New Classification Models for Detecting Hate and Violence Web Content

Shuhua Liu and Thomas Forss
Arcada University of Applied Sciences, Jan-Magnus Janssonin Aukio 1, 00560, Helsinki, Finland

Keywords: Web Content Classification, Topic Extraction, Topic Similarity, Sentiment Analysis, Imbalanced Classes, LDA Topic Models.

Abstract: Today, the presence of harmful and inappropriate content on the web still remains one of the most primary concerns for web users. Web classification models in the early days are limited by the methods and data available. In our research we revisit the web classification problem with the application of new methods and techniques for text content analysis. Our recent studies have indicated the promising potential of combing topic analysis and sentiment analysis in web content classification. In this paper we further explore new ways and methods to improve and maximize classification performance, especially to enhance precision and reduce false positives, thorough examination and handling of the issues with class imbalance, and through incorporation of LDA topic models.

1 INTRODUCTION

Today, the presence of harmful and inappropriate content on the web still remains one of the most primary concerns for web users. Our work on web content classification is motivated especially by the fact that certain groups of web pages such as those carry hate and violence content have proved in practice to be much harder to classify with good accuracy than many others. There is a great need for better content detection systems that can more accurately identify excessively offensive and harmful websites. In the mean time, advanced developments in computing methods have brought us many new and better means for textual content analysis such as new methods for topic extraction, topic modeling and sentiment analysis. It is our intention to make use of these new developments to develop better content classification models, especially for the detection of violence, intolerance and hateful web content.

Automatic classification of web pages has been studied extensively, using different learning methods and tools, investigating different datasets to serve different purposes (Qi and Davidson, 2007). Observing that hate and violence web pages often carry strong negative sentiment while their topics may vary a lot, we have in our recent study explored the potential of combining topic analysis and sentiment analysis in improving web content classification (Liu and Forss, 2014a, 2014b). Topic analysis consists of first topic extraction then topic similarity analysis. In topic extraction we apply different word weighting method and the centroid based text summarization tool to determine the topic representation for web pages and web categories. The topic similarity between each web page and a web category is then determined by computing their cosine similarity. In sentiment analysis we assess the sentiment tone and sentiment strength for each web page based on its topic representation, applying lexicon based sentiment analysis method SentiStrength. The extracted topic similarity and sentiment features form the data for learning classification models applying alternative machine learning methods.

Our study so far suggested that incorporating the sentiment dimension can bring much added value in the detection of sentiment-rich web categories such as those carrying hate, violent and racist messages. Classifiers However, there is still much room for performance improvement on completely new test sets. In order to help further improve the performance of the classification models, especially to increase precision and reduce false positives, in this study we develop new models by incorporating topic models, handling class imbalance in data, in order to build the most discriminative binary

Liu, S. and Forss, T..
New Classification Models for Detecting Hate and Violence Web Content.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 487-495
ISBN: 978-989-758-158-8

classifiers for detecting Hate and Violence web content.

In Section 2, we present related research. In Section 3, we describe our approach to web content classification and explain the methods and techniques used in topic extraction, sentiment analysis and topic modeling. In Section 4 we describe our data and experiments. We elaborate on imbalanced learning problem and discuss the methods and techniques for handling the issue. In Section 5 Our results are present and discussed. Section 6 concludes the paper and discusses possible directions for future work.

2 RELATED RESEARCH

Earliest studies on web classification already appeared in the late 1990s soon after the web was invented, exploring anchor description, link structure (Chakrabarti et al, 1998, Cohen, 2002), hierarchical structure (Dumais and Chen, 2000), web page classification with Positive Example Based Learning (PEBL) (Yu et al, 2004) and probabilistic relational models (Getoor et al, 2001; Broecheler et al, 2010; Fersini and Messina, 2013).

Addressing online safety and security problems, Hammami et al (2003) developed a web filtering system WebGuard that focuses on automatically detecting of adult content on the Web. It combines the textual content, image content, and URL. Last et al (2003) developed a system for anomaly detection on the Web using content-based methods, based on clustering analysis of web content accessed by a normal group of users. The content models of normal behavior are then used to reveal deviation from normal behavior at a specific location on the web. Elovici et al (2005) studied terrorist detection system that monitors the traffic emanating from the monitored group of users, issues an alarm if the access information is not within the typical interests of the group, and tracks down suspected terrorists by analyzing the content of information they access. Calado et al (2006) studied link-based similarity measures combined with text-based similarity metrics for the classification of web documents for anti-terrorism applications.

