SDRS: A new lossless dimensionality reduction for text corpora

Iñaki Velez de Mendizabal[1], Vitor Basto-Fernandes[2], Enaitz Ezpeleta[1], José R. Méndez[3,4,5], Urko Zurutuza[1]

[1]Electronics and Computing Department, Mondragon Unibersitatea, Arrasate-Mondragón Spain

[2]Instituto Universitário de Lisboa (ISCTE-IUL), University Institute of Lisbon, ISTAR-IUL, Av. das Forças Armadas, 1649-026 Lisboa, Portugal

[3]Department of Computer Science, University of Vigo, ESEI - Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain

[4]CINBIO - Biomedical Research Centre, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain

[5]SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur). SERGAS-UVIGO. Hospital Álvaro Cunqueiro Bloque técnico, Estrada de Clara Campoamor, 341, 36312 Vigo, Pontevedra, Spain

Abstract.

In recent years, most content-based spam filters have been implemented using Machine Learning (ML) approaches by means of token-based representations of textual contents. After introducing multiple performance enhancements, the impact has been virtually irrelevant. Recent studies have introduced synset-based content representations as a reliable way to improve classification, as well as different forms to take advantage of semantic information to address problems, such as dimensionality reduction.

These preliminary solutions present some limitations and enforce simplifications that must be gradually redefined in order to obtain significant improvements in spam content filtering. This study addresses the problem of feature reduction by introducing a new semantic-based proposal (SDRS) that avoids losing knowledge (lossless). Synset-features can be semantically grouped by taking advantage of taxonomic relations (mainly hypernyms) provided by Babelnet ontological dictionary (e.g. "*Viagra*" and "*Cialis*" can be summarized into the single features "*anti-impotence drug*", "*drug*" or "*chemical substance*" depending on the generalization of 1, 2 or 3 levels).

In order to decide how many levels should be used to generalize each synset of a dataset, our proposal takes advantage of Multi-Objective Evolutionary Algorithms (MOEA) and particularly, of the Non-dominated Sorting Genetic Algorithm (NSGA-II). We have compared the performance achieved by a Naïve Bayes classifier, using both token-based and synset-based dataset representations, with and without executing dimensional reductions. As a result, our lossless semantic reduction strategy was able to find optimal semantic-based feature grouping strategies for the input texts, leading to a better performance of Naïve Bayes classifiers.

Keywords.

Spam filtering; token-based representation; synset-based representation; semantic-based feature reduction; multi-objective evolutionary algorithms

1. Introduction and motivation.

The popularization of the Internet together with the exponential growth in the number of users taking advantage of the most recent applications and services (such as Instant Messaging Applications, Social Networks, e-mail service, etc.) have revolutionized how people communicate. In January 2019, the Internet had more than 4.3 million users, compared to a million users in 2005 (Clement, 2019). Most of these users are permanently connected through their smartphones using permanent broadband 3G/4G connections. In fact, since 2015 there have been more mobile devices connected to the Internet than to personal computers (Enge, 2019). Unfortunately, the scourge of spam content delivered through a wide variety of services (Geerthik, 2013) is limiting the experience of these users and hampers the usability of the Internet. Some services commonly used to distribute junk contents are web 2.0 applications (which originated the concept of spam 2.0) (Chakraborty, Pal, Pramanik, & Chowdary, 2016), search engines (Chandra & Suaib, 2014; Fdez-Glez et al., 2016) (webspam), e-mail (Perez-Diaz, Ruano-Ordás, Fdez-Riverola, & Méndez, 2012), Short Message Service (SMS) (Hidalgo, Bringas, Sánz, & García, 2006) and Instant Messaging (IM) applications (Silva, Alberto, Almeida, & Yamakami, 2017).

The fight against spam has been addressed from different perspectives including legal (for instance the European Directive on Privacy and Electronic Communications (European Parliament and the Council of the European Union, 2002)), networking standards (such as Request for Comments 6376[1] or 7208[2]), collaborative solutions (Gandhi, Mohanraj, Nandhini, Poovarasan, & Prakashraj, 2019) or URIBL[3] for webspam), or content-based schemes (Altınel & Ganiz, 2018; Guzella & Caminhas, 2009; Kastrati, Imran, & Yayilgan, 2019; Vyas, Prajapati, & Gadhwal, 2015). The combination of all of these technologies has severely limited spamming activities but has not definitively solved the problem.

We strongly believe that content-based filtering has not been widely exploited but could provide high quality results through knowledge engineering techniques. Most of the content-based approaches are based on taking advantage of different ML classification schemes (such as the popular Naïve Bayes) by using information about tokens found in the texts as input data. Upon analysing the performance of these approaches, we feel that this type of classifier has reached the utmost efficiency level that could be achieved by exploiting token-based approaches. This hypothesis is also supported by recent studies (analysed in Section 2) which successfully brought the use of concept-based features to spam filtering (Almeida, Silva, Santos, & Gómez Hidalgo, 2016; Bahgat, Rady, Gad, & Moawad, 2018; Méndez, Cotos-Yañez, & Ruano-Ordás, 2019) by using synsets extracted from ontological dictionaries like Wordnet (Miller, 1995). In a previous study (Méndez et al., 2019), the authors take advantage of taxonomic semantic relations between synsets

[1] Available at https://tools.ietf.org/html/rfc6376

[2] Available at https://tools.ietf.org/html/rfc7208

[3] See uribl.com

(specially hypernymy) to generalize semantic concepts found in text and to reduce the dimensionality from any number of features to 181 features (synsets having a maximum distance of 4 from 'entity' using hyponymy relations). Despite not taking advantage of semantic disambiguation schemes or individual features customized semantic generalization levels, these approaches achieved a better performance than simple token-based classification schemas.

This study brings a more innovative approach to improve semantic generalization. Specifically, we have taken advantage of disambiguation to adequately find the synset (semantic concept) that fits each word extracted from the text best. In most cases, synset-properties are optimally generalized in order to identify the best semantic concepts to address the classification, thus to discriminate between spam and ham contents. For instance, in some academic environments, it could probably be considered that all the contents related to chemical substances are spam regardless of their use, form of obtaining and/or nature. However, in a chemical engineering department, only those about specific medicines (such as Viagra or Cialis) should be probably considered spam. This implies that grouping all chemical substances into a single dataset property, including all drugs, will not be an appropriate decision.

In order to determine how many times a synset-property should be defined, we used an optimization approach. The optimization process was done by using the successful MOEA NSGA-II (Deb, Agrawal, Pratap, & Meyarivan, 2000). Despite of the resulting dimensionality reduction achieved by our proposal is higher than the one in a previous work (Méndez et al., 2019), the synset information is not over-generalised and none of the synsets are discarded, but only generalised. Therefore, we are starting a new generation of synset-based algorithms capable of maximizing the performance and avoiding knowledge loss (lossless feature reduction).

The main purpose of this study is to introduce a new way of reducing dimensionality when using synset-based representations (SDRS), based on the aforementioned ideas. The method exploits taxonomic information (i.e. hypernyms and hyponyms) stored in ontologies (in this case Babelnet), and takes advantage of NSGA-II genetic algorithm to find an optimized representation, which is able to obtain a problem representation that minimizes both the dimensionality and the amount of false positive (FP) and false negative (FN) errors.

