IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10434

Learn a method to detect articles with false content

Hoang Tuan Long[1], Nguyen Van Xuan[2], Dinh Viet Hung[3], Ho Nguyen Xuan Thanh[4], Le Minh Tri[5]

People's Police University - Ho Chi Minh City, Vietnam[1,2,3,4,5]

Abstract: This study presents a method to identify articles written in Vietnamese on the internet that contain reactionary viewpoints against the Government of Vietnam and the leadership of the Communist Party of Vietnam. These articles often comprise various errors such as spelling mistakes, typos, misplaced punctuation marks, new and unfamiliar "terms" to Vietnamese people, etc. Hence, it is not appropriate to apply grammatical and vocabulary analysis methods. We propose to use the word orders in triplet form (Subject, Verb, Object) and its variables including doublet form (Subject, Predicate, null) or (Verb, Object, null), and singulet form (Subj, null, null) to screen these articles in accordance with the following principle: if one article has at least one sentence containing the elements of such word orders, the article will be considered as containing reactionary viewpoints. The original triplets are established based on the training corpus (dataset), and then extended using the synonyms in VietWordNet. The extension of triplets is able to increase the accuracy of this algorithm significantly. The Program can help professional security units to reduce human resources and enhance operational effectiveness.

Keywords: document analysis, document classification, reactionary viewpoint, triple, edge triplet, triplet finding, Spark GraphX, VietWordNet.

I. INTRODUCTION

The war that reunited the North and the South of Vietnam ended more than 40 years ago but a number of the defeated parties who are living in many different countries are still trying to incite Vietnamese people inside and outside the country to involve in sabortage activities against the Government of Vietnam. One of the most effective ways to disseminate their reactionary viewpoints is using social media and webpages/weblogs. Therefore, one of the tasks for ensuring the political security and social order for the country is to identify the articiles containing reactionary viewpoints, expose the wrongful allegations and provide the grassroots with clear explanations. This study proposes a method to detect articles containing reactionary viewpoints in order to assist the above mentioned task.

This problem belongs to the field of document classification. A number of related studies are presented below.

1. Professional method

An organ of public security sector is assigned to detect the articles containing reactionary viewpoints. The competent officials enter social media pages or webpages/weblogs where reactionary viewpoints are often expressed to analyze and determine wherether an article is reactionary or not [1] based on one of the following factors:

- Containing phrases that express incitations, defamations, and distortions of the government authority, the Communist Party of Vietnam, the famous people of the Communist Party of Vietnam, or historical details with the purposes of disseminating suspections, discontentments, and mistrusts in the government authority, the Communist Party of Vietnam, the respectful people of the Communist Party of Vietnam, or historical details; or containing incitation phrases that stimulate people to take part in illegal meetings, causing internal riots or overturns on large scales with the supports of external reactionary organizations.

Example 1: Analyzing the following sentence: "Nhưng ở đây, lại có một vấn đề mà tôi muốn làm rõ, chẳng những với cá nhân ông "trí ngủ" Đỗ Văn Xê mà còn với hằng ngàn dư luận viên đang ngày đêm giúp đảng CSVN che giấu sự thật tồi bại của đảng."

The article which has one sentence containing phrases of this type will be considered as containing reactionary elements. As in example 1, the phrase is: "sự thật tồi bại của đảng".

Then, the article will be thoroughly analysed, focusing on the phrases or sentences containing the reactionary elements, in order to identify every reactionary allegation and find out appropriate explanations or disseminations to people;

- The article does not contain such noticeable phrases, but distorts the truths or incites riots or overturns in hilarious or figurative ways

Example 2: Analyzing the following sentence: "Cuộc CCRĐ Hồi Thứ Nhất với mục tiêu lừa đảo là Người Cày Có Ruộng kéo dài từ năm 1953 đến năm 1956 đã diệt chủng long trời lở đất đến phải sửa sai và chấm dứt. Trong khi "bác Hồ" đóng phim nhỏ vài giọt lệ khóc những người chết oan thì Võ Đại tướng phải thay mặt cụ và Tổng bí thư Trường Chinh đứng ra nhận sửa sai."

