STANDER: An Expert-Annotated Dataset for News Stance Detection and Evidence Retrieval

Costanza Conforti[1], **Jakob Berndt**[2], **Mohammad Taher Pilehvar**[1,3],
Chryssi Giannitsarou[2], **Flavio Toxvaerd**[2], **Nigel Collier**[1]

[1] Language Technology Lab, University of Cambridge
[2] Faculty of Economics, University of Cambridge
[3] Tehran Institute for Advanced Studies, Iran
{cc918,jb2088}@cam.ac.uk

Abstract

We present a new challenging news dataset that targets both stance detection (SD) and fine-grained evidence retrieval (ER). With its 3,291 expert-annotated articles, the dataset constitutes a high-quality benchmark for future research in SD and multi-task learning. We provide a detailed description of the corpus collection methodology and carry out an extensive analysis on the sources of disagreement between annotators, observing a correlation between their disagreement and the diffusion of uncertainty around a target in the real world. Our experiments show that the dataset poses a strong challenge to recent state-of-the-art models. Notably, our dataset aligns with an existing Twitter SD dataset: their union thus addresses a key shortcoming of previous work, by providing the first dedicated resource to study multi-genre SD as well as the interplay of signals from social media and news sources in rumour verification.

1 Introduction

Starting from early work by Agrawal et al. (2003), Stance Detection (SD) has gained increasing interest from the research community (Zubiaga et al., 2018a). Recent work in SD has mostly focused on modeling user-generated data (Mohammad et al., 2017; Küçük and Can, 2020). However, SD on complex and articulated texts, such as news articles, has been considerably less studied, mainly due to the scarcity of published datasets (Pomerleau and Rao, 2017; Hanselowski et al., 2019). Moreover, research on user-generated SD and news SD has proceeded on parallel and independent tracks, neglecting the deep mutual influence that exists between social media and news sources (Canter, 2015; Kostkova et al., 2017).

In this paper, we seek to fill this gap, introducing STANDER (STANce Detection & Evidence Retrieval), a new expert-annotated dataset which is labeled for both news SD and fine-grained ER. STANDER collects news articles in English from high-reputation sources which discuss four recent mergers and acquisitions (M&A) operations between major healthcare companies in the US (Table 1). The term *M&A* refers to the process by which the ownership of a company (the *target*) is transferred to another company (the *buyer*). An M&A process (*merger*) ranges from informal talks between the companies to the closing of the deal; high secrecy is involved and discussions are usually not publicly disclosed during its early stages (Bruner and Perella, 2004). Thus, the analysis of the evolution of opinions and concerns about a potential merger is a process similar to rumor verification (Zubiaga et al., 2018b).

Notably, the news articles in STANDER discuss the same targets as in WT–WT (Conforti et al., 2020), a large Twitter SD dataset: thus, their union provides aligned signals from both authoritative (articles) and user-generated (tweets) sources, constituting the first resource of this kind for SD.

In this paper, we make the following contributions:

(1) We construct STANDER, a large expert-annotated news dataset [1] labeled for SD and fine-grained ER. To our knowledge, it is the first news SD dataset to provide evidence snippets, along with their *exact location* in the corresponding article.

(2) We provide detailed statistics of our data, as well as the first diachronic analysis of the sources of disagreement among annotators in a SD paper, shedding light on the potential correlation between uncertainty in the world and increased ambiguity in journalistic prose. This suggests that considering

[1] https://github.com/cambridge-wtwt/emnlp2020-stander-news
Data is released according to Factiva (https://library.princeton.edu/resource/3791) and the University of Cambridge's data policy.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4086–4101
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

Merger	Buyer	Target	Outcome
AET_HUM	Aetna	Humana	rejected
ANTM_CI	Anthem	Cigna	rejected
CI_ESRX	Cigna	Express Scripts	succeeded
CSV_AET	CVS	Aetna	succeeded

Table 1: Mergers considered in this work. Note that two companies appear both as *Buyer* and as *Target*.

SD in a controlled domain, such as mergers, could allow model builders to develop deeper insights into the factors influencing model performance.
(3) We report results obtained for several state-of-the-art models on our dataset, and show that STANDER constitutes a challenging benchmark for future research in SD, ER and multi-task learning.
(4) We provide a correlation analysis of the articles from STANDER and the tweets from WT–WT, observing a moderately strong correlation. While the interplay between social media and news sources has been widely studied in other research fields, such as journalism studies (Johnson et al., 2018; Orellana-Rodriguez and Keane, 2018), very little work exists in computer science (Dredze et al., 2016), and notably, none considering SD.

2 Background

The Task. SD is the task of automatically identifying the opinion expressed in a text with respect to a target (Mohammad et al., 2017). Note that SD constitutes a related, but different task than both *sentiment analysis* and *textual entailment*. The first considers the emotions conveyed in a text (Alhothali and Hoey, 2015; Tang et al., 2016), while in the second, the goal is to predict whether a logical implication exists between two sentences (Bowman et al., 2015). Consider the following example:

Target: *Aetna will merge with Humana*
Text: *Aetna & Humana CEOs met again to talk about deal, can't stand those bla-bla people!!!*

The text's *sentiment* is *negative*, as the author is complaining about the meeting; concerning *entailment*, it is *positive*: the target entails the text because, in order to merge, two companies need to discuss the deal; finally, its *stance* is commenting, as it is just talking about the merger, without expressing the orientation that it will happen (or not).

SD as a Sub-Task. SD is often integrated into *rumor verification* (Zubiaga et al., 2018b), as testified by popular shared tasks (Derczynski et al., 2017; Gorrell et al., 2018). Starting from Vlachos and Riedel (2014), SD has been identified as a key step in *fake news detection* (Lillie and Middelboe, 2019) and *automated fact-checking* (Popat et al., 2017; Thorne and Vlachos, 2018; Baly et al., 2018): in this context, *textual entailment* is sometimes preferred to SD as the penultimate sub-step before verification (Thorne et al., 2018).

Twitter SD. Traditionally, research on SD focused on user-generated data, such as blogs and commenting sections on websites (Skeppstedt et al., 2017; Hercig et al., 2017), apps (Vamvas and Sennrich, 2020), and Facebook posts (Klenner et al., 2017); above all, mainly due to the handiness of its API, Twitter was used as a data source (Mohammad et al., 2016; Zubiaga et al., 2016; Inkpen et al., 2017; Aker et al., 2017; Conforti et al., 2020).

News SD. At the time of writing, a very small number of SD datasets collecting news have been released, usually building on platforms originally developed by professional journalists, like Emergent (Ferreira and Vlachos, 2016) or Snopes (Hanselowski et al., 2019). Note that in Twitter SD, the task consists of defining the stance of a tweet with respect to a short target (usually a named entity like *Hillary Clinton* (Inkpen et al., 2017), or a concept like *feminism* (Mohammad et al., 2016)); in news SD, on the contrary, the input article is much longer than a tweet, and the target is a complete sentence (Hanselowski et al., 2018a).

Comparison with corpora for News SD. EMERGENT (Ferreira and Vlachos, 2016) constitutes the first released corpus for news SD: it collects 300 targets and 2,595 articles (with an average of 8.6 articles/target), labeled using a 3-class classification schema. For the first edition of the *Fake News Challenge* (Pomerleau and Rao, 2017), it was enriched with randomly generated *unrelated* samples. Neither of the two corpora is annotated with evidences.

