TOBB-ETU at CLEF 2019: Prioritizing Claims Based on Check-Worthiness

Bahadir Altun and Mucahid Kutlu

TOBB University of Economics and Technology, Ankara, Turkey
{ialtun, m.kutlu}@etu.edu.tr

Abstract. In recent years, we witnessed an incredible amount of misinformation spread over the Internet. However, it is extremely time consuming to analyze the veracity of every claim made on the Internet. Thus, we urgently need automated systems that can prioritize claims based on their check-worthiness, helping fact-checkers to focus on important claims. In this paper, we present our hybrid approach which combines rule-based and supervised methods for CLEF-2019 Check That! Lab's Check-Worthiness task. Our primary model ranked 9^{th} based on MAP, and 6^{th} based on R-P, P@5, and P@20 metrics in the official evaluation of primary submissions.

Keywords: Fact-Checking · Check-Worthiness · Learning-to-Rank

1 Introduction

With the fast developing information retrieval (IR) technologies and social media platforms such as Twitter and Facebook, it is very easy to reach any kind of information we are looking for and share it with other people. Therefore, any information can be popular worldwide by the incredible sharing behavior of the Internet users. As a worrisome finding, a recent study [7] reports that false news spread eight times faster than true news.

Obviously, false news can have many undesired consequences and be very harmful for our daily life. Many researchers and journalists are trying to minimize the spread of misinformation and its negative impacts. For example, fact-checking websites, such as Snopes[1], investigate the veracity of claims spread on the Internet and publish their findings. However, these valuable efforts are not enough to effectively combat the false news because fact-checking is a very time-consuming task, taking around one day for a single claim [6].

Considering the vast amount of claims spread on the Internet on a daily basis and high cost of fact-checking, there is an urgent need to develop systems that

[1] https://www.snopes.com

Algorithm 1 Claim Ranking Algorithm

Input: Training Data TrD
Input: Test Data TD
$F_{TrD} = extract\_features(TrD)$
$model = train\_MART(F_{TrD})$
$F_{TD} = extract\_features(TD)$
$ranked\_claims = test(model, F_{TD})$
$ranked\_claims = apply\_rules(ranked\_claims, RULES)$
return $ranked\_claims$

assist fact-checkers to focus on the important claims instead of spending their precious time for less important claims.

In this paper, we present our method for CLEF-2019 Check That Lab!'s[3] Check Worthiness task[1]. We use a hybrid approach in which claims are ranked using a supervised method and then reranked based on hand-crafted rules. In particular, we use MART [5] learning-to-rank algorithm to rank the claims. Features we investigate include topical category of claims, named entities, part-of-speech tags, bigrams, and speakers of the claims. We also develop rules to detect the statements that are not likely to be a claim, and put those statements at the very end of our ranked lists. Our primary model ranked 9^{th} based on MAP, and 6^{th} based on R-P, P@5, and P@20 metrics in the official evaluation of primary submissions.

2 Proposed Approach

We propose a hybrid approach which uses both rule-based and supervised methods to rank claims based on their check-worthiness. Now we explain our approach in detail.

2.1 Hybrid Approach

We first rank claims using a supervised method and then rerank the claims using a rule-based method. **Algorithm 1** shows the steps in our proposed hybrid approach. For the supervised ranking phase, we investigated logistic regression (LR) and various learning-to-rank (L2R) algorithms including MART [5], RankBoost [4], and RankNet[2]. In our not-reported initial experiments, we observed that MART outperforms other L2R methods. Thus, we focus on only MART and LR in developing our primary and contrastive systems.

In the training dataset, we observed that many sentences contain phrases that are not likely to be a part of a check-worthy claim such as *thanks*. In addition, speaker of many statements are defined as *system* and these statements contain non-claim phrases such as *applause* indicating the actions of the audience during the debate. Thus, in rule-based reranking phase, inspired from [9], we set score of a sentence to the minimum score if it contains any of the *thanks*, *thank you*, *welcome*, and *goodbye* phrases, or 2) its speaker is *system*.

2.2 Features

The features we use can be categorized into five categories: 1) Named entities, 2) part-of-speech tags, 3) topical category, 4) bigrams, and 5) speakers. Now we explain our features used in our supervised methods.

Named Entities: Claims about people and institutions are likely to be check-worthy because people and institutions can be negatively affected by those claims. In addition, a check-worthy claim should be also verifiable. The location information, numerical values in claims are easy to verify because they provide unambiguous information. Therefore, detecting whether a statement is about a person or an institution, and existence of numerical values, location, and date information might be good indicators for check-worthy claims. Thus, we identify the named entities using Stanford Named Entity Tagger[2] which tags entities such as person, location, organization, money, date, time and percentage. We use a binary vector of size 7 representing the presence of each named entity type separately.

