Modern Management based on Big Data II and Machine Learning and Intelligent Systems III
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210228

Opinion-Aware Retrieval Models Based on Sentiment and Intensity of Lexical Features

Mohammad BAHRANI [a,1], Thomas ROELLEKE [a]
[a] *School of Electronic Engineering and Computer Science, Queen Mary University of London*

Abstract. Sentiment analysis has received much attention in Information Retrieval (IR) and other domains including data mining, machine learning algorithms and NLP. However, when it comes to big data, incorporating sentiment of words into IR models becomes even more important, and as yet no widely accepted standard exists for this task. The contribution of this paper is a framework for quantifying term frequency (TF) variants with sentiments. We propose models derived from the strength of lexical features to improve sentiment-based ranking.

Keywords. Search and Analytics, Information Retrieval, Sentiment Analysis, Modern Ranking Models

1. Introduction and related work

Companies need to analyse customer's general feelings about their products. On the other hand, singular buyers want to know the sentiment of the product reviews before buying [1]. Wherefore the examination of sentiments would be beneficial for many applications specially in the field of big data. Researchers need to analyse the sentiment intensity over time to know about changes in rhetoric [2]. This would benefit analysts in companies, government and political departments that need to track emotions and attitudes. To date, sentiment analysis is mostly applied to the polarity (positive or negative) classification task. However, this may not be sufficient for many domains. Companies need to provide buyers with search engines that are able to retrieve top products based on user queries and sentiment analysis of reviews.

This paper proposes the research in the development of intensity-aware retrieval models in a generalizable framework. Sentiment analysis explores texts containing people's opinions, emotions and attitudes [3]. IR models take features such as term rareness, e.g. Inverse Document Frequency (IDF), into consideration to rank documents. However, they do not capture opinions in the retrieval process. For example, concerning movie reviews, the word *'good'* might occur nearly in every review, and from an IDF-point of view, it is not informative and selective. We expect a query such as *'good comedy movie'* to find

[1]Corresponding Author: Mohammad Bahrani, the school of Electronic Engineering and Computer Science, Queen Mary University of London, UK, Email address: m.bahrani@qmul.ac.uk.

good movies, but IR might process it similar to *'comedy movie'* due to the high frequency of the term *good* in the collection.
The main contribution of this paper is to address this problem by developing and investigating a sentiment-based framework for ranking reviews. This framework incorporates the sentiment and intensity of lexical features into IR units including the term frequency (TF) and IDF. Within this framework two different tasks are employed. Firstly, we add the IDF of sentiment-bearing words as a notion of rareness to the sentiment classification process. Secondly, we generalise IR models by proposing intensity-aware methods which take sentiment intensity into consideration. The idea is to regulate the term frequency by boosting weights of sentiment-bearing words. Such boosting is expected to overcome the problem caused by the rareness of these words with respect to IR models.
Foundations of IR models are expressed in [4]. Additionally, opinion-based retrieval has been studied in a range of publications including [5,6,7,8]. Work related to our approach considers the use of semantics e.g. lexicons in machine learning and IR. [9] discussed the importance of aggregating semantics in IR models. [10] investigated the distribution of the emotions within textbooks, which resulted in the development of a framework for sentiment classification (SenticNet). [11] proposed a sentiment-aware attention method which leverages a three-step strategy to boost the performance sentiment analysis regarding movie reviews. Furthermore, [12] used IR techniques to classify sentences by polarity. Recently sentiment-aware approached received attention in deep learning [13]. An example of these applications is the incorporation of sentiment lexicons into neural networks [14]. VADER (Valence Aware Dictionary for sEntiment Reasoning) is a parsimonious rule-based tool for sentiment analysis concerning social media texts developed by Hutto and Gilbert [15]. It leverages a combination of qualitative and empirical methods using human experts and judgmental evaluations. Moreover, it employs a rich intensity-based lexicon to assign sentiments to sentences.

