, J. (2020). Embeddings in natural language processing: Theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4), 1–175.
- PNUD: La verdadera riqueza de las naciones: caminos al desarrollo humano. Tech. Rep. Reporte del desarrollo humano 2010, Programa de las Naciones Unidas para el Desarrollo, New York (2010). [http://hdr.undp.org/sites/default/files/hdr\\_2010\\_es\\_complete\\_reprint.pdf](http://hdr.undp.org/sites/default/files/hdr_2010_es_complete_reprint.pdf)
- Pokhriyal, N., & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, 114(46), E9783–E9792. <https://doi.org/10.1073/pnas.1700319114>.
- Pokhriyal, N., Zambrano, O., Linares, J., & Hernández, H. (2020). Estimating and forecasting income poverty and inequality in haiti using satellite imagery and mobile phone data. Tech. rep., Inter-American Development Bank. <https://doi.org/10.18235/0002466>. <https://publications.iadb.org/en/estimating-and-forecasting-income-poverty-and-inequality-in-haiti-using-satellite-image-ry-and-mobile-phone-data>
- Prabhakaran, V., Hutchinson, B., & Mitchell, M. (2019). Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5740–5745. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1578>. <https://www.aclweb.org/anthology/D19-1578>
- Pulse, U. G. (2014). *Mining indonesian tweets to understand food price crises*. Jakarta: UN Global Pulse.
- Ravallion, M. (2015). *The economics of poverty: History, measurement, and policy*. Oxford: Oxford University Press.
- Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50. ELRA, Valletta, Malta. <http://is.muni.cz/publication/884893/en>
- Reynolds, D. (2009). Gaussian mixture models. In *Encyclopedia of Biometrics*, pp. 659–663
- Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1–3), 19–41.
- Rezaeina, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117, 139–147.
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. arXiv preprint [arXiv:2002.12327](https://arxiv.org/abs/2002.12327)
- Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., et al. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15), 8398–8403.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513–523.
- Salvatore, C., Biffignandi, S., & Bianchi, A. (2020). Social media and twitter data quality for new social indicators. *Social Indicators Research* pp. 1–30
- Sen, A.: *Commodities and Capabilities*. North-Holland, Amsterdam, (1985). *New Delhi: Oxford University Press, 1987; Italian translation: Giuffrè Editore, 1988* (p. 1988). Japanese translation: Iwanami.
- Sen, A. (1999). *Development as freedom*. Oxford: Oxford University Press.
- Sen, A. K. (2009). *The idea of justice*. United States: Harvard University Press.
- Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Burke, M., Lobell, D., Ermon, S. (2019). Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2698–2706

- Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y. A., Iqbal, A. M., et al. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127), 20160690.
- Stein, R. A., Jaques, P. A., & Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. *Information Sciences*, 471, 216–232.
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188.
- Villatoro, P., & Santos, M. E. (2019). *quiénes son pobres? análisis de su identificación en américa latina*. Revista Latinoamericana de Economía: Problemas del Desarrollo.
- Wijffels, J. (2019). Udpipes: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the udpipes nlp toolkit. R package version 0.8 3
- World Bank: Monitoring Global Poverty: Report of the commission on Global Poverty. World Bank, Washington, D.C. (2017). <https://doi.org/10.1596/978-1-4648-0961-3>. <https://openknowledge.worldbank.org/bitstream/handle/10986/25141/9781464809613.pdf>
- Yu, L.C., Wang, J., Lai, K.R., & Zhang, X. (2017). Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 534–539

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.