Mixture of Language Models Utilization in Score-Based Sentiment Classification on Clinical Narratives

Tran-Thai Dang[1(✉)] and Tu-Bao Ho[1,2]

[1] Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa, Japan
{dangtranthai,bao}@jaist.ac.jp
[2] John Von Neumann Institute, VNU-HCM, Ho Chi Minh City, Vietnam

Abstract. Sentiment classification on clinical narratives has been a groundwork to analyze patient's health status, medical condition and treatment. The work posed challenges due to the shortness, and implicit sentiment of the clinical text. The paper shows that a sentiment score of a sentence simultaneously depends on scores of its terms including words, phrases, sequences of non-adjacent words, thus we propose to use a linear combination which can incorporate the scores of the terms extracted by various language models with the corresponding coefficients for estimating the sentence's score. Through utilizing the linear combination, we derive a novel vector representation of a sentence called language-model-based representation that is based on average scores of kinds of term in the sentence to help supervised classifiers work more effectively on the clinical narratives.

Keywords: Sentiment shifters · Language-model-based representation · Linear combination

1 Introduction

The clinical narratives reflect the patient's health status through observations of symptoms, progress in treatment, and physician's assessments. Therefore, determining such observations and assessments as positive or negative or neutral towards a disease plays an important role in therapeutic assistance and abnormality recognition.

The text in clinical narratives has several particular characteristics that pose some challenges for sentiment classification on such text. Beside lack of domain-specific resources, implicit sentiment mentioned in [1], we have to face with two main challenges as the following:

- The diversity of sentiment shifters used in clinical text.
- The shortness of clinical text.

H. Fujita et al. (Eds.): IEA/AIE 2016, LNAI 9799, pp. 255–268, 2016.
DOI: 10.1007/978-3-319-42007-3_22

Sentiment shifters are known as expressions used to change the sentiment orientation of a sentence such as negation words. The clinical text contains descriptions of patient's health status, to express the improvement of patient status, nurses or doctors often use the negation of symptoms and negative observations. However, the negation is in various variants not only negation words. For instance, we consider the following sentences/clauses:

- "There has significant improvement in pleural effusion." (positive)
- "There is no evidence of pleural effusion." (positive)
- "There has been marked decrease in right pleural effusion." (positive)
- "less nauseous than previous." (positive)

The example shows that the sentiment shifters are not only strong positive/negation words such as "improvement", "no" but also phrases like "less nauseous", or sequences of non-adjacent words as "decrease ... pleural effusion".

The problem of sentiment shifters was mentioned and solved by several methods on product-review domain. Such methods follow one of two main approaches, one is negation words and scope of the negation detection, the other is simple voting for overall sentence's sentiment score by word/phrase scores. The first approach often gives a better performance than the second one due to the intensive analysis of word contexts while the second one is more flexible because of the specific language independence. For our case, the first approach seems to be not effective because it is difficult to exactly capture all variants of sentiment shifters. Therefore, the second one is more appropriate, but it requires some modifications to enhance word's contexts considering instead of individually aggregating scores at word-level or phrase-level. For example, the word "improvement" is a strong positive word, so its score can dominate the other and rule the sentence's score while the word "less" may not due to a weaker positive sense. However, the phrase "less nauseous" with more positive purity volume can make a bigger influence on the sentence's score. It helps us raise an idea that the sentence score does not separately depends on word or phrase score. Thus, we simultaneously sum up word and phrase scores by a linear combination in which the coefficients characterize how words and phrases affect the sentence sentiment orientation. Besides, sequences of non-adjacent words are also used to capture more contexts of words. All words, phrases, sequences of non-adjacent words (terms) are extracted by using different language models.