To our knowledge, the only other news dataset to be annotated for both SD and ER is that of Hanselowski et al. (2019), which annotates *fact-checking instances* from the debunking website Snopes[2]. Our work differs in a number of aspects:

- *Statistics.* While Snopes is larger in size, it provides relatively few samples per target (14,296 samples and 6,422 targets, with an average of 2.22 articles/target); STANDER, in contrast, collects 3,291 articles on 4 targets, with an average

[2] https://www.snopes.com/about-snopes/

of more than 800 articles/target.

- *Annotators.* Snopes is annotated by crowdsourcing, we employ domain-expert annotators.
- *Evidence Annotations.* Snopes provides entire sentences as evidence; importantly, STANDER is the first to provide the exact start and end indices of evidence snippets *inside* the sentences (Figure 4): this will enable future research on more fine-grained evidence extraction.
- *Multi-Genre.* At the time of writing, almost all released SD corpora collect data from one genre, with a prevalence of user-generated content. Snopes constitutes the only exception, as some of the collected documents (11%) come from Facebook or Twitter. Note, however, that they do not provide aligned signals from news and user-generated sources for *all* considered targets, but only for a limited portion of them. In contrast, our news dataset is the first to *completely* align with an existing resource for Twitter SD, providing a relevant amount of samples from two genres for all considered targets (Section 6). This will open a number of interesting research directions: while adversarial domain adaptation – using data of the same news genre, but from another domain – proved to be useful for news SD (Xu et al., 2019), the impact of considering data of the same domain but from another genre has never been studied in SD.

3 Building the Dataset

In this section, we report on data collection and annotation, and provide a detailed analysis of the findings from the annotation process.

3.1 Data Retrieval

We consider four recent mergers involving US companies in the healthcare industry (Table 1). To retrieve news articles related to the mergers, we used Factiva (Johal, 2009), a database by Dow Jones which collects more than 32,000 general and finance-specific sources, including newspapers, journals and magazines.

For each merger, we searched for the involved companies and selected articles in English tagged as *Acquisitions/Mergers/Shareholdings*. We retrieved articles from one month before the first contact of the firms up to one month after any final decision on the merger. Refer to Appendix A for details on the crawl settings and the crawling timeline.

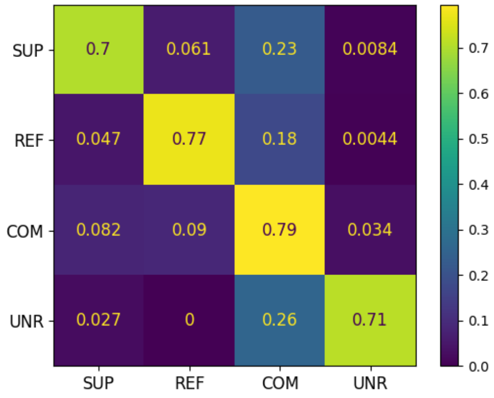


Figure 1: Inter-rater agreement (normalized).

3.2 Annotation Guidelines

The annotation process was initiated by a pilot, after which the annotation guidelines were written in close collaboration with three domain experts. Extracts from the annotation guidelines are reported in Appendix A.

Stance Annotation. Following Pomerleau and Rao (2017), we consider four stance labels:

1. *Support*: the article is voicing confidence that the two companies will merge.
2. *Refute*: the article is voicing doubts that the two companies will merge.
3. *Comment*: the article is talking about the merger, neither directly supporting, nor refuting it.
4. *Unrelated*: the article is unrelated to the merger. Note that the article might be talking about one or both the considered companies, but without discussing their merger.

Evidence Annotation. In addition to the stance label, annotators were asked to select the text snippets or sentences from the article which were determinant for them to classify its stance, which we refer to as *evidence* (Thorne and Vlachos, 2018).

3.3 Data Annotation

In line with previous work on news SD (Vlachos and Riedel, 2014; Ferreira and Vlachos, 2016), in which data was labeled by professional journalists, we rely on domain experts for annotation. Specifically, we provided articles to eight economists[3] in batches and asked them to annotate no more than 100 articles per day[4]; the annotation process lasted 4 months. Each article was independently labeled

[3] Six PhD students and two lecturers in Economics (Faculty of Economics, University of Cambridge)

[4] Reported annotation speed is ∼55 articles/hour; annotators were asked not to spend more than 2 hrs/day on the task.

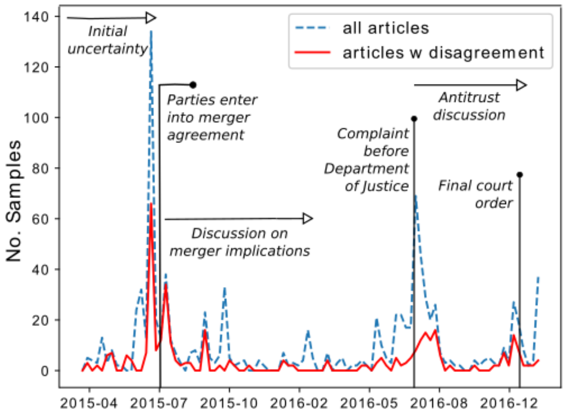


Figure 2: Timestamp of publication of articles whose stance annotators disagreed on (AET_HUM merger).

by 2 to 4 annotators.

To aggregate stance labels, we used majority voting. For evidence snippets, we merged the provided snippets to obtain a list of selected evidences; a further annotator, who did not take part in the first phase, manually checked the overlapping snippets.

3.4 Analysis of Annotators' Disagreement

The most common source of disagreement between annotators is on *support*/*comment* (Figure 1): note that, sometimes, the given stance depends on subtle nuances in the article's argumentative structure and it is therefore somehow subjective; such samples are difficult to discriminate for ML systems as well (Riedel et al., 2017). With respect to datasets with randomly generated *unrelated* samples (Pomerleau and Rao, 2017; Hanselowski et al., 2018b), we report a slightly higher *unrelated/comment* disagreement between annotators, which reflects the higher complexity of the task in our setting.

To further understand the sources of disagreement between annotators, we perform a diachronic analysis of the samples which received different labels and their time of publication. As shown in Figure 2, a correlation exists between some relevant events (such as the first joint press release) and the number of articles published. However, a higher volume of articles does not always correlate with higher disagreement rates between annotators: interestingly, it seems that some events (such as the merger agreement) spread more uncertainty around the merger than others (such as the start of the antitrust trial). This uncertainty is transmitted to the press, resulting in journalists writing speculative articles: such articles seem to be more prone to the reader's subjective biases, eventually producing a higher inter-annotator disagreement.

The interplay of different layers of uncertainty until the resolution of the event (i.e. confirmation of merger talks or the complaint before the DOJ) makes our domain choice particularly insightful for model builders.

3.5 Quality Assessment

To assess the quality of our dataset, we asked a domain expert to annotate a random 10% of the samples, which are used as an upper bound for evaluation. First, she received targets together with the gold evidence snippets selected in the first annotation round; in a second phase, the same annotator received the complete articles and was asked to re-annotate the samples. In the former setting and similar to Hanselowski et al. (2019), we wanted to assess whether the selected evidence snippets alone are sufficient to provide the correct stance: the Cohen's κ between those labels and the gold is 75.2, which is *substantial* (Cohen, 1960) and reflects the good quality of the extracted snippets. Cohen's κ obtained when considering the entire article texts is 59.5 (*moderate*).

