Web Content Classification based on Topic and Sentiment Analysis of Text

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

Keywords: Web Content Classification, Text Summarization, Topic Similarity, Sentiment Analysis, Online Safety Solutions.

Abstract: Automatic classification of web content has been studied extensively, using different learning methods and tools, investigating different datasets to serve different purposes. Most of the studies have made use of the content and structural features of web pages. However, previous experience has shown that certain groups of web pages, such as those that contain hatred and violence, are much harder to classify with good accuracy when both content and structural features are already taken into consideration. In this study we present a new approach for automatically classifying web pages into pre-defined topic categories. We apply text summarization and sentiment analysis techniques to extract topic and sentiment indicators of web pages. We then build classifiers based on combined topic and sentiment features. A large amount of experiments were carried out. Our results suggest that incorporating the sentiment dimension can indeed bring much added value to web content classification. Topic similarity based classifiers solely did not perform well, but when topic similarity and sentiment features are combined, the classification model performance is significantly improved for many web categories. Our study offers valuable insights and inputs to the development of web detection systems and Internet safety solutions.

1 INTRODUCTION

Web content classification, also known as web content categorization, is the process of assigning one or more predefined category labels to a web page. Classification models are built through training and validation using a set of labeled data, and are then applied to label new web pages, or in other words, to detect if a new webpage falls into certain predefined categories.

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). However, practical experience has shown that certain groups of web pages, such as those containing hatred and violence, are much harder to classify with good accuracy even when both content and structural features are already taken into consideration. There is a great need for better content detection systems that can accurately identify excessively offensive and harmful websites.

Hate and violence related web pages often carry strong negative sentiment while their topics may vary a lot. In the meantime, advanced developments in computing methodologies and technology have brought us many new and better means for text content analysis such as topic extraction, topic modeling and sentiment analysis. In our research we set out to explore the effectiveness of combined topic and sentiment features for improving automatic classification of web content. We present a new approach for automatically classifying web pages into pre-defined topic categories, which first applies text summarization and sentiment analysis techniques to extract topic and sentiment indicators of web pages, and then builds classifiers based on the extracted topic and sentiment features. Our results offer valuable insights and inputs to the development of web detection systems and Internet safety solutions.

The rest of the paper is organized as follows: in Section 2, we give an overview of related research in web content classification and Internet safety and security solutions. In Section 3, we describe our approach and explain the methods and techniques used in topic extraction and sentiment analysis. In Section 4, we describe our data and three sets of

Liu S. and Forss T..
Web Content Classification based on Topic and Sentiment Analysis of Text.
DOI: 10.5220/0005101803000307
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2014), pages 300-307
ISBN: 978-989-758-048-2

experiments including solely sentiment-based classifiers on three web categories, classifiers based on combined topic and sentiment features while extending our work first to eight web categories, then further to all web categories. Our results are presented and analyzed. Section 5 concludes the paper.

2 RELATED RESEARCH

Earliest studies on web classification already appeared in the late 1990s soon after the web was invented. Chakrabarti et al (1998) studied hypertext categorization using hyperlinks. Cohen (2002) combined anchor extraction with link analysis to improve web page classifiers. The method exploits link structure within a site as well as page structure within hub pages, and it brought substantial improvement on the accuracy of a bag-of-words classifier, reducing error rate by about half on average (Cohen, 2002).

Dumais and Chen (2000) explored the use of hierarchical structure for classifying a large, heterogeneous collection of web content. They applied SVM classifiers in the context of hierarchical classification and found small advantages in accuracy for hierarchical models over flat (non-hierarchical) models. They also found the same accuracy using a sequential Boolean decision rule and a multiplicative decision rule, with higher efficiency. Yu et al (2004) presented a framework for web page classification without negative examples, called Positive Example Based Learning (PEBL), to address issues related with the acquisition of negative training examples and to avoid bias. They found that given the same set of positive examples, their Mapping-Convergence algorithm outperforms one-class SVMs, and was almost as accurate as the traditional SVMs.