The shortness of text requires a particular representation method instead of popular methods such as bag-of-words, bag-of-n-grams because the short length of text does not provide enough word co-occurrence or shared context for good similarity measures [11]. Through the idea of using the linear combination of different kinds of term extracted by their corresponding language models in estimating the sentence's score, it is clear to see that the sentence score depends on the score of each kind of term. Thus, that raises an idea of a novel vector representation for a sentence based on the average scores of such kinds of term called language-model-based representation to deal with the problem of representing short text. Different from the strategy of using topic model to enhance the co-occurrence of words in sentences for improving the similarity measures that are

based on the appearance of common words, language-model-based representation measures the similarity between two sentences by comparing the scores of such sentences according to each kind of term.

In this paper, we present two our contributions for sentiment classification on clinical narratives in case of lack of sentiment resources for medical domain:

- Effectively using a linear combination of different kinds of term extracted by various language models to estimate the sentence's score.
- Deriving a novel vector representation of a sentence called language-model-based representation to deal with the problem of short text representation.

2 Related Work

The techniques for sentiment classification on clinical text are mainly based on available techniques used in product review domain. In [21,24], the authors made a review of sentiment classification techniques that follow three approaches: machine learning-based approach, lexicon-based approach, hybrid approach. In machine learning-based approach, the feature set is determined through part-of-speech, n-gram, or sentiment words [23] before applying classification methods such as Naive Bayes, Support Vector Machine [22]. Besides, with a simpler way, in [2,3], the authors just summed up sentiment scores of words, phrases to estimate a sentence score for decision making.

In product review domain, sentiment shifters are mainly indicated via negation terms, so several works attempt to detect such terms, and the scope of negation in the sentence. In [4], Polanyi et al. described how the base attitudinal valence of lexical item can be modified by context and proposed a simple "proof of concept" implication for some context shifters. In other work, Li et al. [5] presented a shallow semantic parsing approach to learn the scope of negation. Ikeda et al. [6] proposed a method that models polarity shifters better than simple voting by sentiment word method. The effect of valence shifters on classification was examined in [7]. The parser and some heuristic rules was used to identify the scope of negation [8]. In [9], Li et al. proposed a feature selection method to generate scale polarity shifting training data, and a combination of classifiers to improve the performance. In [14], Kiritchenko et al. determined the sentiment of words in the presence of negation by detecting negation context via computing two scores of term in two parts: affirmative context, and negated context.

To improve similarity measures for short text, the probabilistic topic model is commonly utilized. LSA, pLSA, LDA have been widely applied to discover the latent topics for short text representation [10,11,15]. In addition, PMM-based classifier based on conditional probabilities of upcoming symbol given several previous symbols was applied for topic and non-topic classification [12]. Dai et al. [13] proposed cluster-based representation method named CREST to deal with the shortness and sparsity of text.

Several works also try using sentiment classification on clinical text, but on nurse/doctor narratives, the ontained results are not good enough. Ali et al. [16]

applied the methods such as Naive Bayes, SVM, Logistic-R to classify the posts in medical forums. Additionally, SVM and Naive Bayes were also used in [20] to determine the watchlist of drugs as positive or negative in drug surveillance. In [17], Deng et al. applied dictionary-based method to classify nurse letters, radiology reports in the MIMIC II database, they also presented some difficulties when doing classification on such data set. Besides, Na et al. [18] did clause-level sentiment classification using pure linguistic approach.

3 Mixture of Language Models Utilization in Sentiment Classification on Clinical Narratives

3.1 Our Approach

The core of our solution for sentiment classification on clinical text is to simultaneously sum up the score of words, phrases, sequences of non-adjacent words extracted by different language models by a linear combination. The linear combination is a simple and efficient model for voting sentiment score of the sentence with low computational cost that characterizes the importance of its components via the corresponding coefficients. Moreover, relying on such linear combination, we are able to derive a novel vector representation of a sentence called language-model-based representation. The proposed idea is formulated as the following.

