arXiv:2009.06402v2 [cs.CL] 10 Feb 2021

# Time-Aware Evidence Ranking for Fact-Checking

Liesbeth Allein[a,b,*], Isabelle Augenstein[c], Marie-Francine Moens[b]

[a] *Text and Data Mining Unit, European Commission Joint Research Centre, Via E. Fermi 2749, 21027 Ispra (VA), Italy*
[b] *Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium*
[c] *Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100 Copenhagen, Denmark*

**Abstract**

Truth can vary over time. Fact-checking decisions on claim veracity should therefore take into account temporal information of both the claim and supporting or refuting evidence. In this work, we investigate the hypothesis that the timestamp of a Web page is crucial to how it should be ranked for a given claim. We delineate four temporal ranking methods that constrain evidence ranking differently and simulate hypothesis-specific evidence rankings given the evidence timestamps as gold standard. Evidence ranking in three fact-checking models is ultimately optimized using a learning-to-rank loss function. Our study reveals that time-aware evidence ranking not only surpasses relevance assumptions based purely on semantic similarity or position in a search results list, but also improves veracity predictions of time-sensitive claims in particular.

*Keywords:* automated fact-checking, temporal relevance, temporal semantics, document ranking, learning to rank

## 1. Introduction

While some claims are incontestably true or false at any time (e.g. *"Smoking increases the risk of cancer"*), the veracity of others are subject to time indications and temporal dynamics (e.g. *"Face masks are obligatory on public transport"*) [1]. Not only their veracity, but also their semantics are time-sensitive or time-dependent as connotations and real-world references can change over time. Evidence supporting or refuting such time-sensitive claims are likewise time-dependent. The relevance of these documents, which reflects both their semantic relatedness to the claim and suitability for accurate claim veracity prediction, is thus relative to a claim's publishing time and/or the documents' publishing time. A fact-checking model's inability to correctly frame both claim and evidence in time, and its inability to rank evidence pages by relevance, can result in inaccurate semantic representations, truth predictions and relevance estimations. Nonetheless, automated fact-checking research has paid little attention to the temporal dynamics of truth, semantics and relevance. In this work, we focus on the temporal relevance of Web documents, which serve as evidence to a given claim. We introduce four temporal ranking methods, which constrain evidence ranking relying on diverse hypotheses for evidence relevance, and explore how time-aware evidence ranking impacts the veracity prediction performance of three fact-checking models (Figure 1).

*Corresponding author
*Email address:* `Liesbeth.ALLEIN@ec.europa.eu` (Liesbeth Allein)

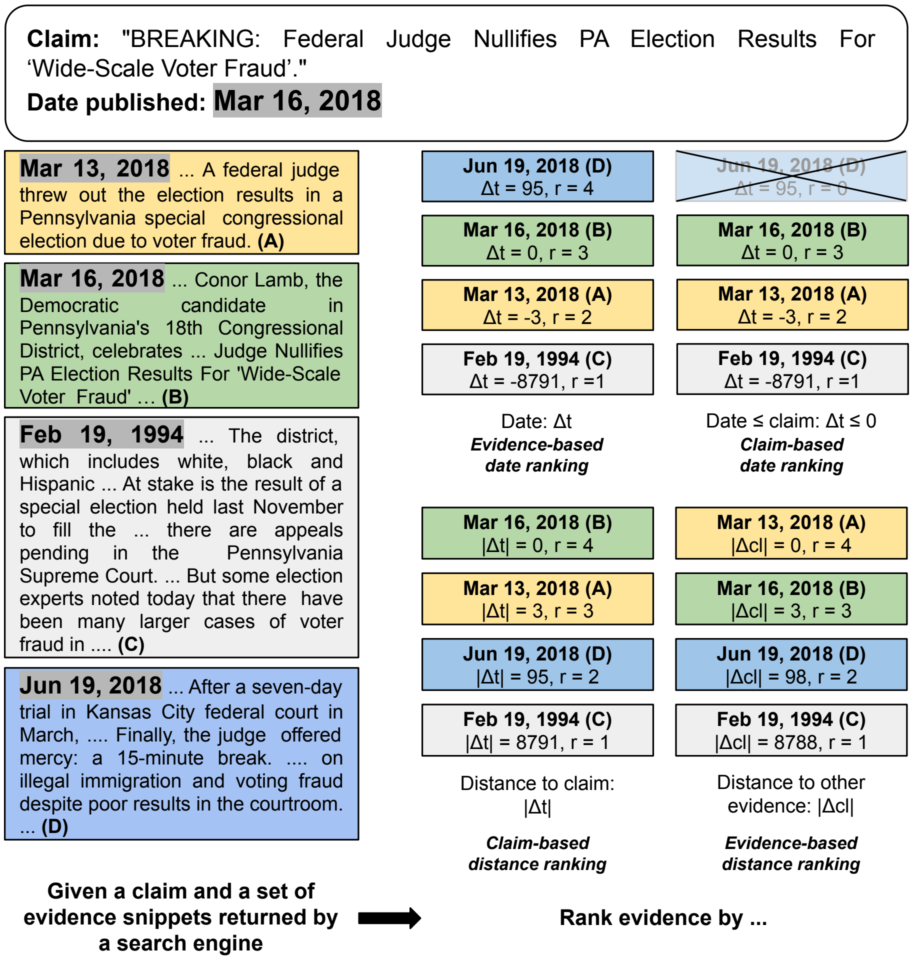



Figure 1: Overview of our temporal ranking methods that constrain evidence ranking following different hypotheses for temporal relevance: *evidence-based date ranking*, *claim-based date ranking*, *claim-based distance ranking*, *evidence-based distance ranking*. Although the methods actually rank up to ten snippets per claim in the experiments, we display four snippets for simplicity. The evidence timestamps are first extracted and normalized to their distance to the claim in days ($\Delta t$). In the fourth ranking method, *evidence-based distance ranking*, the mediod and the snippets' distance to that mediod is computed ($\Delta cl$). Based on the $\Delta t$ or $\Delta cl$ values, the temporal ranking methods assign a ranking score ($r$) to each evidence snippet - with a higher score denoting a higher degree of relevance.

The first two methods simply rank evidence by date in descending order: *evidence-based date ranking* sorts all given evidence, while *claim-based date ranking* only ranks evidence published before and at claim time. The other two methods rank evidence by distance in days to either the claim (*claim-centered distance ranking*) or the other evidence in the same set (*evidence-centered distance ranking*), both in ascending order. These methods then simulate method-specific ground-truth rankings given the timestamp of each evidence snippet. Ultimately, the evidence ranking module of a fact-checking model is directly optimized using a dedicated learning-to-rank loss function which measures the agreement between the model's ranking output and the simulated ground-truth rankings.

In summary, the **contributions** of this work are as follows.

- We propose to model the temporal dynamics of evidence for content-based fact-checking and show that it outperforms a ranking of Web documents based purely on semantic similarity – as used in prior work – and search engine ranking.
- We test various hypotheses for evidence relevance using timestamps and explore the performance differences between those hypotheses.
- We train evidence ranking by optimizing a learning-to-rank loss function; this elegant, yet effective approach requires only a few adjustments to the model architecture and can be easily added to any fact-checking model.
- Optimizing evidence ranking using a dedicated, learning-to-rank loss function is, to our knowledge, novel in automated fact-checking research.

## 2. Related Work

Previous work on content-based fact-checking exploiting both claim and evidence has differentiated between pieces of evidence in various manners. Some consider different evidence documents to be equally important [2, 3]. Others weigh or rank evidence according to its assumed relevance to the claim. Liu et al. [4], for instance, link evidence relevance to node and edge kernel importance in an evidence graph using neural matching kernels. Li et al. [5, 6] define evidence relevance in terms of evidence position in a search engine's ranking list, while Wang et al. [7] relate it with source popularity. However, evidence relevance has been principally associated with semantic similarity between claim-evidence pairs and is computed using language models [8], textual entailment/inference models [9, 10, 11], cosine similarity [12], or token matching and sentence position in the evidence document [13]. In contrast to previous work, we hypothesize that the timestamp of a piece of evidence and reasoning with the temporal information is crucial to how its relevance should be defined and how it should be ranked for a given claim.

The dynamics of time in fact-checking have not been widely studied yet. Yamamoto et al. [14] incorporate the idea of temporal factuality in a fact-checking model. Uncertain facts are input as queries to a search engine, and a fact's trustworthiness is determined based on the detected sentiment and frequency of alternative and counter facts in the search results in a given time frame. However, the authors point out that the frequency of a fact can be misleading, with incorrect claims possibly having more hits than correct ones

[6]. Hidey et al. [15] recently published an adversarial dataset that can be used to evaluate a fact-checking model's temporal reasoning abilities. In this dataset, arithmetic, range and verbalized time indications are altered using date manipulation heuristics. Zhou et al. [16] study pattern-based temporal fact extraction. By first extracting temporal facts from a corpus of unstructured texts using textual pattern-based methods, they model pattern reliability based on time cues such as text generation timestamps and in-text temporal tags. Unreliable and incorrect temporal facts are then automatically discarded. However, relying on the above method a large amount of data is needed to determine a claim's veracity, which might not be available for new claims yet.

## 3. Time-Aware Evidence Ranking

We encourage a content-based fact-checking model to reason about the temporal semantics and time dependency of both claim and evidence by constraining the model's ranking module, which assigns a relevance/ranking score to each Web document serving as evidence. During training, the ranking module is optimized using a learning-to-rank loss function, which measures the agreement between the learned ranking output and the expected ranking. As ground-truth evidence rankings are lacking, we introduce four ranking methods relying on several hypotheses on temporal relevance and use these methods to simulate ground-truth rankings.

We first discuss the three fact-checking models whose ranking module we will optimize on the simulated ground-truth evidence rankings. Next, we elaborate on the specific learning-to-ranking loss used during training. We then explain how timestamps for a given claim and evidence set are extracted and normalized, and introduce the four temporal ranking methods which: (a) constrain evidence ranking following several hypotheses for evidence relevance; and (b) simulate hypothesis-specific ground-truth evidence rankings. From Section 3.3 onwards, we illustrate the time extraction/normalization process and the temporal ranking methods with the sample evidence set and claim from Figure 1.

