ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Topic-based Historical Information Selection for Personalized Sentiment Analysis

Siwen Guo, Sviatlana Höhn and Christoph Schommer

University of Luxembourg - CSC, ILIAS Research Lab
6, Avenue de la Fonte, L-4364 Esch-sur-Alzette - Luxembourg

Abstract. In this paper, we present a selection approach designed for personalized sentiment analysis with the aim of extracting related information from a user's history. Analyzing a person's past is key to modeling individuality and understanding the current state of the person. We consider a user's expressions in the past as historical information, and target posts from social platforms for which Twitter texts are chosen as exemplary. While implementing the personalized model PERSEUS, we observed information loss due to the lack of flexibility regarding the design of the input sequence. To compensate this issue, we provide a procedure for information selection based on the similarities in the topics of a user's historical posts. Evaluation is conducted comparing different similarity measures, and improvements are seen with the proposed method.

1 Introduction

Personalized sentiment analysis is a challenging research area that is relatively under-studied compared to other tasks in sentiment analysis. Analyzing the individuality in expressions is crucial for understanding the writer's standpoint. Such an aspect has been mostly researched targeting product reviews [1, 2], where the texts are domain-specific with a distinct structure. In contrast, PERSEUS [3] focuses on texts from social platforms which are domain-independent, and employs methods that can be used for general social texts. PERSEUS analyzes historical information from the previous texts of a user, and applies a recurrent neural network (RNN) with long short-term memory [4] to facilitate the ability of looking into the past. Each node in the input layer of the RNN corresponds to a post of the targeted user at a certain time point. Additionally, attention mechanism is used on top of the RNN in order to generate more flexible representations at the output layer [5]. However, for users with various lengths of history, there can be problems with regard to the lengths of the input sequences in each training batch. Practically, people use 'padding' or 'bucketing' to handle this issue. The former sets the length of the input sequence to the maximal length observed in a corpus, and the shorter ones are padded with zeros. This method is not feasible in our task, because such a representation can be very sparse given the number of a user's posts ranging from a few to a few thousand. The latter groups the input sequences by ranges of lengths which can be seen as a relaxation of the former method, but zero padding is still needed depending on the size of each 'bucket'. Previously, PERSEUS chose the length of the input sequence empirically that was used for all the users — the same number of

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

historical posts are considered no matter how frequently the user posts. In this paper, we perform a selection technique that is specially developed for this task and provides a more flexible solution for this problem. The proposed approach is based on an extension of the assumption about opinions on related topics in [3]: We believe that the current opinion is affected more by a past opinion on related topics than on unrelated ones. To leverage this assumption, the relatedness of topics between posts of a user is analyzed by calculating the distance between the topic embeddings. With the algorithm, the network is able to take selected posts from the entire history of a user based on the similarity. Experimented with Twitter text generated by frequent users, the performance of the algorithm is compared using five distinct distance measures.

2 Personalized Sentiment Model PERSEUS

The main goal of PERSEUS is to capture the individuality in the expression from the previous texts of a user so that the user's current sentiment can be better predicted. In the system, each post is represented by a set of concepts[1], negations, topics and the user identifier. Negations are extracted given a set of terms; Topics are extracted from the texts based on grammatical rules, and each post can contain a small number of topics. However, only the explicit topics are concerned. In some cases, there are no topics can be extracted from the texts that convey a general sense of a status. For instance, the sentence *'Everything is so bright and shinny!'* can be a statement about an implicit topic *'life'* or *'cleaning'* based on interpretations. Moreover, the nature of social texts being highly informal makes the topic extraction a very challenging task. After a series of preprocessing steps, an embedding layer is applied to produce a representation for each post. Then, the representations of a number of posts from the same user are ordered by the creation time in order to generate one input sequence for the RNN. Since the length of an input sequence (the time steps T) is fixed, only the recent T posts are considered. For the users with history longer than T, the information before is lost. There are cases that the recent posts of a user are unrelated to the current one while related ones have appeared long before – the historical posts taken by the system only provide noise.

3 Historical Information Selection

The selection procedure is designed to overcome the problem with the information loss when a user has a history longer than the pre-defined number of time steps. All the previous posts of a user are considered in the process which provides the RNN a number of posts that are related to the current topics.

[1]We use SenticNet5: http://sentic.net/downloads/, last seen on February 15, 2019

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

3.1 Topic Embeddings

In order to compare the similarities between the topics, we apply a shallow neural network to generate a high-dimensional representation for the topics. The network takes concepts extracted from each post as input and topics of the post as output. The co-occurrences between the topics and concepts are analyzed so that two topics appear close to each other in the high-dimensional space if they are associated with similar sets of concepts. After training the network, the weights between the hidden layer and the output layer are used as the embeddings for the topics. Note that we differentiate our task of identifying topics or entities in the post with the topic modeling task as in [6].

