A Semantics-Aware Approach to Automated Claim Verification

Blanca Calvo Figueras
Barcelona Supercomputing Center
blanca.calvo@bsc.es

Montse Cuadros
Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
mcuadros@vicomtech.org

Rodrigo Agerri
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
rodrigo.agerri@ehu.eus

Abstract

The influence of fake news in the perception of reality has become a mainstream topic in the last years due to the fast propagation of misleading information. In order to help in the fight against misinformation, automated solutions to fact-checking are being actively developed within the research community. In this context, the task of Automated Claim Verification is defined as assessing the truthfulness of a claim by finding evidence about its veracity. In this work we empirically demonstrate that enriching a BERT model with explicit semantic information such as Semantic Role Labelling helps to improve results in claim verification as proposed by the FEVER benchmark. Furthermore, we perform a number of explainability tests that suggest that the semantically-enriched model is better at handling complex cases, such as those including passive forms or multiple propositions.

1 Introduction

With the rise of digital channels that disseminate all kinds of information, misinformation has become a big challenge for a healthy society (Hermida, 2010). Fake news has been defined as a news article or message published through media that carries false information (Kshetri and Voas, 2017). Although this is not a new phenomenon, the current absence of control systems in social media facilitates the fast spreading of misinformation, arriving to a large number of users and greatly influencing their perception of real world events (Zubiaga et al., 2018). Recent work has shown that fake news spread faster in social media than factual news (Vosoughi et al., 2018), which is why researchers from different fields have proposed using automated solutions to help dealing with this situation (Zhou and Zafarani, 2020; Oshikawa et al., 2020).

Claim verification is the task of assessing the veracity of a statement by finding evidence about the claimed facts. This work is usually done manually by fact-checkers, who use their trusted sources to label the claims as true, false or other assessments. Automated Claim Verification, as proposed by Thorne et al. (2018), consists in, given a claim, finding the evidence regarding the veracity of that claim to then infer its truth-label. Systems for Automated Claim Verification have been trained both using synthetic data (Thorne et al., 2018; Jiang et al., 2020), and crawling datasets from fact-checking websites (Augenstein et al., 2019; Wang, 2017). These datasets have enabled the development of models for the three tasks involved in the claim-verification pipeline: document retrieval (Chen et al., 2017a; Nogueira and Cho, 2020), sentence retrieval (Danesh et al., 2015; Hanselowski et al., 2018), and natural language inference (Parikh et al., 2016; Chen et al., 2017b). In this work, we focus on the last module: natural language inference (NLI).

Given the right pieces of evidence, a fact-checking system will have to reason over all the utterances involved in order to determine if the claim can be supported, refuted, or whether there is not enough info to do so. In Figure 1, for instance, it should recognize that the *Rodney King riots* is the same entity in the claim and in evidence 1. Then, it should identify that the location of this event is *Los Angeles County*, and understand that evidence 2 confirms that this happens to be *the most populous county in the USA*.

As illustrated in Figure 1, this reasoning process requires a deep understanding of the semantics of all the utterances involved. In this work, we propose to introduce explicit semantic knowledge in order to improve the systems for Automated Claim Verification. We hypothesize that this information might guide the natural language inference model in claims that have complex semantics.

The linguistic information we use in this work is Semantic Role Labelling (SRL, Palmer et al., 2005) and Open Information Extraction (OpenIE,

Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER), pages 37 - 48

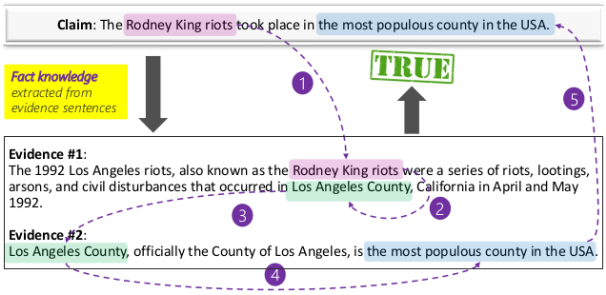

Figure 1: Natural language inference reasoning example, by Zhong et al. (2020)

