Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 124 (2017) 77–85

Procedia Computer Science

www.elsevier.com/locate/procedia

4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia

Effects of Word Class and Text Position in Sentiment-based News Classification

June Ling Ong Hui, Gan Keng Hoon*, Wan Mohd Nazmee Wan Zainon

School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia

Abstract

In news domain, sentiments captured in the form of sentiment labels (emoticon) give a quick feedback of reactions towards the contents of the news. As these reactions are valuable indicators for social and political well beings, we are motivated to automate the classification of news texts based on these indicators, e.g. happy, sad, angry, amused etc. Unlike other review texts that contain more explicit words which can be interpreted directly for sentiment classification, news texts mostly report facts and figures. This resulted in needs to identify whether contents of news can be exploited for classification or otherwise. Hence, in this work, a study is conducted to analyze and determine the relevant key parts of news contents that can be to be used for sentiment-based classification. Two criteria, i.e. text Part of Speech and text position, which could possible influence the training of the classifier are studied. Evaluations are conducted on the collection of 250 English news texts labelled with sentiments from sentiment voting system. The results for sentiment-based category has recorded F score of 0.422 whereas for polarity-based category has recorded F score of 0.837. The study has shown that when finer categories (e.g. happy, sad etc.) are used, the inspected criteria are less effectively; however, when these categories are based on polarity orientations, the outcomes show potentials of the proposed criteria especially for text positioned at headlines and text using adjective words.

Keywords: News Sentiment; Text Classification; Text Analysis

1. Motivation

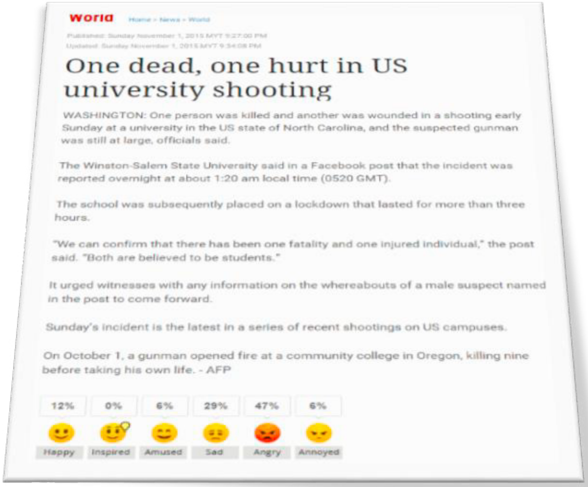
Publication of news involves events, stories, reports from all walks of life normally triggers reactions from the readers. This trigger does not limit to paper-based news medium but also digital as well. People tend to have their

* Corresponding author. Tel.: +604-6534634.
E-mail address: khgan@usm.my

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 4th Information Systems International Conference 2017
10.1016/j.procs.2017.12.132

feeling expressed by talking and debating the topic of interest at the coffee shop on in the neighborhood during the days when the internet has not been widely used yet. Such scenario of offline discussion is often emotionally oriented and still can be seen anywhere until today. In this digital age, the online platform provides another alternative for the expression of thoughts and feelings. Some of these online platforms are social media sites and blogs such as Twitter, Facebook and Blogger where the online users usually express their feelings in text or comment.

Among these, one of the interesting ways for gathering feedback from online users is via the sentiment voting system. As a news article may provoke mix emotions and different readers may feel differently, the sentiment voting features provided in some websites are believed to have contributed greatly in understanding the readers' reaction after reading the news. This special feature captures the reader's reaction or attitude towards a situation by using emotion representation. Fig. 1 shows a news excerpt taken from TheStar online that shows the number of votes for each sentiment label. The six reactions captured are Happy, Inspired, Amused, Sad, and Annoyed, with three positive and three negative sentiments. Number of votes generated by users for a particular sentiment suggest that the news has higher tendency to provoke such emotion after reading it.



World Home > News > World

Published: Sunday November 1, 2015 MYT 9:27:00 PM
Updated: Sunday November 1, 2015 MYT 9:34:08 PM

One dead, one hurt in US university shooting

WASHINGTON: One person was killed and another was wounded in a shooting early Sunday at a university in the US state of North Carolina, and the suspected gunman was still at large, officials said.