Assume that $\mathbf{L} = \{L_1, L_2, ..., L_m\}$ is a set of m language models used to extract terms. $\mathbf{T} = \{L_1(s), L_2(s), ..., L_m(s)\}$ where $L_i(s), i = 1, 2..., m$, is a set of terms extracted from the sentence s according to the language model L_i. For each term $t \in L_i(s)$ compute $Score(t)$. An average score over all terms belonging to $L_i(s)$ is computed by the following equation:

$$Score(L_i(s)) = \frac{\sum_{t \in L_i(s)} Score(t)}{N_i} \tag{1}$$

where N_i is the number of terms in $L_i(s)$.

The sentiment score of the sentence s is defined as a linear combination over $Score(L_i(s))$ as the following:

$$Score(s) = \sum_{i=1}^{m} w_i \times Score(L_i(s)) \tag{2}$$

$$\begin{cases} Score(s) > 0 \Rightarrow positive \\ Score(s) < 0 \Rightarrow negative \end{cases}$$

In the linear combination, the coefficients $(w_1, w_2, ..., w_m)$ characterize how the sentence's score depends on each $Score(L_i(s))$. If the sentence's score is strongly related to a kind of term, its coefficient is larger, that means there is a bias for such kind of term. Besides, some kinds of term contribute to sentence's score identification with equal roles. Therefore, we pose three assumptions regarding the coefficient's values:

- Assumption 1: The value of coefficients $(w_1, w_2, ..., w_m)$ are different. That means there is a bias in the voting process.
- Assumption 2: The value of coefficients are equal, and set as 1.
- Assumption 3: That incorporates assumption 1 and assumption 2. There exists a subset of language models following assumption 1, and the rest is appropriate with assumption 2. In this case, the sentence's score is computed as the following:

$$Score(s) = \sum_{i=1}^{k} w_i \times Score(L_i(s)) + \sum_{i=k+1}^{m} Score(L_i(s)) \tag{3}$$

where k, $m - k$ are the number of language models following assumption 1, assumption 2 respectively.

Through the experiments and interpretations, we assess that if the components $Score(L_i(s))$ have a weak linear relationship, assumption 1 is more appropriate to obtain a better performance because in this case, there will has a conflict when aggregating such components, so we need to adjust the aggregation by a priority setting via adding the different weights for the components. Otherwise, in case such components have a strong linear relation that means we can use one of them to make the aggregation to make the decision, and we do not need to adjust them, thus assumption 2 is more appropriate. The detail and explanation are presented in Subsect. 4.2.

Equation 2 gives an idea of a vector representation for a sentence that is different from most of previous works using topics of words. In this equation, the sentence's score depends on the concurrent contribution of the components $Score(L_i(s))$, thus the set $\mathbf{S} = \{Score(L_1(s)), Score(L_2(s)), ..., Score(L_m(s))\}$ could be considered a feature set to represent the sentence that is called language-model-based representation. By such method, the similarity measure of two sentences is based on the comparison between the sentence's scores which are decomposed into the components $Score(L_i(s))$ instead of enhancing the co-occurrence of common words like using the topic model.

3.2 Proposed Method

Relying on the proposed approach mentioned above, we propose a method that includes three main steps: language-model-based terms extraction, sentiment score measure, and feature derivation and linear combination coefficients estimation.

Language-Model-Based Terms Extraction. In our work, language models including n-gram and skip-gram models play a role as templates in terms extraction. More general than n-gram models that help to extract sequences of adjacent words, skip-gram [19] models can capture not only sequences of adjacent words but also sequences of non-adjacent words. For example, we consider the following sentence:

"There is no evidence of pleural effusion."

Various language models such as unigram, bigram, trigram, 1-skip-bigram, 2-skip-bigram, 3-skip-bigram, 4-skip-bigram, 1-skip-trigram, 2-skip-trigram are used in this step. Table 1 shows an example of language model utilization for term extraction.

