JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Implementing Naïve Bayes Classification to Identify Intrusive and Offensive Text on Social Media Platform

[1]Prashant Jagtap, [2]Asst. Prof. Lalita Randive
[1,2] Department of Computer Science & Engineering, MIT College, Aurangabad

Abstract— One of the fundamental challenges faced by Social Media Platforms (SMP) is the ability to restrict messages posting, in order to avoid unwanted messages from getting displayed. Another challenge is give users the ability to control the contents posted on their own private space to. Today's Social Media Platforms provide very little when it comes to controlling of contents. In this paper, we propose a method to be implementing in SMP allowing it's users to have a direct control on the messages posted on their walls. This is achieved through a flexible rule-based system, that allows users to customize the filtering criteria to be applied to their walls, and Machine Learning based soft classifier automatically labeling messages in support of content-based filtering.

Index Terms— Social Media Platforms, Information Filtering, Short Text Classification, Policy-based Personalization

I. INTRODUCTION

Online networks like MySpace, Facebook, Bebo, and LinkedIn characterize some of the most dynamic and capable manifestations of social media up till now. These sites permit networking on a grand scale, where individuals can connect with others based on offline friendships, common interests, universal professional objectives, or mutual acquaintances. When users join a Social Media site, it provides a page to user on which they can create a profile. They are urged to enter personal information such as hometown, work history, hobbies, favorite movies, interests, etc. They can then upload photos or link to other Web pages that interest them. This information is displayed on their Profile page, and users are given the option of making the page public or viewable only to those within their network. Profile pages serve as launching pads from which users explore these Social Media sites. They can search for other individuals, or find people with common interests. Users who identify others they want as part of their networks invite one another to be "friends", and such networks are displayed for others to see and browse. In this way, global networks of people with friends or interests in common are born. Social Media Platforms (SMPs) [19] are popularly used to share daily lives, contents, keep in touch with friends and share thoughts and information. Thus sharing of data includes images, text, audio and video formats. According to a survey by Facebook [20] statistics average user creates 90 pieces of content each month, whereas more than 30 billion pieces of content (web links, news stories, blog posts, notes, photo albums, etc.) are shared each month. Content filtering usually works by specifying character strings that, if matched, indicate undesirable content that is to be screened out. Content is typically screened for pornographic content and sometimes also for violence- or hate-oriented content. Critics of content filtering programs point out that it is not difficult to unintentionally exclude desirable content. However, the aim of the majority of these proposals is mainly to provide users a classification mechanism to avoid they are overwhelmed by useless data. In SMPs, information filtering can also be used for a different, more sensitive, purpose. This is due to the fact that in SMPs there is the possibility of posting or commenting other posts on particular public/private areas, called in general walls. Information filtering can therefore be used to give users the ability to automatically control the messages written on their own walls, by filtering out unwanted messages. We believe that this is a key SMP service that has not been provided so far. Indeed, today SMPs provide very little support to prevent unwanted messages on user walls. For example, Facebook allows users to state who is allowed to insert messages in their walls (i.e., friends, friends of friends, or defined groups of friends). However, no content-based preferences are supported and therefore it is not possible to prevent undesired messages, such as political or vulgar ones, no matter of the user who posts them. Providing this service is not only a matter of using previously defined web content [7], [9], [10] mining techniques for a different

application, rather it requires to design ad-hoc classification strategies. This is because wall messages are constituted by short text for which traditional classification methods have serious limitations since short texts do not provide sufficient word occurrences.

The goal of this work is to introduce an automated system as Filtered Wall (FW), which percolates useless and unwanted messages from SMPs. We process using Machine Learning (ML) [11], [13], [17] to grouped text which automatically attaches each short text depends on its contents to a set of categories of text.