In detail, this study aims specifically to the following objectives: (*i*) reduce the dimensionality of synset-based datasets minimising the information loss (loossless), (*ii*) check the performance of Naïve Bayes classifiers, both on the original datasets and on their combination with our SDRS lossless dimensional reduction technique, (*iii*) compare the performance of Naïve Bayes classifiers on both token-based and synset-based representations.

The remainder of the work is structured as follows: Section 2 presents the state of the art in the use of semantic information extracted from ontological dictionaries to reduce

dimensionality in classification problems; Section 3 presents our algorithm to address lossless semantic-aided feature reduction. The data used as input corpus, the experimental protocol and the results achieved during the empirical benchmark are included in Section 4. Finally, Section 5 summarizes the main outcomes and future research lines.

2. Related work

The popularity of ML approaches has experienced an amazing growth due to their ability to solve a wide variety of problems by using past experiences and related information as input data. Particularly, in the context of binary classification, we can take advantage of a lot of well-tested approaches in order to successfully deal with problems such as spam filtering (Hall et al., 2009; Pedregosa et al., 2011). According to the vast literature in this area, the performance that these classifiers can achieve is not in doubt (Vyas et al., 2015). However, if the classifier does not receive relevant and clean information that can be generalized, the classification performance is poor. One of the key aspects when applying ML schemes is the preparation of input data (Zhang, Zhang, & Yang, 2003); improving the performance achieved by ML techniques in the context of filtering spam is directly dependent on this aspect.

The text mining research field (Tandel, Jamadar, & Dudugu, 2019) emerged as the prevalent form of exploiting token information to solve problems such as text classification, information retrieval, etc. Furthermore, the first well-known ML proposal for spam filtering[4] introduced by Paul Graham, and many others introduced later (Méndez, Glez-Peña, Fdez-Riverola, Díaz, & Corchado, 2009; Pérez-Díaz, Ruano-Ordás, Fdez-Riverola, & Méndez, 2016), take advantage of token-based information. In this manner, topic models (Grün & Hornik, 2011) emerged with the aim of studying terms that are usually found together in texts and therefore allow for topic detection without using semantic information about the connections between terms. This type of model was able to statistically find "topics" and allowed documents to be represented using this information.

However, during the last few years a new method of representing textual content has been introduced: the concept. Concepts have been recently modelled by using synsets from ontology dictionaries; the growing interest in their use is based on the capability of handling different human language representations of the same information (Almeida et al., 2016). Particularly the following text pieces "cheap vehicle for sale", "I'm selling my car" or "automobile offer" all stand for the same information but are composed of different tokens.

One of the most relevant problems when representing text contents using tokens or concept-based representations is the problem of dimensionality (Méndez et al., 2019). Often the dimensionality on this kind of problem is overwhelming, and classifiers cannot properly operate on them. Moreover, some features in the dataset could be irrelevant,

[4] See http://www.paulgraham.com/spam.html

redundant, dependent on others, or inconsistent and their identification and dropping would be necessary. To deal with this issue, different types of feature selection (reduction) schemes were introduced. It has been suggested (Chandrashekar & Sahin, 2014; Deng, Li, Weng, & Zhang, 2019) that classical feature reduction schemes can be classified into three groups: (*i*) filter (*ii*) wrapper and (*iii*) embedded.

Filter feature selection methods (*i*) can assess the relevance of each feature included in input data to solve the problem. These results can be quickly computed and used to rank features. The top *n* best features, or those better evaluated than a threshold value, are selected. The main advantage of filters is their simplicity, while their main limitation is the difficulty of avoiding duplicate information (or select highly independent variables). Wrapper methods (*ii*) combine a strategy of grouping input features. The quality of each group of features is assessed by running an ML classifier using group features to solve the target problem. These approaches are slower than filters but can successfully address the elimination of dependent variables. Their main limitations are the dependency of using an ML classifier that could result in overfitting issues. Finally, embedded feature selection schemes (*iii*) bring together some feature reduction methods that are intimately related to specific ML algorithms (Lal, Chapelle, Western, & Elisseeff, 2006).

Although classical feature selection methods (filters, wrappers or embedded) could be applied over concept features, the most recent advances in the use of semantic information from ontological dictionaries allows for the definition of alternative feature selection methods. In a study by Almeida et al. (Almeida et al., 2016) synonyms of words found in SMS messages are added as artificial features to cope with their small size (and contents) and improve classification performance. Additionally, in a study by Bahgat *et al*. (Bahgat et al., 2018) the authors successfully take advantage of synonymy relations to bring similar terms into the same feature. Although the achieved reduction was very limited, this work introduced the concept feature to address e-mail classification. A more recent study focused on semantic feature selection (Méndez et al., 2019) introduced the use of hypernymy/hyponymy relations to achieve better dimensionality reduction (from any number of features down to 181). Using hypernymy relations allows for semantic generalizations until one of the synsets of the first 4 levels of Wordnet (Miller, 1995) ontology is reached. This idea could bring together words such as "viagra", "cialis", “tadalafil” or "xanax" under concept features such as “drug” or “chemical_substance”. One of the most relevant benefits of this proposal is the reduced loss of information because terms are not removed (they are only generalized).

Despite the relevance of the contribution of a previous study (Méndez et al., 2019), using only 4 levels of Wordnet does not seem to be adequate for all situations, because the above mentioned words would be represented as “agent” (the synset in the 4th level of Wordnet) with other words such as “coolant” (which could be connected with engines or computers), “antifungal” (which is connected with agricultural treatments), “diluent” (such as those used for paints), etc. We strongly believe that the concepts found in a text content should be generalized as much as required to bring together texts of interest to

the target user while keeping them apart from those that are not. In fact, a user may not be interested in texts about virility drugs but interested in other kinds of drugs (relaxing, antibiotics, etc.). This means that we should use "anti_impotence_drug" and "drug" as features to represent a target dataset to handle the information required for text classification. Additionally, the correct application of disambiguating schemes (as applied in a previous work (Almeida et al., 2016)) is required in this article.

We addressed the research challenges identified in previous semantic-based feature selection approaches and reused ideas from the classical feature selection methods (wrappers) and the latest advances in concept-based text representation. These ideas led us to the definition of Semantic Dimension Reducing Scheme (SDRS) introduced in the next section.

3. Introducing SDRS

For the application of our feature selection proposal, text extracts from several Internet communication mechanisms (e-mail, SMS, Instant Messaging, tweets, etc.) are represented using concepts (synsets from an ontological dictionary). Following the recommendations included in a study by Almeida et al. (Almeida et al., 2016), we used Babelnet as ontological dictionary. Additionally, each token extracted from text was disambiguated using Babelfy[5] to find the synset that best fit it. Using this scheme, a corpus was transformed into a dataset (D) where each content is represented as a row. Each content is represented with its id (# column), the target class (target) and a large set of binaries (yes/no) concept features (synsets, $S=\{s_1,s_2,s_3,..,s_n\}$) that indicate whether or not the concept is included in the content. Figure 1 contains an example of output data generated through the preprocessing of the dataset.

