Deep approach based on user's profile analysis for capturing user's interests

Randa Besnkhelifa, Nasria Bouhyaoui

Universite Kasdi Marbah, Ouargla, BP 511 Route de Ghardaia, Ouargla 30 000, Algérie.
randa.benkhelifa@univ-ouargla.dz
nasria.bouhyaoui@univ-ouargla.dz

Abstract. Capturing user's interests and preferences by analyzing and interpreting the daily-shared contents in online social networks offer a unique information source for several domains such as business, marketing and politics.

User's profile describes its owner's characteristics, where it contains several important personal information such as (age, sex, job title, level of education, etc.), which can help to improve the process of user's interests identification. This information can typically represent a range of values representing only one user profile. Hence, the shared posts, the reactions on other posts and their circle of friends can help to reflect their interests. However, exploiting all this information through the analysis of user profiles can help to enhance user's interests identification performances.
In this paper, we propose a deep learning user's profile analysis based approach that relies on users' personal information and textual content for detecting user's interests and preferences. We experimented our approach using a large Facebook dataset, and show how the deep learning approach perform significantly better than the classical algorithms such as SVM.

Keywords: Deep learning, Personal information attributes, profiling, Text classification, Social text, Online Social Networks, User's interests.

1 Introduction

Online social network (OSN) has become a part of the daily routine of a huge number of persons. OSNs users can fill information that represents their personal information attributes, posts, and comments covering several topics. A large number of fields including: trendsetting, future prediction, recommendation systems, community detection, business marketing, and sentiment classification are interested in the automatic use of this information. User's personal attributes such as gender, age, location, marital status, education, career, etc. are the personal static information that describes user's profile in OSN. Some of the previous researches have interested to the categorization of author's characteristics relying on textual content (textual features) generated by this OSNs user, a part of them is interested in the detection of the author's gender [22], [23], [33], and [25]. Other works, focused on the identification of the author's age [15] and [17]. The mentioned works relied on textual content generated by the user to detect what are their personal information attributes. The previous works prove the existence of strong relation, which led as consequence, to use the text generated by users in order to identify user's personal attributes. Hence, we suggest exploiting personal information attributes textual content generated by the user for identifying user's interests.

We notice from the existing studies on OSNs text classification, that only user-generated content was taken in concern for the aim of improving social text classification performances. In the work in [9] and others have relied on sentiment bags, and some others on embedding models [31]. Whereas, there are a studies that have proposed approaches which deal with the specific characteristics of the social short text [31] and [7]. In The research proposed by (use the name of author here) [3], the first order Markov model for hierarchical Arabic text classification has been used. In addition, [2] has relied on the singular value decomposition method in order to identify textual feature, where [5] has presented an improved method of Chi-square feature selection method. The research in [25] has explored the use of Natural Language Processing techniques in a gender classification system, where the results have determined that word embedding models have significantly performed better using multiple machine learning techniques in opposite of the traditional Bag of Words model. Both of personal information information and user-generated content are available in online social networks. All the existing studies on OSNs text classification depend only on textual content generated by the user. We suggest including the personal information attributes in addition of user-generated content. That step can play a crucial role in improving the classification performances. Hence, it is feasible to leverage these attributes to build a smarter classifier and achieve better performance. In this work, we investigate how both user-generated content (textual data) and personal attributes are exploited to categorize textual content by topics of interests.

Inspired by the recent success of deep learning techniques in many NLP tasks, we further propose a deep personal information-content-based approach relied on both of textual features and personal information features for the classification of user textual content by topics of interests. To evaluate our approach, we propose to compare between the 17 well-known classical classifier SVM that got the best results in [6] and the deep learning classifiers results.

We discussed the previous related works. In section 3, we present our methodology including our proposed approach. In section 4, we represent the details of the used dataset. In section 5, we detailed the results applied to the dataset extracted from Facebook, and finally in section 6, we present a conclusion and perspectives.

