Extract and Aggregate: A Novel Domain-Independent Approach to Factual Data Verification

Anton Chernyavskiy and **Dmitry Ilvovsky**
Faculty of Computer Science
National Research University Higher School of Economics
Moscow, Russia
aschernyavskiy_1@edu.hse.ru, dilvovsky@hse.ru

Abstract

Triggered by Internet development, a large amount of information is published in online sources. However, it is a well-known fact that publications are inundated with inaccurate data. That is why fact-checking has become a significant topic in the last 5 years. It is widely accepted that factual data verification is a challenge even for the experts. This paper presents a domain-independent fact checking system. It can solve the fact verification problem entirely or at the individual stages. The proposed model combines various advanced methods of text data analysis, such as BERT and Infersent. The theoretical and empirical study of the system features is carried out. Based on FEVER and Fact Checking Challenge test-collections, experimental results demonstrate that our model can achieve the score on a par with state-of-the-art models designed by the specificity of particular datasets.

1 Introduction

With the development of online technologies, people tend to receive information mainly through the Internet. Nevertheless, Internet sources have a tendency to spread unauthentic information. In some cases, it can be done intentionally. So that to achieve, for instance, some political advantages, or to obtain a financial benefit through advertising or product promotion. In particular, the analysis conducted by Shao et al. (2017) demonstrated that, during the 2016 US presidential election on Twitter, social bots spread a lot of misinformation. Moreover, even statements about the falseness of some information in its turn can appear to be fake claims.

This paper discusses how modern approaches to the analysis of text information, such as BERT (Qiao et al., 2019), CatBoost[1] and pre-trained contextual embeddings, can assist in a fact-checking problem. We developed a model that is universal in relation to the data to be checked. Our model is based on the automatic information extraction from sources and combines best techniques from the modern approaches. Verified information can be either confirmed or refuted by each source subject to the presence of the necessary data. The collection of such results allow us to make a general conclusion about the truth or falsity of the fact.

Investigated sub-tasks are the following:

- extract qualitative information from the authoritative sources
- find the relationship between the extracted information and the verifiable claim

Due to the domain-independence of the proposed system, the problem of determining any fake information can be solved both completely or with the further study by experts. In this aspect, the task will be significantly simplified (in fact, experts just need to make the right conclusion based on the model predictions).

In our work, we combine the most successful ideas of solving each step of the fact-checking problem to build a *domain-independent pipeline* that surpasses all of the previous ones. We additionally focus on the *independence of the components* in its development (each component is not allowed to use the scores of the others). We also analyze in details the effect of natural language preprocessing (stemming, stop-words filtering, normalization, keyword highlighting, coreference resolution) and text embeddings selection. Based on this, we make some improvements at each stage[2].

[1] https://catboost.ai

[2] The source code is available online at https://github.com/aschern/FEVER

Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 69–78
Hong Kong, November 3, 2019.

The paper is organized as follows. First, we review the relevant methods and approaches used in recent fact-checking studies and shared tasks. Then, the baseline model architecture is presented. After that, the components of the developed model are described. This is followed by quantitatively comparative analysis with the state-of-the-art models on the several datasets (FEVER and Fake News Challenge). The paper ends with a summary and directions for further research.

2 Related Work

The fact-checking problem can be solved by various approaches. The majority of the most successful ones are based on information extraction from the authoritative sources. All of them were proposed in the framework of various competitions. Approaches that do not consider any additional information, achieve significantly lower results (Oshikawa et al., 2018).

FEVER competition for factual data verification with the help of information extraction from Wikipedia was held in 2018 (Thorne et al., 2018b).

A 3-stage model consisting of a sequential application of document retrieval (DR), sentence retrieval (SR) and natural language inference (NLI) components was proposed as a baseline (Thorne et al., 2018a). The first and second components select relevant articles from Wikipedia and sentences from them respectively using the part of DrQA (Chen et al., 2017) system combined with the TF-IDF metric. Then the Decomposable Attention Model (DAM) (Parikh et al., 2016) is used as the Recognizing Textual Entailment (RTE) module.