There is a huge amount of research on text classification in general. However, web content classification differs from general text categorization due to its special structure, meta-data and its dynamics. Shen et al (2004, 2007) studied web-page classification based on text summarization. They gave empirical evidence that web-page summaries created manually by human editors can indeed improve the performance of web-page classification algorithms. They proposed sentence-based summarization methods and showed that their summarization-based classification algorithm achieves an approximately 8.8% improvement as compared to pure-text-based classification algorithm, and an ensemble classifier using the improved summarization algorithm achieves about 12.9% improvement over pure-text-based methods. Our approach differs in that we take a word-based instead of a sentence-based approach.

In recent years, there have been many studies on text classification techniques for social media analysis (e.g. customer reviews, twitter), sentiment analysis, etc. For example, an interesting study by Zhang et al (2013) investigated the classification of short texts using an information path to deal with the less informative word co-occurrences and sparseness with such texts. Their method makes use of ordered subsets in short texts, which is termed “information path”. They found the classification based on each subset to result in higher overall accuracy than classifying the entire data set directly.

Related with online safety solutions, Hammami et al (2003) developed a web filtering system WebGuard that focuses on automatically detecting and filtering adult content on the Web. It combines the textual content, image content, and URL of a web page to construct its feature vector, and classify a web page into two classes: Suspect and Normal. The suspect URLs are stored in a database, which is constantly and automatically updated in order to reflect the highly dynamic evolution of the Web.

Last et al (2003) and Elovici et al (2005) developed systems for anomaly detection and terrorist detection on the Web using content-based methods. Web content is used as the audit information provided to the detection system to identify abnormal activities. The system learns the normal behavior by applying an unsupervised clustering algorithm to the content of web pages accessed by a normal group of users and computes their typical interests. The content models of normal behavior are then used in real-time to reveal deviation from normal behavior at a specific location on the web (Last et al, 2003). They system can thus monitor 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 (Elovici et al, 2005).

In more recent years, Calado et al (2006) studied link-based similarity measures as well as a combination with text-based similarity metrics for the classification of web documents for Internet safety and anti-terrorism applications (Calado et al. 2006). Qi and Davidson (2007) presented a survey of features and algorithms in the space of web content classification.

3 WEB CONTENT CLASSIFICATION BASED ON TOPIC AND SENTIMENT ANALYSIS

Our approach to web content classification is illustrated in Figure 1. Exploring the textual information, we apply word weighting, text summarization and sentiment analysis techniques to extract topic features, content similarity features and sentiment indicators of web pages to build classifiers.

In this study we only take into consideration the page attributes that are text-related. Our focus is on added value to web classification that can be gained from textual content analysis. We should point out that structural features and hyperlink information capture the design elements of web pages that may also serve as effective indicators of their content nature and category (Cohen, 2002). They contain very useful information for web classification. In addition, analysis of images contained in a web page would provide another source of useful information for web classification (Chen et al, 2006; Kludas, 2007). However, these topics are dealt with in other projects.

3.1 Topic Extraction and Text Summarization

The Topic Extraction step takes web textual information as input and generates a set of topic terms. We start with extracting topics from each web page and then each of the collections of web pages belonging to same categories. The extracted topics hopefully give a good representation of the core content of a web page or a web category.

Text summarization tools have the capability to distil the most important content from text documents. However, most of the text summarization systems are concerned with sentence extraction targeted for human users (Radev et al, 2004a, 2004b). The summarization approach that Shen et al (2004, 2007) took is also sentence based. To help web content classification, we believe simple term extraction could already be a sufficiently effective and more efficient approach, as the extracted content (terms) are only used as cues for classifying the content instead of presenting it to human users. Thus, in this study, we used the time-tested tf-idf weighting method (Salton and Buckley, 1988) to extract topic terms from web pages and their collections.

3.1.1 Topic Extraction for Individual Pages

For each webpage, we make use of its different content attributes (full page or meta-content) as input. By applying different compression rates, we obtained different sets of topic words (for example top 30, top 50, top 20%, 35%, 50%, 100%).