Figure 1: *Example of input dataset (D) for feature selection*

$S=\{s_1,s_2,s_3,..,s_n\}$

#	bn:00071570n (viagra)	bn:00019048n (cialis)	bn:00015620n (Madrid)	bn:00007309n (car)	bn:00045229n (hungry)	target
1	1	0	0	0	0	spam
2	0	1	0	0	0	spam
3	0	0	1	1	0	ham
4	0	0	1	0	1	ham
5	0	0	0	0	1	ham

While SDRS is inspired in wrapper feature selection, it successfully takes advantage of semantic information by using ontologies. Following a previous work proposal (Méndez et al., 2019), hypernym relations of the synsets are primarily considered to generalize

[5] See http://babelfy.org

concepts. However, each synset is generalized as necessary to adequately identify the subject of interest (or not) to a target user (or user group). To identify the best configuration, we formulated a multi-objective optimization problem and used NSGA-II (Non-dominated Sorting Genetic Algorithm) (Deb, Pratap, Agarwal, & Meyarivan, 2002) to solve it. NSGA-II is a MOEA (Multi-Objective Evolutionary Algorithm) that has been successfully used to solve many different multi-objective optimization problems.

A chromosome $C=\{c_1,c_2,c_3,..,c_n\}$ in SDRS represents how much each synset-feature $s_i \in S$ (the set of concept-features used to represent texts) will be generalized. Therefore c_i $(i=1..n)$ is an integer value in the interval $[0..\gamma]$ where γ represents the maximum generalization level and c_i represents the number of steps the synset s_i is generalized.

To compute the transformation of a dataset D with the chromosome $C, T(D,C)$, we should take into consideration that if s_i is generalized *m* times into another synset s_i', all features from S that are hyponyms of s_i' ($\left\{s_j \in S \middle| s_j \in hyponyms\left(s_i'\right)\right\}$) are deleted and represented by the new feature s_i' (thus achieving a dimensionality reduction). Hence, in a chromosome $C=\{2,1,0,0,0\}$ that is evaluated for the example included in Figure 1, the feature 'bn:00071570n (Viagra)' is generalised two times into 'bn:00004605n (anti-impotence drug)' (first generalization) and 'bn:00028872n (drug)' (second time). As long as the feature 'bn:00019048n (Cialis)' is a hyponym of 'bn:00028872n (drug)' they are both merged into the last feature. As long as s_1 and s_2 share the same direct hypernym the same effect would be observed when $s_1=k$ and s_2 has a value within the range $[0..k]$. Given the chromosome $C=\{2,1,0,0,0\}$, the example included in Figure 1 would be transformed into the one represented in Figure 2.

Figure 2: *Transformation of input dataset* D *for the chromosome* $C=\{2,1,0,0,0\}, T(D,C)$

#	bn:00028872n (drug)	bn:00015620n (Madrid)	bn:00007309n (car)	bn:00045229n (hungry)	target
1	1	0	0	0	spam
2	1	0	0	0	spam
3	0	1	1	0	ham
4	0	1	0	1	ham
5	0	0	0	1	ham

In addition to the problem of representation, the definition of the optimization objective functions (i.e. fitness functions) is mandatory to run any MOEA. Following wrapper approaches, we considered the use of performance classification results achieved when a classifier is run (10-fold cross validation scheme (Kohavi, 1995)) over the reduced dataset. As done in previous spam filter optimizations (Basto-Fernandes et al., 2016;

Ruano-Ordás, Basto-Fernandes, Yevseyeva, & Méndez, 2017; Ruano-Ordás, Fdez-Riverola, & Méndez, 2018; Yevseyeva, Basto-Fernandes, Ruano-Ordás, & Méndez, 2013), the minimization of false positives (FP, ham texts classified as spam) and false negatives (FN, spam texts classified as ham) could be successfully introduced as objectives for this problem. Additionally, as the main goal of the process is to reduce the dimensionality of the initial dataset, we introduced these minimizations as a fitness function. Equation 1 shows the fitness functions used to solve the current problem

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

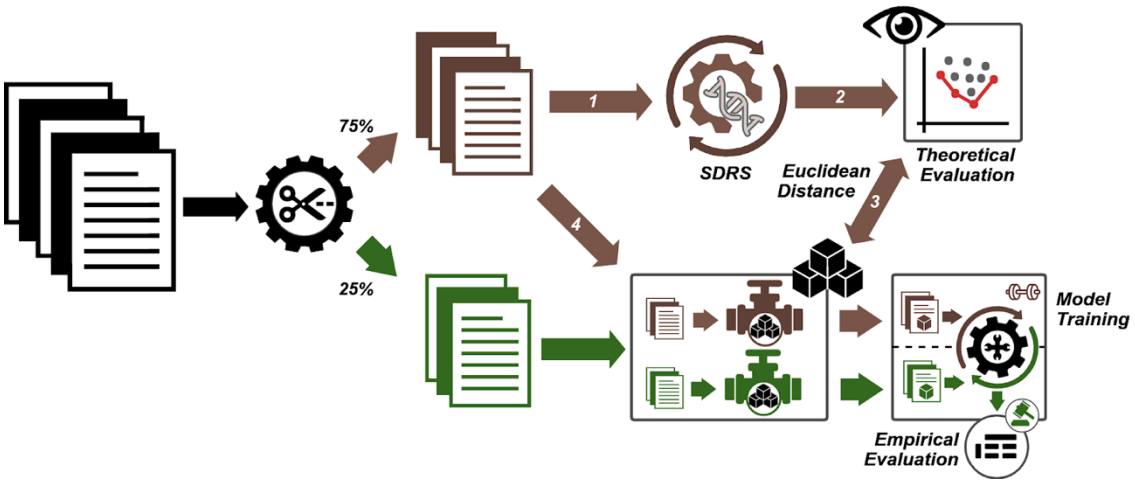
where $10\,\text{xval\_eval.FPr}\left(c,D\right)$ and $10\,\text{xval\_eval.FNr}\left(c,D\right)$ represents the false positive and false negative ratios achieved when running a 10-fold cross validation test using the classifier c and the dataset D, c is the classifier selected for evaluation, $T\left(D,C\right)$ represents the transformation of the dataset D with the chromosome C and $\text{num\_cols}\left(D\right)$ is the number of columns in the dataset D.

Using this configuration, the NSGA-II genetic algorithm is executed to discover the highest reduction that can be achieved minimizing the number of FP and FN errors. The next section shows the configuration and results of the experiments carried out to evaluate SDRS

4. Experimental evaluation

In order to test our proposal, we carefully designed an experimental protocol. One of the key features of the protocol is the inclusion of two different performance evaluation points and the possibility of detecting overfitting phenomena. The proposed experimental protocol is shown in Figure 3.

Figure 3. *Experimental protocol*



The evaluation protocol comprises the use of a text corpus as input data. The corpus is divided into two stratified splits comprising 75% and 25% of input texts. The biggest split is used to execute SDRS and determine the best form of representing the dataset using the MOEA problem formulation (SDRS feature selection method). The outputs of the MOEA (a set of non-dominated solutions representing the Pareto front) are evaluated in the first stage of the experimental protocol (see Theoretical Evaluation in Figure 3). Finally, we represent SDRS solutions as a point in the form (FPr, FNr) and use Euclidean distance to select the closest solutions to the origin of coordinates (0, 0). The four best solutions are used to transform the dataset splits (75%, 25%) and run a simple train/test experiment to perform a benchmark with other dimensionality reduction schemes.

