PeerJ Computer Science

Multi-objective evolutionary optimization for dimensionality reduction of texts represented by synsets

Iñaki Vélez de Mendizabal[1,2], Vitor Basto-Fernandes[2], Enaitz Ezpeleta[1], José R. Méndez[3,4,5], Silvana Gómez-Meire[5] and Urko Zurutuza[1]

[1] Electronics and Computing Department, Mondragon Unibertsitatea, Arrasate-Mondragón, Gipuzkoa, Spain
[2] University Institute of Lisbon ISTAR-IUL, Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
[3] Galicia Sur Health Research Institute (IIS Galicia Sur), Hospital Álvaro Cunqueiro, Bloque técnico, SING Research Group, Vigo, Pontevedra, Spain
[4] CINBIO-Biomedical Research Centre, Lagoas-Marcosende, Vigo, Pontevedra, Spain
[5] Department of Computer Science Universidade de Vigo, Ourense, Spain

ABSTRACT

Despite new developments in machine learning classification techniques, improving the accuracy of spam filtering is a difficult task due to linguistic phenomena that limit its effectiveness. In particular, we highlight polysemy, synonymy, the usage of hypernyms/hyponyms, and the presence of irrelevant/confusing words. These problems should be solved at the pre-processing stage to avoid using inconsistent information in the building of classification models. Previous studies have suggested that the use of synset-based representation strategies could be successfully used to solve synonymy and polysemy problems. Complementarily, it is possible to take advantage of hyponymy/hypernymy-based to implement dimensionality reduction strategies. These strategies could unify textual terms to model the intentions of the document without losing any information (*e.g.*, bringing together the synsets "viagra", "ciallis", "levitra" and other representing similar drugs by using "virility drug" which is a hyponym for all of them). These feature reduction schemes are known as lossless strategies as the information is not removed but only generalised. However, in some types of text classification problems (such as spam filtering) it may not be worthwhile to keep all the information and let dimensionality reduction algorithms discard information that may be irrelevant or confusing. In this work, we are introducing the feature reduction as a multi-objective optimisation problem to be solved using a Multi-Objective Evolutionary Algorithm (MOEA). Our algorithm allows, with minor modifications, to implement lossless (using only semantic-based synset grouping), low-loss (discarding irrelevant information and using semantic-based synset grouping) or lossy (discarding only irrelevant information) strategies. The contribution of this study is two-fold: (i) to introduce different dimensionality reduction methods (lossless, low-loss and lossy) as an optimization problem that can be solved using MOEA and (ii) to provide an experimental comparison of lossless and low-loss schemes for text representation. The results obtained support the usefulness of the low-loss method to improve the efficiency of classifiers.

Submitted 20 October 2022
Accepted 13 January 2023
Published 8 February 2023

Corresponding author
José R. Méndez,
moncho.mendez@uvigo.es

Academic editor
Bilal Alatas

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj-cs.1240**

How to cite this article Vélez de Mendizabal I, Basto-Fernandes V, Ezpeleta E, Méndez Reboredo JR, Gómez-Meire S, Zurutuza U. 2023. Multi-objective evolutionary optimization for dimensionality reduction of texts represented by synsets. **PeerJ Comput. Sci. 9:e1240 DOI 10.7717/peerj-cs.1240**

Subjects Artificial Intelligence, Computational Linguistics, Natural Language and Speech, Optimization Theory and Computation
Keywords Spam filtering, Synset-based representation, Semantic-based feature reduction, Multi-Objective Evolutionary Algorithms

INTRODUCTION

In a very few years, the Internet has established itself as one of the fast-growing and most transformative technologies, changing the way we do business and the way we communicate. In 2021, more than 4.6 billion people were taking advantage of the large number of online tools available to make our lives easier, a huge increase from only 1.1 billion connected (*Statista, 2022*) users in 2005. In a study (*Ali, 2020*) analyzing online Internet traffic in 1 min, it can be clearly identified that the applications most frequently used by users are those related to text messages. The resulting statistics in 60 s are clear: Facebook users upload 147,000 photos, 208,333 hosts participate in Zoom meetings, people make 1.4 billion calls and WhatsApp users share around 42 billion messages. However, just as legitimate users have derived substantial benefits from the use of the Internet, many malicious users have also abused the network for their own profit at the expense of the user experience of others. In particular, a large amount of inappropriate content (spam) has been distributed through communication services based on the exchange of text messages, including instant messaging (*Cabrera-León, García Báez & Suárez-Araujo, 2018*; *Silva et al., 2017*), email (*Chakraborty et al., 2016*; *Suryawanshi, Goswami & Patil, 2019*) or social networks (*Chakraborty et al., 2016*; *Xu, Sun & Javaid, 2016*).

Machine learning (ML) has been very helpful in the fight against spam, mainly for being able to use past experiences and relative information as input to classify messages. To take advantage of ML techniques, texts should be represented in a matrix in which instances are arranged as rows (feature vector), and specific features as columns. Selecting an adequate way to represent the texts (*i.e.*, to select specific features) is very important, since this determines the filtering efficiency and may reduce the computational capacity required for running the classifier.

Some generic text representation methods have been successfully exploited in different text analysis problems. In particular, some studies have used Bag of Words (BoW) representations (*Novo-Lourés et al., 2020*; *Sahin, Aydos & Orhan, 2018*). These models represent each text as a feature vector using words that are included in the form of term frequencies (TF), inverse document frequencies (IDF) or even binary forms such as the term presence (TP). Instead of using words/tokens as features (columns), some studies have explored the use of other type of characteristics including: (i) character n-grams (*Aiyar & Shetty, 2018*), (ii) word n-gram (*Lopez-Gazpio et al., 2019*), (iii) word embeddings (*Barushka & Hajek, 2019*), (iv) topic-based schemes (*Li et al., 2021*) and (v) synset based representations (*Almeida et al., 2016*; *Bahgat & Moawad, 2017*; *Méndez, Cotos-Yañez & Ruano-Ordás, 2019*; *Vélez de Mendizabal et al., 2020*).

Synset-based representation methods are the most recent advance in the context of text representation and take advantage of the synset (synonym set) concept. In this case, each

word (or sequence of words) included in the text is replaced by a synset identifier. A synset identifier represents any synonym of a specific word or sequence of words and achieves similar representations for sentences such as "take an aspirin every 8 h for cephalalgia" and "for headache you should take a salicylic acid tablet every 8 h" (because "salicylic acid" and "aspirin" are synonyms and therefore are included in the same synset). Synset information used for representation is obtained from semantic graphs/ontologies available online (such as WordNet (*Princeton University, 2010*) or BabelNet (*Sapienza, 2012*). Recent improvements in NLP have helped us to deal with the translation of polysemic words into synsets with a high degree of accuracy (*Scozzafava et al., 2015*) through word sense disambiguation (WSD) (*Moro, Raganato & Navigli, 2014*).

Several recent publications have shown that synset-based systems achieve better performance than token methods (*Vélez de Mendizabal et al., 2020*) and topic representations (*Méndez, Cotos-Yañez & Ruano-Ordás, 2019*). In fact, unlike other semantic-based representations (based on word embeddings or topics), synset-based schemes take advantage of semantic knowledge created and manually revised by research communities. Finally, synset-based representations overcome two of the main issues of natural language processing (NLP): polysemy and synonymy (*Bahgat & Moawad, 2017*).

