An Empirical Study on Word Sense Disambiguation for Adult Content Filtering

Igor Santos, Patxi Galán-García, Carlos Laorden Gómez, Javier Nieves, Borja Sanz, Pablo García Bringas and Jose Maria Gómez

DeustoTech Computing
Universidad de Deusto
isantos@deusto.es, patxigg@deusto.es, claorden@deusto.es, jnieves@deusto.es,
borja.sanz@deusto.es, pablo.garcia.bringas@deusto.es, jmgomez@deusto.es

Abstract. It is obvious that Internet can act as a powerful source of information. However, as happens with other media, each type of information is targeted to a different type of public. Specifically, adult content should not be accessible for children. In this context, several approaches for content filtering have been proposed both in the industry and the academia. Some of these approaches use the text content of a webpage to model a classic bag-of-word model to categorise them and filter the inappropriate content. These methods, to the best of our knowledge, have no semantic information at all and, therefore, they may be surpassed using different attacks that exploit the well-known ambiguity of natural language. Given this background, we present the first semantics-aware adult filtering approach that models webpages, applying a previous word-sense-disambiguation step in order to face the ambiguity. We show that this approach can improve the filtering results of the classic statistical models. *abstract* environment.

Key words: information filtering, content filtering, machine learning, web categorisation

1 Introduction

Information available in the Internet, sometimes, may not be shown or is not appropriate for every person. There are several examples of this type of media like gambling, dating, violence, racism or adult content [1]. Although these websites are sometimes illegal, they can be legal or barely legal and also easily accessible. However, there are some collectives, specially children, to whom this content is unacceptable to be shown.

An important amount of work has been performed in this problem using images as source of information e.g., [2] or the POESIA filter[1]. Another approach, is to use the textual information of the webpage to filter, which usually contains very explicit words that make the discrimination possible and easier than relying only on data from the images. These techniques are sometimes combined with

[1] `http://www.poesia-filter.org`

image classification as a further step if image filtering is not possible. In particular, there are several works [3,4,5] that use the classic bag-of-words model with a Vector Space Model (VSM) to weight the terms within the textual information.

The approach that both the academia and the industry have followed in order to filter these not appropriate contents is web filtering. These filters are broadly used in workplaces, schools or public institutions [1]. Information filtering itself can be viewed as a text categorisation problem (or image categorisation problem if images are used). In particular, in this work, we focus on pornographic site filtering. An important amount of work has been performed to filter these contents using the image information [2]. VSM, an algebraic approach for *Information Filtering* (IF), *Information Retrieval* (IR), indexing and ranking, represents the natural language documents in a mathematical manner through vectors in a multidimensional space. As in any other IR system, the VSM is affected by the characteristics of the text, with one of those features being *word sense ambiguity* [6]. The use of ambiguous words can confuse the model, permitting some webpages to bypass the filters.

In light of this background, we have performed an empirical study on Word Sense Disambiguation (WSD) for pornographic filtering and how this technique affects the categorisation results. In this way, our approach pre-processes webpages disambiguating the terms, using three different approaches, before constructing the VSM. Thereafter, based on this representation, we train several supervised machine-learning algorithms to detect and filter adult pages.

In summary, we advance the state of the art through the following contributions:

- We adopt a method to disambiguate terms in webpages.
- We conducted an empirical validation of WSD for adult filtering with an extensive study of several machine-learning classifiers.
- We show that the proposed method improves filtering rates; we discuss the weakness of the model and explain possible enhancements.

The remainder of this paper is organised as follows. Section 2 introduces our method to improve detection rates by using WSD. Section 3 provides an empirical evaluation of the experiments performed and presents the results. Section 4 presents the conclusions and outlines the avenues for future work.

2 Our Word Sense Disambiguation Approach

Today's attacks against Bayesian filters attempt to keep the content visible to humans, but obscured to filters. For instance, in spam filtering, attackers circumvent the filters by replacing suspicious words by innocuous terms with the same meaning [7,8]. In a similar vein, these filtering systems do not take into account the possible existence of ambiguous terms within the text [9]. This could lead to misclassified legitimate contents and attackers evading filtering, since it is expected that incorrectly disambiguated words may entail noise [10] and decrease the classification accuracy [11]. To solve this issue, we apply WSD to adult content filtering, a pre-processing procedure that is able to disambiguate confusing terms, to improve the capabilities of these filtering systems.