More efforts are carried out to construct a robust text classifier which extract and select a set of characteristics and segregate properties. Proposed work based on earlier work from which we obtained the learning model and process for collecting pre-organized data. The main set of properties is build from the features of short text and is extended with the reference of information relevant to the context form which the message is derived. In this work, we make use of neural learning model which is proven more robust and dynamic solution in text classification technique. Our proposed method based on Radial Basis Function Networks (RBFN) because it holds some facilities of Radial Basis Function Networks (RBFN) such as acting as soft classifiers, in managing noisy data and intrinsically vague classes. We attempt to use two level hierarchical classification strategies. In the first hierarchical level, the RBFN separates short messages into Neutral and Non-Neutral category; in the second stage, Non-Neutral messages are organized into the group producing gradual estimates of appropriateness to each of the considered category.

Apart from classification capabilities, proposed system assures robust rule layer, which adventure a very flexible language to determine Filtering Rules (FRs), with the help of which user can decide the messages to be displayed and which should not be displayed on their walls. Filtering Rules (FRs) handles wide range of different filtering principles that can be associates according to user requirements. FRs accomplishes user profiles, relationships with the other users as friends, friends of friends, or defined groups of friends and the outcome of the ML partition process to describe the filtering principles to be required. Proposed system also provides advantage of Backlists which are specified by users and includes user names that are restricted to post any type of messages on a user wall for some time span.

II. RELATED WORK

The present work defines the construction of a system which supports content-based message filtering for SMPs, depending on Machine Learning techniques. Proposed system has relationships with the state of the art in content-based filtering, and with the field of policy-based personalization for SMPs and, generally in web contents.

A. Content-based filtering

Generally Information filtering systems are constructed to analyze a flow of effectively developed information dispatched asynchronously using information manufacturer producer and deliver to the user those information that are likely to satisfy his/her needs [5].

Assumption for content-based filtering is we have to consider operations of each user individually. As an outcome, system depending upon content-based filtering prefers items based on interaction between the content of the items and the user preferences as resisted to collaborative filtering [1], [6] system which selects items depending upon interaction between people with identical preferences. Documents refined using content-based filtering are mostly text documents and thus content-based filtering comes nearer to text classification. The process of filtering can be modeled as a case of single label, binary classification, dividing incoming documents into related and non-related types. Multi-label text categorization which tags messages is used by more complicated filtering systems.

Working of Content-based filtering depends on functions of ML paradigm with reference to which classifier is naturally motivated by learning from a set of pre-classified examples. A noticeable range of related work has newly appeared which conflict they accept property extraction methods, model learning, and collection of samples. The property extraction process plans text into a compact production of its content and is consistently applied to training and generalization phases.

B. Policy-based personalization of SMP contents

The efficiency of a learning method does play an important role in the decision of which technique to select. The most important aspect of efficiency is the computational complexity of the algorithm, even though storage necessities can also turn into a problem as many user profiles have to be maintained. Neural networks and genetic algorithms are much limited in speed as compared to other learning methods as several iterations are needed to determine whether or not a document is relevant [4]. Instance based methods slow down performance as more training cases turn out to be accessible because each and every example has to be analyzed in contrast to all the unseen documents. However, such systems do not offer a filtering strategy level with help of which user can develop the result of the classification process to elect how and to which level filtering process is carried out to remove unnecessary and useless information. In contrast, proposed filtering policy language allows the setting of FRs conferring to a range of benchmarks that do not scrutinize only the results of the classification process but also the relationships of the wall owner with other SMP users as well as information on the user profile [7]. Moreover, our system is complemented by a flexible mechanism for BL management that provides a further opportunity of customization to the filtering procedure.

In the field of SMPs, the majority of access control models proposed so far enforce topology-based access control, conferring to which access control necessity are articulated in terms of relationships that the requester should establish with the source proprietor. Proposed system utilizes similar concept to identify the users to which a FR applies. However, our filtering policy language enhances the languages recommended for access control policy specification in SMPs to cope with the extended requirements of the filtering domain [18]. Indeed, since we are dealing with filtering of unwanted contents, one of the important factors of our system is the availability of a description for the message contents to be accomplished by the filtering mechanism

[16]. In contrast, no one of the access control models specified previously enhances the content of the resources to enforce access control. Moreover, the notion of BLs and their management are not considered by any of the above-mentioned access control models.