This drop testifies that: (1) SD on long, unstructured texts is complex and more prone to subjective biases than SD on evidence snippets; interestingly, a similar low inter-annotator agreement (Fleiss' κ of 0.55, (Fleiss, 1971)) has been observed also for the *related* news articles in the *Fake News Challenge* dataset (Hanselowski et al., 2018a), which does not contain annotation of evidences; unfortunately, Hanselowski et al. (2019) do not report on the agreement considering the entire sample texts; (2) therefore, providing evidence annotation is fundamental to building a reliable dataset that can be used to train supervised stance classifiers.

4 Corpus Analysis

4.1 Desiderata and Challenges

Notably, STANDER satisfies all four desired properties outlined in Mohammad et al. (2017):

1. Topics should be *commonly understood* by a wide number of people. We consider some of the major US healthcare providers, with which almost everyone has interacted at different levels (insurers, pharmacy chains, ...): thus, not only finance experts (example ***(a)*** in Table 3) and local sources ***(b)***, but also politicians ***(c)***, physicians ***(d)***, policymakers ***(e)*** and the general

public are interested in their outcome, resulting in a dataset which collects different registers.

2. The topics convey *different opinions*, producing a significant amount of data *for all stance labels*. The considered mergers are controversial, because their outcome might change the US healthcare landscape; moreover, as they happened during the change from the Obama to the Trump administration, with the introduction and partial rollback of Obamacare, there is considerable interference with politics ***(f)***.
3. The dataset contains *indirect references* to the targets, as when the involved companies are not explicitly mentioned: for example, given a merger between A and B, if a source states that A is in talk with C, this implicitly undermines the likelihood of the A-B merger to happen ***(g)***.
4. The dataset contains samples where the *target of opinion is different*. This is the case of articles that discuss about one or both companies, without taking a stance on their merger ***(h)***.

(a)	Aetna shares rose 1.3% premarket after climbing 10% just before the market closed Thursday following a Wall Street Journal report that CVS Health is in talks to buy the insurer [...]
(b)	Commercial real estate office experts [...] agree that the [...] planned acquisition [...] by Connecticut-based Aetna could have a significant negative impact on Humana's major office footprint in [...] Louisville.
(c)	Rep. EG and state Senator DH are asking the state insurance commissioner to receive a guarantee of zero job reductions within Humana's state locations if its proposed merger with Aetna proceeds.
(d)	February survey of physicians [...] found that 28% are so concerned by the potential merger that they would be likely to retire early [...] said the CMS *[Colorado Medical Society]* president.
(e)	Justice Department attorneys, arguing before a federal court judge on Monday, contended that Aetna's (AET) planned acquisition of Humana (HUM) violated antitrust law [...]
(f)	The second thing [...] Aetna must persuade Bates of is that [...] the merger won't harm individuals who receive their health coverage through Obamacare.
(g)	Meanwhile, UnitedHealth is said to be interested in scooping up Aetna.
(h)	Aetna, with eye on regulators, sells Medicare drug business to WellCare [...]
(i)	Besides the possible Anthem deal, Humana is considering a sale, possibly to Cigna or Aetna.
(j)	Even amid the Anthem talks [...] Cigna continues to examine a potential purchase of Louisville-based Humana Inc., people familiar with the matter said.

Figure 3: Example snippets from STANDER.

Stance	**support**
Target	AET_HUM
Title	Aetna to Acquire Humana for $37 Billion
Body	Aetna (NYSE: AET) and Humana Inc. (NYSE: HUM) today announced that they have entered into a definitive agreement under which Aetna will acquire all outstanding shares of Humana for a combination of cash and stock [...]
Stance	**comment**
Target	CI_ESRX
Title	Cigna's Purchase of Expres Scripts Unlikely to Affect Workers' Comp
Body	According to Joe Paduda, principal of Health Strategy Associates, these kinds of purchases don't really impact worker's comp stakeholders. "Health plans and PBMs are merging to better control care delivery and cost," [...]
Stance	**refute**
Target	CVS_AET
Title	Health Care up Amid Deal Activity
Body	A federal judge voiced concern about the Justice Department's decision to allow CVS Health's nearly USD 70 billion acquisition of Aetna, and said he may require CVS to hold Aetna's assets separately while he considers the settlement between the companies and the government [...]

Figure 4: A *supporting*, a *commenting* and a *refuting* sample from STANDER (evidence snippets underlined).

Moreover, as the mergers happened simultaneously, there is considerable interplay between companies; a successful classifier thus requires the modeling of the deep relationship between the target merger and the article, not just simple keyword matching ***(i, j)***.

In addition, the task is challenging as the underlying argumentative structure is needed in order to correctly classify the article. Considering the *support* example in Figure 4, both the title and the body contain the same information. In the *comment* example, the evidence is in the title, while the body provides additional information. In the *refute* example the evidence is in body while the title does not contain information regarding the stance.

These characteristics contribute to making STANDER a challenging benchmark for news SD.

4.2 Corpus Statistics

Dataset Statistics. The final dataset collects 3,291 labeled news articles from heterogeneous news outlets (Figure 5): while finance-specific publications constitute the majority of the most frequent sources, the corpus also contains many general newspapers (such as *Reuters News* or *The New York Times*) as well as local journals (such as the *Louisville Business First*). News articles present an asym-

avg articles/source	11.2
avg paragraph/body	24.4
avg tokens/title	12.8
avg tokens/paragraph	24.3
avg tokens/body	592.1
avg evidence/article	1.9
avg tokens/evidence	22.7

Table 2: Relevant statistics from the STANDER corpus.

	support		*refute*		*comment*		*unrelated*		Total
	#samples	%	#samples	%	#samples	%	#samples	%	
CVS_AET	372	46.4	104	12.9	294	36.7	31	3.8	831
CI_ESRX	207	59.8	64	18.4	70	20.2	5	1.4	376
ANTM_CI	367	31.4	537	46.0	248	21.2	14	1.2	1,199
AET_HUM	463	47.3	313	32.0	197	20.1	5	0.5	1,009
Total	1409		1018		809		55		

Table 3: Label distribution across different mergers in the STANDER corpus (refer to Table 1).

metric and hierarchical structure: they are formed by a concise and short *title* and a (usually) long *body*, which in turn is composed of ordered *paragraphs* (Table 2). Note that, while articles might be very long (Figure 8), evidences are usually located in the title or in the first few paragraphs (Figure 6). This is in line with the *inverted pyramid* (Scanlan, 2000; Pöttker, 2003) or *summary news lead* (Errico et al., 1997) style – widely adopted in modern journalistic prose – in which the most relevant information is concentrated at the beginning of the article.

Label Distribution. A clear correlation can be observed between the merger's outcome (blocked/succeeded) and the relative proportion of *supporting* and *refuting* samples (Table 3). Contrary to many popular SD datasets (Derczynski et al., 2017; Pomerleau and Rao, 2017; Hanselowski et al., 2018b), the *related* labels present a relatively balanced distribution: this is in line with property (2) (Section 4.1); however, in contrast to Mohammad et al. (2017), who employed query keywords to "force" it, such a balanced distribution arose naturally from our data.