Studies most directly related with ours appeared more recently. Warner and Hirschberg (2012) presented an approach for detecting hate speech on the web and developed a mechanism for detecting some commonly used methods of evading common "dirty word" filters. Kwok and Wang (2013) developed classification models to detect Tweets against Blacks. Djuric et al (2015) studied hate speech detection in online user comments, and proposed a method to learn distributed low-dimensional representations of comments using neural language models, which are then fed to a classification algorithm as inputs. They address the issues of high-dimensionality and sparsity.

3 TOPIC ANALYSIS AND SENTIMENT ANALYSIS FOR WEB CONTENT CLASSIFICATION

3.1 Topic Extraction

3.1.1 Web Page Representation and Topic Representation

One of the most popularly used text representation methods is fixed-length vector space model of bag-of-words, due to its simplicity, efficiency and often surprisingly good level of accuracy and robustness (although theoretically it also has some widely acknowledged major flaws such as loss of word ordering in text, and ignorance of word semantics). We experimented with bag-of-n-grams in another study and found that the models show certain performance improvements in some cases, but not always in a consistent way (Liu and Forss, 2014b). Considering the much heavier computation related with n-gram models and the uncertainty with performance, in this study we only adopt the bag-of-words representation, and treat each web page or web category as a vector, with each column of the vector is a word, with tf-idf weight as its value.

3.1.2 Topic Extraction for Individual Pages

For the purpose of web content classification, we consider tf-idf term weighting based approach be a sufficiently effective and more efficient approach for topic extraction from a web page, as the extracted content (terms) are only used as cues for classifying the content instead of presenting to human users.

Textual information on a web page includes multi-types. From our database, we found 18 out of the 30 text attributes could be useful sources of important content and will be used as raw text input for features extraction. They belong to four groups: (1) full page; (2) text paragraphs (a less noisy version of the full page); (3) meta-content (url, title, description, headings, highlights, special fonts, table

and list elements); (4) title and description of outgoing links. These four types of content attributes are used as raw data of the feature extraction process, individually and jointly (details in section 4). By applying different compression rates, we obtained different sets of topic words (top 500, top 1000 and top 15,000).

3.1.3 Topic Extraction for Collection of Web Pages

From a collection of web pages that belong to one web category, we obtain a topic representation of the category through summarization of all the web pages in the collection. Here we apply the Centroid method based MEAD summarization tool (Radev et al, 2004) to the Hate and Violence web page collections separately. MEAD has been a benchmarking multi document text summarization tool. By applying different compression rate, different sets of topic terms can be obtained for each category. In our case, we try to match up the number of extracted terms for each web category with the number of extracted terms for each web page, that is top 500, top 1000, and top 15,000 respectively.

3.2 Page vs. Category Topic Similarity

We use topic similarity to measure the content similarity between a web page and a web category. Our web page-category similarity is simply implemented as the cosine similarity between topic terms of a web page and topic terms of each web category. The Cosine similarity measure is generic and robust. We consider it as a good starting choice for our purpose.

3.3 Sentiment Feature Extraction

Sentiment analysis methods generally fall into two categories: (1) the lexical approach - unsupervised, use direct indicators of sentiment, i.e. sentiment bearing words; (2) the learning approach - supervised, classification based algorithms, exploit indirect indicators of sentiment that can reflect genre or topic specific sentiment patterns (Pan and Lee, 2008; Liu, 2012; Thelwall et al, 2012).

SentiStrength (Thelwall et al, 2010, 2012) takes a lexical approach to sentiment analysis, making use of a combination of sentiment lexical resources, semantic rules, heuristic rules and additional rules. While most opinion mining algorithms attempt to identify the polarity of sentiment in text - positive, negative or neutral, SentiStrength gives sentiment measurement on both positive and negative direction with the strength of sentiment expressed on different scales. To help web content classification, we use sentiment features to get a grasp of the sentiment tone of a web page. This is different from the sentiment of opinions concerning a specific entity, as in traditional opinion mining literature.

