NATURAL LANGUAGE PROCESSING THROUGH DIFFERENT CLASSES OF MACHINE LEARNING

Harsh Jain and Keshav Mathur

Department of Information Science and Engineering,
BMS College of Engineering, Bangalore, India
harshjain30@gmail.com
keshavmathur@outlook.com

ABSTRACT

The internet community has been benefitting tremendously from the works of various researchers in the field of Natural Language Processing. Semantic orientation analysis, sentiment analysis, etc. has served the social networks as well as companies relying on user reviews well. Flame identification has made the internet less hostile for some users. Spam filtering has made the electronic mail a more efficient means of communication. But with the incessant growth of the internet, NLP using machine learning working on massive sets of raw and unprocessed data is an ever-growing challenge. Semi-supervised machine learning can overcome this problem by using a large set of unlabeled data in conjunction with a small set of labeled data. Also, focusing on developing NLP systems that can contribute to developing a unified architecture could pave the way towards General Intelligence in the future.

KEYWORDS

Semantic Orientation, Social Networks, E-mail, Sentiment Analysis, Unsupervised Machine Learning

1. INTRODUCTION

Natural Language Processing (NLP) is a field of Artificial Intelligence that deals with the interactions between natural languages and computers. The purpose of NLP is to enable machines to process an input in natural or human language and derive meaning from it.

Modern NLP algorithms are based on machine learning and are designed to automatically learn rules for generating an output from a given input through the analysis of large corpora of labelled or unlabelled examples. Many different classes of machine learning algorithms have been applied to NLP tasks. They can be organized based on the type of input for training in the following way:

- Supervised learning algorithms, where the input comprises of labelled examples. The output for previously unseen inputs is generated by using a general function or mapping from inputs to outputs created with the help of the training data.
- Unsupervised learning algorithms, where the input comprises of unlabelled examples. The desired output is unknown and produced by discovering structure in the data.
- Semi-supervised learning, which uses a combination of both labelled and unlabelled examples.

David C. Wyld et al. (Eds) : CCSIT, SIPP, AISC, PDCTA, NLP - 2014
pp. 307–315, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4226

With the internet boom and increasing popularity of social networks, analysis of the written word has become more significant. Tools that can perform various tasks related to the internet, such as identifying spam, analysing the semantic orientation of words and sentiment analysis, have become highly useful. In this paper, various supervised, unsupervised and semi-supervised machine learning approaches for performing such tasks have been outlined. It also identifies their application and possible improvements in the domain of social networks.

2. DISCUSSIONS

According to UNESCO as well as BBC, there are about 7,000 different languages in the world. With copious amounts of documents produced in a many languages every day, a mechanism that could relate documents in different languages and indicate their similarity would serve as a great tool. A language-independent supervised approach to controlled vocabulary keyword assignment using the EUROVOC thesaurus, whose multi-lingual nature allows cross-language document comparison was presented by Ralf Steinberger [1]. In the first step, which is keyword assignment, if a certain word occurs significantly more often in a given text than it occurs, on average, in a large selection of 'normal' texts (the reference corpus), it was identified as a keyword. Both the documents are lemmatised for this. Lists of *stop words* and multi-word expressions were also included. Then, using the EUROVOC thesaurus, frequent words were mapped to associates, which were multiplied by their *keyness* or relevance to the document to generate ranks and identify the descriptors of the document. Although the author could not conduct extensive tests on his algorithm, the initial tests conducted on the corpus gave satisfactory results. One inherent drawback of automated keyword generation is that there is no standard way of testing the accuracy of the results produced by such an algorithm, since it is well-known that no set generated manually can be claimed as the right set for a document. Also, the corpus that was used by the author pertains to a very specific legalistic and administrative sublanguage. This algorithm with some optimization has vast potential for use in cross-language document search and comparison.