One of the most important advances in the context of synset-based representation is the introduction of new semantic-based feature reduction approaches that are able to combine two or more related synset features into a single one (*Méndez, Cotos-Yañez & Ruano-Ordás, 2019*; *Bahgat & Moawad, 2017*). Although the first proposals had serious shortcomings, a recent study (*Vélez de Mendizabal et al., 2020*) has demonstrated that it is possible to reduce the dimensionality on datasets represented in this way without losing any information (lossless). The idea of a lossless information dimensionality reduction system seems appropriate but we realized that actually some noise and/or irrelevant synsets found in documents should probably be discarded (low-loss). Therefore, we introduce a (low-loss) approach capable of preserving valuable information but also allowing the elimination of undesirable data. We also design and execute an experimental protocol to check whether the low-loss scheme is better for classification tasks and, in particular, for spam filtering.

The goals of this study are three-fold: (i) to discuss and analyze the limitations of lossless dimensionality reduction methods for spam filtering, (ii) to introduce of three wrapper sementic-based feature reduction schemes (lossless, low-loss and lossy) as optimization problems, and (iii) to execute an experimental benchmark to compare the previous successful lossless approach with the low-loss approach introduced in this work.

The remainder of the study is structured as follows: the next section presents the state of the art in the use of semantic information to reduce the dimensionality of synset-based representation of datasets. The Problem Formulation section presents the formulation of the different dimensionality reduction schemes as an optimization problem and provides a discussion about the differences between low-loss and lossless approaches. The Experimental Protocol section describes the dataset used as input *corpus*, the preprocessing configuration and the parameter settings. Then, the Results and Discussion section

analyses the results of the experiments performed. Finally, last section summarizes the main results and future research directions.

RELATED WORK

Although there are multiple and different types of ML classifiers for filtering spam messages, all of them are affected by the problem of the high dimensionality of the feature space (*Shah & Patel, 2016*). Text messages often contain a large number of words or n-grams (both could be represented as synsets) which are considered features for the classification process. Many of these features can be redundant, less informative or noisy making the process of classifying messages more difficult.

Feature selection (*Kalousis, Prados & Hilario, 2007*) is one of the most commonly used techniques for removing relevant features from text during data pre-processing. Feature selection methods consist of reducing the total number of input variables by selecting the subset of variables that equal or better describe the underlying structure of the text (*Salcedo-Sanz et al., 2004*). Some well-known feature selection methods (*Trivedi & Dey, 2016*), such as the popular Information Gain (IG) and/or Document Frequency (DF), have been widely used to identify and eliminate low quality features. However, there are a wide variety of methods that can be used to address feature selection. According to the classification proposed in (*Chandrashekar & Sahin, 2014*), these can be divided into three main categories: (i) filtering methods (*Blum & Langley, 1997*), used to calculate the relevance of each variable according to an evaluation function that is based only on the properties of the data; (ii) wrapped methods (*Kohavi & John, 1997*), which use the performance of a classification algorithm as a quality criterion; and finally, (iii) embedded methods, which integrate the selection process into the learning of the classifier.

Filter-based methods are a suitable mechanism for extracting features from big datasets with a large number of features. Although filtering methods obtain fast and reliable generalisation results, discarding features based strictly on their significance value can lead to a reduction in classification performance methods. Wrapper-based methods use the performance of a (possibly non-linear) classification algorithm as an objective function to evaluate the amount of relevant information collected by a subset of features. These methods have the ability to outperform filtering strategies in terms of classification error and to take feature dependencies into account. However, wrapper methods are usually computationally demanding. Finally, embedded methods emerged to combine the benefits of filter-based and wrapper-based methods. To do so, they act as a trade-off between these two models by including feature selection in the model generation process. This improves the results obtained by the filtering methods and, at the same time, reduces the computational cost of the wrapper methods by performing multiple runs of the learning model to evaluate the features. The main disadvantage of these methods is their dependence on the learning model due to their use within the feature selection process.

The use of synsets is quite new and only a few synset-based dimensionality reduction schemes are available to deal with dimensionality reduction in text datasets. The first dimensionality reduction approach was able to take advantage of synsets to group synonym words into single features (*Bahgat & Moawad, 2017*). Using this scheme, when

two or more terms from the same text are included in a synset, they are represented within the same feature. This procedure slightly reduces the number of features compared to a token-based representation.

Later, a new dimensionality reduction method that was also able to take advantage of hypernymy/hyponymy relationships was introduced (*Méndez, Cotos-Yañez & Ruano-Ordás, 2019*). They exploited Wordnet taxonomic relations to generalize words into more abstract concepts (for instance, Viagra could be generalized into "anti-impotence drug", "drug" or "chemical substance"). Considering Wordnet as a taxonomy (*i.e.*, a tree of synsets) and "entity" as its root, the authors used the 181 synsets in the first four levels (168 level-4, 10 level-3, 2 level-2 and 1 level-1) as features for concept identification. Synsets extracted from the text were generalized using hypernymy relations until they match with some of the selected features. This approach reduced the number of features used for representing text to a maximum of 181, which is effective for associating a text with a subject or concept. The main weakness of this algorithm is that large texts, even with thousands of words, are reduced to a fixed number of features (181), which could lead to the loss of relevant information.

SDRS (Semantic Dimensionality Reduction System) has recently been introduced (*Vélez de Mendizabal et al., 2020*). Instead of using Wordnet, it takes advantage of the BabelNet ontological dictionary, supports multiple size n-gram matching and is able to adjust the dimensionality to optimize the performance without loss of information. The SDRS dimensionality reduction method uses multi-objective evolutionary algorithms (MOEA) to identify the maximum level to which each synset can be generalized, while preserving and even increasing the classification performance. The NSGA-II (non-dominated sorting genetic algorithm) algorithm (*Verma, Pant & Snasel, 2021*) was adopted to optimize a problem that was represented as follows: (i) each chromosome represents a possible reduction configuration, thus an integer vector of size n (n = number of features) that defines how many levels should be generalized each synset; (ii) the synsets to be generalized are replaced by the corresponding hypernyms based on the optimization results; (iii) three fitness functions are applied to evaluate each configuration from the perspective of classification performance and dimensionality reduction. The main limitation of this approach is that some unhelpful synsets (sometimes without hypernyms and/or adjectives that do not add information) remain as features. Such words do not provide information in the classification process or might even introduce noise. This fact suggests that the performance of SDRS could be improved by allowing the removal of these features instead of using a pure lossless approach. The following section presents a formulation of the above dimensionality reduction strategies so that they can be implemented as optimization problems.

FORMULATING DIMENSIONALITY REDUCTION PROBLEM AS AN OPTIMIZATION PROBLEM

As stated above, in order to address text classification tasks, each text $T = (a_1, a_2, \ldots, a_n, d)$ is represented as a vector containing integer values (number of occurrences) for a list of synset-based attributes ($A = \{a_1, a_2, \ldots, a_n\}$) and a value d which represents the target

class. A *corpus* can be represented as a matrix M in which each row contains the representation of a text together with the target class attribute. This representation of information is suitable for performing classification tasks on the synset-based attributes.

A feature reduction scheme may involve (i) eliminating some irrelevant attributes, (ii) grouping two or more related attributes or (iii) both. The first strategy corresponds to a lossy approach, while the second one is a lossless scheme and the last one corresponds to a low-loss method. Derived from previous works (*Méndez, Cotos-Yañez & Ruano-Ordás, 2019*; *Vélez de Mendizabal et al., 2020*) the grouping attributes strategy (ii) is guided by taxonomic relations (hypernym/hyponym) between synsets. In this section we introduce a formulation to address the dimensionality reduction approaches mentioned above as an optimization problem. In particular, we formulated a multi-objective optimization problem that could be solved by evolutionary algorithms.