Table 1. Terms extraction by language models

Language model	Extracted terms
unigram	there, is, no, evidence, of, pleural, effusion
bigram	there is, is no, no evidence, evidence of, of pleural, pleural effusion
1-skip-bigram	there is, there no, is no, is evidence, no evidence, no of, evidence of, evidence pleural, etc.
2-skip-bigram	there is, there no, there evidence, is no, is evidence, is of, no evidence, no of, no pleural, etc.
trigram	there is no, is no evidence, no evidence of, evidence of pleural, of pleural effusion.
1-skip-trigram	there is no, there is evidence, there no evidence, is no evidence, is evidence of, is no of, etc.

As the definition in [19], k-skip-n-grams consider k or less skips to construct n-gram. For example, 3-skip-bigram includes 3 skips, 2 skips, 1 skip, 0 skips (bigram). Relying on number of tokens in terms, the language models are divided into three groups as the following:

- Group 1: occurrence of words individually (unigram)
- Group 2: co-occurrence of two words (bigram, 1-skip-bigram, 2-skip-bigram, 3-skip-bigram, 4-skip-bigram).
- Group 3: co-occurrence of three words (trigram, 1-skip-trigram, 2-skip-trigram).

Term's Sentiment Score Measure. Sentiment score of a term measures the related volume between the term and the sentence's sentiment label. We use the following equation to compute the term's sentiment score as in [3]:

$$Score(t) = \frac{p(t|positive) - p(t|negative)}{p(t|positive) + p(t|negative)} \tag{4}$$

$p(t|positive)$ is computed by taking number of times term t appears in positive sentences then dividing it by the total number of terms in the positive sentences. $p(p|negative)$ is also computed in the similar way. The term's score $Score(t)$ ranges from -1 to 1. If $Score(t) > 0$ the sentiment orientation of the term is likely positive, and vice versa.

Language-Model-Based Feature Derivation and Coefficient Estimation. As we mentioned in Subsect. 3.1, the simultaneous contribution of various kinds of term to the sentence sentiment orientation is characterized by a linear combination of their score as Eq. 2, in which each coefficient indicates how each kind of term gives its influence on the sentence score. Therefore, identifying such influence is equivalent to estimating such coefficient. We need to estimate coefficients in case of assumption 1, 3.

Algorithm 1. Linear combination coefficients learning

$\mathbf{L} = \{L_1, L_2, ..., L_m\}$ is a set of language models used.
for *each sentence s in training set* **do**
  vector := *empty*
  for *each* $L_i \in \boldsymbol{L}$ **do**
    Extracting a set of terms $L_i(s)$ in the sentence s according to L_i
    for *each term t in* $L_i(s)$ **do**
      Compute $Score(t)$ by Eq. 4
    Compute score average $Score(L_i(s))$ by Eq. 1
    Append $Score(L_i(s))$ to *vector*
if $\boldsymbol{L}$ *follows assumption 1* **then**
  Train with Support Vector Machine to to identify $(w_1, w_2, ..., w_m)$
if $\boldsymbol{L}$ *follows assumption 2* **then**
  Set $w_1 = w_2 = ... = w_m = 1$
if $\boldsymbol{L}$ *follows assumption 3* **then**
  if $\boldsymbol{L1} \subset \boldsymbol{L}$ *follows assumption 1* **then**
    Train with Support Vector Machine to identify coefficients
  if $\boldsymbol{L2} \subset \boldsymbol{L}$ *follows assumption 2* **then**
    Set the coefficients as 1

The most likely coefficients estimation is based on the training data. Each sentence in the training set is converted into the corresponding linear combination like Eq. 2, and then if the label of the sentence is positive the linear combination is greater than 0, and if it is negative, the combination is smaller than 0. For example, we assume that we convert n sentences in the training data into a set of inequalities as the following:

$$\begin{cases} s_1 : \sum_{i=1}^{m} w_i \times Score(L_i(s_1)) < 0 \\ s_2 : \sum_{i=1}^{m} w_i \times Score(L_i(s_2)) > 0 \\ ... \\ s_n : \sum_{i=1}^{m} w_i \times Score(L_i(s_n)) > 0 \end{cases}$$

We see that determining the most likely $(w_1, w_2, ..., w_m)$ is equivalent to finding a hyperplane as a linear boundary of a data set represented by the set of vectors $\{Score(L_1(s_k)), Score(L_2(s_k)), ..., Score(L_m(s_k))\}, k = 1, 2, ..., n$. Thus, this problem can be solved by using Support Vector Machine (SVM) technique. We propose algorithm 1 to for coefficients learning. In Algorithm 1, to determine which assumption $\mathbf{L}$ should follow, we base on assessment 2 presented in Subsect. 4.2.

