# The University of Sheffield at CheckThat! 2020: Claim Identification and Verification on Twitter

Thomas McDonald, ZiQing Dong, Yingji Zhang, Rebekah Hampson,
James Young, Qianyu Cao, Jochen L. Leidner, and Mark Stevenson

Department of Computer Science, University of Sheffield, UK
{tmmcdonald1,zdong3,yzhang320,rgparnham1}@sheffield.ac.uk
{jrjyoung1,qcao8,mark.stevenson}@sheffield.ac.uk, leidner@acm.org

**Abstract.** The spread of misinformation online has been gathering pace in recent years which has led to research into automatic methods for claim verification. The COVID-19 pandemic presents a unique challenge due to the large amount of inaccurate information being shared on social media platforms. This paper describes the University of Sheffield's entry to the CLEF 2020 CheckThat! Lab, which focuses on the problems of determining check-worthiness and verification of claims found in tweets, including those related to COVID-19. For the Tweet Check-Worthiness Task (Task 1), we found that TF-IDF term weightings used by a Random Forest model outperformed more complex approaches employing Word2Vec embeddings and recurrent neural networks, and for the Claim Retrieval Task (Task 2), we found that BM25 similarity score weightings based on TF-IDF term weightings with a Support Vector Machine classifier scoring model outperformed other methods making use of cosine and Euclidean similarity metrics, and regression-based scoring models.

## 1 Introduction

The rise of social media platforms has changed the way in which people across the world consume media, including news: these platforms allow individual users to publish and disseminate content, including disinformation and so called 'fake news'. Social media platforms can act as a breeding ground for disinformation and allow controversial posts to spread to millions of users in a very short space of time. Sites such as Instagram, Facebook and Twitter are popular platforms for individuals looking to spread disinformation [18]. The 2020 *CheckThat! Lab* took place in the context of the COVID-19 pandemic which swept across the globe during the year. The spread of disinformation in the midst of such a global health emergency can cause physical harm to members of the public through increased panic, unsafe actions and confusion [15]. Research surrounding the spread of misinformation pertaining to COVID-19 has concluded that different social media platforms contain different levels of misinformation, although misinformation spreads in a similar way on each platform [5].

The primary focus of the 2020 CheckThat! Lab was verification of tweets pertaining to the COVID-19 pandemic [2, 21]. The University of Sheffield participated in Tasks 1 and 2. The aim of Task 1, entitled ‘Tweet Check-Worthiness’, is to create a system which can rank a set of tweets according to how likely they are to require manual fact-checking. The aim of task 2, entitled ‘Claim Retrieval’, is to rank a set of previously verified claims given an input claim such that those which verify the input claim are ranked highly.

## 2 Related Work

### 2.1 Task 1

Task 1 focuses on ranking tweets based on whether the claims they contain need to be fact checked. Similar tasks have been run at previous versions of the CheckThat! Lab and entries have applied a range of approaches with varying levels of complexity. The winning approach to Task 1 at CheckThat! 2019 [9] used Word2Vec embeddings [19] and syntactic dependencies as features, fed into a recurrent neural network, with Long Short Term Memory (LSTM) units [12]. Other entries to the 2019 task included the TOBB ETU team’s hybrid approach that combined supervised learning with handcrafted rules to rank the statements several times [1]. More traditional approaches were implemented by other teams, such as UAICS, who tested multiple ML algorithms including Decision Trees, Naïve Bayes, SVM, Random Forest, Logistic Regression and Multi-Layer Perceptron models [17]. Tokenisation and stop-word removal were both performed as preprocessing steps before TF-IDF weighting was used to extract features for training. Their Naive Bayes model reached first place in the Lab for the P@1, P@3, P@10 and RR evaluation metrics.

### 2.2 Task 2

Task 2 focuses on identifying whether previously verified claims contain information that can help support the verification of a new claim. Approaches are given an input claim and a set of claims that have previously been verified. The goal is to produce a ranking of the previously verified claims such that the ones that verify the input claim appear early in the ranking. This is a new task and was not included in the two previous iterations of CheckThat! Nevertheless, there are many similarities between it and Task 2A from CheckThat! 2019 as both tasks focus on document ranking, therefore many techniques utilised by teams during Task 2A last year are relevant to this task. Previous entries included two main approaches to the problem: learning-to-rank (L2R) and text classification for ranking problems [11].