The Winston-Salem State University said in a Facebook post that the incident was reported overnight at about 1:20 am local time (0520 GMT).

The school was subsequently placed on a lockdown that lasted for more than three hours.

"We can confirm that there has been one fatality and one injured individual," the post said. "Both are believed to be students."

It urged witnesses with any information on the whereabouts of a male suspect named in the post to come forward.

Sunday's incident is the latest in a series of recent shootings on US campuses.

On October 1, a gunman opened fire at a community college in Oregon, killing nine before taking his own life. - AFP

12%	0%	6%	29%	47%	6%
Happy	Inspired	Amused	Sad	Angry	Annoyed

Fig. 1. News article with sentiment voting (The Star Online http://www.thestar.com.my/).

As emotions are often one of the influencing factors during decision making, knowing what others think and feel is essential for various group of people in the society. Marketer and advertiser can track public responses of an advertising campaign; politician and public relation officials can get insights from public emotions reacted on certain event especially on the political situation of a country; stock traders are able to capture the market behaviors for stock trading etc. Hence, this results in the needs to improve existing news categorization based on emotional labels. Handling emotion-based categories is challenging, and requires different treatment in terms of contents analysis compared to related works that have mostly focused on classification based on general categories like sports, politic, entertainment etc. Contents analysis based on measures like term frequency, n-grams, may not be able to differentiate those sentiments associated terms then general keywords. Thus, we believe that weighting should be given to certain type of terms so that they can better reflect the concept of emotion-based category. For example, news headlines could be more significant compared to news bodies in catching readers' attention, hence useful to be associated to the effect of an emotion. Likewise, type of word classes may also have different effect in terms of arousing the emotion of reader which resulting to a vote decision.

As such, in this paper, we focus on the analysis of textual contents in the setting of sentiment-based news classification. Our research scope is set to discover appropriate word classes (nouns, verbs, and adjective) as well as the type of news contents (headlines, first paragraph, last paragraph) in learning the model for classifying emotion label for news.

2. Related works

In a classification process, text analysis is an important preprocessing steps that leads to good features and then to finally building a good classifier. The importance of understanding text types in classification is presented in [1]. It is obvious that different text types need certain fine tuning to achieve better classification results. For example, verb analysis has been used for review text [2]. A similar work is presented by [3], whereby various types of textual representation such as named entities, noun phrases and bags of words has been carried out for financial news text sentiment analysis. In [4] and [5], the effects of adjective words are studied based on their positive and negative orientation.

The research on news classification has started as part of the text classification problem, where news collections like Reuters-21578 [6], TagMyNews [7], BBC Datasets [8] and Yahoo News Feed [9] have been used as main text categorization dataset. In these data sets, pre-labelled data based on topic-based categories like sports, entertainment, technology, health etc. are provided for learning and evaluating various processed involved in text classification, especially text preprocessing and learning [10] approaches. Reviews presented in [11] shows that micro averaged breakeven point up to 0.920 has been achieved on one of the learning approaches based on topic related articles.

Nevertheless, as the need for more facets of classification arises, e.g. to classify by personal needs, to classify by user general sentiment preference (like/dislike rather than topic), to classify by popularity (most read) and lately to classification by emotional responses, the newer trend of researches extended to these areas of news classification. For example, [12], [13] and [14] focus on modelling and personalization based on user preferred categories, whereas [15] focuses on collaborative effort in similar goal of personalization. Classification of emotional responses in news by readers were presented by [16], [17], [18] and [19].

As the influences of users' feedback on social media are getting more prominent, new type of classification trends can be seen in texts, and one of this trend is classification based on text polarity. Classification by sentiment is popularly performed on opinionated texts, e.g. reviews, whereby a review or some discussed entities can be classified by whether it is liked or disliked/good or bad based on some analysis of contents polarity. In [20] news related posting in tweets form are classified based on user's like and dislike, whereby [21] reinforces the news contents with tweets.