When it comes to application of NLP to the web and social networks, mechanisms to identify the semantic orientation of words are particularly useful. In a supervised approach by Hatzivassiloglou and McKeown [2], it was proposed that there are constraints on the semantic orientations of conjoined adjectives. Using corpus data, the hypothesis was verified and the results were found to be extremely significant statistically, except for a few cases where the sample was small. Next, these constraints were used to construct a log-linear regression model automatically. This model, with the combination of supplementary morphology rules, predicted whether two conjoined adjectives are of same or different orientation. An appreciable accuracy of 82% was achieved. Several sets of adjectives were classified according to the links inferred in this way and labelled as positive or negative. A remarkable 92% accuracy was achieved on the classificationtas k for reasonably dense graphs and 100%accuracy on the labelling task. But in the classification of adjectives, the number of times the method accurately classified a test set dropped when the data was sparser, ranging from 92% to 78%.A strong point of this method is that decisions on individual words are aggregated to provide decisions on how to group words into a class and whether to label the class as positive or negative.

Hatzivassiloglou and McKeown's work was extended by Hatzivassiloglou with Janyce M. Wiebe [16]**.** Subjectivity in natural language refers to aspects of language used to express opinions and evaluations. This paper studied the effects of dynamic adjectives, semantically oriented adjectives and gradable adjectives on a simple subjectivity classifier and established that they are strong predictors of subjectivity. Since the mere presence of one or more adjectives is useful in predicting that a sentence is subjective, this paper investigated the effects of additional lexical semantic features of adjectives. It considered two such features: *semantic orientation*, which

represents an evaluative characterization of a word's deviation from the norm for its semantic group; and *gradability*, which characterizes a word's ability to express a property in varying degrees. For gradability, two indicators were used: grading modifiers such as *very* and inflected forms of adjectives. In the experiment all adjectives with a frequency of 300 or more from the 1987 Wall Street Journal corpus (21 million words) were extracted; producing a list of 496 words. 453 of the 496 adjectives (91.33%) were assigned gradability labels by hand (using the designations of the Collins COBUILD dictionary), while the remaining 53 words were discarded. The method automatically assigned labels to the entire set of 453 adjectives, using 4-fold cross-validation with resulted precision of 94.15%, recall of 82.13%, and accuracy of 88.08% for the entire adjective set. In the experiment they measure the precision of a simple prediction method for subjectivity: a sentence is classified as subjective if at least one member of a set of adjectives S occurs in the sentence and objective otherwise. By varying the set S (e.g., all adjectives, only gradable adjectives, only negatively oriented adjectives, etc.), it is noted that all sets involving dynamic adjectives, positive or negative polarity, or gradability are better predictors of subjective sentences than the class of adjectives as a whole. This kind of analysis could help efforts to encode subjective features in ontologies. The authors had begun to incorporate the features developed into systems for recognizing flames and mining reviews in Internet forums. They were seeking ways to extend the orientation and gradability methods so that individual word occurrences, rather than word types, are characterized as oriented or gradable. They also planned to incorporate the new features in machine learning models for the prediction of subjectivity and test their interactions with other proposed features.