L2R methods [16] can be divided into three distinct approaches: point-wise, pair-wise and list-wise. The pairwise approach is perhaps the most popular currently due to the fact that it balances performance with computational cost.

This technique approximates the L2R problem using binary classification between pairs of documents within the collection. Two teams utilised the pairwise L2R technique at CheckThat! 2019 [8], [10]. One achieved the highest normalised Discounted Cumulative Gain at cutoff 20 (nDCG@20) score of 0.55, although this was still lower than the score achieved by the baseline provided by CLEF (0.61).

Text classification algorithms can also be applied to ranking problems. Such algorithms can be broadly split into two main categories: those based on traditional machine learning models and those which employ deep learning. Traditional approaches typically exhibit better performance on smaller datasets, whereas deep learning models tend to be the optimal approach when a large amount of data is provided. In last year's task, one team utilised the BERT pre-trained neural language model [7], achieving an nDCG@20 score of 0.14 which was much lower than the baseline, as well as the best result obtained by any team using L2R.

## 3 Task 1: Tweet Check-worthiness

### 3.1 Methodology

The focus of Task 1 was the ranking of a collection of tweets about a given topic based on their check-worthiness [3]; in this case, the topic was the COVID-19 pandemic. 'Check-worthy' in this context simply means that the content of the tweet is of questionable veracity, could mislead a large number of people, and should be fact-checked manually. A system which can classify tweets as such and automatically disregard tweets which are not check-worthy could be extremely useful, as manual fact-checking is a time intensive process.

Multiple different models and approaches were tested for Task 1, but the first step taken was to preprocess the English data provided. Firstly, all URLs were stripped from the tweets using a regular expression, emojis and punctuation were removed and the text was converted to lowercase before tokenizing each tweet. Natural Language Toolkit (NLTK) [4] was then used to remove frequently used stopwords from the list of tokens, before Porter stemming and lemmatization were applied in order to conflate similar terms [20].

In order to establish a baseline, a number of different classifications models were tested, including Naïve Bayes, Random Forest, Gradient Boosting, AdaBoost, k-means and a Support Vector Machine model. Said models were trained on term frequency-inverse document frequency (TF-IDF) weights obtained from the preprocessed text, using unigrams, bigrams and trigrams. After evaluation on the validation set provided, the best performing model from the selection tested was the Random Forest classifier. Random Forest classifiers are versatile models, but the main drawback associated with them is that they are difficult to interpret due to the fact they make predictions using a large ensemble of individual decision trees.

Following the development of this initial system, a pre-trained Word2Vec embedding based on the GoogleNews dataset was tested for constructing the document vectors in attempt to improve the accuracy of our baseline classifier. A deep learning approach was also tested, employing an LSTM recurrent neural network. The network consisted of several layers including an input layer, word embedding layer, LSTM layer and fully connected layer as well as an output layer. After constructing the neural network, Adam [14] and BCELoss[1] were used as the optimizer and loss function respectively, and dropout was used to avoid overfitting. In an attempt to overcome the issues associated with training a very complex, deep model on a limited amount of data, a fastText classification model was also tested [13].

**Table 1.** Task 1 Prediction Example

| | |
|---|---|
| **Original Tweet** | I just landed at JFK after reporting on #coronavirus in Milan and Lombardy —the epicenter of Italy's outbreak— for @vicenews. I walked right through US customs. They didn't ask me where in Italy I went or if I came into contact with sick people. They didn't ask me anything. |
| **Preprocessed Tweet** | land jfk report coronavirus milan lombardy —the epicenter italy outbreak— vicenews walk right us custom ask italy go come contact sick people ask anything |
| **FastText Score** | 0.512 |
| **FastText Prediction** | 1 (i.e. check-worthy) |
| **True Label** | 1 (i.e. check-worthy) |

Table 1 illustrates an example of how our fastText system arrives at a prediction. The original tweet undergoes a series of preprocessing before being fed into the model, which outputs the probability of the tweet in question being check-worthy. In this particular case, the predicted probability is above 0.5, therefore the system determines that the tweet is indeed check-worthy, and this aligns with the true label given in the validation dataset.

### 3.2 Task 1 Results

Based on validation set performance, the approach based on fastText was submitted as our primary system with the Random Forest classifier trained on TF-IDF term weightings acting as our contrastive submission. Other approaches based

[1] `https://pytorch.org/docs/master/generated/torch.nn.BCELoss.html` (accessed 2020-07-14)

on Word2Vec embeddings and deep recurrent neural networks such as LSTMs yielded poor results on the validation set and were not submitted.