III. FILTERED WALL ARCHITECTURE

Architecture which supports to SMP services depends on 3-tier architecture as shown in above figure. Goal of first layer is to deliver basic SMP functionalities. First layer is called as Social Network Manager (SNM). Second layer is known as Social Network Applications (SNAs). Third layer is called as Graphical User Interfaces (GUIs) which is additional layer to support some needed SNAs. User collaborate with system by using GUI for the purpose of setting and managing their FRs/BLs. GUI provides the functionality of Fire Walls (FWs), on which only certified messages are displayed according to their FRs/BLs rules.

The basic parts of our implemented system are Content-Based Messages Filtering (CBMF) and the Short Text Classifier (STC). Goal of these parts is to organize messages in to set of groups depends on their nature. With the help of STC module, first part accomplishes message separation.
The procedure followed by a message, can be summarized as follows and illustrated in figure:

1) After arriving into the private wall of one of the contacts in frendlist, the user tries to post a message, which is intercepted by FW.
2) Role of ML-based text classifier is to abstract metadata from the content of the message.
3) This abstracted data by classifier is further used by FW along with social graph and users profiles, to enforce the filtering and BL rules.
4) According to the generated outcome of step 3, either message will be published on wall or filtered by FW.

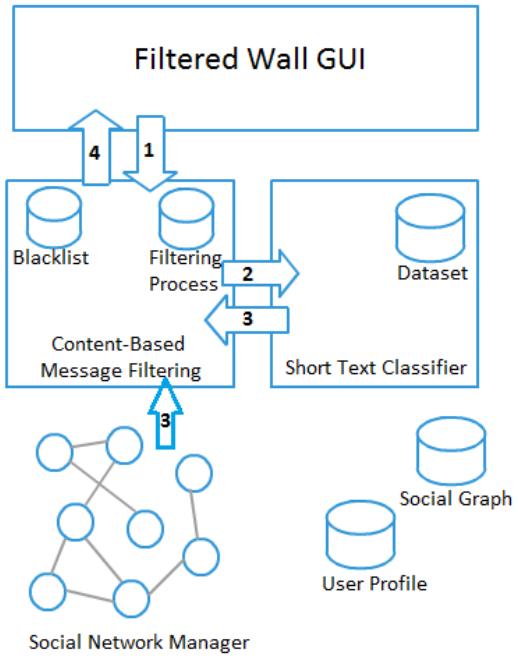


Figure: 1. Filtered Wall Conceptual Architecture and the flow messages follow, from writing to publication

IV. SHORT TEXT CLASSIFIER

Existing strategies such as newswires corpora functions well with huge documents but troubles with short documents. There are some complex conditions in restoring characteristics and discriminant features which describes essential concepts with combination of a complete and consistent group of supervised examples. Goal of proposed work is to design and evaluate some representation strategies along with a neural learning technique to semantically partition short texts.

We advent proposed job by illustrating a two level hierarchical technique, from a ML point of view, for that we consider that it is preferable to determine and terminate "neutral" sentences, then separate "non-neutral" sentences from the class of interest instead of doing everything in one step [7]. This technique is inspired from the related strategies which show benefits in partitioning text and/or short texts with the help of a hierarchical strategy. First level step is to group short texts according to labels with crisp Neutral and Non-Neutral labels. In the second stage, soft classifier works on crisp group of non-neutral short texts. For each short text, it produces estimated appropriateness or "gradual membership", without taking any "hard" decision on any of them. This list of ratings is then used by the subsequent phases of the filtering process. Later on phases of the filtering process uses such a list of grades.

A. Text Representation

The process of extracting a proper group of properties which describes texts of given document is critical, which can also harmful for the performance of overall classification technique. Some strategies were invented for text categorization procedure but accurate or more proper feature set and feature representation has not yet been investigated. Depending on these, we had taken into accounts three different properties as BoW, Document properties (DP) and Contextual Features (CF) [17]. First two properties are fully based on information contained within the text of the message.