3.1.2 Topic Extraction for Web Categories

The topic terms of a web category are obtained through summarization of the collection of all web pages in the same category. For each collection, we apply the Centroid method of the MEAD summarization tool (Radev et al, 2004a; 2004b) to create summaries of it. Through this we try to extract topics that are a good representation of a specific web category. The Centroid method has been a benchmarking text summarization method. Given a document or a collection of documents to be summarized, it creates a cluster and all sentences in the cluster are represented using a tf-idf weighted vector space model. A pseudo sentence, which is the average of all sentences in the cluster, is then calculated. This pseudo sentence is regarded as the centroid of the document (cluster); it is a set of the most important/informative words of the whole cluster, and thus can be regarded as the best representation of the entire document collection. By applying different compression rates, different sets

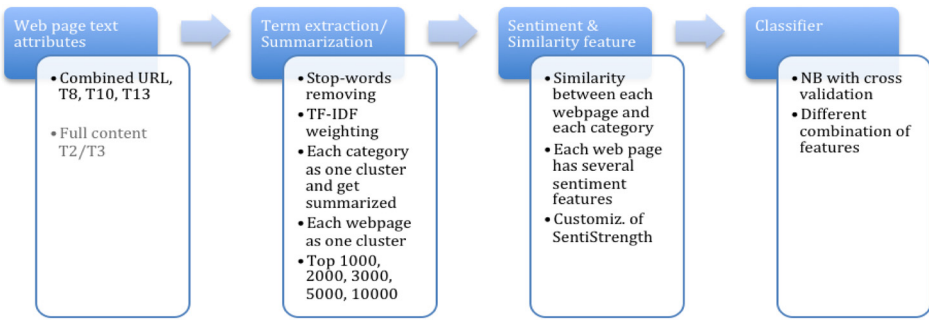

Figure 1: Web content classification based on topic and sentiment analysis.

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.

3.2 Sentiment Feature Extraction

Sentiment analysis is the process of automatic extraction and assessment of sentiment-related information from text. Sentiment analysis has been applied widely in extracting opinions from product reviews, discovering an affective dimension of the social web (Pang and Lee, 2008; Thelwall et al, 2010; Liu, 2012).

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. Performance of supervised methods and unsupervised methods vary depending on text types (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. It contains an EmotionLookupTable of 2,310 sentiment words and word stems taken from the Linguistic Inquiry and Word Count (LIWC) program (Pennebaker et al, 2003), the General Inquirer list of sentiment terms (Stone et al, 1966) and ad-hoc additions made during testing of the system. The SentiStrength algorithm has been tested on several social web data sets such as MySpace, Twitter, YouTube, Digg, Runners World, BBC Forums. It was found to be robust enough to be applied to a wide variety of social web contexts.

While most opinion mining algorithms attempt to identify the polarity of sentiment in text – positive, negative or neutral, SentiStrength gives sentiment measurement to both positive and negative directions with the strength of the 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.

As a starting point, we apply an unsupervised approach with the original SentiStrength system. 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. We tested with different selections of key topic terms: top 20%, 30%, 50%, and 100% (which represents the raw text after stop-word removing).

3.3 Page Vs. Category Topic Similarity

We use topic similarity to measure the content similarity between a web page and a web category. There are several different approaches for text similarity analysis: (1) Lexical/word based similarity analysis making use of hierarchical lexical resources such as WordNet – words are considered similar if they are synonyms, antonyms, used in the same way, used in the same context, or one is a type of another; (2) the vector space model and Cosine similarity analysis; (3) corpus based word semantic similarity analysis by SVD (Singular Value Decomposition) supported Latent Semantic Analysis methods (Landauer et al 1998; Laudauer and Dumais, 2008); (4) explicit semantic analysis: using external resources such as Wikipedia concepts to define vector space (Gabrilovich and Markovich 2007); (5) Language model based similarity measures; (6) graph based similarity analysis.

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.

4 DATA AND EXPERIMENTS

Our dataset is a collection of over 165,000 web pages in 20 categories, with each web page only labeled with one class (instead of multi-class). Each webpage is represented by a total of 31 attributes including full page, URL, Title and other meta-content, plus structural info and link information.

