Semantically Enriched Machine Learning Approach to Filter YouTube Comments for Socially Augmented User Models

Ahmad Ammari , Vania Dimitrova , Dimoklis Despotakis

School of Computing, University of Leeds
{A.Ammari, V.G.Dimitrova, scdd} @leeds.ac.uk

Abstract. Social media are media for social interaction that allow creating and exchanging user-generated content. The massive social content can provide rich resources for deriving social profiles that can augment user models and improve adaptation in traditional applications. However, potentially valuable social contributions can be buried within highly *noisy* content that is irrelevant or spam. This paper sketches a research roadmap toward augmenting user models with key user characteristics derived from social content. It then focuses on the first step: identifying and filtering noisy content to create data corpus about a specific activity. A novel, semantically enriched machine learning approach to filter the noisy content from social media is described. This is applied to a specific social source and activity: public comments on YouTube job interview videos. A potential application of the approach to augment user models in simulated experiential learning environments is discussed.

1. Introduction

The Social Web, or Social Media, includes a range of public data sources that are becoming an inevitable part in our life. Since their introduction, social media sharing sites such as YouTube[1], Flickr[2], and delicious[3] have attracted millions of users, many of whom have integrated these sites into their daily practices. An inspection of the social video sharing platform YouTube reveals a high amount of community feedback through user comments on the published videos. These comments often include 'authentic stories' of people's experiences of a particular activity. Pre-processing and mining these comments could provide a highly rich resource of real world activity descriptions based on individual's and societies' cognitive and social states, such as interests, knowledge, and experiences within that activity domain [14]. These identified features can be further mined to discover correlations between them that could then be used to augment existing and limited user models used to adapt many applications.

However, an important research challenge is how viable it is to extract the relevant content from within the huge amount of social media data that is likely to contain *noisy* content (*i.e. content irrelevant to the activity of interest*). The broad objective of our research is to evaluate whether social media content that is relevant to an activity

[1] http://www.youtube.com/
[2] http://www.flickr.com/
[3] http://www.delicious.com/

of interest can be identified, mined, and used as an efficient source to augment user models used to adapt simulated learning environments.

The rest of the paper will include the following: In Section 2, we present a research roadmap towards achieving our broad objective. In Section 3, we describe a novel methodology to filter the noisy content from the social media data that we use to achieve our objective, which is the user comments on videos found on YouTube that describe a particular activity of interest. In Section 4, we position our work in the relevant literature on finding good quality content on the social Web by filtering the noisy content. In Section 5, we present and discuss the experimental results of our preliminary implementation. Finally, in Section 6, we discuss various considerations for subsequent implementations.

2. Socially Augmented User Models: Research Roadmap

Existing simulated learning environments suffer from the limited understanding of the learner because they are disconnected from the learners' real job experiences. This often hinders learners' engagement and motivation to undertake training since the skills developed in the simulated learning environment are not effectively connected to the skills used in the real job practice. Augmented User Modelling; *i.e. enriching existing user models with additional information mined from other data sources not considered previously*, is perceived as an approach to effectively help in aligning the learning experience in the simulated environments with the real world context and the day-to-day job practice. The key advantage is that the user models become aware of a range of aspects that cannot be captured from merely analysing the user interaction with the learning application.

Toward achieving the user model augmentation, we introduce a research roadmap, describing the research phases and the key research challenges that will be addressed.

Phase 1: Identifying social media content that represents real world user experiences. The key research challenge in this phase is how to filter the noise from the data sets retrieved from a given social media data source. By noise we mean those instances in the data sets that are *highly irrelevant to a particular activity domain*, thus not valuable for deriving significant features that can be used to augment existing user models with real world learning experiences.

Phase 2: Deriving key user characteristics from the *clean* relevant social content identified in phase 1. The key research challenge in this phase is how to derive *social user profiles* from the identified relevant content.

Phase 3: Using the social user profiles derived in phase 2 to augment an existing limited user model used to adapt a simulated learning environment. The key challenge in this phase is how to align the user in the existing user model with the social user profiles derived from the relevant social media content.