3.2 Similarity measures

Five measures are concerned to calculate the similarity or distance between two sets of topics. Euclidean distance (**ED**) measures the straight-line distance between two terms; Manhattan distance (**MD**) measures the sum of the absolute differences of the coordinates between two terms; Cosine similarity (**CS**) measures the cosine of the angle between two terms. These three measures are calculated dimension-wise after finding the centroid of the topics in each set. The earth mover's distance (**EMD**) [7] measures the cost to transport a term to another. In our experiment, EMD is calculated in two ways that one is the same as the aforementioned measures while the other is to compute directly between both sets since it is capable of processing documents with different lengths. Furthermore, the word mover's distance (**WMD**) [8] is used as well, which is a special case of EMD implemented with GloVe word vectors [9]. When implementing EMD and WMD, euclidean distance is chosen as the ground distance.

3.3 The Selection Procedure

Algorithm 1 shows the selection procedure. The distance measures are as listed in the last section, while for the cosine similarity, a reverse of the value is used since the more similar two terms are, the closer they are in the vector space. Note that the preceding posts are used without a selection when the number of preceding posts by the same user is less or equal to the length of the input sequence, or when no topics can be extracted for the current post.

In the case that the number of the selected posts is smaller than the length of the input sequence, the algorithm takes the preceding posts of the current one according to the creation time. As shown in **Figure 1**, the user has 30 posts, and the ones that are close in time are chosen when no selection method is used. With the algorithm, 7 posts from the past – with similarity above the threshold – are selected while t_{-1} and t_{-3} are added as well to fill the empty slots in the input sequence. Therefore, the posts that are created recently with unrelated topics are dropped and accommodated with related posts generated further before. The selection procedure is executed more frequently with a smaller value of T, and the number of execution will keep growing while the user continues posting on the platform in the future — reshaping the model is not required.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Algorithm 1 Input Sequence Generation

Input: Corpus with attributes: [user, time, topic, content];
Distance measure $D(a, b)$ and the threshold for the measure B
Output: Input sequences X with shape:
[length of the corpus N, number of time steps T]
Initialization: $X[:][-1] = \text{corpus}[:][\text{'content'}]$, $k = 1$
for $i = 0$ to N **do**
  if corpus$[i]$['user'] = corpus$[i-1]$['user'] **then**
    $k = k + 1$
    if $k \leq T$ **then**
      $X[i][:-1] = X[i-1][1:]$
    else if corpus$[i]$['topic'] = [''] **then**
      $X[i][:-1] = X[(i-T+1):i][-1]$
    else
      distances = [D(corpus$[i]$['topic'], corpus$[i-j]$['topic'])
        for $j = 1$ to $k-1$]
      $l = min(len([m \textbf{ for } m \text{ in distances } \textbf{if } m \leq B]), T)$
      selected = $argsort$(distances)$[:l]$
      if len(selected)$\neq T$ **then**
        selected.$extend$([$i-n$ **for** $n = 1$ to T **if** $i-n$ **not in**
          selected][:$(T - len(\text{selected}))$])
      end if
      $X[i][:-1] = X[\text{selected}][-1]$
    end if
  else
    $k = 1$
  end if
end for

4 Experiments and Discussions

In the experiments, we use the Sentiment140[2] corpus as used before with PERSE-US. The corpus contains 122,000 tweets with frequent users who have at least 20 tweets. The corpus is split for training, validation and testing given preset timestamps. The topics are represented by vectors of dimension 100 (the size of the hidden layer) while the same size is chosen for the GloVe vectors. The thresholds of the distance measures are chosen individually by comparing their performance when applied in the system. Other settings remain the same with before as described in [3] without considering time gaps between the posts.