Part-of-speech (POS) Tags: Sentences without informative words are less likely to be check-worthy. POS tags of words might help us to detect the amount of information in a sentence. For instance, a sentence without any noun such as "That is great!" does not contain any information that can be fact-checked. Thus, we detect POS tags in the statements using Stanford POS toolkit[3]. For this set of features, we again use a binary feature vector of size 36 in which each feature represents existence or absence of a particular POS tag in the corresponding statement.

Topical Category: Topic of a claim might be an effective indicator for check-worthy claims [8]. For instance, a claim about celebrities might be less check-worthy than a claim about economics or wars. Therefore, we detect categories of statements using IBM-Watson's Natural Language Understanding Tool[4]. This categorization classifies every statement into topics such as finance, law, government, and politics, where each topic could be branched up to two levels of subtopics. We only use the main topic and their immediate subtopics, yielding 298 features in total. The topics are represented as binary features as we do for named entities and POS tags.

Bigram: Instead of using a large list of bigrams, we use a limited but powerful set of bigrams. In this set of features, we use bigrams that appear at least N times only in check-worthy or only in not check-worthy claims in the training set. We set N to 50 empirically based on our initial (not-reported) experiments, yielding 47 features. Some of these bigrams are "All right", "I know", and "go

[2] https://nlp.stanford.edu/software/CRF-NER.html
[3] https://nlp.stanford.edu/software/tagger.html
[4] https://www.ibm.com/watson/services/natural-language-understanding/

ahead".

Speaker List: A claim's check worthiness may depend on the person who makes it. Claims of influential people are more important than claims of random people. In addition, some people might make check-worthy claims more frequently than others. For instance, in the training dataset, we found that 3.8% of Donald Trump's statements and 2.6% of Hilary Clinton's statements are labeled as check-worthy, showing that Donald Trump is more likely to make check-worthy claims than Hilary Clinton. Therefore, we use the 33 speakers appearing in the training data as binary features.

3 Experimental Results

In this section, we present experimental results on both training and test data using different sets of features and machine learning algorithms.

The training and test datasets contain 19 and 7 files at varying lengths, respectively, where each file corresponds to transcription of a debate for the US 2016 presidential election. We perform 19-fold cross validation on the training data where each fold is a separate debate, and calculate mean average precision (MAP). We use Ranklib[5] library in our L2R approach and Scikit toolkit[6] for LR. We set the number of trees and the number of leaves parameters of MART to 50, and 2, respectively based on our initial experiments with different parameter configurations. We use default parameters for LR.

We first investigate the impact of each feature group using LR and MART. In particular, for each feature group we use all other features to build the model and observe how the performance would change without the corresponding feature group. **Table 1** shows MAP scores for various feature sets.

Table 1. MAP Scores using k-Fold Cross Validation on Training Data

Features	**MART**	**LR**
All	0.2193	**0.2198**
All - {*Named Entities*}	0.1897	0.2197
All - {*POS tags*}	0.2132	0.1967
All - {*Topical Category*}	0.1965	0.2116
All - {*Bigrams*}	0.2193	0.2083
All - {*Speaker List*}	**0.2196**	0.2094

From the results in Table 1, we can see that features have different impacts on the performance of MART and LR. Not using named entities and topical

[5] https://sourceforge.net/p/lemur/wiki/RankLib/
[6] https://scikit-learn.org/

category features cause around 2.96%, and 2.28% decrease in the performance of MART, respectively. However, for LR, POS tags seem the most effective features while named entities and topical categories have very small impact on the performance. Nevertheless, we achieve the best results when we use LR with all features (0.2198) while LR with All-{*Named Entities*}, MART with All-{*Speaker List*}, and MART with all features have also similar results.

For our primary submission, we selected MART algorithm with all features except speaker list. This is because its performance is very close to our best performing model and it does not use any speaker list which makes it more debate-independent model. For our contrastive-1 and contrastive-2 submissions, we elected MART and LR using all features.

For the evaluation on test data, we train our selected models with all training data. **Table 2** presents official performance scores of all our submissions based on various evaluation metrics. 12 groups submitted 25 runs in total. Our primary submission ranked 9^{th} among primary models based on MAP which is the official evaluation metric of the task. Our primary submission is also ranked 6^{th} based on R-P, P@5, and P@20 metrics, among primary models.

