A Hybrid Sentiment Analysis Framework for Large Email Data

Sisi Liu
Information Technology Academy
College of Business, Law and Governance
James Cook University, PO Box 6811, Cairns,
QLD 4870, Australia
Sisi.Liu1@my.jcu.edu.au

Ickjai Lee
Information Technology Academy
College of Business, Law and Governance
James Cook University, PO Box 6811, Cairns,
QLD 4870, Australia
Ickjai.Lee@jcu.edu.au

Abstract*—Sentiment analysis for online text documents has been a *burgeoning field of text mining among researchers and scholars for the past few decades. Nevertheless, sentiment analysis on large Email data, a ubiquity means of social networking and communication, has not been studied thoroughly. This paper proposes a framework for Email sentiment analysis using a hybrid scheme of algorithms combined with Kmeans clustering and support vector machine classifier. The evaluation for the framework is conducted through the comparison among three labeling methods, including SentiWordNet labeling, Kmeans labeling, and Polarity labeling, and five classifiers, including Support Vector Machine, Naïve Bayes, Logistic Regression, Decision Tree and OneR. Empirical results indicate a relatively high classification accuracy with proposed framework in comparison with other approaches.

Keywords-Sentiment analysis; text mining; Kmeans; Support Vector Machines

I. INTRODUCTION

As online social networking and communication is increasingly appealing to the public, extracting information and patterns from online documents and web content has become a burgeoning research focus recently [1] [2]. Meanwhile, with the widespread of Internet and digital devices, communication and social networking through Electronic mail (Email) is ubiquitous. From 2011 to 2015, statistics indicates an increase of 3% in the number of global Email users with an average of 1.7 Email accounts per user counted in 2015 [3]. Furthermore, business Email communication accounts for the majority of the total Email traffic with over 108.7 billion Emails exchange every day [3], and Email remains the most common way of business workspace communication.

Byron [4] states that it is unavoidable to reveal feelings, either intentionally or unintentionally, when people communication through sending Emails. Referring to as emotional polarity computation, sentiment analysis, originally derived from text mining, is a mining method specifically developed for the determination of writers' or speakers' emotional state or attitude towards certain topics or domains [5]. Sentiment analysis on large business Emails could reveal valuable patterns for business intelligence [6]. As the applications of sentiment analysis and opinion mining have been extended to various fields for the last decade, the techniques and algorithms for sentiment analysis have been developed and improved as well. Lexicon-based approach and machine learning approach outweigh hybrid approach in terms of methods; while reviews and social media rank high in the domain options [1].

In this research paper, a hybrid sentiment analysis schema of approaches using combined *K*means clustering and Support Vector Machine (SVM) classifier algorithms for large Email data is presented. The main contributions of this paper reflect in the following three aspects:

1) Introducing a hybrid scheme of algorithms for systematic and rigorous sentiment analysis for massive Email data;
2) Searching for the most appropriate labeling method for unlabeled data through the comparison among *SentiWordNet* (SWN) labeling, *Kmeans* labeling, and *Polarity* labeling;
3) Examining the feasibility and efficiency of combined *K*means clustering and SVM classification algorithm for sentiment classification in the proposed framework.

The outline of this paper is epitomized as follows. Section 2 reviews previous research on Email classification. Section 3 defines the flow work of the proposed scheme of algorithms. Section 4 displays empirical experimental results. And finally, Section 5 summarizes findings and points out possible future directions.

II. RELATED WORK

In Pang and Lee [6]'s survey for opinion mining and sentiment analysis, the essentiality and essence of extracting opinions and sentiments from decision-making process lies in social reliance and dependence on online recommendations and suggestions with the widespread of World Wide Web and network technologies. The idea of applying sentiment analysis for online text could be dated back to 2002 when the first paper addressing the market sentiments published in the Association for Computational Linguistics (ACL) [7].