This paper focuses on the **first phase** in the roadmap. It presents a novel approach to filter the noise identified in the social media data. This hybrid approach combines machine learning, data mining, and semantics to address the challenge of this phase, which is the extraction of social media content that is highly relevant to a given real world activity of interest. The problem is narrowed down by considering a **specific activity** that is being practiced in the simulated environments. We use **Job Interviews** as the **target activity**, which is represented by videos selected from the social video

sharing site YouTube. The user comments found on these videos represent the corpus that will be processed by the approach to reduce the noisy content by filtering out those comments that are irrelevant to the particular activity domain of interest.

3. The Social Noise Filtering Approach

3.1 Filtering Noisy YouTube Comments: Methodology

In order to achieve a significant improvement in the relevance degree of the YouTube comments that are sufficiently good to derive key user characteristics for user model augmentation, we present a semantically enriched machine learning noise filtering approach. Figure 1 shows a flowchart representing the methodology for the approach.

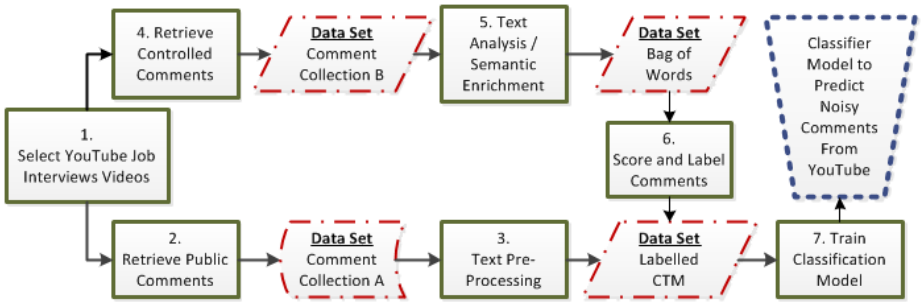


Figure 1. Filtering Noisy Comments: Methodology

Step 1. Select **video corpus** from YouTube about job Interviews. This was conducted as part of a research study to extract individual viewpoints from user comments in social spaces [4]. To illustrate the job interview activity, videos published on YouTube were selected as content source, and a thorough search and classification of different video types was performed. In particular, four different category types were identified to classify each retrieved video including: guides (explanations of best practices), interviewees' stories, interviewers' stories and interview mock examples. It was decided to focus on examples, as these resources can be closely connected to real world context representing the activity.

Step 2. For each selected video, retrieve the public comments on the video from YouTube. We call this **Comment Collection A**. Because this collection is retrieved from a very crowded and open social media sharing site, it contains a considerable rate of *noisy* comments. By noisy, we mean those comments whose text content is highly irrelevant (*e.g* spam, abuse, etc) to the activity illustrated by the videos.

Step 3. Pre-process the Comment Collection A to build a **Comment-Term Matrix** (CTM) to train a supervised classification model. The goal is to represent each comment in the collection by a comment term vector. The pre-processing step is described in Section (3.2).

Step 4. Use the experimentally-controlled, relatively *clean* collection of YouTube comments collected and analyzed by the research study described in [4]. By clean, we mean comments whose text content is highly relevant to the job interview activity. We call this **Comment Collection B**.

Step 5. Analyze the Comment Collection B to build a **semantically enriched Bag of Words** (BoW). The resulted BoW forms a *ground truth vocabulary* that is highly relevant to the job interview activity domain. The selection and pre-processing of this comment collection are further described in Section (3.3).

Step 6. For each comment in Comment Collection A, **compute a relevance score** for the comment. Using the scores of the comments, label a new class attribute, i.e. a binary class attribute, with the distinct values: *relevant*, *noisy*, to supervise the learning of the classification model. This is further described in Section (3.4).

Step 7. Using the labelled Comment-Term matrix, train a **supervised classification model** that will learn the underlying classification rules to predict the *relevance*, i.e. *relevant, noisy*, of each new comment retrieved from the same data source, i.e. YouTube in the current case study, thus filter out those noisy comments deriving little-to-no key user characteristics for social user profiling.

