2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) | 978-0-7381-4259-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCWAMTIP51612.2020.9317342

A NON-PARAMETRIC MULTI-LINGUAL CLUSTERING MODEL FOR TEMPORAL SHORT TEXT

JAY KUMAR[1], RAJESH KUMAR[1], AMIN UL HAQ[1], SIDRA SHAFIQ[2]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China
[2]Department of Computer Science, The Women University Multan, Punjab, Pakistan
E-MAIL: jay@std.uestc.edu.cn, rajakumarlohano@gmail.com

Abstract:

Short text data is being continuously generated by many social streams such as Facebook and Twitter. Clustering the temporal text, data for the sake of identifying new topics, over huge volume of data has become very challenging task recently. Apart from supervised approaches, most of the existing clustering approaches assume that the input data belong to one language. Whereas, generally it has been observed that multi-lingual short text on social media exist in bulk amount. In this paper, we propose a model to cluster unknown number of topics in temporal environment for multi-lingual data. The proposed framework integrates non-parametric dirichlet model with language translation component (NDML) to cluster the temporal stream of short text data, and transforms the cluster feature into uniform language vector representation. We conducted experiments on real time crisis data to evaluate the accuracy of our proposed model.

Keywords:

Dirichlet model; Temporal data; Stream clustering; Multi-lingual Text

1. Introduction

During the past decade, social media and digital news platforms have become vital source of information for many global events or emergencies. A huge volume of text data is generated everyday containing critical events over time. The task of clustering such stream of text data has attracted attention of researchers in recent years. However, due to temporal dependencies of text data and unknown number of events make this task very challenging. Along with these constraints, generally a massive amount of generated text data belongs global events such as disaster, crisis and safety warnings thus contain multi-lingual information [1].

A series of models to deal with multi-lingual data has been proposed in the literature. Each model has its own benefits and disadvantage. [2] proposed a cross-lingual tweet classification M-BERT model with employing mixup manifold. However, it is necessary to pre-define the number of classes to train M-BERT for downstream task. In addition, class specific features highly rely while training phase. Usually previous models used original terms to fed the classification model, but [3] explored exploited semantic and statistical features to improve the classification accuracy on multi-lingual text data. Previous research on event discovery from news using evolution graph has been introduced [4], but it is not fit for short text [12]. Additionally, it needs a pre-defined number of classes to train the model. Most approaches are supervised and do not consider temporal dependency [5]. Where, [6] and [7] proposed short text clustering model, however their approach deal with monolingual text data. Therefore, there exist need a general framework to overcome mentioned issues together.

In this paper, we propose a framework which builds a bridge between clustering model and language translation component. We employ non-parametric dirichlet clustering model (NDML) to cluster the temporal short text data and build a bridge to identify the events based on change in probability distribution. The temporal data is splitted into fixed window, then dirichlet model update the existing model either by merging each instance into existing cluster or create a new cluster. The continuous updated model having active clusters is then processed to label the cluster related to different language using translation component. Furthermore, we exploit the word hypernym to expand the cluster features for performance improvement. The main contributions of our work are defined as follows.

- We propose a model to cluster multi-lingual temporal text data.
- Proposed approach can identify evolving number of events from generated text from social streams.
- We conducted empirical study to evaluate and prove the significance of our model.

2. Proposed Model

To solve the problem of identifying continuous arrival of events from temporal text data, we propose a framework (NPML), a non-parametric dirichlet model based clustering for multi-lingual data. Figure 1 shows our proposed architecture. In the initial step, streaming1 data is sliced into chunks. These chunks are either based on time-window or fix-sized window. The former type of chunk consider instances in time unit i.e., hour, day, or week. The latter type of chunk consider static number of instances for each window. Then, each chunk is fed into dirichlet model to group the short text instances. Within the existing model, each instance of chunk either merged into existing cluster or create a new cluster. Afterwards, each cluster' features are passed into translation process where clusters having same topics (based on similarity discussed in Section III) with different language are assigned with common label. We define each component in detail.

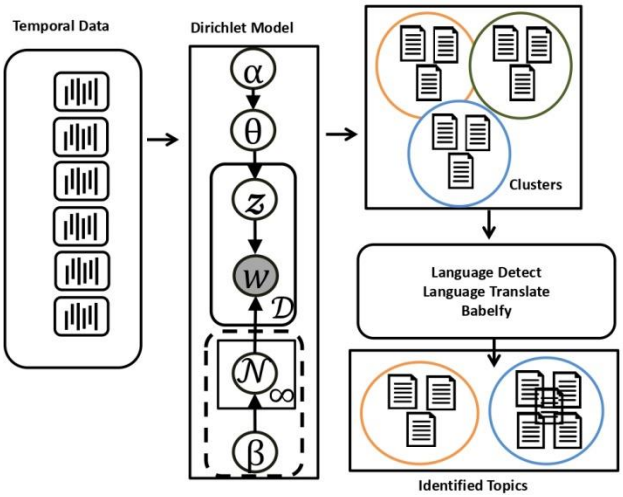

Fig.1 The proposed model for multi-lingual text clustering.