Other experimental protocol details, which include selecting the target corpus (subsection 4.1), configuring the pre-processing steps (subsection 4.2) and selecting the SRDS parameters (subsection 4.3), are shown below. Finally, the empirical results are shown and analysed in subsection 4.4.

4.1. Selecting a dataset from available corpora

A wide variety of corpora is available on the Internet for benchmarking spam filtering techniques applied to different Internet services. Table 1 compiles a collection of well-known corpora classified according to type of content, language, ham/spam ratio and size.

Table 1: *Available corpora for spam filtering*

dataset	type of content	language	ham/spam ratio	size
Spam Corpus	email	English	34% spam	4,027
TREC 2007 Public Corpus	email	English	66% spam	75,419
SpamAssassin	email	English	31% spam	6,047
Enron email	email	English	0% spam	619,446
Bruce Guenter spam collection	email	English	100% spam	>3,000,000
Ling spam	email	English	16% spam	2,893
SMS Spam Collection v.1	SMS	English	13% spam	5,574
British English SMS corpora	SMS	English	48% spam	875

Webspam-uk 2007	Web (html) pages	English	unknown	105,896,555
Websmap-uk 2011	Web (html) pages	English	53% spam	3,766
DC 2010 / EU 2010	Web (html) pages	English, French and German	unknown	23M
Webb spam 2011	Web (html) pages	unknown	unknown	330.000
Clueweb 09	Web (html) pages	10 languages	unknown	1,040M
Clueweb 12	Web (html) pages	English	unknown	870M
Common Crawl Data	Web (html) pages	multilingual	100% spam	9 Billion in 2014 and increasing
YouTube Comments Dataset	Youtube comments	multilingual	7% spam	6M
YouTube Spam Collection Dataset	Youtube comments	English	49% spam	1,956
HSpam14.s2	Twitter messages (tweets)	unknown	unknown	14M

There is a wide list of available corpora to execute spam-filtering experiments over different Internet services. Due to the high execution load of a wrapper method (which is usually slow), we believe that using a small corpus would be particularly suitable for our experimental protocol. Therefore, we selected the YouTube Spam Collection available at the UCI ML repository.

The selection of a rather small dataset will be adequate for the inner properties of SDRS, which involves running a 10-fold cross validation test for the instances selected for optimization (75% of the input dataset as detailed in the experimental protocol). Additionally, an effective use of stochastic methods, such as evolutionary algorithms, requires significant computational resources and computation time, typically implying the execution of dozens of thousands of objective function evaluations and dozens of independent runs of the algorithms, to deal with their stochastic nature.

4.2. Pre-processing configuration

Given the raw nature of the Youtube Spam Collection Dataset, we pre-processed the dataset with the Big Data Pipelining for Java (BDP4J[6]) and Natural Language Pre-processing Architecture (NLPA) projects[7]. The semantic information was extracted from Babelnet ontological dictionary and the disambiguation facilities provided by Babelfy[8].

The pre-processing of the corpus instances includes the extraction of the comment body text using YouTube API[9]. The HTML tags are then removed while entities are replaced by their corresponding plain text (CSS tags, JavaScript and URIs are also removed). Moreover, emoticons and emojis identified in (Wikipedia contributors, 2019) and (Unicode Inc, 2018) are removed from the text and registered in an instance property. The language of the text is then determined using a Java library[10] to remove stop words, interjections and onomatopoeias. Additionally, by taking advantage of the language information, contractions, abbreviations and slang expressions are replaced by their meaning.

In order to identify the synsets that best match each term in the resulting test, we performed a disambiguation. In cases where different BabelNet synsets (meanings/concepts) can be associated with a word, Babelfy was used to select which of the meanings was the most appropriate given the specific context of the word. For example, the occurrence of the word "bank" can lead to different synsets of the BabelNet network, “sloping land”, “depository financial institution”, “a long ridge or pile”, and more. Each of these meanings has a different synset id on BabelNet, so it is very important to identify which meaning/synset match the (text) context.

In the data pre-processing step, the text is sent to Babelfy, whose service eliminates the stop words, performs disambiguation tasks and returns the list of synsets corresponding to the text. This list needs to be further processed in order to deal with situations where groups of words must be considered together. For example, if we try to disambiguate "neural network", Babelfy returns three possible answers (synsets): one synset that corresponds to the word "neural"; another synset that corresponds to the word "network"; and a synset that corresponds to "neural network". In our case, we always look for the synset that groups the largest set of words.

For experimental purposes, the dataset was also represented using token features. This process was made using NLPA and maintaining the same pre-processing steps used for the case of the synsets, but using token-based features to represent the instances.

---

[6] Available at https://github.com/sing-group/bdp4j
[7] Available at https://github.com/sing-group/nlpa
[8] Available at babelfy.org
[9] See https://www.youtube.com/intl/es/yt/dev/api-resources/
[10] Available at https://github.com/optimaize/language-detector

Finally, Weka (Witten, Frank, Hall, & Pal, 2016) was used to execute the ML part of the experimentation.

4.3. SDRS and NSGA-II configuration

In order to run SDRS, the inner classifier and γ (maximum number of generalization steps) parameter should be defined. Due to its popularity in spam filtering, we selected Naïve Bayes (MultinomialNaïveBayes implemented in Weka) as the inner classifier for SDRS. The lower requirements on computational resources, together with its prevalent use on the domain, suggest that it would provide interesting results (Ezpeleta, Garitano, Arenaza-Nuno, Hidalgo, & Zurutuza, 2017; Ezpeleta, Garitano, Zurutuza, & Hidalgo, 2017). Additionally, due to the low performance of the Naïve Bayes classifier in the presence of dependent variables, its use would also select more independent features.

In order to find an adequate value for γ, we ran an empirical evaluation of different configuration values (1-5) using the complete Youtube Comments Dataset (see Table 1). We specifically ran SDRS and discovered how many configurations from Pareto are better than keeping the original dimensionality. To evaluate the performance of each configuration we used a geometric distance between the 3-D points (FPr, FNr) and the coordinate origin (0, 0). Table 2 summarizes the results achieved for different configurations.

Table 2: *Analysis of different γ configurations using the geometric distance criterion*

γ	Minimum distance
0 (without using SDRD)	0,394946621
1	0,322903916
2	0,319093112
3	0,312350806
4	0,319875171

The best configuration to obtain the minimum value of distance for γ is 3. Keeping in mind these results, we selected this value to use throughout the entire experimentation process.

Additionally, we used the NSGA-II implementation provided in JMetal framework. JMetal framework was configured with 25 runs and a maximum of 25.000 function evaluations. NSGA-II default settings were used, with population size 100, integer SBXCrossover and PolynomialMutation operators with 1.0 crossover probability and 1/NumberOfVariables mutation probability.

Once the parameter selection was completed, we executed the experimental protocol. The experimental results achieved during the execution are detailed and analyzed in subsection 4.4.

Table 3: *Tokens and Synset classification results without using SDRS*

Classification using tokens				Classification using synsets			
%OK	%FP	%FN	dim	%OK	%FP	%FN	dim
0.884	0.048	0.068	2279	0.828	0.02	0.152	1684

The results indicate that the number of features (the dimension) is smaller when using synsets than when using tokens. This situation happens because of (i) different synonymous words that are represented in one single feature, (ii) misspelled words or words subjected to obfuscation tricks (commonly used by spammers to avoid spam filters) that are discarded, and (iii) the existence of words that are not found in the semantic network (Babelnet).