A multi-objective optimization problem can be presented as a simultaneous optimization of i objective functions $f = (f_1, f_2, \ldots, f_i)$, such that $f_k, k \in 1, \ldots, i$ is a real-valued function evaluated in decision space (minimization of all functions is assumed). Some constraints of equality or inequality type can be imposed to the optimization problem on the decision variables y by the domain definition of objective functions or on the objective functions range: $f_k(y) \leq c_k$, where f_k is a real valued function of a vector of decision variables y, and c_k is a constant value.

In our study, the optimization problem consists of finding a vector of integers $V = \{v_1, v_2, \ldots, v_n\}$ that determines the optimal action that should be carried out for each synset attribute a_i to minimize the number of features (goal 1) and ensure that the classifiers achieve the best performance (goal 2). However, due to its complexity, the last one has been broken down into two simpler objectives: the reduction of false positive (FP) and false negative (FN) errors. Each action v_i can have one of the following values: (i) -1 to remove the attribute, (ii) 0 to keep the attribute without change, or (iii) another integer value ($0 < m < \gamma$) that implies replacing the synset feature by its hypernym, after m generalization steps. γ is a parameter that specifies the maximum number of generalization steps to be done.

The grouping of objectives has been designed taking into account that optimizing (minimizing) dimensionality would lead to higher error rates (both for FP and FN), as they are conflicting objectives (*Basto-Fernandes et al., 2016*). In order to address feature reduction as an optimization problem, we defined three fitness functions to be minimized simultaneously, which are detailed in Eq. (1).

$$\begin{aligned} f_1 &= \frac{num\_cols(T(M,V))}{num\_cols(D)} \\ f_2 &= 10xval\_eval(c, T(M,V)).FPr \\ f_3 &= 10xval\_eval(c, T(M,V).FNr \end{aligned} \tag{1}$$

where $T = (M, V)$ is the transformation of the dataset matrix M using the vector of changes V, $10xval\_eval(c, T(M,V)).FPr$ and $10xval\_eval(c, T(M,V)).FNr$ represent the false positive ratio (FPr) and false negative ratio (FNr), respectively. FPr and FNr are

calculated using a 10-fold cross validation scheme of the classifier c, applied to the dataset represented as a matrix (M). To improve the readability of the text, dimension ratio ($DIMr$), FPr and FNr are used in the rest of the manuscript to denote the functions f_1, f_2 and f_3 respectively.

Transforming the dataset Matrix M according to a vector of transformations V involves the following steps: (i) removing the columns that are marked for deletion (-1), (ii) performing attribute generalizations according to the transformation vector and (iii) adding columns and merging them where necessary. In the first step, any attribute a_i marked with -1 (*i.e.*, $v_i = -1$), causes v_i column removal for all instances of the dataset in the matrix M. Then, in the second step, the remaining attributes a_i marked for generalization ($v_i > 0$) are replaced by their hyponyms, corresponding to v_i generalization steps. This causes attributes semantically close to the original form of become direct or indirect hyponyms of its transformation. During the last step, a set of attributes $A' \subseteq A, A' \cap \{a_i\} = \varnothing$ are merged into the same attribute a_i if and only if $\forall A'_j \in A'$, $A'_j \in hyponyms(a_i)$, where $hyponyms(a_i)$ is the set of direct and indirect hyponyms of a_i.

The formulation provided in this study allows the representation of different dimensionality reduction schemes: a lossy approach (which is based on feature removal) could be implemented using only -1 and 0 values for v_i (that is $-1 \leq v_i \leq 0$), a lossless scheme is obtained when the removal of features is avoided ($0 \leq v_i \leq \gamma$) and finally, there are no limitations on the low-loss strategy ($-1 \leq v_i \leq \gamma$). Additionally, we are exploring whether pure lossless approaches are the best way to address dimensionality reduction in synset-based text classification approaches. Before this study, lossless schemes (*Vélez de Mendizabal et al., 2020*) seemed to be the best alternative. However, we have found that some synsets, even when combined with others, do not provide relevant information for the classification process and may even be noisy. This finding led us to think that a lossless reduction process might not be the most appropriate solution and that it would probably be worthwhile to allow some synsets (features) to be removed in cases where this action would improve the classifier results. The idea behind this new approach (low-loss) is that the optimization process should include not only the possibility to group features, but also to remove them. The next section provides a detailed comparison between the lossless method and low-loss feature reduction approaches.

EXPERIMENTAL DESIGN

This section describes the experimentation performed to determine whether a low-loss configuration is more adequate than the traditional lossless ones in the context of spam filtering. The following subsections provide additional configuration details of the experimental protocol, including the selection of the target *corpus*, the configuration of the preprocessing steps, the details of experimental protocol and the selection of the optimization process configuration parameters.

Selecting a dataset from available corpora

As stated in recent works (*Vélez de Mendizabal et al., 2020*; *Novo-Lourés et al., 2020*; *Vázquez et al., 2021*), there are a large number of public available corpora that can be used for testing new spam classification proposals. After reviewing the datasets reported in them, we have selected the Youtube Spam Collection dataset (*Alberto & Lochter, 2017*) which was also the one chosen for the preceding SDRS study. Computing two of the fitness functions defined in our proposal involves running a 10-fold cross-validation test on the instances selected for optimization (75% of the input *corpus*). Moreover, our dimensionality reduction schemes are time-consuming processes performed using a stochastic method (genetic algorithm) configured for 25 executions and 25,000 fitness function evaluations. With these considerations in mind, it is necessary to use a small dataset. Thus, the Youtube Spam Collection dataset seems to be suitable for the experimental process as it contains only 1,956 short messages.

Preprocessing configuration

Considering the raw nature of the messages within the YouTube Spam Collection Dataset, we preprocessed the dataset using the Big Data Pipelining for Java (BDP4J) (*Novo-Lourés et al., 2021*) and the Natural Language preprocessing Architecture (NLPA) (*Novo-Lourés et al., 2020*) projects.

The preprocessing of the instances was started by extracting the content of the messages through the YouTube API. These messages were then cleaned up by removing all of their HTML, CSS, URL and JavaScript tags. Emojis and emoticons were also deleted from the messages, as well as stop words, interjections, contractions, abbreviations and slang expressions. Figure 1 contains a detailed representation of the preprocessing pipeline used in our experimentation.

One of the major problems in Natural Language Processing is the polysemy of words included in texts (*e.g.*, the word "break" can refer to an "interruption", "fracture", "recess" and 50 more meanings/synsets). Therefore, disambiguation is a critical preprocessing step to improve classification accuracy. With this in mind, the translations of texts into synsets were carried out using the Babelfy (*Moro & Navigli, 2010*) tool, which identifies the right BabelNet synset for each word/n-gram. For each word or sequence of words in particular, the tool selects the synset that best fits in the context of the target sentence. As Babelfy identifies the whole list of tokens and n-grams (*e.g.*, "computer network" is usually transformed into "computer", "network" and the bi-gram "computer network"), the identification of n-grams requires an additional post-processing of the output information. Thus, when the list of identified synsets contains n-grams, the largest n-gram is selected. Furthermore, we should bear in mind that the translation of two or more different tokens could result in the same synset, leading to an initial dimensionality reduction.

After the preprocessing and the synset translation process (Fig. 1) 1,684 columns were obtained for the representation of the target dataset. This result is better than the one achieved if we replace the synonym translation process (last two stages of preprocessing) by a tokenisation scheme. Using this scheme, the dataset would be represented using 2,279 columns.

