Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification

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: This research concerns the development of web content detection systems that will be able to automatically classify any web page into pre-defined content categories. Our work is motivated by practical experience and observations that certain categories 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 account. To further improve the performance of detection systems, we bring web sentiment features into classification models. In addition, we incorporate n-gram representation into our classification approach, based on the assumption that n-grams can capture more local context information in text, and thus could help to enhance topic similarity analysis. Different from most studies that only consider presence or frequency count of n-grams in their applications, we make use of tf-idf weighted n-grams in building the content classification models. Our result shows that unigram based models, even though a much simpler approach, show their unique value and effectiveness in web content classification. Higher order n-gram based approaches, especially 5-gram based models that combine topic similarity features with sentiment features, bring significant improvement in precision levels for the Violence and two Racism related web categories.

1 INTRODUCTION

This study concerns the development of web content detection systems that will be able to automatically classify any web page into pre-defined content categories. Previous experience and observations with web detection systems in practice has shown that certain groups of web pages such as those carry hate and violence content prove to be 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 web pages often carry strong negative sentiment while their topics may vary a lot. 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 recent work we have developed topic similarity and sentiment analysis based classification models (Liu and Forss, 2014), which bring encouraging results and suggest 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. In addition, our results also highlight the effectiveness of integrating topic similarity and sentiment features in web content classifiers.

Meanwhile, we observed from our earlier experiments that topic similarity based classifiers alone perform rather poorly and worse than expected. To further improve the performance of the classification models, in this study we develop new models by incorporating n-gram representations into our classification approach. Our assumption is that n-grams can capture more local context information in text, thus could help to enhance topic similarity analysis. N-grams are commonly applied in many text mining tasks. However, their effects are uncertain and very much depending on the nature of the text and the purpose of the task. It is our goal to investigate the effects of adopting higher order n-grams in web content classification. Unlike most

Liu S. and Forss T..
Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification.
DOI: 10.5220/0005170305300537
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (SSTM-2014), pages 530-537
ISBN: 978-989-758-048-2

applications that only consider presence (Boolean features) or frequency counts of n-grams, in this study we will explore the use of tf-idf weighted n-grams in topic analysis of web pages and web categories. We then build our content classification models by integrating topic similarity analysis with sentiment analysis of web pages.

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). The earliest studies on web classification appeared already 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 to 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 much more efficiency.

There exists 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 a sentence-based summarization method 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 sentence-based approach.

In recent years, there have been many studies of 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 classification of short text using information paths to deal with the less informative word co-occurrences and sparseness of such texts. Their method makes use of ordered subsets in short texts, which is termed “information path”. They found classification based on each subset resulted in higher overall accuracy than classifying the entire data set directly.

Related to 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 can thus monitor the traffic emanating from the monitored group of users, issue an alarm if the access information is not within the typical interests of the group, and track 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 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.

Fürnkranz (1998) and Fürnkranz et al (1999) are the earliest studies on n-grams in text classification. They studied the effect of using n-grams and linguistic phrases for text categorization. They found that bigram and trigrams were most useful when applied to a 20 newsgroups data set and 21,578 REUTERS newswire articles. Longer sequences were found to reduce classification performance. Fürnkranz et al (1999) and Riloff et al (2001) then revealed that linguistic phrase features can help improve the precision of learned text classification models at the expense of coverage.

The rest of the paper is organized as follows. In Section 2, we describe our approach for web content classification and explain the methods and techniques used in topic extraction, topic similarity analysis and sentiment analysis. In Section 3 we describe our data and experiments for the classification of Hate, Violence and Racist web content. We compare the performance of different models based on unigram, trigram and 5-grams. Section 4 concludes the paper.

2 COMBINING N-GRAM BASED CONTENT SIMILARITY ANALYSIS WITH SENTIMENT ANALYSIS IN WEB CONTENT CLASSIFICATION

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.

2.1 Content Representation and Topic Extraction

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 the same categories. The extracted topics hopefully give a good representation of the core content of a web page or a web category.

We use the tf-idf weighted vector space model of n-grams (where n=1, 3, 5) to represent the original text content and extracted topics of web pages and web categories. When n=1, it is a feature vector that contains one weight attribute (instead of Boolean or simple frequency count) for each unique term that occurs in a web page or collection and their topics. In other words, each web page or collection or topic is represented by the set of unique words it consists of. Similarly, when n>1, each web page or collection or topic is represented by the set of unique n-grams it contains. We planned an experiment testing the possibilities with n= 1 to 6. However, due to time constraints we have to scale down the number of experiments, and choose to test only unigram, trigram and 5-grams.