Wiebe also independently came up with a supervised approach for subjectivity [17].The author claims that subjectivity tagging is about distinguishing sentences that are *used to* present opinions and evaluations from sentences used to present factual information. Her previous work on subjectivity *(Wiebe, Bruce, & O'Hara* 1999; Bruce&Wiebe 2000) had established a relation between subjectivity and the presence of adjectives in a sentence. This paper identified higher quality adjective features using the results of a method for clustering words according to distributional similarity (Lin 1998), seeded by a small amount of detailed manual annotation. These features were then further refined with the addition of lexical semantic features of adjectives, specifically *polarity* and gradability, which could be automatically learned from corpora. For the experiment a corpus of 1,001 sentences of the Wall Street Journal Treebank Corpus (Marcus *et al.* 1993) was manually annotated with subjectivity classifications and also the strength of subjective elements was manually rated on a scale of 1 to 3. The experiment, with a simple adjective feature where a sentence is subjective if one adjective is found, had a precision of 55.8%. With the improvements in adjective features using Distributional Similarity where a sentence is labelled as subjective if an adjective from the seed sets is found, it had a precision of 61.2% (a significant improvement). Further improvements can be made with addition of lexical semantic features of adjectives (polarity and gradability). Both features used together showed an accuracy of 71%, displaying together they are more precise. Previous work on subjectivity had been focused on many applications like recognizing flames, mining Internet forums for product reviews, and clustering messages by ideological point of view. Wiebe's approach in this paper was directed at supplementing such endeavors by developing a repository of potential subjective elements to enable us to exploit subjective language. The adjectives learned by the experiment were being incorporated into a system for recognizing flames in Internet forums. In addition; the author planned to apply the method to a corpus of Internet forums, to customize knowledge acquisition to that genre.

Hatzivassiloglou's, McKeown's and Wiebe's work was based on supervised machine learning and required training data. Peter D. Turney [3]presented a paper in 2002 that proposed an unsupervised learning algorithm for semantic orientation. It also wasn't limited to just adjectives. The general strategy in this paper was to infer semantic orientation from semantic association. The semantic orientation of a given word was calculated from the strength of its association with

some chosen positive words, minus the strength of its association with some chosen negative words. This general strategy is called SO-A (Semantic Orientation from Association). This paper also examined SO-PMI-IR (Semantic Orientation from Point-wise Mutual Information and Information Retrieval) and SO-LSA (Semantic Orientation from Latent Semantic Analysis).The experiments suggested that the SO-PMI-IR can reach the same level of accuracy as Hatzivassiloglou and McKeown's, given a sufficiently large corpus. The results also hinted that SO-LSA is able to use data more efficiently than SO-PMI-IR, and SO-LSA might surpass the 80% accuracy attained by SO-PMI-IR, given a corpus of comparable size.

PMI-IR was proposed, and compared with LSA, by Turney in his previous work [6].In this paper, PMI-IR was presented as a new unsupervised algorithm for recognizing the closest synonym of a given problem word from a given option set. It achieved this by using PMI algorithm that analyzed statistical data collected by Information Retrieval (IR). The performance was evaluated by testing it on 80 synonym test questions from the Test of English as a Foreign Language (TOEFL) and 50 synonym test questions from a collection of tests for students of English as a Second Language (ESL). It was then compared to that of Latent Semantic Analysis (LSA) on the same question sets. LSA is a statistical algorithm based on Singular Value Decomposition (SVD). A variation on this algorithm has been applied to information retrieval, where it is known as Latent Semantic Indexing (LSI).[5] The paper also evaluated four different versions of PMI-IR, using four different kinds of queries to the Alta Vista search engine. The first described co-occurrence as words being present in the same document, second one used the NEAR operator, the third removed the problem of antonyms being rated same as synonyms and the fourth one took context into account.

The PMI-IR algorithm, like LSA, is based on the concept of co-occurrence. The core idea is that "a word is characterized by the company it keeps". There are many well-known lexical databases that include synonym information. But these have a problem of sparse data and the need for extensive manual labor for each language. Several hybrid approaches have been proposed which combine statistical and lexical information. PMI-IR addresses the sparse data problem by using a huge data source: the Web.

The results with the TOEFL questions showed that PMI-IR (in particular for the query which removes the antonym problem) can score almost 10% higher than LSA. However, the interpretation of the results was difficult, due to two factors: (1) PMI-IR was using a much larger data source than LSA. (2) PMI-IR was using a much smaller chunk size than LSA. A future work that emerges from this is seeing how LSA would perform for such a large data source like the one used for PMI and also test the hypothesis by Landauer and Dumais [4] who claim that mutual information analysis would obtain a score of about 37% on the TOEFL questions, given the same source text and chunk size as they used for LSA. PMI-IR may prove to be suitable as a tool to aid in the construction of lexical databases and automatic keyword extraction. It might also be useful for improving IR systems.