3.2 Pre-Processing the YouTube Comments

Pre-processing the Comment Collection A is necessary to transform the textual corpus into a Comment Term Matrix (CTM) to be used as input data set to train classification models. A thorough description of the text pre-processing techniques to build Document – Term matrices to train machine learning models is found in [5]. The pre-processing steps to build the CTM are summarized in the following steps:

1. Remove all non-content bearing *stop words* like “a”, “an”, “the”, etc, which should not contribute to neither the representation of the comment nor to the scoring mechanism of each comment. A standard stop word list by Google[4] has been used by this study.
2. Stem the words to retain the roots and discard common endings. The Iterated Lovins Stemmer [13] has been used widely for stemming unstructured data for machine learning and is therefore used by this study.
3. Rank the words based on their *tfidf* scores [1]. The *tfidf* score consists of two parts: term frequency *tf*, and inverse document frequency *idf*. A *tfidf* score is normalized between “0” and “1”.
4. Represent each comment by a Comment Term Vector, forming a Comment Term Matrix (CTM) representation of the Comment Collection. Each row in the matrix is a comment and each column represents a term and the value is the term *tfidf* score for that particular comment.

3.3 Building the Semantically Enriched Vocabulary

A clean, semantically enriched vocabulary / Bag of Words (BoW) that well represent the context of the job interview activity domain is needed to score each comment in Collection A. For this, we parse part of the corpus of study described in [4]. In that study, the selected YouTube videos were used in a system developed within the research context, and a research study is being conducted to collect video comments from participating users. The usage scenario for each participant includes: watching

[4] http://www.ranks.nl/resources/stopwords.html

the video; identifying useful video snippets; writing free text comments for each snippet indicating whether the comment corresponds to the activity presented in the video or a personal experience/opinion, and whether the comment concerns the interviewer or the interviewee. These comments provide examples of good (focused) corpus collected in experimental settings.

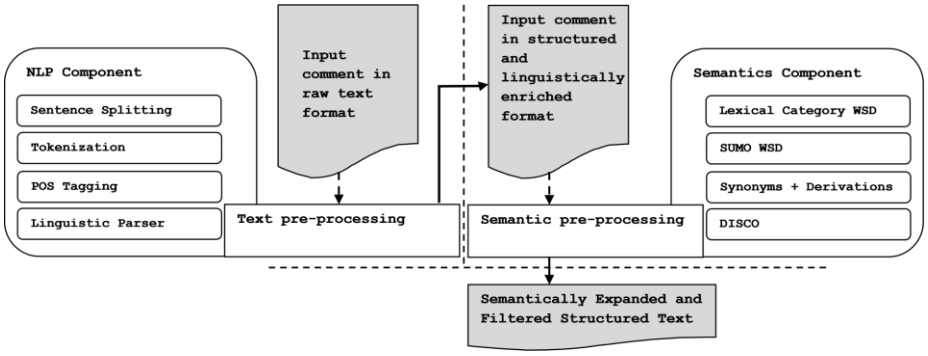


Figure 2. Comment Analysis and Semantic Enrichment of Collection B

Figure 2 illustrates the corpus analysis phase. Each comment was handled as a separate document. The first step includes NLP techniques for text analysis using the Antelope NLP framework[5], i.e. sentence splitting, tokenization, Part of Speech tagging and syntactic parsing using the Stanford parser for linguistic analysis. This enables the extraction of a structured form text representation to empower further analysis using semantics. The second step consists of the semantic analysis layer, representing Ontology based word sense disambiguation and linguistic semantic text expansion. The first filter applied concerns the selection of specific lexical categories implemented within the WordNet Lexicon English language thesaurus[6] to directly exclude non-significant terms for the job interview activity. For the words remained, the Suggested Upper Merged Ontology (SUMO) [3] has been exploited, which provides direct mappings of WordNet English word units to concepts in the ontology. The resulted concepts were used as word sense disambiguation indicators (second filter). In this context, WordNet Lexicon queries were performed to retrieve synonyms, antonyms and word lexical derivations to expand the word set. Furthermore, DISCO [8] has been exploited to retrieve distributionally similar words from the Wikipedia corpus, and the filters discussed above have been applied, i.e. lexical category and SUMO concept mapping.