2. Opinion-Aware TF

2.1. Opinion-aware TF_{total}:

We introduce micro and macro models for sentiment-based retrieval. These models are built upon VADER scores. Based on the VADER lexicon, a lexical feature could have a score between +1 and -1. Equation (1) shows how VADER calculates the compound sentiment of a sentence:

$$\text{sentiment}_{\text{vader}}(s) := \frac{\sum_{t \in s} \left(\sum_{i=1}^{n_L(t,s)} \text{W}_{\text{sentiment}}(t,i,s) \right)}{\sqrt{\left(\sum_{t \in s} \left(\sum_{i=1}^{n_L(t,s)} \text{W}_{\text{sentiment}}(t,i,s) \right) \right)^2 + \alpha}} \tag{1}$$

$\text{W}_{\text{sentiment}}(t,i,s)$ is the sentiment score of the i_{th} occurrence of term t, and s is a sentence, and α is a normalization parameter. The algorithm is reinforced by five heuristics including punctuation marks, capitalization, modifiers, negation and 'but' checker. It regulates the score of a word depending on the distance between the word and its degree modifier. The primary sentiment of a term is transformed into a sentence-dependent sentiment $\text{W}_{\text{sentiment}}(t,i,s) := \acute{\text{W}}_{\text{sentiment}}(t) + \text{seed}(i,s)$, where $\text{seed}(i,s) := \sum_{j=0}^{i-1} \alpha(j,i,s)$. For α, the cases are:

$$\alpha(j,i,s) := \begin{cases} \mathrm{W_f}(t_j(s), i-j) & \text{if } t_j(s) \text{ is an influencer (e.g. modifier)} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$\acute{\mathrm{W}}_{\text{sentiment}}(t)$ is the primary sentence-independent sentiment of term t, and $\text{seed}(t,s)$ is a weight to be added to this score in order to bring the heuristics into consideration. $\mathrm{W_f}$ estimates the weight of a single heuristic parameter in relation to the term by considering the distance and the constant weight of the parameter which is defined by VADER.
The intensity of a lexical feature is the sum of the absolute value of the sentiment and it's corresponding seed weight $\mathrm{W}_{\text{intense}}(t,i,s) := \left|\acute{\mathrm{W}}_{\text{sentiment}}(t)\right| + \text{seed}(i,s)$. Subsequently, to rank the documents using VADER, we use $\mathrm{RSV}_{\text{vader}}(d) := \sum_{s \in d} \text{sentiment}_{\text{vader}}(s)$.
The RSV shall be high if the document contains many positive opinion words. This is the rationale for the following opinion-aware TF variants:

Definition 1 (Opinion-Aware Total TF Variants)
Let d be a document, s a sentence, and t a term (word). Let $n_L(t,s)$ be the number of locations (positions) at which term t occurs in sentence s; the notation $n_L(t,d)$ applies to a document rather than a sentence.

$$\mathrm{TF}_{\text{total-sentiment-Macro}}(t,d) := \sum_{s \in d} \left(\sum_{i=1}^{n_L(t,s)} \mathrm{W}_{\text{sentiment}}(t,i,s) \right) \quad (3)$$

$$\mathrm{TF}_{\text{total-sentiment-Micro}}(t,d) := n_L(t,d) \cdot \acute{\mathrm{W}}_{\text{sentiment}}(t) \quad (4)$$

$$\mathrm{TF}_{\text{total-intense}}(t,d) := \sum_{s \in d} \left(\sum_{i=1}^{n_L(t,s)} \mathrm{W}_{\text{intense}}(t,i,s) \right) \quad (5)$$

Equation (3) shows the sentiment-based macro-TF $\mathrm{TF}_{\text{total}}(t,d)$, where $n_L(t,s)$ is the number of locations in which term t appears in sentence s and $\mathrm{W}_{\text{sentiment}}(t,i,s)$ returns the VADER score of the i_{th} occurrence of term t in the sentence.
In micro $\mathrm{TF}_{\text{total}}$, Equation (4), we determine the term frequency independently and then multiply the result by the corresponding primary sentiment. Therefore, this model does not consider the impact of degree modifiers.
The opinion-aware TF can also be adopted from intensity or force of lexical features. In this paper, we considered the combination of strength and the corresponding seeds to determine the intensity weight $\mathrm{W}_{\text{intense}}(t,i,s)$ as expressed in Equation (5).

2.2. *Opinion-Aware TF_{BM25}:*

A pivoted term frequency has been shown consistently to be important for retrieval (for the BM25 retrieval model, which can be viewed as a particular TF-IDF model). Therefore, we need pivoted term frequencies that are built upon sentiments in order to obtain a notion of the opinion-aware $\mathrm{TF}_{\text{BM25}}$. The following definition lists new pivoted tf variants (lower-case tf indicates the pivoted TF variants).

