ClaimHunter: An unattended tool for automated claim detection on Twitter

Javier Beltrán
R&D Department
Newtral
Madrid, Spain
javier.beltran@newtral.es

Rubén Míguez
R&D Department
Newtral
Madrid, Spain
ruben.miguez@newtral.es

Irene Larraz
Fact-checking
Newtral
Madrid, Spain
irene.larraz@newtral.es

ABSTRACT

As political campaigns have moved from traditional media to social networks, fact-checkers must also adapt how they are working. The explosion of information (and disinformation) on social networks makes impossible to manually fact-check each piece of data. With this reality in mind, Newtral, a fact-checking organization, has developed its own automated monitor tool for Twitter: ClaimHunter.

Recently, deep learning approaches have obtained very high performance across many different NLP tasks. Automated claim detection is not an exception. These models are showing promising results on fact-checking scenarios without task-specific feature engineering. Based on the BERT architecture, ClaimHunter AI models shown a 80% F1 score tested on real-life scenarios with expert fact-checkers. Through a simple UI interface deployed on Slack, ClaimHunter notifies journalists and gets feedback from their day-to-day work to improve the final performance of the algorithm. Launched 6 months ago, ClaimHunter has processed more than 130.000 tweets expanding Newtral's operation beyond national politicians to the regional and local level. The number of reviewed claims per day have increased by a multiplicative factor of 10 since the adoption of the ClaimHunter tool by Newtral's fact-checking team.

This paper focuses on explaining the challenges of building such a system inside of a fact-checking organization: data labelling and fact-checker alignment, system architecture, testing and deployment of underlying models, continuous feedback and model refinement.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; Natural language processing; Information extraction;

• **Human centered computing → HCI**; HCI design and evaluation methodologies; Field studies;

• **Information systems → Information retrieval**; Evaluation of retrieval results; Presentation of retrieval results;

KEYWORDS

Automated fact-checking, Claim detection, BERT

1 Introduction

In the journalism domain we define fact-checking as the task of assessing if a claim made by a public figure is true or not. This is a complex activity normally performed by trained professionals (fact-checkers) that must evaluate known facts and data published by official institutions to reach a final verdict.

The fact-checking process involves four main steps: 1) monitoring of relevant sources; 2) spotting facts; 3) data verification; and 4) publication [1]. This paper focuses on the first two steps.

Monitoring and spotting claims is a time-consuming task which inevitably must be automated. Despite the importance of this component for fact-checking, automated claim detection algorithms are still in an early stage [2][3]. However, the increased demand for fact-checking and the recent advances of deep learning techniques for NLP has stimulated a rapid progress in developing tools and systems to automate parts of this task [4][5].

To automatically decide which information is fact-checked, a definition of a check-worthy claim must be agreed. However, the concept of check-worthiness lacks of an agreed definition on scientific literature resulting in inconsistent and unreliable datasets [6]. Frequently researchers build training datasets based only on claims published by fact-checkers on their web sites. This conceptualization is flawed by design. Per each published content, fact-checkers have to spot and review dozens of potentially check-worthy claims. Only by having access to this internal work large non biased datasets can be built. Besides, to precisely identify a check-worthy statement is not an easy task. Expert knowledge is needed. Check-worthy claims normally are: objective, do not contain information that is common knowledge, establish some sort of comparison and are verifiable with data. As the political discourse moved to Twitter, tweets have become the object of study for fact-checkers across the world. ClaimHunter is our solution to automate the detection of relevant tweets. In this work, we define a tweet as check-worthy if there is at least one check-worthy claim on it.

The structure of this paper is as follows: We begin (in Section 2) by reviewing the most relevant initiatives in the automated claim detection field. We then explore (in Section 3) how ClaimHunter was built, including the training dataset, the system

KnOD'21 Workshop - April 14, 2021

architecture and an overall description of its UI. Further (in Section 4) we describe our learning model, evaluate it with performance metrics and explore its evolution through time. Finally (in Section 5) we draw our main conclusions after testing the system for 6 months and outline some research lines for the near future.