Although news texts are highly related to sentiments and emotions, nevertheless, classification of news by sentiment is challenging as the contents itself is often objective. For example, a news with statement "the bus caught fire", "9 people dead" are linked to a negative sentiment although the text does not contain opinionated keywords, but merely reporting facts. As news texts may not correlate to sentiment-based classes directly, we need to examine whether certain aspects of news texts could result in better classification. In the next section, we explain the research methodology of this study.

3. Sentiment-based news classification

3.1. Background

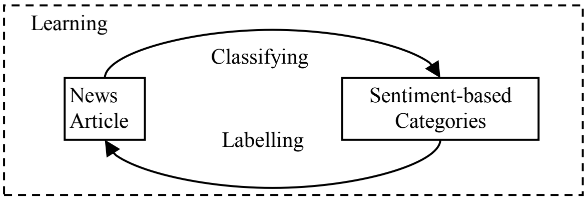


Fig. 2. Relations between labelling, learning and classifying.

In general, text classification involves three main processes, i.e. labelling, learning and classifying. In Fig. 2, we show the relationships between these processes in the context of sentiment-based news classification, where learning is depicted as the background process supporting both labelling and classifying processes for news article and sentiment-based categories. In the labelling process, news article is annotated with sentiment-based categories. The result of this process would be a set of labelled articles. Normally, for domain specific classification problem, labeling or annotating task are carried out by expert of the domain. However, recent trend of crowd sourcing or social feedbacks has resulted in huge number of articles labelled by end users, especially for reviews and news texts. These labelled articles are then used in the machine learning process to generate classification models.

3.2. News text classification

News text classification involves steps as depicted in Fig. 3. The labeled article first goes through text processing step, i.e. POS analysis and position analysis. The processed text tokens are also filtered and cleansed using stop words. From there, relevant features are selected based on the two criteria, i.e. text POS and text position, for the supervised learning process. This creates a classification model that will be used to classify the new unlabeled article with the most relevant label. The outcome will be evaluated using standard performance metrics.

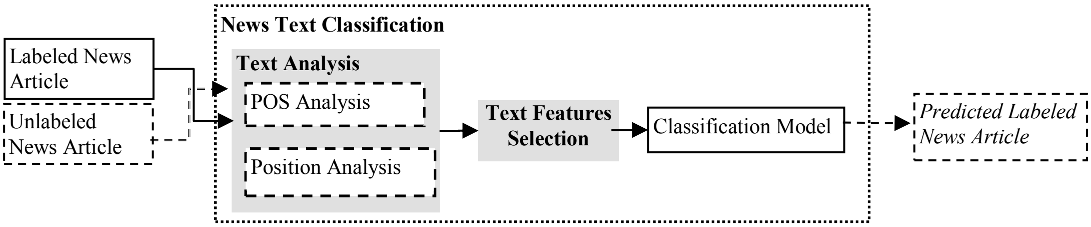
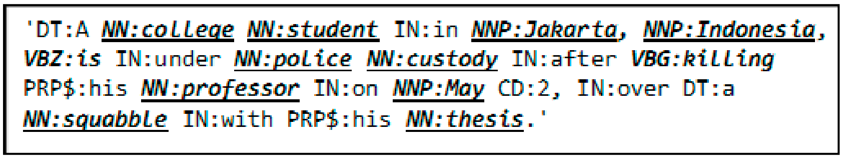


Fig. 3. The Process of News Text Classification

3.2.1. Text part of speech analysis

For a given news article, the news content is parsed using the Stanford NLP parser to obtain the part of speech tagging. Fig. 4 shows an example of a tagged headlines. In this research, three word-classes are used to investigate their impact in influencing the overall sentiment engaged to the reader. These word classes are nouns, verbs, and adjective. In other words, not all words in the article are useful for the training and testing purpose.



'DT:A NN:college NN:student IN:in NNP:Jakarta, NNP:Indonesia,
VBZ:is IN:under NN:police NN:custody IN:after VBG:killing
PRP$:his NN:professor IN:on NNP:May CD:2, IN:over DT:a
NN:squabble IN:with PRP$:his NN:thesis.'