Turney's algorithm [3] could have many potential applications, such as filtering "flames" for newsgroups, improving Tong's system for generating *sentiment timeline*[7], in the analysis of survey responses to open ended questions, in an automated chat system to help decide whether a positive or negative response is most appropriate or to classify reviews.

Turney used his PMI-IR algorithm in an approach for classification of reviews [8]. The simple unsupervised learning algorithm took as input a written review and generated a classification (thumbs up or thumbs down) as the output. The classification was based on the average semantic orientation of the phrases in a review that contained adjectives or adverbs. The first step was to use a part-of-speech tagger to identify phrases in the input text that contain adjectives or adverbs. The second step was to evaluate the *semantic orientation* of each extracted phrase. This was done

using the PMI-IR algorithm. A phrase was considered to have a positive semantic orientation when it had good associations(e.g., "subtle nuances") and a negative semantic orientation when it had bad associations (e.g., "very cavalier"). The final step was the classification of a review as recommended(thumbs up) or not recommended(thumbs down) based on the average semantic orientation of all the extracted phrases.

The algorithm was evaluated on 410 reviews from Epinions (all from unique users) from four different domains (automobiles, banks, movies, and travel destinations).It achieved an average accuracy of 74% ranging from 84% for automobile reviews to 66% for movie reviews.

Turney's algorithm is useful in labelling an input text as positive or negative. But to some domains in the online community, this is not enough. Companies around the world have turned to the internet to obtain reviews for their products. They require much more information than just the overall sentiment about the topic. For example, a smart phone manufacturing company might want to analyse its public discussion forums and people's reviews on its social network fronts to find out the specific features of the phone that are favoured by its users. This makes the tools for extraction of sentiments about a given topic almost indispensable to this community. A "Sentiment Analyzer" was proposed by Jeonghee Yi et al [9]**.** Sentiment Analyzer (SA) extracted sentiment (or opinion) about a subject from online text documents. The authors believe that although the overall opinion about a topic is useful, it is only a part of the information of interest. So, instead of classifying the sentiment of an entire document about a subject, SA detected all references to the given subject, and determined sentiment in each of the references. The paper anticipated shortcomings of the purely statistical approaches and showed that the analysis of grammatical sentence structures and phrases based on NLP techniques mitigated some of the shortcomings. To extract the feature terms form a document, only the nouns were selected from the document. Then, one of the two feature term selection algorithms, one based on the mixture language model by Zhai and Laffertry [10]; and the other based on the likelihood-ratio test by Dunning [11], were applied. The extracted feature terms by each of the algorithms were manually examined by two human subjects and their precision and accuracy were tabulated. The Likelihood Test method consistently performed better than the Mixture Model algorithm. The next step was identifying the sentiment phrase and the assignment of the sentiment to a subject. For this, *SA* used sentiment terms defined in the *sentiment lexicon* and sentiment patterns in the *sentiment pattern database*. The *sentiment lexicon* contained the sentiment definition of individual words collected from various sources. The sentiment pattern database contained sentiment extraction patterns for sentence predicates. For each sentiment phrase detected, *SA* determined its *target* and final polarity based on the sentiment pattern database. *SA* first identified the *Ternary-expression* for the statement, and tried to find matching sentiment patterns. Once a matching sentiment pattern was found, the *target* and sentiment assignment were determined as defined in the sentiment pattern.*SA* consistently demonstrated high quality results of 87% for review articles, 86 ~91% (precision) and 91 ~93% (accuracy) for the general web pages and news articles. The results on review articles were comparable with the state-of-the-art sentiment classifiers, and the results on general web pages were better than those of the state of the art algorithms by a wide margin.The more advanced sentiment patterns required a fair amount of manual validation. In the future, full parsing could provide better sentence structure analysis, thus better relationship analysis.