2 Related work

One of the first and most well-known approaches to claim detection is ClaimBuster [7]. It is built on a large annotated dataset of factual sentences taken from American presidential debates. It uses a machine learning algorithm based on SVM combining TF-IDF, POS tagging and NER features. Another known approach to claim detection is ClaimRank [4] supporting both English and Arabic. It is built on a dataset of factual sentences published by 9 different fact-checking organizations and applies a multi-task learning setup. Main novelty on this paper is the inclusion of a variety of contextual and discourse-based features. ClaimRank can mimic the claim selection strategies of each of them or the union of them all.

In [8] authors collaborated with FullFact, an independent fact-checking organization, to create a classifier based on universal sentence representations. The system was tested with real fact-checkers through a live feed of transcripts from TV called “Live”. Squash [9], system developed by Duke Reporters’ Labrom, is another proposal to live fact-checking where ClaimBuster algorithms are combined with ElasticSearch to promote real-time search of previously verified claims.

From a brief review of system architectures for automated fact-checking in the scientific literature we observe that a combination of deep neural networks (DNNs), non-DNNs and heuristic approaches are commonly employed. The Fact Extraction and VERification (FEVER) dataset [10] enables the development of data-driven neural approaches to the automatic fact checking task. Additionally, the FEVER Shared Task [11] introduced a benchmark, the first of this kind, to evaluate both evidence retrieval and claim verification tasks. The CheckThat! Lab at the Conference and Labs of the Evaluation Forum (CLEF), different research groups compete to create claim verification models. The workshop proposes four complementary tasks, offered in English, Arabic and Spanish. One of those tasks focuses on identifying which tweets in a Twitter stream are worth fact-checking [12]. Latest studies on this topic explore the fusion of syntactic features and BERT embeddings, to classify check-worthiness of tweets with promising results [13].

In ClaimHunter, we have fine-tuned a BERT model for the check-worthiness task using a large dataset (+30K tweets) annotated by expert fact-checkers. As far as we know this is the first work where such a model has been developed by fact-checkers and tested on real-life scenarios for a 6 month-period. Next sections describe how this system works.

3 System architecture

ClaimHunter is a monitoring tool for Twitter which accelerates the traditional fact-checking process by automatically detecting check-worthy content. The behavior of ClaimHunter consists of the following steps, also summarized in Figure 1:

1. Expert fact-checkers establish the Twitter accounts to monitor based on their public relevance.
2. ClaimHunter retrieves tweets from selected accounts in real time via the Twitter API.
3. ClaimHunter’s detector classifies tweets as positive if they contain a check-worthy claim or negative otherwise.
4. Positive tweets are sent to Slack. A small fraction of negative tweets, chosen randomly, is also sent (see Section 3.1).
5. Fact-checkers review the tweets classified as check-worthy and confirm or reject the predicted labels.
6. Both predicted label and manual feedback from fact-checkers are stored in the database.
7. New labeled samples from step 6 are used to retrain the model.

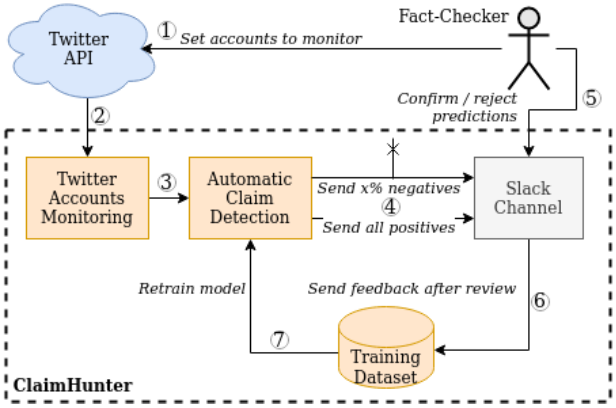

Twitter API
① Set accounts to monitor
Fact-Checker
Confirm / reject predictions
⑤
②
Twitter Accounts Monitoring
③
Automatic Claim Detection
Send x% negatives
④
Send all positives
Slack Channel
Retrain model
Send feedback after review
⑥
⑦
Training Dataset
ClaimHunter

Figure 1: High-level diagram of ClaimHunter

3.1 Dataset

ClaimHunter follows a supervised learning approach. Supervised machine learning requires annotated datasets for the objective task but human labeling is expensive and slow. Journalism expertise is needed to create a good quality dataset. Besides, claim detection is a highly unbalanced classification problem where not check-worthy claims (negative class) are by far more common than check-worthy ones (positive class). Our empirical estimate shows that only 10-15% of our Twitter feed were check-worthy tweets.