2 Related Works

This section emphasizes previous studies relevant to this research, and pinpoints the considered features in each work for improving the performance of the social text classification. Usually, the previous studies have been classified the social text considering only the textual content. The authors in [18] have introduced a text-based

hidden Markov models which utilizes word orders without the need of sentiment lexicons. A posts classification model based on neural network with incorporating user tastes, topic tastes, and user comments has been developed by [11]. The authors in [31], have described a Twitter election classification task that aims to detect election-related tweets, this work is based on embedding models for improving the classification performance. [4] has proposed an approach which deals with the limitations of the social short texts for improving the classification performance. In addition, [7] have introduced a novel preprocessing methods adapted to the special characteristics of the social text in order to improve the classification performances. A real-time system has been developed by [9] to extract and classify the YouTube cooking recipes reviews automatically in order to improve the performance this system some sentiment bags, based on emoticons and injections have been constructed. The authors in [3] have proposed a text classification method which is a space efficient method that utilizes the first order Markov model for hierarchical Arabic text classification. The work in [2] has used the singular value decomposition method to extract textual features; they compare between some of the well-known classification methods.
The research [5] has presented an improved method of Chi-square feature selection method to minimize the data and produce higher classification accuracy. The authors in [20] have focused on the evaluation of feature selection methods for improving text classification performance. The authors in [26] have used the structure of social media opinions for enhancing the sentiment classification performance.

All these previous works mentioned above have focused only on content-based methods and especially, the textual features for analyzing and classifying the social text. They totally neglect the personal information aspect of the user, which can play a crucial role toward the content classification. Moreover, we remarked a strong link between user-generated content and his/her personal information attributes trough several previous studies that have exploited this text shared in OSNs to detect one or more personal information attributes of the user (author's text).

Many studies in OSNs have focused on the task of capturing author's gender. The authors [23] have considered a set of text features such as function words and part of speech n-grams for providing a system of gender classification. The authors [33] have presented an efficient gender classification model to predict the gender values of specified users crawled from a Chinese micro-blog service. The problem of gender classification has been treated by [22] which have proposed a typical surface-level text classification approach by identifying differences between genders in the ways they use the same words. Other works have focused on the age detection, the authors [15] have proposed an approach considering the writing style and both users' history and profile for determining the age groups of Twitter users. Other research have focused on several other characteristics, where a hybrid text-based and community-based method for the personal information estimation of Twitter users has been proposed by [17].The authors in [25] have achieved a significantly better results using word embedding models with multiple machine learning techniques than the traditional Bag of Words model in a gender classification system.

The authors in [6] have considered the personal information attributes with proposing a content-personal information-based approach which use not only the textual features, but both the textual content shared by the user and his/her personal information attributes inorders to improve the classification of the textual content by a topic of interests. Here, the authors have used only the classical algorithm SVM.

In this paper, we propose a deep content-personal information-based approach which uses both of textual content and user's personal information attributes. This approach is based on Deep Learning algorithms. To evaluate the proposed approach, we propose to compare between the well-known classical classifier SVM and the deep learning classifiers results.

3 Proposed methodology

User-generated text in OSNs is exploited in order to discover user's preferences and interests, which represent an important piece of information. Hence, in the work [8] the authors have been discover that even users' personal information attributes can be useful in predicting users' interests, and improving the performance of the social text categorization. To understand the link between users personal information attributes and user's generated text, they have proposed a quantitative analysis by computing the number of users with similar personal information attribute in each topic of interest. Hence, they have deduced that the link between users personal information attributes and user-generated text is strong or weak relatively to the topic of interest in question. "each attribute has a strong impact on some categories (topic of interest) and weak impact on others" [8]. basing on this interesting study proposed by the authors in [8] and due to the power of deep learning in NLP and text classification, we propose a new approach based on deep learning and the users' personal information attributes.