In pre-processing, we apply stemming and stop-words removing for obtaining unigrams, but no stop-words removing when obtaining higher order n-grams. However, we remove n-grams with stop words as the beginning or ending. We build our own IDF databases using the entire data collection of over 165,000 web pages of 20 categories. The tf-idf weight of an n-gram is adjusted using the weight of unigrams it contains, i.e. add to the tf-idf value of the n-gram the tf-idf value of the words it contains. This is done independently for unigrams, trigrams and 5-grams.

For each webpage, we make use of its multiple content attributes as raw data input for the term/

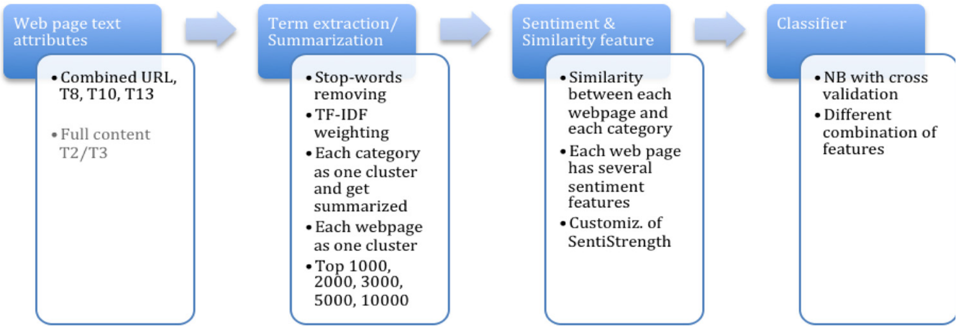
Web page text attributes
• Combined URL, T8, T10, T13
• Full content T2/T3
Term extraction/ Summarization
• Stop-words removing
• TF-IDF weighting
• Each category as one cluster and get summarized
• Each webpage as one cluster
• Top 1000, 2000, 3000, 5000, 10000
Sentiment & Similarity feature
• Similarity between each webpage and each category
• Each web page has several sentiment features
• Customiz. of SentiStrength
Classifier
• NB with cross validation
• Different combination of features

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

n-gram weighting process. The content attributes include full-page content as well as URL and three other meta-contents (T8, T10, T13).

The topic of a web page is obtained simply based on tf-idf n-gram weighting. For each webpage, by applying different compression rates, we can obtain different sets of topic words (for example top 10%, 20%, 50%, 100%). The topic of a web category is obtained through summarization of all the web pages in the same category. For each web page collection, we apply the Centroid method of the MEAD summarization tool (Radev et al, 2003; 2004) to make summaries of the document collection. Through this we try to extract topics that are a good representation of a specific web category.

Based on our earlier experiments, we set a cut-off level at top 15000 highest weighted unigrams, top 100,000 highest weighted trigrams, and top 120,000 highest weighted 5-grams to represent the topic content of a web page or web category.

We should point out that there exists a large body of literature on topic extraction. For example, LDA based topic models are another popular means for topic detection, especially deriving hidden topics in large text collections. It is our intention to develop and test such topic models in web content classification soon, but it is out of the scope of this article.

2.2 Web-Page vs. Web-Category Topic Similarity

We use topic similarity to measure the content similarity between a web page and a web category. Topic similarity is implemented as the cosine similarity between topic terms of a web page and topic terms of each web category. For each web page, we compute its cosine similarity to each web category in different n-grams vector space. We compare results from using only unigrams and unigram-weight adjusted n-grams.

2.3 Extracting Sentiment Features

Based on the extracted topic content for each web page, we make assessments of its sentiment strength through using the SentiStrength (Thelwall et al, 2011, 2012) sentiment analysis tool.

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 (Thelwall et al, 2011). SentiStrength 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 a EmotionLookupTable of 2310 sentiment words and wordstems taken from 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, Runner's 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 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.

Sentiment features are extracted by using the key topic terms extracted from the topic extraction process as input to the SentiStrength. This gives sentiment strength values for each web page in the range of -5 to +5, with -5 indicating strong negative sentiment and +5 indicating strong positive sentiment. We found that negative sentiment strength value was a better discriminator of web content than positive sentiment strength value at least for the three web categories Hate, Violence and Racism. Thus, in our first set of experiments we only use negative sentiment strength value as data for learning and prediction. Corresponding to the six sets of topic words for each web page, six sentiment features are obtained.

In sentiment analysis, we only apply topic content in the form of unigrams (stemmed, with stop-words removing). In addition to applying the original SentiStrength tool, we also tried to customize the SentiStrength algorithm in two ways: (1) Counts of positives and negative sentiment words in a web page; (2) Sum of word sentiment value weighted by word frequency, normalized on total word counts, value between -5 and 5.