### *3.1. Fact-Checking Model Architecture*

In this section, we describe the model architecture of three fact-checking models. In a training setup without any ranking constraints, all model parameters are optimized using a loss function on the verification classification task. Evidence ranking is then learned implicitly. In a training setup with our ranking constraints, a model's ranking parameters are directly optimized using a learning-to-rank loss on the yielded evidence rankings while the other model parameters are optimized using a loss on the predicted veracity labels. By applying our temporal ranking methods to various neural architectures, we show their advantage over time-unaware approaches in a transparent manner. We take the Joint Veracity Prediction and Evidence Ranking model presented in the MultiFC dataset paper [17] as base model architecture, and experiment with different neural architectures for the other two models. For the sake of simplicity, we name the models by their sentence encoder architecture. An overview of the model architectures is given in Figure 2.

#### 3.1.1. BiLSTM

This is the Joint Veracity Prediction and Evidence Ranking model as presented in the dataset paper [17]. The BiLSTM model takes as input claim sentence $c_i$, evidence set $E_i = \{e_{i_1}, ..., e_{i_K}\}$ and claim metadata[1] $m_i$ - with $K \leq 10$ the total number of evidence snippets in evidence set $E_i$. The sentence encoder - a bidirectional LSTM with skip-connections - transforms $c_i$ and $e_{i_j}$ into their respective hidden representation $h_{c_i}$ and $h_{e_{i_j}}$, while the metadata encoder - a CNN - transforms $m_i$ into $h_{m_i}$. Each $h_{e_{i_j}}$ is then fused with $h_{c_i}$ and $h_{m_i}$ into a joint representation $f_{i_j}$ following a natural language inference matching method proposed by Mou et al. [18]:

$$f_{i_j} = [h_{c_i}; h_{e_{i_j}}; h_{c_i} - h_{e_{i_j}}; h_{c_i} \circ h_{e_{i_j}}; h_{m_i}] \quad (1)$$

where semi-colon denotes vector concatenation, "$-$" element-wise difference and "$\circ$" element-wise multiplication. All $f_{i_j}$ are sent through two modules: a label scoring and a ranking module. In the label scoring module, similarity between $f_{i_j}$ and all labels across all domains are scored by taking the dot product between $f_{i_j}$ and all label embeddings, which are updated during model training. In this way, relationships between labels across all fact-check domains are learned. This results in label similarity matrix $S_{i_j}$. As the model needs to predict a domain-specific label, a domain-specific mask is applied over $S_{i_j}$, masking all out-of-domain label similarity scores. Ultimately, a fully-connected layer computes domain-specific label score vector $l_{i_j}$. In the ranking module, a two-layer fully-connected layer computes a ranking score $r(f_{i_j})$ for each $f_{i_j}$. The ranking score reflects the relevance of $e_{i_j}$: the higher the score, the higher the evidence snippet's relevance to the claim. The dot product between each label score vector $l_{i_j}$ and its respective ranking score $r(f_{i_j})$ is taken, the resulting vectors are summed and domain-specific label probabilities $p_i$ are obtained by applying a softmax function. Finally, the model outputs the domain-specific veracity label with the highest probability.

#### 3.1.2. RNN

This model's architecture is similar to that of the BiLSTM model, but differs in terms of sentence encoder architecture, fusion mechanism and number of fully-connected layers in the ranking module. Instead of a bidirectional LSTM with skip-connections, a two-layer unidirectional RNN encodes $c_i$ and $e_{i_j} \in E_i$ into their hidden representations $h_{c_i}$ and $h_{e_{i_j}}$. These representations are then fused with the hidden metadata representation $h_{m_i}$ using a simple concatenation operation instead of the natural language inference matching method: $f_{i_j} = [h_{c_i}; h_{e_{i_j}}; h_{m_i}]$. Lastly, we halve the number of fully-connected layers in the ranking module so that the ranking module is a shallower network.

#### 3.1.3. DistilBERT

In this model, a DistilBERT Transformer model with a sequence classification head on top[2] jointly takes as input claim sentence $c_i$ and evidence snippet $e_{i_j}$, and returns the '[CLS]' embedding from its final contextual layer and a probability distribution over all labels $p_{i_j}$. The '[CLS]' embedding represents the joint claim-evidence representation $f_{i_j}$. All joint representations $f_{i_j}$ are then sent through a two-layer fully-connected layer that computes a ranking score $r(f_{i_j})$ for each $f_{i_j}$.

[1] Metadata contains information on speaker, tags and categories.

[2] We adopt the pre-trained DistilBertForSequenceClassification model from the Hugging Face Transformers library [19].

As DistilBERT already yields a probability distribution over all labels $p_{i_j}$, we simply apply a domain-specific mask over $p_{i_j}$ to obtain label score vector $l_{i_j}$. The dot product between each label score vector $l_{i_j}$ and its respective ranking score $r(f_{i_j})$ is taken, the resulting vectors are summed and domain-specific label probabilities $p_i$ are obtained by applying a softmax function. Finally, the model outputs the domain-specific veracity label with the highest probability.

### *3.2. Learning-to-Rank Loss*

In order to optimize evidence ranking, we need a loss function that measures how correctly an evidence snippet is ranked with regard to the other snippets in the same evidence set. For this, the **ListMLE** loss [20] is computed:

$$ListMLE(r(E), R) = -\sum_{i=1}^{N} logP(R_i|E_i; r) \quad (2)$$

$$P(R_i|E_i; r) = \prod_{u=1}^{K} \frac{exp(r(E_{R_{i_u}}))}{\sum_{v=u}^{K} exp(r(E_{R_{i_v}}))} \quad (3)$$

ListMLE is a listwise, non-measure-specific learning-to-rank algorithm that uses the negative log-likelihood of a ground-truth permutation as loss function [21]. It is based on the Plackett-Luce model, that is, the probability of a permutation is first decomposed into the product of a stepwise conditional probability, with the $u$-th conditional probability standing for the probability that the snippet is ranked at the $u$-th position given that the top $u$ - 1 snippets are ranked correctly. ListMLE is used for optimizing the evidence ranking with each of the four temporal ranking methods.

In case an evidence snippet is excluded from the ground-truth evidence ranking $R_i$ or lacks a timestamp, we apply a mask over the predicted evidence ranking vector and compute the ListMLE loss over the ranking scores of the included evidence snippets with timestamps. We assume that the direct optimization of these evidence snippets' ranking scores will indirectly influence the ranking score of the others, as they may contain similar explicit time references further in their text or exhibit similar patterns.

### *3.3. Temporal Relevance and Ranking Methods*

We explain how temporal information for both claim and evidence is extracted and normalized, and introduce the four temporal ranking methods that rank the Web documents in the evidence set. We use the claim and evidence set from Figure 1 for illustration purposes:

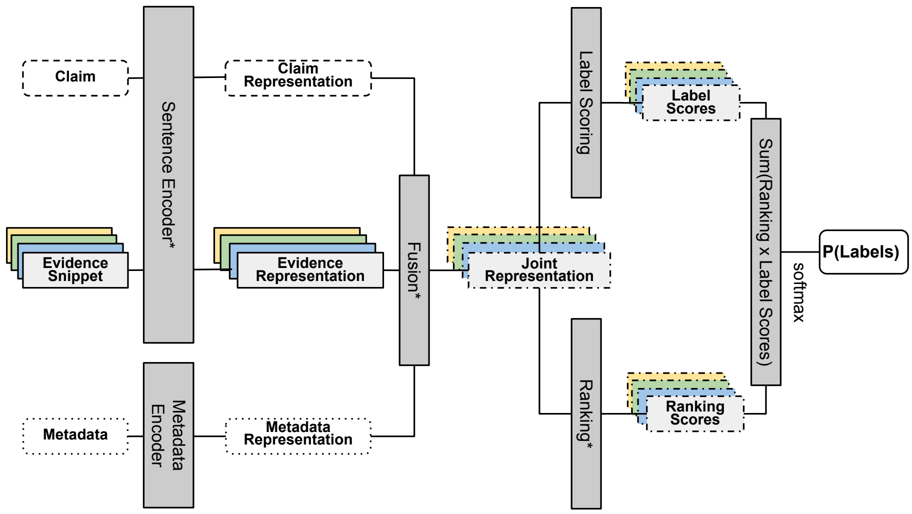
$E_s$ = {

$e_{s_1}$ *"Mar 13, 2018 ... A federal judge threw out the election results in a Pennsylvania special congressional election due to voter fraud."*

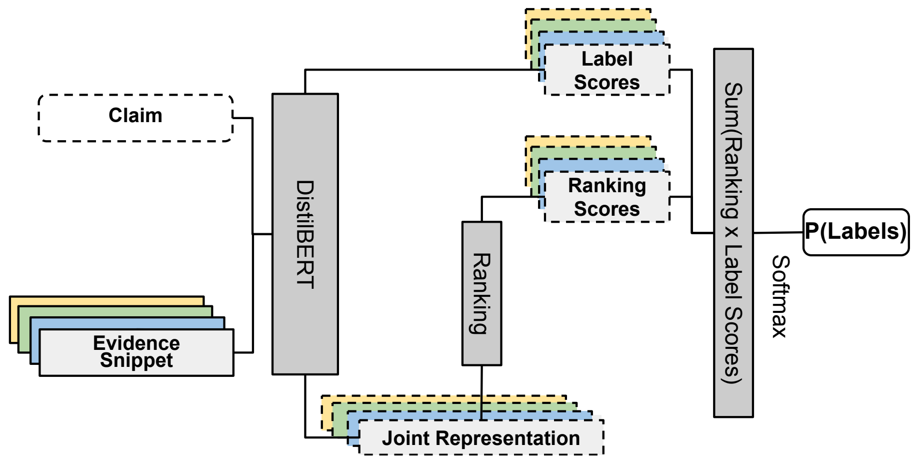
$e_{s_2}$ *"Mar 16, 2018... Conor Lamb, the Democratic candidate in Pennsylvania's 18th Congressional District, celebrates ... Judge Nullifies PA Election Results For 'Wide-Scale Voter Fraud' ..."*

$e_{s_3}$ *"Feb 19, 1994... The district, which includes white, black and Hispanic ... At stake is the result of a special election held last November to fill the ... there are appeals pending in the Pennsylvania Supreme Court. ... But some election experts noted today that there have been many larger cases of voter fraud in ...."*

$e_{s_4}$ *"Jun 19, 2018... After a seven-day trial in Kansas City federal court in March, .... Finally, the judge offered mercy: a 15-minute break. .... on illegal immigration and voting fraud despite poor results in the courtroom. ..."*




(a) BiLSTM/RNN




(b) DistilBERT

Figure 2: Overview of the fact-checking models. The BiLSTM and RNN model share a similar model structure (a), but differ in sentence encoder architecture (BiLSTM vs RNN), fusion mechanism (inference matching vs. concatenation) and number of fully-connected layers in the ranking module (two vs. one).