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10434

This type of article, though normally does not contain such keyphrases, still contains reactionary elements. The second example has no keyphrase but still implies defamation of the leaders causing people's mistrusts

The articles in this form will be analyzed carefully in every aspect and implication to identify reactionary allegation in order to find out suitable explanations and disseminations.

In reality, most of the reactionary articles fall into the first type, i.e. containing reactionary phrases. This type accounts for more than 95% of the total number of reactionary articles.

Since the competent organ is still using the manual method without the assistance of computerized programs, the works are costly and labor-intensive while detection of reactionary phrases by computerized programs are feasible and not very complicated. The development of programs to detect articles containing implicative distortions or riot incitations is more challenging, even infeasible under the current study situations. The following overview will only present the analyses of the studies in order to detect the articles that contain reactionary phrases, not the ones that contain implicative reactionary elements.

2. Keyphrase finding method

The above professional analysis shows us one of the simplest methods for analysing a single sentence to detect the appearance of keyphrases – hereinafter referred to as the first method. If the sentence contains a reactionary element, it is determined that the article contains reactionary elements. Otherwise, if none of the sentences in the article contains any reactionary elements, it is determined that the article does not contain reactionary elements.

Some phrases, however, can be rephrased into various grammatical elements like S-V-O, for instant: "chính quyền đán áp nhân dân" can be rephrased into "chính quyền" as S, "đàn áp" as V, and "nhân dân" as O, then it might not be feasible to detect the entire phrase in a sentence if the writer added another subject/adverbal phrase/predictate/ or object in the middle. A sentence which is added more words but still follows S-V-O structure will not be screened by simple finding algorithims. Thus, the accurary of the algorithms will be significantly reduced.

Example 3: Analyzing the following sentence: "Hiện nay, chính quyền ra sức đàn áp đối với nhân dân tham gia các cuộc biểu tình chống đối lại quyết định của họ."

In this example, the S-V-O structure is used: "chính quyền, đàn áp, nhân dân". However, the additional words in the middle makes it impossible to detect the reactionary elements for the entire phrase.

3. Document classification using semantic analysis method

The next method for solving the problem is the use of gramamatical analysis algorithms to detect reactionary phrases - hereinafter referred to as the second method. There are not many studies on grammatical analyses for Vietnamese language. Most of the studies focus on separations of words and phrases [2]; some others study on grammatical functions of words and phrases in sentences [3-7].

The second method will provide more accurate results in comparison to the first one since it is able to determine exactly whether the elements of a phrase appearing in a sentence have semantic connections or not.

Example 4: Analyzing the following sentence: "Dưới sự lãnh đạo của Đảng Cộng sản thì chúng ra mới thấy được sự dã man, tàn độc của các thế lực thù địch." then the phrase "dã man, tàn độc" does not modify the phrase "Đảng Cộng sản".

In fact, the percentage of sentences containing all elements of a reactionary phrase without any grammatical connections is very low.

The second method also has some shortcomings when being applied, i.e. if a sentence contains spelling mistakes, new/unpopular/borrowed terms and words, grammatical errors due to wrong or missing punctuation marks and conjunctions, the algorithm does not work or can produce a wrong result while these cases are very common in reality.

Example 5: Analyzing the following sentence: "Nhưng đó chỉ là kế hoãn binh của những kẻ câm quyền, cộng sản, hệ thống đảng, chuyên lừa, lọc dối trá." (when being correctly written, it should mean: "But that is only the temporization of the power holders, the communist people, the party system who are deceptive and cheating"). We can see that the algorithm provides an incorrect result due to the wrong punctuation mark between "lừa, lọc" and the spelling mistake of the word "câm quyền", thus the accuracy of the algorithm is lowered.

4. Document classification using statistical machine learning methods

Such statistical machine learning methods as Bayes, LDA,… [8-12], hereinafter referred to as the third method, help to classify texts, firstly by identification of the words representing each classification, and then relevant functions will be applied to screen the text to determine the classification that each keyphrase belongs to. This method is more advantageous over the first and second methods because it automates the establishment of key words and phrases while it is still able to avoid the shortcomings of the first method. Even though, for separation of words and phrases, the third method also has similar shortcomings to the second method including spelling mistakes, grammatical errors, appearance of borrowed and unpopular terms, etc. These shortcomings significantly reduce the efficiency of this method.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10434

II. PROPOSED METHOD USING THE TRIPLET

1. Proposed of Data Structure and Algorithm

According to the above analyses, the first method, though very simple, is effective in solving this problem. Hence, we propose to use this method with a minor change: each keyphrase will be analysed as a set of elements with semantic relations. Each set will have maximum 3 elements called a triplet. The forms of semantic relations can be Subject -Verb -Object, Noun- Modified Adjective, Verb – Modified Noun, Noun, etc.