Most of the participants used the same multi-stage model structure. An additional aggregation step was often added at the last stage instead of combining all sentences into one paragraph as the entrance of the RTE module (Hanselowski et al., 2018b; Luken et al., 2018).

For relevant documents selection search API was widely used (Wikipedia Search, Google, Search, Solr, Lucene, etc.). UCL Machine Reading Group (Yoneda et al., 2018) and Athene UKP (Hanselowski et al., 2018b) teams searched for the noun phrases extracted from the statement; Columbia NLP (Chakrabarty et al., 2018) and GESIS Cologne (Otto, 2018) teams searched for the named entities.

So far, various techniques have been proposed for sentence retrieval: Word Mover's Distance and TF-IDF (Chakrabarty et al., 2018), supervised models such as logistic regression purposefully trained on the specific features (for instance, sentence numbers accounting has a big impact for the FEVER dataset – evidence is often placed at the beginning of the documents) (Yoneda et al., 2018). Thus, the model presented by Yoneda et al. (top-2 result in the competition) is not domain-independent.

Leaders of the competition UNC-NLP reformulated all of the sub-tasks in terms of neural semantic matching and solved each of them with the same architecture, based on bi-LTSM (Nie et al., 2018). Their NLI component used scores from the SR component and the SR used scores from the DR step. For this reason, this model is not task-independent.

UCL Machine Reading Group, Athene UKP, Columbia NLP used Enhanced Sequential Inference Model (ESIM) or DAM as RTE module and their variations as SR component. Sweeper team conducted joint SR and RTE components training, adapting ESIM (Hidey and Diab, 2018).

At present, other current and completed competitions related to fact-checking are also held: RumourEval[3], Fact Checking in Community Question Answering Forums[4], Fake News Challenge[5], Fast & Furious Fact Check Challenge[6].

In Fake News Challenge participants used conventional well-established machine learning models: gradient boosting, Multilayer Perceptron (MLP). These models were applied to the set of features, based on TF-IDF and word embeddings (Riedel et al., 2017; Sean Baird and Pan, 2017). Masood and Aker (2018) proposed a new state of the art model after the competition. It also utilized standard machine learning methods for manually extracted features (n-grams; similarity of embeddings, tf-idf and WordNet[7]; BoW; length of sentences, etc.).

3 Model Description

The implemented model comprises four components, like the FEVER competition baseline (it is illustrated in Figure 1).

First, document retrieval selects the set of relevant documents $\{d_1, ..., d_m\}$ for each claim c from

[3] http://alt.qcri.org/semeval2017/task8/
[4] https://competitions.codalab.org/competitions/20022
[5] http://www.fakenewschallenge.org/
[6] https://www.herox.com/factcheck/community
[7] https://mitpress.mit.edu/books/wordnet

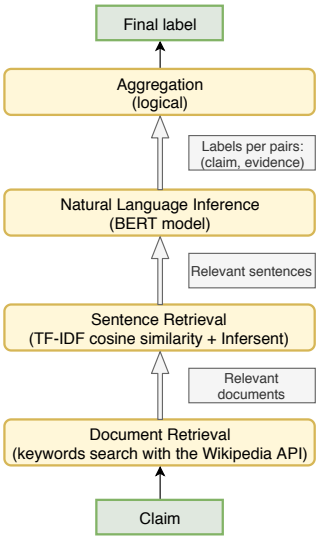

Figure 1: Four-stage model structure. Outputs of the model in each step are shown in grey boxes near the arrows.

the corpus D (if it is not initially specified). Then sentence retrieval extracts sentences $\{s_1, ..., s_n\}$ from these documents, which will help in verification. Afterwards, the NLI model f analyzes the extracted sentences in pairs with the statements and issues a verdict for each pair. Ultimately, aggregation step is implemented to obtain the final forecast: $agg(f(c, s_1), ..., f(c, s_n))$.