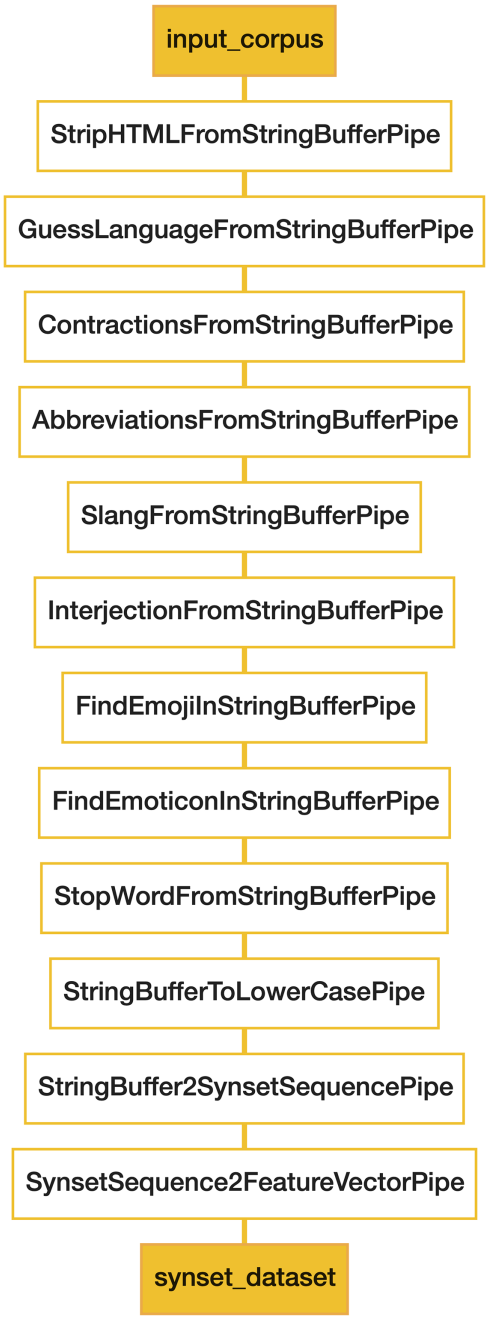

Figure 1 **Preprocessing configuration.** Full-size DOI: 10.7717/peerj-cs.1240/fig-1

Experimental protocol

This subsection introduces the experimental protocol designed to compare the performance of low-loss and lossless feature reduction approaches, and to check that no overfitting occurs. Compared to the original lossless approach, where only synsets can be generalized, the low-loss dimensionality reduction scheme introduced in this work allows the generalization and removal of synsets. The most serious drawback is that both approaches could lead to an overfitting of the result when the classifier adjusts too closely

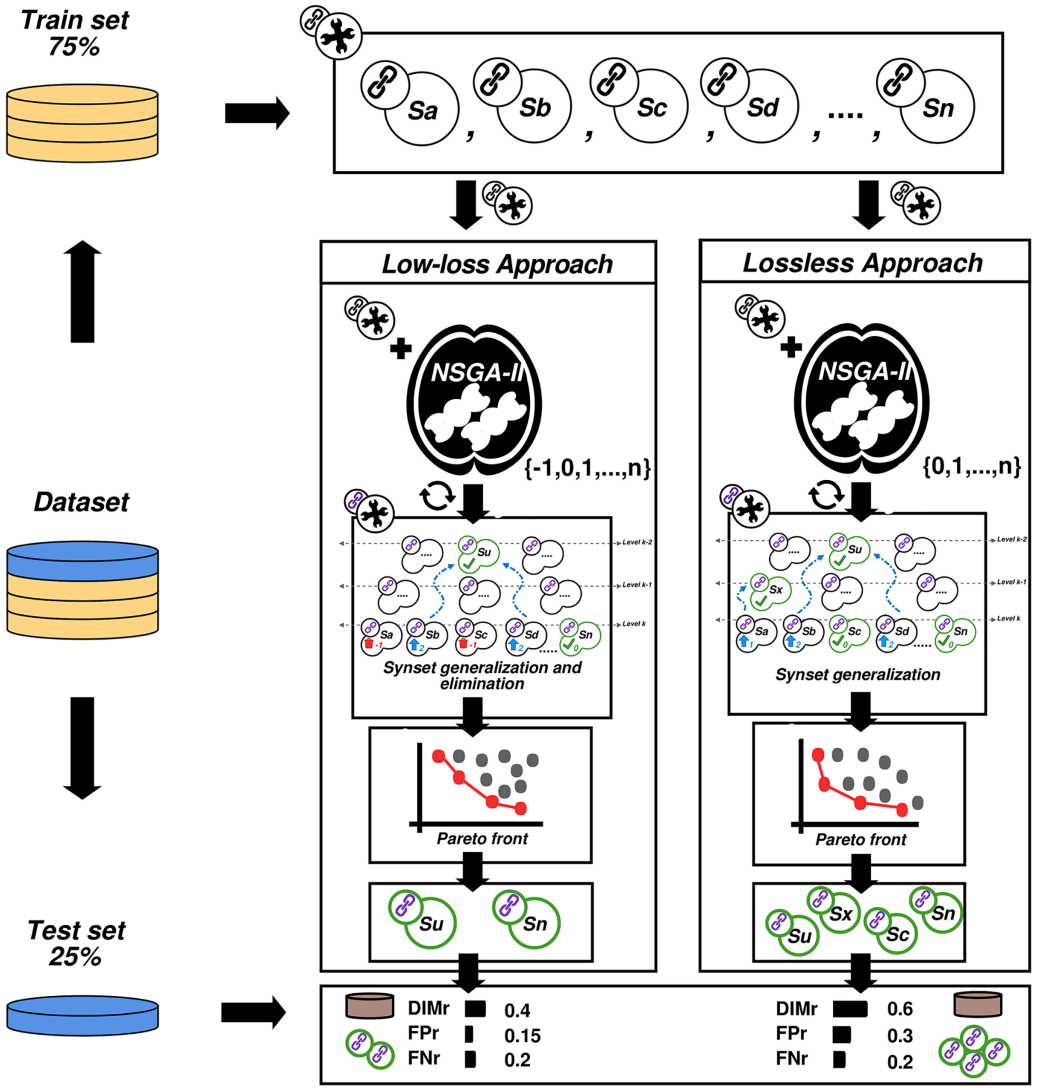

Figure 2 **Experimental protocol.** Full-size DOI: 10.7717/peerj-cs.1240/fig-2

to the curve of the input data. To evaluate and avoid the problem of overfitting in machine learning, 25% of the input data was reserved for a final test while the remaining 75% was used to run the optimization process. Figure 2 provides a graphical representation of the designed experimental protocol.

As shown in Fig. 2, the designed experimental protocol computes the dimensionality reduction for the low-loss and lossless approaches and compares their performances. The same dataset was used to analyze both approaches (using a 75/25% split as explained above). Among the set of solutions (set of chromosomes provided by the selected MOEA), the five that achieved the lowest error rates (*FPr*, *FNr*) were selected to be used in the next step of the experiments. Using the best 10 solutions obtained in the first stage of the experimental protocol (five low-loss and five lossless), we executed a new training/testing experiment using the whole dataset. The 25% of the not yet used instances were taken as

Table 1 Analysis of different configurations using the Euclidean distance criterion.

Values for γ	Minimum Euclidean distance
0 (without optimization)	0.3949
[−1,0] (lossy approach)	0.3157
[−1,1]	0.3241
[−1,2]	0.312
[−1,3]	0.3148
[−1,4]	0.3685

test data and the results of the experiment allowed us to detect overfitting and other possible weaknesses in the processes.