The basic system uses Vector Space Model (VSM) to represent text. In this method a text document d_j is defined as a vector of binary or real weights dj $=w_{1j},\ldots..,w_{|\mathcal{T}|j}$, where the term $\mathcal{T}$ is the collection of terms which occurs at least once in at least one collected documents $\mathcal{T}r$, and $w_{kj} \in [0; 1]$ denotes the contribution of the t_k in to the semantics of document d_j [20]. Terms are described with words using BoW representation. In the case of non-binary weighting, the weight wkj of term tk in document dj is computed according to the standard term frequency - inverse document frequency (*tf-idf*) weighting function, defined as

$$tf - idf(\text{tk}, \text{dj}) = \#(\text{tk}, \text{dj}) log + \frac{|\mathcal{T}r|}{\#|\mathcal{T}r|(\text{tk})}$$

where $\#(t_k; d_j)$ represents the number of times t_k occurs in d_j, and $\# \mathcal{T} r(t_k)$ stands for the document frequency of term t_k, i.e., the number of documents in $\mathcal{T}r$ in which t_k occurs.

1) Correct words: it expresses the amount of terms $t_k \in \mathcal{T} \cap K$, where t_k denotes a term of the considered document dj and K is a set of known words the domain language. This value is normalized by

$$\sum_{k=1}^{|\mathcal{T}|} \#(t_k, d_j)$$

2) Bad words: Bad words are calculated similarly to the correct words feature, where the set K is a collection of "dirty words" for the domain language.

3) Capital words: it expresses words written in capital letters, calculated by the percentage of words existing in message containing more characters in capital case.

4) Punctuations characters: it is calculated as the percentage of the punctuation characters over the total number of characters in the message.

5) Exclamation marks: it is calculated as the percentage of exclamation marks over the total number of punctuation characters in the message.

6) Question marks: it is calculated as the percentage of question marks over the total number of punctuations characters in the message.

B. Filtering rules

While describing language for filtering rules, we have to consider three issues that can affect decision of message filtering as follows: 1) In SMP, one message can hold several different meanings. To avoid such situation FR should able to allow users defining of constraints for message author [14]. 2) We can apply some criteria for selection of author imposing conditions on their profile's attributes. By using this method, it is possible to define rules applying only to young creators or to creators with a given religious/political view. 3) In SMPs, with the service provided by social graph, one can find the activities of creator. So, we are able to design conditions deepening on type, depth and trust values of the relationship wall owner having with its friends.

A FR is therefore formally defined as follows.

Definition. (Filtering rule).

A filtering rule (FR) is a tuple consisting (author, creatorSpec, contentSpec, action), where: author stands for the user who describes the filtering rules; creatorSpec is a creator specification implicitly denotes a set of SMP users; contentSpec is a Boolean expression defined on content constraints of the form (C; ml), where C is a class of the first or second level and ml is the minimum membership level threshold [15] required for class C to make the constraint satisfied; action $\in$ {block; notify} denotes the action to be performed by the system on the messages matching contentSpec and created by users identified by creatorSpec.

C. Blacklists

The concept of Blacklist Management is used to bypass messages from unwanted peoples, no matter what they exactly consists of. BL are explicitly administered by the system. BL has ability to regulate the peoples in which user is interested and decide when users retention in the BL is finished [17]. This information is submitted to the system with the help of rules often called as BL rules. Rules of Blacklists may vary from person to person, so our system allows user to describe BL list and to decide who has to be banned from their walls and for how long. Therefore, a user might be banned from a wall, by, at the same time, being able to post in other walls [13].

BL rules allows wall holder to take decision to block users according to theirs profiles and relationships in the SMP [10], [15], [19]. Through BL rules, wall holder is capable to block unknown persons, peoples with which wall holder have only indirect relationships

or peoples about whom wall holder have some cheap opinion. This restriction can be endorsed for specific time period or for undecided time period. Restriction can depend upon the user's behavior in the SMP.