We apply unsupervised approach with the original SentiStrength system and modifications (Liu and Forss, 2014a). For each web page, sentiment features are extracted by using the key topic terms obtained from the topic extraction process as input to SentiStrength. This gives sentiment strength value for each web page in the range of -5 to +5, with -5 indicating strong negative sentiment and +5 indicating strong positive sentiment. Considering that the number of strong sentiment words also has a large effect on the overall sentiment strength, we defined an additional sentiment feature, which is weighted sum of each of strong SentiStrength value (-3, -4, -5 and +3, +4, +5) normalized on the total number of words of the page.

3.4 LDA Topic Models

LDA (Latent Dirichlet Allocation) topic modeling and its variations offer unsupervised methods for extracting topics from a collection of web pages. Topic models are probabilistic models for discovering the hidden thematic structure in large document collections based on a hierarchical Bayesian analysis method (Blei et al, 2003; Blei, 2012). Topic modeling offers a more sophisticated treatment of the topic extraction problem with an unsupervised approach. By discovering patterns of word use and connecting documents that embrace similar patterns, topic models prove to be a powerful technique for finding topic structure in text collections. Topic models can help us answer questions such as what topics are contained in the document collection, what subject matters each document discusses, and how similar one document is to another. Topics are defined as a distribution over a fixed vocabulary of terms; documents are defined as a distribution over topics; with the distributions all automatically inferred from analysis of the text collection. Document distribution over the topics gives a concise summary of the document. When labeled data are available, topic modeling can also be applied to build document classification models, in which the topic features instead of word features are used for learning with much reduced dimensionality.

Topic models have been applied to many kinds

of documents, including email, scientific abstracts and newspaper archives. Here we apply LDA topic modeling to web pages. Raw data of each web page is represented as bag-of-words, but the output representation for each web page will be a distribution over hidden topics in the web collection, and the output representation of hidden topics will be distribution over words. Unsupervised topic models are very useful way to understand the overall thematic composition and distribution of web pages and collection. Supervised topic modeling on the other hand can be applied to text classification directly (Blei and McAuliffe, 2007).

4 DATA AND EXPERIMENTS

4.1 Multiclass Classifier vs Binary Classifier

Our effective dataset for training is a collection of about 80,000 web pages in 20 categories (multi class single label). Many research have approached web classification as a multiclass classification problem, to learn a classification model that can assign any new data to one of the multi- mutually exclusive classes. Alternatively, multiclass classification can also be mapped into a series of simpler binary classification problems, and then the subsequent combination of the outcomes to derive the multiclass prediction (Rocha and Goldenstein, 2013).

Mapping multiclass problems onto a set of simpler binary classification problems will run into efficiency problems when there are hundreds or even thousands of classes. However our problem has only 20 classes, which is easily manageable with 20 binary classifiers. Especially in this study our main concern are the two classes Hate and Violence. So binary classifiers are a natural choice for us.

There are in general three approaches to reduce multiclass to binary classification problems: One-vs-All, One-vs-One, and Error Correcting Output Codes (ECOC). We take a practical approach and develop binary classifiers using OVA, which we consider most closely resemble the real practice.

4.2 Class Imbalance, Sampling Strategies, Covariate Shift

Changing from multiclass classification to binary classification brings an issue of class imbalance. Table 1 shows the distribution of Hate and Violence classes in our training and test sets.

Table 1: Data sets.

Class	Size and Sampling of Data Sets	
	Training set	*Test set*
Complete data set	67212/73107 (Hate/Violence)	3153/3086 (Hate/Violence)
Hate	1733 (2.58%)	184 (5.84%)
Violence	1400 (1.92%)	135 (4.37%)

Highly skewed class distributions like ours are not uncommon in real world applications. However, learning algorithms usually assume that the ratios of each class are close to equal and the errors associated with each class have the same cost. With imbalanced dataset, the cost gets skewed in favor of the majority class, and models built with imbalanced dataset will cause the underrepresented class to be overlooked or even ignored. Thus, a common belief is that we should balance the class prevalence before training a classifier to improve performance.

Solutions for imbalanced learning thus include sampling based, cost sensitive methods and active learning methods. Resampling tries to achieve dataset balance artificially so that the prevalence of the minority class is enriched. Generally two sampling strategies can be employed for balancing the classes: oversampling (adding instances to the minority class) and under-sampling (removing instances from the majority class).