Etzioni et al., 2008). In our experiments, these semantic structures are used as additional input to the BERT contextualized word embeddings (Devlin et al., 2019). We integrate this information using the SemBERT architecture presented in Zhang et al. (2020a).
The contributions of this work are the following:

- We perform a qualitative analysis to compare synthetic datasets and naturally-occurring datasets for claim verification. We find that synthetic claims are semantically more simple.
- We improve the widely used BERT language model to address the inferential component of the task by adding explicit semantic information. We also make publicly available our model to the community.
- We perform explainability tests to understand the influence of the additional semantic information. The performed tests suggest that the semantically-enriched model is better at handling complex cases.

In the following sections, we introduce previous work on datasets, systems and semantic structures (Section 2), we explain our experiments (Section 3) and expose the primary results (Section 4), we perform explainability tests to qualitatively assess the influence of semantic structures (Section 5), and finally we draw our conclusions and future work (Section 6).

2 Related Work

Automated Claim Verification is a relatively new task, and a lot of effort have been put on how to develop datasets to train automated systems for this task. In the following subsections we introduce some of these efforts and the systems that have been developed on these datasets. We also present previous work using semantic structures.

2.1 Datasets

Ideally, a claim verification system should be able to take sentences from naturally-occurring texts (e.g. news articles, social media posts or political speeches) and assess their veracity. However, developing training data for this task has some complexities, such as defining the ground truth and creating a knowledge database with boundaries, which allows the annotators to know for sure that the ground truth is right. For this reason, there have been several attempts to approximate the task by creating domain-specific datasets (Scifact, Wadden et al., 2020) and synthetic datasets (FEVER and HoVer, Thorne et al., 2018; Jiang et al., 2020). These datasets consist of a set of claims annotated with their ground truth, together with a knowledge base, in which the truth labels are based (e.g. a set of scientific abstracts or a set of Wikipedia articles). The labels are usually Supports, Refutes and NotEnoughInfo. Due to its size and popularity, FEVER has become a benchmark for Automated Claim Verification and has been used in the organization of several shared tasks.

Other datasets exist containing naturally-occurring claims (Augenstein et al., 2019; Wang, 2017). These are generally scraped from fact-checking websites, and sometimes include the justification of the fact-checker for the given label. However, these datasets do not contain a fixed database of evidence. This makes it very difficult to use them to train inference systems, as the ground truth at the moment of fact-checking can be different from the current one. Additionally, there is a high heterogeneity in the inventory of labels across different fact-checking platforms.

2.2 Systems

In the first FEVER shared task (2018), Nie et al. (2019) obtained the highest label accuracy by adding the sentence similarity score between claim and evidence to the embedding representation of evidences. Hanselowski et al. (2018) (UKP-Athene) won the task by using noun phrases to query the Wikipedia search API in the retrieval module.

After the shared task, better results were achieved using transformer-based models (Soleimani et al., 2019). Further improvements came from rethinking the interaction between the pieces of evidences. Zhou et al. (2019) (GEAR)

developed a graph approach that uses an attention layer to propagate the information within the evidences. And Zhong et al. (2020) (DREAM) used semantic information to break the evidences into arguments, which then interacted with each other in a graph approach. These two last approaches both used transformer-based models and helped to advance the state-of-the-art on this task. Finally, a recent work (Krishna et al., 2021) developed a system (ProoFVer) based on sequences of natural language logic relations, where the proofs are generated from the claims and corresponding evidence by a seq2seq model (Lewis et al., 2020) and represented as triples. The last inferential step is performed using natural logic proofs only. ProoFVer is the current state-of-the-art on the FEVER benchmark.

Finally, Augenstein et al. (2019) developed a multi-task learning system to deal with a dataset of naturally-occurring claims. They accounted for the multiple labels by creating embeddings for each of these labels, and combining those with the evidence-claim embedding.