Example 6:

The one - element triplet (“Hồ động chủ”, “”, “”)

The two - element triplet (“nhà nước”, “độc tài”, “”), (meaning “government”, “dictatorship”, “”), when being combined as a phrase, will express the “dictatorship” policy of the “government” that the reactionists want to broadcast.

Regarding the three - element triplet of (“chính quyền”, “đàn áp”, “nhân dân”), (meaning “authority”, “suppress”, “people”), when these elements are combined together, the phrase expresses the cruel policy of the “authority” toward their “people”, which incites people and induces them to involve in wrongful activities.

Proposed algorithm will finding the appearance of triplet elements in all sentences. If a sentence contains all the triplet elements following their orders in the triplet, such text will be marked as containing reactionary viewpoints, and vice versa. As a result, the proposed algorithm can limit the shortcomings of the first method because it still helps to identify the sentences containing original keyphrases (when the triplet elements have not been split) with additional words/phrases in the middle. However, this algorithm also has some disadvantages that the first method encountered, i.e. the cases that the elements of a triplet appear in a sentence in good orders without any semantic relations. In fact, among the sentences that contains keyphrases with additional words, the percentage of sentences having elements with grammatical relations is considerably higher than the percentage of sentences having unrelated elements without relations among its elements. Because the language is very rich and diversified, one word or phrase often has one or more synonyms. For a comprehensive coverage of various ways to express a reactionary viewpoint – a triplet, we propose to extend the triplet by using the synonyms described in VietWordNet [13].

Example 7: for the manually supplemented triplet of (“chính quyền”, “đàn áp”, “nhân dân”), if the synonyms of the word “chính quyền” are taken from VietWordNet, we can come up with following additional triplets: (“chính phủ”, “đàn áp”, “nhân dân”), (“nhà nước”, “đàn áp”, “nhân dân”), etc.

2. Realization of the triplet finding algorithm based on the Spark GraphX platform

The above mentioned algorithm has been realized using the graphic data structure in Spark GraphX [14, 15]. The first and third elements of the triplet (can be empty) are displayed as the vertexes of the graph with the main features being a string of characters corresponding to such elements. The middle element of the triplet (can be empty) is presented by edge of the graph with the main features being a string of characters corresponding to such element.

Example 8:

Triplet (“”,””, “Hồ tệ”) is represented as vertexes, edge as follows:
Vertex(0L, “”), Vertex (1L, “Hồ tệ”), Edge(1L, 0L, “”)
Triplet (“chính quyền”, “đàn áp”, “nhân dân”) is represented as vertexes, edge as follows:
Vertex(2L, “chính quyền”), Vertex(3L, “nhân dân”),, Edge(2L, 3L, “đàn áp”)

The algorithm to detect the appearance of triplets will be dispersedly conducted on Spark platform as described below:

- The set of triplets will be generated from all edge triplets of the graph and converted into a RDD structure to distributed to the Spark cluster nodes;
- The set of text sentences will be extracted from the article and converted into a broadcast structure and sent to all nodes of cluster;
- At each nodes, every sentence of the article will be compared with the triplet elements received by the subcomputer. If a sentence contains sufficient elements of a triplet following their triplet orders, it is considered as a match and noted in the relevant results for the text.
- Once the process ends at all the computers, the matching data will be reviewed. If there is at least one match detected, the text is marked as containing reactionary elements, and vice versa.

3. Realization of algorithm to extend the triplets based on VietWordNet

VietWordNet [13] has been based on the wordnet Princeton version 3.0. The compilation of data has been inherited and developed from the WNMS tools, a web protocol developed under the AsianWordNet project by Thai Computational Linguistic (TCL) Laboratory in cooperation with Japan’s National Institute of Information and Communications Technology (NICT). Viet WNMS has been customized and improved to accommodate Vietnamese language.