Definition 2 (Pivoted term frequencies)

$$\mathrm{tf}_{\mathrm{piv,sentiment-Macro}}(t,d) := \frac{\mathrm{TF}_{\mathrm{total-sentiment-Macro}}(t,d)}{(k_1(b\cdot \mathrm{pivdl}+1-b))} \tag{6}$$

$$\mathrm{tf}_{\mathrm{piv,sentiment-Micro}}(t,d) := \mathrm{tf}_{\mathrm{piv}}(t,d)\cdot \acute{\mathrm{W}}_{\mathrm{sentiment}}(t) \tag{7}$$

$$\mathrm{tf}_{\mathrm{piv,intense}}(t,d) := \frac{\mathrm{TF}_{\mathrm{total-intense}}(t,d)}{(k_1(b\cdot \mathrm{pivdl}+1-b))} \tag{8}$$

The rationale is as follows. A scaling of the total TF $\mathrm{tf}(t,d)$ is not advisable since this would have implications on the document length. A scaling of the $\mathrm{TF}_{\mathrm{BM25}}$ would just equate to a linear scaling of the TF.IDF weight. The determination of the opinion-aware $\mathrm{TF}_{\mathrm{BM25}}$ is presented in the following definition.

Definition 3 (Opinion-Aware $\mathrm{TF}_{\mathrm{BM25}}$ Variants)

$$\mathrm{TF}_{\mathrm{BM25-sentiment-Macro}}(t,d) := \frac{\mathrm{tf}_{\mathrm{piv,sentiment-Macro}}(t,d)}{\left|\mathrm{tf}_{\mathrm{piv,sentiment-Macro}}(t,d)\right|+1} \tag{9}$$

$$\mathrm{TF}_{\mathrm{BM25-sentiment-Micro}}(t,d) := \frac{\mathrm{tf}_{\mathrm{piv,sentiment-Micro}}(t,d)}{\left|\mathrm{tf}_{\mathrm{piv,sentiment-Micro}}(t,d)\right|+1} \tag{10}$$

$$\mathrm{TF}_{\mathrm{BM25-intense}}(t,d) := \frac{\mathrm{tf}_{\mathrm{piv,intense}}(t,d)}{\left|\mathrm{tf}_{\mathrm{piv,intense}}(t,d)\right|+1} \tag{11}$$

For the scientific study of intensity, in this paper we focused on opinion words. All of the proposed models consider the neutral terms within queries as stop-word (words to be ignored). However, future studies are needed to explore the integration of topical retrieval with opinion words.

3. Proposed Models

3.1. Term-Frequency – Inverse-Document-Frequency (TF.IDF)

The proposed TF.IDF consists of the well-known IDF as the notion of rareness and opinion-based term frequencies $\mathrm{W}_{\mathrm{TF.IDF-x}}(t,d) := \mathrm{TF}_{\mathrm{x}}(t,d)\cdot \mathrm{IDF}(t)$. x is a generic type which can be any of different forms of opinion-aware approaches including total – sentiment – Macro, total – sentiment – Micro, total – intense and the corresponding BM25 approaches. The Retrieval Status Value (RSV) is the sum of TF.IDF weights across document terms $\mathrm{RSV}_{\mathrm{TF.IDF-x}}(d,c) := \sum_{t\in d}\mathrm{W}_{\mathrm{TF.IDF-x}}(t,d)$.

3.2. Language Modelling (LM)

For LM, we need an approach that considers sentiment when estimating the within-document term probability $p(t|d)$ and the collection-wide term probability $p(t|c)$, respectively. For reflecting the fact that we apply negative values (because of the polarity), we introduce the notation $\pi(t|d)$ and $\pi(t|c)$. We hire opinion-aware term frequencies introduced in the previous section and incorporate them into the probabilities. Therefore, we determine the new parameters as follows:

$$\pi_{\mathrm{x}}(t|d) := \frac{\mathrm{TF}_{\mathrm{total-x}}(t,d)}{|d_{\mathrm{x}}|} \qquad (12) \qquad\qquad \pi_{\mathrm{x}}(t|c) := \frac{\mathrm{TF}_{\mathrm{total-x}}(t,c)}{|c_{\mathrm{x}}|} \qquad (13)$$