**Table 2.** Task 1 Results for both validation and official test data. Validation results were generated using the initial release of training and validation data while the submitted models were trained on the second release[2]

| Model | Validation | | | Test | | |
|---|---|---|---|---|---|---|
| | **MAP** | **P@10** | **P@20** | **MAP** | **P@10** | **P@20** |
| **fastText** (Primary) | 0.812 | 0.800 | 0.900 | 0.475 | 0.200 | 0.350 |
| **Random Forest** (Contrastive) | 0.810 | 0.900 | 0.900 | 0.646 | 0.800 | 0.600 |

Results in Table 2 show that our primary submission performed poorly, ranking last out of all systems submitted for Task 1. The fastText model exhibited a large drop in performance when used to evaluate the test set used for ranking, compared with the validation set used for model selection, which leads us to conclude that a considerable amount of overfitting occured during training. This could have been mitigated by choosing a simpler form of fastText model, perhaps only considering bi-grams instead of 4-grams. Conversely, our contrastive submission performed much better than anticipated, outperforming a number of BERT-based models amongst others.

# 4 Task 2: Claim Retrieval

## 4.1 Methodology

For Task 2 our objective was to construct a system that could rank all verified claims according to their relevance to a given tweet, and return the most relevant claim with its score generated by the system. Learning-to-rank (L2R) techniques are a primary candidate for solving such a problem. In this task, we chose to take the point-wise approach as the Task 2 dataset includes binary labels ('relevant' and 'not relevant').

In order to apply the L2R algorithm, firstly several features measuring similarity were generated from the raw data: cosine similarity (Eq. 1), BM25 score (Eq. 2) and simple Euclidean distance. Cosine similarity is computed as

$$\cos(Q, D) = \frac{\overrightarrow{Q} \times \overrightarrow{D}}{\|Q\|\|D\|} \tag{1}$$

where $D$ represents a verified claim and $Q$ a tweet. The BM25 score is computed as

[2] `https://github.com/sshaar/clef2020-factchecking-task1` (accessed 2020-07-14)

$$BM25(Q, D) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i, D) \times (k_1 + 1)}{f(q_i, D) + (1 - b + b \times \frac{|D|}{|D|_{avg}})} \tag{2}$$

where $q_i$ are the terms in claim $Q$, $f(q_i, D)$ the term frequency of $q_i$ in $D$, $|D|$ is the number of terms in the verified claim and $|D|_{avg}$ is the average length of a verified claim within the collection. $b$ and $k_1$ are free parameters.

Two combinations of these features were explored. The first combination consisted of Euclidean distance and cosine similarity based on word embeddings. For the second combination TF-IDF was used to generate cosine similarities between tweets and verified claims then combined using BM25 scores. In addition, both contents and titles of the verified claims were used to generate the introduced features against the tweets to avoid losing any potentially useful information.

Multiple different machine learning models were implemented as a scoring model, such as Logistic Regression, Random Forest, Gradient Boosted Trees, Linear SVM and Linear Regression. The sum of multiplying each introduced feature's value by its corresponding coefficient was considered as the final score. Class weighting was applied during the training process to accommodate for the high class imbalance in the training data. Approaches based on a linear Support Vector Machine (SVM), logistic regression and linear regression were submitted.

### 4.2 Task 2 Results

**Table 3.** Task 2 Results. All models were trained using the third version of the Task 2 data released by the task organisers[3]

| Model | Validation | | Test | |
|---|---|---|---|---|
| | **MAP@5** | **P@5** | **MAP@5** | **P@5** |
| **Linear SVM** (Primary) | 0.622 | 0.125 | 0.807 | 0.162 |
| **Logistic Regression** (Contrastive-1) | 0.612 | 0.123 | 0.772 | 0.155 |
| **Linear Regression** (Contrastive-2) | 0.586 | 0.118 | 0.767 | 0.154 |

The models which made use of Word2Vec embeddings and Euclidean distance exhibited very poor performance on the validation set provided, far below the random baseline, therefore all three of our submissions for Task 2 made use of TF-IDF weightings and BM25 scores, and simply employed different scoring models. Table 3 provides a summary of the performance of these techniques.

[3] `https://github.com/sshaar/clef2020-factchecking-task2` (accessed 2020-07-14)

Whilst none of the systems ranked within the top half of the leaderboard, the overall results from Task 2 were more promising than Task 1, with all three systems exhibiting reasonable performance. As we predicted based on validation set testing, the Linear SVM scoring model outperformed both the Logistic and Linear Regression scoring models.

