IRIT at CheckThat! 2018

Romain Agez[1], Clement Bosc[1], Cedric Lespagnol[1],
Josiane Mothe[2][0000−0001−9273−2193], and Noemie Petitcol[1]

[1] Université P. Sabatier de Toulouse, UPS, France
FirstName.LastName@univ-tlse3.fr
[2] ESPE, IRIT, UMR5505, CNRS & Université de Toulouse, France
Josiane.Mothe@irit.fr

Abstract. The 2018 CLEF CheckThat! is composed of two tasks: (1) Check-Worthiness and (2) Factuality. We participated to task (1) only which purpose is to evaluate the check-worthiness of claims in political debates. Our method to achieve this goal is to represent each claim by a vector of five computed values that correspond to scores on five criteria. These vectors are then used with machine learning algorithms to classify claims as check-worthy or not. We submitted three runs using different machine learning algorithms. The best result we achieved using the official measure MAP ranks our run that uses non linear SVM the 12th over the 16 submitted runs. Our run that uses linear SVMis ranked 2nd with the Mean Precision@1 measure.

Keywords: Information retrieval · fact-checking · information nutritional label.

1 Introduction

The CLEF CheckThat! first task aims at predicting which claims in political debates should be prioritized for fact-checking. All the background and detailed information about the task are available on the task description paper provided by the organizers of the task [8].

To achieve this goal, the task organizers released several textual transcripts of political debates with each sentence being annotated according to whether it is check-worthy or not.

This paper describes the participation of the Université de Toulouse team (official name RNCC) at CLEF 2018 CheckThat! pilot task for check-worthiness.

We preprocessed the data by representing each sentence corresponding to a transcription of what a speaker said in the debate by a vector containing the score of this sentence for five different criteria. We then trained three classifiers using these vectors to submit three different runs.

The remaining of this paper is organized as follows: Section 2 gives a description of the pilot task. Section 3 details the model we developed and the submitted runs. Then Section 4 details the results we obtained. Finally, Section 5 concludes this paper.

2 Task Description

2.1 Objectives

The Check-Worthiness task aims to predict which statements in a political debate should be fact-checked. Indeed, nowadays, information objects are spreading faster and faster on the Internet and especially on social networks. This spreading is named the *virality* of the information [1].

During a political debate, any of the statements made by the participants can be reused without checking its factuality and it even can become viral. CheckThat! aims at providing journalists with a list of statements members of the debate made that should be checked before they are reused by others.

2.2 Dataset

There are two datasets : one to train the model and one to test it. Both sets consist of political debates transcribed into texts.

They are annotated so that each row indicates the sentence number, the speaker, the transcription of the sentence that the speaker said. The training dataset includes in addition a label that indicates whether this sentence is to be fact-checked or not. The training set contains three political debates while the test set contains seven debates [8].

2.3 Evaluation metric

The task has been evaluated according to different measures. The official measure is MAP which calculates the usual mean of the average precision. Then, other measures were used as Mean Reciprocal Rank which allows to obtain reciprocals of rank of the first relevant document as well as Mean Precision at x which performs the average of x best candidates. Details on the measures used can be found in the task overview [8].

Evaluations are carried out on primary and contrastive runs. Primary run corresponds to the results file of the participant's main model ; the decision of the main run was the participant's decision. Contrastive runs match the secondary models the participant used.

3 Method and runs

We computed five of the criteria from the Information Nutritional Label for Online Documents proposed by [3]. These criteria and the methods we developed in this work to calculate their score are as follows:

- **Factuality and Opinion** : Determines whether a sentence represents a fact or a personal opinion. These two features are based on the same algorithm. Each value is the opposite of the other, it is either 0 or 1. We use a Multi-layer Perceptron classifier, using LBFGS gradient descent [10]. This neural network is composed of 500 neurons in the first hidden layer and 5 neurons in the second hidden layer. The activation function used is the rectified linear unit function (”*relu*”). We used a MLP classifier because it was the best performing classifier over Random Forest, Support Vector Machine and Linear Regression. The datasets to train the neural network come from various Wikipedia articles[3] for factual sentences and from Opinosis[4] for opinion sentences. The features used to classify a sentence are fine-grained part-of-speech tags extracted with spaCy[5].
- **Controversy** : Determines the degree of controversy in a text. We count the number of controversial issues in the text based on the Wikipedia Article List_of_controversial_issues[6]. For each issue referenced in the wiki article, we also take in account the anchor text labels[7] to find the synonyms and other appellations of the issues in all of the Wikipedia database. For example : Donald Trump is in the list of controversial issues. Other names can link to his Wikipedia page such as ”45^{th} President of America”. These names are called anchor text labels and will be recognized as a controversial issue.
- **Emotion** : Determines the intensity of emotion in a sentence. We use the list of $2,477$ emotional words and valuation from AFINN[8] [9] (ex : abusive = -3, proud = 2). We sum the absolute value of the positive and negative valuations of the emotional words found in the sentence and we divide it by the total number of words in the sentence :

$$(\sum posWordValue + \sum |negWordValue|)/totalNumberWords$$

- **Technicality** : Determines the degree of technicality in a text. We count the number of domain-specific terms in the text. For that, we use NLTK[9] [2] to perform part of speech tagging (adjective = JJ, name = NN, etc.). Then, we use the RE library[10] to match, from tags, a regular expression defined in [6] which identifies the *terminological noun phrases* (NPs). NPs represent domain-specific terms in the text. We extract all the NPs from the text and

[3] Each of the following URL should be preceded by https://en.wikipedia.org/wiki /World_War_I, /Industrial_Revolution, /October_Revolution, /Fermi_paradox, /Steam_engine, /Barack_Obama, /Amazon_(company), /Netherlands, /Triangular_trade, /Song_dynasty, /Nanking_Massacre, /The_Holocaust

[4] http://kavita-ganesan.com/opinosis/

[5] spaCy is a library for Natural Language Processing in Python. It provides NER, POS tagging, dependency parsing, word vectors and more.
https://spacy.io/

[6] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

[7] https://en.wikipedia.org/wiki/Anchor_text

[8] http://www2.imm.dtu.dk/pubdb/p.php?6010

[9] Natural Language ToolKit, https://www.nltk.org/

[10] Regular Expression, https://docs.python.org/3/library/re.html

keep those which appear more than once. We then calculate the ratio of the number of these NPs over the number of words in the text.

$$(\sum NPs)/totalNumberWords$$

We decided to use only these criteria as features because our goal was to test the Information Nutritional Label on a concrete task.

3.1 Models

Each of our three runs uses its own model to compute a check-worthiness score. For each of our models, we preprocessed the data using the criteria previously described. We computed the five features for each sentence that has to be evaluated for check-worthiness. These sentences are then represented by a vector containing five features, one for each criterion score.

For our INL_SVM_RBF (primary run) and INL_SVM_Lin (first contrastive) runs, we decided to use the Support Vector Machine in sklearn [11] with the probability setting set to "True". We used a RBF kernel for INL_SVM_RBF run and a linear kernel for the INL_SVM_Lin run. For our INL_RF (second contrastive) run, we used the random forest classifier in sklearn.

To train our models, we used the three annotated debates provided by the clef2018-factchecking github repository[12].

To obtain a score of check-worthiness, we computed the probability for each sentence to be check-worthy using the classifiers. The score of a sentence was then normalized by the highest score obtained for this sentence divided by the highest probability computed, so that the scores are between 0 and 1.

4 Results

Seven teams submitted runs to this task for a total of 16 runs.

Table 1 presents the results of our three runs and the best submitted run according to the MAP measure, which is from the Copenhagen team [4].

Table 1. Results for each of our runs and the best run submitted. Values in parenthesis correspond to the ranks of our runs over the 16 that were submitted.

Name	**MAP**	**MRR**	**Mean Prec@1**
INL_SVM_RBF	.0632 (16)	.3775 (9)	.2857 (6)
INL_SVM_Lin	.0886 (12)	.4844 (5)	**.4286 (2)**
INL_RF	.0747 (15)	.2198 (15)	.0000 (14)
Copenhagen [4]	.1810 (1)	.6224 (1)	.5714 (1)

[11] http://scikit-learn.org/stable/modules/svm.html

[12] https://github.com/clef2018-factchecking/clef2018-factchecking/tree/master/data/task1/English

Overall, the INL_SVM_Lin run obtained better results than the INL_SVM_RBF run; that was somehow unexpected since non linear kernel have been shown to work better in other information retrieval applications. The INL_SVM_Lin run has been ranked twelfth according to the main measurement (Mean Average Precision), but obtained better rank when considering other measures: it is ranked fifth according to the Mean Reciprocal Rank and second according to the Mean Precision@1. These ranks mean that our INL_SVM_Lin run would be good if the purpose of the task was finding the most check-worthy claim instead of finding all the check-worthy claims. However, we need to deeper analyse the results to understand why.