Fig. 4. A tagged headline.

3.2.2. Text position analysis

Apart from word classes, the research also considers if the position of the text can affect the sentiment of the reader toward the news article. There are a few parts that can be considered from a news article. At this point, the parts are news title, first paragraph and last paragraph. The title of a news report or also known as news headline often summarizes the content of the news report. This means that the news headline, in fact, has a high relationship to its report content. Previous research conducted found out that the polarity of the title is generally consistent with the polarity of the whole news text report. Meanwhile, the first paragraph of a news can also be delivering the summary of the text, making it similar to contents of the title as it often highlighted the main point. In most cases, it is the part that has the highest relativity to the news content as it often describes the main idea of the whole content. However, they may be neutral in some cases. On the other hand, the last paragraph is typically describing more about the event in a detailed contain a summary of the topic or event.

3.2.3. Classification process

In terms of classification approach, this research proposed methodology to adopt supervised learning technique which in this case the k-Nearest Neighbour classification. It assumes that when given a set of instances in a training set, the class of an uncertain occurrence is more likely to be classified into the instance of the majority of its k closest "neighbour" belongs to from the training set. The classification process is then conducted using the classifier available in the WEKA tool [22]. In this research, the selected classifier is KNN approach. The learning and testing is conducted with 10 folds cross-validation, meaning that 90 percent of the data is used for training and the remaining 10 percent will be used for testing.

4. Evaluation and discussion

4.1 Data benchmarking

In this evaluation, we prepare our own annotated data set as lack of available benchmark data sets that capture the sentiment-based category for news corpus. The data set of English news articles is obtained from The Star online, which is one of the popular online news platform in Malaysia. The articles extracted come from four different domains which are World, Nation, Sport and Regional. For each article, different sentiment categories are available for the readers to vote via the sentiment voting system. The six sentiment labels are Happy, Inspired, Sad, Amused, Angry and Annoyed. The percentage of the vote reflected the general reaction of the readers to the specific news event. In addition, the captured news text is also segmented into parts like news headline or news title, introduction news body and conclusion of news body.

As each news article may have been voted for more than one labels, with different number of votes, only significant labels are selected for the classification process. The selection is based on the following rules.

Rules of Label(s) Selection

Given the set of sentiment labels as S_i where i = {happy, inspired, amused, sad, angry, annoyed}, different rules are applied by comparing the percentage votes recorded among label S_i. Let V be the S sorted by number of votes, where votes for $V_i \geq V_{i+1}$.

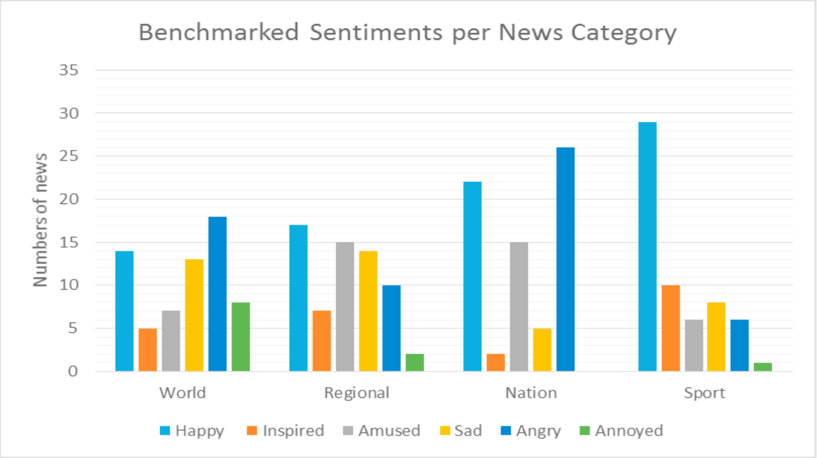


Fig. 5. Breakdowns of data sets by labels and news type.

Rule i. If vote of V_i is equal or higher than threshold:

a. V_i is chosen when it has the maximum value.

b. V_{i+1} is also chosen if it both sentiments are meeting the same threshold value which is 50%.