We can also find important classification differences by analysing percentages of FP and FN errors included in Table 3. These differences occur due to the procedure designed to convert words into Synsets and the loss of information (words that are not represented as synsets) described before. However, the results indicate that the use of synsets performs better on legitimate texts (less FP errors) where misspelling errors and obfuscation tricks are not used. This suggests that removing terms that do not exist in a dictionary improves the identification of legitimate contents (ham).

Additionally, we used the traditional and popular Information Gain (IG) feature selection filter method (Méndez, Cid, Glez-Peña, Rocha, & Fdez-Riverola, 2008; Méndez et al., 2019) to reduce the dimensionality of datasets represented as synsets or tokens. To compare performance impact when the dimensionality is reduced with IG, we ran a Multinomial Naïve Bayes classifier with different dimensionalities including all (2279), 2000, 1500 and 1000 features. The results are shown in Table 4.

Table 4: *Tokens classification result applying Information gain.*

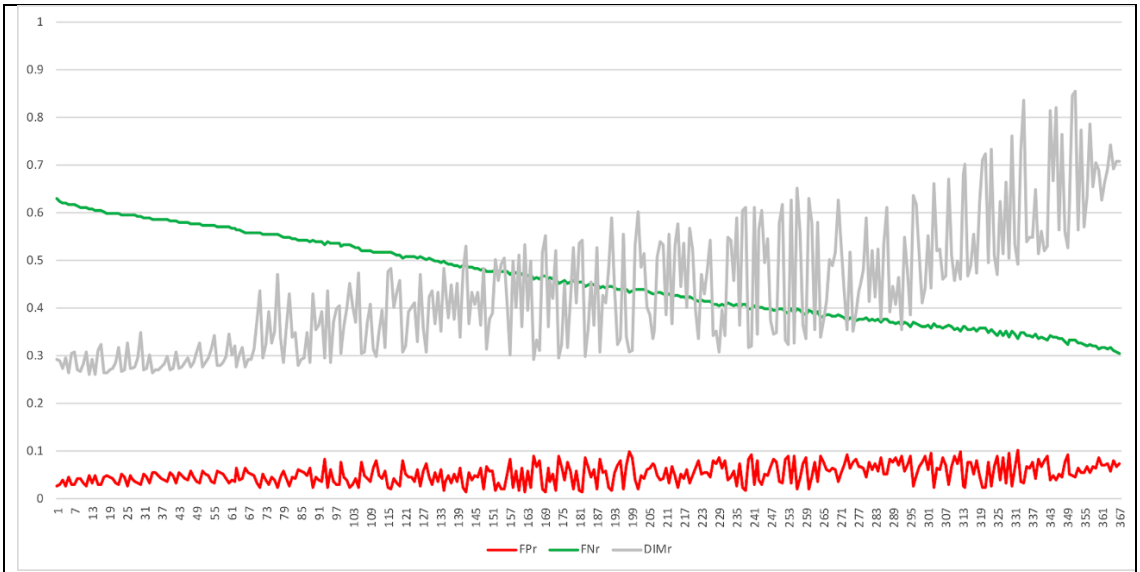
Dimension	**Token representation**			**Synset representation**		
	%OK	**%FP**	**%FN**	**%OK**	**%FP**	**%FN**
2279 (=number of tokens)	0.884	0.048	0.068			
2000	0.9	0.048	0.052			
1684 (=number of synsets)	0.888	0.056	0.056	0.828	0.02	0.152
1500	0.884	0.064	0.052	0.832	0.028	0.14

1000	0.884	0.072	0.044	0.888	0.012	0.1
500	0.896	0.056	0.048	0.868	0.02	0.112

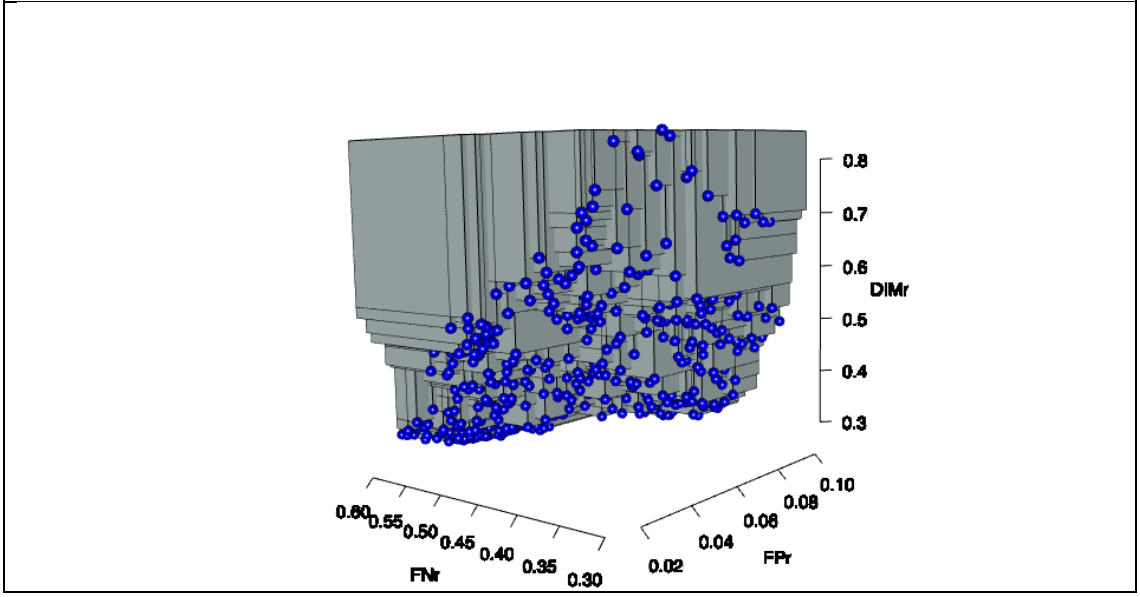
For all configurations analyzed, synset representations achieve a larger percentages of FN errors. However, the use of synset-based representations makes it possible to significantly reduce the number of FP errors.

In order to test the utility of our feature reduction protocol (SDRS), the results obtained after optimization were analysed in detail. The results comprise the set of non-dominated solutions achieved in all the 25 executions. These results were represented by plotting a multiple line chart and a 3D Paretto front, as shown in Figures 4a and 4b respectively.

Figure 4: *Optimization performance analysis*



a) Multiple line chart representing solutions



b) 3D Pareto front

As shown in Figure 4a, the number of FP errors is close to 0 for all evaluated configurations. However, the FNr achieved values within the range 0.3-0.7. Additionally, a wide variety of dimensions is achieved within the range of 25% and 85% of the original size of the dataset. Some of the solutions in Figure 4a are clearly relevant, such as solution 367 (last one), which represents a good compromise with good evaluation in all objectives.

The results from Figure 4a indicate that an improvement of 32% on FNR (from 0.62 to 0.30) has only an 8% degradation effect on FPr (from approximately 0.2 to 0.8). Additionally, we can see that dimensionality reduction can be achieved at the expense of FNR degradation, while FPr seems relatively indifferent to dimensionality variations.