Generally, sentiment analysis is composed of feature selection and sentiment classification. Features such as terms and their frequency and negation are extracted through techniques including N-gram model, Bag of Words (BoWs), Term Frequency-Inverse Document Frequency (TF-IDF), Pointwise Mutual Information (PMI) [1] [7] [8]. Sentiment

978-1-4673-9323-2/15 $31.00 © 2015 Crown Copyright
DOI 10.1109/ISKE.2015.91

classification algorithms are categorized into machine learning approaches, lexicon-based approaches and hybrid approaches, among which hybrid approaches are more prevalent recently. For example, Feng, Wang, Yu, Yang and Yang [9] combine clustering approach with SWN lexicon for blogs sentiment analysis; Li and Wu [5] utilize *K*means clustering for hotspot detection and SVM for sentiment classification and prediction.

Current research on text mining and sentiment analysis mainly addresses large scale social media data, such as Facebook data, Twitter corpus, and blogs. For instance, Li and Wu [5] conduct a study on hotspot detection through online forum. Additionally, Balasubramanyan, Routledge and Smith [10] propose an algorithm model for the prediction of poll results using public opinion mining. As for Email data analysis, most studies focus on the identification of spam mails, discarded mails, the study of social networking among Emails, and priority issues. [11] [12] [13]. However, less research has been conducted on Email sentiment analysis. Mohammad and Yang [14] study the gender difference in sentiment axis among set of sentiment labeled Email data, and Hangal, Lam and Heer [15] design a system for visualizing archived Email data with sentiment words tracking. Despite of these studies, a systematic and structured framework for Email sentiment classification has not been investigated yet.

III. FRAMEWORK AND METHODOLOGY

Figure 1 displays the proposed sentiment analysis framework for unlabeled Email datasets. The framework is composed of various techniques including data extraction, preprocessing, feature extraction, sentiment lexicon and sentiment classification. Preprocessing incorporates the following five steps: duplication removal, tokenization, stop words removal, stemming, and POS tagging. Feature selection method utilizes BoWs approach with opinion words as features. For pre-labeling process, three methods including lexicon labeling using SWN, *K*means labeling, and polarity labeling using Liu, Li, Lee and Yu [16]'s English opinion lexicon have been compared and contrasted in terms of classification accuracy. As for sentiment classification algorithms, SVM, NB, Logistic Regression (LR), Decision tree (J48), and OneR classifiers have been tested. Detailed flow of work will be explained in the following subsections.

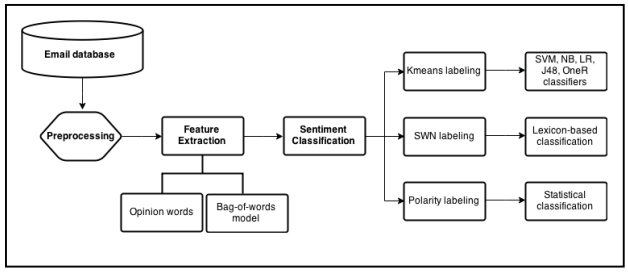

Figure 1. Hybrid sentiment analysis framework for Email data.

A. Preprocessing

As Email data are unstructured and noisy, preprocessing is essential for generating relevant data and passing the refined Email messages for experiments. Generally, standard preprocessing steps for text mining tasks involve stemming, lemmatization, stop word removal and etc. [1] [2] [9] [11]. However, Email data cleaning is a complex task due to the limitation of noisy types of data, such as headers and quotations, that could be cleaned [17]. As mentioned previously, only Email content has been queried from the database. Herein, for resolving and refining unstructured Email text, the five pre-processing steps practiced in this research paper are:

- ***Remove duplication***: through the implementation of "DISTINCT" keyword and subjects without "RE" or "FW" notations in SQL query statements, this process aims at minimizing the retrieval of duplicated data;
- ***Tokenization***: converting the entire Email messages into separate tokens for further cleaning;
- ***Stop words removal***: removing words with special characters or common words, such as "the", "a" and etc., that are meaningless for classification;
- ***Stemming****:* converting words into their original format. For instance, "going" into "go", "friends" into "friend";
- ***POS tagging****:* identifying each token as "adjective", "noun", "verb" or "adverbs" and assigning a POS tag to each one.