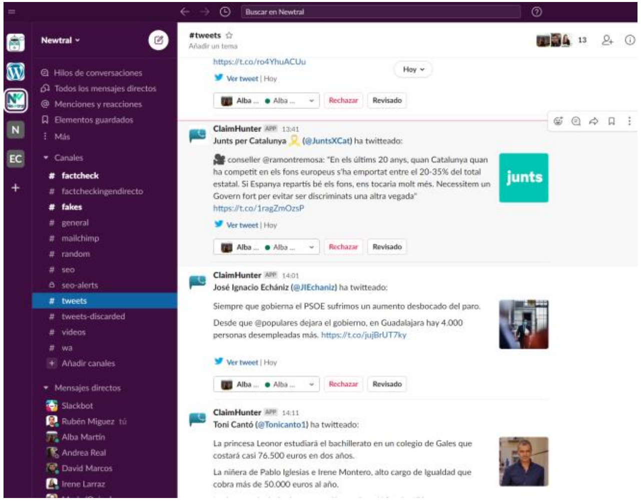
Our initial model was built on a dataset of 5.000 tweets manually annotated by 3 fact-checkers independently. Only tweets considered check-worthy by at least 2 fact-checkers were labeled as check-worthy. We followed an iterative approach launching new model releases when new data was available through the feedback loop integrated in the system. After 6 months of ClaimHunter running at Newtral, our dataset has increased from 5K to more than 30K tweets.

Beside positive tweets, ClaimHunter sends to Slack a small random fraction of the negative predictions that is also reviewed by fact-checkers. This prevents from biasing the dataset by adding only tweets originally predicted as positive. Only a fraction is sent because negative tweets are the majority class, so we are progressively undersampling it by making the check-worthy tweets more representative for training. The negative fraction is a predefined parameter set to 30%. Fact-checkers' selection of which claim to review and which to ignore can be biased to their interest, expertise, workload and other contextual factors. To minimize this issue, journalists were asked to label claims based on their factuality and not their journalistic relevance.

3.2 Automated ClaimDetection in the newsroom

Developed as a Slack app, ClaimHunter monitors Twitter accounts and sends alerts to a private channel (#tweets by default) when a new tweet is classified as check-worthy. Newtral's fact-checkers review its content and give feedback on its check-worthiness as part of their daily workflow. The UI offers three options:

- **Rejected**: Prediction was wrong. The selected tweet does not contain any claim.
- **Reviewed**: Prediction was right. The selected tweet contains a claim but there is no interest in publishing a fact-check on it.
- **Selected**: Prediction was right. The selected tweet contains a claim and the fact-checker proposes it to publish a fact-check.



https://t.co/ro4YhuACUu

Ver tweet | Hoy

Alba ... ● Alba ... Rechazar Revisado

ClaimHunter APP 13:41
Junts per Catalunya (@JuntsXCat) ha twitteado:

conseller @ramontremosa: "En els últims 20 anys, quan Catalunya quan ha competit en els fons europeus s'ha emportat entre el 20-35% del total estatal. Si Espanya repartís bé els fons, ens tocaria molt més. Necessitem un Govern fort per evitar ser discriminats una altra vegada" https://t.co/1ragZmOzsP

Ver tweet | Hoy

Alba ... ● Alba ... Rechazar Revisado

ClaimHunter APP 14:01
José Ignacio Echániz (@JIEchaniz) ha twitteado:

Siempre que gobierna el PSOE sufrimos un aumento desbocado del paro.

Desde que @populares dejara el gobierno, en Guadalajara hay 4.000 personas desempleadas más. https://t.co/jujBrUT7ky

Ver tweet | Hoy

Alba ... ● Alba ... Rechazar Revisado

ClaimHunter APP 14:11
Toni Cantó (@Tonicanto1) ha twitteado:

La princesa Leonor estudiará el bachillerato en un colegio de Gales que costará casi 76.500 euros en dos años.

