Hindawi
Scientific Programming
Volume 2021, Article ID 9781923, 3 pages
https://doi.org/10.1155/2021/9781923

Editorial

Novel Tools for the Management, Representation, and Exploitation of Textual Information

David Ruano-Ordás,[1,2,3] **Jose R. Méndez**,[1,2,3] **Vítor Basto Fernandes**,[4] **and Guillermo Suárez-Tangil**[5,6]

[1]*Department of Computer Science, University of Vigo, ESEI-Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain*
[2]*CINBIO-Biomedical Research Centre, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain*
[3]*SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, 36312 Vigo, Pontevedra, Spain*
[4]*Instituto Universitário de Lisboa (ISCTE-IUL), University Institute of Lisbon, ISTAR-IUL, Av. Das Forças Armadas, 1649-026 Lisboa, Portugal*
[5]*IMDEA Networks Institute, Av. Del Mar Mediterraneo, 22, Leganes, Spain*
[6]*Department of Informatics, King's College London, Faculty of Natural and Mathematical Science, Strand Campus, London, UK*

Correspondence should be addressed to David Ruano-Ordás; drordas@uvigo.es

Received 28 July 2021; Accepted 28 July 2021; Published 15 August 2021

Over the last decade, the explosive growth of social media and instant message applications together with the improvement of computer performance, networking infrastructures, and storage capabilities have led to the advent of the information age. Particularly, most people in industrialized countries have permanent and unlimited access to the Internet via mobile devices (smartphones, tablets, etc.). This infrastructure allows users to generate and send massive data to Internet servers from everywhere [1].

As long as human communications are made through language, most of the information gathered from mobile devices is textual. Common examples are instant messages, e-mails sent/received, updates sent to microblogs and/or social networks, opinions on products/apps, etc. This scenario has propitiated the massive dissemination and collection of massive and diverse textual information from multiple (and almost unlimited) data sources. However, the deluge of raw text data is neither meaningful nor useful without the use of proper methodologies to detect and extract valuable knowledge. To this end, we have selected some works that clearly contribute by providing relevant ideas, methodologies, and models on the topic of this special issue. A summary of these works and their specific findings are included below.

Knowing relevant information about Internet users in the current digital era is a crucial competitive advantage for industries. Consequently, several works have been developed and introduced in recent years focusing on the management, representation, and exploitation of the knowledge available from textual sources. One of the most interesting works in this line is the profiling and clustering of users using information shared through social networks [2]. Concretely, they propose a data collection framework to obtain specific data on individuals to explore user profiles and identify segments based on these profiles. This information is particularly relevant in the customization of user-oriented websites (advertisements, news referrals, etc.). Next, the work presented in [3] addresses the age prediction problem by combining social media-specific metadata and language-related features. To accomplish this task, the authors combine (i) part-of-speech *N*-gram features, (ii) stylometry features (average sentence length, average word length, etc.), and (iii) features from lexicons that correlate words/phrases with specific age and sentiment scores together using deep learning schemes via Keras (available at https://keras.io) framework achieving a good performance.

Another interesting area which is also related with social media textual data exploitation is the detection of breaking

news and trending histories in social networks to avoid spreading rumours (unverified stories or statements) or fake news (misleading information presented as news). These kinds of situations produce serious damage in different areas such as personal or professional life, corporate image of a company, or even a stock market turmoil. The work described in [4] extracts word-embedding features from social networks to train a deep learning model (recurrent neural network) to automatically identify rumours and mitigate topic shift issues. Concerning the fake news detection, we can highlight the works described in [5, 6]. The former proposes a fake news detector (FNDNET) based on the usage of a deep convolutional neural network (CNN). Achieved results applied over the Kaggle fake news dataset show that FNDNet clearly outperforms the results gathered by other well-known alternatives. The latter explores different textual properties (useful to distinguish between fake and real contents) to train a combination of different machine learning algorithms using various ensemble methods. Experimental evaluation carried out over four different real-world datasets confirms the superior performance of the proposed ensemble in comparison to individual learners.