x is the type of opinion-aware model, d_{x} is opinion-based document length and c_{x} denotes collection length. The determination of the length parameters is dependent on the used model-type x. The calculation of opinion-aware LM would result in issues related to negative values within logarithm. To address this issue, we apply the logarithm to the absolute result of the division and deliver the polarity of term sentiment into the formula by the use of $\mathrm{TF}_{\mathrm{x}}(t,q)$ parameter which determines the sign. In LM, the TF quantification is for the *query*; in other words, on the TF-IDF side of IR, it is more straight-forward to generalise the TF regarding sentiment. Below we show the opinion-aware LM:

$$\mathrm{W}_{\mathrm{LM-x}}(t,d,q,c) := \mathrm{TF}_{\mathrm{x}}(t,q) \cdot \log\left(\left|\frac{(1-\sigma_d)\cdot\pi_{\mathrm{x}}(t|c) + \sigma_d \cdot \pi_{\mathrm{x}}(t|d))}{\pi_{\mathrm{x}}(t|c)}\right|\right) \qquad (14)$$

The document is ranked by dividing the smoothed version of the multinomial probability of the query given the document by the probability of the query in the collection. Therefore, the corresponding RSV is defined as $\mathrm{RSV}_{\mathrm{LM-x}}(d,q,c) := \sum_{t\in q} \mathrm{W}_{\mathrm{LM-x}}(t,d,q,c)$.

4. Experiments

The basic models process all of query-terms regardless of the sentiments, whereas the intensity-aware models only consider the intensity values of sentiment-bearing terms within queries and ignore neutral words. To perform the evaluation task, we used the IMDB-review collection as the primary dataset. For confirming the results, we used all 2000 DVD reviews exists in the Amazon Sentiment Dataset [16]. Each query set consists of 50 reviews which are correspondent to the query set label in terms of polarity. As an example, *'There are scenes which make you gulp with sudden emotion, and those which even put a smile on your face through ...'* is a snippet of a positive query that we used.
To evaluate the intensity-aware models and their corresponding basic models, we hired Mean Average Precision (MAP) and Reciprocal Rank as shown in table 1. All of the novel models provided higher MAP scores than the basic models for both query sets. The models were more effective for negative reviews than the positive ones concerning IMDB dataset whereas, they provided much higher scores for the positive set compared to the negative queries when applied on Amazon DVD reviews. $\mathrm{LM}_{\mathrm{intense}}$ and $\mathrm{TF.IDF}_{\mathrm{BM25-intense}}$ achieved the highest MAP and Reciprocal Rank values. Although the intensity-aware LM provided a higher MAP score than the macro version of the $\mathrm{TF.IDF}_{\mathrm{BM25}}$, the variance is extremely small.
we performed a query-based analysis to capture the statistics of queries which are more compatible with the novel models for both positive and negative query sets.
We calculated the average of APs for both basic and intensity-aware models separately and subsequently for each query set, plotted the distribution of the differences ordered by descending. As can be seen in Figure 1, the intensity aware models were more effective

Table 1. Evaluation of Intensity-Aware Retrieval Models: Intense methods worked better than their corresponding basic models.

Model	MAP				Reciprocal Rank
	pos	*neg*	*avg*		*total*
IMDB Reviews					
$\text{TF.IDF}_{\text{total-basic}}$	0.286	0.237	0.261		0.80
$\text{TF.IDF}_{\text{total-intense}}$	0.320	0.345	0.332	(+27.2%)	0.86
$\text{TF.IDF}_{\text{BM25-basic}}$	0.293	0.234	0.263		0.87
$\text{TF.IDF}_{\text{BM25-intense}}$	0.351	0.322	<u>0.336</u>	(+27.76%)	<u>1.00</u>
LM_{basic}	0.259	0.241	0.250		0.60
$\text{LM}_{\text{intense}}$	0.305	0.369	<u>0.337</u>	(+34.8%)	<u>1.00</u>
Amazon DVD Reviews					
$\text{TF.IDF}_{\text{total-basic}}$	0.288	0.111	0.199		0.826
$\text{TF.IDF}_{\text{total-intense}}$	0.478	0.115	0.296	(+48.74%)	0.895
$\text{TF.IDF}_{\text{BM25-basic}}$	0.280	0.125	0.202		0.886
$\text{TF.IDF}_{\text{BM25-intense}}$	0.471	0.138	<u>0.304</u>	(+50.49%)	<u>1.000</u>
LM_{basic}	0.281	0.090	0.185		0.750
$\text{LM}_{\text{intense}}$	0.465	0.145	<u>0.305</u>	(+64.86%)	<u>0.995</u>

Table 2. Sentiment Polarity Classification: Sentiment models had higher F1 scores than the baseline. Sentiment-aware macro $\text{TF.IDF}_{\text{BM25}}$ was more effective than other models.