Table 2. Coefficients assumptions with groups of language models investigation

Method		MIMIC II	Movie-Review
Our method			
1	Assumption 1 with group 2	**0.823**	**0.736**
2	Assumption 1 with group 3	0.69	0.507
3	Assumption 1 with group 1 + group 2	0.799	0.747
4	Assumption 1 with group 1 + group 3	**0.827**	**0.754**
5	Assumption 1 with group 2 + group 3	0.807	0.605
6	Assumption 1 with group 1 + group 2 + group 3	0.811	0.723
7	Assumption 2 with group 2	0.817	0.732
8	Assumption 2 with group 3	0.68	0.594
9	Assumption 2 with group 1 + group 2	**0.836**	**0.756**
10	Assumption 2 with group 1 + group 3	0.823	0.738
11	Assumption 2 with group 2 + group 3	0.813	0.723
12	Assumption 2 with group 1 + group 2 + group 3	0.832	0.751
13	Assumption 3 with group 1 + group 2 + group 3 (*)	**0.836**	**0.764**
Individually sum up term's scores of each language model			
14	Terms from unigram	0.827	0.747
15	Terms from bigram	0.769	0.688
16	Terms from trigram	0.579	0.464
17	Terms from 1-skip-bigram	0.799	0.709
18	Terms from 2-skip-bigram	0.81	0.717
19	Terms from 3-skip-bigram	0.812	0.721
20	Terms from 4-skip-bigram	0.818	0.727
21	Terms from 1-skip-trigram	0.644	0.556
22	Terms from 2-skip-trigram	0.678	0.599
Bag-of-words			
23	SVM + bag-of-words	0.698	0.503

(*): The sentence's score is computed by the following equation:
$Score(s) = \sum_{i=1}^{k} w_i \times Score(L_i(s)) + \sum_{j=1}^{h} Score(L_j(s))$
where $L_i \in$ group 1 and group 3, $L_j \in$ group 2.

4 Experimental Evaluation

4.1 Data Preparation

In the experiment, the MIMIC II data set that contains the information of more than 32,000 patients are used for our method evaluation. 6000 sentences that are manually annotated with two labels "1" (positive) and "–1" (negative) are obtained from "NOTEEVENTS" records.

For evaluation method, the annotated data is randomly divided into 10 parts then 6 parts are used for training, and the rest for testing. This process is repeated 10 times, then we take an average of precision.

We aim to build a classifier that can work well on clinical narratives in case sentiment resources for medical domain are not available, so the classification method should not depend on a specific domain. Therefore, to investigate whether our proposed method with the derived assessments is robust and can be applied on other data set or not, we additionally use movie review data[1] for evaluation due to some fairly similar points. The text in movie review data set is also separated into sentences/snippets (short text), and also contains some kinds of sentiment shifters like the MIMIC II data set.

In case of assumption 1 and 3, we use scikit learn, a python package implementing SVM algorithm with kernel functions[2] to determine coefficients.

4.2 Experiment Results and Interpretation

Coefficient's Assumption for Language Models of Groups. The experiments aim to determine which assumption is appropriate to a given language model. In the experiments, we consider the features generated from the language models in three groups and in the combination of such groups. All sentences are represented according to the language-model-based representation method. The classification results of three assumptions with three groups are showed in Table 2.