From a medical perspective, textual information posted in social networks could be an important thermometer to both detect and suggest medical diseases/treatments and measure the quality of healthcare services. Particularly, in [7], natural language processing (NLP) and sentiment analysis are used to analyse patient experiences shared through the Internet to assess healthcare performance. Moreover, the work described in [8] presents how latent Dirichlet allocation (LDA) and random forests (RF) were adequately combined to find latent topics of healthcare and show the utility of social media forums to automatically detect healthcare issues in patients.

Another important area lies in the automatic identification, detection of interpersonal and gender-based violence. In this sense, the work of [9] proposed a methodology to detect and associate fake profiles on Twitter used for defamatory purposes based on analysing the content of the comments generated by troll and victim profiles. In their methodology, they used text, time of publication, language, and geolocation as features. They compared different machine learning (ML) classifiers including random forests, J48, K-nearest neighbour (KNN), and sequential minimal optimization (SMO) for assessing the probability of a user being the author of a tweet. The experimentation carried out used the false-positive/false-negative ratio and area under receiver operating characteristic (AUC) curves to demonstrate the suitability of the proposal.

Finally, the identification of influencers in social networks has also been addressed [10]. Identifying hot blogs and opinion leaders allows marketers to determine if the opinions shared are favourable to sell their products. The work of Li and Du introduces a framework to identify opinion leaders using the information retrieved from blog content, authors, readers, and their relationships. Blog contents are used to automatically learn an ontology. This ontology is used to measure expertise, find readers, and assess relationships between readers and blog authors. This data is used to find hot blogs and assess the influence of bloggers.

This special issue brings together several papers showing different utilities of text mining and NLPA. It comprises four high-quality works submitted by researchers from China and India, which were selected from the submitted ones. In general, published studies address the following problems: (i) clustering web services, (ii) identification of hotspots (current hot topics), (iii) prediction of user behaviour for the early fetch of interesting web pages, and (iv) the application of named entity recognition (NER) in medical documents written in Chinese.

The first study included in the special issue authored by C. Shan and Y. Du [11] presents a method to automatically group similar web services. Web services are usually Representational State Transfer (REST) Application Programmers Interfaces (API) and provide a collection of tags that are used for clustering. To this end, the authors propose two algorithms. The former is in charge of computing the semantic similarity between the tags of different online available APIs by using the WordNet lexicon database. The latter one is responsible for grouping the APIs according to the values previously computed. To analyse the performance of the proposal, the authors grouped the services available in ProgrammableWeb (available at https://www.programmableweb.com) and measured the performance using recall, precision, and F-score. The proposal outperforms the other five popular and classical clustering approaches.

The work presented by H.R. Cao et al. [12] proposes a hybrid solution for identifying online hotspots, assessing their importance, and enabling their monitoring. They integrate different mechanisms for filtering invalid user posts and replies and design an algorithm to extract keywords from hotspots. Experimental evaluation showed that the method could effectively filter out invalid data, improve the representation of datasets, and reflect changes in hotspot trends.

The proposal included by S. Setia et al. [13] introduced a methodology to accurately model the behaviour of web users. To this end, web browsers can fetch web pages in the background before the user explicitly demands them to improve the experience. The model used to predict the behaviour uses *N-gram* parsing and the click-counter of queries to improve the prediction of web pages. Experimental results have shown that the proposed strategy can significantly reduce the fetching time.

Finally, the work by Y. Wang et al. [14] present a new method for data augmentation based on Masked Language Model (MLM). They compare the performance of NER in the Chinese medical literature using the data augmentation method, pretraining models (such as Bidirectional Encoder Representations from Transformers (BERT), ERNIE (available at https://github.com/PaddlePaddle/ERNIE), or RoBERTa [15]), common deep learning models (such as Bidirectional Long Short-Term Memory (BiLSTM) [16]), and downstream models with different structures (FC, CRF, LSTM-CRF, and BiLSTM-CRF). Their experiments showed the utility of their data augmentation method to improve the

performance of entity recognition, which can also be used to increase the performance of pretraining models.

Despite the abovementioned works in the context of the management, representation, and exploitation of textual information, this field of computer science includes major challenges that have yet to be resolved. We sincerely hope that readers enjoy the special issue and find it worthy for understanding the real value of textual information compiled worldwide.