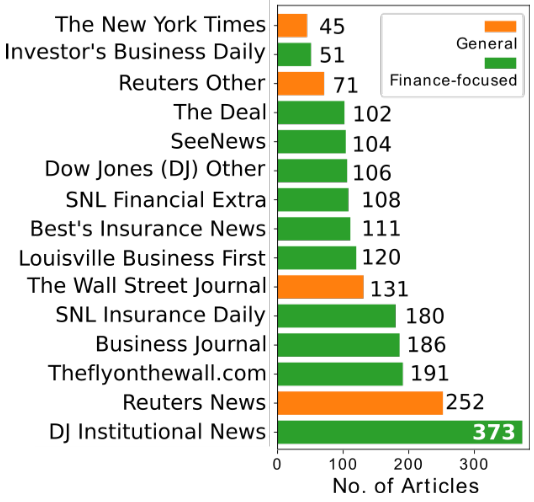


Figure 5: 15 most frequent news sources in the dataset.

5 Baselines and Discussion

This section provides results for a number of recent techniques. While more complex models could possibly achieve better results, our aim was to set baselines for our dataset with a number of strong models. Detailed description of the experimental setting is provided in Appendix B and C for replication.

5.1 Experiments

Models. We consider two dummy baselines – a *random* and a *majority vote* baseline – and, following Hanselowski et al. (2019), three neural baselines: *BertEmb*, an MLP leveraging sentence-BERT embeddings (Reimers and Gurevych, 2019); *UseEmb*, an MLP leveraging Universal Sentence Encoder's sentence embeddings (Cer et al., 2018); and a *BiLSTM* over Glove embeddings (Pennington et al., 2014). As upper bound, we consider the performance of a domain expert against the aggregated gold data (see Section 3, *Quality Assessment*, for further details).

Experimental Setting. We first test the models' ability to perform SD given the *correct set* of sentences which contain an evidence snippet (*SD in isolation*). Secondly, we consider both SD and ER: while the tasks could be approached with a

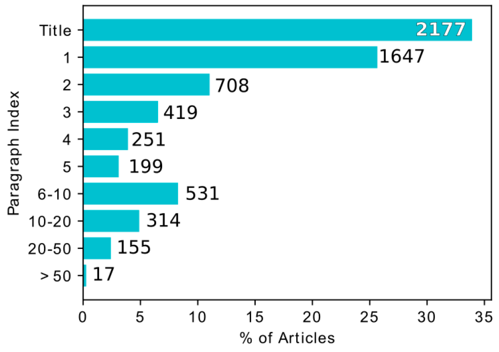


Figure 6: Distribution of the evidence locations.

		Stance Detection: F_1 across mergers				**avg Stance Detection**			**avg Evidence Retrieval**	
	Model	CVS_AET	CI_ESRX	ANTM_CI	AET_HUM	$avgP$	$avgR$	$avgF_1$	$avgP$@5	$avgR$@5
3 classes (only *related*)	*Random Base*	25.0	24.3	26.0	24.5	25.3	25.3	25.0	15.3	08.2
	Majority Base	15.2	15.2	15.2	15.2	10.9	25.0	15.2	58.3	46.1
	BiLSTM	44.1	67.2	46.5	60.2	64.0	56.6	52.7	–	–
	UseEmb	44.0	59.6	55.2	56.4	59.3	55.5	53.3	–	–
	BertEmb	47.4	55.6	50.1	59.4	56.6	56.6	52.8	–	–
	BiLSTM (+ER)	46.4	60.8	**56.5**	55.5	60.9	57.0	54.2	54.6	57.7
	UseEmb (+ER)	47.8	54.4	48.3	58.1	57.6	54.8	51.8	**56.4**	**58.5**
	BertEmb (+ER)	**54.2**	**70.0**	52.8	**60.3**	**63.5**	**59.6**	**57.3**	54.1	53.8
4 classes (+*unrelated*)	*Random Base*	17.5	17.4	17.1	16.5	19.6	19.8	17.1	15.1	07.9
	Majority Base	12.0	12.0	12.0	12.0	8.6	20.0	12.0	58.0	46.7
	BiLSTM	38.8	42.9	42.8	43.8	46.2	43.9	42.1	–	–
	UseEmb	35.8	33.2	39.7	43.3	44.0	40.6	39.1	–	–
	BertEmb	42.5	33.2	46.4	43.9	50.5	45.6	43.2	–	–
	BiLSTM (+ER)	40.2	35.1	41.1	**43.8**	44.4	42.4	41.0	55.4	57.1
	UseEmb (+ER)	31.8	36.1	35.5	43.0	41.6	39.6	36.9	**56.9**	**57.4**
	BertEmb (+ER)	**47.3**	**53.6**	**45.3**	41.8	**51.7**	**47.8**	**45.7**	54.2	55.0
	Upper Bound	72.3	85.2	64.2	75.6	72.9	73.2	71.9	–	–

Table 4: Results of baseline experiments on Stance Detection (SD), both in isolation and jointly with Evidence Retrieval (+ER). We consider both SD on the completed stance tagset (*4 classes*) and on only *related* classes (*3 classes*; note that in this case the sample distribution is balanced). Macro F_1 refers to testing on the target merger while training on the other three. Performances over all operations are averaged weighting by merger's size.

pipelined strategy (as in Thorne et al. (2018)), we follow a multi-task training approach, which has proven to be more effective (Yin and Roth, 2018). When jointly training, we employ a simple ER strategy, by taking the title and the first 4 paragraphs from each article as candidates.

We train in a cross-target setting (train on three mergers, test on the fourth), and consider two training settings: first, we select only *related* samples, which present a balanced distribution (Table 3); then, we consider all stances: this is more difficult because *unrelated* samples are very infrequent, resulting in a skewed distribution as in RumourEval (Derczynski et al., 2017; Gorrell et al., 2018). To account for performance fluctuations (Reimers and Gurevych, 2017), we run 5 simulations for each model and take the average of the results. We leave the identification of the evidence's indices in the sentences, as well as the usage of more sophisticated ER methods, to future work.

Evaluation. We follow recent work (Thorne et al., 2018; Hanselowski et al., 2018b, 2019) and consider macro-averaged precision, recall and F_1 for SD, and precision and recall on the 5 selected evidence candidates (P@5 and R@5) for ER.

5.2 Results and Error Analysis

Results of the experiments are reported in Table 4. As expected, we observe a drop in performance when considering only *related* vs all classes. While all considered models obtain significant gains over the two dummy baselines, the BertEmb model – as observed also in Hanselowski et al. (2019) – obtains the best results overall for SD. Note, however, the wide gap between BertEmb performance and the upper bound, which confirms the difficulty of our dataset. Considering ER results, we observe a smaller gap in performance between models, with UseEmb obtaining the best results overall.

Interestingly, we observe a gain in stance classification when BertEmb is jointly trained to perform both SD and ER: this seems to indicate that, by learning to classify whether an input sentence constitutes an evidence snippet or not, the system is indirectly gathering knowledge which is also useful to solve the SD task. An error analysis of BertEmb's predictions shows that most mis-classifications happen between the *comment* the *support* labels: this is in line with findings from both previous work (Riedel et al., 2017) and the analysis of the inter-annotator agreement (Section 3). A relatively high number of *comment* samples are also mis-classified as *refute*: note that – while in news SD corpora *refuting* samples coming from popular newspapers can sometimes be easily spotted by the presence of words such as *fake*, *hoax*, or similar – STANDER contains articles from high-reputation sources, which usually do not use sensationalist language.