To complement this information, we also measured the time required to execute SDRS feature selection algorithm. The results of the measurements are included in Table 5.

Table 5: *Time required to execute SDRS feature selection algorithm with the selected dataset*

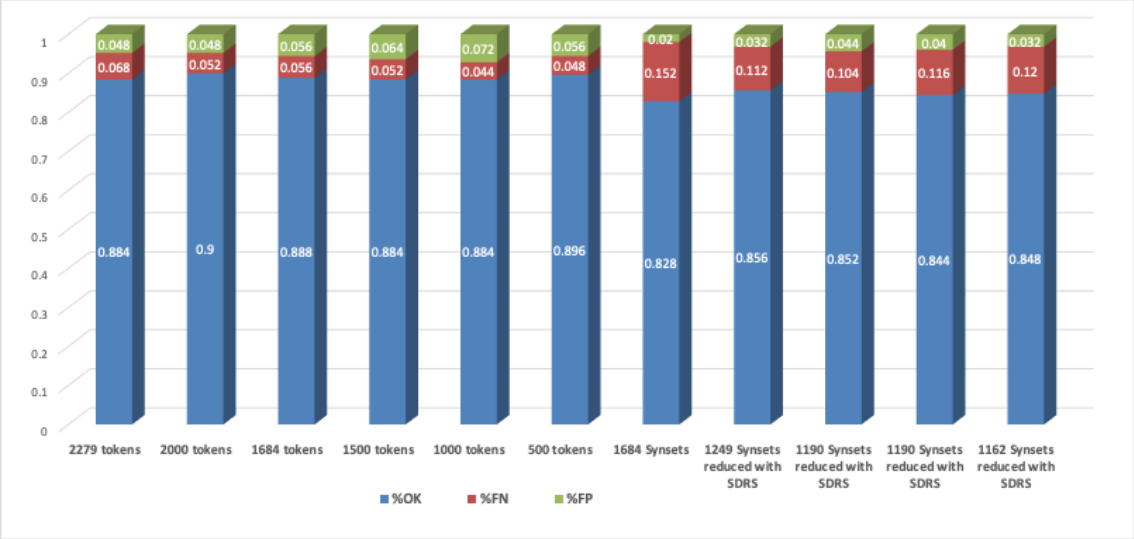
Computer specifications	Execution time
2 x Intel Xeon X5675 (Q1/2011) 3.07GHz with (6 cores/12 threads). 128 GB RAM	13 days
4 x Intel Xeon E7-8890 (Q2/2016) v3 2.5 GHz (18 cores/48 threads) 1 TB RAM	10 days
2 x Intel Xeon E5-2640 (Q1/2012) v3 2.6 GHz (8 cores/12 threads) 128 GB RAM	13 days

These results reveal the high computational requirements of the SDRS process even when the input dataset presents a small dimensionality. Fortunately, this process can be executed only once to identify the best form of grouping features for the target organization.

The best feature reduction configurations identified by SDRS results were evaluated using unseen data (25% of the whole dataset). We compared the performance achieved when using 4 best configurations of SDRS and 4 IG-based dimensionality reducing configurations applied on tokens. A Multinomial Naïve Bayes model was built, using tokens/synsets with the same input data (75%) we had previously selected for optimization purposes. Thus, the remaining 25% of the Youtube comments were used to execute the test to produce results that are comparable with our SDRS system. The configurations selected from the multiple configurations generated by SDRS are those achieving (FNR, FPR) evaluations closest to the origin of coordinates (*FPr*=0.074/*FNr*=0.303; *FPr*=0.066/*FNr*=0.306; *FPr*=0.080/*FNr*=0.309 and

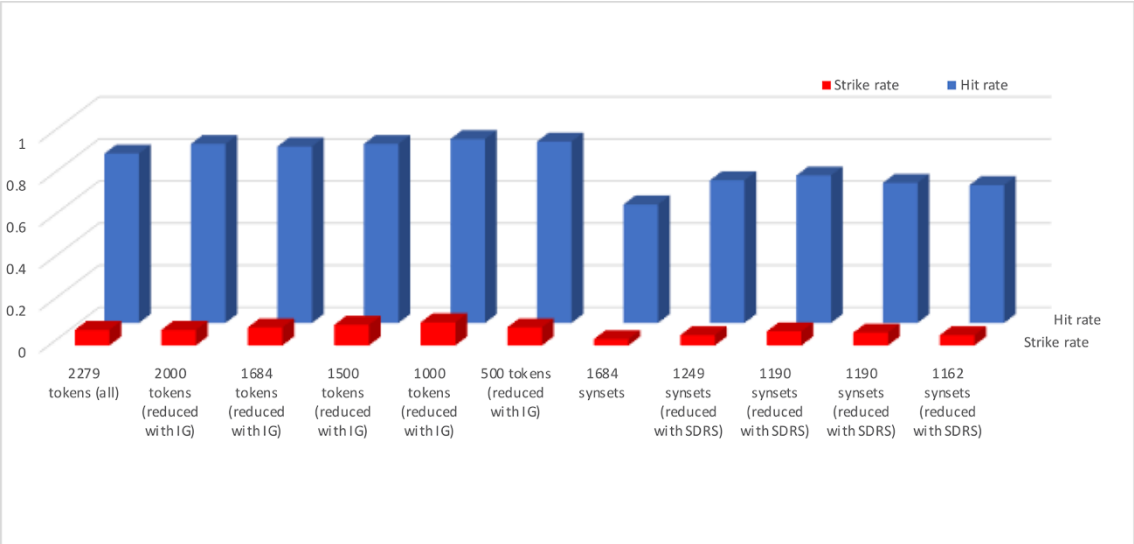
FPr=0.057/*FNr*=0.3015). Firstly, we evaluated the results in a percentage form. Figure 5 shows a percentage evaluation between a token based scheme and SDRS.

Figure 5: *Percentage evaluation of different feature-reduction schemes*



When using a synset representation, the number of FN errors is higher, but the number of FP errors achieves an important reduction. This observation can be appreciated more clearly in batting average scores (see Figure 6). A batting average measures the proportion of successfully detected spam contents (hit rate) and the proportion of FP errors with regard to the number of ham instances (strike rate). This allows us to deduce that the best classifiers are those achieving highest hit rate and lowest strike rate scores.