$\}$ and $c_s$ = *"BREAKING: Federal Judge Nullifies PA Election Results For 'Wide-Scale Voter Fraud'."* with timestamp *"Mar 16, 2018"* given as claim metadata.

### *3.4. Timestamp Extraction and Normalization*

The dataset contains $N$ claim sentences $c_i \in C$ and $N$ evidence sets $E_i \in E$, each providing $K \leq 10$ evidence snippets $e_{i_j} \in E_i$. A timestamp for each claim $c_i$ is included as metadata. For the evidence snippets $e_{i_j}$, however, we need to extract the timestamp from the evidence text itself. A publishing timestamp is frequently given at the beginning of the text, which facilitates timestamp extraction. We then normalize all extracted timestamps to *year-month-day*, resulting in:

$$t : C, E \rightarrow C^t = \{t(c_1), ..., t(c_N)\}, E^t = \{E_i^t, ..., E_N^t\} \tag{4}$$

with $E_i^t = \{t(e_{i_1}), ..., t(e_{i_K})\}$. For each evidence snippet $e_{i_j} \in E_i$, the temporal distance in days between claim time $t(c_i)$ and evidence time $t(e_{i_j})$ is calculated by subtracting $t(c_i)$ from $t(e_{i_j})$. This results in:

$$\Delta t : E^t \rightarrow \Delta E^t = \{\Delta E_i^t, ..., \Delta E_N^t\} \tag{5}$$

with $\Delta E_i^t = \{\Delta t(e_{i_1}), ..., \Delta t(e_{i_K})\}$. Positive and negative $\Delta t(e_{i_j})$ denote evidence snippets that are published later and earlier than $t(c_i)$, respectively. These integer time values are subsequently used for simulating the ground-truth evidence rankings.

**Example:** For $E_s = \{e_{s_1}, e_{s_2}, e_{s_3}, e_{s_4}\}$ and $c_s$:

- Extract timestamps from evidence texts:
  $E_s = \{$*"Mar 13, 2018"*, *"Mar 16, 2018"*, *"Feb 19, 1994"*, *"Jun 19, 2018"*$\}$
- Normalize all timestamps to *year-month-day* (1):
  $t : c_s, E_s \rightarrow t(c_s) =$ *2018–3–16*,
  $E_s^t = \{$*2016–3–13*, *2018–3–16*, *1994–2–19*, *2018–6–19*$\}$
- Calculate the temporal distance in days to $t(c_s)$ for each $t(e_{s_j}) \in E_s^t$ (2):
  $\Delta t : E_s^t \rightarrow \Delta E_s^t = \{-3, 0, -8791, 95\}$

### *3.5. Temporal Ranking Methods*

All four temporal ranking methods assign ranking scores to the evidence snippets following a method-specific ranking constraint.

$$r : E \rightarrow R = \{R_i, ..., R_N\} \tag{6}$$

with $R_i = \{r(e_{i_1}), ..., r(e_{i_K})\}$. In all methods, an evidence snippet's ranking score depends on its relative position in time compared with the other evidence snippets in the same evidence set. The ranking scores are proportional to relevance, i.e., a higher ranking score denotes higher relevance. The ranking methods fall into two categories: date-based and distance-based ranking [22, 23]. The date-based ranking methods follow the general hypothesis that more information on a subject becomes gradually available over time, and thus newer evidence should be ranked higher than older evidence [24]. Both *evidence-based date ranking* and *claim-based date ranking* rank evidence by date in descending order, but the latter excludes evidence that is published after the claim date. While these date-based ranking methods consider relevance as a property proportional to recency, the distance-based ranking methods define relevance in terms of temporal distance to the claim (*claim-centered distance ranking*) or the other evidence in the evidence set (*evidence-centered distance ranking*). In these ranking methods, evidence is ranked by distance in ascending order. Table 1 provides an overview of all methods illustrated with simulated ground-truth rankings for sample evidence set $E_s$.

*3.5.1. Evidence-based date ranking*

**Hypothesis 1.** *Evidence published later in time is more relevant than evidence published earlier in time, and should thus be ranked accordingly.*

The evidence-based date ranking method ranks the evidence snippets in a given set by their publishing date in descending order. The simulated ground-truth ranking $R_i = \{r(e_{i_1}), ..., r(e_{i_K})\}$ satisfies the following constraint:

$$\begin{aligned} &\forall e_{i_j}, e_{i_k} \in E_i : \Delta t(e_{i_j}) \leq \Delta t(e_{i_k}) \\ &\implies r(e_{i_j}) \leq r(e_{i_k}) \end{aligned} \tag{7}$$

For two evidence snippets in the same evidence set, the constraint imposes a higher ranking score $r(e_{i_k})$ for $e_{i_k}$ and a lower ranking score $r(e_{i_j})$ for $e_{i_j}$ if $\Delta t(e_{i_k})$ is larger than $\Delta t(e_{i_j})$. If $\Delta t(e_{i_k})$ and $\Delta t(e_{i_j})$ are equal, $e_{i_k}$ and $e_{i_j}$ obtain the same ranking score.

**Example**: For $E_s = \{e_{s_1}, e_{s_2}, e_{s_3}, e_{s_4}\}$ with $\Delta E_s^t = \{-3, 0, -8791, 95\}$, the simulated ground-truth ranking $R_s = \{r(e_{s_1}), r(e_{s_2}), r(e_{s_3}), r(e_{s_4})\}$ should satisfy the method-specific constraint (4): $\forall e_{s_j} \in E_s : (\Delta t(e_{s_3}) = -8791) < (\Delta t(e_{s_1}) = -3) < (\Delta t(e_{s_2}) = 0) < (\Delta t(e_{s_4}) = 95) \implies (r(e_{s_3}) = 1) < (r(e_{s_1}) = 2) < (r(e_{s_2}) = 3) < (r(e_{s_4}) = 4)$. Hence, $R_s = \{2, 3, 1, 4\}$.

*3.5.2. Claim-based date ranking*

**Hypothesis 2.** *Although newer evidence is more relevant than older evidence, fact-checkers can only base their veracity estimation on information which was available at fact checking time. In this case, evidence published after the claim date is not relevant and should be excluded from the documents to be ranked.*

In this method, we mimic the information accessibility and availability at claim time $t(c_i)$. As a result, only evidence that had been published before or at the same time as the claim is considered in this approach. For ground-truth evidence ranking $R_i$, all $e_{i_j}$ with $\Delta t(e_{i_j}) \leq 0$ are ranked according to the following constraint:

$$\begin{aligned} &\forall e_{i_j}, e_{i_k} \in E_i : \\ &(\Delta t(e_{i_j}) \leq \Delta t(e_{i_k})) \wedge (\Delta t(e_{i_j}) \wedge \Delta t(e_{i_k}) \leq 0) \\ &\implies r(e_{i_j}) \leq r(e_{i_k}) \end{aligned} \tag{8}$$

For two evidence snippets in the same evidence set, the constraint imposes a higher ranking score $r(e_{i_k})$ for $e_{i_k}$ and a lower ranking score $r(e_{i_j})$ for $e_{i_j}$ if $\Delta t(e_{i_k})$ is larger than $\Delta t(e_{i_j})$ and both $\Delta t(e_{i_k})$ and $\Delta t(e_{i_j})$ are negative or zero. If $\Delta t(e_{i_k})$ and $\Delta t(e_{i_j})$ are equal, $e_{i_j}$ and $e_{i_k}$ obtain the same ranking score. This is the only temporal ranking method that does not necessarily rank all given $e_{i_j} \in E_i$ because it excludes all $\Delta t(e_{i_j}) > 0$ from simulated ground-truth ranking $R_i$.

**Example**: For $E_s = \{e_{s_1}, e_{s_2}, e_{s_3}, e_{s_4}\}$ with $\Delta E_s^t = \{-3, 0, -8791, 95\}$, the simulated ground-truth ranking $R_s = \{r(e_{s_1}), r(e_{s_2}), r(e_{s_3}), r(e_{s_4})\}$ should satisfy the method-specific constraint (5): $\forall e_{s_j} \in E_s : ((\Delta t(e_{s_3}) = -8791) < (\Delta t(e_{s_1}) = -3) < (\Delta t(e_{s_2}) = 0)) \wedge (\Delta t(e_{s_1}) \wedge \Delta t(e_{s_2}) \wedge \Delta t(e_{s_3}) \leq 0) \implies (r(e_{s_3}) = 1) < (r(e_{s_1}) = 2) < (r(e_{s_2}) = 3)$. The ranking scores for $\Delta t(e_{s_j}) > 0$ are automatically set to 0. Hence, $R_s = \{2, 3, 1, 0\}$.

*3.5.3. Claim-centered distance ranking*

**Hypothesis 3.** *Assuming that a topic and its related subtopics are discussed around the same time [25], ev-*

| Temporal Ranking Method | Ranking Constraint $\forall e_{i_j}, e_{i_k} \in E_i$ : | Ranking $R_s$ |
|---|---|---|
| *Evidence-based* date ranking | $\Delta t(e_{i_j}) \leq \Delta t(e_{i_k}) \implies r(e_{i_j}) \leq r(e_{i_k})$ | $\{2, 3, 1, 4\}$ |
| *Claim-based* date ranking | $(\Delta t(e_{i_j}) \leq \Delta t(e_{i_k})) \wedge (\Delta t(e_{i_j}) \wedge \Delta t(e_{i_k}) \leq 0) \implies r(e_{i_j}) \leq r(e_{i_k})$ | $\{2, 3, 1, 0\}$ |
| *Claim-centered* distance ranking | $\|\Delta t(e_{i_j})\| \geq \|\Delta t(e_{i_k})\| \implies r(e_{i_j}) \leq r(e_{i_k})$ | $\{3, 4, 1, 2\}$ |
| *Evidence-centered* distance ranking | $\|\Delta cl(e_{i_j})\| \geq \|\Delta cl(e_{i_k})\| \implies r(e_{i_j}) \leq r(e_{i_k})$ | $\{4, 3, 1, 2\}$ |

Table 1: Overview of the four temporal ranking methods and their ranking constraints. The last column provides sample ground-truth rankings for a given evidence set $E_s = \{e_{s_1}, e_{s_2}, e_{s_3}, e_{s_4}\}$ with $\Delta E_s^t = \{-3, 0, -8791, 95\}$ and $\Delta E_s^{cl} = \{0, 3, 8788, 98\}$. A higher ranking score denotes higher relevance.