Table 1 shows the performance of the system. We can see that euclidean distance performs the best when T is 10, and Manhattan distance outperforms others when T is 20. It is unexpected that WMD has a performance that is not better than the Basic model where no selection is used. The reason can be that

[2]http://help.sentiment140.com/for-students, last seen on February 15, 2019

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

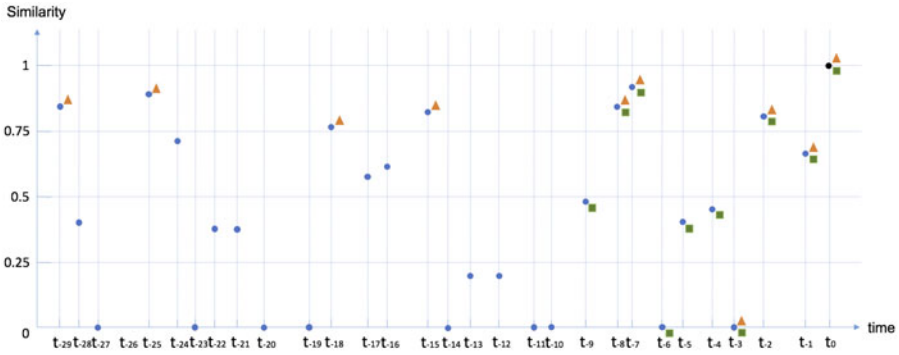

Fig. 1: An example of constructing an input sequence with the selection method (labeled by orange triangle) and without (labeled by green square). The current post is marked by the black dot at t_0, the number of time steps is 10, and the threshold of the similarity measure is set to 0.75.

Model	$T = 10$			$T = 20$		
	Pos. F1	Neg. F1	Accuracy	Pos. F1	Neg. F1	Accuracy
Basic	0.7362	0.7439	0.7401	0.7463	0.7566	0.7517
ED	0.7419	**0.7568**	**0.7496**	0.7473	0.7619	0.7548
MD	**0.7420**	0.7471	0.7447	0.7496	**0.7667**	**0.7585**
CS	0.7369	0.7480	0.7426	0.7472	0.7624	0.7550
EMDc	0.7379	0.7566	0.7476	0.7480	0.7619	0.7552
EMDt	0.7391	0.7525	0.7461	**0.7508**	0.7632	0.7572
WMD	0.7330	0.7444	0.7389	0.7454	0.7572	0.7515

Table 1: Performance of the system before (Basic model) and after implementing the selection technique with different distance measures for time steps 10 and 20. **EMDc** denotes the model in which EMD is calculated between the two centroids of the topics, and **EMDt** denotes the model in which EMD is calculated between the two sets of topics.

WMD takes external vectors while the topic embeddings for other measures are learned by considering the surrounding concepts which capture the affective information. There are no significant differences between other distance measures, however there are significant improvements between the Basic model and the method that provides the best results. This shows that considering the longer history by topic similarity has a positive effect on the performance of the system. It is also significant to increase the number of time steps from 10 to 20 so that more information can be related to by the RNN. Although the procedure is executed fewer times with $T = 20$, it is still effective implementing the selection method compared to the Basic model. Therefore, we believe that the selection procedure generally improves the performance, but the choice of distance measure used in the algorithm can vary depending on the value of T. Models with

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

smaller values of T can be more sensitive to the selection given the number of executions, while at the same time, greater improvements can be observed.

5 Conclusion and Future Work

In this paper, we introduce a selection method for personalized sentiment analysis that is able to relate to all the history of a user and prevent significant information loss. Experiments have shown positive results by leveraging the assumption that the historical posts with related topics have more impact on the current opinion than the ones with unrelated topics. The improvements provide more evidence on the significance of introducing individuality in sentiment analysis and the effectiveness of considering longer history. The approach offers flexibility for the scenario when the number of posts that a user publishes can not be anticipated, which aligns with the situation in reality. With the selection procedure, the generated input sequences span over longer time periods, which makes it more important to evaluate how information decays over time [3]. Thus, a more comprehensive evaluation of the whole system is planned for the future work. Moreover, a refining of the method for topic extraction can help analyze the implicit topics of the posts. This will potentially increase the number of executions of the selection method and further boost the performance. We also propose to develop a more sophisticated function for measuring topic similarities that takes into consideration the individuality, because the relativeness between topics can vary from person to person as well.

References

[1] Lin Gong, Mohammad Al Boni, and Hongning Wang. Modeling social norms evolution for personalized sentiment classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 855–865, 2016.

[2] Zhen Wu, Xin-Yu Dai, Cunyan Yin, Shujian Huang, and Jiajun Chen. Improving review representations with user attention and product attention for sentiment classification. *arXiv preprint arXiv:1801.07861*, 2018.

[3] Siwen Guo, Sviatlana Höhn, Feiyu Xu, and Christoph Schommer. Personalized sentiment analysis and a framework with attention-based Hawkes process model. In *International Conference on Agents and Artificial Intelligence*, pages 202–222. Springer, 2018.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] David M Blei and John D Lafferty. Topic models. In *Text Mining*, pages 101–124. Chapman and Hall/CRC, 2009.

[7] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, September 2009.

[8] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.