Optimization process configuration

Since the dimensionality reduction schemes to be compared were defined as optimization problems, we have selected NSGA-II to address it. NSGA-II is a MOEA algorithm that has been used in many different types of problems (*Verma, Pant & Snasel, 2021*; *Goldkamp & Dehghanimohammadabadi, 2019*; *Robles, Chica & Cordon, 2020*; *Turk, Özcan & John, 2017*). The JMetal Framework (*Durillo & Nebro, 2008*) implementation of NSGA-II was chosen for experimentation. The experiments were configured to execute 25 independent runs (optimization process) with a maximum of 25,000 function evaluations. For the remaining settings, default JMetal configurations were used. In particular, the NSGA-II population size was set to 100 and the integer operators SBXCrossover and PolynomialMutation were configured with a crossover probability of 1.0 and a mutation probability of 1/NumberOfVariables respectively.

MultinomialNaïveBayes implementation from Weka (*Witten et al., 2016*) was selected as the classifier to compute fitness functions (*FPr*, *FNr*).

To select the most suitable value for γ parameter used in the low-less approach, an empirical evaluation of its performance was carried out using configuration values in the range of $[1..4]$. For this evaluation, we only considered *FPr* and *FNr* objectives (*DIMr* was not considered). These values (intervals) were tested using the whole Youtube Spam Collection dataset. Table 1 includes the Euclidean distance of the closest solution to the origin or coordinates $(0, 0)$ for different configurations of γ.

The selected configuration was the one which fitness evaluation (FPr, FNr) has the smallest Euclidean distance to the origin of coordinates. Using this criterion, Table 1 shows that the best configuration found corresponds to $\gamma = 2$ (*i.e.*, $-1 \leq v_i \leq 2$). For the configuration of the γ value for the lossless approach (SDRS), the value defined in the original study ($\gamma = 3$) was used, ensuring that values of v_i were included in the interval $[0..3]$.

RESULTS AND DISCUSSION

Once the parameters were set as detailed in the previous section, we executed the designed experimental protocol (Fig. 2) and analyzed in detail the solutions generated by NSGA-II

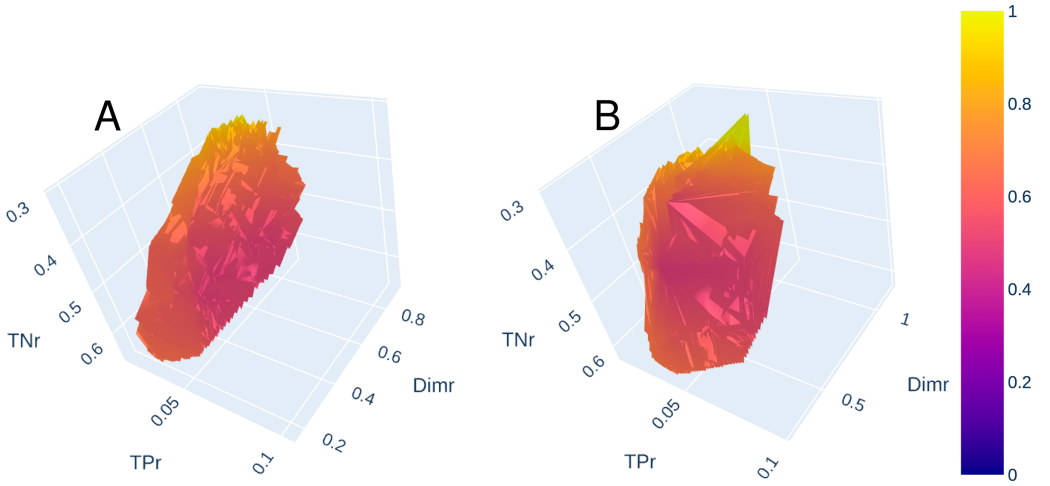

Figure 3 **Pareto fronts for low-loss (A) and lossless (B) approaches.**
Full-size DOI: 10.7717/peerj-cs.1240/fig-3

in the decision space (best generated chromosomes). Considering only the feature reduction objective, the number of columns resulting after applying the solutions achieved with low-loss and lossless processes are within the ranges (318–1,541) and (438–1,434) respectively. However, these solutions should not be analysed alone, since a drastic reduction in dimensionality could lead to an excessive number of errors. Therefore, the Pareto curves have been plotted so that we have a view of the trade-offs between the different objectives. Figures 3A and 3B show the Pareto fronts achieved by low-loss and lossless approaches respectively. Regions plotted in a color close to red are the nearest ones to the origin of coordinates.

Comparing both figures, we can observe that non-dominated solutions are more distributed in the low-loss approach. Therefore, through the 25 executions and 25,000 evaluations in the optimization phase of the low-loss approach, a larger space of solutions was explored and identified. The most significant difference is on the *DIMr* axis, where solutions under the value of 0.2 are found. In addition, non-dominated solutions reached by both proposals were analyzed in depth, sorted by *DIMr* and plotted in a multiple line chart (Fig. 4).

As can be observed in Fig. 4, *FPr* values of all evaluated configurations are close to 0 (most of them are under 0.1) for both approaches. *FNr* values follow the same behavior and are included in the interval 0.3 and 0.7. However, the low-loss approach clearly achieves better *DIMr* values (up to values under 0.2), which is associated with a limited impact on *FNr* evaluations. Furthermore, *FPr* shows a relatively independent behavior with respect to *DIMr*, revealing that optimality conditions can be preserved by *DIMr* and *FNr* trade-offs.

The 10 configurations that obtained the best *DIMr* for lossless and low-loss approaches were compared using the test part of the *corpus* (25%). In all cases, the Naïve Bayes Multinomial classifier model was built on the training data set (75% input data) and then