***Rule ii. If vote of V_i is lower than threshold but highest among the six sentiments**:*
Two labels, V_i and V_{i+1} will be selected for each of the cases below

a. $V_i = V_{i+1}$.

b. Difference between V_i and V_{i+1} is not greater than 10% and value of V_{i+1} is larger than the others.

c. If $V_{i+1}=V_{i+2}$ ..., the polarity of the respective sentiments will be considered as the selection criteria. If V_i is of positive orientation, the second label chosen must be positive as well.

A total of 260 news texts are collected across different news type and a bar chart is plotted to show the sentiment distribution of the data set after the Label Selection process to create annotated news (see Fig. 5). In general, the chart shows that Sport News is mainly happy and inspiring news as the numbers are scoring the highest compared to others. Meanwhile, Nation news and World news tend to stimulate anger while the "Angry" emotion topped the chart for these two categories. On the other hand, Regional and Nation news have a relatively comparable amount of "Amused" news.

4.2 Performance metric

Precision and recall are used as performance measures to determine if the news classified to a sentiment is accurate. The classification of the news sentiment adopts the four cases of classification i.e. true positive (TP), false positive (FP), true negative (TN) and false negative (FN) as shown in Table 1. Given a piece of presumably happy news report, the predicted result can appear as follows:

Table 1: Cases of TP/TN/FP/FN for sad/happy news.

Predicted case Actual case	-ve case: Sad	+ve case: Happy
-ve case: Sad	True negative	False positive
+ve case: Happy	False negative	True positive

As in [19], precision is given as the fraction of predicted cases (positive and negative) which are correct. On the other hand, recall is defined as fraction of positive cases that are correct classified.

4.3 Results and discussions

The evaluation of sentiment-based news classification is carried out in two manners, first to study the effect of different word classes and news contents in classification involving sentiment type of categories. Second, as the category concept is a great influence to classification, we also further evaluate whether the intended six sentiment-based categories are able to differentiate the classified news. Hence, we divide the evaluations into the original six sentiments categories (Section 4.3.1) and two categories based on polarity by grouping three positive sentiments and three negative sentiments (Section 4.3.2).

4.3.1 Sentiment-based category

The findings depicted in Table 2 shows that among the three word-classes, higher F-score is obtained while training the nouns (0.302) rather than verbs (0.208) and adjectives (0.209). This outcome is generally not as convincing as nouns are usually terms without emotion orientation. However, it is not impossible for the learning process in relating the occurrence of the specific words to the sentiment learned within the news text collected. For example, when come across with the noun "terrorist", this may classify the news into angry emotion. Meanwhile, the findings also suggesting that emotion classification of news using adjective is much prominent compared to verbs. This indicates that usage of verbs in classification may need further treatment during to pick up significant verbs.

In terms of text position, it is noticed that the headlines indicate greater importance in producing better result for classification. In this set of experiments, the word class of the news instance is not significance as it is focusing more on the content of the text in different position of the set of news to be studied. From the evaluation, the news headlines show the best results (0.422) among all the content of different text position trained and tested, followed by the first paragraph (0.321). This may be due to the fact that both text position has close word relevancy in its content. On the contrary, the last paragraph of news text has lowest performance among all the metrics. Often, contents obtained from the last paragraph is less relevance to the detailed description on events of the news reported. Instead, it is typically elaborating on general information to be conveyed to the readers. The low magnitude recorded for the classification process maybe due to several reasons. One of the possible factors affecting the accuracy of the classification model is the size of data sets and the implementation of specific classes for training and testing of the data. Increasing the number of data sets could be a solution to improve the accuracy.

Table 2. Classification results for sentiment-based category by word classes.

Word Class	True Positive	False Positive	Precision	Recall	F-measure
Noun	0.367	0.293	0.342	0.367	0.302
Verb	0.298	0.290	0.241	0.298	0.208
Adjective	0.322	0.312	0.371	0.322	0.209

Table 3. Classification results for sentiment-based category by news contents.