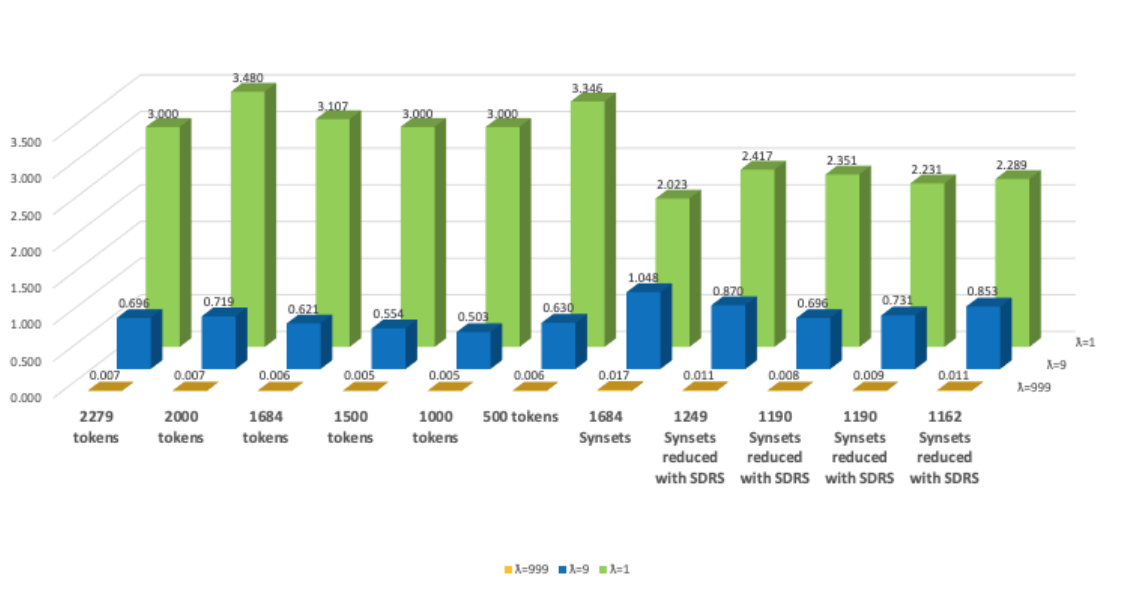
Figure 6: *Batting average comparison of analysed configurations*



Although token-based configurations detect more spam texts (better hit rate), the amount of FP errors they achieve (strike rate) indicates serious trouble, suggesting its use should be discarded for some applications. In order to determine the kind of applications that should not be used with token-based representations, we ran a cost-sensitive analysis of performance using TCR (Total Cost Ratio). TCR scores consider the asymmetric cost of

FP and FN errors in most environments by using a λ parameter that indicates how many times an FP error is more serious than an FN error. In our comparison we used popular values for λ (1, 9 and 999) which model common account situations (services used for exchanging jokes; services used sometimes to receive information about e-commerce, e-banking operations; services used to exchange professional and commercial information, or contact customers). TCR evaluations are shown in Figure 7.

Figure 7: *TCR benchmarking results*



Given these results, we have deduced that synset-based schemes are a bit poor for situations where the cost of FP and FN errors is the same (λ=1). Unfortunately, the popularization of some Internet services such as e-banking, e-commerce or Internet job searching, makes improper considering error cost symmetric. Furthermore, when considering asymmetric costs of errors, the use of synset-based representations, and SDRS feature selection methods in particular, is more appropriate. A quick analysis of the results shown in this work suggests the adoption of synset-based representation methods and semantic-based feature reduction methods. The next section these findings in greater detail and outlines the main strategies for future research.

5. Conclusions and future work

This study has provided an experimental comparison of the performance that can be achieved by using token and semantic based representations of texts for spam filtering purposes. Moreover, we have introduced and shown the performance of a feature reduction scheme based on an MOEA optimization scheme (SDRS). Results show that semantic-based approaches perform much better in contexts where the cost of FP and FN errors is assymetric (almost every situation in which we currently use Internet).

From a practical point of view, the decreases in the number of redundant words (grouping words taxonomically related into an unique feature) implies a time and computational

resources requirements reduction for training. Moreover, from a theoretical perspective, by combining highly dependant features, in-between features gets lower, thus increasing the performance of some techniques, e.g. Naïve Bayes. Additionally, grouping words with similar meaning allows the identification of the subject topics related with the texts. As an example, texts about "anti_impotence_drugs" could be connected in the same group (as one of the topics connected with spam concept). The identification of problematic topics for a company/user is essential to handle the isolated nature of spam (including textual content matching different topics that are irrelevant for the company/user).

Experimental results allow us to conclude that the use of a synset representation is better for the detection of ham contents rather than spam contents (the number of FP errors is reduced while we achieve a slight increase in FN errors). Despite this, results have clearly shown that, when accounting for the asymmetric costs of FP and FN errors, synset representations perform better than those that are token based. Moreover, this fact shows that ham contents (usually with no obfuscation and no spelling errors) allow the successful translation of a greater number of words into synsets and a better classification on this kind of instance. In contrast, most words included in spam contents (with many misspelled words, obfuscated tokens or URLs) cannot be successfully represented into synset-based representations, resulting in the compilation of less information for this kind of text (and less performance to classify them). However, this limitation (the difficulty of translating synset words into spam content) can be used as a feature to improve the entire process (future work). A new feature containing the number of words that could not be translated into synsets could be added to the spam filtering process, which would probably result in a new increase in performance.

SDRS can be also categorized as a lossless feature selection method. This means that the knowledge present in words is not discarded (just grouped into features that bring together more or less generic concepts). The identification of the lossless feature selection method could also easily derive into a non lossless feature selection, where the loss of information can be parametrized (future work). The study with loss of information implies that the least relevant features (most of them either noise or entirely irrelevant) will be identified and removed from the classification process. Comparing the results of lossless and non lossless feature selection methods will raise new research findings to address both the spam classification and intentionality analysis problems.

The introduction of SDRS has achieved two objectives: (*i*) to reduce the amount of features required to represent textual contents; and (*ii*) to detect the words that should not be generalized and provide useful information about the purpose of the content. The main limitation of SDRS is its long execution time caused by the evolutionary metaheuristics involved in the optimization process. However, we are convinced that the use of SDRS will facilitate the identification of the relevant features of the semantic lossless feature selection method required for spam filtering. This knowledge will be enough to design an iterative feature selection method able to achieve the required (or even better)

performance levels. The discovery of iterative semantic-based feature selection methods should also be covered in future works.

Acknowledgements

This work was partially supported by the project Integración de Conocimiento Semántico para el Filtrado de Spam basado en Contenido, subprojects TIN2017-84658-C2-1-R and TIN2017-84658-C2-2-R, from the Spanish Ministry of Economy, Industry and Competitiveness (SMEIC), State Research Agency (SRA) and the European Regional Development Fund (ERDF).

Intelligent Systems for Industrial Systems group is supported by the Department of Education, Language policy and Culture of the Basque Government.

SING group thanks CITI (C*entro de Investigación, Transferencia e Innovación*) from University of Vigo for hosting its IT infrastructure.

References

Almeida, T. A., Silva, T. P., Santos, I., & Gómez Hidalgo, J. M. (2016). Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering. *Knowledge-Based Systems*, *108*, 25–32. https://doi.org/10.1016/j.knosys.2016.05.001

Altınel, B., & Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. *Information Processing & Management*, *54*(6), 1129–1153. https://doi.org/10.1016/j.ipm.2018.08.001

Bahgat, E. M., Rady, S., Gad, W., & Moawad, I. F. (2018). Efficient email classification approach based on semantic methods. *Ain Shams Engineering Journal*, *9*(4), 3259–3269. https://doi.org/10.1016/j.asej.2018.06.001

Basto-Fernandes, V., Yevseyeva, I., Méndez, J. R., Zhao, J., Fdez-Riverola, F., & T.M. Emmerich, M. (2016). A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification. *Applied Soft Computing Journal*, *48*, 111–123. https://doi.org/10.1016/j.asoc.2016.06.043

Chakraborty, M., Pal, S., Pramanik, R., & Chowdary, C. R. (2016). Recent developments in social spam detection and combating techniques: A survey. *Information Processing & Management*, *52*(6), 1053–1073.