We use two measures based on user's bad behavior as: 1) if user has been injected into blacklist for more times than some defined threshold, then that user will remain into blacklist unless user's behavior is not improved. But this mechanism works on only those users which are already injected into blacklist at least one time. 2) Relative Frequency (RF) is used to catch bad behaviors of users. The task of RF is to find out those users whose messages always try to break down the filtering rules. These measures can be used locally or globally, as dealing with messages and BL of the user describing the BL rule or walls of all SMP users.
A BL rule is therefore formally defined as follows.

Definition (BL rule). A BL rule is a tuple consists of (author, creatorSpec, creatorBehavior, T), where: author is the SMP user who specifies the rule, i.e., the holder of wall; creatorSpec is a creator specification; creatorBehavior holds two components as RFBlocked and minBanned. RFBlocked = (RF, mode, window) is defined such that:

$$\mathrm{RF} = \frac{\#bMessages}{\#tMessages}$$

where #tMessages is the total number of messages that each SMP user identified by creatorSpec has tried to publish in the author wall (mode = myWall) or in all the SMP walls (mode = SN); whereas #bMessages is the number of messages among those in #tMessages that have been blocked; window is the time interval of creation of those messages that have to be considered for RF computation; minBanned = (min, mode, window), where min is the minimum number of times in the time interval specified in window that SMP users identified by creatorSpec have to be inserted into the BL due to BL rules specified by author wall (mode = myWall) or all SMP users (mode = SN) in order to satisfy the constraint. T denotes the time period the users identified by creatorSpec and creatorBehavior have to be banned from author wall.

V. RESULTS AND DISCUSSION

Let us suppose that the system applies a given rule on a certain message. As such, Precision reported is the probability that the decision taken on the considered message (that is, blocking it or not) is actually the correct one. In contrast, Recall has to be interpreted as the probability that, given a rule that must be applied over a certain message, the rule is really enforced. Let us now discuss, with some examples, the results presented, which reports Precision and Recall values. The second column represents the Precision and the Recall value computed for FRs with (Neutral; 0:5) content constraint. In contrast, the fifth column stores the Precision and the Recall value computed for FRs with (Vulgar; 0:5) constraint.

By trial and error we found a quite good parameter configuration for the Naive learning model. The best value for the M parameter, that determines the number of Basis Function, is heuristically addressed to N=2, where N is the number of input patterns from the dataset. The value used for the spread , which usually depends on the data, is = 32 for both networks M1 and M2
Network M1 has been evaluated using the OA and the K value. Precision, Recall and F-Measure were used for the M2 network because, in this particular case, each pattern can be assigned to one or more classes.