Our approach utilises *FreeLing* [12], a linguistic tool that includes a WSD approach. The WSD algorithm in FreeLing is known as the UKB algorithm [13], that relies on a semantic relation network to disambiguate the most likely senses for words in a text using the well-known PageRank algorithm [14]. Because the WSD needs a pre-processing stage in which the text is annotated with part-of-speech (PoS) tags, our e-mail message dataset was previously tagged using *Freeling* [15], a suite of analysis tools based on the architecture of [16].

In this way, we formally define a webpage $\mathcal{W}$ as a set composed of n terms t_i, $\mathcal{W} = \{t_1, t_2, \ldots, t_{n-1}, t_n\}$, where each term corresponds to a word (although we are aware of the possibility of applying WSD to collocations, we decided to leave this strength for future improvements of our system). Each t_i has a set of n senses s_i, $s = \{s_1, s_2, \ldots, s_{n-1}, s_n\}$. WSD selects the corresponding s_i for each term and generates a new relation of term-sense $t_{i,j}$, where i indicates the term and j denotes its corresponding sense.

Our method builds a model with term-sense relations, which we use to train several machine-learning classification algorithms. In order to perform this training, we first create an *ARFF* file (attribute relation file format) that describes the shared attributes (e.g., term-sense) for each instance (e.g., document). Secondly, we use the *Waikato Environment for Knowledge Analysis* (WEKA) [17] to build the desired classifiers. Finally, we test different machine-learning classification algorithms with WEKA.

The output of the WSD algorithm is a ranked list of senses for each actual word in a text, according to their probability as estimated by the machine-learning classifier. The main approach we follow is to attach the top scoring sense to a word, the way each ambiguous word is replaced by its word-form plus the predicted sense. In consequence, ambiguous terms with different senses in different occurrences represent different indexing tokens for the representation of the Web pages.

As WSD is not perfect in terms of accuracy, we have tested two additional disambiguation algorithms as control methods or baselines:

– The "most frequent" sense approach, which is a typical baseline in WSD evaluations. This approach consists on selecting the most frequent sense for a word according to a tagged reference corpus. In fact, WordNet senses for each word are sorted according to this criterion (using the corpus SemCor), so this method algorithmically corresponds to select the first sense provided by WordNet for each word.
– The "soft WSD" approach. Instead of taking the first predicted sense by our WSD module, we attach all possible senses but sorted by probability. In this way, two different occurrences of a word may be incorrectly disambiguated using the first selected sense, but they may lead to different sequences of senses. For instance, the word "jugar" ("to play") is incorrectly disambiguated with the same first sense in these real sentences extracted from our corpus of adult/non adult Web pages, but the sequence of senses

is different for each one, leading to different indexing terms: "A Isabelle le encanta jugar con su chico" ("Isabelle loves to play with her boy") vs. "Quisiera jugar con esas tetitas ricas" ("I would like to play with those yummy tits"). In these cases, we get the following sorted synsets respectively: (01072949-v, 02418686-v, 01076615-v, 01079480-v), and (01072949-v, 01076615-v, 01079480-v, 02418686-v). We call this approach "soft" because a hard decision about the sense is not taken, but in fact it augments the granularity of the different word references and, in consequence it is harder that two occurrences of the same word have exactly the same sorted synsets attached.

While designed as control methods, these algorithms improve classification accuracy in comparison with our primary WSD algorithm.