VietWordNet comprises 40,788 synonyms with 67,344 word units including 40,788 common Vietnamese words. It is possible to extend the triplets with the assistance of the language data provided by VietWordNet.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10434

III. TESTING AND ASSESSMENT

The proposed algorithm was tested as follows:

- Firstly, development of test datasets. The articles in reactionary webpages as well as in other (non-reactionary) webpages were extracted and analyzed for being tagged as containing non-reactionary or reactionary viewpoints, and the keyphrases expressing reactionary viewpoints (if any) were recorded. This resulted in the identification of 100 articles containing reactionary viewpoints and 100 articles containing non-reactionary viewpoints.

Next, the articles tagged as containing reactionary viewpoints were classified into two sets according to the timely orders: Training Corpus consisting 50 earlier posts was used for training in order to develop triplets based on manual method; Test Dataset A consisting 50 later posts was used for testing. The set of articles containing non-reactionary viewpoints were gathered into Test Dataset B.

- Development of triplets. For the Training Corpus, all the keyphrases were extracted and developed into various elements of triplets by manual method. This resulted in 400 triplets including 60 one- element triplets, 110 two-element triplets, and 230 three-element triplets.
- Testing triplets developed by manual method. The algorithm was tested with the triplets that were developed by the manual method for tree datasets: Training Corpus, Test Dataset A and Test Dataset B. The test results were presented in Table 1. According to the obtained results, the accuracy of the algorithm was 100% for the Training Corpus, 100% for Test Dataset B since the articles in this set did not contain keyphrases, and only 56% for Test Dataset A since this set contained different keyphrases than the training process. Analysis Test Dataset A only has 28 articles contain keyphrases in triplets developed by manual method.

Table 1. Testing triplets developed by manual method

Data		Number of bookmarks	Analysis results	Precision	Recall
Training Corpus	The article has a reactionary element	50	50		
	The article has no reactionary elements	0	0		
Test Dataset A +Dataset B	The article has a reactionary element	50	28	100%	56%
	The article has no reactionary elements	100	122		

- Proceeding to extension of triplets. The algorithm for extension of triplets using the synonyms in VietWordNet increased the number of triplets to 521,200, including 60 one-element triplets, 5,910 two-element triplets, and 515,230 three-element triplets.
- Testing extended triplets. The algorithm was also tested with the extended triplets for the same datasets used for the previous test. The test results were presented in Table 2. It was obvious that the accuracy for Test Dataset A was increased significantly to 78%. This could be explained by the presents of some keyphrases in the extended triplets thanks to the synonyms in VietWordNet that had not appeared in the original triplets for Test Dataset A. Thus, it was proved that the extension of triplets using VietWordNet helped to increase the accuracy of the algorithm.

Table 2. Testing extended triplets

Data		Number of bookmarks	Analysis results	Precision	Recall
Training Corpus	The article has a reactionary element	50	50		
	The article has no reactionary elements	0	0		
Test Dataset A +Dataset B	The article has a reactionary element	50	39	100%	78%
	The article has no reactionary elements	100	111		

The testing extended triplets results for Test Dataset A is higher than the testing triplets developed by manual method results of 22%. After Test Dataset A analyzed, it includes 28 articles contain keyphrases in triplets developed by manual method, 11 articles contain keyphrases in extended triplets, 10 articles contain new keyphrases and 1 article does not contain keyphrases, but containing reactionary viewpoints.

The results of both tests showed that the accuracy for the "prospective" dataset was still low (less than 80%). This is the shortcoming of the algorithm. For practical application, it requires updated triplets to express new and unpopular reactionary viewpoints. However, such new and unpopular reactionary viewpoints rarely appear, accounting for about

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10434

22% of the cases, and the existing viewpoints expressed in corresponding terms are prevailing, accounting for approximately 78%.

IV. CONCLUSION

The study considered different solutions for the problem of determining whether the articles in webpages contain reactionary viewpoints or not as a tool to improve the effectiveness of information dissemination activities. The study proposes the use of triplets and the algorithm of extended triplets with the synonyms in VietWordNet to solve this problem and to realize the algorithm using graph data structure in Spark GraphX. The algorithm gains high accuracy for the training corpus and acceptable accuracy for the test dataset with extended triplets.