Taking into account the missing entries for different attributes, we selected a subset of the content features as the raw data for our study. In our first round of experiments (Section 4.1) we utilized full-page free text content. In our second and third round of experiments (Section 4.2 and Section 4.3), we used mainly the textual meta-content of web pages including URL words, title words, and several meta-description words (CobraMetaDescription, CobraMetaKeywords, TagTextA and TagTextMetaContent).

In the following, we report three sets of experiments and results of our study. In our first set of experiments, we focused on the three most problematic web categories i.e. Hate, Violence and Racism, and investigated classification models solely based on web pages' sentiment features. In our second set of experiments, we extend our modeling work into eight web categories (including Hate, Violence and Racism), developed classification models based on combined topic similarity and sentiment features of the web pages. In our third set of experiments, we expand our work further to include twenty web categories and built classification models based on combined topic similarity and sentiment features.

4.1 Sentiment based Classifier for Detecting Hate, Violence and Racism Web Content

To build classifiers for Hate, Violence and Racism web content, we sampled three datasets from the full database. The datasets contain training data with balanced positive and negative examples for the three web categories. Each dataset makes maximal use of the positive examples available. Negative samples are distributed evenly in the other 19 web categories.

Through some quick exploration, we found negative sentiment strength a better discriminator of web content than positive sentiment strength at least for the three web categories Hate, Violence and Racism. Thus, in our first set of experiments we only used negative sentiment strength value as data for learning and prediction. Here strong negative sentiment strength (e.g. -3, -4, -5) represents something with "bad sentiment" instead of "lack of sentiment". Lack of sentiment is when the sentiment strength is on the weak side (e.g. -1, -2).

Corresponding to the six sets of topic words for each web page, six sentiment features are obtained. Features for learning include a number of negative sentiment strength values of each web page, based on the different sets of topic terms (top 30, top 50, top 20%, 35%, 50% and 100%). Eventually we found that 30% cut-off seems to be a good level, as the average sentiment strength for each of the web categories starts to have larger variations and would not be at its best as a distinctive indicator for different web categories.

We built a classification model using the NäiveBayes (NB) method with cross validation, as three binary classifiers: c = 1, belongs to the category, (Violence, Hate, Jew-Racism), c = 0 (does not belong to the category). NB Classifier is a simple but highly effective text classification algorithm that has been shown to perform very well on language data. It uses the joint probabilities of features and categories to estimate the probabilities of categories given a document. Support Vector Machines (SVM) is another most commonly used algorithm in classification and foundation for building highly effective classifiers to achieve impressive accuracy in text classification. We experimented with both NB and SVM methods, and found that they achieved similar results, while SVM training takes much longer.

We tested with different combinations of the sentiment features. The best results show fairly good precision and recall levels for all three categories.

Table 1: Sentiment based NB classifiers.

	Model Performance	
Category	Precision	Recall
Hate	71.38%	77.16%
Jew-Racism	63.29%	72.79%
Violence	81.91%	73.92%

4.2 Combining Topic Similarity with Sentiment Analysis in Web Content Classification

Following our first set of experiments, we try to find ways to further improve the classification performance, while in the meantime extend our study from 3 to 8 web categories. Eight datasets were sampled from the full database, in the same way as described earlier. Each dataset contains training data with balanced positive and negative examples for a particular web category. In this round of experiments we made use of combined metadata of web pages instead of full page content as the raw data.

4.2.1 Topic Similarity based Classifier

To derive topic similarity features for each web page, we first summarize the data collection of each web category (as described in 3.1.2) and then calculate the cosine similarity between the topic vectors of each web page and each web category (as described in Section 3.3).

For each web category under study, we built NäiveBayes classification models (with cross validation) using only topic similarity features. It turns out that the results were very disappointing for most categories, low on both precision and recall measures. We thus conclude that our topic similarity

based classifiers solely do not perform well.

Next we seek to improve the classification performance through combined use of topic similarity and sentiment features. The results are very encouraging and the classification performance is significantly improved for most categories.