Post-hoc experiments showed that the least important criterion is Technicality. This may be due to the fact that the method we use to compute this feature was meant to work with large texts and it is not appropriate for a single sentence. The most important criterion is Emotion. We can assume that a claim has greater chances to be check-worthy if it is highly emotional. The speaker thinks less about what he says and it is more likely that his claims are not fully accurate. We will check this hypothesis in future work.

Table 2 presents the weight of the 5 features for our INL_SVM_Lin model. The weights of the features for our INL_RF model are similar.

Table 2. Weights for the features used in our INL_SVM_Lin model.

Feature	**Weight**
Controversy	-2.08e-05
Factuality and Opinion	-1.03e-05 and 1.03e-05
Technicality	2.22e-06
Emotion	2.56e-05

5 Conclusion and perspectives for future works

In this paper we proposed three models to solve the CLEF2018 CheckThat! challenge (task 1 Check Worthiness) which deals with the evaluation of the check-worthiness of statements in political debates. We used random forest and support vector machine to learn models that make use of the Information Nutritional Label features [3]. We show that these models perform pretty well when considering the Mean Precision@1 measure, which ranks our run that uses a support vector machine with a linear kernel 2nd over 16 submitted runs.

We are currently working on better calculation of the five features. We would like to complete the representations of the texts by using content-based components like it is done in [5]. While the objective is different (virality prediction), some of the features may also be useful for the task tackled by CheckThat!. To improve more our models, we would also like to investigate the use of word-embedding since we are using successfully this approach in other tasks [7] and this approach also worked well according to Hansen et al. [4] in the CheckThat! context. As future work, we will also take in consideration the sentences around the one to be classified and who said these sentences.

Finally, we will test these models on other datasets such as social networks. For example, we will consider a Twitter-based dataset where each tweet would have a score indicating its worthiness for fact-checking taking into account hashtags and tweet sources.

References

1. Berger, J., Milkman, K.L.: What makes online content viral? Journal of marketing research **49**(2), 192–205 (2012)
2. Bird, Steven, E.L., Klein, E.: Natural language processing with python (2009)
3. Fuhr, N., Giachanou, A., Grefenstette, G., Gurevych, I., Hanselowski, A., Jarvelin, K., Jones, R., Liu, Y., Mothe, J., Nejdl, W., et al.: An information nutritional label for online documents. In: ACM SIGIR Forum. vol. 51, pp. 46–66. ACM (2018)
4. Hansen, C., Hansen, C., Simonsen, J., Lioma, C.: The Copenhagen Team Participation in the Check-Worthiness Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 Fact Checking Lab. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Working Notes, Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org, Avignon, France (September 2018)
5. Hoang, T.B.N., Mothe, J.: Predicting information diffusion on twitter–analysis of predictive features. Journal of Computational Science (2017), https://doi.org/10.1016/j.jocs.2017.10.010
6. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering **1**(1), 9–27 (1995)
7. Mothe, J., Ramiandrisoa, F.: IRIT at TRAC. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC), 27th International Conference on Computational Linguistics, COLIN 18. International Committee on Computational Linguistics (2018)
8. Nakov, P., Barron-Cedeno, A., Elsayed, T., Suwaileh, R., Marquez, L., Zaghouani, W., Atanasova, P., Kyuchukov, S., Da San Martino, G.: Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J.Y., Soulier, L., SanJuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Nineth International Conference of the CLEF Association (CLEF 2018). Lecture Notes in Computer Science (LNCS) 11018, Springer, Heidelberg, Germany (2018)
9. Nielsen, F.Å.: Afinn (mar 2011), http://www2.imm.dtu.dk/pubdb/p.php?6010
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)