Algorithm: Email data preprocessing

Define E as the whole Email set containing Email message $e_1,e_2,\ldots,e_n$, W as a set of tokens in each Email message containing word $w_1,w_2,\ldots,w_n$

```
1:  Connect to database through JDBC
2:  Query from database using SQL statement
3:  Remove duplicates from E
4:  While E has next, do
5:     For each Email message e_n , do
6:        Tokenization
7:        For each word w_i in e_n, do
8:           Remove stop words
9:           Stemming
10:          Remove special characters
11:          POS tagging
12:       End for
13:    End for
14: End
```

Figure 2. Pseudo-code for Email data pre-processing.

Figure 2 shows the pseudo code for preprocessing algorithm used in our research, in which tokenization, stop words removal, stemming and POS tagging are coded using Apache Lucene libraries [18]. Note that POS tagging process is exclusively used for SWN labeling purpose.

B. Feature Extraction

Feature selection process is divided into two phases: feature detection and extraction. Current features found in most sentiment analysis studies include terms frequency, n-gram, POS tags, and negation [1] [8]. As indicated by Mejova and Srinivasan [19], feature extraction techniques, in particular for sentiment analysis, are generally domain specific. Although quite a few methods have been proposed, strengths and weaknesses have been examined in all of them [1] [2] [5] [20]. As discussed in the previous section, the research on Email sentiment analysis is limited, thus the most appropriate feature selection method for Email data remains for further studies. Therefore, in accordance with the characteristics of Email data being informative and communicative, the proposed framework implements the direct and baseline approach of **BoWs** model with **opinion words** as features [7] [21].

BoWs: a simple and direct word representation model utilized in Information Retrieval (IR) for document classification, where word frequency is counted regardless of grammar and order. A sample dictionary generated with BoWs model is presented in Fig. 3.

BoWs Model: $\begin{Bmatrix} [happy = 2], [awful = 0], [maybe = 4], \\ [likes] = 3], [anger = 0] \end{Bmatrix}$

Figure 3. Sample modeling text document using BoWs.

Opinion words: are words commonly defined and used as opinion expressions. Positive opinion words including good, like, happy, excellent and etc.; Negative opinion words including bad, hate, dislike and etc. [1]. This paper implements Liu, Li, Lee and Yu [16]'s English opinion lexicon particularly established for social media sentiment analysis. The list contains 2006 positive words and 4784 negative words (see Table 1).

TABLE I. SAMPLE POSITIVE WORDS AND NEGATIVE WORDS FROM [16]'S OPINION WORDS LIST.

Positive words	Negative words
achieve, adore, confidence, enjoy, fine, success	anger, amiss, bitter, disgrace, irritate, lousy, sick

Let PL be a collection of positive word, NL be a collection of negative words, and A be a set of attributes generated from E, containing feature $a_1, a_2, \dots \dots, a_i$,

$$a_i = frequency * e_n \ (e_n \in PL \cup NL). \quad (1)$$

For each feature attribute a_i, a min-max normalization to $[0,1]$ interval for the comparison to Weka results [22].

$$Normalized_{(a_i)} = \frac{a_i - Min(a_i)}{Max(a_i) - Min(a_i)}. \quad (2)$$

In the above Equation (2), a_i represents i^{th} row of attribute A. $Min(a_i)$ represents the minimum value in attribute range $A[i]$, while $Max(a_i)$ represents the maximum value in attribute range $A[i]$.

C. Sentiment Classification

Since the dataset chosen for conducting experiments is unlabeled, it is essential to have pre-labeling process for providing training data for classification algorithms. Three labeling methods have been selected for comparison purpose. **SWN labeling** is a lexicon labeling approach; ***K*means labeling** is an unsupervised labeling approach; and **Polarity labeling** is a heuristic labeling approach. Variables used in the equation in this section are correspondence with those defined in Fig. 3.