The SMOTE algorithm (Synthetic Minority Over-sampling Technique) developed by Chawla, Hall and Kegelmeyer in 2002, is a commonly used over-sampling technique. It oversamples the minority class by generating new minority-class instances (e.g. creating artificial synthetic examples of k nearest class neighbors), combines with random under-sampling of the majority class. It has many variations as well. Under-sampling strategy then selects a subset of majority class samples randomly thus increase the share of the target class. The Wilson's Editing approach removes majority-class instances which are too close to the majority-minority class boundary. The EasyEnsemble and BalanceCascade makes informed under-sampling integrated with boosting (He and Garcia, 2009).

Resampling may or may not help, depending on the problem. Zumel investigated if balancing classes improves performance for logistic regression, SVM, and Random Forests. She found that, “balancing class prevalence before training a classifier does not across-the-board improve classifier performance. In fact, “it is contraindicated for logistic regression models”, although it may help random forest and SVM classifiers (Nina Zumel, 2015, KDNuggets).

Another set of solutions is cost based learning. Cost sensitive modeling weights the costs of misclassifying the majority class (false negatives) and the minority class (false positives) separately. By training the learner to minimize overall cost it gives more incentive for more true positives. Cost-sensitive bootstrap sampling uses misclassification costs to select the best training distribution. In addition, there are also cost-sensitive ensembles, and cost-sensitive functions incorporated directly into classification methods: Cost-Sensitive Bayesians, Cost-Sensitive SVMs etc. (He and Garcia, 2009).

A different view of the class imbalance problem is that poor performance is caused by there not being enough patterns belonging to the minority class, not by the ratio of positive and negative patterns itself. Generally when there is enough data, the "class imbalance problem" doesn't arise (He and Garcia, 2009). Thus the essence of issues with imbalanced learning is not only the disproportion of positive and negative classes, but also the poor representativeness of the minority class due to its small size.

The same issue could happen to majority class as well. For example, in many classification tasks on web scale, positive and unlabeled data are widely available, whereas collecting a reasonable and representative sampling of the negative examples could be challenging, because the negative data set, as the complement of the positive one, should uniformly represent the universal set excluding the positive class, but such probability distribution can hardly be approximated (Yu et al, 2004).

When a dataset is artificially balanced, it often implies that there is close to equal prior probability of positive and negative patterns. When that is not the case, the model could make poor predictions by over-predicting the minority class. In practice it often happens that the training set and test set not only both have highly skewed class priors, but the class distribution also very different from each other. This is referred to as Covariate Shift issue, which often cause model performance degradation in the test set, which can be especially obvious with Naive Bayes and Logistic Regression based classifiers (Bickel et al, 2007).

4.3 Data Preparation: Preprocessing and Feature Set

4.3.1 Raw Text and Pre-Processing

To prepare the text input for pre-processing, we extract the textual content from the database and create five alternative raw content: (1) full page + meta-content, (2) full page + meta-content with up-weighting (meta content applied twice), (3) text paragraph + meta-content; (4) text paragraphs + meta-content with up-weighting; (5) meta-content only. The different text inputs will be tested out to understand the effect of different text attributes.

Preprocessing include tokenization and stop-words removing, no stemming. Same process executed for the test set.

4.3.2 Feature Set

Topic similarity features contains cosine similarity and its transformation and expansion: (1) Cosine similarity with the target category; (2) Log cosine similarity, (3) power to 1/2 1/3, ¼, 1/5, 1/6 of Cosine similarity; (4) Cosine similarity adjusted by effect of semantics on word frequency; (5) cosine similarity adjusted by effect of semantics on tf-idf.

Sentiment features: (1) Counts of words with SentiSrength value as -3, -4, -5, +3, +4, +5; (2) weighted sum of word frequency and SentiStrength value normalized on page length.

We also tried to explore the effect of outgoing links through features based on count of number of links in a page, as well as number of links that overlaps with list of links from a collection of pages for a web category (Hate and Violence).

These features are only relevant for classifiers that combine topic and sentiment analysis. The LDA topic model based classifiers only rely on the distribution over topics of web pages.

4.4 Modeling for Detecting Hate and Violence

To handle the data imbalance issues, we adopted an under-sampling strategy, combined with cost sensitive methods and threshold control to adjust the model more in favor of the minority class. We experimented with several different sampling strategies including the natural distribution (close to 5_95, highly skewed), 20_80 (unbalanced) and 50_50 (balanced). This means we trained our models using a training set where the target was present at its native prevalence, as well as enriched by a large multiplier to ten times and twenty-five times its native prevalence or simply balancing the classes. In cases of resampling, the negative samples are uniformly drawn from the other 19 classes in the original database.