2.1. Non-parametric Dirichlet Model (NDM)

It is a type of stochastic model which models the random procedure of generated objects from a pre-defined distribution. A variety of dirichlet models have been proposed in the past, however, recently, [7] proposed a model based on Chinese Restaurant Process (CRP) to deal with temporal data to capture infinite number of topics. We exploit this model to cluster the incoming documents from temporal stream. This model is specifically designed for short text stream clustering task. The model specifies the probability to calculate incoming document and existing cluster, defined in Equation (1).

$$p\left(z_d = z \mid \vec{d}\right) = \left(\frac{n_z}{D-1+\alpha D}\right) \cdot \left(\frac{\prod_{w \in d}\prod_{j=1}^{N_d^w}\left(v_z^w\right)+\beta+j-1}{\prod_{i=1}^{N_d^z} l_z+(V\beta)+i-1}\right) \quad (1)$$

To create a new cluster, the defined probability is give n in Equation (2).

$$p\left(z_d = z_{new} \mid \vec{d}\right) = \left(\frac{\alpha D}{D-1+\alpha D}\right) \cdot \left(\frac{\prod_{w \in d}\prod_{j=1}^{N_d^w}\beta+j-1}{\prod_{i=1}^{N_d}(V\beta)+i-1}\right) \quad (2)$$

Here, n_z represents total number of documents in the cluster, N_d^w total unique terms in the document, D rep resents the current number of clusters in the model, V specifies the total vocabulary of current cluster, N_d is the length of document in terms of words, and v_z^w re present frequency count of term w in cluster z. The n otation α and β are model parameters. The incoming d ocument can create new cluster if probability in Equati on (1) is greater than Equation (2) for all existing clus ters.

2.2. Cluster Feature Set

The vector space model (VSM) is generally used for clustering task. However, to represent a cluster into subspace, we employ cluster feature set (CF) [8]. A CF is defined as Each cluster is represented by words of related documents. In our model, we followed [7] and define CF set as a 3-tuple $\left\{n_z, v_z^w, l_z\right\}$. Here, n_z represents number of documents in the cluster, d_z^w represents frequency of each unique term in cluster and l_z is total words in cluster.

2.3. Language Translation Component

We build a translation component by cascading detectlanguag[1], google-translate[2] and Babelfy[3]. The first library detects the language to transfer the input for translation into English language. Afterwards, Babelfy library is used to fetch hypernyms of key terms of cluster features. The extracted hypernyms of terms help to calculate

[1] https://detectlanguage.com
[2] https://translate.google.com
[3] https://babelfy.org

similarity between two different clusters of different language [9]. The reason to identify the language is to avoid calculating distance between clusters of same topics. We calculate both probabilities in Equation [1] and Equation [2] to calculate similarity to relate two clusters of same topics. Here, instead of using model vocabulary, V represent the cluster local vocabulary.

3. Experiments

This section discusses the experimental setup of evaluate our model. We define the process of constructing dataset. We also describe selected state-of-the-art algorithms to compare the performance with detail discussion. As we already mentioned that we employ NDM for clustering the short text, which relies on two parameters. Therefore, we also demonstrate the effect of these parameters in terms of three different evaluation measures.

3.1. Data set

For our empirical study, we downloaded SOSItalyT4[4], ChileEarthquakeT1, and CrisisLexT26 from the CrisisLex platform to construct a dataset. It contains 26 crisis or disaster events occurred during 2009 to 2014 around the world, hence documents are related to different languages. Along with removing the special characters, we deducted all duplicate instances from the individual datasets to reduce content redundancy. Table 1 shows the statistics of constructed dataset.

Table 1 Dataset Statistics

Language	Documents	Vocabulary	Topics
English	5171	18140	26
Italian	3239	11301	26
Spanish	3788	12133	26
Total	12198	41574	26

3.2. Models for Comparison

To compare with our model, we choose two supervised deep neural network model, M-BERT and FastText [10]. M-BERT is multi-lingual pre-trained BERT [11] model on multi-lingual data from wikipedia. We set $batchsize = 32$, a learning rate between 10-3 and a fine-tuning rate of 2×10^{-5}. We run each selected model five times and then calculate the average performance. As we already mentioned, these models are fully supervised, therefore we employ 3-fold cross validation to evaluate the models. Although, our proposed model follows fully unsupervised setting.

3.3. Results Discussion

We trained the chosen models on two types of feature sets. The VSM representation on actual terms of text, referred as "ofs" and extended feature set using hypernyms of key terms referred as "hyper". As the results of all the models can be seen in Table 2. In terms of accuracy and recall, our model outperformed the BiLSTM and FastText. The reason behind is the limited training dataset to learn the network weight in multi-lingual instance. Whereas, M-BERT is specifically pre-trained to multi-lingual data that is the reason it performed well with limited training instances. It can be clearly observed that using hypernym as additional information improves the model performance.