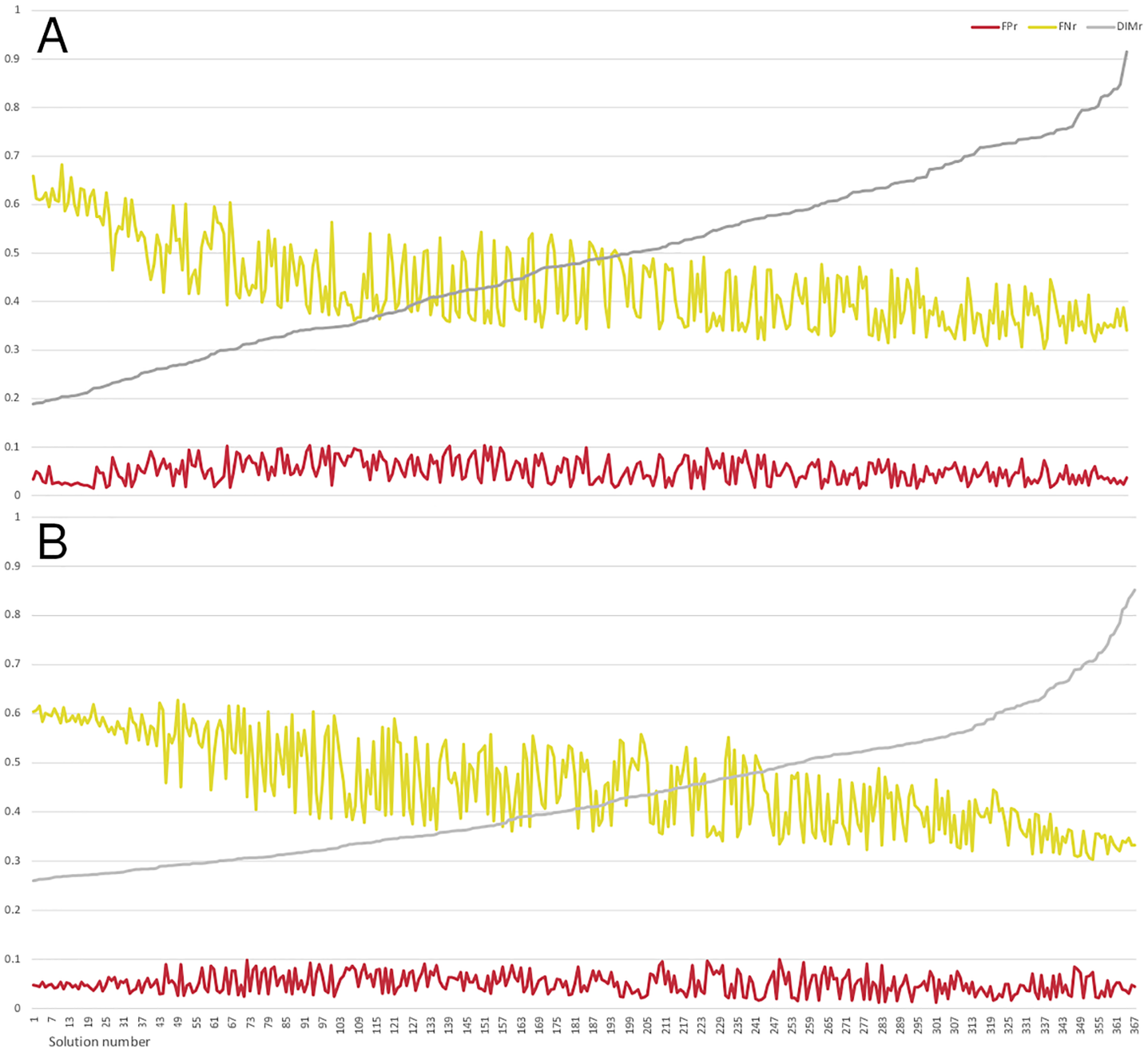

Figure 4 **Multiple line chart representing low-loss (A) and lossless (B) approach solutions.** Full-size DOI: 10.7717/peerj-cs.1240/fig-4

applied to the test set (25%) to get the results. Figure 5 shows the results of this comparison.

The performance achieved by both approaches in terms of *FPr* and *FNr* is quite similar. However, the *DIMr* evaluation function obtains lower values for the low-loss approach.

Moreover, we have analyzed the deleted synsets and their relevance (or influence) in generating optimal solutions to ensure that the low-loss approach worked properly and the

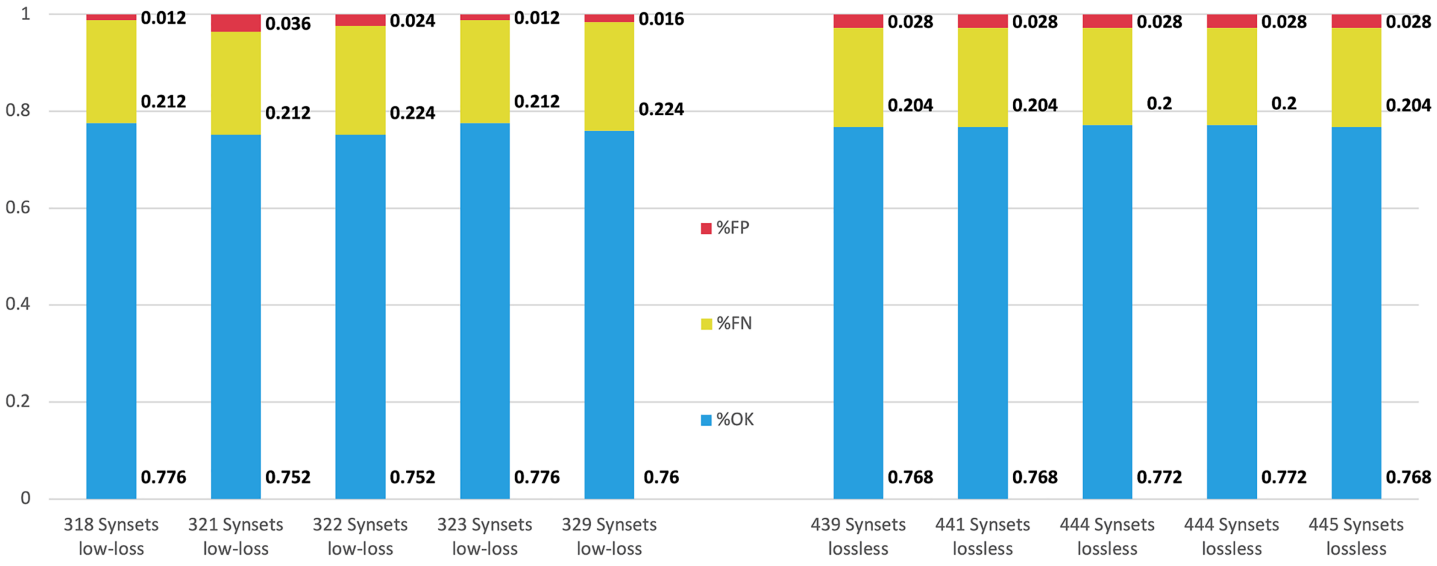

Figure 5 **Performance comparison for the top five configurations achieved by low-loss and lossless approaches.**
Full-size DOI: 10.7717/peerj-cs.1240/fig-5

Table 2 **Information Gain for the most often removed synsets.**

Synset	% of solutions where the synset is marked for removal	Information Gain	Meaning
bn:00110036a	0.2361	0	Sad
bn:14838845n	0.2278	0	Patrik
bn:00104384a	0.2115	0	Human
bn:00082684v	0.2153	0	Dress
bn:00085250v	0.2112	0	Code
bn:00061695n	0.2029	0	Perry
bn:00104562a	0.1988	0	Illegal
bn:00086717v	0.1988	0	Detest
bn:00083286v	0.1988	0	Happen
bn:00076203n	0.1946	0	Tatto

removed synsets were indeed irrelevant or noisy. Table 2 shows the rate of solutions where the synset is selected for removal, the Information Gain (IG) value and the synset meaning for each of the 10 synsets that were marked for removal most often in the solutions included in the Pareto front of the low-loss approach.

Furthermore, Table 2 reveals that the most frequently removed synsets correspond to useless terms (terms that have a similar probability of being part of spam and ham contents and therefore have a poor IG evaluation). Among the most highlighted synsets, some names of people (*e.g.*, "Patrik") and adjectives (*e.g.*, "sad", "human", "illegal") were identified. In the opposite, and to check that the algorithm has kept the relevant synsets (those that obtain a high evaluation for IG metric), Table 3 shows the same information as in Table 2 but for the 10 most usually remaining synsets.

Table 3 Analysis of the synsets achieving best IG evaluation.

Synset	% of solutions where the synset is marked for removal	Information Gain	Meaning
bn:00017681n	0.0455	0.0903	Channel
bn:00094545v	0	0.087	Subscribe
bn:00008378n	0.0289	0.0354	Cheque
bn:00088421v	0.0165	0.0233	Follow
bn:00032558n	0.0248	0.0226	Eyeshot
bn:00066366n	0.0414	0.0184	Subscriber
bn:00103299a	0.0372	0.017	Free
bn:00094547v	0.0414	0.0155	Take
bn:00042306n	0.0124	0.0129	Guy
bn:00055644n	0.0331	0.0123	Money

Table 4 POS analysis of results achieved by the low-loss approach.

Synset type	Composition dataset %	Probability of maintenance	Accumulated IG %
a (adjective)	14.01	0.1175	7.78
r (adverb)	2.25	0.0219	2.39
n (noun)	63.24	0.2474	62.27
v (verb)	20.48	0.1630	27.54

The rates of solutions where these synsets are marked for removal are clearly lower than those reported in Table 2. In fact, the bn:00094545v synset (which corresponds to the verb "subscribe" that is most likely to identify spam messages) is never removed.