2.3 Semantic Structures

Natural Language Inference can be framed as a relation extraction task: in order to know if a sentence is entailed by another sentence, it is necessary to identify the semantic relation between the verb and the arguments of both the premises and hypothesis. For this reason, early approaches used semantic information to approach tasks that required NLI. He et al. (2015) introduced the possibility of annotating semantic roles as a question-answering task, showing that predicate-argument structures can be extracted from natural language questions. In the same direction, Stanovsky et al. (2015) demonstrated the contribution of semantic structures, such as OpenIE, when performing text comprehension with a simple unsupervised lexical matching algorithm.

The creation of more extensive datasets (Bowman et al., 2015; Williams et al., 2018) enabled the development of systems based on neural networks (Wang and Jiang, 2016). Later, the release of transformer-based language models (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019) revolutionized the performance of many NLP tasks, which also was reflected in NLI.

Recently, a new research direction has suggested using information that had been helpful for NLI models before the arrival of deep learning, in order to guide the self-attention mechanisms (Zhang et al., 2020b). Zanzotto et al. (2020) designed a system that explicitly embeds syntax parse trees into sentence embeddings using distributed tree kernels, and can visualise the decisions made (KERMIT). Zhang et al. (2020a) introduced a modified BERT architecture (SemBERT), that maps semantic role labels (SRL) to embeddings in parallel and integrates the text representation with the contextual explicit semantic embedding to obtain a joint representation. In automated claim verification, Zhong et al. (2020) used SRL tuples to structure information graphs.

A variety of lexical resources have been developed to structure the semantics of sentences with different focus (Baker et al., 1998; Kipper et al., 2000). Semantic roles (SRL), for instance, represent the different arguments that a predicate might have. These semantic categories are relations between noun phrases and verbs. An ideal set of roles should be able to concisely label the arguments of any relation. Nonetheless, the exact set of these relations remains an open discussion inside the linguistic community (Bonial et al., 2011).

SRL in PropBank (Palmer et al., 2005) was designed to be used in automated tasks. The goal of this framework is to create a shallow but broad representation that covers every instance of every verb in a corpus to allow representative statistics to be calculated. PropBank defines semantic roles on a verb-by-verb basis: individual verb's semantic arguments are numbered, beginning with zero. In the example in Figure 2, the agent of the verb *bought* is Arg0, the theme is Arg1, the location Arg2, and the price Arg3.

[Mr. Bean]$_{Arg0}$ [bought]$_{V}$ [the sweater]$_{Arg1}$ [from the second hand store]$_{Arg2}$ [for 400 pounds]$_{Arg3}$.

Figure 2: PropBank semantic roles example

Open Information Extraction (OpenIE) was first introduced as an extraction paradigm to tackle an unbounded number of relations (Etzioni et al., 2008). Systems based on OpenIE extract relational tuples from text by identifying relation phrases and the arguments associated to these relations (Mausam et al., 2012). Stanovsky et al. (2015) were the first to propose this task as an intermediate structure for other semantic tasks, similar to what was already being done with other linguistic

	Supports	Refutes	NEI
Training	80,035	29,775	35,639
Development	3,333	3,333	3,333
Test	3,333	3,333	3,333

Table 1: Number of claims in the FEVER dataset

information, such as semantic roles, syntactic dependencies or lexical representations. An example of the difference between SRL in PropBank and OpenIE is shown in Figure 3.

PropBank:
[John]$_{Arg0}$ [refused]$_{V}$ [to visit a Vegas casino]$_{Arg1}$
[John]$_{Arg0}$ refused to [visit]$_{V}$ [a Vegas casino]$_{Arg1}$
OpenIE:
[John]$_{Arg}$ [refused to visit]$_{V}$ [a Vegas casino]$_{Arg}$

Figure 3: Example of the representations extracted with OpenIE and SRL in PropBank from Stanovsky et al. (2015)

3 Experiments

In this work, we use the FEVER dataset (Thorne et al., 2018). We first develop a baseline using the BERT model (Devlin et al., 2019), and then introduce two types of semantic information to the model (SRL and OpenIE) by using the SemBERT architecture (Zhang et al., 2020a).