Table 2. Official Results for our Submitted Models

Model	MAP	RR	R-P	P@1	P@3	P@5	P@10	P@20	P@50
MART w/ All-{Speaker List} (Primary)	.0884	.2028	**.1150**	.0000	.0952	**.1429**	**.1286**	.1357	**.0829**
MART w/ All (Contr. -1)	.0898	.2013	**.1150**	.0000	**.1429**	.1143	**.1286**	**.1429**	**.0829**
LR w/ All (Contr. -2)	**.0913**	**.3427**	.1007	**.1429**	**.1429**	.1143	.0714	.1214	**.0829**

4 Qualitative Analysis

In this section, we take a deeper look into our results and conduct a qualitative analysis. We manually investigate the top 10 claims in our primary system's output in each test file. **Table 3** shows a sample of claims with their labels and ranks in our system's output.

Comparing Row 1 and 2, while both of the statements are about the change in unemployment, one of them is labeled as check-worthy while the other one is not. We also observe a similar issue in statements shown in Row 3 and 4 where both statements are about the amount of trade and use similar words with similar sentence structure but their labels are different. This might be due to many reasons. First, check-worthy claims can be subjective, making the task even more challenging. Second, we might need background information about the

Table 3. A sample of statements and their ranks in our primary system's output

Row	Speaker	Statement	Label	Rank	Debate
1	Trump	And Hispanic American unemployment has also reached the lowest levels in history.	1	2	20180131_state_union
2	Trump	Unemployment claims have hit a 45-year low.	0	3	
3	Trump	They do $531 billion with us.	1	9	20181015_60_min
4	Trump	We do $100 billion with them.	0	10	
5	Christie	He's lost $4,000 in the last seven years in his income because of this administration.	0	3	20160129_7_gop
6	Trump	But they had to pay over a billion dollars.	0	1	trump_emergency
7	Trump	I mean, were taking in billions and billions of dollars in tariffs from China.	0	2	
8	Rubio	But you still have hundreds of billions of dollars of deficit that you're going to have to make up.	0	9	20160311_12_gop
9	Trump	And we have a $505 billion trade deficit right now.	1	10	

claims because a claim which has been already discussed among people is more check-worthy than the one which is not discussed by anyone. The requirement of background information suggests that context the claim made and external resources such as web pages, social media posts about the claim can be useful to detect check-worthiness of claims.

Regarding the claims in Row 5 and 6, the statements contain verifiable claims and their topic is about economics which can be considered as an important topic. In addition, the statement in Row 5 has also an explicit accusation regarding to an administration. However, they are both labeled as not check-worthy. These samples provide further evidence that detecting check-worthiness of a claim requires utilizing external resources and its context instead of just using the claim itself.

Regarding the claims in Row 7, 8, and 9, we can see that all three claims are about trade but only claim in Row 9 is labeled as check-worthy. This might be because the amount of money mentioned in Row 7 and 8 are not exact numbers, making them hard to verify. On the other hand, the claim in Row 9 mentions an exact number about the trade deficit. While our features take account whether a claim has a numerical value or not, this suggest that we have to handle special cases where numbers are not exact.

5 Conclusion

In this paper, we described our methods for CLEF-2019 Check That! Lab's Check-Worthiness task and discuss results conducting a qualitative analysis. We proposed a hybrid approach which combines rule-based and supervised methods together. We first rank claims using MART algorithm, and then change the scores of some statements based on our hand-crafted rules. We investigated various features including topical category, POS tags, named entities, bigrams, and speakers of claims. Our primary model ranked 9^{th} based on MAP, and 6^{th} based on R-P, P@5, and P@20 metrics, among primary models. In our qualitative analysis we discussed that detecting check-worthiness is a subjective task and linguistic analysis of claims are not enough, requiring utilization of other resources to capture the background information about claims.

In the future, we plan to investigate linguistic features to capture the contextual information about the claims, and develop more effective hand-crafted rules.

References

1. P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, and G. Da San Martino. Overview of the clef-2019 checkthat! lab on automatic identification and verification of claims. task 1: Check-worthiness. CEUR Workshop Proceedings, Lugano, Switzerland, 2019. CEUR-WS.org.
2. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine learning (ICML-05)*, pages 89–96, 2005.
3. T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, and P. Atanasova. Overview of the clef-2019 checkthat!: Automatic identification and verification of claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, LNCS, Lugano, Switzerland, September 2019.
4. Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.
5. J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
6. N. Hassan, C. Li, and M. Tremayne. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th acm international on conference on information and knowledge management*, pages 1835–1838. ACM, 2015.
7. S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
8. K. Yasser, M. Kutlu, and T. Elsayed. bigir at clef 2018: Detection and verification of check-worthy political claims. In *CLEF (Working Notes)*, 2018.
9. C. Zuo, A. I. Karakas, and R. Banerjee. A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In *CLEF 2018 Working Notes. Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum*, volume 8995, pages 171–175. CEUR-WS. org, 2018.