3.1 Document Retrieval

Search in the corpus (Wikipedia): Here we have implemented the Document Retrieval stage from (Hanselowski et al., 2018b). We applied Python Wikipedia API[8] to retrieve relevant documents from Wikipedia corpus. The following list of keywords and phrases from the claim has been taken to construct search queries: noun phrases, named entities, part of the sentence up to the "head" word. For each query, the top-k results were selected for the final list. Because sometimes there are many search queries for each claim, additionally, the filtering of results was performed. We applied Porter Stemmer to all titles of the found documents. Then we selected those documents that fully contained an initial query.

Determining document relevance: We proposed the following algorithm. Initially, the keywords (noun phrases and named entities) are highlighted from the claim. If the document contains none of them (after stemming), it is considered as "unrelated". Otherwise, an additional examination is conducted. The cosine distance between TF-IDF embeddings of the claim and each sentence in the document is calculated. If the maximum is lower than some fixed bound, the document is also considered as "unrelated".

3.2 Sentence Retrieval

We chose the combination of the TF-IDF approach and Infersent[9] for the SR stage. To find the similarity between two texts we calculated the cosine between their TF-IDF representations with the weight 0.45 and the cosine between Infersent embeddings (built on the Glove) with the weight 0.55. These weights were selected using the validation. The set of top-k sentences closest to the statement by this measure was selected.

We have also experimented with other encoding options (Glove, Word2vec), ranked by variations of BM25 (Trotman et al., 2014) and further reranked with BERT. But the final quality for these options was lower (see chapter 5 for the details).

3.3 Natural Language Inference

NLI component determines a relationship between the statement and each retrieved sentence from the previous step. Bidirectional Embedding Representations from Transformers (BERT) model was employed, as it had high results for several Glue dataset tasks (Devlin et al., 2018). Sentences from the evidence set (combined into one paragraph or stand-alone) and claim statement were involved as the "premise" and the "hypothesis" in terms of RTE. The evidence set here is the set of sentences from the SR stage.

3.4 Aggregation

In case of training BERT model on separate sentences, we applied an additional aggregation step to obtain the final prediction.

CatBoost gradient boosting model was applied as the main algorithm at this step. It was trained on the stacked predictions from the NLI step.

It is also possible to use the logical aggregation (if there is not enough training data). If all predicted labels are "NOT ENOUGH INFO", the result is the same. Otherwise, a vote between the number of "SUPPORTS" and "REFUTES" labels is taken. In the case of equality, the answer is given according to the label with the highest NLI component score. Another variant is to use the sum of

[8] https://wikipedia.readthedocs.io/en/latest/

[9] https://github.com/facebookresearch/InferSent

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los_Angeles_Riots]
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los_Angeles_County]
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

Figure 2: Example of FEVER task from (Thorne et al., 2018a). The required evidence set for the claim consists of two sentences.

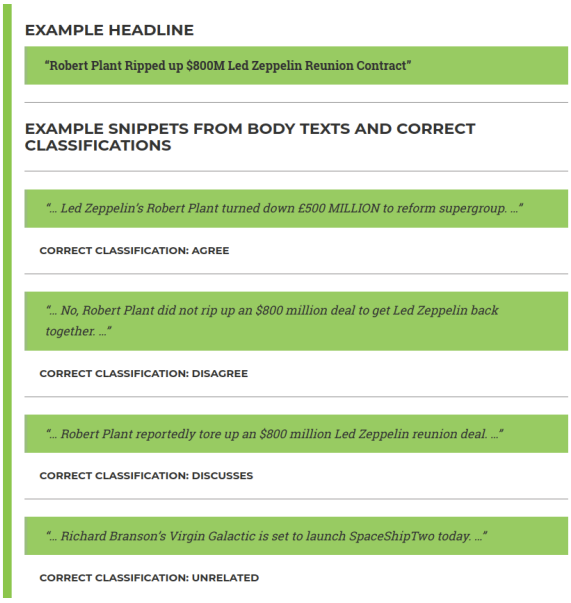

Figure 3: Example of Fake News Challenge task

class probabilities for voting.