News Content	True Positive	False Positive	Precision	Recall	F-measure
Headline	0.449	0.235	0.449	0.449	0.422
First Paragraph	0.376	0.276	0.353	0.376	0.321
Last Paragraph	0.322	0.319	0.320	0.322	0.214
First + Last Paragraph	0.335	0.311	0.306	0.335	0.230
Headline + First Paragraph	0.404	0.267	0.374	0.404	0.358
Headline + Last Paragraph	0.380	0.292	0.442	0.380	0.285

4.3.2 Polarity-based Category

For polarity-based evaluation, the six sentiments are grouped by sentiment polarity, i.e. positive (for Happy, Inspired and Amused) or negative (for Sad, Angry and Annoyed). From the evaluation, findings (in Table 4 and Table 5) show that broader class has higher accuracy in classification of the datasets compared to when it is focusing on specific scope of sentiments. For example, usage of adjectives records a relatively high F-measure with 0.844 whereas headline records similar performance as well with 0.837. This result is expected as it is often hard to someone to differentiate between categories like happy, inspired and amused for a positive sentiment. Therefore, there could be inconsistency in the voting when people give feedback on these specific kind of feelings as it is more abstract. Similar situation happens for the categories like angry, sad and annoyed. However, when these categories are grouped by polarity, the voting should be more consistent as it is easier to differentiate between something one likes and dislikes, hence, contributing to a better accuracy of classification results. Next, the overall performance recorded for the effect of text position in polarity-based category are consistent with the results recorded in sentiment-based category, in which the content of Headline plays critical role in classification with the highest precision and recall compared to other content elements (both Table 3 and Table 5).

Table 4. Classification results for polarity-based category by word classes.

Word Class	True Positive	False Positive	Precision	Recall	F-measure
Noun	0.686	0.613	0.714	0.686	0.588
Verb	0.751	0.398	0.743	0.751	0.743
Adjective	0.853	0.268	0.864	0.853	0.844

Table 5. Classification results for polarity-based category by news contents.

News Content	True Positive	False Positive	Precision	Recall	F-measure
Headline	0.837	0.203	0.837	0.837	0.837
First Paragraph	0.792	0.359	0.796	0.792	0.775
Last Paragraph	0.682	0.633	0.785	0.682	0.568
First + Last Paragraph	0.682	0.633	0.785	0.682	0.568
Headline + First Paragraph	0.727	0.507	0.739	0.727	0.677
Headline + Last Paragraph	0.686	0.619	0.740	0.686	0.583

Several limitations have been observed during the evaluation. For example, the votes obtained from the sentiment voting system may not be comprehensive at the time when the data is taken as more votes could be generated later. Nevertheless, we have tried our best to filter news that do not record any votes (or extremely low votes) from the readers. Lastly, as the background of readers are unknown, the quality of dataset may also be compromised.

5. Conclusion

In conclusion, the outcomes suggest the word classes and news contents do play significant roles in affecting the news classification process even though it recorded a generally low average weighted score. According to the results obtained, it can be concluded that the most influential word class in classifying the news is adjective as it recorded a steady performance while different class features are being used for classification. As for the text position, we can conclude that headlines and first paragraph of the news are comparably important in classifying news by their sentiments. However, this is an early verdict made according to the findings of the relatively small test size. The low accuracy of the classification result may be due to several reasons such as the type of categories selected and the size and source of the datasets. With respect to the research impact, our findings could benefit of both society and government considering the fact that sentiments play a big part in decision making. With people tends to share out their thoughts and expression on the things that they concerned or interested in, there is a demand in the field of media monitoring for ensuring to understand the underlying emotions from different perspective such as political, economic and social. For example, the predicted results can also bring great influence in planning national policy as political news sentiments can reflects the public responses on governments' intention to implement new laws or policy. Lastly, classification of news based on specific sentiments can be made available as predictions. This is believed to provide an interesting way of browsing news in return whereby one can expect to see news article being categorized into various types of emotion such as happy, sad, and angry and many more.

Acknowledgements

This research was supported by USM Research University Grant (1001/PKOMP/811335: Mining Unstructured Web Data For Tour Itineraries Construction), Universiti Sains Malaysia.