*idence is more relevant when it is published around the same time as the claim and becomes less relevant as the temporal distance between claim and evidence grows.*

This ranking method assigns ranking scores to evidence snippets in terms of their temporal vicinity to the claim. For ground truth ranking $R_i$, we rank all $e_{i_j}$ based on $|\Delta t(e_{i_j})|$ in ascending order, satisfying the following constraint:

$$\forall e_{i_j}, e_{i_k} \in E_i : |\Delta t(e_{i_j})| \geq |\Delta t(e_{i_k})| \implies r(e_{i_j}) \leq r(e_{i_k}) \tag{9}$$

For two evidence snippets in the same evidence set, the constraint imposes a higher ranking score $r(e_{i_j})$ for $e_{i_j}$ and a lower ranking score $r(e_{i_k})$ for $e_{i_k}$ if $|\Delta t(e_{i_j})|$ is smaller than $|\Delta t(e_{i_k})|$. If $|\Delta t(e_{i_j})|$ and $|\Delta t(e_{i_k})|$ are equal, $e_{i_j}$ and $e_{i_k}$ obtain the same ranking score.

**Example**: For $E_s = \{e_{s_1}, e_{s_2}, e_{s_3}, e_{s_4}\}$ with $\Delta E_s^t = \{-3, 0, -8791, 95\}$, the simulated ground-truth ranking $R_s = \{r(e_{s_1}), r(e_{s_2}), r(e_{s_3}), r(e_{s_4})\}$ should satisfy the method-specific constraint (6): $\forall e_{s_j} \in E_s : (|\Delta t(e_{s_3})| = 8791) > (|\Delta t(e_{s_4})| = 95) > (|\Delta t(e_{s_1})| = 3) > (|\Delta t(e_{s_2})| = 0) \implies (r(e_{s_3}) = 1) < (r(e_{s_4}) = 2) < (r(e_{s_1}) = 3) < (r(e_{s_2}) = 4)$. Hence, $R_s = \{3, 4, 1, 2\}$.

*3.5.4. Evidence-centered distance ranking*

**Hypothesis 4.** *Analogous to the assumption that relevant documents have a tendency to cluster in a shared document space [26], relevant evidence snippets also cluster in time. Therefore, evidence snippets that are clustered in time are more relevant than evidence snippets that are temporally distant from the others.*

This ranking method assigns ranking scores to evidence snippets in terms of their temporal vicinity to the other snippets in the same evidence set. We first detect the medoid of all $\Delta t(e_{i_j}) \in \Delta E_i^t$ by computing a pairwise distance matrix, summing the columns and finding the argmin of the summed pairwise distance values. Then, the Euclidean distance between all $\Delta t(e_{i_j})$ and the detected medoid is calculated.

$$\Delta cl : \Delta E_i^t \to \Delta E_i^{cl} = \{\Delta cl(e_{i_j}), ..., \Delta cl(e_{i_K})\} \tag{10}$$

We rank all $\Delta cl(e_{i_j}) \in \Delta E_i^{cl}$ in ascending order, resulting in ground-truth ranking $R_i$ which satisfies the following constraint:

$$\forall e_{i_j}, e_{i_k} \in E_i : |\Delta cl(e_{i_j})| \geq |\Delta cl(e_{i_k})| \implies r(e_{i_j}) \leq r(e_{i_k}) \tag{11}$$

For two evidence snippets in the same evidence set, the constraint imposes a higher ranking score $r(e_{i_j})$ for $e_{i_j}$ and a lower ranking score $r(e_{i_k})$ for $e_{i_k}$ if $|\Delta cl(e_{i_j})|$ is smaller than $|\Delta cl(e_{i_k})|$. If $|\Delta cl(e_{i_j})|$ and $|\Delta cl(e_{i_k})|$ are equal, $e_{i_j}$ and $e_{i_k}$ obtain the same ranking score.

**Example**: In order to obtain $\Delta E_s^{cl} = \{\Delta cl(e_{s_1}), \Delta cl(e_{s_2}), \Delta cl(e_{s_3}), \Delta cl(e_{s_4})\}$:

- Compute the pairwise distance matrix for $\Delta E_s^t = \{-3, 0, -8791, 95\}$:
$$PD = \begin{Bmatrix} 0 & 3 & 8788 & 98 \\ 3 & 0 & 8791 & 95 \\ 8788 & 8791 & 0 & 8886 \\ 98 & 95 & 8886 & 0 \end{Bmatrix}$$
- Sum the matrix columns:
$SD = \begin{Bmatrix} 8889 & 8889 & 26465 & 9079 \end{Bmatrix}$
- Find the argmin and mediod[3]:
$argmin(SD) = 8889, mediod = \Delta t(e_{s_1}) = -3$
- Calculate the Euclidean distance between all $\Delta t(e_{s_j})$ and the detected mediod:
$\Delta cl : \Delta E_s^t \rightarrow \Delta E_s^{cl} = \{0, 3, 8788, 98\}$

For $E_s = \{e_{s_1}, e_{s_2}, e_{s_3}, e_{s_4}\}$ with $\Delta E_s^{cl} = \{0, 3, 8788, 98\}$, the simulated ground-truth ranking $R_s = \{r(e_{s_1}), r(e_{s_2}), r(e_{s_3}), r(e_{s_4})\}$ should satisfy the method-specific constraint (8): $\forall e_{s_j} \in E_s : (|\Delta cl(e_{s_3})| = 8788) > (|\Delta cl(e_{s_4})| = 98) > (|\Delta cl(e_{s_2})| = 3) > (|\Delta cl(e_{s_1})| = 0) \implies (r(e_{s_3}) = 1) < (r(e_{s_4}) = 2) < (r(e_{s_2}) = 3) < (r(e_{s_1}) = 4)$. Hence, $R_s = \{4, 3, 1, 2\}$.

[3] In $\Delta E_i^t$ with an even number of $\Delta t(e_{i_j})$, the algorithm finds two mediods instead of one. For sake of consistency, we set the first mediod in the set as general mediod.

## 4. Experimental Setup

**Dataset**. We opt for the MultiFC dataset [17], as it is the only large, publicly available fact-checking dataset which provides temporal information for both claims and evidence pages, and follow their experimental setup. The dataset contains 34,924 real-world claims extracted from 26 different fact-check websites. The fact-check domains are abbreviated to four-letter contractions (e.g., Snopes to *snes*, Washington Post to *wast*). For each claim, metadata such as speaker, tags and categories are included and a maximum of ten evidence snippets crawled from the Internet using the Google Search API are used to predict a claim's veracity label.

Regarding temporal information, the dataset provides an explicit timestamp for each claim as structured metadata. For the evidence snippets, however, we need to extract their timestamp from the evidence text itself. The publishing date is often contained in the document text, most frequently at the beginning of the text and immediately followed by an ellipsis (i.e., '...'). We split the text at the ellipsis and take the left part as timestamp. If Python's datetime module recognizes the extracted timestamp as a real timestamp, it is regarded as ground-truth evidence timestamp. Otherwise, the timestamp is not included in the dataset. Both claim and evidence date are then automatically formatted as *year-month-day* using Python's datetime module.[4] If datetime is unable to correctly format an extracted

[4] We randomly extracted 150 timestamps from claims and evidence

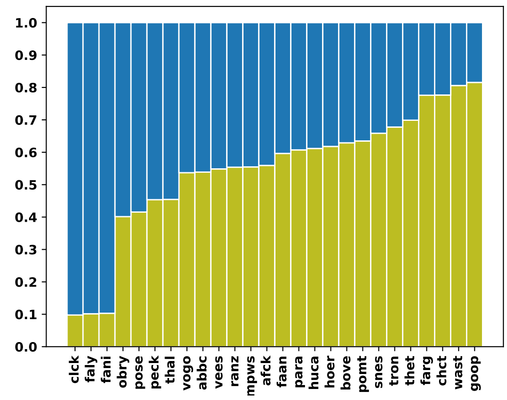



Figure 3: The share of evidence snippets with retrievable timestamps per domain.

timestamp, we do not include it in the dataset. Figure 3 displays the distribution of evidence snippets with and without a retrievable timestamp per domain. Finally, the dataset is split in training (27,940), development (3,493) and test (3,491) set in a label-stratified manner.

**Pre-Training and Fine-Tuning.** The models are first pre-trained on all domains before they are fine-tuned for each domain separately. Pre-training the models on all domains is advantageous, as some domains contain few training data. In the pre-training phase, batches from each domain are alternately fed to the model during each epoch with the maximum number of batches for all domains equal to the maximum number of batches for the smallest domain so that the model is not biased towards the larger domains. For the veracity classification task, cross-entropy loss is computed over the label probabilities, and RMSProp (BiLSTM/RNN) or AdamW (DistilBERT) optimizes

snippets in the dataset and manually verified whether datetime correctly parsed them in *year-month-day*. As zero mistakes were found, we consider datetime sufficiently accurate.

| Phase | BiLSTM | RNN | DistilBERT |
|---|---|---|---|
| Only PT | .5998/.3436 | .5425/.2960 | .5891/**.3282** |
| Only FT | .6072/.3525 | .5862/**.3513** | .5424/.2859 |
| PT + FT | **.6265/.3673** | **.5899**/.3453 | **.5921**/.3135 |

Table 2: Overview of the aggregated test results (Micro/Macro F1) when optimized on the temporal ranking methods. The methods are integrated during pre-training only (Only PT), during fine-tuning only (Only FT) or during both pre-training and fine-tuning (PT + FT).

all the model parameters except the evidence ranking parameters. For the evidence ranking task, ListMLE loss is computed over the ranking scores, and Adam optimizes only the evidence ranking parameters. We use the ListMLE loss function from the allRank library [27]. An overview of all hyperparameter settings is included in the Appendix. In the fine-tuning phase, we select the best performing pre-trained model based on the development set for each domain individually and fine-tune it on that domain.

We found that directly optimizing the evidence ranking on the temporal ranking methods in both the pre-training and fine-tuning phase yields the highest results for all models (Table 2).

## 5. Results

We take the best performing model per domain based on the development set and report the average over all domain-specific test results on the veracity prediction task (Table 3). We use Micro and Macro F1 score as evaluation metrics. The results of our BiLSTM base model are comparable[5] to those of Augenstein et al.