La niñera de Pablo Iglesias e Irene Montero, alto cargo de Igualdad que cobra más de 50.000 euros al año.

Figure 2: Basic example of ClaimHunter UI

The category "Reviewed" allows to include a bigger number of check-worthy claims to the training dataset, independently of their potential for publication.

When a fact-checker checks one tweet as "Selected", a copy of the tweet is automatically sent to a new channel (#fact-check by default). The head of the fact-checking unit reviews content in this channel and make a final decision on which content is promoted to a fact-check story. Rejected and reviewed tweets are removed from the Slack feed.

4 Claim Detection experiments

ClaimHunter tackles the claim detection problem as a binary classification problem with a high unbalance on the positive class. However, this issue can be attenuated via the feedback mechanism described above. We follow standard experimentation practices, splitting the dataset into training (80%) validation (10%) and test set (10%). The test set is used for evaluation only, being our main evaluation metric the F1 score of the positive class, which corresponds to the harmonic mean between Precision and Recall.

We propose a transfer learning approach to build our claim detection classifier. We leveraged our dataset to progressively fine-tune a pre-trained XLM-RoBERTa[1] [15] model for the claim identification task. This model is based on BERT and hence we follow the recommended practices for fine-tuning a BERT architecture for text classification [16]. The following hyperparameters were adjusted on the validation dataset: epochs = 2, batch_size = 32 and learning_rate = 2e-5. Adam with weight decay was used as optimizer.

Our final dataset contains 31.883 tweets by Spanish representatives and political parties, 32% of which are labeled positive. While most tweets are written in Spanish, there is a small fraction in Catalan, Galician and Basque, due to the co-existence of several co-official languages in Spain. We are not filtering these out because: 1) XLM-RoBERTa is a multilanguage model and 2) they are valuable for our use case.

To better assess the quality of our model, we compare it with two different baselines:

- **LR-NNLM**: A logistic regression model which uses NNLM sentence embeddings[2] as a feature extraction strategy [17].
- **SVM-TFIDF**: A SVM with linear kernel where the features are a bag-of-words model limited to the 20.000 most common unigrams and the feature values are scaled with the TF-IDF transformation.

Hyperparameters such as the vocabulary size and the degree of regularization were adjusted on the validation set and the best model was selected during training, so they can be compared fairly. Table 1 summarizes the main results achieved during our test, comparing our final candidate model with the two proposed baselines.

[1] We have used the model xlm-roberta-base available on the Huggingface repository through the library transformers==4.2.2.

[2] These embeddings are available at Tensorflow Hub https://tfhub.dev/google/nnlm-es-dim128/2

Model	Precision	Recall	F1
LR-NNLM	66.95%	53.45%	59.44%
SVM-TFIDF	69.99%	64.68%	67.23%
XLM-RoBERTa	75.23%	85.41%	80.00%

Table 1: Precision, Recall and F1 scores for the proposed model and the baselines on the test set

Our fine-tuned XLM-RoBERTa model overcomes proposed baselines in both Precision and Recall metrics. This means that we recover more check-worthy tweets and at the same time we are making fewer mistakes. Additionally, to this evaluation on a fixed test set, we have been monitoring the precision achieved by the classifier through time over the period we have being using ClaimHunter at Newtral. Figure 3 shows the evolution of the precision metric over the last 20 weeks of 2020.

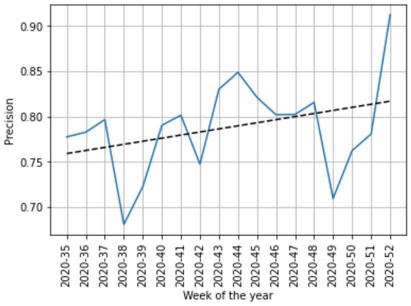


Figure 3: Evolution of the precision on the positive class over the last 20 weeks of the project