Conflicts of Interest

The Guest Editors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

David Ruano-Ordás acknowledges Xunta de Galicia for its support under its fellowship program (ED481D-2021/024). José R. Méndez acknowledges the funding support of the Spanish Ministry of Economy, Industry and Competitiveness (SMEIC), State Research Agency (SRA), and the European Regional Development Fund (ERDF) (Semantic Knowledge Integration for Content-Based Spam Filtering, TIN2017-84658-C2-1-R). Vitor Basto-Fernandes acknowledges FCT (Fundação para a Ciência e a Tecnologia), I.P., for its support in the context of projects UIDB/04466/2020 and UIDP/04466/2020. We would also like to thank all authors for their contributions to this special issue and the reviewers for their generous time in providing detailed comments and suggestions that helped us to improve the quality of this special issue.

David Ruano-Ordás
José R. Méndez
Vitor Basto-Fernandes
Guillermo Suárez-Tangil

References

[1] M. Anshari and Y. Alas, "Smartphones habits, necessities, and big data challenges," *The Journal of High Technology Management Research*, vol. 26, no. 2, pp. 177–185, 2015.

[2] J.-W. Van Dam and M. Van de Velden, "Online profiling and clustering of Facebook users," *Decision Support Systems*, vol. 70, no. 2, pp. 60–72, 2015.

[3] A. Pandya, M. Oussalah, P. Monachesi, and P. Kostakos, "On the use of distributed semantics of tweet metadata for user age prediction," *Future Generation Computer Systems*, vol. 102, no. 1, pp. 437–452, 2020.

[4] S. A. Alkhodair, S. H. H. Ding, B. C. M. Fung, and J. Liu, "Detecting breaking news rumors of emerging topics in social media," *Information Processing & Management*, vol. 57, no. 2, Article ID 102018, 2020.

[5] R. K. Kaliyar, A. Goswami, P. Narang, S. Sinha, and S. Sinha, "FNDNet-a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, no. 6, pp. 32–44, 2020.

[6] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity*, vol. 2020, no. 10, 11 pages, Article ID 8885861, 2020.

[7] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson, "Harnessing the cloud of patient experience: using social media to detect poor quality healthcare: table 1," *BMJ Quality and Safety*, vol. 22, no. 3, pp. 251–255, 2013.

[8] H. Jelodar, Y. Wang, M. Rabbani, G. Xiao, and R. Zhao, "A collaborative framework based for semantic patients-behavior analysis and highlight topics discovery of alcoholic beverages in online healthcare forums," Journal of Medical Systems, vol. 44, no. 5, p. 101, 2020.

[9] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying," in *Proceedings of Advances in Intelligent Systems and Computing,International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*, pp. 419–428, Salamanca, Spain, September 2014.

[10] F. Li and T. C. Du, "Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs," *Decision Support Systems*, vol. 51, no. 1, pp. 190–197, 2011.

[11] C. Shan and Y. Du, "A web service clustering method based on semantic similarity and multidimensional scaling analysis," *Scientific Programming*, vol. 2021, no. 5, 12 pages, Article ID 6661035, 2021.

[12] H. Cao, X. Li, S. Lian, and C. Zhan, "A hotspot information extraction hybrid solution of online posts' textual data," *Scientific Programming*, vol. 2021, no. 4, 11 pages, Article ID 6619712, 2021.

[13] S. Setia, V. Jyoti, and N. Duhan, "HPM: a hybrid model for user's behavior prediction based on N-gram parsing and access logs," *Scientific Programming*, vol. 2020, no. 11, 18 pages, Article ID 8897244, 2020.

[14] Y. Wang, Y. Sun, Z. Ma, L. Gao, and Y. Xu, "Named entity recognition in Chinese medical literature using pretraining models," *Scientific Programming*, vol. 2020, no. 9, 9 pages, Article ID 8812754, 2020.

[15] Y. Liu, M. Ott, N. Goyal et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, https://arxiv.org/abs/1907.11692>.

[16] X. Chen, C. Ouyang, Y. Liu, and Y. Bu, "Improving the named entity recognition of Chinese electronic medical records by combining domain dictionary and rules," International Journal of Environmental Research and Public Health, vol. 17, no. 8, p. 2687, 2020.