6 Integrating News and Twitter Signal

As outlined in the Introduction, STANDER contains the same targets as the Twitter SD WT–WT corpus (Conforti et al., 2020). The union of both corpora thus provides a great opportunity for studying the interplay between authoritative and user-generated signals: the first refers to long and articulated texts written by professional journalists, while the second refers to a very abundant but potentially noisy stream of posts, which are published without any editorial review. While a detailed time series analysis (Lim and Tucker, 2019) is beyond the scope of this paper, we provide a first data description and a correlation analysis, which show the potential of the obtained aligned corpus and the challenges it may pose to future research.

6.1 Statistical Analysis.

The relative frequency of samples between mergers is similar in both the news and the Twitter signals (Figure 7), with CI_ESRX being the less popular target (refer to Conforti et al. (2020) for a detailed analysis of the WT–WT corpus). The same holds true for the relative distribution of *related* labels, with *refuting* samples being more frequent in the case of blocked mergers.

However, there are a number of differences between the two signals: notably, the Twitter signal presents a high number of noisy *unrelated* samples, which is not surprising when dealing with user-generated data (Zubiaga et al., 2015); we also observe a higher proportion of *commenting* samples, which has often been observed in financial microbloggings (Žnidaršič et al., 2018). On the contrary, the news signal is cleaner, but around one order of magnitude less abundant (Figure 7). Apart from this asymmetry in label distribution, a further asymmetry in length can be observed between the corpora: tweets tend to be short and compact, while pieces of news are long and articulated (Figure 8), thus posing interesting challenges for future work on multi-genre SD.

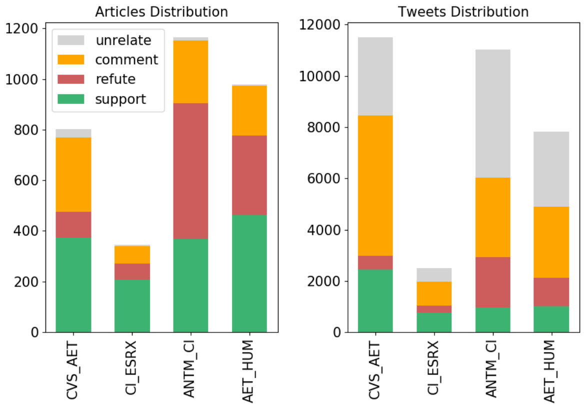


Figure 7: Label distribution across the considered mergers in the news (left) and Twitter dataset (right).

6.2 Signal Correlation

A diachronic analysis of the volume of tweets and articles discussing CVS_AET (Figure 9) shows a relatively similar distribution between the two signals, with some notable differences. While the Twitter signal presents some constant but minor activity from the very beginning of the process, the news signal remains completely silent until the companies' views are reported by a major news outlet. For some of the mergers, we even observe a notable spike in the Twitter activity before, but close to the first merger's mention in the press (see the analysis of the ANTM_CI merger in Appendix D). This is in line with studies on the usage of social media, especially Twitter, as sources of information for journalists (Van Leuven and Deprez, 2017; Rony et al., 2018; Johnson, 2019).

As reported in Table 5, the two signals exert moderate levels of correlation, which is further increased when only considering *related* tweets. This follows from the observation that large spikes in both the Twitter and the news signal are around dates of milestones within the merger process (Bruner and Perella, 2004; Piesse et al., 2013) and that many of the *unrelated* tweets occur before the first news article is published, when no activity is

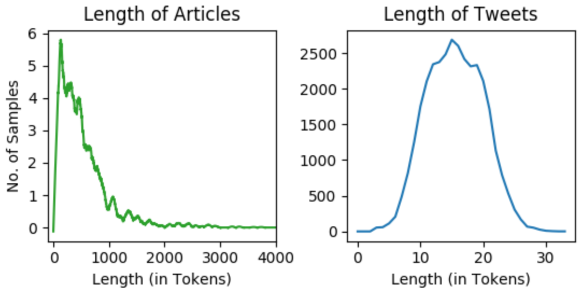


Figure 8: Asymmetry in length in STANDER (left) and in the WT–WT Twitter SD corpus (right).

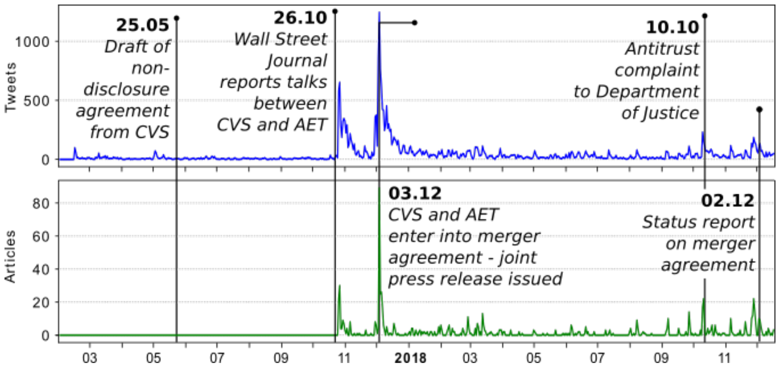


Figure 9: Volume of tweets and news over time for the CVS_AET merger (for further visualizations see Appendix D).

Merger	all stances	only *related*	obs. (days)
AET_HUM	0.5527 (0.0244)	0.6116 (0.0220)	815
ANTM_CI	0.4793 (0.0230)	0.5535 (0.0207)	1,124
CI_ESRX	0.4878 (0.0350)	0.5398 (0.0326)	475
CSV_AET	0.6260 (0.0235)	0.6470 (0.0225)	671

Table 5: Spearman correlation and approx. standard errors between the twitter and the news signals.

registered for the news signal (see Appendix D for further discussion).

7 Conclusions and Future Work

We presented STANDER, a new expert-annotated resource for news SD and ER. We provided a detailed description of the annotation process and corpus statistics, as well as of the findings from the annotation process. Our experiments with a set of strong models indicated a consistent (up to 30%) performance gap between SoA and human upper bound: this proves that our corpus constitutes a strong challenge and leaves plenty of room for future work on news SD, ER, domain adaptation and multi-task training.

Moreover, our corpus enables future research in a number of new areas, including: fine-grained ER for news SD – where the goal is not only to retrieve evidence snippets, but also their exact location in the text – which goes in the direction of improving interpretability of a model's predictions; and multi-genre SD – due to the fact that our corpus aligns with an existing resource for Twitter SD – which would open new interesting scenarios in the wider field of rumour verification.

Acknowledgments

We thank the anonymous reviewers of this paper for their efforts and for the constructive comments and suggestions. We gratefully acknowledge funding from the Keynes Fund, University of Cambridge (grant no. JHOQ). CC is grateful to NERC DREAM CDT (grant no. 1945246) for partially funding this work. CG and FT are thankful to the Cambridge Endowment for Research in Finance (CERF).

References

Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of the Twelfth International World Wide Web Conference, WWW 2003, Budapest, Hungary, May 20-24, 2003*, pages 529–535. ACM.

Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, Anna Kolliakou, Rob Procter, and Maria Liakata. 2017. Stance classification in out-of-domain rumours: A case study around mental health disorders. In *Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II*, volume 10540 of *Lecture Notes in Computer Science*, pages 53–64. Springer.