4.2.2 Extracting New Sentiment Features

For each web page, using the new raw data (meta content of web pages), we extract topic terms based on the tf-idf term weighting method and then the sentiment features using the original SentiSrength as well as the customized algorithms we defined. We tried different ways to customize the SentiStrength algorithm: (1) Counts of the amount of positives and negative sentiment words in a web page; (2) NewScale: sum of word SentiStrength value weighted by word frequency, normalized on total word counts, value between 5 and -5; (3) update the EmotionLookupTable. We found only few novel terms comparing with the original EmotionLookupTable, so we did not pursue it further as the effect would be minor. Figure 2 gives an overview of sentiment strength variations corresponding to changes of compression rate when extracting topic terms for each web collection.

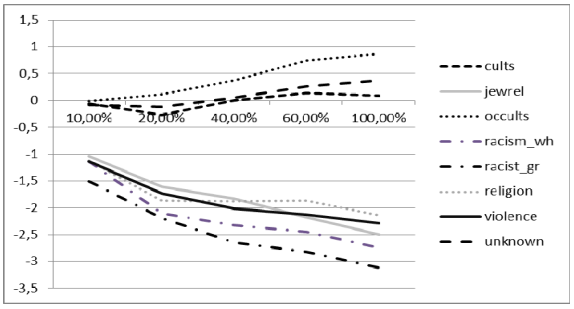

Figure 2: Sentiment strength value (NewScale) for each web category as compression rate changes.

Each category seems to keep their relative sentiment position rather consistently (with Violence being a bit different). We tested the new sentiment features by building NB classifiers for a few web categories. We found the classifiers to not necessarily perform better than the earlier sentiment based classifier. The performance varies from category to category, some slightly better, some not.

4.2.3 Classification using Combined Features

Next, we built NäiveBayes models (with cross validation) for eight web categories, using combined topic similarity features and sentiment features (as listed in Table 2). The results are very encouraging and the model performances are significantly improved for almost all categories when compared with solely sentiment based or solely topic similarity based classifiers, as is shown in Table 3.

Table 2: List of combined features for each web page.

Features of Web Page for Classification Models			
Page-Category topic similarity		Sentiment features	
Sim1 to Sim8	Topic similarity between a web page and web category #1 to web category #8	Pos1 to Pos5	Counts of +1 to +5 SentiStrength values
Note: In the experiments for Section 4.3, there will be 20 similarity features Sim1 to Sim20 as we work on all the 20 web categories.		Neg1 to Neg5	Counts of -1 to -5 SentiStrength values
		NS	New scale

Table 3: Classifiers making use of combined sentiment and topic similarity features (8 web categories).

Model Performance (combined features)		
Category	Precision	Recall
C1: Cults	75.80%	90.55%
C2: Occults	**87.08%**	**91.84%**
C3: Racism	**98.26%**	**96.30%**
C4: Racist	69.96%	91.82%
C5: JewRel	**64.43%**	**96.28%**
C6: Religion	67.01%	92.81%
C7: Violence	**93.69%**	**82.75%**
C8: Unknown	**89.59%**	**93.31%**

The best performing models are for Category C3 (RacWh) and C7 (Violence), for which both of the precision levels are over 93% and the recall levels are also very good. For web category C5 (JewRel) and C8 (Unknown), the models achieve very good recall levels (over 96%), but precision is lower. The C2 model performance also reaches a good level in terms of a good balance of precision and recall.

4.3 Extending to All Web Categories

Finally, in trying to gather more understanding about the effect of combined topic and sentiment analysis on classification performance, we extend the application of our approach to the other 12 web categories as well. The results are shown in Table 4.

From the table we can see that, the classification

models perform on a relatively good level only on two web categories: C9 (Adult) and C15 (Marijuna), for which the performance level is comparable to the previous 8 web categories. The recall levels for all the web categories are quite good, with category C13 and C20 models achieving the highest recall level of about 96%. However, the precision levels are in general much lower than the previous 8 web categories. This is interesting, and it seems to confirm that our approach works better on detecting negative web content than more neutral content.

Table 4: Classifiers making use of combined features (additional 12 web categories).