3.1 Data

The FEVER dataset consists of 185,445 generated claims with its truth label and the evidence for that label, divided between a train, a development and a test set. The statistics can be seen in Table 1.

The claims were generated manually by annotators, using the June 2017 Wikipedia dump. They were given sentences at random and were asked to generate variations of the claims, altering them in ways that may or may not change their truth label. The types of mutations were: paraphrasing, negation, substitution of entity/relation, and making the claim more general or specific. In a second phase, these claims were labelled as Supports, Refutes or NotEnoughInfo (NEI), and the evidences used for the labelling were recorded (Thorne et al., 2018).

FEVER has been criticized for missing some of the complexity that naturally-occurring claims have, such as claims that contain rich semantics in long and complex sentences (Thorne and Vlachos, 2019). For this reason, we decided to perform a

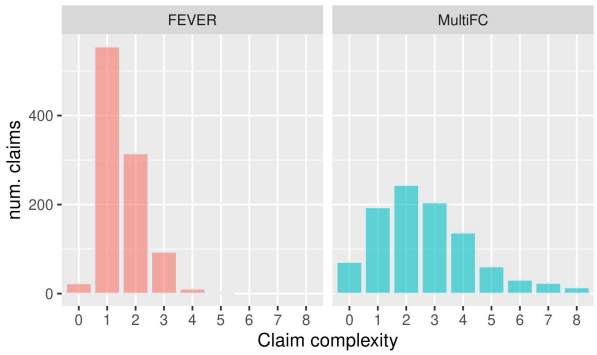

Figure 4: Comparison of claim complexity between FEVER and MultiFC. Axis x indicates the number of verbs per claim.

comparison between the claims in FEVER and in a naturally-occurring claims dataset (MultiFC, Augenstein et al., 2019), we used a sample of 1000 claims of each dataset. As a proxy to measure semantic complexity, we counted the number of verbs per claim[1]. As can be observed in Figure 4, while claims in FEVER are almost always simple (contain 1-2 verbs), MultiFC follows a Benford distribution, in which the number of claims decreases when complexity increases.

This complexity difference lead our attention towards building a system that improves the performance of the semantically complex examples present in FEVER, in order to be able to use these systems in naturally-occurring data.

3.2 Experimental setup

As this work focuses on the NLI module of claim verification, we do not perform evidence retrieval, and instead, we use the evidences retrieved by the system that had the highest evidence recall in the FEVER shared task (Hanselowski et al., 2018). We take the top 5 evidences for each claim.

Given that transformer-based architectures, such as BERT (Devlin et al., 2019), have given state-of-the-art results in the task of NLI (Soleimani et al., 2019), we use this architecture as our baseline, and add the semantic information to it. BERT is designed to be given plain natural text as input. However, recent work suggests that it could benefit from additional linguistic knowledge (Zanzotto et al., 2020; Zhong et al., 2020). Zhang et al. (2020a) proposed an architecture that is able to encode both natural text and semantic information: SemBERT.

[1]Measured with the Universal pos-tags of the nltk package.

At a first step, SemBERT encodes text in the same way that BERT does: tokenizing the text into sub-tokens and computing contextualized embeddings for each of these sub-tokens. In parallel, SemBERT takes the semantic representation that it is given, which should have one tag per word (SRL tags in the original paper), and computes tag embeddings. Given that a single sentence can have several predicates, and consequently several argument-predicate structures (propositions), Zhang et al. (2020a) allow for up to three different representation vectors. A linear layer aggregates the three semantic representation vectors (for the three propositions per sentence allowed) into one final semantic embedding. Then, the BERT word representation and the final semantic representation are concatenated. According to the authors, SemBERT outperforms BERT in NLI tasks, increasing the final accuracy between 1 and 3 percentage points (Zhang et al., 2020a).