Areej Alhothali and Jesse Hoey. 2015. Good news or bad news: Using affect control theory to analyze readers' reaction towards news articles. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1548–1558, Denver, Colorado. Association for Computational Linguistics.

Ramy Baly, Mitra Mohtarami, James R. Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 21–27. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Robert F Bruner and Joseph R Perella. 2004. *Applied mergers and acquisitions*, volume 173. John Wiley & Sons.

Lily Canter. 2015. Personalised tweeting: The emerging practices of journalists on twitter. *Digital Journalism*, 3(6):888–907.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Costanza Conforti, Jakob Berndt, M. Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 69–76.

Mark Dredze, Prabhanjan Kambadur, Gary Kazantsev, Gideon Mann, and Miles Osborne. 2016. How twitter is changing the nature of financial news discovery. In *Proceedings of the Second International Workshop on Data Science for Macro-Modeling, DSMM@SIGMOD 2016, San Francisco, CA, USA, June 26 - July 1, 2016*, pages 2:1–2:5. ACM.

Marcus Errico, J April, A Asch, L Khalfani, M Smith, and X Ybarra. 1997. The evolution of the summary news lead. *Media History Monographs*, 1(1).

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. Rumoureval 2019: Determining rumour veracity and support for rumours. *CoRR*, abs/1809.06683.

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, Avinesh P. V. S., Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018b. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1859–1874. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Tomáš Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. 2017. Detecting stance in czech news commentaries. In *Proceedings of the 17th ITAT: Slovenskoceský NLP workshop (SloNLP 2017)*, volume 1885, pages 176–180.

Diana Inkpen, Xiaodan Zhu, and Parinaz Sobhani. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 551–557. Association for Computational Linguistics.

Rajiv Johal. 2009. Factiva: Gateway to business information. *Journal of Business & Finance Librarianship*, 15(1):60–64.

Michiel Johnson. 2019. *Sourcing Twitter: a multi-methodological study on the role of Twitter in economic journalism*. Ph.D. thesis, University of Antwerp.

Michiel Johnson, Steve Paulussen, and Peter Van Aelst. 2018. Much ado about nothing? the low importance of twitter as a sourcing tool for economic journalists. *Digital Journalism*, 6(7):869–888.

Manfred Klenner, Don Tuggener, and Simon Clematide. 2017. Stance detection in facebook posts of a german right-wing party. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, LSDSem@EACL 2017, Valencia, Spain, April 3, 2017*, pages 31–40. Association for Computational Linguistics.

Patty Kostkova, Vino Mano, Heidi J. Larson, and William S. Schulz. 2017. Who is spreading rumours about vaccines? influential user impact modelling in social networks. In *Proceedings of the 2017 International Conference on Digital Health*, DH '17, page 48–52, New York, NY, USA. Association for Computing Machinery.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Anders Edelbo Lillie and Emil Refsgaard Middelboe. 2019. Fake news detection using stance classification: A survey. *CoRR*, abs/1907.00181.

Sung Hoon Lim and Conrad S. Tucker. 2019. Mining twitter data for causal links between tweets and real-world outcomes. *Expert Syst. Appl. X*, 3:100007.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Techn.*, 17(3):26:1–26:23.

Claudia Orellana-Rodriguez and Mark T. Keane. 2018. Attention to news and its dissemination on twitter: A survey. *Computer Science Review*, 29:74 – 94.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Jenifer Piesse, Cheng-Few Lee, Lin Lin, and Hsien-Chang Kuo. 2013. Merger and acquisition: Definitions, motives, and market responses. *Encyclopedia of Finance*, pages 411–420.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 1003–1012.

Horst Pöttker. 2003. News and its communicative quality: The inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 338–348. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1707.03264*.

Md Main Uddin Rony, Mohammad Yousuf, and Naeemul Hassan. 2018. A large-scale study of social media sources in news articles. *arXiv preprint arXiv:1810.13078*.

Christopher Scanlan. 2000. *Reporting and writing: Basics for the 21st century*. Oxford University Press.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2017. Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media, DDDSM@IJCNLP 2017, Taipei, Taiwan, November 27, 2017*, pages 1–8.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective lstms for target-dependent sentiment classification. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3298–3307. ACL.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3346–3359. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 809–819. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. *CoRR*, abs/2003.08385.

Sarah Van Leuven and Annelore Deprez. 2017. 'to follow or not to follow?': How belgian health journalists use twitter to monitor potential sources. *Journal of Applied Journalism & Media Studies*, 6(3):545–566.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the Workshop on Language Technologies and Computational Social Science@ACL 2014, Baltimore, MD, USA, June 26, 2014*, pages 18–22. Association for Computational Linguistics.

Brian Xu, Mitra Mohtarami, and James Glass. 2019. Adversarial domain adaptation for stance detection. *CoRR*, abs/1902.02401.

Wenpeng Yin and Dan Roth. 2018. Twowingos: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 105–114. Association for Computational Linguistics.

Martin Žnidaršič, Jasmina Smailović, Jan Gorše, Miha Grčar, Igor Mozetič, and Senja Pollak. 2018. Trust and doubt terms in financial tweets and periodic reports. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018a. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2):32:1–32:36.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, Rob Procter, and Peter Tolmie. 2015. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *CoRR*, abs/1511.07487.

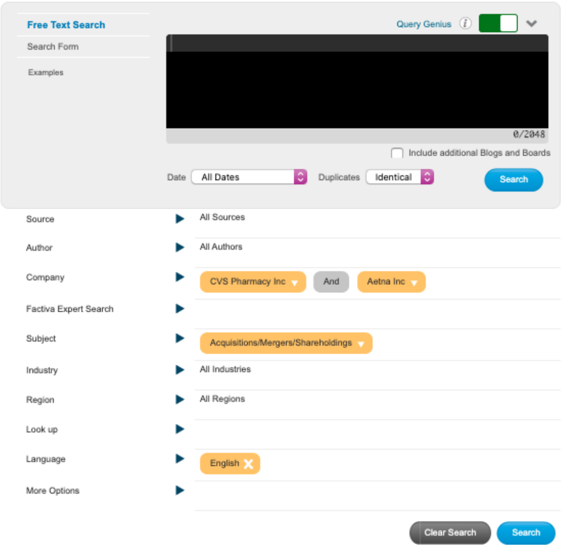
Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *COLING 2016, 26th International Conference on Computational Linguistics, December 11-16, 2016, Osaka, Japan*, pages 2438–2448. ACL.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018b. Discourse-aware rumour stance classification in social media using sequential classifiers. *Inf. Process. Manage.*, 54(2):273–290.

A Corpus-related Specifications

A.1 Screenshot from Factiva

Below, we report a screenshot from the Factiva interface (Johal, 2009), while crawling for the CVS_AET merger:



[redacted]

0/2048

□ Include additional Blogs and Boards

Date: All Dates

Duplicates: Identical

Source: All Sources

Author: All Authors

Company: CVS Pharmacy Inc And Aetna Inc

Factiva Expert Search

Subject: Acquisitions/Mergers/Shareholdings

Industry: All Industries

Region: All Regions

Look up

Language: English

More Options

A.2 Crawling Timelines

Table 6 gives an overview of the considered M&A operations, their respective crawling timelines and the total number of articles.