Most social networks allow the users to communicate verbally. Although, a lot of them let the users decide whether something is inappropriate for public discussions (flagging), an intelligent system could help the system determine basic conspicuous inappropriate content. Such messages could contain the use of abusive language. Some of them could be directed towards a person, perhaps in a private message, not using any abusive language, but of hostile nature. These messages could be identified and filtered through supervised machine learning, thus reducing the need for user involvement in the process. Ellen Spertus presented aprototype system

"Smokey"[12], which analyses the syntax and semantics of each sentence for 47 pre-listed features and builds a vector accordingly. It then combines the vectors for the sentences within each message. For training, a set of 720 messages was used by Quinlan's C4.5 decision-tree generator to determine the feature-based rules used to categorize the messages as "flame" or "okay". The test set consisting of 460 messages was then categorized by the system, as well as manually for accuracy analysis. Smokey was able to correctly categorize 64% of the flames and 98% of the non-flames in the test. There are certain limitations that this system was unable to overcome, such as recognizing sarcasm and innuendo and making sense of complex sentences and mistakes in grammar, punctuation, and spelling. In the future, this system could learn from dictionaries and thesauri, user feedback, or proximity to known insults. It could also benefit from morphological analysis, spelling and grammar correction and analysing logical parse trees of sentences.

A system like Smokey could also be used to prioritize mail and help in the identification of spam mail. Internet subscribers world-wide are unwittingly paying an estimated €10 billion a year in connection costs just to receive "junk" emails, according to a study undertaken for the European Commission [13]. Many machine learning approaches have already been suggested to help identify and get rid of spam emails. A lot of them rely on the keyword-based approach, where certain keywords found in spam emails are used to identify other spam emails. But there is an inherent flaw in using keyword-based approach: if the keywords that are considered for a mail to be marked as spam are known, spammers could actively work to avoid their inclusion in their mails. It could be seen as an "arms race", where the spammers continuously identify and avoid the keywords that the anti-spam systems consider and the system engineers continuously try to find and add new keywords to stay one step ahead of the spammers. Ion Androutsopoulos et al investigated the performance of two machine-learning algorithms, the Naïve Bayesian approach and a memory-based classification approach in the context of anti-spam filtering, and demonstrated how both are better than the keyword-based filter [14]. An experiment was conducted for the same, which used a benchmark corpus consisting of both spam as well as legitimate mails. An important note was the cost of a mistake: a legitimate mail marked as spam is much more undesirable than a spam passing as legitimate. This difference factor was denoted by λ. In the Naive Bayesian approach, the corpus was pre-processed in which Baye's theorem and the theorem of total probability were employed. In the memory-based classification approach, a variant of the simple *k*-nearest-neighbour (*k*-nn) algorithm was used. The experiment used the algorithm implemented in the TiMBL software. Next, formulae were derived for parameters like Weighted Accuracy and Weighted Error, as well as Total Cost Ratio, which were used to compare a filter's performance with the baseline (when no filter is used). λ was varied creating three different scenarios and in each scenario, the number of selected attributes was varied between 50 and 700. In all three scenarios, the two aforementioned approaches performed better than the keyword-based approach, and except the scenario in which λ was very high, also better than the no filter approach. In the future, alternative learning methods for the same task could be examined, including attribute-weighted versions of the memory-based algorithm. Alternative attribute selection techniques including term extraction methods to move from word to phrasal attributes can also be explored.