- A comparison between group 2 and group 3

We consider language models in the same group, and make a comparison between language models in group 2 and group 3. Line 1, 2, 7, 8 in Table 2 show that the features of group 2 provide remarkably higher performance than those of group 3 with both assumption 1 and 2. To explain why there is a significant difference between the features of group 2 and group 3, we visualize the training set and testing set in Fig. 1, then observe the distribution of data points.

We observe that the language models in a same group often generate their features with similar value, so the points in Fig. 1 almost fluctuate around the bisector $y = x$ with close distance.

Figures 1a and b show a difference of the points distribution between group 2 and group 3. The data points of group 2 tend to spread along the bisector while the data points of group 3 tend to converge at the corners. The reason is that sentiment orientations of terms extracted by language models of group 3 is almost pure with very high absolute value of score because the probability of co-occurrence of three words in a sentence is very small that gives poor information for prediction. In addition, the sentences in testing set are represented through the lexicon extracted from training set, so the terms of group 3 appearing together in a training sentence have a less chance to co-occur in the testing sentence that makes the testing set significantly different from the training set. In contrast to group 3, due to the higher probability of co-occurrence of two words, features of group 2 make our method get better accuracy. We also obtain

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/
[2] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

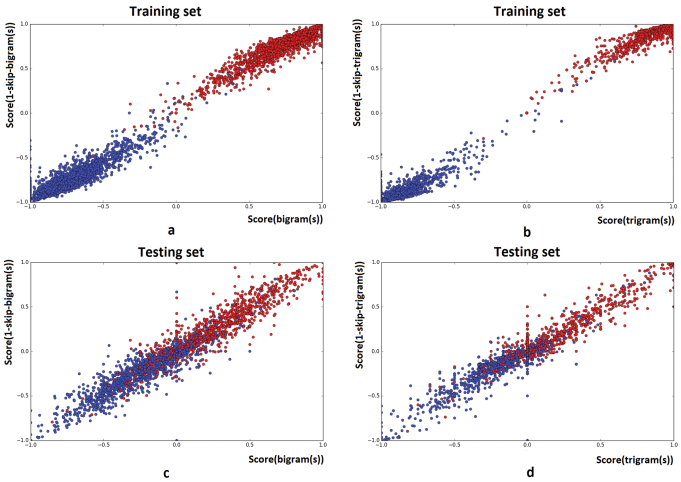

Fig. 1. A visualization of training and testing set with two features of group 2 ($Score(bigram(s))$, $Score(1 - skip - bigram(s))$) and group 3 ($Score(trigram(s))$, $Score(1 - skip - trigram(s))$). The blue, and red points indicate negative sentences, positive sentences respectively. a and c show the data points with the features of group 2, and b and d show the data points with the features of group 3. (Color figure online)

a similar result when doing classification on movie-review data set. Therefore, we have an assessment of using language models in a same group as the following:

Assessment 1: When building the feature set by language models in a same group, the language models considering the co-occurrence of two words provide better performance than ones considering the co-occurrence of three or more words.

- A comparison among different combinations of groups

Line 3, 4, 5, 6 show the accuracy when using assumption 1 with different combinations of three groups. We obtained the highest precision by incorporating language models of group 1 and group 3 (line 4), and get lower accuracy on other combinations. The quality of features depends on the linear relationship among them. If the features have a strong linear relation, there is less information to make the decision because they are considered as duplicated features, and the decision is just based on one of them. The volume of linear relationship between two features can be measured via correlation coefficient. In case the correlation coefficient is close to 1 or -1, the linear relation is strong. Table 3 shows the correlation coefficient of features generated by incorporating groups. For each group, we take a language model to generate the feature because other ones also generate the similar feature.

From Table 3, we observe that the features generated by language models of group 1, and group 3 have lowest correlation coefficient on both MIMIC and movie-review data that explains why such features get high performance of classification with assumption 1.