Merger	Crawl start	Crawl end	Articles
CVS_AET	15/02/2017	17/12/2018	831
CI_ESRX	27/05/2017	17/08/2018	376
ANTM_CI	01/04/2014	28/04/2017	1,199
AET_HUM	01/09/2014	23/01/2017	1,009

Table 6: Crawling specifications.

A.3 Metadata Included in the Corpus

We provide a sample of the data in the Supplementary material. Each sample in the dataset is associated with the following fields:

- *Target merger*; one from {CVS_AET, CI_ESRX, ANTM_CI, AET_HUM}.
- *Stance* of the article with respect to the target merger; one from {*support, refute, comment, unrelated*}.
- *Title* of the article, followed by a ordered list of the article's *Paragraphs*.

- A list of *Evidence Snippets*, indicating
  1) the index of the paragraph in the article where the evidence is located; and
  2) the exact start and end indices of the snippet in the corresponding paragraph.

A.4 Annotation Guidelines

The following is an extract from the annotation guidelines sent to the annotators. Each label description was correlated with a number of examples, which we don't report due to space limitation.

You will be sent a number of news articles. The annotation process consists of choosing one of 4 possible labels for each article and marking which part of the article (e.g., the title or a specific sentence, phrase, paragraph) led you to your assessment.

The four labels to choose from are *Support*, *Comment*, *Refute*, and *Unrelated*.

Label: Support This label should be chosen if the article is supporting the theory that the merger is happening. That is, after reading the article the reader feels more confident that the two companies will merge. Articles that mention the merger as a fact and then talk about e.g. the implications or consequences of the merger should not be labelled as supporting but as commenting.

Label: Refute This label should be chosen if the article is refuting the theory that the merger is happening. That is, after reading the article the reader feels less confident that the two companies will merge. Articles that are voicing doubts or mention potential roadblocks (such as antitrust issues) should be labelled refute as well.

Label: Comment This label should be chosen if the article is commenting on one of the mergers. The article should neither directly state that the merger is happening, nor refute that it will be completed successfully. Articles that mention the merger as a fact and then talk about e.g. the implications or consequences of the merger should also be labelled as commenting. For articles that are long, presenting both positive and negative evidence, annotators should weigh the evidence and conclude whether the article is 'mostly' positive or negative. Only of the assessment of the annotator is that the evidence is equal should the article be labelled as commenting.

Label: Unrelated This label should be chosen if the article is unrelated to the merger in question. Since the articles have been collected from a news aggregation service, some of them may not in fact be about one of the mergers. This label will only have few articles and should be the easiest to identify. Note that an article that is mainly about a different topic/merger, but talks about the relevant merger in one paragraph or just a sentence, annotators should choose the label based on this paragraph or sentence.

B Baselines-Related Specifications

Below, we report on the implementations details for the baselines presented in Section 5. SD stands for *Stance Detection* and ER for *Evidence Retrieval*.

B.1 Dummy Baselines

Two dummy baselines have been considered as lower bound.

- *Random Baseline*.
  SD: outputs a random stance;
  ER: outputs two random sentences chosen from the title and all body's paragraphs.
- *Majority Baseline*.
  SD: always outputs *support* (the most frequent label in the corpus);
  ER: always outputs the title and the first paragraph (the most frequent locations of evidences in the dataset, Figure 6).

B.2 Neural Baselines

Three strong neural baselines, which obtained state-of-the-art results in previous work (Hanselowski et al., 2019), are considered for future reference.

Inputs. The models receive as input $n + 1$ sequences $\{t, s_1, ..., s_n\}$, where t is the target and $\{s_1, ..., s_n\}$ is the list of n sentences from the articles. If training for SD in isolation, such sentences are the *gold evidences*; if jointly training for SD+ER), they are the *evidence candidates*: as a simple sentence retrieval method, we always retrieve the title and the first four paragraphs of the article, where evidence snippets are most frequently located in the corpus (Figure 6). For a target merger between companies *A* and *B* (with acronyms *a* and *b*), we employ as target a string containing the text: `"A (a) will merge with B (b)."`

Encoders. We employ three neural encoders to obtain a target-aware representation h_i of each input sentence s_i:

- *BiLSTM.* We employ 300-dimension word embeddings to encode each input token. The embedding matrix is initialized with Glove[5] embeddings (Pennington et al., 2014), which are kept fixed over training to prevent overfitting. We concatenate each input evidence with the target, and we obtain a hidden representation for each pair of inputs with a BiLSTM network with size of 128 hidden units.

- *UseEmb*. We obtain sentence embeddings for each input sequence with the Universal Sentence Encoder (Cer et al., 2018), and we concatenate each input sentence with the input target. We use the *large* model for English[6]. We then pass the obtained encoded representation through a position-specific dense layer with 128 hidden units.
- *BERTEmb*. We follow the same principle as above, but using Sentence-BERT (Reimers and Gurevych, 2019) to obtain sentence embeddings for each input sentence. We use the `bert-base` model trained on the SNLI and MultiNLI datasets[7].

Decoders. After encoding, we obtain n representations $\{h_1, ..., h_n\}$, where h_i is the target-aware representation of the sentence at position i. Inspired by Yin and Roth (2018), we obtain a probability $\alpha_i \in (0, 1)$ of the sentence s_i being an evidence as:

$$\alpha_i = sigmoid(v \cdot h_i) \tag{1}$$

where v is a learned parameter vector. To model the entire set of input sentences as a whole, we construct their joint representation e as:

$$e = \sum_{i=1}^{n} \alpha_i \cdot h_i \tag{2}$$

We then consider two decoders, depending on the task(s) we are training for (only SD, or SD+ER):

- *Only SD*. We predict the stance label with a softmax operation over the stance tagset on e.
- *ER and SD*. If we jointly perform both ER and SD with a multi-task training setting, we binarize the probability vector $\alpha = [\alpha_1, \alpha_n]$ by rounding at 0.5; we consider all input sentence s_i where $\alpha_i > 0.5$ as an evidence snippet.

C Experimental Setting Specifications

Data Preprocessing. We perform minimal data preprocessing. The following refers to the BiLSTM model: we include all types in the corpus without selecting any minimal frequency; for tokenization, we use NLTK's `word_tokenize` tokenizer (Loper and Bird, 2002)[8]; we pad/cut input sentences up to 10 tokens (in the case of the article's title) or 25 tokens (in the case of the article's paragraphs).

(Hyper)-Parameters and Runtime Specifications. Refer to Appendix B for a description of the considered models' architectures (completed with embedding size and number of hidden units per layer). We train all models with Adagrad setting the learning rate to 0.02. We train with batches of 32 samples for a maximum of 70 epochs, using Early Stopping with a patience of 10. To prevent overfitting, dropout of 0.2 has been used during training on all layers of the models.

Note that, given that this is a resource paper, our goal is to provide a set of robust baselines for future research. For this reason, we don't perform extensive hyper-parameter tuning on the selected models.

Table 7 reports on the total number of (trainable) parameters for each considered model.

Model	#parameters	#trainable parameters
3 classes		
BiLSTM	1,701,832	201,832
UseEmb	657,032	657,032
BertEmb	984,712	984,712
4 classes		
BiLSTM	1,701,969	201,969
UseEmb	657,161	657,161
BertEmb	984,841	984,841

Table 7: Number of (trainable) parameters for all considered models and training settings.

This resulted in the average runtime/step reported in Table 8 (the average runtime is calculated over five different runs of the same model, trained on the ANTM_CI, AET_HUM and CVS_AET mergers).