In this work, we adapt SemBERT to fit the requirements of Automated Claim Verification. Since we use 5 pieces of evidence per claim, the input to the model consists of 6 sentences. Given that we can have many propositions per instance, we allow up to 12 propositions per instance and implement different sets of tags. Both the SRL tags and the OpenIE tags are extracted with the AllenNLP toolkit (Gardner et al., 2018; Shi and Lin, 2019; Stanovsky et al., 2018) and mapped to the different sets.
To summarise, the model has two separate inputs of the exact same length:

1. The claim plus the 5 concatenated evidences (given to the model as represented in the left part of Figure 5).

2. The semantic tags for each word in the claim and evidences (given to the model as represented in the right part of Figure 5).

Our experiments include a BERT baseline and 5 other models that interact with different sets of semantic tags. All the models have a maximum input length of 250 tokens, and are trained for 4 epochs with a batch size of 20, an AdamW optimizer (Loshchilov and Hutter, 2019) with the learning rate set to 2e-5, and a linear scheduler.

SemBERT_base On first instance, we train a model with all the semantic roles (from now on we will call them tags) retrieved by the AllenNLP parser. This results in a tags-vocabulary of size 19, so the encoding layer contains 19 contextualized embeddings (plus 3 BERT-special tokens) of length 10 (see the tags in Appendix A).

Provided that the set of tags is quite large, the sparsity of the SRL data could be preventing the model from learning patterns. We make additional experiments reducing the set of tags by doing two different mappings.

SemBERT_tags1 One mapping reduces the amount of tags by removing the positional part of the tags, which is given in BIO notation (e.g. I- B-), and reducing the amount of modifier arguments to just *temporal*, *location* or *other modifiers*, leaving a total of 10 tags. The correspondence with the tags of the first model are in Appendix A.

SemBERT_DREAM The second tag set comes from using the mapping of the DREAM system (Zhong et al., 2020), which additionally reduces all the ARG tags to a single *argument* tag, leaving a total of 5 tags. The correspondence can be seen in Appendix A.

SemBERT_Attention The original SemBERT model uses a linear layer to squeeze all the 12 predicates into one. That is needed to remove the multiple predicates dimension and be able to concatenate the representation coming from the SRL to the one produced by BERT. We hypothesized that this linear layer could be replaced by an attention mechanism that allowed evidences to reason between them, inspired by the self-attention mechanism from Zhou et al. (2019).

This self-attention mechanism concatenates the vectors of each predicate in pairs, to then compute self-attention between them and use that information to reshape the 12 representations into one, using a linear layer. To train this model, we used the mapping of SemBERT_tags1.

SemBERT_OpenIE In order to get the OpenIE tags we have also used the AllenNLP parser (Gardner et al., 2018). Then, we have kept the tags *argument*, *verb* and *O – O* meaning that the word is not part of the predicate. This makes a tag vocabulary of size 3.

4 Results

Table 2 reports the accuracy of the predictions of all these models in the development set. We observe that all the SemBERT experiments have a