4 Evaluation Setup

To assess the quality and verify domain-independence of our approach we tested the proposed model on several datasets and several tasks.

4.1 Datasets

Fact Extraction and VERification: The dataset from the FEVER competition (Thorne et al., 2018a) was selected as the main collection for the analysis of the presented model. Its corpus includes approximately 5.4M Wikipedia articles. All statements (about 220K) are split into 3 classes: "SUPPORTS", "REFUTES", "NOT ENOUGH INFO", depending on the presence of the corresponding evidence in the corpus. Evidence is a sentence (or set of sentences), which allows making a conclusion about the truth or falsity of the claim.

The organizers of the competition proposed the special "FEVER score" metrics. It awards points for accuracy only if the correct evidence is found. Thus, the goal is not only to identify the label correctly but also to highlight relevant evidence. Nowadays, FEVER collection is the only large collection for fact-checking with the usage of additional information.

Fake News Challenge: The Fake News Challenge competition was held in 2017 with the aim of automating the Stance Detection task. It contains 4 classes of headers paired with the articles' bodies: "agrees" (the text is in agreement with the title), "disagrees" (the text is in disagreement with the title), "discusses" (the text describes the same topic, but does not take any position related with the title), "unrelated" (the text and the title describe different topics). The dataset consists of around 75k such pairs for about 2587 texts.

In this competition, a special metrics was developed. It awards 0.25 for the correct separation of the class "unrelated" from "related" (the rest) and an additional 0.75 for the correct assignment of the first three labels. The maximum score on the test part is 11651.25.

4.2 Implementation and Training Details

Fact Extraction and VERification: The following model hyperparameters were fixed: Wikipedia API returns top-3 results for each query; the Sentence Retrieval selects top-20 sentences.

All words and phrases utilized to identify relevant documents were extracted using the Constituency Parsing, Named Entity Recognition, and Dependency Parsing implemented in the AllenNLP library. We applied Porter Stemmer from NLTK for stemming.

The first two parts (Sentence Retrieval and Document Retrieval) do not require a training step. We trained NLI component on examples of classes "SUPPORTS" and "REFUTES" from the training sample. As for statements with the "NOT ENOUGH INFO" label there is no ground truth evidence, we took the top-3 sentences from the retrieval part of the model. This number was chosen to balance "NOT ENOUGH INFO" and "SUPPORTS" classes. BERT Large was trained from

the official baseline[10] for 1 epoch on mini-batches of size 32 with the learning rate 3e-5.

We used a part of the validation sample (random 70%) to train the CatBoost aggregation method. CatBoost was trained on trees of depth 9 for 500 iterations (other parameters were taken by default).

Fake News Challenge: In this case, we applied the second variant of the Document Retrieval (determining the relevance of a particular document). Keywords for filtering were selected with the Constituency Parsing and Named Entity Recognition modules from AllenNLP. The filtering threshold for TF-IDF in the Document Retrieval component was chosen 0.05. The Sentence Retrieval highlighted top-5 sentences for each title.

The dataset was divided into training and validation samples according to the official competition repository[11].

To train BERT we used all three classes ("agrees", "disagrees", "discusses"). We chose the BERT Base version because the dataset is small. In contrast to FEVER, here the full paragraph composed of 5 separate sentences for each statement was submitted as the input because there is no ground-truth markup for the correct evidence. Thus, the aggregation stage is not required (the final result is obtained directly from BERT). The model was trained for 5 epochs on mini-batches of size 32 with the learning rate 2e-5.

5 Results and Analysis

5.1 Fact Extraction and VERification

The proposed model has many modifications: hyperparameters of TF-IDF (binarization, stop-words filtering, lower case conversion, idf usage, sublinear tf usage); application of coreference resolution (replacement of pronouns on representational entities or their addition to the beginning of the corresponding sentences); aggregation variants (boosting or logical).