The three labeling method practiced in the paper are described as follows:

SWN Labeling: SentiWordNet 3.0 [23] is an English sentiment term lexicon based on WordNet. The entire lexicon consists of more than 0.1 million sets of cognitive synonyms (synsets) with two values, positivity and negativity, assigned to each synset. SWN labeling approach is implemented by searching the matching keyword with POS tag assigned to it (as discussed in Section 3.2).

Let $ObjScore\ o$ be the final score of each synset, o is calculated with the below equation

$$o = 1 - (PosScore + NegScore). \quad (3)$$

Let $Score_{(e)}$ be the final score of each Email message, $Score_{(e)}$ is calculated as follows:

$$Score_{(e)} = \frac{\sum_{i=1}^{n} o_i}{n}. \quad (4)$$

Define class label as $c_{(e)}$, label results are parametrized by δ:

$$\text{Class } c_{(e)} = \begin{cases} -1, & S < \delta \\ 0, & S = \delta \\ 1, & S > \delta \end{cases} \qquad \delta \in |0,0.01| \quad (5)$$

Kmeans Labeling: using Weka SimpleKMeans [22] clustering algorithm, a range of k values from 3 to 7 has been chosen for k-sensitivity test to be discussed in the next section. The clustering criteria Sum of Squared Errors (SSE) is calculated as below:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{x} \sqrt{(x_j - \bar{x})^2}. \quad (6)$$

The clustering results of SSE with corresponding number of clusters is shown in the table below:

TABLE II. KMEANS CLUSTERING RESULTS.

numClusters	3	4	5	6	7
SSE	722.686	628.271	626.471	616.639	585.063

Therein, the result of clustering assignments using 3 clusters is to be opted for the comparison with SWN and polarity labeling methods.

Polarity Labeling: polarity for each Email message is simply defined by the occurrence of opinion words extracted from Liu, Li, Lee and Yu [16]'s word list. -1 is assigned if the total number of occurrence of positive words is less than that of negative words; 1 if otherwise. If the time of occurrence of positive words and negative words is equal, or there are neither positive words nor negative words appear, 0 will be assigned.

$$\text{Class } c_{(e)} = \begin{cases} -1, \sum_{k=1}^{i}(w_i \in PL) > \sum_{k=1}^{i}(w_i \in NL) \\ 0, \sum_{k=1}^{i}(w_i \in PL) = \sum_{k=1}^{i}(w_i \in NL) \\ \quad or\ w_i \notin (PL \cup NL) \\ 1, \sum_{k=1}^{i}(w_i \in PL) < \sum_{k=1}^{i}(w_i \in NL) \end{cases} \quad (7)$$

Followed by labeling process, post-labeled dataset is utilized as training data for *Machine learning classifiers, Lexicon-based classification and statistical classification* [1] [24]. A *k*-sensitivity test with comparison among five supervised classifiers has been undertaken with 10 folds cross-validation experiments. Lexicon-based classification is conducted on the basis of SWN score for each instance. And statistical classification results from the polarity labeling process.

IV. EMPIRICAL RESULTS

The Enron Email corpus is a large public set of Email data adapted for the majority scientific research and experiments on Email classification task. The original version of Enron corpus (available at (https://www.cs.cmu.edu/~./enron/) contains more than 0.5 million messages retrieved from about 150 users. As the fundamental purpose of this research is to classify sentiments from Email content, a database version of the corpus [25] has been used for easier removal of unstructured data components, such as "from", "to", and "content-type".

Considering the feasibility and validity of the research, a total number of 200 Email messages have been generated as the dataset with proper preprocessing (which will be discussed in the next section) for the examination and evaluation of the framework. Data is accessed through J connector linking MySQL database and Eclipse IDE using Java programming language (see Fig. 2).

After data cleaning and feature selection processes, 6790 feature words have been extracted as attributes in numeric value representations. The minimum and the maximum value among all attributes for each Email data are used for normalization. The final format for feature selection process is represented as follows in Fig. 4:

Format: #feature: value #feature: value #feature: value ……

745:0.8 913:0.6 964:0.6 998:0.6 1011:0.6 1012:0.8 1073:0.6 1074:0.6 1238:0.5 1315:0.6 1412:0.6 1430:0.6 1701:1.0 1727:0.6 1936:0.6 2784:-0.2

Figure 4. Fragments of feature represented Email data.