3 CLASSIFICATION MODELS FOR THE DETECTION OF HATE, VIOLENCE AND RACISM WEB PAGES

3.1 Data and Experiments

Our dataset is a collection of over 165,000 single labeled web pages in 20 categories. As described earlier, in our study we selected a subset of the content features as the raw data, taking into account missing entries for different attributes. More specifically we utilized full-page free text content, in combination with the textual meta-content of web pages including URL words, title words (TextTitle) and meta-description terms (CobraMetaDescription, CobraMetaKeywords, TagTextA and TagTextMeta Content).

To build classifiers for identifying violence, hate and racism web pages, four datasets are sampled from the entire database. The datasets contain training data with balanced positive and negative examples for the four web categories: Violence, Racism, Racist and Hate. Each dataset makes maximal use of positive examples available, with negative samples distributed evenly in the other 19 web categories.

Features for learning in the data for each web page include topic similarity to ten web categories (including the 4-selected categories) and a number of sentiment strength values of each web page. A summary of the features is given in Table 1.

In a series of experiments we develop three types of classification models and compare their performance: topic similarity based, sentiment based, and the combined models over unigrams, trigrams and 5-grams. We choose to apply NäiveBayes (NB) method with cross validation to build binary classifiers: c = 1, belong to the category, (Violence, Hate, Racism, Racist), c = 0 (not belong to the category). NB Classifier is simple but has been shown to perform very well on language data. Support Vector Machines (SVM) as another of the most commonly used algorithms in classification often achieves similar results while training takes much longer time.

Table 1: List of extracted features for web pages.

Page-Category topic similarity	Sentiment strength features
Sim1- Sim10: Topic similarity between a web page and web category #1 to #10	Pos3-Pos5 and Neg3-Neg5 (Counts of SentiStrength values as +3, +4, +5 and -3, -4, -5)
	NewScale1
	NewScale2

3.2 Results and Discussion

The results of our experiment with unigram, trigram and 5-gram based models are summarized in Table 2, Table 3 and Table 4.

3.2.1 Unigram based Models

Among the unigram models, topic similarity based models perform surprisingly well when compared to our earlier studies, especially with the three web categories Hate, Racism and Racist. One reason could be that the raw data input had a big effect here. Another could be that stemming and a customized IDF database helped very much in content similarity analysis.

The sentiment based models alone, on the other hand, did not perform as well as the similarity based models, also a bit lower level than results from our previous experiments. The reason lies mainly in the differences in the negative samples of the training set. We could still try to improve on the sentiment-based models by looking more into the sentiment features.

In all four web-categories, we were able to develop combined classification models with rather decent performance.

3.2.2 Trigram based Models

When compared to using unigrams, in the case of trigram models, the topic similarity based method did not gain on precision level but rather on recall level. This is counter-intuitive. One reason could be that our selection of cut-off level for category centroid (the dimension of trigrams vectors) is not suitable, which would have limited the performance on similarity-based models. We will try to improve the results by enlarging the vector space, which will have an effect on the precision of the classifiers.

When topic similarity and sentiment features are combined, we are able to build classifiers that are slightly better than in the case of unigrams, especially for the Racist category, which has increased precision level by a big margin. This is interesting and seems to tell us that there is something genre-specific about the Racist category.

3.2.3 5-gram based Models

Comparing with the unigram and trigram approaches, 5-gram based combined models show significantly improved precision levels for the Violence and two Racism related web categories. However, when comparing the trigram and 5-grams results, it seems the effect of high n-grams on topic similarity based models is indeed minor. We need to look further into this to understand if the large amount of computation needed in processing higher-order n-gram adds much to improve the classification performance.

Table 2: Unigram: similarity based classifiers vs sentiment based vs combined.

Unigram based classification models						
Web Category	Topic similarity based		Sentiment based		Best combined	
	Precision	Recall	Precision	Recall	Precision	Recall
Violence	63.64%	83.17%	69.96%	53.87%	87.11%	60.16%
Hate	79.51%	75.97%	65.86%	65.49%	81.06%	76.25%
Racism	76.96%	77.35%	70.89%	58.83%	78.17%	79.77%
Racist	79.31%	80.50%	68.42%	62.22%	86.23%	88.41%

Table 3: Trigram: similarity based classifiers vs sentiment based vs combined.