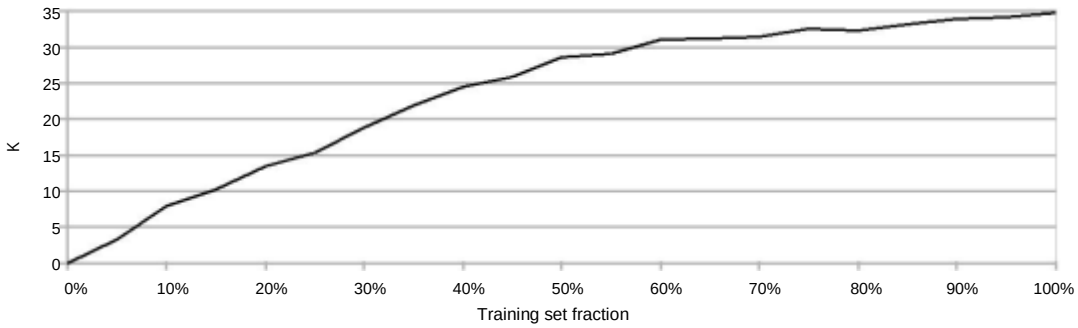
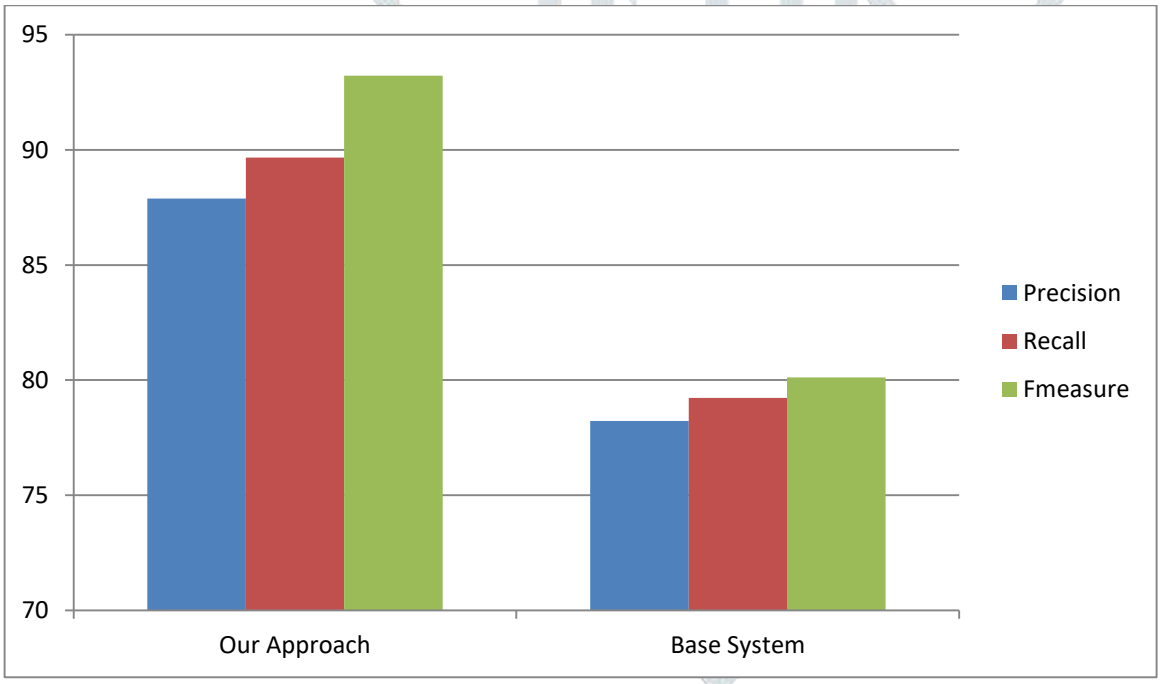


Figure Line Graph for Training using MAGS and Labeled

Results achieved by the content-based specification component, on the first level classification, can be considered good enough and reasonably aligned with those obtained by well-known information filtering techniques [51]. Results obtained for the content-based specification component on the second level are slightly less brilliant than those obtained for the first, but we should interpret this in view of the intrinsic difficulties in assigning to a messages a semantically most specific category. However, the analysis of the features shows that the introduction of contextual information (CF) significantly improves the ability of the classifier to correctly distinguish between non-neutral classes. This result makes more reliable all policies exploiting non-neutral classes, which are the majority in real-world scenarios.



It is evident that the proposed algorithm was able to successfully identify an post as intrusive or neutral with 87.9% precision. The algorithm's fmeasure 66of detection of mags was higher (89.1%) as compared to that of detection of spammers as suggested in our base reference (68.4%). The results showed that Naïve Bayes algorithm performs better in detection of intrusive posts accounts. Our algorithm was able to maintain the high accuracy of naïve bayes algorithm in detecting neutral posts and at the same time, retain the accuracy in detecting intrusive poses thereby, increasing the overall accuracy.