3.1 Traditional classification approach

Exploiting both content and personal information information for social text classification have been studied in [6] using traditional machine learning algorithms such as support vector machine (SVM). In this work, we use SVM model, which achieves the best performance in [6] for our social text classification task. The SVM model is based on the following objective function for classification: for each post p_i. The vector v_i is the corresponding content and personal information attributes vector.

$$f(p_i)=w.v_i+b \tag{1}$$

where w is the normal vector to the hyperplane, and v_i is the input vector. The post class is attributed by the sign of f :

$$h=sing(f(p_i))=\begin{cases} +1, if\ f(pi) > 0 \\ -1\ , else \end{cases}$$

For multi class SVM model, a multi class categorization can be obtained by combining a set of binary classifiers f_1, f_2,... f_m for M classes, and each classifier is trained to differentiate class from the rest. The combination is carried out according to the maximal output before applying the sign function [27].

3.2 Deep learning approach

Due to the recent success of deep learning techniques in text classification [19], we propose a content-personal information-based approach, which use deep neural networks models to classify social text.

Convolutional Neural Network based on Personal Attributes (CNN-PA). Inspired by the good results of convolutional neural network model [19] in text classification task, we further propose to use CNN model in our proposed approach based on the personal information information (CNN-PA). In the training phase, in order to produce a new feature f_i , a convolution operation uses a filter w which has a window h of words sequence $x_{i:i+h-1}$. Usually, the feature fi in the content-based CNN is defined as follows:

$$f_i = g(w.x_{i:i+h-1} + b) \tag{2}$$

where b is a bias term and g is a non-linear function such as hyperbolic tangent. This filter is applied to each possible window of words in the sentence $\{x_{1:h}, x_{2:h+1}, x_{n-h+1:n}\}$ to produce a feature map $f = \{ f_1, f_2,..., f_{n-h+1}\}$. Then a max-pooling operation [12] is applied on the feature map to capture the most important feature by taking the maximum value $\hat{f} = max\{f\}$. Finally, we use a softmax function to generate the probability distribution over labels. Now, we introduce the personal information-content based approach CNN-PA, for each post p_i, v is the combination between words sequence $x_{i:i+h-1}$ and personal attributes A, $V = (x_{i:i+h-1}, A)$.

$$f_i = g(w.v+b) \tag{3}$$

Recurrent Neural Network based on Personal Attributes (RNN-PA): Recurrent neural networks (RNNs) are a rich class of dynamic models. RNNs can be trained for sequence generation by processing real data sequences one step at a time and predicting what comes next [14]. An input vector sequence $x = (x_1,..., x_T)$ is passed through weighted connections to a stack of N recurrently connected hidden layers to compute first the hidden vector sequences $h^n = (h_1^n,..., h_T^n)$ and then the output vector sequence $y = (y_1,..., y_T)$. Each output vector y_t is used to parameterize a predictive distribution $Pr(x_{t+1}|yt)$ over the possible next inputs x_{t+1}.

$$h_t^1 = H(W_{ih^n} x_t + W_{h^1h^1} h_{t-1}^1 + b_t^1) \tag{4}$$

$$h_t^n = H(W_{ih^n} x_t + W_{h^{n-1}h^n} h_t^{n-1} + W_{h^nh^n} h_{t-1}^n + h_t^n) \tag{5}$$

where the W denote weights, the b denote bias vectors and H is the hidden layer function. Given the hidden sequences, the output sequence is computed as follow:

$$\hat{y} = b_y + \sum_{n+1}^{N} w_{h^n y}\, h_t^n \qquad (6)$$

$$y_t = Y(\hat{y}_t) \qquad (7)$$

where Y is the output layer function. The output vectors y_t are used to parameterize the predictive distribution $Pr(x_{t+1}|y_t)$ for the next input. The probability given by the network to the input sequence x is:

$$\Pr(x) = \prod_{t=1}^{T} \Pr(x_{t+1}|y_t) \qquad (8)$$

Usually, for classifying a textual post p the input vector sequence x is constructed only from the content d of this post. Here, in our model called Recurrent Neural Network based on Personal information information (RNN-PA) the input vector sequence x corresponds to both the post-content d and the user's personal attributes A.