3.4 Computing the Relevance Scores and Labelling the Comments

We present a mathematical model, using the Comment Collection A and the derived BoW in Section (3.3), to compute a numerical score for each public comment in

[5] www.proxem.com/Default.aspx?tabid=119

[6] http://wordnet.princeton.edu/

collection A, which represents the relevance of the comment to the job interview activity domain. Let C be the set of all n comments in the YouTube public comment collection A. For each comment $c_x \in \{ c_1, c_{2, \dots}, c_n \}$, there is a set w_{cx} of unique tokenized and stemmed m non-stopwords, where m is the number of these words in comment c_x. Let B be the set of all the stemmed and unique words in the BoW derived in Section 4.3. We then define a ***relevance score*** S_{cx} for the comment c_x to be:

$$S_{c_x} = \frac{\left|w_{c_x} \cap B\right|}{\left(\sum_{k=1}^{n}\left|w_{c_k} \cap B\right|\right) / n}$$

where $\left|w_{c_x} \cap B\right|$ is the number of words that exist in the intersection between the sets w_{cx} and B, and the denominator is the average number of words that exist in the intersections between each set w_{ck} and B, where $k \in \{1, 2, \dots, n\}$.

In order to train a binary classification model, we define a target class attribute ,$CLASS_{cx}$, which contains a nominal value $\in \{$ *noisy (0)*, *relevant (1)* $\}$, based on the value of the score S_{cx} for the comment c_x:

$$CLASS_{C_X} = \begin{cases} noisy\ (0) & if \quad S_{c_x} < 1.00 \\ relevant\ (1) & if \quad S_{c_x} \geq 1.00 \end{cases}$$

The class value for each comment is then assigned as the target class attribute value to the term vector representation of the comment, forming a supervised training corpus for building machine learning classification models that learn the underlying classification rules to predict the class value of new comments.

Table 1. Example Noisy and Relevant Comments with their Computed Scores

Comment Labelled Noisy	Score	Comment Labelled Relevant	Score
what if you never had a job	0.34	To be honest, I probably wouldn't hire either one of them. The girl is obvious, but the guy's leg twitching bothered me, as did his leaning forward in the chair, and he focused too much on his past. I want to hear what he's going to do with the job available, not so much what he has done.	5.08
LOL	0.0		
Interview on wednesday hope it goes well	0.68	that part when she answers her phone was just retarded, AHHHHHH! someone's calling me! the person giving the interview must think she's psychopathic	1.13
come see my job interview come see my job interview come see my job interview called Boss Boss Baby Boss Boss Baby	0.79		

To give a sense of the reasonability of the scores and labels assigned to the comments based on our model, table 1 shows four example comments on the left that have been labelled as noisy by the scoring model. Obviously, the first three ones do not comment on the job interview video being watched, whereas the fourth one is a spam. The scoring mechanism was reasonable in labelling them as noise even while containing a considerable number of words, i.e. comment 4. The two comments on the right clearly describe actions occurring within the activity watched in the video, thus potentially can derive user characteristics related to the activity. Again, it was reasonable labelling them as relevant.

4. Related Work

There have been a few attempts in the literature to create information filtering mechanisms for adaptation in the social Web, which can be linked to the research challenge addressed in our study. For example, the work in [11] presents CompleXys, a system that accesses a variety of social data sources, including social networks and blogs, and semantically annotates and categorizes the retrieved content based on a filtering layer and displays only the relevant content to the user. The filtering layer takes the output of a content annotator component that annotates the retrieved content using a domain ontology. The expanded taxonomy is then meant to decide whether a given resource is relevant to the list of topics stored in the filtering layer. The frequency of occurring annotations can then be used as a simple indicator for the relevance of a certain topic. We have further expanded this mechanism by introducing the mathematical model in Section (3.4), which computes a relevance score for each retrieved content observation, i.e. YouTube video comment, and then labels the observation, i.e. relevant or noisy, accordingly.

Works on filtering spam blogs (or *splogs*) [15] as well as filtering blog spam comments [6] could also be linked to this study. In [15], blogs and their connections are represented as a graph and then various graph statistics, i.e. degree distribution, clustering coefficient, are computed. It is shown that these statistics are considerably different between splogs and legitimate blogs, and therefore could be leveraged to identify splogs. The work in [6] presents a similar approach to identifying spam comments irrelevant to the discussion by generating a blogger network based on the blogger's commenting behaviour. However, social comments in general contain no (or very little) hyperlinks between them. This leads to a highly sparse adjacency matrix with very few non-zero values that represent the link strength between the comments [1]. Computing content-based similarities between the comments could be used to fill the matrix in addition to the direct links to reduce sparsity. However, since comments usually do not contain much text, content-based estimation of the comment linkage is not a good alternative and the underlying noise filtering approach is likely to perform poorly in noisy comments identification.