V. CONCLUSION

Proposed method represents system to filter unwanted messages from walls of SMP users. The system uses ML soft classifier to implement FRs and BL to boost filtering preference. FRs should allow users to state constraints on message creators. Proposed system allows user to decide to describe BL list and to decide who has to be banned from their walls and for how long. Therefore, a user might be banned from a wall, by, at the same time, being able to post in other walls. By analyzing the user's behavior in the past, learning methods applied for content-based filtering in proposed system find out the proper and relevant documents. This

technique yields to restrain to user to prepare documents similar to those already seen. So, the approach is recognized as over-specialization problem.

REFERENCES:

[1] P. J. Denning, "Electronic junk," Communications of the ACM, vol. 25, no. 3, pp. 163–165, 1982.

[2] S. Pollock, "A rule-based message filtering system," ACM Trans- actions on Office Information Systems, vol. 6, no. 3, pp. 232–254, 1988.

[3] P. S. Jacobs and L. F. Rau, "Scisor: Extracting information from on- line news," Communications of the ACM, vol. 33, no. 11, pp. 88–97, 1990.

[4] P. J. Hayes, P. M. Andersen, I. B. Nirenburg, and L. M. Schmandt, "Tcs: a shell for content-based text categorization," in Proceedings of 6th IEEE Conference on Artificial Intelligence Applications (CAIA- 90). IEEE Computer Society Press, Los Alamitos, US, 1990, pp. 320–326.

[5] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: Two sides of the same coin?" Communications of the ACM, vol. 35, no. 12, pp. 29–38, 1992.

[6] P. W. Foltz and S. T. Dumais, "Personalized information delivery: An analysis of information filtering methods," Communications of the ACM, vol. 35, no. 12, pp. 51–60, 1992.

[7] P. E. Baclace, "Competitive agents for information filtering," Communications of the ACM, vol. 35, no. 12, p. 50, 1992.

[8] D. D. Lewis, "An evaluation of phrasal and clustered representa- tions on a text categorization task," in Proceedings of 15th ACM International Conference on Research and Development in Information Retrieval (SIGIR-92), N. J. Belkin, P. Ingwersen, and A. M. Pejtersen, Eds. ACM Press, New York, US, 1992, pp. 37–50.

[9] C. Apte, F. Damerau, S. M. Weiss, D. Sholom, and M. Weiss, "Automated learning of decision rules for text categorization," Trans- actions on Information Systems, vol. 12, no. 3, pp. 233–251, 1994.

[10] H. Schutze, D. A. Hull, and J. O. Pedersen, "A comparison of classifiers and document representations for the routing problem," in Proceedings of the 18th Annual ACM/SIGIR Conference on Resea. Springer Verlag, 1995, pp. 229–237.

[11] M. J. Pazzani and D. Billsus, "Learning and revising user profiles: The identification of interesting web sites," Machine Learning, vol. 27, no. 3, pp. 313–331, 1997.

[12] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in Proceedings of Seventh International Conference on Information and Knowledge Management (CIKM98), 1998, pp. 148–155.

[13] G. Amati and F. Crestani, "Probabilistic learning for selective dissemination of information," Information Processing and Management, vol. 35, no. 5, pp. 633–654, 1999.

[14] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," in Proceedings of the Fifth ACM Conference on Digital Libraries. New York: ACM Press, 2000, pp. 195–204.

[15] R. E. Schapire and Y. Singer, "Boostexter: a boosting-based system for text categorization," Machine Learning, vol. 39, no. 2/3, pp. 135– 168, 2000.

[16] Y. Zhang and J. Callan, "Maximum likelihood estimation for filtering thresholds," in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 294–302.

[17] F. Sebastiani, "Machine Learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[18] A. Adomavicius, G.and Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transaction on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.

[19] M. Chau and H. Chen, "A Machine Learning approach to web page filtering using content and structure analysis," Decision Support Systems, vol. 44, no. 2, pp. 482–494, 2008.

[20] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, "Content-based filtering in Social Media Platforms," in Proceedings of ECML/PKDD Workshop on Privacy and Security issues in Data Mining and Machine Learning (PSDML 2010), 2010.