To our surprise, although using different raw data inputs resulted in different models, the differences on performance level are not that

significant. So we only continued with the option that uses least amount of attributes.

We try to make full use of all the effective positive samples of Hate and Violence. However, it is unavoidable that the classes may be incompletely sampled and falls short in coverage to represent a complete reality. To address this and the covariate shift issue, we incorporated semantic information into the topic representation of the two classes. We also apply threshold control on test set.

Semantics is incorporated into modeling process through up-weighting of concept terms for Hate and Violence, which are collected from ConceptNet (http://conceptnet5.media.mit.edu), by extracting terms that have associative relations (DefinedAs, SymbolOf, IsA, HasA, UsedFor, CapableOf, Causes, RelatedTo) with seed words such as Violence, Hate, Racism, Discrimination.

Several model types are considered: NaiveBayes, NaiveBayes Kernel, SVM, NeuralNet, in combination with the different training sets utilizing different raw text input. Cost-sensitive classifiers are developed in which misclassifying a negative as positive has a larger cost than misclassifying a positive as negative. Different parameters in the cost function are tested.

For LDA topic modeling based classifiers, we adopted Mallet (http://mallet.cs.umass.edu/) sLDA, NaiveBayes models. The sLDA algorithm is a supervised version of the LDA, where the number of topics is set as the number of classes.

Models are evaluated against the same test set (one for Hate, one for Violence) with the classes at their native prevalence. Accuracy is not a good evaluation metrics for imbalanced learning. So our performance metrics focuses on Precision, Recall, F-measure, PR curves.

4.5 Results and Discussion

Large amounts of experiments were conducted to develop different types of classifiers for Hate and Violence. Our experiences and key results are summarized below.

4.5.1 Topic Vector and Raw Data Input

Our earlier experience indicated that the longer the vector, the results are better. However, comparing the three alternative topic vector lengths: 500, 1000, 15000, we found that the vector length 1000 showed better results in this setting. So we continued our experiments with topic length equals to 1000 words.

4.5.2 Sampling Strategies

When we build models the straightforward way, with the training and test sets at native prevalence, the results are not as good as the more balanced datasets. The changing of model types does not make much difference. Although the validation results are decent, performance on test set very poor. Enriching the target class prevalence during training helped improve performance on both the validation and test result (with the target at its native prevalence).

All models performance degrades as target prevalence decreases. Performance at natural prevalence is worse than at the enriched prevalence. May be oversampling can result in different experience.

4.5.3 Modeling Methods

NaiveBayes models are calibrated to the training distribution, thus changes in the distribution will naturally affect model performance. SVM models tend to stay more stable as SVM's training procedure is not strongly dependent on the class distributions. However our SVM and NeuralNet based models do not perform better than Kernel based NaiveBayes models. And they take much longer time in training the models, especially if we consider cost-sensitive models.

With cost-sensitive models, setting the cost function as TP (0), FP (1.5), TN (0), FN (1) seems produce best result in terms of higher precision and acceptable recall.

Threshold: tuning the threshold can help improve precision and lower recall on positive class or the other way around. We only apply threshold control when tune the model performance on test set (range 0.1-0.4). Overall, we found that the effect of threshold control is still rather limited (around max 3% gain in precision, with tradeoff on recall), can't bring performance up in a significant way.

However, LDA topic models based classification performance is significantly better than all other models – although the precision level for positive class can be similar, recall level is much higher, which gives room to bring up the precision level at trade-off of recall level.

4.5.4 Performance

Lots of results were generated from extensive experiments. The best performing models are summarized in Table II. What we should note is that our test set is collected totally separately from the

training set. It is thus very different from performance results on a hold-out set in a more common data mining sense.

Table 2: Best Performance Models.