In this section, two types of experiments have been conducted for the comparison among different labeling methods and classifiers in terms of classification accuracy. The first experiment on *k*-sensitivity assists in the justification of the option of SVM as the classification algorithm for the proposed framework. And the second experiment on different labeling methods assists in the justification of the option of *K*means clustering as labeling method for the proposed framework.

A. K-sensitivity Test

The fundamental purpose for conducting *k*-sensitivity test is to compare the performance of five selected classification algorithms: SVM, NB, LR, J48 and OneR using Weka machine learning environment [22]. Table 3 below defines a sample confusion matrix using three classes for displaying classification assignments.

TABLE III. CONFUSION MATRIX FOR CLASSIFICATION RESULTS (*K*=3).

Classified as / **Labelled as**	**Confusion Matrix**		
	Positive	*Neutral*	*Negative*
Positive	tPos	fNeuP	fNegP
Neutral	fPNeu	tNeu	fNegNeu
Negative	fPNeg	fNeuNeg	tNeg

In Table 3, the elements in diagonal represent data that are correctly classified into labeled classes. The performance of various classifiers is measured by *Precision, Recall, F-measure and Accuracy*. For each class, take "Positive" as an example, precision is calculated using the fraction of true positive assignments against all classified positive instances; recall is calculated using the fraction of true positive assignments against all labeled positive instances; and *F*-measure is calculated using the harmonic mean of the former two values. Accuracy measures the correctly classified rate. The mathematical equations are presented below:

$$Precision(Pos) = \frac{tPos}{tPos + fPNeu + fPNeg}. \quad (8)$$

$$Recall(Pos) = \frac{tPos}{tPos + fNeuP + fNegP}. \quad (9)$$

$$F - measure(Pos) = 2 * \frac{Precision(Pos) * Recall(Pos)}{Precision(Pos) + Recall(Pos)}, \quad (10)$$

$$Accuracy = \frac{tPos + tNeu + tNeg}{total}. \quad (11)$$

On the basis of Equation (8)-(11), the results for each experiment are evaluated on precision, recall and *F*-measure rate for each class and an average accuracy rate. An eventual classification results with confusion matrix and results matrix using SVM is depicted as a representative:

TABLE IV. CLASSIFICATION RESULTS USING SVM (*K*=3).

Classified as / **Labeled as**	**Confusion Matrix**			**Results**			
	Pos	*Neu*	*Neg*	*Precision*	*Recall*	*F-Measure*	*Accuracy*
Positive	9	1	0	75.0 %	90. 0%	81. 8%	97. 7%
Neutral	2	187	0	99.5 %	98. 9%	99. 2%	
Negative	1	0	0	0.0%	0.0 %	0.0 %	

The overall performance curve (see Fig. 5) indicates a prominence of LibSVM over other four classifiers with an average accuracy of 96.44% using different clustering results as labels. Furthermore, in comparison with other classification algorithms, LibSVM has relatively low sensitivity to the changes of *k* values. Herein, LibSVM is an appropriate option for Email sentiment classification and further comparison of different labeling methods for our study.

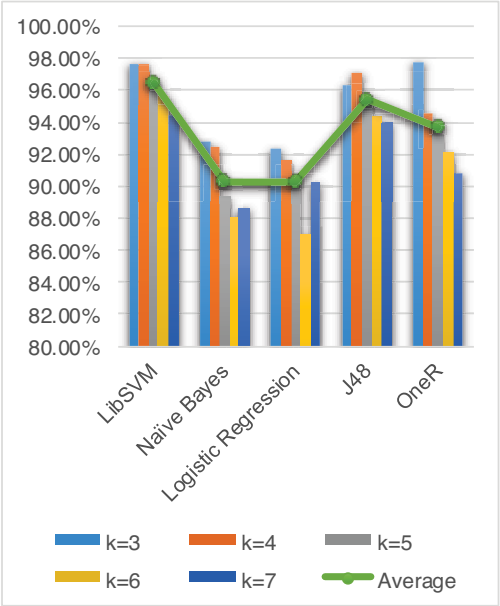

Figure 5. Performance comparison among five classification algorithms.