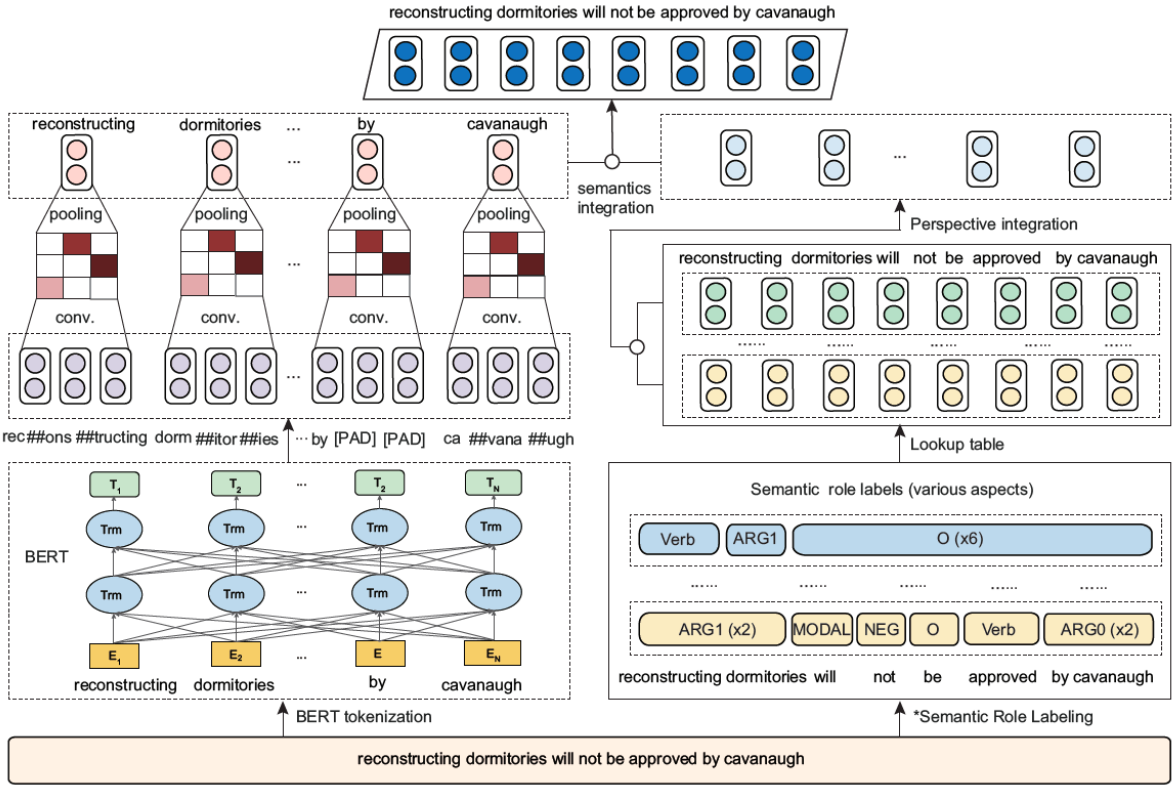

Figure 5: SemBERT architecture by Zhang et al. (2020a)

better performance than the BERT baseline. This difference is of 1 to 2 percentage points. Our best model is the SemBERT model with the SRL set tags1 (SemBERT_tags1).

Going back to our hypothesis that claim complexity will be better understood by using models that include SRL, we calculate the accuracy separately for claims with more (and with less) than 5 verbs. The SemBERT_tags1 model improves 6.5 points on complex claims over BERT, while it just improves 1.5 points on simple claims. However, since FEVER has few complex claims (only 62), further experiments with more complex claims should be used to confirm our hypothesis.

	Label Accuracy
BERT_base (baseline)	73.82
SemBERT_base	75.06
SemBERT_tags1	**75.37**
SemBERT_DREAM	75.12
SemBERT_Attention	74.92
SemBERT_OpenIE	74.34

Table 2: Results from all the models in the FEVER development set

The evaluations on the test set can be seen in Table 3. In the unseen data, the SemBERT model also outperforms the BERT baseline by 1.3 percentage points in label accuracy. Both models drop around 3 percentage points with respect to the development set. Additionally, we also report the results on the test set of previous work such as UKP-Athene (Hanselowski et al., 2018), GEAR (Zhou et al., 2019), DREAM (Zhong et al., 2020), and ProoFVer (Krishna et al., 2021). For our model, we used the evidences extracted by UKP-Athene, and some pre-processing scripts from GEAR, which explains why all three models have (almost) the same F1 for evidence retrieval. Our model outperforms both of these models in the inference module.