Table 3. Correlation coefficient among features generated by different combinations of three groups

Group pair	Correlation coefficient on MIMIC II	Correlation coefficient on movie review
Group 1 + group 2	0.901	0.893
Group 1 + group 3	0.844	0.837
Group 2 + group 3	0.977	0.983

Although the combinations of group 1 and group 2 or group 2 and group 3 do not produce the high performance with assumption 1 on the MIMIC data set and movie-review set, they get better results with assumption 2 (line 9, 11).

Through the results from line 1 to 12, we have an assessment to select the appropriate assumption for language models as the following:

Assessment 2: Assumption 1 is appropriate for language models whose generated features have a weak linear relation. In case such features have a strong relation, assumption 2 is more appropriate.

There has an interesting meaning inside this assessment. In case the features have a weak linear relationship, it will raise a conflict when aggregating, so we need a referee to judge which features are important then give such features a priority. In our method, the priority is characterized through the coefficients. Otherwise, if such features strongly linearly depend on each other, no conflict happens, so the referee is not necessary.

Line 13 shows the best result when we use assumption 3 with a combination of three groups, in which the features of group 1 and group 3 are aggregated with the different coefficients. We obtained 83.6 % on MIMIC and 76.4 % on movie-review data.

From line 14 to line 22, we show the results when using each language model to extract terms then make their score summing up. By this method, unigram has the highest performance (82.7 % on MIMIC and 74.7 % on movie-review), but it is not better than our method with assumption 3 that considers the interaction among terms extracted from different language models in voting for sentence's score.

Influence of Balance and Imbalance Training Set on Classification Performance. The experiment aims to examine the influence of balance and imbalance training data on the classification performance. A balance set contains an equal number of positive and negative sentences while a imbalance set is in contrast. The proportion between positive sentences and negative sentences impacts the term's score measure in Eq. 4. Table 4 shows how the proportion affects the classification performance.

Table 4. Influence of balance and imbalance training set on classification performance

	Method		MIMIC II	Movie-review
1	Sum up score (unigram)	B	0.827	0.747
2	Sum up score (unigram)	IB-P	0.805	0.72
3	Sum up score (unigram)	IB-N	0.813	0.726
4	Assumption 1 with group 2	B	0.823	0.731
5	Assumption 1 with group 2	IB-P	0.715	0.585
6	Assumption 1 with group 2	IB-N	0.783	0.582
7	Assumption 2 with group 1 + group 2	B	0.836	0.756
8	Assumption 2 with group 1 + group 2	IB-P	0.799	0.695
9	Assumption 2 with group 1 + group 2	IB-N	0.82	0.711

- B: Balance data set
- IB-P: Imbalance data set with greater number of positive sentences.
- IB-N: Imbalance data set with greater number of negative sentences.

Table 4 shows that imbalance sets make the accuracy reduce on both MIMIC and movie-review data set. The difference between number of positive sentences and negative sentences makes the term's score measure not fair, thus the scores are not precise.

5 Conclusion and Future Work

The paper presents our work on sentiment classification on clinical narratives. In this work, we propose a classification method to deal with two challenges of such text: the diversity of sentiment shifters, and the shortness of text. Our method uses a mixture of language models to extract terms, then estimate the sentiment score of sentences by a linear combination of such term's scores. In addition, we also derive a novel vector representation according to the language models used to extract terms that can work better on short text. Moreover, this method is flexible and independent with a specific language. The experimental results show the improvement of classification performance by using our method.

Beside the advantages, our method still has some drawbacks. The exist of sentiment shifters in training data makes the estimation of term's score sometimes is not precise. We also have to face with the problem of sparse data when using language models in group 3. Therefore, we plan to overcome these drawbacks to improve the performance of our method in the future work.

Acknowledgments. This work is partially funded by Vietnam National University at Ho Chi Minh City under the grant number B2015-42-02, and Japan Advanced Institute of Science and Technology under the Data Science Project.