Model	P@1000		R@1000		F1	
	pos	*neg*	*pos*	*neg*	*pos*	*neg*
IMDB Reviews						
$\text{sentiment}_{\text{vader}}$	0.893	0.724	0.0714	0.0579	0.132	0.107
$\text{TF.IDF}_{\text{total-sentiment-Macro}}$	0.905	0.764	0.0724	0.0611	0.134 (+1.5%)	0.113 (+5.6%)
$\text{TF.IDF}_{\text{total-sentiment-Micro}}$	0.898	0.762	0.0718	0.0610	0.132 (+0.0%)	0.112 (+4.7%)
$\text{TF.IDF}_{\text{BM25-sentiment-Macro}}$	0.953	0.804	0.0762	0.0643	<u>0.141</u> (+6.8%)	<u>0.119</u> (+11.2%)
$\text{TF.IDF}_{\text{BM25-sentiment-Micro}}$	0.943	0.801	0.0754	0.0641	0.139 (+5.3%)	0.118 (+10.3%)
$\text{LM}_{\text{sentiment-Macro}}$	0.926	0.773	0.0741	0.0618	0.137 (+3.8%)	0.114 (+6.5%)
$\text{LM}_{\text{sentiment-Micro}}$	0.919	0.780	0.0735	0.0624	0.136 (+3.0%)	0.115 (+7.5%)

on more than 95% of the queries compared to the basic approach. Figure 2 shows the positive correlation between the quality of the intensity-aware models and the ratio of the query intensity to the query length. Interestingly, the correlation is stronger for positive queries which shows that the polarity of the queries could impact effectiveness of the models. The Pearson correlation coefficient for positive query set is 0.21 and for the negative query set is estimated as weak positive.

Moreover, we applied the plain sentiment-based VADER as well as the proposed sentiment-aware IR models on the IMDB dataset containing 25000 highly polar reviews [17]. The dataset is divided equally into negative and positive parts. We used the data and their labels as the gold standard for this task. Top 1000 reviews retrieved by models are labelled as *pos* and accordingly top 1000 reviews from the bottom of the reversed

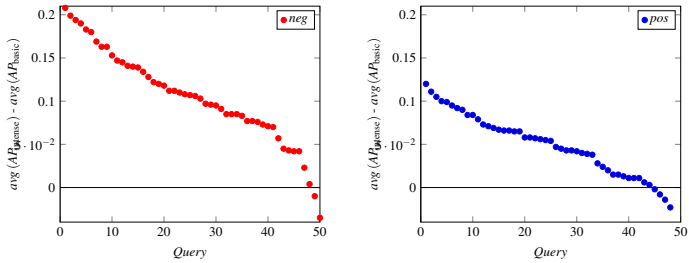

Figure 1. Distribution of Avg-AP Differences between Intense and Basic Models considering 100 Queries (in descending order): Query Analysis shows intense models were more effective for roughly 96% of queries compared to the basic models.

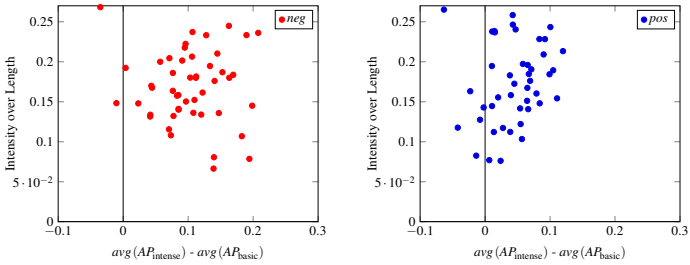

Figure 2. Pearson Correlation between Avg-AP Differences and the Ratio of Query-Intensity to Query Length: The correlation value for positive queries is 0.21 while the relationship between the parameters is weak positive regarding negative queries.