The experiment has proved that the proposed algorithm is able to increase the effectiveness of text classification as it can detect a considerable percentage (more than 70%) of the articles containing reactionary viewpoints by using keyphrases. For more effective application of the solution in reality, the rest articles of less than 30% tagged as containing non-reactionary viewpoints must be reanalyzed. In case it finds out reactionary phrases, the triplets should be duly updated.

This study can be further developed by the combination with the text classification method using statistical machine learning to establish original triplets, and/or with the grammatical analyzing method to check the grammatical relations among the elements of a triplet in a sentence.

REFERENCES

[1] Đào Thanh Quyên (2016), "Đổi mới nội dung, phương thức đấu tranh chống quan điểm sai trái trên mạng internet hiện nay", http://www.tuyengiao.vn/Home/ Bao-ve-nen-tang-tu-tuong-cua-Dang/87873/Doi-moi-noi-dung-phuong-thuc-dau-tranh-chong-quan-diem-sai-trai-tren-mang-internet-hien-nay

[2] Phuong Le Hong, Huyen Nguyen Thi Minh, Vinh Ngo Tuong (2008), "A Hybrid Approach to Word Segmentation of Vietnamese Texts", 2nd International Conference on Language and Automata Theory and Applications.

[3] L. N. Thi, L. H. My, H. N. Viet, H. N. T. Minh and P. L. Hong, "Building a treebank for Vietnamese dependency parsing," In Proceedings of the 10th IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future, pp. 147-151.

[4] D. Nguyen, D. Nguyen, S. Pham, P.-T. Nguyen and M. L. Nguyen (2014), "From treebank conversion to automatic dependency parsing for Vietnamese," In Proceedings of 19th International Conference on Application of Natural Language to Information Systems, pp. 196-207.

[5] C. Vu-Manh, A. T. Luong and P. Le-Hong (2015), "Improving Vietnamese Dependency Parsing Using Distributed Word Representations," Proceedings of the Sixth International Symposium on Information and Communication Technology, pp. 54- 60.

[6] D. Q. Nguyen, M. Dras and M. Johnson (2016), "An empirical study for Vietnamese dependency parsing," In Proceedings of Australasian Language Technology Association Workshop, pp. 143-149.

[7] K. V. Nguyen and N. L.-T. Nguyen (2016), "Vietnamese Dependency Parsing with Supertag Features," 2016 Seventh International Conference on Knowledge and Systems Engineering (KSE).

[8] T. T. T. Trần, C. T. Vũ và N. Tạ (2012), "Xây dựng hệ thống phân loại tài liệu Tiếng Việt", Khoa Công nghệ Thông tin, Trường ĐH Lạc Hồng, Biên Hòa.

[9] Đ. C. Trần, K. N. Phạm (2012), "Phân loại văn bản với máy học vector hỗ trợ và cây quyết định", Tạp chí khoa học, Trường Đại học Cần Thơ.

[10] T. T. V. Nguyễn (2012), "Nghiên cứu một số thuật toán học máy có giám sát và ứng dụng trong lọc thư rác", Luận văn thạc sỹ kỹ thuật, Học viện công nghệ bưu chính viễn thông, Hà Nội.

[11] D. Blei, A. Ng and M. Jordan (2003), "Latent Dirichlet Allocation", Journal of Machine Learning Research 3.

[12] P. N. Trần, V. T. Phạm, X. C. Phạm, Q. V. D. Nguyễn (2013), "Phân loại nội dung tài liệu web tiếng Việt", Tạp chí Khoa học và Công nghệ 51.

[13] Bộ Khoa Học Công Nghệ, (Accessed 1 July 2017) http://viet.wordnet.vn

[14] .X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, R. Xin, M. Franklin, R. Zadeh, M. Zaharia and A. Talwalkar† (2016), "MLlib: Machine Learning in Apache Spark", Journal of Machine Learning Research 17.

[15] J. Gonzalez, R. Xin, A. Dave, D. Crankshaw, M. Franklin and I. Stoica (2014), "GraphX: Graph Processing in a Distributed Dataflow Framework".