Finally, we carried out an analysis of the probability of deleting a synset depending on its part of speech (POS). Table 4 shows the presence rates of each POS in the *corpus*, the probability of keeping a synset when it has a given POS and the distribution of the Information Gain for each POS.

From Table 4 we can conclude that the probability of maintaining the synset type is clearly aligned with the Information Gain (*e.g.*, nouns are kept more frequently and their accumulated IG is clearly greater). Moreover, the results also reveal that nouns and verbs are the most influential POS for spam filtering.

Theoretical and practical implications

Dimensionality reduction when using synset-based approaches has been explored in some recent works (see Related Work section). The complexity of the solutions proposed for this task has evolved over time and the application of optimization strategies to solve the problem has been introduced recently (*Vélez de Mendizabal et al., 2020*). However, all previous studies have focused on obtaining lossless strategies in which at most two or more features (columns) could be combined. The present study assumes that, in this context, a purely lossless strategy may not be the best way to reduce dimensionality as there may be synsets identified in the text that are simply useless or even confusing.

The conclusion of this study is that the combination and the elimination of features should be mixed to increase effectiveness in reducing the number of features required to represent texts in classification tasks. Effectiveness improvements refer to improving the accuracy of the classifiers to be used as well as to reducing dimensionality as far as possible. In detail, to address dimensionality reduction as an optimization problem (*i.e.*, following the approach presented in (*Vélez de Mendizabal et al., 2020*)), it is desirable that the choice of the form for representing the problem (chromosomes) allows the application of both strategies (combining/removing). In this way, we are allowing the algorithms to find a solution with an adequate balance between features to be remained, combined and/or eliminated. This conclusion is supported by results of the experiments carried out.

Despite the small size of the dataset used, the execution of the proposed experimental protocol has required the use of a lot of computational resources for a significant amount of time. Moreover, we dedicated special attention to the identification of overfitting situations so that the results can be suitable for this study. The next section summarizes the conclusions drawn from this research and outlines the direction for future work.

CONCLUSIONS AND FUTURE WORK

This study has introduced the formulation of three different dimensionality reduction strategies to use when texts are represented by using synsets (an earlier lossless one called SDRS, a new low-loss one and a lossy one). The strategies were defined as optimization problems. Moreover, we have experimentally compared the use of lossless and low-loss dimensionality reduction approaches. To this end, we have studied the performance of a low-loss dimensionality reduction schemes based on MOEA, and we have validated that the features marked for removal are adequate using IG classic metric. The achieved results reveal that lossless feature reduction schemes can be successfully complemented with approaches for the identification of irrelevant or noisy synsets in order to reduce the computational costs of classification.

The findings of this study are two-fold: (i) pure lossless feature reduction schemes are not the most appropriate in the context of anti-spam filtering and (ii) allowing the feature reduction to discard irrelevant features can contribute to achieve better results.

With regard to the first finding, it should be noted that the main mechanism of pure lossless synset-based feature reduction schemes is the use of taxonomic relations (hypernymy). However, some synsets cannot be merged because they do not have hypernyms. In particular, BabelNet only defines hypernyms for nouns and verbs, so adjectives and adverbs cannot be reduced using these approaches. Despite this, these words can sometimes be successfully removed without affecting the classification results. Moreover, although the low-loss dimensionality reduction scheme does not take advantage of feature evaluation metrics such as IG, the results show that the synsets selected for elimination have, in most cases, a low value for this metric.

The second finding is supported by experimental results which show that a low-loss feature reduction scheme introduced in this study can further reduce the dimensionality achieved by a lossless reduction, without affecting the classification performance (in terms of FP and FN errors). We have also observed that allowing the removal of columns from

the datasets (low-loss approach) contributes to exploring a wider solution space (see Pareto fronts in Fig. 3). This is because combining noisy features with other relevant ones (which have not been previously removed) would probably cause a reduction in classification performance. On the other hand, feature removal leads to a further reduction in dimensionality while preserving the accuracy of the classifier.

The major limitation of lossless and low-loss algorithms is the time required for their execution. In fact, to run our experimentation, each optimization process took 13 days in a computer with 128 gigabytes of RAM memory and 2 × Intel Xeon E5-2640 (Q1/2012) v3 microprocessors (2.6 GHz 8 cores/12 threads). Therefore, the main line of future research should focus on the development of improvements to drastically reduce the required execution time. This will allow dimensionality reduction in larger datasets with lower computational costs. Moreover, we are currently developing support for using other MOEA algorithms including hypervolume-based approaches, decomposition-based algorithms and other recent MOEAs (*Tanabe & Ishibuchi, 2020*).

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by SMEIC, SRA and ERDF (TIN2017-84658-C2-1-R and TIN2017-84658-C2-2-R subprojects of Semantic Knowledge Integration for Content-Based Spam Filtering) and by the Conselleria de Cultura, Educación e Universidade of Xunta de Galicia (Competitive Reference Group—ED431C 2022/03-GRC). The Intelligent Systems for Industrial Systems research group of Mondragon Unibertsitatea (Iñaki Vélez de Mendizabal, Enaitz Ezpeleta, and Urko Zurutuza) is supported by the department of Education, Universities and Research of the Basque Country (IT1676-22). Vitor Basto Fernandes was supported by FCT (Fundação para a Ciência e a Tecnologia) I.P. (UIDB/04466/2020 and UIDP/04466/2020). There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
SMEIC, SRA and ERDF: TIN2017-84658-C2-1-R and TIN2017-84658-C2-2-R.
Conselleria de Cultura, Educación e Universidade of Xunta de Galicia: ED431C 2022/03-GRC.
Universities and Research of the Basque Country: IT1676-22.
FCT (Fundação para a Ciência e a Tecnologia): UIDB/04466/2020 and UIDP/04466/2020.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Iñaki Vélez de Mendizabal conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Vitor Basto-Fernandes conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, coordination, and approved the final draft.
- Enaitz Ezpeleta conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, coordination, and approved the final draft.
- José R. Méndez conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Silvana Gómez-Meire analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Urko Zurutuza conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, coordination, Project funding, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The YouTube Spam Collection Data Set is available at the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection.

The source code used for running the experiments is available at GitHub: https://github.com/sing-group/moea4sdr; Iñaki Velez de Mendizabal, Vitor Basto Fernandes, Enaitz Ezpeleta, José Ramón Méndez Reboredo, Silvana Gómez Meire, & Urko Zurutuza. (2022). Multi-Objective Evolutioary Algorithms for Synset Dimensionality Reduction [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7441851.

REFERENCES

Aiyar S, Shetty NP. 2018. N-gram assisted youtube spam comment detection. *Procedia Computer Science* **132(6)**:174–182 DOI 10.1016/j.procs.2018.05.181.

Alberto T, Lochter J. 2017. YouTube spam collection. UCI machine learning repository. *Available at https://archive.ics.uci.edu/ml/datasets/Youtube+Spam+Collection.*

Ali A. 2020. Here's What Happens Every Minute on the Internet in 2020 (Visual Capitalist). *Available at https://www.visualcapitalist.com/every-minute-internet-2020/* (accessed 19 October 2022).

Almeida TA, Silva TP, Santos I, Gómez Hidalgo JM. 2016. Text normalization and semantic indexing to enhance instant messaging and SMS spam filtering. *Knowledge-Based Systems* **108(3)**:25–32 DOI 10.1016/j.knosys.2016.05.001.

Bahgat EM, Moawad IF. 2017. Semantic-based feature reduction approach for e-mail classification. In: *AISI 2016: Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016,* 53–63.