References

1. Denecke, K., Deng, Y.: Sentiment analysis in medical settings: new opportunities and challenges. Artif. Intell. Med. **64**(1), 17–27 (2015)
2. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424 (2002)
3. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th International Conference on World Wide Web, pp. 519–528 (2003)
4. Polanyi, L., Zaenen, A.: Contextual valence shifters. In: Shanahan, J.G., Qu, Y., Wiebe, J. (eds.) Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series, vol. 20, pp. 1–10. Springer, Netherlands (2006)
5. Li, J., Zhou, G., Wang, H., Zhu, Q.: Learning the scope of negation via shallow semantic parsing. In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, pp. 671–679 (2010)
6. Ikeda, D., Takamura, H., Ratinov, L.-A., Okumura, M.: Learning to shift the polarity of words for sentiment classification. In: IJCNLP, pp. 296–303 (2008)
7. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Comput. Intell. **22**(2), 110–125 (2006)
8. Jia, L., Clement, Y., Meng, W.: The effect of negation on sentiment analysis and retrieval effectiveness. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 1827–1830 (2009)
9. Li, S., Lee, Sophia Yat Mei Chen, Y., Huang, C.-R., Zhou, G.: Sentiment classification and polarity shifting. In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, pp. 635–643 (2010)
10. Quan, X., Liu, G., Zhi, L., Ni, X., Wenyin, L.: Short text similarity based on probabilistic topics. Knowl. Inf. Syst. **25**(3), 473–491 (2010)
11. Song, G., Ye, Y., Du, X., Huang, X., Bie, S.: Short text classification: a survey. J. Multimedia **9**(5), 635–643 (2014)
12. Bobicev, V., Sokolova, M.: An effective and robust method for short text classification. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, pp. 1444–1445 (2008)
13. Dai, Z., Sun, A., Liu, X.-Y.: Crest: cluster-based representation enrichment for short text classification. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.) PAKDD 2013, Part II. LNCS, vol. 7819, pp. 256–267. Springer, Heidelberg (2013)
14. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. J. Artif. Intell. Res. **50**, 723–762 (2014)
15. Chen, M., Jin, X., Shen, D.: Short text classification improved by learning multi-granularity topics. In: IJCAI, pp. 1776–1781 (2011)
16. Ali, T., Schramm, D., Sokolova, M., Inkpen, D.: Can I hear you? sentiment analysis on medical forums. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 667–673 (2013)
17. Deng, Y., Stoehr, M., Denecke, K.: Retrieving attitudes: sentiment analysis from clinical narratives. In: Medical Information Retrieval Workshop at SIGIR, p. 12 (2014)
18. Na, J.-C., Kyaing, W.Y.M., Khoo, C.S.G., Foo, S., Chang, Y.-K., Theng, Y.-L.: Sentiment classification of drug reviews using a rule-based linguistic approach. In: Chen, H.-H., Chowdhury, G. (eds.) ICADL 2012. LNCS, vol. 7634, pp. 189–198. Springer, Heidelberg (2012)

19. Guthrie, D., Allison, B., Liu, W., Guthrie, L., Wilks, Y.: A closer look at skip-gram modelling. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006), pp. 1–4 (2006)
20. Chee, B.W., Berlin, R., Schatz, B.: Predicting adverse drug events from personal health messages. In: American Medical Informatics Association, pp. 217–226 (2011)
21. Madhoushi, Z., Hamdan, A.R., Zainudin, S.: Sentiment analysis techniques in recent works. In: Science and Information Conference (SAI), pp. 288–291. IEEE (2015)
22. Neethu, M.S., Rajasree, R.: Sentiment analysis in twitter using machine learning techniques. In: Computing Communications and Networking Technologies (ICC-CNT), pp. 1–5. IEEE (2013)
23. Veeraselvi, S.J., Saranya, C.: Semantic orientation approach for sentiment classification. In: Green Computing Communication and Electrical Engineering (ICGC-CEE), pp. 1–6. IEEE (2014)
24. Cambria, E.: Affective computing and sentiment analysis. IEEE Intell. Syst. **31**(2), 102–107 (2016)