Best Performing Models (P: Precision, R: Recall, for target class)		
Models Types	***Classifiers for Hate***	
	P	R
All features (with semantics, outgoing link features), cost-sensitive, threshold on test set	52.29%	30.98%
With links but no semantics, cost-sensitive, threshold	53.26%	26.63%
No links, no semantics, cost-sensitive, threshold	46.23%	26.63%
Topic models	52.00%	92.00%
All combined = with semantics, threshold, cost-function; Topic models = no semantics, no threshold, no cost-function, no links (details in a separate article)		

Cosine similarity features considering the effect of ConceptNet terms on tf-idf seems have no positive effect on the classifier performance. But features related with outgoing links seem to contribute to improve precision and reducing false positives.

Performance on Violence is rather disappointing, and even consistently inferior to performance on Hate. One obvious reason could be the size of the samples, and the even weaker prevalence of positive samples in the data set (both training and testing). The other is that, there are more differences between the training and test set for Violence. Up-weight of meta-data brings down the recall level, increases precision a little bit. We are continuing with modeling for better Violence classifiers.

5 CONCLUSIONS

In this study we seek to improve content classifiers for detecting Hate and Violence on the web. We explored new ways and methods to enhance precision and reduce false positives. We give thorough examination of the issues with the dataset reliability, class imbalance and covariate shift.

To handle class imbalance issue we looked at different strategies and adopted the under-sampling scheme. Our experiments indicate that artificial balancing of the classes brings a positive effect on the model performance. The balanced learning overall produced models outperform the unbalanced learning in terms of both the validation and testing results. This is especially evident with validation results. However, in neither case the classification performance on the test set reached a satisfactory level using current methods. We need to try the oversampling approach.

We also noticed a rather big difference for models' performance on Hate and Violence. Classifiers for Hate consistently outperform the classifiers for Violence when same methods for modeling are employed. This may be partly explained by the difference in the size of their training samples, as well as the difference in degree of covariate shift. The training samples for Hate are more similarity to its test samples than in the case of Violence. This also means the covariate shift issue has not been effectively handled yet. We need add other methods for dealing with the shift of prevalence of the target class.

In terms of the modeling methods, we find NaiveBayes still a very competitive method in terms of both efficiency and performance. When added cost-sensitive learning, the model training process become very intensive for other models such as SVM and NeuralNet. Even if there may be some gains of validation performance with more complicated models, there is no guarantee that such gains can be passed onto the test results as well. In fact, the test results very often inferior to NaiveBayes (Kernel) models. This again confirms that a simple approach could often be competitive by building very large models, and outperform more sophisticated methods.

In our application, the feature set seems to have a much bigger influence on performance than the alternative modeling methods and threshold control, which indicates that selection of good features is critical to the classification results regardless of the algorithms. With our previous models we were able to achieve higher validation performance, may be due to the reason that we included additional topic similarity features (Liu and Forss, 2014a). So our immediate next step improvement will incorporate the rich topic similarity features we have identified from our earlier studies, to include the topic similarity of a web page to other 19 web categories as well into the models for detecting Hate and Violence.

LDA topic modeling based classification models already proved to be very promising approach and we already started exploring more. In addition, another natural extension is to integrate LDA topic modeling based approach with similarity-based approach. Other possibilities for future studies include modeling using positive and unlabeled data,

classifier ensemble, as well as integration of text based and image based classifiers.

ACKNOWLEDGEMENTS

This research is supported by the Tekes funded DIGILE D2I research program, Arcada Research Foundation, and our industry partners.

REFERENCES

Blei, D, Ng, A., and Jordan, M. I. 2003. *Latent dirichlet allocation*. Advances in neural information processing systems. 601-608.

Blei, D. M. and J. D. McAuliffe, Supervised Topic Models, Neural Information Processing Systems 21, 2007

Blei, D. 2012. Probabilistic topic models. Communications of the ACM, 55(4):77–84, 2012

Bickel S., M. Bruckner and T. Scheffer, Discriminative Learning for Differing Training and Test Distributions (ICML 2007), Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007.

Broecheler M., L. Mihalkova, L. Getoor. Probabilistic similarity logic. In Proc. of Uncertainty in Artificial Intelligence, 2010

Calado, P., Cristo, M., Goncalves, M. A., de Moura, E. S., Ribeiro-Neto, B., and Ziviani, N. 2006. Link-based similarity measures for the classification of web documents. *Journal of the American Society for Information Science and Technology* (57:2), 208-221.