[5] Augenstein et al. [17] train the model differently: they train the multi-task learning model over all domains and perform early stop-

| | BiLSTM | | RNN | | DistilBERT | |
|---|---|---|---|---|---|---|
| | Micro F1 | Macro F1 | Micro F1 | Macro F1 | Micro F1 | Macro F1 |
| **Base model** | **.5521** | **.3185** | **.5053** | **.2691** | **.5558** | **.2909** |
| Search ranking | .5468 | .2864 | .4782 | .2372 | .4793 | .2264 |
| Evidence-based date ranking | .5794 | .3227 | .5359 | .2905 | .5424 | .2660 |
| Claim-based date ranking | .5668 | .3215 | .5271 | .3082 | .4791 | .2391 |
| Claim-centered distance ranking | .5532 | .3029 | .4981 | .2700 | .5257 | .2644 |
| Evidence-centered distance ranking | .4756 | .2417 | .5356 | .3019 | .5759 | .2962 |
| **Time-aware evidence ranking** | **.6265** | **.3673** | **.5899** | **.3453** | **.5921** | **.3135** |

Table 3: Aggregated test results for veracity prediction task, with improvements over base model performance underlined.

[17]. As an additional ranking baseline, we optimize evidence ranking on the evidence snippets' position in the Google search ranking list given in the dataset (*search ranking*): the higher in the list, the higher ranked in the simulated ground-truth evidence ranking. For the *time-aware evidence ranking* results, we take for each domain the best performing temporal ranking model and report again the average over all domain-specific test results. We present domain-specific test results for the BiLSTM model in Table 4 and include the domain-specific test results for the RNN and DistilBERT model in the Appendix.

Firstly, the temporal ranking methods generally outperform search engine ranking, which consistently performs worse than the base models (-0.53/-2.71/-7.65% Micro F1; BiLSTM/RNN/DistilBERT). Secondly, the temporal ranking methods affect model performance to various extents. Constraining evidence ranking in all domains using a date-based ranking method positively influences the BiLSTM and RNN model performance - with *evidence-based date ranking* (+2.73/+3.06% Micro F1) leading to slightly higher results than *claim-based date ranking* (+1.47/+2.18% Micro F1). We observe similar performance differences between the date-based ranking methods in the DistilBERT model, however, both methods fall behind the base model (-1.34/-7.67% Micro F1; *evidence-based/claim-based date ranking*). For this model, only *evidence-centered distance ranking* yields increased test results (+2.01% Micro F1). Although this ranking method returns higher test results for the RNN model as well, it is the only temporal ranking method that decreases the BiLSTM model performance (-7.65% Micro F1). In contrast to the other three temporal ranking methods, *claim-centered distance ranking* does not lead to a substantial performance gain in any of the three fact-checking models (+0.11/-0.72/-3.01% Micro F1; BiLSTM/RNN/DistilBERT). While performance gains are often limited when applying a single temporal ranking method to all domains, higher results can be obtained by selecting the best performing temporal ranking method per domain based on the development set. As a result, *time-aware evidence ranking* increases model performance by 7.44% for the BiLSTM model, 8.46% for the RNN model and 3.63%

ping based on accuracy on the target domain development set. We pre-train the model over all domains, select the best performing model per domain based on the development set, and fine-tune each domain-specific model individually. Hence the difference in test results.

| | Base Model | | Search Ranking | | Evidence Date | | Claim Date | | Claim Distance | | Evidence Distance | | Time-Aware Ranking | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro |
| abbc | **.3148** | **.2782** | .5370 | .2329 | .5000 | .2222 | .4630 | .2137 | .4074 | .2578 | .5370 | .2358 | **.5370** | **.2358** |
| afck | **.2222** | **.1385** | .1667 | .0706 | .2778 | .0725 | .2778 | .0741 | .2778 | .0725 | .2222 | .0571 | **.2778** | **.0741** |
| bove | **1.** | **1.** | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | **1.** | **1.** |
| chct | **.5500** | **.2967** | .4250 | .2294 | .4750 | .2956 | .4000 | .2234 | .4500 | .1552 | .4000 | .2267 | **.4750** | **.1959** |
| clck | **.5000** | **.2500** | .1667 | .0952 | .8333 | .6061 | .6667 | .5333 | .5000 | .2500 | .5000 | .2500 | **.8333** | **.6061** |
| faan | **.5000** | **.5532** | .1364 | .1216 | .5455 | .4545 | .5000 | .3266 | .5000 | .4256 | .4091 | .2837 | **.5455** | **.4545** |
| faly | **.6429** | **.2045** | .6429 | .1957 | .8571 | .4521 | .6429 | .1957 | .4286 | .1667 | .2857 | .3000 | **.8571** | **.4521** |
| fani | **1.** | **1.** | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | **1.** | **1.** |
| farg | **.6923** | **.1023** | .6923 | .1023 | .6923 | .1023 | .6731 | .1006 | .6923 | .1023 | .7308 | .2003 | **.7308** | **.2003** |
| goop | **.8344** | **.1516** | .8344 | .1516 | .8344 | .6222 | .8245 | .1512 | .8278 | .1512 | .8344 | .1516 | **.8344** | **.1516** |
| hoer | **.4104** | **.2091** | .4403 | .1975 | .4627 | .2200 | .3955 | .1474 | .4701 | .2854 | .3209 | .1106 | **.4701** | **.2854** |
| huca | **.5000** | **.2857** | .5000 | .2857 | .8333 | .6222 | .5000 | .2857 | .6667 | .5333 | .6667 | .5926 | **.8333** | **.6222** |
| mpws | **.8750** | **.6316** | .8750 | .6316 | .5000 | .2222 | .8750 | .6316 | .5000 | .2424 | .2500 | .1333 | **.8750** | **.6316** |
| obry | **.5714** | **.3520** | .3571 | .1316 | .3571 | .1316 | .4286 | .1500 | .5000 | .2982 | .3571 | .1316 | **.5000** | **.2982** |
| para | **.1875** | **.1327** | .2500 | .0667 | .3125 | .1720 | .2813 | .1584 | .1563 | .0490 | .1875 | .0556 | **.3125** | **.1720** |
| peck | **.5833** | **.2593** | .9167 | .6154 | .5833 | .2593 | .6667 | .8213 | .9167 | .6316 | .2500 | .1724 | **.9167** | **.6316** |
| pomt | **.2143** | **.2089** | .2071 | .1015 | .1732 | .0604 | .1643 | .0539 | .1732 | .0329 | .1813 | .0594 | **.1813** | **.0594** |
| pose | **.4178** | **.0987** | .4178 | .0987 | .3836 | .1112 | .4247 | .1330 | .4041 | .1990 | .4110 | .1422 | **.4247** | **.1330** |
| ranz | **1.** | **1.** | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | .5000 | .3333 | **1.** | **1.** |
| snes | **.5646** | **.0601** | .5646 | .0614 | .5662 | .1026 | .5646 | .0753 | .5646 | .0601 | .5646 | .0602 | **.5662** | **.1026** |
| thal | **.5556** | **.2756** | .5556 | .1786 | .4444 | .1538 | .5556 | .1852 | .5556 | .1852 | .5556 | .1852 | **.5556** | **.1852** |
| thet | **.3750** | **.2593** | .6250 | .3556 | .6250 | .3556 | .6250 | .3556 | .6250 | .3556 | .4375 | .1014 | **.6250** | **.3556** |
| tron | **.4767** | **.0258** | .4767 | .0325 | .4738 | .0258 | .4767 | .0258 | .4622 | .0289 | .4767 | .0258 | **.4767** | **.0366** |
| vees | **.7097** | **.2075** | .7097 | .2075 | .6613 | .1990 | .6129 | .2179 | .7097 | .2075 | .6935 | .2885 | **.7097** | **.2075** |
| vogo | **.5000** | **.2056** | .5000 | .2158 | .4219 | .2531 | .5000 | .2262 | .3750 | .1234 | .3750 | .1114 | **.5000** | **.2262** |
| wast | **.1563** | **.0935** | .2188 | .0667 | .2500 | .1429 | .2188 | .0718 | .2188 | .1412 | .2188 | .0757 | **.2500** | **.1429** |
| avg. | **.5521** | **.3185** | .5468 | .2864 | .5794 | .3227 | .5668 | .3215 | .5532 | .3029 | .4756 | .2417 | **.6265** | **.3673** |

Table 4: Overview of classification test results (Micro F1 and Macro F1) for **BiLSTM model - PT + FT**, with improvements over base model results underlined. The last column contains the highest results returned by one of the four temporal ranking methods.

for the DistilBERT model (Micro F1).

## 6. Discussion

Introducing time-awareness in a fact-checking model by directly optimizing evidence ranking using temporal ranking methods positively influences classification performance. Moreover, time-aware evidence ranking consistently outperforms search engine evidence ranking. This suggests that the temporal ranking methods themselves - and not merely the act of direct evidence ranking optimization - lead to higher results.

However, one could ask to which extent the temporal ranking methods actually change the evidence ranking order. To quantify the difference in returned evidence rankings between the base model and the model optimized on one of the temporal ranking methods, we compute the Spearman's rank correlation coefficient $r_s$. We find that in the BiLSTM model, the temporal ranking methods consistently change ranking orders, as the time-aware evidence rankings are rather weakly correlated with the base rankings ($r_s = .24/.18/.22/.17$; for *evidence-based date ranking*, *claim-based date ranking*, *claim-centered distance ranking* and *evidence-centered distance ranking*, respectively). Although the changes in ranking order are more drastic for the BiLSTM model, the temporal ranking methods have a weaker but still considerable impact on evidence ranking for the RNN model ($r_s = .56/.58/.61/.54$). The lower impact could be attributed to the ranking module's depth: the BiLSTM model's ranking module consists of twice as many nonlinear fully-connected layers than the RNN's ranking module. As a result, the deeper ranking module is able to learn more detailed and abstract representations.

While the correlations between time-aware and base ranking order are mainly positive in the BiLSTM and RNN model, evidence rankings are either negatively or completely uncorrelated in the DistilBERT model ($r_s = -.23/-.35/-.02/-.07$). In the latter model, the sentence encoder yields both a label probability distribution and a joint representation for a given claim and evidence snippet. For the other two models (BiLSTM and RNN), the sentence encoder outputs separate hidden representations for claims and evidence snippets, which are later fused to joint representations. Label scores for each joint representation are then inferred by the label scoring module. It can thus be argued that the joint representations in the BiLSTM model are more label-aware, which in turn lead to a more label-biased ranking module that possibly runs counter to a time-aware one. Hence the negative correlations between time-aware and base rankings. We can conclude that the temporal ranking methods indeed cause changes in evidence ranking order, although their impact varies with each model.