3 Empirical Validation

To validate our approach, we downloaded 4,500 web pages of both adult content and non-adult content such as technology, sports and so on. The dataset contained 2,000 adult and 2,500 non adult Spanish websites. The collection was conformed by gathering different adult websites and sub-pages within them. A similar approach was used to conform the non adult data. We generated two datasets with these data. The first dataset corresponded to the raw contents with no modification. The second dataset had a pre-processing step of the three different WSD methods. To model the content, we used the *Term Frequency – Inverse Document Frequency* (TF–IDF) [18] weighting schema, where the weight of the i^{th} term in the j^{th} document, denoted by $weight(i,j)$, is defined by $weight(i,j) = tf_{i,j} \cdot idf_i$. The *term frequency* $tf_{i,j}$ is defined as $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ where $n_{i,j}$ is the number of times the term $t_{i,j}$ appears in a document d, and $\sum_k n_{k,j}$ is the total number of terms in the document d. The inverse term frequency idf_i is defined as:

$$idf_i = \log \left(\frac{|\mathcal{D}|}{|\mathcal{D} : t_i \in d|} \right) \qquad (1)$$

where $|\mathcal{D}|$ is the total number of documents and $|\mathcal{D} : t_i \in d|$ is the number of documents containing the term t_i.

Once we parse the HTML code from all the web pages, we conducted the following methodology:

- **Cross Validation.** We have performed a *K-fold cross validation* with k=10. In this way, our dataset is 10 times split into 10 different sets of learning (90% of the total dataset) and testing (10% of the total data).
- **Learning the model.** For each fold we have performed the learning phase of the DMC. In this way, we added to the DMC model every website contained in each training dataset, adapting the compression model with each website.

- **Testing the model.** For each fold, we have used different criteria to select the class: Cross-Entropy and MDL. In this way, we measured the True Positive Ratio (TPR), i.e., the number of adult websites correctly detected, divided by the total number of adult webs $TPR = \frac{TP}{TP+FN}$ where TP is the number of adult websites correctly classified (true positives) and FN is the number of adult websites misclassified as non adult sites(false negatives).

  We also measured the False Positive Ratio (FPR), i.e., the number of non adult sites misclassified as adult divided by the total number of not adult sites $FPR = \frac{FP}{FP+TN}$ where FP is the number of not adult websites incorrectly detected as adult and TN is the number of not adult sites correctly classified.

  Furthermore, we measured the accuracy, i.e., the total number of the classifier's hits divided by the number of instances in the whole dataset $Accuracy(\%) = \frac{TP+TN}{TP+FP+FN+TN} \cdot 100$

  Besides, we measured the Area Under the ROC Curve (AUC) that establishes the relation between false negatives and false positives [19]. The ROC curve is obtained by plotting the TPR against the FPR.

Table 1. Results without WSD and with WSD using UKB algorithm.

	Normal DataSet				UKB WSD			
	Accuracy	TPR	FPR	AUC	Accuracy	TPR	FPR	AUC
Naïve Bayes	99.05%	0.99	0.01	0.99	98.89%	0.99	0.01	0.99
BN: K2	99.53%	0.99	0.00	1.00	99.52%	0.99	0.00	1.00
BN: TAN	99.61%	0.99	0.00	1.00	99.58%	0.99	0.00	1.00
Knn K=1	99.39%	0.99	0.00	0.99	99.24%	0.98	0.00	0.99
Knn K=2	99.11%	0.98	0.00	0.99	98.86%	0.97	0.00	0.99
Knn K=3	98.53%	0.97	0.00	0.99	98.24%	0.96	0.00	0.99
Knn K=4	98.29%	0.96	0.00	0.99	98.10%	0.96	0.00	0.99
Knn K=5	98.08%	0.96	0.00	0.99	97.94%	0.95	0.00	0.99
SVM: PolyKernel	99.85%	1.00	0.00	1.00	99.87%	1.00	0.00	1.00
SVM: Norm. PolyKernel	99.83%	1.00	0.00	1.00	99.83%	1.00	0.00	1.00
SVM: PUK	99.73%	1.00	0.00	1.00	99.82%	1.00	0.00	1.00
SVM: RBF	99.72%	0.99	0.00	1.00	99.71%	0.99	0.00	1.00
DT: J48	99.73%	1.00	0.00	1.00	99.72%	1.00	0.00	1.00
DT: RF N=10	99.84%	1.00	0.00	1.00	99.84%	1.00	0.00	1.00
DT: RF N=20	99.85%	1.00	0.00	1.00	99.85%	1.00	0.00	1.00
DT: RF N=30	99.84%	1.00	0.00	1.00	99.84%	1.00	0.00	1.00
DT: RF N=40	99.84%	1.00	0.00	1.00	99.84%	1.00	0.00	1.00
DT: RF N=50	99.84%	1.00	0.00	1.00	99.84%	1.00	0.00	1.00

Tables 1 and 2 show the obtained results. In this way, we can notice that the results are enhanced by WSD when using both Soft WSD and the Most Frequent Sense approach for most of the classifiers. In particular, the best results were obtained by the SVM with the Polynomial Kernel and using the Soft WSD

Table 2. Results using Soft WSD and the Most Frequent Sense methods.