result-lists fell into the *neg* class. Table 2 column 4 lists F1 scores for our runs on the dataset. The data in this column indicates that all of the sentiment-aware models outperformed the baseline $\text{sentiment}_{\text{vader}}$ concerning both negative and positive classifications. The macro instance of the $\text{TF.IDF}_{\text{BM25}}$ achieved the highest score among the models. The advantage of the macro group is the consideration of influencers in retrieval. As we expected, they provided higher scores than the micro models although the differences are not statistically significant. In contrary with the baseline, the performance of the novel models was higher for the negative class compared to the positive one.

5. Conclusion and Future Work

In this paper, we presented two novel families of opinion-aware models, namely sentiment-aware and intensity-aware models to deal with the problem of the opinion words with low IDF and high intensity. This study also explored the consideration of a notion of IDF in sentiment classifications. To investigate the use of sentiment intensity in retrieval, we applied both basic and intensity-aware models to movie reviews and tested to find out if items of a specific polarity retrieve similar items. All of the intensity-aware models outperformed their corresponding basic models. It turned out that the effectiveness of the novel models is consistent across a range of positive and negative queries. The deficiency of the approach is a lack of advanced natural language understanding. Future work could improve the proposed framework by applying a Natural Language Pro-

cessing (NLP) tool that takes into consideration synonyms, antonyms and slang phrases. Moreover, further studies are required to enhance the quality of the determination LM parameters. This paper paves the way to achieving standards for considering sentiment in data. Those standards will be important for the many applications that process big data.

References

[1] Wawre SV, Deshmukh SN. Sentiment classification using machine learning techniques. International Journal of Science and Research (IJSR). 2016;5(4):819–821.

[2] Wilson T, Wiebe J, Hwa R. Just how mad are you? Finding strong and weak opinion clauses. In: aaai. vol. 4; 2004. p. 761–769.

[3] Liu B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. 2012;5(1):1–167.

[4] Roelleke T. Information retrieval models: foundations and relationships. Synthesis Lectures on Information Concepts, Retrieval, and Services. 2013;5(3):1–163.

[5] Zhang W, Yu CT, Meng W. Opinion retrieval from blogs. In: Silva MJ, Laender AHF, Baeza-Yates RA, McGuinness DL, Olstad B, Olsen ØH, et al., editors. Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007. ACM; 2007. p. 831–840. Available from: https://doi.org/10.1145/1321440.1321555.

[6] Gerani S, Carman MJ, Crestani F. Proximity-based opinion retrieval. In: Proceedings of the conference on research and development in information retrieval. ACM; 2010. .

[7] He B, Macdonald C, He J, Ounis I. An effective statistical approach to blog post opinion retrieval. In: Proceedings of the 17th ACM conference on Information and knowledge management. ACM; 2008. p. 1063–1072.

[8] Huang X, Croft WB. A unified relevance model for opinion retrieval. In: Proceedings of the 18th ACM conference on Information and knowledge management. ACM; 2009. p. 947–956.

[9] Bahrani M, Roelleke T. FDCM: Towards Balanced and Generalizable Concept-based Models for Effective Medical Ranking. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management; 2020. p. 1957–1960.

[10] Bisio F, Meda C, Gastaldo P, Zunino R, Cambria E. Sentiment-oriented information retrieval: Affective analysis of documents based on the senticnet framework. In: Sentiment analysis and ontology engineering. Springer; 2016. p. 175–197.

[11] Lei Z, Yang Y, Yang M. SAAN: A sentiment-aware attention network for sentiment analysis. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval; 2018. p. 1197–1200.

[12] Dragoni M. Shellfbk: an information retrieval-based system for multi-domain sentiment analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); 2015. p. 502–509.

[13] Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018;8(4):e1253.

[14] Wu C, Wu F, Liu J, Huang Y, Xie X. Sentiment Lexicon Enhanced Neural Sentiment Classification. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management; 2019. p. 1091–1100.

[15] Hutto CJ, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media; 2014. .

[16] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th annual meeting of the association of computational linguistics; 2007. p. 440–447.

[17] Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning Word Vectors for Sentiment Analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics; 2011. p. 142–150. Available from: http://www.aclweb.org/anthology/P11-1015.