Model Performance (combined features) P: Precision R: Recall					
Category	P	R	Category	P	R
C9 (AD)	**78.69%**	**80.24%**	**C15 (M.**	**74.90%**	**87.97%**
C10	58.25%	93.48%	C16	54.75%	87.56%
C11	57.42%	94.81%	C17	62.81%	90.21%
C12	62.96%	86.14%	C18	53.57%	93.23%
C13	**57.38%**	**96.43%**	C19	57.61%	93.24%
C14	64.42%	90.65%	**C20**	**60.85%**	**95.85%**

5 CONCLUSIONS

In this research we set out to develop sentiment-aware web content detection system. We proposed a new approach for automatically classifying web pages into pre-defined topic categories. Word weighting, text summarization and sentiment analysis techniques are applied to extract topic and sentiment indicators of web pages. NäiveBayes classifiers are developed based on the extracted topic similarity and sentiment features. A large amount of experiments were carried out.

Overall, our experiment results are very encouraging and suggest that incorporating the sentiment dimension can indeed bring much added value to web content classification. Topic similarity based classifiers solely do not perform well, but when topic similarity and sentiment features are combined, the classification model performance is significantly improved for many web categories under study. Our results offer valuable insights and inputs to the development of web detection systems and Internet safety solutions.

The main contributions of this paper are: (1) investigation of a new approach for web content classification to serve online safety applications, through an integrated use of term weighting, text summarization and sentiment analysis methods; (2) customization of SentiStrength algorithm; (3) large amount of feature extraction and model developing experiments contributes to better understanding of text summarization, sentiment analysis methods, and the learning models; (4) several sets of analytical results that directly benefit the development of cyber safety solutions.

Our future work would include the incorporation of n-grams and multi-granularity topics, probabilistic topic models (Blei et al, 2003; Blei, 2012; Lu et al, 2011b), revisiting topic-aware sentiment lexicons (Lu et al, 2011a), word ontology, structural features, and fine-tuning the models with different learning methods. We will also look into new topic similarity measures and refining language processing during topic extraction. We believe there is still much room for improvement and some of these methods will hopefully help to enhance the classification performance to a new level. Our goal will be on improving precision and reducing false positives.

ACKNOWLEDGEMENTS

This research is supported by the Tekes funded DIGILE D2I research program, Arcada Research Foundation, and our industry partner. We thank the reviewers for their helpful comments.

REFERENCES

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

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

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.

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.

Gabrilovich, E., and Markovich, S. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07),* Hyderabad, India.

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.

Kludas, J. 2007. Multimedia retrieval and classification for web content, Proc. of the 1st BCS IRSG conference on Future Directions in Information Access, British Computer Society Swinton, UK ©2007.

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.

Landauer, T. K., Foltz, P. W., and Laham, D. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes* (25), 259-284.

Landauer, T. K., and Dumais, S. T. 2008. Latent semantic analysis. *Scholarpedia 3(11*): 4356.

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

Lu, Y., M. Castellanos, U. Dayal, C. Zhai. 2011a. "Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach", *In Proceedings of the 20th international conference on World wide web* (WWW'2011) Pages: 347-356.

Lu, Y., Q. Mei, C. Zhai. 2011b. "Investigating Task Performance of Probabilistic Topic Models - An Empirical Study of PLSA and LDA", Information Retrieval, April 2011, Volume 14, Issue 2, pp 178-203.

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

Pennebaker, J., Mehl, M., & Niederhoffer, K. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology, 54(1),* 547–577.

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).

Radev, D., Jing, H., Styś, M., and Tam, D. 2004b. Centroid-based summarization of multiple documents. *Information Processing and Management* (40) 919–938.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management,* 24(5), 513-523.

Shen, D., Z. Chen, Q. Yang, H. Zeng, B. Zhang, Y. Lu, W. Ma: Web-page classification through summarization. SIGIR 2004: 242-249.

Shen, D., Qiang Yang, Zheng Chen: Noise reduction through summarization for Web-page classification. Information Processing and Management 43(6): 1735-1747 (2007).

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. 1966. *The general inquirer: a computer approach to content analysis.* The MIT Press, Cambridge, Massachusetts, 1966. 651.

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.

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.

Zhang, S., Xiaoming Jin, Dou Shen, Bin Cao, Xuetao Ding, Xiaochen Zhang: Short text classification by detecting information path. CIKM 2013: 727-732.