5.1.1 Document Retrieval

We achieved the quality 0.908 on the validation set for the Document Retrieval component. Here the predicted set of the documents was considered as correct if it contained full evidence for the examined claim.

Sentence Retrieval	*Score*
Jaccard	0.8574
Glove	0.8548
Infersent (on Glove)	0.9025
TF-IDF, n-grams range (1, 2)	0.8930
+ lowercase	0.8934
+ max df (0.85)	0.8947
+ sublinear tf	0.8976
+ traditional stop-words filtration	0.8889
TF-IDF, n-grams range (1, 1)	0.8926
+ lowercase	0.8930
+ max df, sublinear tf	0.8997
+ binary	0.9024
+ weighted Infersent	**0.9081**

Table 1: Results of Sentence Retrieval on the validation set for the selection of the top-5 sentences. For TF-IDF cumulative results for applied techniques are provided. tf/df - term/document frequency, sublinear tf = 1 + log(tf), max df - all words with df higher, than threshold are considered as stop-words.

5.1.2 Sentence Retrieval

The results of the Sentence Retrieval for finding top-5 sentences are presented in Table 1. The most successful variant was the TF-IDF search by unigrams with the filtering of stop-words selected in each document, binarization and lower case conversion in the combination with Infersent embeddings. Again, the predicted set was considered as correct if it contained entire evidence set.

We considered all words whose proportion in a particular document is higher than 0.85 as stop-words. The importance of using such stop-words follows from the fact that in case of determining the most significant sentences *inside* the document, they do not play an important role. The term frequency binarization has a significant impact because only the availability of information is important but not the number of references.

We also experimented with FastText embeddings, but Glove achieved higher results in all cases (see Figure 4).

In addition, we tried different BM25 modifications: BM25L, BM25+, BM25Okapi. The optimal combination was stop-words filtering, lower case conversion and Krovetz stemming. The results for the selection of the top-20 sentences are presented in Table 2.

[10] https://github.com/google-research/bert

[11] https://github.com/FakeNewsChallenge/fnc-1-baseline

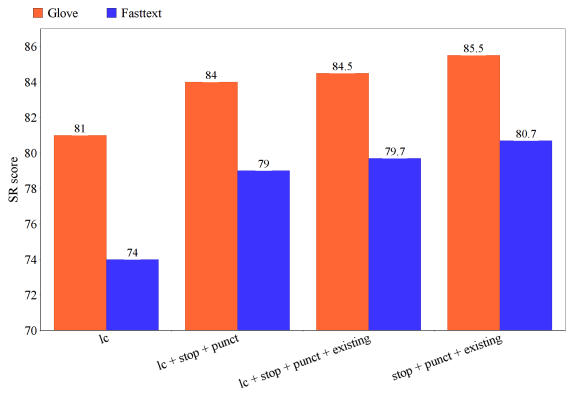

Figure 4: SR-score for Glove and FastText embeddings. lc – lowercase; stop + punct – filtering stop-words and punctuation for embeddings calculation; existing – averaging only by words from the dictionary (otherwise zero vectors were considered for OOV).

algorithm	*lc + stop*	*lc + stop + Krovetz*
BM25Okapi	0.93389	0.93414
BM25+	0.93314	0.93419
BM25L	0.94124	0.94224

Table 2: The results of SR component on the validation sample for the top 20 sentences selection. lc – lowercase, stop – stop-words filtering, Krovertz – Krovetz stemming.

As it was mentioned above, we utilized the top-20 extracted sentences for each claim in our solution (the results are fully correlated with Table 1 for top-5 sentences). We chose this value for two reasons: a relatively high quality (∼94.7) should be achieved, and the number itself should not be very large to simplify further analysis and aggregation. Thus, the quality changed faintly starting with the top-20 and reached ∼95.1 for the top-50.