4 Dataset collection and construction

The dataset provided in [8] consists of 72900 Facebook posts, these posts contain only the textual content, which are sufficient for training our models .The posts are collected from July to September 2016 using Facebook API. They are collected from 300 specific users' profiles, which are public, active and real. This dataset are made of the posts accompanied by their authors' personal information attributes. These posts are annotated to eight (8) different classes, which are (art, fashion, sport, technology, business, news, science and education, and other). Considering the privacy reasons, the users' IDs and names are deleted. Table 1 summarizes the basic statistics of our data (i.e., the average number of posts and the average number of users in each category).

Table 1. Statistics of Our Data.

Data	Statistics
Posts	72,900
Users	300
Categories	News (8748), Business (8311), Technology (3413), Art (9112), Sport (6561), Mode& Fashion (6780), Science & Education (8165), Other (11810).
Gender	Female (42.66%), Male (57.34%)
Age	13–17 (19.4%), 18–27 (36.8%), 28–37 (27.8%), 38–60 (16%)
Work	Worker(60.3%), Unempoloyed (20.3%), Not concerned (19.4%)

Edudcation	University (yes, 44.1%), (No, 55.9%)
Marital status	Married (46.1%), Not married (34.5%), Not concerned (19.4%)

5 Results and discussion

In this section, we compare between the results obtained applying the different algorithms based only on textual content and the results obtained by these different algorithms based on both textual content and personal attributes (gender, age, marital status, work, and education). These results are summarized in Table 2. Here, we evaluate all our ranking models using 10-fold cross validation to reliably compute e statistical significance values.

Table 2. Comparison between classifiers results, Precision, Recall, F-Measure, and Accuracy, using 10- folds cross validation..

Method	Content	Personal Attributes	Accuracy (%)	Precision	Recall	F-Measure
SVM [8]	*		75.55	0.774	0.755	0.761
CNN	*		78.6	0.786	0.786	0.787
RNN	*		75.7	0.778	0.757	0.757
SVM [8]	*	*	90.01	0.896	0.9	0.897
CNN-PA (ours)	*	*	90.7	0.909	0.907	0.908
RNN-PA (ours)	*	*	94.9	0.949	0.949	0.949

By observing the results showed in table 2 we remark that the personal attributes have an impact on the classification performances. Especially, deep neural network models have shown a more powerful classification a model using both content and personal information attributes features. Comparison between classifiers results: As we remark from the table 2, the highest results we achieved are: 94,9% in accuracy, 0.949 in precision, 0.949 in recall and 0.949 in F-measure using the textual content and users characteristics (users personal information attributes) by CNN-PA classifier. Firstly, using only the textual content, we got the following results for the accuracies, 75.55% using SVM, 78,6% using CNN, 75,7% using RNN, the best accuracy obtained here are given by CNN classifier. In term of precision, CNN got the best result with 0.786, SVM with 0.774, and RNN with 0.778. In terms of recall and F-Measure, we got these results respectively, CNN with (0.786 and 0.787) which represent the best results followed by RNN with (0.757 and 0.757) and SVM with (0.755 and 0.761). Now, we discuss the results obtained by adding personal information information of users in the classification process. Where, we got the following results, accuracies with 90.01% using SVM, 94.9% using CNN-PA and 90.7% using RNN-

PA, here we notice that the best accuracy is achieved by CNN-PA classifier. In term of precision, CNN-PA got the best result with 0.949, SVM with 0.896, and RNN-PA with 0.909. In terms of recall and F-Measure, we got these results respectively; CNN-PA classifier with (0.949 and 0.949) represents the best results followed by RNN-PA with (0.907 and 0.908) and SVM with (0.9 and 0.897).