Few works have used machine learning to find quality contents from the user comments on the social space. The work in [2] used binary classification models to automatically identify high quality content in a large community-driven question/answering portal; Yahoo! Answers. We further extend this work by introducing semantic enrichment in Section (3.3) in order to classify the data set used for training the binary classification models. The work in [12] used a supervised

classification approach to analyze a corpus of YouTube comments in order to discover correlations between the user views and sentiments extracted from these comments, and the comment ratings by the readers of these comments. Such correlations may help to automatically structure and filter comments for users who show malicious behaviour such as spammers and trolls. However, relying on a comment rating needs a huge corpus of these comments because just a small fraction of the comments on YouTube is rated by the YouTube community. This large size of corpus is not always available when addressing a particular domain activity. For example, a total of 17 high quality YouTube videos on the "Job Interview" activity selected for the work of this paper did not retrieve more than 1159 comments. Instead of relying on comment ratings, the approach presented in our work creates a semantically enriched taxonomy by analyzing a clean corpus of experimentally-controlled user comments and enriching this vocabulary with semantic annotations to form a *ground truth* Bag of Words (BoW) that is highly relevant to the activity domain of interest, i.e. job interview. The retrieved YouTube comment corpus is then *scored* and *labelled*, using the mathematical model and the semantically-enriched BoW, and then used to train a supervised classification model that predicts and filters out the noisy comments.

5. Experimental Results

A preliminary implementation to the approach has been done to evaluate the classification performance in filtering the noisy comments from the training / testing corpus. Table 2 shows a summary description of the two comment collections before and after being pre-processed.

Table 2. Data Description

Comment Collection A		**Comment Collection B**	
No of Videos	17	No of Videos	5
No of Comments	1159	No of Comments	193
Min Intersection Size with BoW Set	0.0	No of Original Words	6398
Max Intersection Size with BoW Set	48.0	No of Synonyms	25606
Avg Intersection Size with BoW	8.85	No of Antonyms	1978
Min Relevance Score	0.0	No of Derivations	17604
Max Relevance Score	5.55	No of DISCO Entries	79204
Avg Relevance Score	1.00	Total after Stemming & Removing Duplicates	4382

17 YouTube videos have been selected to retrieve 1159 comments for collection A. Five of these videos have been used so far to collect 193 user-guided comments for collection B. Analyzing these comments has derived 4382 unique words relevant to the job interview activity, forming our semantically enriched BoW. For the trial of this paper, we chose to expand the original words of the comments with synonyms,

antonyms, derivations, and DISCO entries. In future implementations, we aim to further expand the vocabulary with the remaining resources as described in Section (3.3). Applying the relevance scoring and labelling model described in Section (3.4) on collection A comments have assigned 724 comments as noisy and 435 comments as relevant. Text pre-processing these comments has derived a CTM matrix having 1159 comment term vectors and 903 predictor attributes representing the *tfidf* term weights, in addition to the target binary attribute containing the class value (noisy or relevant) of each training comment.

We have used the labelled CTM as a training corpus to train two types of classifiers widely used for document classification, C4.5 Decision Tree [9] and Naïve Bayes Multinomial [7], to evaluate predicting noisy comments that should be filtered out when retrieving further YouTube comments to be used for deriving key user characteristics directly relevant to the job interview activity. The C4.5 algorithm has the ability to auto-detect those predictors most contributing to the target class and use them in the underlying classification rules. Naïve Bayes Multinomial (NBM), on the other hand, is a probabilistic classifier that has achieved good prediction results in spam filtering [10]. We used three different training / testing corpus variations to train three models from each classifier to test the prediction stability performances. In the first variation, we test the classifiers on the same full dataset the classifiers are trained on, whereas in the second and third variations, we trained the classifiers on 80% and 60% of the full dataset, respectively, and tested on the remaining instances.