Barushka A, Hajek P. 2019. Review spam detection using word embeddings and deep neural networks. In: MacIntyre J, Maglogiannis I, Iliadis L, Pimenidis E, eds. *Artificial Intelligence Applications and Innovations*. Vol. 559. Cham: Springer International Publishing, 340–350.

Basto-Fernandes V, Yevseyeva I, Méndez JR, Zhao J, Fdez-Riverola F, Emmerich MTM. 2016. A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification. *Applied Soft Computing* **48(4)**:111–123 DOI 10.1016/j.asoc.2016.06.043.

Blum AL, Langley P. 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence* **97(1)**:245–271 DOI 10.1016/S0004-3702(97)00063-5.

Cabrera-León Y, García Báez P, Suárez-Araujo CP. 2018. Non-email spam and machine learning-based anti-spam filters: trends and some remarks. In: *EUROCAST 2017: Computer Aided Systems Theory–EUROCAST 2017*. Vol. 10671. Cham: Springer, 245–253.

Chakraborty M, Pal S, Pramanik R, Ravindranath Chowdary C. 2016. Recent developments in social spam detection and combating techniques: a survey. *Information Processing and Management* **52(6)**:1053–1073 DOI 10.1016/j.ipm.2016.04.009.

Chandrashekar G, Sahin F. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* **40(1)**:16–28 DOI 10.1016/j.compeleceng.2013.11.024.

Durillo JJ, Nebro AJ. 2008. jMetal Web site. *Available at https://jmetal.sourceforge.net/* (accessed 19 October 2022).

Goldkamp J, Dehghanimohammadabadi M. 2019. Evolutionary multi-objective optimization for multivariate pairs trading. *Expert Systems with Applications* **135(21)**:113–128 DOI 10.1016/j.eswa.2019.05.046.

Kalousis A, Prados J, Hilario M. 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems* **12(1)**:95–116 DOI 10.1007/s10115-006-0040-8.

Kohavi R, John GH. 1997. Wrappers for feature subset selection. *Artificial Intelligence* **97(1)**:273–324 DOI 10.1016/S0004-3702(97)00043-X.

Li J, Lv P, Xiao W, Yang L, Zhang P. 2021. Exploring groups of opinion spam using sentiment analysis guided by nominated topics. *Expert Systems with Applications* **171**:114585 DOI 10.1016/j.eswa.2021.114585.

Lopez-Gazpio I, Maritxalar M, Lapata M, Agirre E. 2019. Word n-gram attention models for sentence similarity and inference. *Expert Systems with Applications* **132(Feb)**:1–11 DOI 10.1016/j.eswa.2019.04.054.

Méndez JR, Cotos-Yañez TR, Ruano-Ordás D. 2019. A new semantic-based feature selection method for spam filtering. *Applied Soft Computing* **76**:89–104 DOI 10.1016/j.asoc.2018.12.008.

Moro A, Navigli R. 2010. Babelfy | Multilingual Word Sense Disambiguation and Entity Linking together! *Available at https://babelfy.org* (accessed 19 October 2022).

Moro A, Raganato A, Navigli R. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics* **2(22)**:231–244 DOI 10.1162/tacl_a_00179.

Novo-Lourés M, Lage Y, Pavón R, Laza R, Ruano-Ordás D, Méndez JR. 2021. Improving pipelining tools for pre-processing data. *International Journal of Interactive Multimedia and Artificial Intelligence* DOI 10.9781/ijimai.2021.10.004.

Novo-Lourés M, Pavón R, Laza R, Ruano-Ordas D, Méndez JR. 2020. Using natural language preprocessing architecture (NLPA) for big data text sources. *Scientific Programming* **2020**:1–13 DOI 10.1155/2020/2390941.

Princeton University. 2010. WordNet. *Available at https://wordnet.princeton.edu* (accessed 19 October 2022).

Robles JF, Chica M, Cordon O. 2020. Evolutionary multiobjective optimization to target social network influentials in viral marketing. *Expert Systems with Applications* **147(5439)**:113183 DOI 10.1016/j.eswa.2020.113183.

Sahin E, Aydos M, Orhan F. 2018. Spam/ham e-mail classification using machine learning methods based on bag of words technique. In: *26th IEEE Signal Processing and Communications Applications Conference, SIU 2018*. Piscataway: IEEE, 1–4.

Salcedo-Sanz S, Camps-Valls G, Perez-Cruz F, Sepulveda-Sanchis J, Bousono-Calzon C. 2004. Enhancing genetic feature selection through restricted search and walsh analysis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **34(4)**:398–406 DOI 10.1109/TSMCC.2004.833301.

Sapienza NLP. 2012. BabelNet®, the largest multilingual encyclopedic dictionary and semantic network. *Available at https://babelnet.org* (accessed 19 October 2022).

Scozzafava F, Raganato A, Moro A, Navigli R. 2015. Automatic identification and disambiguation of concepts and named entities in the multilingual wikipedia. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9336. Cham: Springer, 357–366.

Shah FP, Patel V. 2016. A review on feature selection and feature extraction for text classification. In: *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. Piscataway: IEEE, 2264–2268.

Silva RM, Alberto TC, Almeida TA, Yamakami A. 2017. Towards filtering undesired short text messages using an online learning approach with semantic indexing. *Expert Systems with Applications* **83**:314–325 DOI 10.1016/j.eswa.2017.04.055.

Statista Inc. 2022. Number of internet and social media users worldwide as of july 2022. *Available at https://www.statista.com/statistics/617136/digital-populationworldwide/* (accessed 19 October 2022).

Suryawanshi S, Goswami A, Patil P. 2019. Email spam detection: an empirical comparative study of different ML and ensemble classifiers. In: *Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing, IACC 2019*. Piscataway: IEEE, 69–74.

Tanabe R, Ishibuchi H. 2020. A review of evolutionary multimodal multiobjective optimization. *IEEE Transactions on Evolutionary Computation* **24(1)**:193–200 DOI 10.1109/TEVC.2019.2909744.

Trivedi SK, Dey S. 2016. A comparative study of various supervised feature selection methods for spam classification. In: *ACM International Conference Proceeding Series*. Vol. 04-05-Marc. New York: ACM Press, 1–6.

Turk S, Özcan E, John R. 2017. Multi-objective optimisation in inventory planning with supplier selection. *Expert Systems with Applications* **78**:51–63 DOI 10.1016/j.eswa.2017.02.014.

Vázquez I, Novo-Lourés M, Pavón R, Laza R, Méndez JR, Ruano-Ordás D. 2021. Improvements for research data repositories: the case of text spam. *Journal of Information Science* DOI 10.1177/0165551521998636.

Vélez de Mendizabal I, Basto-Fernandes V, Ezpeleta E, Méndez JR, Zurutuza U. 2020. SDRS: a new lossless dimensionality reduction for text corpora. *Information Processing and Management* **57(4)**:102249 DOI 10.1016/j.ipm.2020.102249.

Verma S, Pant M, Snasel V. 2021. A comprehensive review on NSGA-II for multi-objective combinatorial optimization problems. *IEEE Access* **9**:57757–57791 DOI 10.1109/ACCESS.2021.3070634.

Witten IH, Frank E, Hall MA, Pal CJ. 2016. *Data mining: practical machine learning tools and techniques*. Amsterdam Elsevier: Data Mining: Practical Machine Learning Tools and Techniques.

Xu H, Sun W, Javaid A. 2016. Efficient spam detection across online social networks. In: *Proceedings of 2016 IEEE International Conference on Big Data Analysis, ICBDA 2016*. Piscataway: IEEE, 1–6.