	Soft WSD				Most Frequent Sense			
	Accuracy	TPR	FPR	AUC	Accuracy	TPR	FPR	AUC
Naïve Bayes	99.07%	0.99	0.01	0.99	99.09%	0.99	0.01	0.99
BN: K2	99.51%	0.99	0.00	1.00	99.51%	0.99	0.00	1.00
BN: TAN	99.62%	0.99	0.00	1.00	99.62%	0.99	0.00	1.00
Knn K=1	99.24%	0.98	0.00	1.00	99.21%	0.98	0.00	1.00
Knn K=2	98.81%	0.97	0.00	1.00	98.81%	0.97	0.00	1.00
Knn K=3	98.29%	0.96	0.00	1.00	98.32%	0.96	0.00	1.00
Knn K=4	98.11%	0.96	0.00	1.00	98.12%	0.96	0.00	1.00
Knn K=5	97.96%	0.95	0.00	1.00	97.97%	0.95	0.00	1.00
SVM: PolyKernel	99.90%	1.00	0.00	1.00	99.89%	1.00	0.00	1.00
SVM: Norm. PolyKernel	99.85%	1.00	0.00	1.00	99.83%	1.00	0.00	1.00
SVM: PUK	99.82%	1.00	0.00	1.00	99.80%	1.00	0.00	1.00
SVM: RBF	99.74%	0.99	0.00	1.00	99.73%	1.00	0.00	1.00
DT: J48	99.75%	1.00	0.00	1.00	99.77%	1.00	0.00	1.00
DT: RF N=10	99.86%	1.00	0.00	1.00	99.85%	1.00	0.00	1.00
DT: RF N=20	99.88%	1.00	0.00	1.00	99.86%	1.00	0.00	1.00
DT: RF N=30	99.89%	1.00	0.00	1.00	99.86%	1.00	0.00	1.00
DT: RF N=40	99.89%	1.00	0.00	1.00	99.86%	1.00	0.00	1.00
DT: RF N=50	99.89%	1.00	0.00	1.00	99.87%	1.00	0.00	1.00

approach, that attaches the list of possible senses of the words, ordered by the probability in the given context. Surprisingly, UKB algorithm did not enhance the results of both the Soft approach and the Most Frequent Sense approach did.

These results suggest that the there is an important difference in the text of adult webpages with respect to non adult pages and, therefore, there is no need for a complex WSD approach such as UKB, whereas simpler approaches like selecting the most common sense do enhance the filtering accuracy. Besides, the Soft Approach that includes every possible sense ordered by probability given a context, enhances the results, showing that the semantics of the words are important as happens in other text categorisation problems. This approach provides more senses besides the most common one and, thus, enhances the semantic information present in the model.

4 Conclusions

It's clear that porn is one of the most profitable business on the Internet, but, taking into consideration the content promoted in this topic, many entities are devoted to create tools to filter this kind of sources. For this reason, porn site webmasters need to circumvent all the firewalls deployed, in order to increase the reach of their adult content to increase their earnings. The approaches to avoid those filters may vary in many ways, but, in this work, we have focused on attacks that try to avoid text-based filters adding a layer of ambiguity to the textual content of the site. This attack is similar to the one found in spam filtering [9].

In light of this background, we have proposed a new approach to filter porn websites using Word Sense Disambiguation. The results obtained with this approach show improvements on the filtering rates, reaching a 98% of successful filtering with a simple disambiguation of each term found within the websites' text.