## 5 Conclusion and Future Work

This paper described our submissions to Tasks 1 and 2 of the CLEF 2020 Lab "CheckThat!". The aim in constructing a system for Task 1 was to evaluate the check-worthiness of a series of COVID-19 related tweets. Our approach was inspired by the 2019 edition of the Lab, applying a series of preprocessing and feature engineering steps followed by the implementation of Random Forest and fastText models. Unexpectedly, the best performing model we obtained was a simple approach consisting of TF-IDF term weightings and a Random Forest model, which ranked 17th out 27 systems. We also participated in Task 2 of the Lab, whose aim was to identify previously verified claims that contain information that could support the verification of a new claim. As with Task 1, methods used by teams in the 2019 Lab led us to our chosen approach, a point-wise L2R system. This involved similar preprocessing steps to those used in Task 1, followed by the extraction of BM25 scores and TF-IDF term weightings from the data, which were used to train a range of different forms of scoring model. Our highest scoring model employed a SVM, ranking 13th out of 22 systems.

Perhaps the simplest avenue for improvement of our systems would be to supplement the training data with external Twitter data, or other labeled text. The datasets provided for the task were relatively small, containing around 1,000 tweets or less, and it is challenging to train complex models on datasets of this size. Other possibilities could include exploring more descriptive features to capture more information about the tweets, such as syntactic dependencies, numbers of mentioned named entities, links and hashtags, or specific indicators for credibility previously proposed for blogs [6]. Finally, we expect replacing LSTMs by a pre-trained language model would would likely result in superior performance, and such techniques tend to perform well even with limited amounts of data.

## References

1. Altun, B., Kutlu, M.: TOBB-ETU at CLEF 2019: Prioritizing claims based on check-worthiness. In: CLEF (2019)
2. Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., Sheikh Ali, Z.: Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. LNCS (12260), Springer (2020)
3. Barron-Cedeno, A., Elsayed, T., Nakov, P., Martino, G.D.S., Hasanain, M., Suwaileh, R., Haouari, F.: CheckThat! at CLEF 2020: Enabling the Automatic Identification and Verification of Claims in Social Media (2020)

4. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit. O'Reilly, Sebastopol, CA, USA (2009)
5. Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A.: The COVID-19 Social Media Infodemic (2020)
6. Conrad, J.G., Leidner, J.L., Schilder, F.: Professional credibility: Authority on the web. In: Proceedings of the 2nd ACM workshop on Information credibility on the web. pp. 85–88. WICOW (2008)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL 2019. pp. 4171–4186. ACL, Minneapolis, MN, USA (2019). https://doi.org/10.18653/v1/N19-1423
8. Favano, L., Carman, M.J., Lanzi, P.L.: TheEarthIsFlat's Submission to CLEF'19CheckThat! Challenge. In: CLEF (2019)
9. Hansen, C., Hansen, C., Simonsen, J., Lioma, C.: Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss. In: 20th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2019 Conference and Labs of the Evaluation Forum. vol. 2380 (2019)
10. Haouari, F., Ali, Z.S., Elsayed, T.: bigIR at CLEF 2019: Automatic Verification of Arabic Claims over the Web. In: CLEF (2019)
11. Hasanain, M., Suwaileh, R., Elsayed, T., Barrón-Cedeño, A., Nakov, P.: Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 2: Evidence and Factuality. In: CLEF (2019)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
13. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification (2016)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
15. Krause, N.M., Freiling, I., Beets, B., Brossard, D.: Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19. Journal of Risk Research (2020, to appear)
16. Liu, T.Y.: Learning to Rank for Information Retrieval, Foundations and Trends in Information Retrieval, vol. 3. Now Publishers, Delft, The Netherlands (2009)
17. Lucia-Georgiana Coca, Ciprian-Gabriel Cusmuliuc, A.I.: CheckThat! 2019 UAICS. In: CLEF (2019)
18. Marwick, A., Lewis, R.: Media Manipulation and Disinformation Online (2017), Data & Society Research Institute
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119 (2013)
20. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
21. Shaar, S., Nikolov, A., Babulkov, N., Alam, F., Barrón-Cedeño, A., Elsayed, T., Hasanain, M., Suwaileh, R., Haouari, F., Da San Martino, G., Nakov, P.: Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media (2020)