Model	ofs	hyper	ofs	hyper
	Accuracy		Recall	
M-BERT	81.39	83.49	83.59	84.33
FastText	73.79	74.90	74.83	75.55
BiLSTM	72.40	73.10	73.44	74.20
NDML	79.22	83.51	80.32	84.48

3.4. Parameter Sensitivity

We analyzed the two hyper-parameters of NDM for further investigation of clustering quality. The Figure 2 and Figure 3 show the sensitivity of α and β, respectively, with respect to Normalized mutual information (NMI), accuracy, and homogeneity (Ho). From the the given plots we can clearly observe that the model in the streaming environment does not show a high variance over a significant range of values.

4. Conclusion

In this paper, we propose short text clustering framework by integrating non-parametric dirichlet model (to cluster short text) with language translation module to identify event in multi-lingual social stream. In contrast to existing approaches, NDML do not require pre-defined number of events and cluster continuous arrival of text instances in temporal dependency scenario. More importantly, we tried to explore the hypernym using Babelfy to expand the core terms of each cluster. The given empirical results demonstrate the significance of our model.

[4] https://crisislex.org

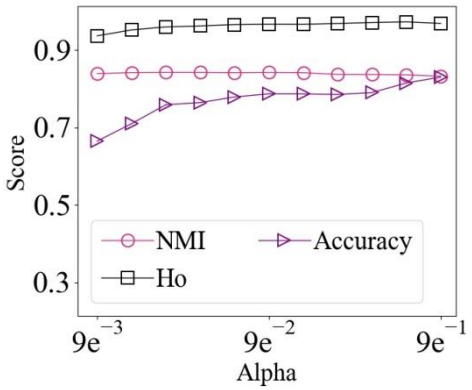

Fig.2 The sensitivity of α parameter.

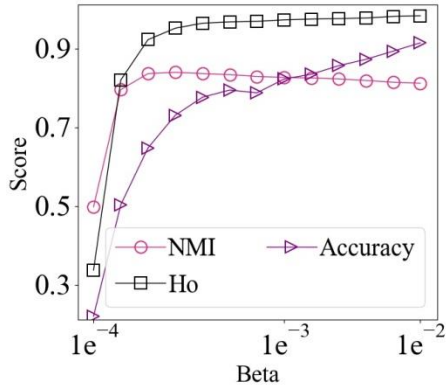

Fig.3 The sensitivity of β parameter.

References

[1] N. D. Doulamis, A. D. Doulamis, P. C. Kokkinos, and E. M. Varvari- gos, "Event detection in twitter microblogging," IEEE Trans. Cybern., vol. 46, no. 12, pp. 2810–2824, 2016.

[2] J. R. Chowdhury, C. Caragea, and D. Caragea, "Cross-lingual disaster- related multi-label tweet classification with manifold mixup," in Pro- ceedings of the 58th Annual Meeting of the Association for Computa- tional Linguistics: Student Research Workshop, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, 2020, pp. 292–298.

[3] P. Khare, G. Burel, D. Maynard, and H. Alani, "Cross-lingual clas- sification of crisis data," in The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I, vol. 11136. Springer, 2018, pp. 617–633.

[4] Y. Liu, H. Peng, J. Li, Y. Song, and X. Li, "Event detection and evolution in multi-lingual social streams," Frontiers Comput. Sci., vol. 14, no. 5, p. 145612, 2020.

[5] K. K, Z. Wang, S. Mayhew, and D. Roth, "Cross-lingual ability of multilingual BERT: an empirical study," in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

[6] J. Chen, Z. Gong, and W. Liu, "A nonparametric model for online topic discovery with word embeddings," Information Sciences, vol. 504, pp. 32–47, 2019.

[7] J. Yin, D. Chao, Z. Liu, W. Zhang, X. Yu, and J. Wang, "Model-based Clustering of Short Text Streams," in ACM International Conference on Knowledge Discovery and Data Mining, 2018, pp. 2634–2642.

[8] S. U. Din and J. Shao, "Exploiting evolving micro-clusters for data stream classification with emerging class detection," Information Sci- ences, vol. 507, pp. 404–420, 2020.

[9] H. Gong, T. Sakakini, S. Bhat, and J. Xiong, "Document similarity for texts of varying lengths via hidden topics," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, 2018, pp. 2341–2351.

[10] Joulin, E. Grave, P. Bojanowski, M. Douze, H. Je´gou, and T. Mikolov, "Fasttext.zip: Compressing text classification models,"CoRR, 2016. [Online]. Available: http://arxiv.org/abs/1612.03651

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[12] Z. Saeed, R. A. Abbasi, A. Sadaf, M. I. Razzak, and G. Xu, "Text stream to temporal network - A dynamic heartbeat graph to detect emerging events on twitter," 2018, doi: 10.1007/978-3-319-93037-4_42.