Coreference resolution gave us 0.9041 for the top-5 sentences extraction. We used Stanford NLP Coreference parser (we also experimented with the Co-reference Resolution module from AllenNLP). Here, we appended representative mentions of pronouns to the beginning of the sentences. But it did not improve the quality. This can be explained by the fact that the fixed document refers to exactly one entity (mentioned in its title) very often. Therefore, additional mentioning does not make sense for the relevancy evaluation.

5.1.3 Natural Language Inference

The quality of the BERT model according to the accuracy metrics was 0.834 (classification of the individual sentences into 3 classes) on a balanced subset of the validation sample. In this case, to solve the coreference problem, we added the titles of the documents to the beginning of the sentences through the separator. In contrast with the relevance assessment, it is important to have a comprehension of what entity is considered.

5.1.4 Aggregation

For 30% of the validation sample, we achieved accuracy 74.81 for the CatBoost aggregation and 73.47 for the logical aggregation. In the case of CatBoost, the model was trained on 70% of the validation set and was tested on the remaining 30

Confusion matrices for logical aggregation on the full validation sample and CatBoost aggregation on its test part are presented in tables 3 and 4 respectively. In the first case, the “NOT ENOUGH INFO” label is the greatest difficulty for the model. In the second case, the classes have approximately equal complexity, but the main fraction of errors also occurs due to the separation of “NOT ENOUGH INFO” from the rest.

For the second variant of logical aggregation (voting by the sum of the class probabilities predicted by BERT model), the maximum accuracy was 72.98. It is lower than 73.47 for the first case.

Our model achieves the accuracy 71.72 for labels and F1-score 70.20 for retrieved evidence on the test set (the results are presented in Table 5).

We tried two prediction options for the evidence. In the first case, only those sentences whose labels match with the final prediction were added to the answer. In the second case, we complemented this set to 5 sentences, according to the ranking of the Sentence Retrieval. This raises the FEVER score (a key metrics for the competition) on the test set from 66.69 to 67.68. However, precision falls significantly (from 71.66 to 41.36), and, respectively, the F1 score for the evidence decreases too.

Additionally, we trained BERT for binary classification into classes “NOT ENOUGH INFO”/“ENOUGH INFO” and re-ranked sentences by the probability of “ENOUGH INFO” label. Thus, the order of relevant sentences from the Sequence Retrieval component was replaced by the order according to this BERT model. However, it did not give positive results – the FEVER score on the test sample even slightly decreased (down to 67.62).

	predicted labels		
	SUPPORTS	REFUTES	NOT ENOUGH INFO
SUPPORTS	**5734**	229	703
REFUTES	599	**4856**	1211
NOT ENOUGH INFO	1465	1238	**3963**

Table 3: Confusion matrix for logical aggregation

	predicted labels		
	SUPPORTS	REFUTES	NOT ENOUGH INFO
SUPPORTS	**1595**	65	347
REFUTES	110	**1395**	434
NOT ENOUGH INFO	250	346	**1458**

Table 4: Confusion matrix for CatBoost aggregation

Team name	*Evidence F1 (%)*	*Label Accuracy (%)*	*FEVER Score*
DREAM (MSRA+MSNews)*	39.33	**76.42**	**69.76**
a.soleimani.b*	38.61	71.86	69.66
abcd_zh*	39.14	72.81	69.40
cunlp*	37.65	72.47	68.80
dominiks*	36.26	71.54	68.46
own*	36.80	72.03	67.56
GEAR*	36.87	71.60	67.10
UNC-NLP	52.96	68.21	64.23
UCL Machine Reading Group	34.97	67.62	62.52
Athene UKP TU Darmstadt	36.97	65.46	61.58
Papelo	64.85	61.08	57.36
Our model	70.20	71.72	66.69
Our model (all 5)	53.21	71.72	***67.68***

Table 5: Results on the FEVER test dataset (top teams)
* - after competition (up to 19.08.2019)

5.1.5 Error Analysis

Document Retrieval: Errors in the DR component are often caused by the misspelling of entities in statements: "Homer *Hickman* wrote some historical fiction novels." vs. "Homer *Hickam*" or "2015 was the year of the Disaster *Aritst* film (film) started." vs. "The Disaster *Artist*".