B. *Different Labeling Methods Comparison*

As details described in Section 3.4.1, this research has implemented and evaluated three labeling methods: SWN, *K*means and polarity labeling. In accordance with the previous experimental results, LibSVM algorithm is the option for the comparison and contrast of the three labeling approaches. The classification performance is measured using confusion matrix with precision, recall and *F*-measure, the same as *k*-sensitivity test. The classification result using polarity labeling approach is listed as an example in Table 5.

TABLE V. CLASSIFICATION RESULTS USING POLARITY LABELING.

Classified as / **Labelled as**	**Confusion Matrix**			**Results**			
	Pos	*Neul*	*Neg*	*Precision*	*Recall*	*F-Measure*	*Accuracy*
Positive	80	18	8	84. 2%	75. 5%	79. 6%	65.8 %
Neutral	1	43	2	48. 9%	93. 5%	64. 2%	
Negative	14	27	7	41. 2%	14. 6%	21. 5%	

The comparison results of the performance are shown in Fig. 6 below:

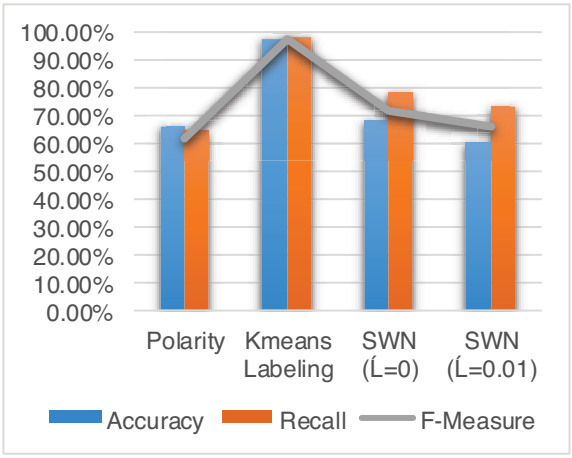

Figure 6: Performance comparison among different labeling methods.

As demonstrated in Fig. 6, *K*means labeling is observed to gain the best accurate results, with 97.7% in accuracy, 98% in recall and 97.8% in *F*-measure, outperforming SWN and polarity labeling. In terms of SWN, two parameters $\delta = 0$ and $\delta = 0.01$ have been examined considering lexicon sensitivity. Results indicate that $\delta = 0$ has performed better than $\delta = 0.01$.

V. Conclusion and Future Work

Email is the most widely used communication method in business. Understanding hidden patterns in Email communication is of great importance in business intelligence. A hybrid sentiment analysis scheme of combined *K*means clustering with SVM algorithm for Email data has been proposed in this research. The proposed hybrid sentiment classification algorithm implements the BoWs model, a baseline feature extraction method with opinion words as features [26]. Through several systematic experiments on the comparison among SVM, NB, LR, J48 and OneR classifiers using *K*means clustering results as labels ranging from 3 to 7, and among SWN, *K*means and polarity labeling methods using SVM classifier. On the basis of the experimental results (see Fig. 5 and Fig. 6), the combined *K*means and SVM algorithm is observed to achieve relatively high accuracy in comparison with other approaches. Therein lies the great opportunity in the application of the proposed framework to sentiment analysis in other fields and datasets.

However, as indicated in Fig. 7, in comparison with lexicon-based approach using SWN and statistical approach using polarity algorithm [1][7], the classification assignments obtained by proposed hybrid classification algorithm are comparatively unbalanced as 188 out of 200 Email messages are classified into neutral. One of the reasonable explanation for the circumstance would be the incompleteness and limitations of the data cleaning.