Chakrabarti, S., B. Dom and P. Indyk. 1998. Enhanced hypertext categorization using hyperlinks. *Proceedings of ACM SIGMOD 1998.*

Chen, Z., Wu, O., Zhu, M., and Hu, W. 2006. A novel web page filtering system by combining texts and images. *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 732–735. Washington, DC IEEE Computer Society.

Cohen, W. 2002. Improving a page classifier with anchor extraction and link analysis. In S. Becker, S. Thrun, and K. Obermayer (Eds.), *Advances in Neural Information Processing Systems* (Volume 15, Cambridge, MA: MIT Press) 1481–1488.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V. and N. Bhamidipati, Hate Speech Detection with Comment Embeddings. In Proceedings of the 24th International World Wide Web Conference, May 2015

Dumais, S. T., and Chen, H. 2000. Hierarchical classification of web content. *Proceedings of SIGIR'00,* 256-263.

Elovici, Y., Shapira, B., Last, M., Zaafrany, O., Friedman, M., Schneider, M., and Kandel, A. 2005. Content-based detection of terrorists browsing the web using an advanced terror detection system (ATDS), *Intelligence and Security Informatics* (Lecture Notes in Computer Science Volume 3495, 2005), 244-255.

Fersini, E. and E. Messina, Web page classification through probabilistic relational models. Int. Journal of Pattern Recognition and Artificial Intelligence, 27(04), 2013

Fersini, E., Messina, E., & Pozzi, F. A. (2014). Sentiment analysis: Bayesian Ensemble Learning. Decision Support Systems, 68, 26-38.

Getoor L., E. Segal, B. Taskar and D. Koller (2001), Probabilistic models of text and link structure for hypertext classification, in Proc. Int. Joint Conf. Artificial Intelligence, Workshop on Text Learning: Beyond Supervision, pp. 24-29.

Hammami, M., Chahir, Y., and Chen, L. 2003. WebGuard: web based adult content detection and filtering system. *Proceedings of the IEEE/WIC Inter. Conf. on Web Intelligence* (Oct. 2003), 574 – 578.

He H. and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. on Knowledge and Data Engineering, vol. 21, pp. 1263-1284, 2009.

Kwok, I. and Y. Wang, Locate the Hate: Detecting Tweets against Blacks. In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, June 2013

Last, M., Shapira, B., Elovici, Y., Zaafrany, O., and Kandel, A. 2003. Content-Based Methodology for Anomaly Detection on the Web. *Advances in Web Intelligence,* Lecture Notes in Computer Science (Vol. 2663, 2003), 113-123.

Liu, B. 2012. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers 2012

Liu S. and T. Forss, Improving Web Content Classification based on Topic and Sentiment Analysis of Text, Proceedings of KDIR2014, the 6th Int. Conf. on Knowledge Discovery and Information Retrieval, October 21-24, 2014 Rome, Italy

Liu S. and T. Forss, Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification, Proceedings of KDIR2014 Special Session on Text Mining (SSTM), October 21-24, 2014 Rome, Italy

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2), 1-135, July 2008

Qi, X., and Davidson, B. 2007. *Web Page Classification: Features and Algorithms.* Technical Report LU-CSE-07-010, Dept. of Computer Science and Engineering, Lehigh University, Bethlehem, PA, 18015

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., and Zhang, Z. 2004a. MEAD-a platform for multidocument multilingual text summarization. *Proeedings of the 4th LREC Conference* (Lisbon, Portugal, May 2004)

Rocha A. and S. Goldenstein, Multiclass from Binary: Expanding One-vs-All, One-vs-One and ECOC-based Approaches. IEEE Transactions on Neural Networks and Learning Systems, August 2013

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61(12),* 2544–2558.

Thelwall, M., Buckley, K., and Paltoglou, G. 2012. Sentiment strength detection for the social Web. *Journal of the American Society for Information Science and Technology, 63(1),* 163-173.

Warner, W. and J. Hirschberg. Detecting hate speech on the world wide web, Proceedings of the 2nd Workshop on Language in Social Media (pp. 19-26). Association of Computational Linguistics, June 2012

Yu, H., Han, J., and Chang, K. C.-C. 2004. PEBL: Web Page Classification without Negative Examples. *IEEE Trans. on Knowledge and Data Eng. (16:1),* 70-81.

Yu H., Jiawei Han, and Kevin Chen-Chuan Chang, PEBL: Web Page Classification without Negative Examples IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 1, January 2004