Another popular mistake is the lack of keywords from the title in the claim. For example, the evidence set for the statement "Christian Gottlob Neefe was an *opera writer*" includes the document "Composer".

The third type of error is dividing one entity into several. For instance, in the claim "The Food Network is a channel that ran *Giada at Home*." our model highlights two entities: "Giada" and "Home" and selects documents with that titles.

Sentence Retrieval: The SR component works mostly correctly since 20 sentences are selected for each claim. Errors often occur in the case of composite evidence where one sentence clarifies some information from another.

Natural Language Inference: The main source of errors is cases with very similar concepts. For example, claim "Wildfang is a US-based women's apparel company featuring *pants* that are tomboyish in style" has "NOT ENOUGH INFO" label. But the model selects evidence "Wildfang is a US-based women 's apparel company featuring *clothing* that is tomboyish in style" and classifies this claim as "SUPPORTS".

There are also opposite cases where words with

different meanings don't have a key impact. So, for the claim "Michigan is a *stop destination* for recreational boating within the U.S." the correct label is "SUPPORTS" with the evidence "Michigan # As a result, it is one of the leading *U.S. states* for recreational boating.". Our model predicts "REFUTES" due to the words "stop destination" vs. "state". Another interesting example: the statement "Seohyun was *only born on July* 28, 1991." has ground truth label "SUPPORTS" with the corresponding evidence "Seo Ju-hyun -LRB- *born June* 28 , 1991 -RRB- ...". Our model predicts the label "REFUTES" focusing on the words "June" and "July" and not the word "only".

Also, the BERT model makes predictions for separate sentences. For the claim "Papua comprised all of a country" the correct label is "SUPPORTS" with the evidence comprising "Papua is the largest and easternmost province of Indonesia, comprising most of western New Guinea" and the document "Western_New_Guinea". But this evidence separately is not enough to make the right conclusion.

5.2 Fake News Challenge

The TF-IDF approach calculated by unigrams and bigrams with filtering of standard stop-words was optimal for the relevance determination (after evaluation by keywords). These parameters differ from the TF-IDF parameters in the Sentence Retrieval. In this case, we filtered standard stop-words (we utilized the list from NLTK), as they do not affect the global assessment of the complete document.

BERT achieved 0.822 accuracy for the classification into one of three classes. We also tried to apply coreference resolution. However, as for the FEVER dataset, no improvement was received. We achieved accuracy 0.815 as maximum among all the cases under consideration (unrepresentative mentions replacement, addition to beginning of the sentences, using of pronouns only). This can be explained by the fact that all 5 sentences are submitted to the NLI component as a single text. And this text already contains representative references to the pronouns with a high probability.

We also estimated the contribution of the features of retrieval components. It was detected that the filtering of the documents by keywords for the binary definition of the type "related"/"unrelated" improves the quality of the final model from 9430 to 9592 points. The reason is that the method has high precision 0.9776 for the class "unrelated". This approach has a relatively small recall 0.5581, but combining with TF-IDF rises it to 0.9668 (it is higher than 0.95 for the separate TF-IDF usage). This observation demonstrates that a preliminary analysis of the presence of the keywords is important for document relevance determination. Discarding the traditional stop-words increases the total score from 9592.0 to 9808.0 (or 0.8417 of max) with the total accuracy at 0.883. The results are presented in Table 6.