Md. Saiful Islamet al [15] in a study investigated the possibilities of modelling spammer behavioural patterns instead of vocabulary as features for spam email categorization, as they believe that keyword-based approach will eventually be less successful as spammers will try and circumvent the filters that such models will employ. The three well-known machine learning algorithms; Naïve Bayes, DTI and SVMs; were experimented to model common spammer patterns. Common spammer techniques were listed and a model was developed exploiting machine learning algorithms to capture common spammer patterns. 21 such patterns were extracted from each of the 1000 mails consisting of equal number of spam and legitimate mails. Accuracy, Precision and Recall were calculated for each of the three methods using the corpus

and it was found that Naïve Bayesian classifier is more efficient than the other two classifiers. Building a perfect data set free from noise or imperfection remains a continuous challenge for spam filtering techniques as noise adversely affect the classifier's performance. Also, most training models of the classifier have limitations on their operations .Naïve Bayes has the advantage of incremental inclusion and/or exclusion of features and DTIoffers best expressive power. So, natural progression will be combining these two MLalgorithms in multi-core architecture, running both classifier simultaneously indifferent cores to minimize time and applying voting mechanism to increase positivity, which will give best opportunity to model spammer common patterns. Multi-classifier based spam filters exploiting spammer behavioural patterns can also be developed.

The very big challenge in pattern recognition task and machine learning process is to train a discriminator using labeled data. However, real world problems have largely raw and unprocessed data and labeling them becomes almost impossible. Semi-supervised learning overcomes this problem by using a small set of labeled data along with a huge set of unlabeled data to train the discriminator.

In a paper, Shafigh Parsazad et al [18] proposed an evolutionary approach called Artificial Immune System (AIS) to determine which data is better to be labeled to get the high quality data. The human immune system was an inspiration for this paper because it is robust, adaptable and automatic. Immune system consists of white blood cells which are responsible for detection and elimination of Antigens. These are called Anti bodies. The immune system has a memory to save its work. In the paper, a modified version of the aiNet algorithm proposed by Younsi (called aiNetC) is used. aiNetC works on the principle that more the similarity between the antibody and the antigen more is the strength of connection between them. The measure of similarity between antibody and antigen cells is called *Affinity*. An affinity threshold in antibody detection process called *Network Affinity Threshold* (NAT) is defined. If the affinity of a given antibody and antigen is lower than NAT, it's assumed the antibody recognized the antigen. It iterates over all antigens till all are recognized or a number of generations reached and after this antibodies too close to each other are eliminated. The main purpose of aiNetC is clustering. The algorithm proposed by the authors modifies aiNetC so that instead of clustering the data it tries to describe them with very few antibodies. They argue that if these antibodies have labels, the accuracy of the clustering will be greatly improved. Another advantage of such a method in semi-supervised learning is that labeling the data will not be random and it is done with the lowest possible information that can be provided. For experimentation, two semi-supervised learning algorithms were included: semi-supervised KMeans as the clustering algorithm and semi-supervised support vector machines as the classification algorithm. Five datasets were used. First of all random set of labels were generated and fed into these algorithms. Then the algorithm was used for analyzing the data and all datasets were fed to aiNetC algorithm. After analyzing, aiNetC algorithm recommended some data to be labeled by the user to achieve the best result. This labeled data was again fed to the learning algorithms as information that we have from dataset. Result of this learning was measured. Experimental results showed the vast improvements in the results.

Most NLP systems focus on individual NLP tasks and very few have of them have characteristics that could contribute to developing a unified architecture. In a paper [19], Ronan Collobert and Jason Weston defined a rather general convolutional network architecture and described its application to many well-known NLP tasks including part-of-speech tagging, chunking, named-entity recognition, learning a language model and the task of semantic role-labeling (SRL) and all these tasks were integrated into a single system which was trained jointly. The main focus of this paper is SRL. Complex tasks like this require large number of complex features which make traditional NLP systems slow and not desirable for large applications. The authors advocated a deep neural network (NN) architecture, trained in an end-to-end fashion. The input sentence is