The temporal ranking methods affect overall model performance to various extents: the BiLSTM and RNN model mainly prefer date-based ranking methods, while the DistilBERT model's performance is only increased using the evidence-centered distance ranking. Moreover, there is not a single temporal ranking method that increases veracity prediction performance across all models. Similar observations can be made on a by-domain level. Some domains benefit from all four temporal ranking methods, while others consistently perform worse or are not even affected at all.

Whether or not time-aware evidence ranking affects a domain's model performance might be ascribed to the share of evidence snippets with retrievable timestamps in that domain. An evidence snippet's computed ranking score can only optimized by the learning-to-rank loss function if its timestamp can be extracted and, in case of *claim-based date ranking*[6], is included in the method-specific ground-truth evidence ranking. A large share of time-grounded evidence snippets enables the model/domain to learn the expected rankings in a more direct and constrained manner, while a small share allows more flexible ranking learning. Consequently, it could be argued that the temporal ranking methods influence domains with a large share of time-grounded evidence snippets (*farg*, *chct*, *wast*, *goop*) more strongly and positively than those with a small share (*clck*, *faly*, *fani*). However, that argument is refuted as several small-share domains have their performance increased by large margins, and large-share domains do not consistently benefit from time-aware evidence ranking to a great extent. However, these findings suggest that the effectiveness of our temporal ranking methods does not rely on a large amount of time-grounded evidence.

Another possible cause of the inter-domain differences in time-aware ranking effect is the time-sensitivity of those domains. We hypothesize that domains tackling claims on time-sensitive subjects benefit more from time-aware evidence ranking than those discussing time-insensitive claims. We retrieve the categories[7] from several domains and analyze their time-sensitivity. The analysis confirms our hypothesis: domains which mainly tackle time-sensitive subjects such as politics, economy, climate and entertainment (*abbc*, *para*, *thet*) benefit more from time-aware evidence ranking than domains discussing both time-sensitive and time-insensitive subjects such as food, language, humor and animals (*snes*, *tron*). We can therefore conclude that relating evidence relevance with time and ranking evidence snippets accordingly is beneficial for time-sensitive claims.

We observe not only inter-model and inter-domain but also inter-method differences. Regarding date-based ranking, a domain or model preference for either evidence-based or claim-based date ranking might depend on the share of evidence posted after the claim date. If an evidence set mainly consists of later-posted evidence, the ranking of only a few evidence snippets is directly optimized with the claim-based date ranking method, leaving the model to indirectly learn the ranking scores of the others. In that case, the evidence-based method might be favored over the claim-based method. However, the share of later-posted evidence is not consistently different in domains preferring evidence-based date ranking than in domains favoring claim-based date ranking.

Concerning distance-based ranking, evidence and claim (claim-centered distance ranking), and evidence and evidence (evidence-centered distance ranking) are more likely to discuss the same topic when they are published around the same time. The distance-based ranking methods would thus increase classification performance for domains in which the dispersion of evidence snippets in the claim-specific evidence sets is small. We measure the temporal dispersion of each evidence set using the standard deviation in domains which mainly favor distance-based ranking over date-based ranking (*afck*, *faan*, *pomt*, *thal*; Group 1), and vice versa (*chct*, *vogo*; Group 2).

We then check whether the domains in these groups

[6]When ranking is constrained using the *claim-based date ranking* method, only evidence snippets with $\Delta t \leq 0$ are ranked in the simulated ground-truth rankings.

[7]Categories are given as claim metadata.

display similar dispersion values. Kruskal-Wallis H tests indicate that dispersion values statistically differ between domains in the same group (Group 1: $H = 258.63, p < 0.01$; Group 2: $H = 192.71, p < 0.01$). Moreover, Mann-Whitney U tests on domain pairs suggest that inter-group differences are not consistently larger than intra-group differences (e.g., *thal-chct*: $Mdn = 337.53, Mdn = 239.26, p = 0.021 > 0.01$; *thal-pomt*: $Mdn = 337.53, Mdn = 661.25, p = 9.95e^{-5} < 0.01$). Therefore, the hypothesis that small evidence dispersion causes a preference for distance-based ranking methods is rejected.

The impact and success of the temporal ranking methods still depends on the informativeness of the given Web documents which serve as evidence to the claim. As the Web documents are automatically crawled from the Internet given the claim as query to the Google Search API, it is not guaranteed that they are all useful for predicting that claim's veracity. If a large number of Web documents in the evidence set do not contain useful information for refuting or supporting the claim, enforcing a temporal ranking will then have little to no effect on model performance. We randomly pick a claim with a large number of evidence snippets with a retrievable timestamp, but for which both base and time-aware models consistently return an incorrect veracity label (Table 5). The veracity and semantics of the claim are time-dependent (i.e., a trade balance varies over time), but the majority of Web documents are irrelevant to the claim. Consequently, a fact-checking model cannot make an informed veracity prediction - independent of how it ranks the evidence.

## 7. Conclusion

Introducing time-awareness in evidence ranking arguably leads to more accurate veracity predictions in fact-checking models – especially when they deal with claims about time-sensitive subjects such as politics and entertainment. These performance gains also indicate that evidence relevance should be approached more diversely instead of merely associating it with the semantic similarity between claim and evidence. By integrating temporal ranking constraints in neural architectures via appropriate loss functions, we show that fact-checking models are able to learn time-aware evidence rankings in an elegant, yet effective manner. To our knowledge, evidence ranking optimization using a dedicated ranking loss has not been studied before in the context of fact-checking. Whereas this study is limited to integrating time-awareness in the evidence ranking as part of automated fact-checking, future research could build on these findings to explore the impact of time-awareness at other stages of fact-checking, e.g., document retrieval or evidence selection, and in domains beyond fact-checking. Alternatively, the analogy with spatial relevance can be explored by adopting similar spatial ranking methods for space-aware evidence ranking.

## Acknowledgements

This work was realised with the collaboration of the European Commission Joint Research Centre under the Collaborative Doctoral Partnership Agreement No 35332 and has been supported by COST Action CA18231.

| Timestamp | **Claim** |
| --- | --- |
| Mar 20, 2018 | 'Claim on Scotland's positive trade balance compared to UK nations' |
| | **Evidence Set** |
| Mar 20, 2018 | Mar 20, 2018 ... Claim on Scotland's positive trade balance compared to UK nations is ... with a new Mostly False verdict and correction note, and re-posted to ... |
| / | Advice, and information so that anyone can check the claims we hear about ... on Scotland's positive trade balance compared to UK nations is Mostly False |
| Mar 10, 2018 | Mar 10, 2018 ... Claim: The US is suffering from a trade imbalance, with a trade ... Reality Check verdict: The President is incorrect about the US trade deficit - it was $566bn (£ 410bn) in 2017. ... declined in most major western industrialised nations over the past ... 9 Two British soldiers injured in Islamic State attack in Syria ... |
| Sep 7, 2016 | Sep 7, 2016 ... Adam Smith's 1776 classic "Wealth of Nations" may have had the ... What was the most important document published in 1776? ... Smith, a Scottish philosopher by trade, wrote the book to upend the mercantilist system. ... fallacies in an argument that is framed as the invisible hand versus the government. |
| Apr 17, 2018 | Apr 17, 2018 ... Everything You've Been Told About Government Debt Is Wrong ... usually presume that the government will run a primary surplus (the excess of ... For example, during the era of relative peace following 1960, decadal ... Using this as a parameter, Barrett estimates a safe debt/GDP level for the U.K. of 140%. |
| Apr 20, 2018 | Apr 20, 2018 ... President Donald Trump's trade policy leaves international ... trade system, which I'd argue has benefited the nation economically overall ... of thought most economists believe Adam Smith extinguished after he ... Second, imposing new and higher tariffs on imports won't make the U.S. trade deficit go away. |
| Jun 29, 2016 | Jun 29, 2016 ... Similar information for devolved administrations are available at Scotland: Fire and Rescue Statistics ... FIRE0103: Fires attended by fire and rescue services by nation and ... FIRE0104: Fire false alarms by reason for false alarm, England (MS .... FIRE1303: Firefighters' pension surplus and deficit (MS Excel... |
| Aug 6, 2017 | Aug 6, 2017 ... After the bitter referendums over Scottish independence and ... Judged on hard metrics, confidence in UK media has fallen .... Remainers detected dangerous instances of false balance, most notoriously when a poll found that 88% of UK ... who said that Brexit would not damage trade and the UK economy. |
| / | ... the failure of people in the rich nations to make any significant sacrifices in ... It is not simply the absence of charity, let alone of moral saintliness: It is wrong, and one cannot claim to be a morally decent ... without thereby sacrificing anything of comparable moral importance, we... |

Table 5: Claim and accompanying evidence set from the dataset, for which both base and time-aware models consistently predict an incorrect veracity label.

## References


[1] T. Halpin, Temporal modeling and orm, in: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", Springer, Berlin, Heidelberg, 2008, pp. 688–698. doi:10.1007/978-3-540-88875-8_93.

[2] R. Mishra, V. Setty, SADHAN: Hierarchical attention networks to learn latent aspect embeddings for fake news detection, in: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, Association for Computing Machinery, New York, 2019, pp. 197–204. doi:10.1145/3341981.3344229.

[3] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, 2018, pp. 809–819. doi:10.18653/v1/N18-1074.

[4] Z. Liu, C. Xiong, M. Sun, Z. Liu, Fine-grained fact verification with kernel graph attention network, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7342–7351. doi:10.18653/v1/2020.acl-main.655.

[5] X. Li, W. Meng, T. Y. Clement, Verification of fact statements with multiple truthful alternatives, in: Proceedings of the 12th International Conference on Web Information Systems and Technologies - Volume 2: WEBIST, SciTePress, 2016, pp. 87–97. doi:10.5220/0005781800870097.

[6] X. Li, W. Meng, C. Yu, T-verifier: Verifying truthfulness of fact statements, in: 2011 IEEE 27th International Conference on Data Engineering, IEEE, Hannover, 2011, pp. 63–74. doi:10.1109/ICDE.2011.5767859.

[7] T. Wang, Q. Zhu, S. Wang, Multi-verifier: A novel method for fact statement verification, World Wide Web 18 (2015) 1463–1480.