Team name	*FNC score*
Zhang et al. (2018)	**10097.00**
Masood and Aker (2018)	9565.70
SOLAT in the SWEN	9556.50
Athene	9550.75
UCLMR	9521.50
Chips Ahoy!	9345.50
CLUlings	9289.50
Our model	***9808.00***

Table 6: Results on the FNC test dataset. FNC-score - relative competition score

	predicted labels			
	unrel.	discuss	agree	disagree
unrelated	**6416**	368	69	45
discuss	123	**1499**	130	48
agree	55	172	**504**	31
disagree	12	50	20	**80**

Table 7: Confusion matrix on the FNC validation set

Confusion matrix (Table 7) on the validation part shows that the class "disagree" is the hardest one for the model. The reason is that its proportion in the training sample is only 2.8%. Nevertheless, the macro-averaged class-wise F1 score is high - 0.709. It is a very important metric in this case (Hanselowski et al., 2018a) and models of competition participants achieve only $\sim$0.60.

It should be noted that fewer examples started to belong to the class "unrelated" when we reduced the hyperparameter in the TF-IDF filtering (that is, the model separated only the most explicit articles). It increased the probability of the correct classification into the remaining 3 classes (which has a great significance in this competition). The highest score of 3799.75 (or 0.8541) on the validation set was obtained with the filtration hyper-

parameter value 0.05. It is also worth to notice that BERT model demonstrated here significantly higher performance (in terms of accuracy) than on the test part: 0.822 vs. 0.783.

6 Conclusion

The paper presents a domain-independent model for checking factual information using automatic information extraction. The presented model was inspired by the FEVER baseline but has significant improvement at all 4 stages (document retrieval, sentence retrieval, natural language inference, aggregation). Experimental and theoretical analysis of all new features was carried out.

The proposed model exploits no data-specific features. Moreover, it can solve all of the subtasks (perform at all 4 steps) independently because none of the components use the scores of the others. We experimentally demonstrated that the model can perform at the same level as the current state-of-the-art models on the two most popular tasks and datasets.

While the model already demonstrates good results, an important further improvement is its integration with the methods that take into account additional linguistic features (for instance, discourse information for an evidence set creation).

7 Acknowledges

The article was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'.

References

Tuhin Chakrabarty, Tariq Alhindi, and Smaranda Muresan. 2018. Robust document retrieval and individual evidence modeling for fact extraction and verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 127–131, Brussels, Belgium. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *CoRR*, abs/1704.00051.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018a. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018b. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Christopher Hidey and Mona Diab. 2018. Team SWEEPer: Joint sentence extraction and fact checking with pointer networks. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 150–155, Brussels, Belgium. Association for Computational Linguistics.

Jackson Luken, Nanjiang Jiang, and Marie-Catherine de Marneffe. 2018. QED: A fact verification system for the FEVER shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 156–160, Brussels, Belgium. Association for Computational Linguistics.

Razan Masood and Ahmet Aker. 2018. The fake news challenge: Stance detection using traditional machine learning approaches.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2018. Combining fact extraction and verification with neural semantic matching networks. *CoRR*, abs/1811.07039.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *CoRR*, abs/1811.00770.

Wolfgang Otto. 2018. Team GESIS cologne: An all in all sentence-based approach for FEVER. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 145–149, Brussels, Belgium. Association for Computational Linguistics.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *EMNLP*.

Yifan Qiao, Chenyan Xiong, Zheng-Hao Liu, and Zhiyuan Liu. 2019. Understanding the behaviors of BERT in ranking. *CoRR*, abs/1904.07531.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *CoRR*, abs/1707.03264.

Doug Sibley Sean Baird and Yuxi Pan. 2017. Talos targets disinformation with fake news challenge victory.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. 2017. The spread of fake news by social bots. *CoRR*, abs/1707.07592.

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

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The Fact Extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, pages 58:58–58:65, New York, NY, USA. ACM.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Ucl machine reading group: Four factor framework for fact finding (hexaf). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102. Association for Computational Linguistics.

Qiang Zhang, Emine Yilmaz, and Shangsong Liang. 2018. Ranking-based method for news stance detection. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 41–42, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.