References

[1] Andrew Dillon and Cliff McKnight (1990) "Towards a Classification of Text Types: A Repertory Grid Approach" *Int. J. of Man-Machine Studies* **33(6)**: 623-636.

[2] Mostafa Karamibekr and Ali A. Ghorbani (2012) "Verb Oriented Sentiment Classification" in *Proc. of WI-IAT* p. 327-331.

[3] Robert P. Schumaker and Hsinchun Chen (2009) "Textual Analysis of Stock Market Prediction using Breaking Financial News: The AZFin Text System" *ACM TOIS* **27(2)**:1-19.

[4] Vasileios Hatzivassiloglou & Kathleen McKeown (1997) "Predicting the Semantic Orientation of Adjectives" in *Proc. of the 35th ACL*

[5] Vasileios Hatzivassiloglou and Janyce M. Wiebe (2000) "Effects of Adjective Orientation and Gradability on Sentence Subjectivity" in *Proc. of the 18th Conf. on Computational Linguistics* p. 299-305.

[6] Chidanand Apt, Fred Damerau and Sholom M. Weiss (1994) "Automated Learning of Decision Rules for Text Categorization" ACM TOIS.

[7] TagMyNews, http://acube.di.unipi.it/tmn-dataset/, Last accessed: 8 Feb 2017.

[8] Derek Greene and Padraig Cunningham (2006) "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering" in *Proc. ICML 2006* p. 377-384.

[9] R10 - Yahoo News Feed dataset, version 1.0 (1.5TB), http://webscope.sandbox.yahoo.com/#datasets, Last accessed: 8 Feb 2017.

[10] Robert Cooley (1999) "Classification of News Stories Using Support Vector Machines" in *IJCAI'99 Workshop on Text Mining*.

[11] Fabrizio Sebastiani (2002) "Machine Learning in Automated Text Categorization" *ACM Computing Surveys* **34(1)**: 1- 47.

[12] Daniel Billsus and Michael J. Pazzani (1999) "A Hybrid User Model for News Story Classification" in *Proc. of UM'09* p. 99-108.

[13] Chee-Hong Chan, Aixin Sun and Ee-Peng Lim (2001) "Automated Online News Classification with Personalization" in *Proc. of the 4th Int. Conf. of Asian Digital Library* p. 320-329.

[14] Talia Lavie, Michal Sela, Ilit Oppenheim, Ohad Inbar and Joachim Meyer (2010) "User Attitudes Towards News Content Personalization" *Int. Journal of Human-Computer Studies* **68(8)**: 483–495.

[15] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram (2007) "Google News Personalization: Scalable Online Collaborative Filtering" in *Proc. of the 16th Int. Conf. on World Wide Web* p. 271-280.

[16] Kai Gao, Hua Xu and Jiushuo Wang (2014) "Emotion Classification Based on Structured" in *Proc. of Int. Conf. on Multisensor Fusion and Information Integration for Intelligent Systems* p. 1-6.

[17] Yuxiang Jia, Zhengyan Chen and Shiwen Yu (2009) "Reader Emotion Classification of News Headlines" in *Proc. of NLP-KE*.

[18] Feng Liangzu, Li Ruifan and Zhou Yanquan "Extracting Key Sentiment Sentences from Internet News via Multiple Source Features" in *Prof. of the 4th IEEE International Conference on Network Infrastructure and Digital Content* p. 126-130.

[19] Kevin Hsin-Yih Lin, Changhua Yang and Hsin-Hsi Chen (2008) "Emotion Classification of Online News Articles from the Reader's Perspective" in *Proc. of WI-IAT '08* p. 220-226.

[20] Lu Weilin and Gan Keng Hoon (2015) "Personalization of Trending Tweets Using Like-Dislike Category Model" in *Proc. of KES-2015, Procedia Computer Science* p. 236-245.

[21] Orhan Demirsoz and Rifat Ozcan (2016) "Classification of News-Related Tweets" *Journal of Information Science*.

[22] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). "The WEKA Data Mining Software: An Update" *SIGKDD Explorations* **(11)1**.