After presenting, the global results obtained using only the textual content and using both textual content and the prsonal attributes, and showing the positive impact of the personal information attributes on the classification performances…

6 Conclusion and future works

In this article, we have presented a deep learning-based approach to sequential social short-text classification. We demonstrate that adding user's personal information improves the performance of the user's interests identification, and the use of deep learning model performs significantly better than the classical algorithms used in the stat of the art, such as the SVM classifier.

Limitations and future work: The main idea of the proposed approach is to include users' personal attributes in the classification process of the social text in order to improve its performances. However, two major limitations are presented here. Firstly, only the explicit attribute are taken. As such, our approach does not work well when user's Social information is not available, or users restrict their personal information. In addition, this work is limited to the textual data where one of the OSNs specificity is the data heterogeneity (posts can be images, videos, text, etc.).

In our future works, we will spread out our dataset and mainly focus on solving the posed limitations: firstly, proposing an automatic model, which can extract the implicit information about the user, existed in his social content. Secondly, we will analyze other multimedia content (such as images).Thirdly, we will exploit other information available on OSNs to build a more powerful classification model, for example, using different interaction information such as the reactions (Like, lovely, etc.), and friendship.

References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. In Proceedings of the 2008 international conference on web search and data mining (pp. 183-194). ACM. DOI: 10.1145/1341531.1341557.
2. Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. Journal of King Saud University-Computer and Information Sciences, 29(2), 189-195. DOI: https://doi.org/10.1016/j.jksuci.2016.04.00.
3. 3. Al-Anzi, F. S.,& AbuZeina, D. (2018). Beyond vector space model for hierarchical Arabic text classification: A Markov chain approach. Information Processing & Management, 54(1), 105-115. DOI: https://doi.org/10.1016/j.ipm.2017.10.003

4. 4. Alsmadi, I., & Hoon, G. K. (2018). Term weighting scheme for short-text classification: Twitter corpuses. Neural Computing and Applications, 1-13. DOI: https://doi.org/10.1007/s00521-017-3298-8
5. 5. Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2018). Feature selection using an improved Chi-square for Arabic text classification. Journal of King Saud University-Computer and Information Sciences. Doi : https://doi.org/10.1016/j.jksuci.2018.05.010.
6. 6. Benkhelifa, R., Bouhyaoui, N., & Laallam, F. Z. (2018). A demographic-based approach for improved content categorization in social networking. In Natural Language and Speech Processing (ICNLSP), 2018 2nd International Conference on (pp. 1-5). IEEE.
7. 7. Benkhelifa, R.,& Laallam, F.Z,. (2016). Facebook Posts Text Classification to Improve Information Filtering. In Proceedings of the 12th International Conference on Web Information Systems and Technologies, 202-207, 2016, Rome, Italy. DOI: 10.5220/0005907702020207. 8.
8. Benkhelifa, R., & Laallam, F. Z. (2018). Exploring demographic information in online social networks for improving content classification. Journal of King Saud University-Computer and Information Sciences.
9. 9. Benkhelifa, R., and Laallam, F. Z. (2018). Opinion Extraction and Classification of Real-Time YouTube Cooking Recipes Comments. In International Conference on Advanced Machine Learning Technologies and Applications, pp. 395-404. Springer, Cham. Doi: https://doi.org/10.1007/978-3-319-74690-6 39.
10. Bilski, A. (2011). A review of artificial intelligence algorithms in document classification. International Journal of Electronics and Telecommunications, 57(3), 263-270. Doi: https://doi.org/10.2478/v10177-011-0035-6
11. Chen W F, Ku L W. (2016) UTCNN: A deep learning model of stance classification on social media text. In Proc. The 26th International Conference on Computational Linguistics, pp.1635-1645.
12. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. (2011) Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011, 12: 2493-2537.
13. Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In Proceedings of the seventh international conference on Information and knowledge management (pp. 148-155). ACM. Doi: 10.1145/288627.288651.
14. Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.
15. Guimaraes, R. G., Rosa, R. L., De Gaetano, D., Rodriguez, D. Z., & Bressan, G. (2017). Age Groups Classification in Social Network Using Deep Learning. IEEE Access, 5, 10805-10816. DOI: 10.1109/ACCESS.2017.2706674.
16. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/BFb0026683.
17. Ikeda K, Hattori G, Ono C et al (2013) Twitter user profiling based on text and community mining for market analysis. Knowl-Based Syst 51:3547. doi:10.1016/j.knosys.2013.06.020
18. Kang, M., Ahn, J.,& Lee, K. (2018). Opinion mining using ensemble text hidden Markov models for text classification. Expert Systems with Applications, 94, 218-227. DOI: 10.1016/j.eswa.2017.07.019
19. Kim Y. (2014) Convolutional neural networks for sentence classification. In Proc. the Empirical Methods in Natural Language Processing, Oct. 2014, pp.1746-1751.

20. Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., &Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. Applied Soft Computing, 86, 105836.
21. Lopes Filho, J. A. B., Pasti, R., & de Castro, L. N. (2016). Gender Classification of Twitter Data Based on Textual Meta-Attributes Extraction. In New Advances in Information Systems and Technologies (pp. 1025-1034). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-31232-3 97.
22. Mihalcea, R.,& Garimella, A. (2016). What Men Say, What Women Hear: Finding Gender-Specific Meaning Shades. IEEE Intelligent Systems, 31(4), 62-67. DOI: 10.1109/MIS.2016.71.
23. Mukherjee, S., & Bala, P. K. (2017). Gender classification of microblog text based on authorial style. Information Systems and e-Business Management, 15(1), 117-138. doi: 10.1007/s10257-016-0312-0
24. Myllymaki, P., & Tirri, H. (1993). Bayesian case-based reasoning with neural networks. In Neural Networks, 1993., IEEE International Conference on (pp. 422-427). IEEE. Doi: 10.1109/ICNN.1993.298594
25. Vashisth, P., & Meehan, K. (2020). Gender Classification using Twitter Text Data. In 2020 31st Irish Signals and Systems Conference (ISSC) (pp. 1-6). IEEE.
26. Vairetti, C., Mart´ınez-Camara, E., Maldonado, S., Luz ´ on, V., & Herrera, F. (2020). Enhancing the clas- ´ sification of social media opinions by optimizing the structural information. Future Generation Computer Systems, 102, 838-846.
27. Weston, J., & Watkins, C. (1998). Multi-class support vector machines. Technical Report CSD-TR98-04, Department of Computer Science, Royal Holloway, University of London, May.
28. Witten, H.A. and Frank, E. (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, Doi: https://doi.org/10.1016/C2009-0-19715-5.
29. Wollmer, M.,Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., & Morency, L. P. (2013). Youtube movie reviews: Sentiment analysis in an audio-visual context. Intelligent Systems, IEEE, 28(3), 46-53.
30. Yan X, & Yan L, (2006). Gender classification of weblog authors. In: AAAI spring symposium series on computational approaches to analysing weblogs, pp 228230
31. Yang, X., Macdonald, C., & Ounis, I. (2018). Using word embeddings in twitter election classification. Information Retrieval Journal, 21(2-3), 183-207. Doi: https://doi.org/10.1007/s10791-017-9319-5
32. Yu, B., Xu, Z. B., & Li, C. H. (2008). Latent semantic analysis for text categorization using neural network. Knowledge-Based Systems, 21(8), 900-904. Doi: https://doi.org/10.1016/j.knosys.2008.03.045.
33. Yu, Y., & Yao, T. (2017, June). Gender Classification of Chinese Weibo Users. In Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government (pp. 5-8). ACM. Doi: 10.1145/3108421.3108423.