Chandra, A., & Suaib, M. (2014). A survey on web spam and spam 2.0. *International Journal of Advanced Computer Research*, *4*(2), 634.

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

Clement, J. (2019). Number of internet users worldwide from 2005 to 2018. Retrieved 2 December 2019, from Statista website: https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/

Deb, K., Agrawal, S., Pratap, A., & Meyarivan, T. (2000). *A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimization: NSGA-II*.

https://doi.org/10.1007/3-540-45356-3_83
Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017
Deng, X., Li, Y., Weng, J., & Zhang, J. (2019). Feature selection for text classification: A review. *Multimedia Tools and Applications*, *78*(3), 3797–3816. https://doi.org/10.1007/s11042-018-6083-5
Enge, E. (2019). Mobile vs Desktop Usage in 2019. Retrieved 19 December 2019, from Perficient, Inc. website: https://www.perficientdigital.com/insights/our-research/mobile-vs-desktop-usage-study
European Parliament and the Council of the European Union. (2002). Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) (L 201. Retrieved 2 December 2019, from Official Journal of the European Communities website: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:NOT
Ezpeleta, E., Garitano, I., Arenaza-Nuno, I., Hidalgo, J. M. G., & Zurutuza, U. (2017). Novel comment spam filtering method on youtube: Sentiment analysis and personality recognition. *International Conference on Web Engineering*, 228–240.
Ezpeleta, E., Garitano, I., Zurutuza, U., & Hidalgo, J. M. G. (2017). Short Messages Spam Filtering Combining Personality Recognition and Sentiment Analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *25*(Suppl. 2), 175–189.
Fdez-Glez, J., Ruano-Ordás, D., Laza, R., Méndez, J. R., Pavón, R., & Fdez-Riverola, F. (2016). WSF2: A Novel Framework for Filtering Web Spam. *Scientific Programming*, *2016*, 6091385:1--6091385:18. https://doi.org/10.1155/2016/6091385
Gandhi, N., Mohanraj, K., Nandhini, K., Poovarasan, K., & Prakashraj, K. (2019). *Cosdes: A Collaborative Spam Detection System With A Novel E-Mail Abstraction Scheme*.
Geerthik, S. (2013). Survey on Internet Spam: Classification and Analysis. *International Journal of Computer Technology and Applications*, *4*(3), 384.
Grün, B., & Hornik, K. (2011). Topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13
Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, *36*(7), 10206–10222.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10. https://doi.org/10.1145/1656274.1656278
Hidalgo, J. M. G., Bringas, G. C., Sánz, E. P., & García, F. C. (2006). Content based SMS spam filtering. *Proceedings of the 2006 ACM Symposium on Document Engineering, DocEng 2006*, *2006*, 107–114. https://doi.org/10.1145/1166160.1166191
Kastrati, Z., Imran, A. S., & Yayilgan, S. Y. (2019). The impact of deep learning on document classification using semantically rich representations. *Information Processing & Management*, *56*(5), 1618–1632.

https://doi.org/10.1016/j.ipm.2019.05.003
Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference of Artificial Intelligence*, 1137–1143. Retrieved from http://dl.acm.org/citation.cfm?id=1643031.1643047
Lal, T. N., Chapelle, O., Western, J., & Elisseeff, A. (2006). Embedded methods. In *Studies in Fuzziness and Soft Computing* (Vol. 207, pp. 137–165). https://doi.org/10.1007/978-3-540-35488-8_6
Méndez, J. R., Cid, I., Glez-Peña, D., Rocha, M., & Fdez-Riverola, F. (2008). A comparative impact study of attribute selection techniques on naïve bayes spam filters. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: *Vol. 5077 LNAI* (pp. 213–227). https://doi.org/10.1007/978-3-540-70720-2_17
Méndez, J. R., Cotos-Yañez, T. R., & Ruano-Ordás, D. (2019). A new semantic-based feature selection method for spam filtering. *Applied Soft Computing Journal*, *76*, 89–104. https://doi.org/10.1016/j.asoc.2018.12.008
Méndez, J. R., Glez-Peña, D., Fdez-Riverola, F., Díaz, F., & Corchado, J. M. (2009). Managing irrelevant knowledge in CBR models for unsolicited e-mail classification. *Expert Systems with Applications*, *36*(2 PART 1), 1601–1614. https://doi.org/10.1016/j.eswa.2007.11.037
Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.
Perez-Diaz, N., Ruano-Ordás, D., Fdez-Riverola, F., & Méndez, J. R. (2012). SDAI: An integral evaluation methodology for content-based spam filtering models. *Expert Systems with Applications*, *39*(16), 12487–12500. https://doi.org/10.1016/j.eswa.2012.04.064
Pérez-Díaz, N., Ruano-Ordás, D., Fdez-Riverola, F., & Méndez, J. R. (2016). Boosting Accuracy of Classical Machine Learning Antispam Classifiers in Real Scenarios by Applying Rough Set Theory. *Scientific Programming*, *2016*, 1–10. https://doi.org/10.1155/2016/5945192
Ruano-Ordás, D., Basto-Fernandes, V., Yevseyeva, I., & Méndez, J. R. (2017). Evolutionary multi-objective scheduling for anti-spam filtering throughput optimization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: *Vol. 10334 LNCS* (pp. 137–148). https://doi.org/10.1007/978-3-319-59650-1_12
Ruano-Ordás, D., Fdez-Riverola, F., & Méndez, J. R. (2018). Using evolutionary computation for discovering spam patterns from e-mail samples. *Information Processing & Management*, *54*(2), 303–317. https://doi.org/10.1016/j.ipm.2017.12.001
Silva, R. M., Alberto, T. C., Almeida, T. A., & Yamakami, A. (2017). Towards filtering undesired short text messages using an online learning approach with semantic indexing. *Expert Systems with Applications*, *83*, 314–325. https://doi.org/10.1016/j.eswa.2017.04.055
Tandel, S. S., Jamadar, A., & Dudugu, S. (2019). A Survey on Text Mining Techniques. *2019 5th International Conference on Advanced Computing & Communication*

Systems (ICACCS), 1022–1026. https://doi.org/10.1109/ICACCS.2019.8728547
Unicode Inc. (2018). Full Emoji List, v11.0. Retrieved 3 December 2019, from https://unicode.org/emoji/charts/full-emoji-list.html
Vyas, T., Prajapati, P., & Gadhwal, S. (2015). A survey and evaluation of supervised machine learning techniques for spam e-mail filtering. *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 1–7. https://doi.org/10.1109/ICECCT.2015.7226077
Wikipedia contributors. (2019). List of emoticons. Retrieved 18 December 2019, from https://en.wikipedia.org/wiki/List_of_emoticons
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. In *Data Mining: Practical Machine Learning Tools and Techniques*. https://doi.org/10.1016/c2009-0-19715-5
Yevseyeva, I., Basto-Fernandes, V., Ruano-Ordás, D., & Méndez, J. R. (2013). Optimising anti-spam filters with evolutionary algorithms. *Expert Systems with Applications*, *40*(10), 4010–4021. https://doi.org/10.1016/j.eswa.2013.01.008
Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180