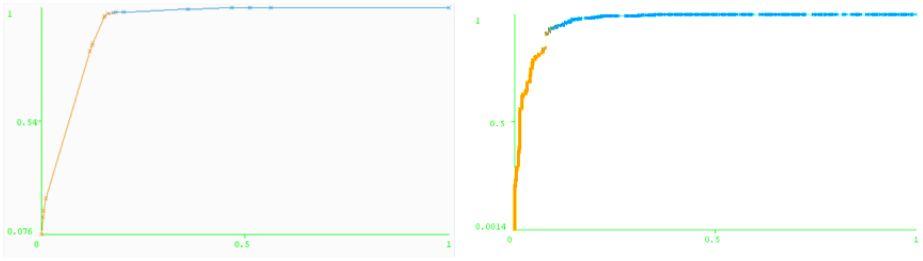
Table 3. Classification Evaluation Metrics

	C4.5 Full	C4.5 80%	C4.5 60%	NBM Full	NBM 80%	NBM 60%
Testing Size	1159	232	464	1159	232	464
Correctly Classified Comments	1070 (92.3%)	194 (83.6%)	390 (84.1%)	1063 (91.7%)	189 (81.5%)	362 (78.0%)
MAE	0.14	0.22	0.22	0.10	0.21	0.24
RMSE	0.26	0.38	0.37	0.27	0.41	0.45
TP Rate	0.98	0.90	0.92	0.92	0.85	0.76
FP Rate	0.16	0.28	0.31	0.09	0.24	0.18
Precision	0.91	0.85	0.85	0.95	0.86	0.89
Recall	0.97	0.9	0.92	0.92	0.85	0.76
ROC Area	0.93	0.84	0.82	0.97	0.87	0.85

Table 3 shows the evaluation metrics for the six trained models. The average correctly classified comments by the C4.5 algorithm is 86.7%, slightly higher than for the NBM algorithm, 83.7%, resulting in a slightly lower Root Mean Squared Error (RMSE) for C4.5 (0.34) than it is for NBM (0.38). However, the average Mean Absolute Error (MAE) for C4.5 and NBM are almost the same, 0.19 and 0.18, respectively. The True Positive (TP) rate is the rate of correctly classified noisy comments to the total number of noisy comments in the testing dataset. On average,

C4.5 is more able to correctly classify noisy comments from within the total available noise than NBM. However, NBM is less likely than C4.5 to misclassify relevant comments that may derive important user characteristics as noise from within the total relevant comments available. This is noticed in the lower False Positive (FP) rate for NBM than it is for C4.5, as well as for the higher Precision rates for NBM.

The Classifier Output also gives the ROC area, which reflects the true positive rate versus the false positive rate. This metric reflects the probability that a randomly chosen noisy comment in the testing data is ranked above a randomly chosen relevant comment, based on the ranking produced by the classifier. The best outcome is that all noisy comments are ranked above all relevant comments, in which case the ROC is 1. In the worst case it is 0. Figure 3 depicts the ROC curves for C4.5 (a) and NBM (b) both tested by the full data set (n = 1159), with FP Rate on the x-axis and TP Rate on the y-axis. NBM shows a slightly larger ROC area (0.90) than C4.5 (0.86). Moreover, NBM needs less costly misclassifications of noise (FP rate) than C4.5 to reach the optimal desired correct predictions of noisy comments (TP rate).



(a) (b)

Figure 3. ROC Curve for the (a) C4.5 and the (b) NBM Classifiers

In general, the output of the experimental study – the classification evaluation metrics – shows that the two classifiers implemented provide good performance in predicting and filtering out the noisy YouTube comments that are irrelevant to the particular activity domain of interest (job interviews). Although the C4.5 decision tree classifier is slightly better in filtering the noisy comments from the total available noise, the Naïve Bayes Multinomial classifier shows less risk in misclassifying relevant comments, which can derive key user characteristics to augment user models, as noise. In addition, the comment relevance scoring and labelling model proposed in Section (3.4) provides a reasonable estimate to whether each comment within the classification training corpus could be considered either noisy or relevant to the sought domain activity, i.e. job interview.

As discussed in Section 2, filtering out the irrelevant content from the noisy social media data is considered as the first phase in our research roadmap toward utilizing social media content to augment existing user models. After removing the identified noise, the remaining relevant YouTube comments will then be used to retrieve additional YouTube content generated by the users who posted these comments. These may include meta-data about any videos that these users upload or mark as favourites, additional comments they may post on YouTube, and explicit information that the users may write about themselves on their YouTube profiles. All these user-

generated contents will then be analyzed further to derive the social user profiles for those YouTube users. These profiles will then be mined to discover interesting associations between the several user characteristics that these profiles consist of. Finally, the revealed associations will be exploited to augment existing user models for similar users who use simulated learning environments for experiential learning.