Training Setting. All models are trained using cross-validation, testing on one merger and training on the other three.

[5] We use 300-dimensional word embeddings pretrained on Wikipedia 2014 + Gigaword 5, https://nlp.stanford.edu/projects/glove/

[6] https://tfhub.dev/google/universal-sentence-encoder-large

[7] https://github.com/UKPLab/sentence-transformers

[8] https://www.nltk.org/api/nltk.tokenize.html

		Training Setting	
	Model	3 classes	4 classes
SD	BiLSTM	33s 11ms	37s 13ms
	UseEmb	0s 147μs	1s 513μs
	BertEmb	0s 161μs	1s 408μs
SD+ER	BiLSTM	41s 14ms	37s 13ms
	UseEmb	0s 167μs	1s 418μs
	BertEmb	0s 163μs	1s 422μs

Table 8: Average runtime/step for each considered model and training setting.

To account for performance fluctuations (Reimers and Gurevych, 2017), we run 5 simulations for each model and take the average of the results, weighting according to the size of the collected articles for each merger.

Table 9 reports the standard deviation between different runs of the same model. Interestingly, UseEmb is the most stable model for SD, while BertEmb is most stable for ER.

		SD			ER	
	Model	P	R	F_1	P@5	R@5
3 classes	BiLSTM	3.039	7.909	9.817	–	–
	UseEmb	1.246	3.681	5.897	–	–
	BertEmb	1.972	2.967	4.700	–	–
	BiLSTM	1.353	2.490	5.273	8.362	10.58
	UseEmb	1.287	2.806	4.295	8.337	10.26
	BertEmb	4.723	6.131	6.770	6.154	11.60
4 classes	BiLSTM	1.214	2.646	1.943	–	–
	UseEmb	0.102	2.876	3.825	–	–
	BertEmb	5.016	3.413	4.986	–	–
	BiLSTM	1.657	2.440	3.148	7.562	9.806
	UseEmb	2.304	3.027	4.064	7.457	10.745
	BertEmb	4.882	4.637	4.311	4.592	11.50

Table 9: Standard deviation between results obtained with the considered models over different runs. For each training setting (3 vs 4 classes) we first report σ on SD in isolation, then on jointly training SD+ER.

Computing Infrastructure. We run experiments on an NVIDIA GeForce GTX 1080 GPU.

Evaluation Specifications. For SD, we use the sklearn's (Pedregosa et al., 2011) implementation of macro-averaged precision, recall and F_1 score[9]. For ER, we use Thorne et al. (2018)'s implementation of P@5 and R@5[10], which has also been used by Hanselowski et al. (2019).

D Correlation Analysis

D.1 Implementation Details

For the correlation analysis in Section 6, we used Panda's implementation of the Spearman correlation[11] (Wes McKinney, 2010).
We calculate the standard error as:

$$\sigma_x = \frac{1 - r_x^2}{\sqrt{n-2}} \tag{3}$$

where r_x is the correlation coefficient and n is the number of observations (i.e. the number of days of observations collected for each mergers).

D.2 The Case of the Anthem/Cigna Merger

Figure 10 shows the distribution of tweets and articles over time for ANTM_CI. Three distinct phases can be distinguished in the timeline of the merger.

The first phase goes from the beginning of the data collection to the first report on the companies' talks which appeared on a major news outlet. During this phase, we observe minor movements in the Twitter signal and some sparse news articles. Considering only *related* samples, most of the tweets and articles in this phase disappear. However, at the end of this phase there are spikes in the Twitter signal. This suggests that during this period the ongoing talks between the companies are not publicly known, but at the very end information may be leaked. The tweet signal spikes during the first phase on 20.05.2015, around one month before the first news report.

The second phase begins with the first report by a major news outlet and lasts until the beginning of the antitrust process. It is characterised by large spikes in both the volume of tweets posted and the number of published articles. The first spike in the news articles occurs on 15.06.2015, when the Wall Street Journal – as it happens for most considered mergers – reports on the ongoing talks between the two companies. The second spike occurs on 24.07.2015, when the companies publicly announce the merger with a joint press release. These two spikes in the news signal are mirrored in the tweets. After the initial reporting about the two companies' intentions, most news articles and

[9] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

[10] https://github.com/sheffieldnlp/fever-scorer/

[11] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html

tweets discuss the implications of the merger, displaying a constant but not heavy activity.

The third phase begins with the antitrust process and lasts until the end of the merger's timeline. Spikes in the volume of tweets and articles can be observed around specific events, such as when the official antitrust complaint is presented to the Department of Justice (DOJ), at the start of the antitrust trial and around the date of the court decision. During this phase, spikes present a very similar distribution for both signals.

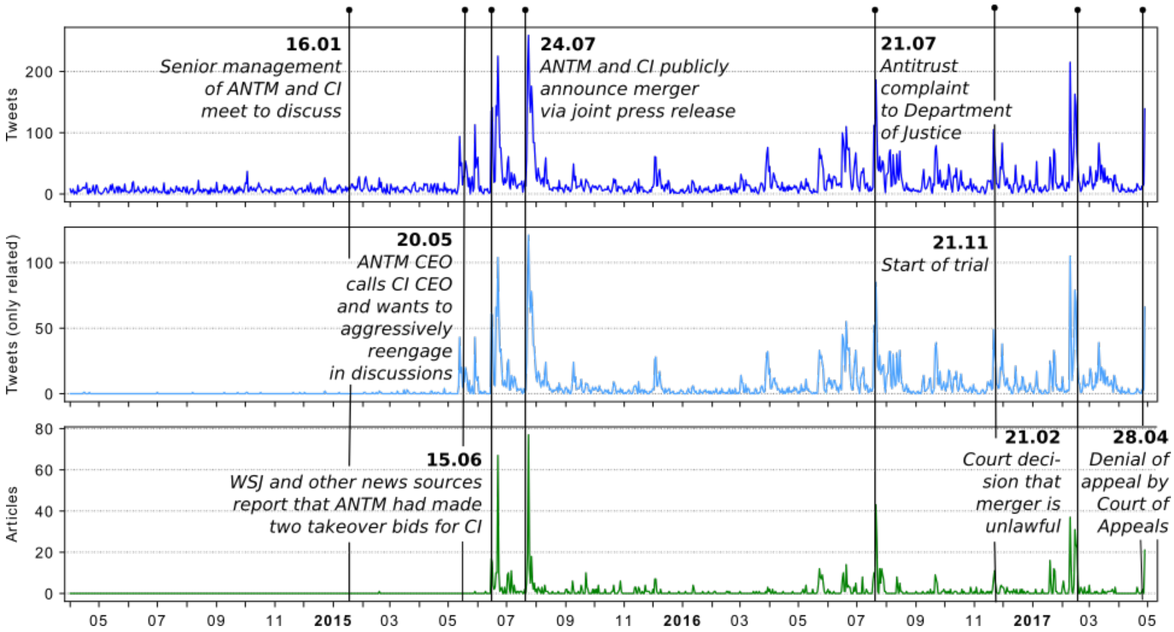


Figure 10: Evolution of the ANTM_CI merger over time. From top to bottom: volume of all posted tweets discussing this target in the WT–WT corpus (Conforti et al., 2020); volume of tweets annotated as *related*; volume of published news articles in STANDER.