[8] W. Zhong, J. Xu, D. Tang, Z. Xu, N. Duan, M. Zhou, J. Wang, J. Yin, Reasoning over semantic-level graph for fact checking, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6170–6180. doi:10.18653/v1/2020.acl-main.549.

[9] A. Hanselowski, H. Zhang, Z. Li, D. Sorokin, B. Schiller, C. Schulz, I. Gurevych, UKP-Athene: Multi-sentence textual entailment for claim verification, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, 2018, pp. 103–108. doi:10.18653/v1/W18-5516.

[10] J. Zhou, X. Han, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Gear: Graph-based evidence aggregating and reasoning for fact verification, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, 2019, pp. 892–901. doi:10.18653/v1/P19-1085.

[11] Y. Nie, H. Chen, M. Bansal, Combining fact extraction and verification with neural semantic matching networks, in: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, Honolulu, 2019, pp. 6859–6866. doi:10.1609/aaai.v33i01.33016859.

[12] S. Miranda, D. Nogueira, A. Mendes, A. Vlachos, A. Secker, R. Garrett, J. Mitchel, Z. Marinho, Automated fact checking in the news room, in: The World Wide Web Conference, Association for Computing Machinery, New York, 2019, pp. 3579–3583. doi:10.1145/3308558.3314135.

[13] T. Yoneda, J. Mitchell, J. Welbl, P. Stenetorp, S. Riedel, UCL machine reading group: Four factor framework for fact finding (hexaf), in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, 2018, pp. 97–102. doi:10.18653/v1/W18-5515.

[14] Y. Yamamoto, T. Tezuka, A. Jatowt, K. Tanaka, Supporting judgment of fact trustworthiness considering temporal and sentimental aspects, in: International Conference on Web Information Systems Engineering, Springer, 2008, pp. 206–220. doi:10.1007/978-3-540-85481-4_17.

[15] C. Hidey, T. Chakrabarty, T. Alhindi, S. Varia, K. Krstovski, M. Diab, S. Muresan, DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8593–8606. doi:10.18653/v1/2020.acl-main.761.

[16] Y. Zhou, T. Zhao, M. Jiang, A probabilistic model with commonsense constraints for pattern-based temporal fact extraction, in: Proceedings of the Third Workshop on Fact Ex-

traction and VERification (FEVER), Association for Computational Linguistics, Online, 2020, pp. 18–25. doi:10.18653/v1/2020.fever-1.3.

[17] I. Augenstein, C. Lioma, D. Wang, L. C. Lima, C. Hansen, C. Hansen, J. G. Simonsen, Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, 2019, pp. 4685–4697. doi:10.18653/v1/D19-1475.

[18] L. Mou, R. Men, G. Li, Y. Xu, L. Zhang, R. Yan, Z. Jin, Natural language inference by tree-based convolution and heuristic matching, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Berlin, 2016, pp. 130–136. doi:10.18653/v1/P16-2022.

[19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[20] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, H. Li, Listwise approach to learning to rank: Theory and algorithm, in: Proceedings of the 25th International Conference on Machine Learning, Association for Computing Machinery, New York, 2008, pp. 1192–1199. doi:10.1145/1390156.1390306.

[21] T.-Y. Liu, Learning to Rank for Information Retrieval, Springer, Heidelberg, 2011.

[22] N. Kanhabua, A. Anand, Temporal information retrieval, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, 2016, pp. 1235–1238. doi:10.1145/2911451.2914805.

[23] R. Bansal, M. Rani, H. Kumar, S. Kaushal, Temporal information retrieval and its application: A survey, in: N. R. Shetty, L. M. Patnaik, H. C. Nagaraj, P. N. Hamsavath, N. Nalini (Eds.), Emerging Research in Computing, Information, Communication and Applications. Advances in Intelligent Systems and Computing, Springer, Singapore, 2019, pp. 251–262. doi:10.1007/978-981-13-6001-5_19.

[24] X. Li, W. B. Croft, Time-based language models, in: Proceedings of the Twelfth International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, 2003, pp. 469–475. doi:10.1145/956863.956951.

[25] F. Martins, J. Magalhães, J. Callan, Modeling temporal evidence from external collections, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, New York, 2019, pp. 159–167. doi:10.1145/3289600.3290966.

[26] M. Efron, J. Lin, J. He, A. de Vries, Temporal feedback for tweet search with non-parametric density estimation, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Association for Computing Machinery, New York, 2014, p. 33–42. doi:10.1145/2600428.2609575.

[27] P. Pobrotyn, T. Bartczak, M. Synowiec, R. Bialobrzeski, J. Bojar, Context-aware learning to rank with self-attention, in: SIGIR eCom '20, Virtual Event, China, 2020.

## Appendix

| Hyperparameter settings | |
|---|---|
| batch size | 32 |
| epochs | 150 |
| word embedding size | 300 |
| word embedding weight initialization | Xavier uniform |
| BiLSTM # layers | 2 |
| BiLSTM skip-connections | concatenation |
| BiLSTM hidden size | 128 |
| CNN out channels | 32 |
| CNN kernel size | 32 |
| CNN activation function | ReLU |
| ranking FC weight initialization | Xavier uniform |
| ranking FC (1) activation function | ReLU |
| ranking FC (2) activation function | Leaky ReLU |
| label embedding size | 16 |
| label embedding weight initialization | Xavier uniform |
| label embedding FC weight initialization | Xavier uniform |
| label embedding FC activation function | Leaky ReLU |
| RMSProp learning rate | 0.0002 |
| Adam learning rate | 0.001 |

Table 6: Hyperparameter settings **BiLSTM model**.

| Hyperparameter settings | |
|---|---|
| batch size | 32 |
| epochs | 150 |
| word embedding size | 300 |
| word embedding weight initialization | Xavier uniform |
| RNN # layers | 2 |
| RNN dropout | 0.1 |
| RNN hidden size | 128 |
| CNN out channels | 32 |
| CNN kernel size | 32 |
| CNN activation function | ReLU |
| ranking FC weight initialization | Xavier uniform |
| ranking FC activation function | Sigmoid |
| label FC weight initialization | Xavier uniform |
| label FC activation function | Leaky ReLU |
| RMSProp learning rate | 0.0002 |
| Adam learning rate | 0.001 |

Table 7: Hyperparameter settings **RNN model**.

| Hyperparameter settings | |
|---|---|
| batch size | 32 |
| epochs | 10 |
| DistilBERT configuration | distilbert-base-uncased |
| DistilBERT model | DistilBERT for sequence classification |
| ranking FC weight initialization | Xavier uniform |
| ranking FC (1) activation function | ReLU |
| ranking FC (2) activation function | Leaky ReLU |
| AdamW learning rate | 0.0005 |
| Adam learning rate | 0.01 |

Table 8: Hyperparameter settings **DistilBERT model**.

| | Base Model | | Search Ranking | | Evidence Date | | Claim Date | | Claim Distance | | Evidence Distance | | Time-Aware Ranking | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro |
| abbc | .4815 | .3111 | .2963 | .2301 | .5000 | .2278 | .5556 | .4233 | .4444 | .2133 | .4444 | .3139 | .5556 | .4233 |
| afck | .2500 | .1191 | .2778 | .0725 | .2222 | .1215 | .2500 | .1252 | .3056 | .1546 | .3056 | .1891 | .3056 | .1891 |
| bove | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. |
| chct | .4000 | .2337 | .4000 | .1429 | .4000 | .2122 | .4250 | .2347 | .6250 | .3365 | .4500 | .2836 | .6250 | .3365 |
| clck | .1667 | .1333 | .1667 | .0952 | .6667 | .4545 | .8333 | .6222 | .5000 | .3590 | .5000 | .2500 | .8333 | .6222 |
| faan | .5000 | .3529 | .4545 | .2083 | .4091 | .2000 | .3636 | .3228 | .5000 | .4442 | .4545 | .3472 | .5000 | .4442 |
| faly | .5714 | .1818 | .5000 | .1750 | .5714 | .1905 | .6429 | .1957 | .5000 | .2321 | .5000 | .2321 | .6429 | .1957 |
| fani | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. |
| farg | .6923 | .1023 | .6923 | .1023 | .7115 | .1803 | .6923 | .1023 | .6923 | .1023 | .6538 | .0988 | .7115 | .1803 |
| goop | .8344 | .1516 | .8344 | .1516 | .8344 | .1516 | .8344 | .1516 | .8344 | .1516 | .8344 | .1516 | .8344 | .1516 |
| hoer | .3657 | .1120 | .3731 | .0446 | .3806 | .1260 | .3731 | .1152 | .3657 | .1292 | .3433 | .0982 | .3806 | .1260 |
| huca | .3333 | .2222 | .1667 | .2222 | .3333 | .1667 | .1667 | .0952 | .1667 | .0952 | .5000 | .2500 | .5000 | .2500 |
| mpws | .5000 | .2222 | .1250 | .0833 | .6250 | .4148 | .3750 | .3241 | .3750 | .1818 | .6250 | .4809 | .6250 | .4809 |
| obry | .3571 | .1471 | .4286 | .2755 | .5000 | .2847 | .4286 | .3056 | .0714 | .0357 | .2857 | .1641 | .5000 | .2847 |
| para | .2188 | .1033 | .1875 | .1104 | .1875 | .1061 | .2813 | .1415 | .2188 | .1643 | .3438 | .2993 | .3438 | .2993 |
| peck | .5000 | .3397 | .5833 | .2456 | .5833 | .3679 | .7500 | .5095 | .5000 | .2222 | .5833 | .3679 | .7500 | .5095 |
| pomt | .1955 | .0792 | .1750 | .0753 | .2196 | .1578 | .2286 | .1651 | .2152 | .1627 | .2402 | .1699 | .2402 | .1699 |
| pose | .4110 | .1335 | .4178 | .0982 | .4178 | .0982 | .4041 | .1908 | .3904 | .1279 | .4178 | .1864 | .4178 | .1864 |
| ranz | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | .5000 | .3333 | 1. | 1. |
| snes | .5646 | .0601 | .5646 | .0601 | .5646 | .0601 | .5646 | .0601 | .5646 | .0601 | .5554 | .0796 | .5646 | .0601 |
| thal | .5000 | .2688 | .5556 | .1786 | .5000 | .1667 | .5556 | .1786 | .3333 | .1691 | .5556 | .1786 | .5556 | .1786 |
| thet | .5625 | .2419 | .4375 | .1167 | .5625 | .1636 | .2500 | .0784 | .5000 | .1333 | .5625 | .1500 | .5625 | .1636 |
| tron | .4767 | .0258 | .4767 | .0258 | .4767 | .0398 | .4767 | .0258 | .4651 | .0326 | .4680 | .0360 | .4767 | .0398 |
| vees | .6613 | .2381 | .7097 | .2075 | .5806 | .2328 | .5806 | .1875 | .7097 | .2075 | .7097 | .2115 | .7097 | .2115 |
| vogo | .3750 | .0779 | .3906 | .1735 | .4375 | .2467 | .4531 | .1979 | .4531 | .2286 | .4063 | .1769 | .4531 | .2286 |
| wast | .2188 | .1387 | .2188 | .0718 | .2500 | .2467 | .2188 | .2595 | .2188 | .0757 | .1875 | .1343 | .2500 | .2467 |
| avg. | .5053 | .2691 | .4782 | .2372 | .5359 | .2905 | .5271 | .3082 | .4981 | .2700 | 5356 | .3019 | .5899 | .3453 |