6. Future Work

For future implementations of the approach, we aim to take several considerations into account to further improve filtering results, as summarized below:

- Further statistical analysis of the comments in the training corpus (collection A) could be conducted, in order to improve the accuracy of the scoring mathematical model. Comparisons with other variations, such as considering the comment size rather than and in addition to the comment intersection with the ground truth bag of words are also aimed. Expert-based evaluation of the computed scores and labels are also important to reduce false learning of the classification rules by the trained classifiers.
- Further semantic enrichment to the ground truth vocabulary by considering the ontologies described in Section (3.3) will be conducted. Weighting the original words derived from the controlled comments as well as the semantic expansions to these words by their importance to the activity domain of interest is also aimed to improve the accuracy of the relevance scoring mechanism.
- Further evaluations and comparisons with more classifiers that provide good classification results with unstructured data. These include variations to the Naïve Bayesian algorithm, Singular Value Decomposition-based algorithm, Support Vector Machines, and Combination algorithms. Moreover, classifier-specific parameter tuning and dimensionality reduction to the training comment-term matrix will be applied to further improve the prediction accuracy.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no ICT 257831 (ImREAL[7] project).

References

1. Agarwal N, Liu, H., Modelling and Data Mining in Blogosphere, In: Synthesis Lectures on Data Mining and Knowledge Discovery, R. Grossman, ed., Morgan & Claypool Publishers, vol. 1 (2009)
2. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G., Finding high-quality content in social media, In: Proceedings of the International Conference on Web Search and Web Data Mining (WSDM), Palo Alto, California, USA (2008)

[7] http://www.imreal-project.eu/

3. Chung, S. F., Kathleen, A., Chu-Ren, H., Using WordNet and SUMO to Determine Source Domains of Conceptual Metaphors, In: Proceedings of 5th Chinese Lexical Semantics Workshop (CLSW-5). Singapore: COLIPS. pp. 91-98, (2004)
4. Despotakis, D., Multi-perspective Context Modelling to Augment Adaptation in Simulated Learning Environments, Submitted and Accepted In**:** The 19th User Modeling, Adaptation, and Personalization UMAP Conference , Doctoral Consortium, Girona Spain (2011)
5. Feldman, R., Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, New York, NY, (2006)
6. Kamaliha, E., Riahi, F., Qazvinian, V., Adibi, J., Characterizing network motifs to identify spam comments. In: IEEE International Conference on Data Mining Workshops, 2008. ICDMW'08, pp. 919–928 (2008)
7. Kibriya, A. M., Frank, E., Pfahringer, B., Holmes, G., Multinomial Naive Bayes for Text Categorization Revisited, In: Lecture Notes in Computer Science, vol. 3339/2005, pp. 235-252 (2005)
8. Kolb, P., DISCO: A Multilingual Database of Distributionally Similar Words. In: Proceedings of KONVENS-08, Berlin, (2008)
9. Kotsiantis, S.B., Supervised Machine Learning: A Review of Classification Techniques, In: Informatica, vol. 31, pp. 249-268 (2007)
10. Metsis, V., Androutsopoulos, I., Paliouras, G., Spam filtering with naive bayes - which naive bayes? 3rd Conference on Email and Anti-Spam CEAS (2006)
11. Schimratzki, O., Bakalov, F., Knoth, A., König-Ries, B., Semantic Enrichment of Social Media Resources for adaptation, In*: Proceedings of International Workshop on Adaptation in Social and Semantic Web (SAS-WEB 2010),* Big Island of Hawaii, pp. 31-41 (2010)
12. Siersdorfer, S., Chelaru, S., Nejdl, W., Pedro, J., S., How useful are your comments?: analyzing and predicting YouTube comments and comment ratings, In: Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA, pp. 26-30, (2010)
13. Turney, P., Learning algorithms for keyphrase extraction. In: Information Retrieval, vol. 2(4), pp. 303–336 (2000)
14. Wang, F. Y., Carley, K. M., Zeng, D., and Mao, W., Social computing: From social informatics to social intelligence, In: IEEE Intell. Syst., vol. 22, no. 2, pp. 79–83 (2007)
15. Zhu, L., Sun, A., Choi, B., Online spam-blog detection through blog search. In: Proceedings of the Seventeenth ACM International Conference on Information and Knowledge Management(CIKM), pp. 1347–1348 (2008)