processed by several layers of feature extraction. They showed how multi-task learning and semi-supervised learning significantly improve the performance of this task in the absence of hand-engineered features. Multitask learning (MTL) is the procedure of learning several tasks at the same time with the aim of mutual benefit. A NN automatically learns features for the desired tasks in the deep layers of its architecture, thus sharing deep layers in these NNs would improve features produced by these deep layers, and hence improve generalization performance. The MTL network jointly trains supervised tasks on labeled data and unsupervised tasks on unlabeled data because unlabeled data is abundant and freely available on the web. The experimental results showed that the deep NN could be applied successfully to various tasks such as SRL, NER, POS, chunking and language modeling. They also demonstrated that learning tasks simultaneously can improve generalization performance. In particular, when training the SRL task jointly with the language model their architecture achieved state-of-the-art performance in SRL without any explicit syntactic features. This was an important result, given that the NLP community considers syntax as a mandatory feature for semantic extraction. In the experiment NER error was not considered and future work would include more thorough evaluations of these tasks.

3. CONCLUSION

With the exponential growth of internet, it's becoming increasingly important for all organizations relying on it in some way to make the most of the information they have at their disposal. Since the internet is grounded in natural language, NLP has become a very important topic. The internet community has been benefitting tremendously from the works of various researchers in this field. But, with the internet's growth has come a hyper exponential growth in the amount of data that humans are producing each year. The fifth annual IDC Digital Universe study released on June 28, 2011 [19] stated that the world's information was doubling every 2 years. It projected that the world would be generating 50 times the amount of information by the end of the decade. Although there are well-developed algorithms in both the supervised and unsupervised classes of machine learning for various NLP tasks, NLP using machine learning working on massive sets of raw and unprocessed data is an intricate challenge. Semi-supervised machine learning can simplify this problem by using a large set of unlabeled data in conjunction with a small set of labeled data and could be the way forward for NLP in the future. Another useful approach would be to focus on developing NLP systems in such a way that they can contribute to developing a unified architecture. Such architecture would be necessary for generic semantic tasks and could pave the way towards General Intelligence in the future.

REFERENCES

[1] Ralf Steinberger. Cross-lingual Keyword Assignment
[2] Vasileios Hatzivassiloglou and Kathleen R. McKeown. 2007. Predicting the Semantic Orientation of Adjectives
[3] Peter D. Turney. May 16, 2002. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus
[4] T.K. Landauer, S.T. Dumais: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104 (1997) 211-240.
[5] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41 (1990) 391-407.
[6] Peter D. Turney: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL.
[7] R.M. Tong. An operational system for detecting and tracking opinions in on-line discussions. Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification (pp. 1-6). New York, NY: ACM. (2001)
[8] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

[9] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, Wayne Niblack. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Technique
[10] C. Zhai and J. Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In Proc.of the 10th Information and Knowledge Management Conf.
[11] T. E. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1).
[12] Ellen Spertus. Smokey: Automatic Recognition of Hostile Messages
[13] Data protection: "Junk" E-mail Costs Internet Users 10 Billion a Year Worldwide -Commission Study. In: http://europa.eu/rapid/pressReleasesAction.do?reference=IP/01/154,last accessed on 15thNov, 2013
[14] Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Georgios Sakkis, Constantine D. Spyropoulosand Panagiotis Stamatopoulos. Learning to Filter Spam E-Mail: A Comparison of a Naïve Bayesian and a Memory-Based Approach
[15] Md. Saiful Islam, Abdullah Al Mahmud, Md. Rafiqul Islam. Machine Learning Approaches for Modelling Spammer Behaviour
[16] Vasileios Hatzivassiloglou, Janyce M. Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity
[17] Janyce M.Wiebe. Learning Subjective Adjectives from Corpora
[18] Shafigh Parsazad, Ehsan Saboori, Amin Allahyar.Data Selection for Semi-Supervised Learning
[19] Ronan Collobert, Jason Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning

ACKNOWLEDGEMENTS

The authors wish to thank the Accendere KMS Research team for the guidance and mentoring which made this work successful.