Table 9: Overview of classification test results (Micro F1 and Macro F1) for **RNN model - PT + FT**, with improvements over base model results underlined. The last column contains the highest results returned by the four temporal ranking methods.

| | Base Model | | Search Ranking | | Evidence Date | | Claim Date | | Claim Distance | | Evidence Distance | | Time-Aware Ranking | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro |
| abbc | .5370 | .4181 | .5370 | .2329 | .5370 | .2329 | .5370 | .2329 | .5370 | .2329 | .5370 | .2329 | .5370 | .2329 |
| afck | .1667 | .0987 | .1994 | .0903 | .1944 | .0903 | .1944 | .0903 | .1389 | .0663 | .2778 | .1273 | .2778 | .1273 |
| bove | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. |
| chct | .5250 | .2854 | .4000 | .1429 | .4500 | .1552 | .4250 | .2121 | .4000 | .2321 | .5250 | .2846 | .5250 | .2846 |
| clck | .1667 | .0952 | .1667 | .0952 | .1667 | .0952 | .1667 | .0952 | .3333 | .1667 | .5000 | .2222 | .5000 | .2222 |
| faan | .4091 | .1667 | .4545 | .2083 | .4545 | .2083 | .1818 | .1159 | .7273 | .5593 | .7273 | .5388 | .7273 | .5593 |
| faly | .8571 | .4762 | .0714 | .0333 | .8571 | .4521 | .2143 | .0882 | .8571 | .4521 | .8751 | .4521 | .8571 | .4521 |
| fani | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. |
| farg | .6923 | .1023 | .6923 | .1023 | .6923 | .1023 | .6923 | .1023 | .6923 | .1023 | .6538 | .0988 | .6923 | .1023 |
| goop | .8344 | .1516 | .8344 | .1516 | .8344 | .1516 | .8344 | .1516 | .8344 | .1516 | .8344 | .1516 | .8344 | .1516 |
| hoer | .5075 | .1863 | .3731 | .0776 | .3731 | .0776 | .3731 | .0776 | .3731 | .0776 | .4851 | .1671 | .4851 | .1671 |
| huca | .5000 | .2857 | .1667 | .0952 | .5000 | .2222 | .3333 | .1667 | .1667 | .0952 | .1667 | .0952 | .5000 | .2222 |
| mpws | .5000 | .2222 | .5000 | .2222 | .5000 | .2222 | .5000 | .2222 | .5000 | .2222 | .6250 | .4578 | .6250 | .4578 |
| obry | .4286 | .1500 | .4286 | .1500 | .4286 | .1500 | .4286 | .2483 | .4286 | .1500 | .4286 | .1500 | .4286 | .2483 |
| para | .2500 | .1435 | .1875 | .0526 | .2500 | .0667 | .2813 | .1530 | .2500 | .0667 | .1250 | .0370 | .2813 | .1530 |
| peck | .6667 | .4652 | .5833 | .2456 | .8333 | .5879 | .5833 | .2456 | .5833 | .2456 | .5833 | .2456 | .8333 | .5879 |
| pomt | .2482 | .1263 | .1938 | .0361 | .2464 | .1635 | .1732 | .0328 | .2536 | .1634 | .2545 | .1650 | .2545 | .1650 |
| pose | .4178 | .0982 | .4178 | .0982 | .4178 | .0982 | .4178 | .0982 | .4178 | .0982 | .4178 | .0982 | .4178 | .0982 |
| ranz | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. |
| snes | .5646 | .0601 | .5646 | .0601 | .5646 | .0601 | .5646 | .0601 | .5646 | .0601 | .5645 | .0601 | .5646 | .0601 |
| thal | .5556 | .0601 | .5556 | .1786 | .3333 | .1043 | .5556 | .1786 | .1111 | .0500 | .5000 | .2273 | .5556 | .1786 |
| thet | .5625 | .1440 | .5625 | .1440 | .5625 | .1440 | .5625 | .1440 | .5625 | .1500 | .5625 | .1440 | .5625 | .1440 |
| tron | .4767 | .0258 | .4767 | .0258 | .4767 | .0258 | .4767 | .0258 | .4767 | .0258 | .4767 | .0258 | .4767 | .0258 |
| vees | .7097 | .2075 | .7097 | .2075 | .7097 | .2115 | .7097 | .2075 | .7097 | .2075 | .7097 | .2075 | .7097 | .2115 |
| vogo | .5313 | .2199 | .1719 | .1567 | .5000 | .2146 | .1250 | .1676 | .5313 | .2199 | .5156 | .2098 | .5313 | .2199 |
| wast | .3438 | .2546 | .2188 | .0800 | .2188 | .0800 | .1250 | .1008 | .2188 | .0800 | .2188 | .0800 | .2188 | .0800 |
| avg. | .5558 | .2909 | .4782 | .2372 | .5359 | .2905 | .5271 | .3082 | .4981 | .2700 | 5356 | .3019 | .5921 | .3135 |

Table 10: Overview of classification test results (Micro F1 and Macro F1) for **DistilBERT model - PT + FT**, with improvements over base model results underlined. The last column contains the highest results returned by the four temporal ranking methods.

| | Base Model | | All Temporal Ranking Methods | | Evidence Date + Claim Date | | Claim Distance + Evidence Distance | | Best Temporal Ranking Method | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro |
| abbc | .3148 | .2782 | .5556 | .4083 | .5185 | .2276 | .5185 | .2276 | .5185 | .2276 |
| afck | .2222 | .1385 | .2500 | .1145 | .2778 | .1280 | .2222 | .1009 | .3333 | .1517 |
| bove | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. |
| chct | .5500 | .2967 | .5250 | .2877 | .4250 | .2297 | .4250 | .2310 | .4750 | .2613 |
| clck | .5000 | .2500 | .5000 | .2500 | .5000 | .2500 | .5000 | .2500 | .5000 | .2500 |
| faan | .5000 | .5532 | .5455 | .1609 | .4091 | .2799 | .3182 | .1728 | .4091 | .2918 |
| faly | .6429 | .2045 | .5000 | .2321 | .6429 | .6429 | .6429 | .2045 | .6429 | .2045 |
| fani | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. |
| farg | .6923 | .1023 | .6923 | .1023 | .6923 | .1023 | .6923 | .1023 | .6923 | .1023 |
| goop | .8344 | .1516 | .8344 | .1516 | .8311 | .1516 | .8013 | .1583 | .8344 | .1516 |
| hoer | .4104 | .2091 | .4179 | .2226 | .3955 | .1742 | .4104 | .1926 | .4552 | .2453 |
| huca | .5000 | .2857 | .5000 | .2857 | .5000 | .2857 | .5000 | .2857 | .6667 | .5333 |
| mpws | .8750 | .6316 | .8750 | .6316 | .8750 | .6316 | .8750 | .6316 | .8750 | .6316 |
| obry | .5714 | .3520 | .3571 | .1316 | .5714 | .3716 | .2143 | .0882 | .5714 | .3716 |
| para | .1875 | .1327 | .3750 | .1692 | .3750 | .1692 | .3750 | .1692 | .4375 | .2010 |
| peck | .5833 | .2593 | .6667 | .4082 | .6667 | .4082 | .6667 | .4082 | .9167 | .6316 |
| pomt | .2143 | .2089 | .1777 | .0579 | .1804 | .0597 | .1804 | .0593 | .1964 | .0637 |
| pose | .4178 | .0987 | .4178 | .1002 | .4178 | .0987 | .4041 | .0964 | .4247 | .2080 |
| ranz | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. |
| snes | .5646 | .0601 | .5631 | .0601 | .5646 | .0602 | .5646 | .0602 | .5646 | .0601 |
| thal | .5556 | .2756 | .3333 | .1143 | .2778 | .1250 | .5556 | .3405 | .5556 | .3600 |
| thet | .3750 | .2593 | .6250 | .3556 | .6250 | .3556 | .6250 | .3007 | .6250 | .3556 |
| tron | .4767 | .0258 | .4767 | .0258 | .4767 | .0258 | .4593 | .0321 | .4767 | .0366 |
| vees | .7097 | .2075 | .7097 | .2075 | .7097 | .2075 | .7097 | .2075 | .7258 | .3021 |
| vogo | .5000 | .2056 | .4375 | .1822 | .4688 | .1883 | .5156 | .2491 | .5469 | .2463 |
| wast | .1563 | .0935 | .2188 | .0757 | .2188 | .0737 | .0938 | .0400 | .3438 | .2785 |
| avg. | .5521 | .3185 | .5598 | .2975 | .5623 | .3172 | .5488 | .2926 | .6072 | .3525 |

Table 11: Overview of classification test results (Micro F1 and Macro F1) for **BiLSTM model - Only FT**, with improvements over base model results underlined. We present three optimization approaches, in addition to the approach used in the main paper. Instead of optimizing the evidence ranking based on single temporal ranking methods, all four temporal ranking methods (evidence-based date ranking, claim-based date ranking, claim-based distance and evidence-based distance; **All Temporal Ranking Methods**) are considered, and the evidence ranking is optimized based on the sum of the four ranking losses. We also explore a combination of the date-based ranking losses (**Evidence Date + Claim Date**), and the distance-based ranking losses (**Claim Distance + Evidence Distance**). The lower results motivate our choice for the optimization approach presented in the main paper (i.e. optimization based on single temporal ranking methods and selecting the best method for each domain; **Best Temporal Ranking Method**).