4.1 Discussion of results

The advantages of a fine-tuning approach go beyond getting good performance metrics with a limited dataset. Fine-tuning a model is faster than training from scratch (our best results are achieved after only 2 epochs) which allow us to retrain often and iterate quickly over the dataset. By keeping low the costs to retrain the model we can integrate early on the process the fact-checkers feedback. Quick iterations help us to correct mistakes in the initial datasets and quickly evolve the model to new topics of conversation. Another advantage comes from the fact that XLM-RoBERTa is pre-trained on a massively multilingual dataset of more than 100 languages. If the model retains its multilingual knowledge after fine-tuning on a single language, zero-shot claim identification in other languages could be feasible. Although we lack datasets of multilingual tweets for a rigorous evaluation, we have observed tweets in English, Catalan and Galician as being correctly detected as check-worthy by our model. The lexical similarity that Catalan and Galician maintain with Spanish makes that the model provides particularly good results for these languages without annotated datasets that could be harder to obtain in these lower resource languages. This is a desirable outcome for use cases like ours. In Spain, the political debate happens in several languages and newsrooms need fact-checking tools capable of working on several languages at a time. Quantitative evaluation is needed to validate our expectations regarding the multilingual capabilities of our developed model.

5 Future directions

Our iterative approach to model development has shown positive results over time in real-life scenarios. ClaimHunter was designed as an internal tool for Newtral but, in February 2021, we have opened it to other fact-checking agencies. Fact-checkers from Chile, Mexico, Ecuador and Colombia are currently testing ClaimHunter in different political contexts. Our goal is to check whether the model generalizes properly and satisfies the mixed criteria of different agencies in different countries. Besides, we are building a more rigorous benchmark to test its multilanguage capabilities on Spanish co-official languages. In the future we also plan to expand ClaimHunter capabilities to other social networks as Facebook and Instagram.

REFERENCES

[1] Mevan Babakar and Will Moy. 2016. The State of Automated Factchecking. Technical report. Retrieved from https://fullfact.org/blog/2016/aug/automated-factchecking/

[2] Naeemul Hassan, Bill Adair, James Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The Quest to Automate Fact-Checking. In *Proceedings of the 2015 Computation + Journalism Symposium*.

[3] Lucas Graves. 2018. Understanding the Promise and Limits of Automated Fact-Checking. Technical report, Reuters Institute, University of Oxford. Retrieved from https://reutersinstitute.politics.ox.ac.uk/our-research/understanding-promise-and-limits-automated-fact-checking

[4] Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting CheckWorthy Claims in Arabic and English. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[5] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking: A Survey. arXiv preprint arXiv:2011.03870.

[6] Liesbeth Allein and Marie-Francine. 2020. Checkworthiness in Automatic Claim Detection Models: Definitions and Analysis of Datasets. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*. Springer, Cham, 1-17.

[7] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. In *Proceedings of the VLDB Endowment 10*, 12 (2017), 1945–1948

[8] Lev Konstantinovskiy, Oliver Price, Mevan Babakar and Arkaitz Zubiaga. 2018. Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. arXiv preprint arXiv:1809.08193.

[9] Bill Adair. 2020. Squash report card: Improvements during State of the Union … and how humans will make our AI smarter for consistent automated claim detection. Retrieved February 9, 2021 from https://reporterslab.org/squash-report-card-improvements-during-state-of-the-union-and-how-humans-will-make-our-ai-smarter/.

[10] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*.

[11] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and verification (fever) shared task. arXiv preprint arXiv:1811.10971.

[12] Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh and Fatima Haouaril. 2018. CheckThat! at CLEF 2020: Enabling the Automatic Identification and Verification of Claims in Social Media. ECIR 2020. Lecture Notes in Computer Science, vol 12036. Springer, Cham. https://doi.org/10.1007/978-3-030-45442-5_65

[13] Gullal Cheema, Sherzod Hakimov and Ralph Ewerth. 2020. Check_square at CheckThat! 2020: Claim Detection in Social Media via Fusion of Transformer and Syntactic Features. arXiv preprint arXiv:2007.10534

[14] Naman Goyal, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.acl-main.747

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). DOI: https://doi.org/10.18653/v1/N19-1423

[16] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin. 2003. A Neural Probabilistic Language Model. Journal of Machine Learning Research 3, 1137-1155.