However, with the addition of Word Sense Disambiguation to the filtering system, there is a problem derived from the use of Natural Language to interpret the textual content: language phenomena. Each language has their own special features and characteristics, i.e., language phenomena, which creates a language dependency. Besides, as in any Information Retrieval approach using supervised techniques, it is complicate to acquire a good amount of carefully labelled data which, in addition to the need of gathering it in different languages, slows down the evolution of the filter. In a similar vein, Machine-learning approaches model the content using the Vector Space Model [20], which represents natural language documents in a mathematical manner through vectors in a multidimensional space, a not completely adequate approach from a linguistic point of view.

In this way, future lines of research include, firstly, the representation of websites using the enhanced Topic-based Vector Space Model (eTVSM) [21] which has proven to be effective in a similar domain as is spam [8]. Secondly, we will adopt some methods to fight attacks against the tokenisation step or statistical attacks such as the *Good Words Attack*. Thirdly, we will expand our knowledge base increasing our dataset, trying to even include different languages. Finally, we will try to reduce the negative impact of supervised learning by adapting semi-supervised approaches to the filtering system.

References

1. Gómez Hidalgo, J., Sanz, E., García, F., Rodríguez, M.: Web content filtering. Advances in Computers **76** (2009) 257–306
2. Choi, B., Chung, B., Ryou, J.: Adult Image Detection Using Bayesian Decision Rule Weighted by SVM Probability. In: 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, IEEE (2009) 659–662
3. Du, R., Safavi-Naini, R., Susilo, W.: Web filtering using text classification. In: Networks, 2003. ICON2003. The 11th IEEE International Conference on, IEEE (2003) 325–330
4. Kim, Y., Nam, T.: An efficient text filter for adult web documents. In: Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference. Volume 1., IEEE (2006) 3–pp
5. Ho, W., Watters, P.: Statistical and structural approaches to filtering internet pornography. In: Systems, Man and Cybernetics, 2004 IEEE International Conference on. Volume 5., IEEE (2004) 4792–4798
6. Sanderson, M.: Wsd and ir. In: Proceedings of the 17^{th} annual international ACM SIGIR conference on Research and development in information retrieval, Springer-Verlag New York (1994) 142–151

7. Nelson, B., Barreno, M., et al.: Misleading learners: Co-opting your spam filter. Machine Learning in Cyber Trust (2009) 17–51
8. Santos, I., Laorden, C., Sanz, B., Bringas, P.G.: Enhanced topic-based vector space model for semantics-aware spam filtering. Expert Systems With Applications (**39**) 437–444 doi:10.1016/j.eswa.2011.07.034.
9. Laorden, C., Santos, I., Sanz, B., Alvarez, G., Bringas, P.G.: Word sense disambiguation for spam filtering. Electronic Commerce Research and Applications **11** (2012) 290–298 doi:10.1016/j.elerap.2011.11.004.
10. Mavroeidis, D., Tsatsaronis, G., Vazirgiannis, M., Theobald, M., Weikum, G.: Wsd for exploiting hierarchical thesauri in text classification. Knowledge Discovery in Databases: PKDD 2005 (2005) 181–192
11. Xu, H., Yu, B.: Automatic thesaurus construction for spam filtering using revised back propagation neural network. Expert Systems with Applications **37** (2010) 18–23
12. Padr, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, ELRA (2012)
13. Agirre, E., Soroa, A.: Personalizing pagerank for wsd. In: Proceedings of the 12^{th} Conference of the European Chapter of the Association for Computational Linguistics. (2009) 33–41
14. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. (1999)
15. Carreras, X., Chao, I., Padró, L., Padró, M.: Freeling: An open-source suite of language analyzers. In: Proceedings of the 4th LREC. Volume 4. (2004)
16. Carreras, X., Padró, L.: A flexible distributed architecture for natural language analyzers. In: Proceedings of the LREC. Volume 2. (2002)
17. Garner, S.R., et al.: (Weka: The waikato environment for knowledge analysis)
18. Salton, G., McGill, M.: Introduction to modern information retrieval. McGraw-Hill New York (1983)
19. Singh, Y., Kaur, A., Malhotra, R.: Comparative analysis of regression and machine learning methods for predicting fault proneness models. Int. J. Comput. Appl. Technol. **35** (2009) 183–193
20. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM **18** (1975) 613–620
21. Becker, J., Kuropka, D.: Topic-based vector space model. In: Proceedings of the 6th International Conference on Business Information Systems. (2003) 7–12