Trigram based classification models						
Web Category	Topic similarity based		Sentiment based		Combined	
	Precision	Recall	Precision	Recall	Precision	Recall
Violence	59.66%	92.67%	69.96%	53.87%	84.57%	82.09%
Hate	61.73%	94.64%	65.86%	65.49%	64.07%	94.75%
Racism	65.85%	92.73%	70.89%	58.83%	78.63%	84.94%
Racist	67.67%	95.97%	68.42%	62.22%	95.85%	81.36%

Table 4: 5-gram: similarity based classifiers vs sentiment based vs combined.

5-gram based classification models						
Web Category	Topic similarity based		Sentiment based		Combined	
	Precision	Recall	Precision	Recall	Precision	Recall
Violence	60.71%	92.57%	69.96%	53.87%	**96.56%**	**70.52%**
Hate	62.05%	95.96%	65.86%	65.49%	67.56%	96.45%
Racism	60.87%	95.43%	70.89%	58.83%	74.34%	90.80%
Racist	63.77%	96.22%	68.42%	62.22%	**97.62%**	**82.62%**

Table 5: Combined models (only meta content as raw data, unigram, no stemming).

Model Performance (combined features)		
Category	Precision	Recall
Violence	93.69%	82.75%
Hate	64.43%	96.28%
Racism	69.96%	91.82%
Racist	98.26%	96.30%

Finally, comparing our results with results from an earlier study (Table 5), the 5-gram models making use of combined topic similarity and sentiment strength measurements outperform on precision levels for the Violence, Hate and Racism groups, but not on the other category (Racist). a much simpler approach we applied earlier actually shows its value as well on the detection of the Racist category content.

4 CONCLUSIONS

In this study we investigated the application of n-gram representation in web content classification models. Our assumption is that n-grams can capture more local context information in text, and thus could help to improve accuracy in capturing content similarity, which will subsequently help further improving the performance of the classification models. N-gram 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 number of experiments were carried out.

Our results reveal that unigram based models, although a much simpler approach, show their unique value and effectiveness in web content classification. Raw data input, stemming, IDF database, all play important roles in determining topic similarity, just like the choice of representation model as uni-gram or higher order n-grams.

Higher order n-gram based approach, especially 5-gram based models in our study, when combined with sentiment features, bring significant improvement in precision levels for the Violence and two Racism related web categories. However, the effect of high n-grams on topic similarity based models seems to be really minor. We need to look into this further to understand if the improvements made in classification models justify the large amount of computation needed in processing n-gram.

The main contributions of our paper are: (1) Investigation of a new approach for web content classification to serve online safety applications; (2) Contrary to most studiesn which only consider presence or frequency count of n-grams in their applications, we make use of tf-idf weighted n-grams in building the content classification models. (3) A large amount of feature extraction and model developing experiments contributes to a better understanding of text summarization, sentiment analysis methods, and learning models; (4) Analytical results that directly benefit the development of cyber safety solutions.

In our future work we will explore the incorporation of probabilistic topic models (Blei et al, 2003; Blei, 2012; Lu et al, 2011b), revisit topic-aware sentiment lexicons (Lu et al, 2011a), and fine-tuning the models with different learning methods. We believe there is still much room for improvements and some of these methods will hopefully help to enhance the classification performance to a new level.

ACKNOWLEDGEMENTS

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

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. 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), 244-255.

Fürnkranz J, Exploiting structural information for text classification on the WWW, Advances in Intelligent Data Analysis, 487-497, 1999

Fürnkranz J., T. Mitchell and E. Riloff, A Case Study in Using Linguistic Phrases for Text Categorization on the WWW, Working Notes of the 1998 AAAI/ICML Workshop on Learning for Text Categorization.

Fürnkranz J, A study using n-gram features for text categorization, Austrian Research Institute for Artifical Intelligence 3 (1998), 1-10

Fürnkranz J, T Mitchell, E Riloff, A case study in using linguistic phrases for text categorization on the WWW, Proceedings from the AAAI/ICML Workshop on Learning for Text Categorization, 5-12, 1999

Gabrilovich, E., and Markovich, S. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *In 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.

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

Liu S. and T. Forss, “Web Content Classification based on Topic and Sentiment Analysis of Text”, accepted by KDIR 2014, Rome, Italy, October 2014

Lu, Y., M. Castellanos, U. Dayal, C. Zhai. 2011a. "Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach", *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 Process. and Management* (40) 919–938.

Riloff E, J Fürnkranz, T Mitchell, A Case Study in Using Linguistic Phrases for Text Categorization on the WWW, AAAI/ICML Workshop on Learning for Text Categorization, 2001

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. Info. Process. and Manage. 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 Sci. 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.

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