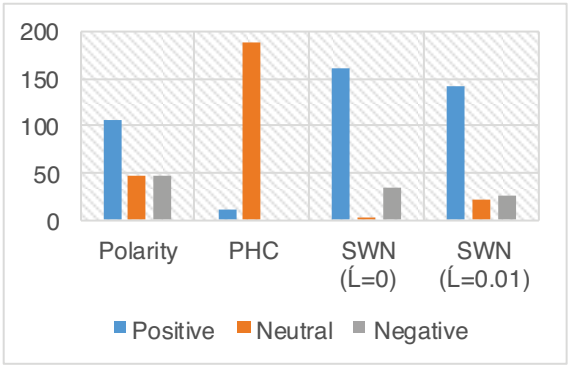

Figure 7: Classification assignments using different approaches.

Herein, promising and potential future research directions are to incorporate well-structured and thorough Email data cleaning procedures [17] and domain-orientated feature selection approaches [7]. With less noisy data and more useful features for classification, it is estimated to achieve more balanced classification results and sentiment patterns.

References

[1] Medhat, W., Hassan, A. and Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5, 4 2014), 1093-1113.

[2] Khan, F. H., Bashir, S. and Qamar, U. TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 572014), 245-257.

[3] Radicati, S. and Hoang, Q. Email statistics report, 2011-2015. *Retrieved May*, 252011), 2011.

[4] Byron, K. Carrying too heavy a load? The communication and miscommunication of emotion by Email. *Academy of Management Review*, 33, 2 2008), 309.

[5] Li, N. and Wu, D. D. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48, 2 2010), 354-368.

[6] Liu, B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

[7] Pang, B. and Lee, L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2, 1-2 2008), 1-135.

[8] Liu, B. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 22010), 627-666.

[9] Feng, S., Wang, D., Yu, G., Yang, C. and Yang, N. *Sentiment clustering: a novel method to explore in the blogosphere*. Springer, City, 2009.

[10] Balasubramanyan, R., Routledge, B. R. and Smith, N. A. From tweets to polls: Linking text sentiment to public opinion time series2010).

[11] Klimt, B. and Yang, Y. *The enron corpus: A new dataset for Email classification research*. Springer, City, 2004.

[12] Sharma, A. K. and Sahni, S. A comparative study of classification algorithms for spam Email data analysis. *International Journal on Computer Science and Engineering*, 3, 5 2011), 1890-1895.

[13] Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. *A Bayesian approach to filtering junk e-mail*. City, 1998.
[14] Mohammad, S. M. and Yang, T. W. *Tracking sentiment in mail: how genders differ on emotional axes*. City, 2011.
[15] Hangal, S., Lam, M. S. and Heer, J. *Muse: Reviving memories using Email archives*. ACM, City, 2011.
[16] Liu, B., Li, X., Lee, W. S. and Yu, P. S. *Text classification by labeling words*. AAAI Press, City, 2004.
[17] Tang, J., Li, H., Cao, Y. and Tang, Z. *Email data cleaning*. ACM, City, 2005.
[18] McCandless, M., Hatcher, E. and Gospodnetic, O. *Lucene in Action: Covers Apache Lucene 3.0*. Manning Publications Co., 2010.
[19] Mejova, Y. and Srinivasan, P. *Exploring Feature Definition and Selection for Sentiment Classifiers*. City, 2011.
[20] Pak, A. and Paroubek, P. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. City, 2010.
[21] Hu, M. and Liu, B. *Mining and summarizing customer reviews*. ACM, City, 2004.
[22] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11, 1 2009), 10-18.
[23] Baccianella, S., Esuli, A. and Sebastiani, F. *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf*. City, 2010.
[24] He, Y. and Zhou, D. Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47, 4 (7// 2011), 606-616.
[25] Schulz, A. H. Enron Email Data, http://www.ahschulz.de/enron-Email-data/enron-mysqldump_v5.sql.gz(Accessed 20 May 2015.
[26] O'Keefe, T. and Koprinska, I. *Feature selection and weighting methods in sentiment analysis*. City, 2009.