	Evidence F1	Label Acc.	Fever Score
UKP-Athene	36.97	65.46	61.58
GEAR	36.87	71.60	67.10
DREAM	39.45	76.85	70.60
ProoFVer	40.03	79.47	76.82
BERT_base	36.87	70.86	65.52
SemBERT_tags1	36.87	**72.18**	67.16

Table 3: Results on the test set of our models and previous work

Our approach is similar to the one in DREAM, as both integrate semantic information to improve the reasoning process. However, instead of using a graph-based approach, we use the SemBERT architecture to incorporate the semantic information. As observed, DREAM performs better than our model, suggesting that graph-based architectures might be a better representation for semantic information. Finally, the highest scoring system is ProoFVer[2]. Furthermore, both DREAM and ProoFVer rely on better evidences, as shown by the F1 in Table 3. Still, while being substantially simpler than a higher-performing work such as ProoFVer, our approach provides an effective method to integrate explicit semantic information with clear benefits in performance. Furthermore, our code and model are publicly available to facilitate research on claim verification and reproducibility of results.

5 Explainability tests

While the accuracy results allow for a comparison between models, they are not enough to understand the contribution of the semantic information to the model. For this reason, we decided to perform qualitative explainability tests based on calculating saliency scores and performing adversarial attacks.

5.1 Saliency Scores

Extracting the saliency of each of the tokens given as input is not a trivial task for deep-learning models. Simonyan et al. (2014) proposed to compute them as the gradient of the output with respect to each input. Later improvements to this technique proposed to then multiply these gradients to the input (*InputX-Gradient*), or to overwrite the gradients of the ReLU functions in order to prevent negative gradients from being propagated (*Guided Backpropagation*, Kindermans et al., 2016; Springenberg et al., 2015).

We will use the saliency scores proposed above to get a better grasp of where the model focuses in order to make its inference decisions. For an interpretable output, we want to have one saliency value for each token. Given that the last layer that we can compute the gradients for is the embedding layer, we will get one gradient for each value in the embedding of each token. In order to aggregate these values and get one single value per token we will use the L2 norm (Atanasova et al., 2020).

[2]Results are those reported in the official FEVER leaderboard, which differ from the performance reported in the paper Krishna et al. (2021)

In Figure 6, we can see an example where both BERT and SemBERT get the output right. The instance looks like:

- **Claim**: Telemundo is an English-language television network.
- **Evidence**: Telemundo is an American Spanish-language terrestrial television network owned by Comcast through the NBCUniversal division NBCUniversal Telemundo Enterprises.

Both models output REFUTES and the saliency scores clearly point towards the words *English-language* in the claim, and *Spanish-language* in the evidence. As an opposite case we display Figure 7. In this case, the instance looks like:

- **Claim**: Easy A is directed by Bert V. Royal.
- **Evidence**: Easy A, stylized as easy A, is a 2010 American teen comedy film directed by Will Gluck, written by Bert V. Royal and starring Emma Stone, Stanley Tucci, Patricia Clarkson, Thomas Haden Church, Dan Byrd, Amanda Bynes, Penn Badgley, Cam Gigandet, Lisa Kudrow and Aly Michalka.

In this instance, BERT gets the inference wrong and outputs SUPPORTS, while SemBERT gets it right and outputs REFUTES. Based on the saliency scores, BERT tries to focus on many different tokens, while SemBERT ignores almost all of them. From this observation, we hypothesize that, with such a semantically-complicated evidence (it contains 5 predicates), SemBERT is relying on the semantic information for its decision, which is not plotted on this figure. We further investigate this hypothesis by creating manual adversarial attacks in the next section.

5.2 Adversarial Attacks

Performing adversarial attacks consists on changing the input in order to assess the influence that it has over the output. This has been done both by removing input tokens systematically (Zeiler and Fergus, 2014), and by altering the input instances to generate adversarial attacks which can show what the model actually understands (Ribeiro et al., 2018; Ebrahimi et al., 2018). In this section, we are going to create some manual adversarial attacks in order to test the capabilities of our models.

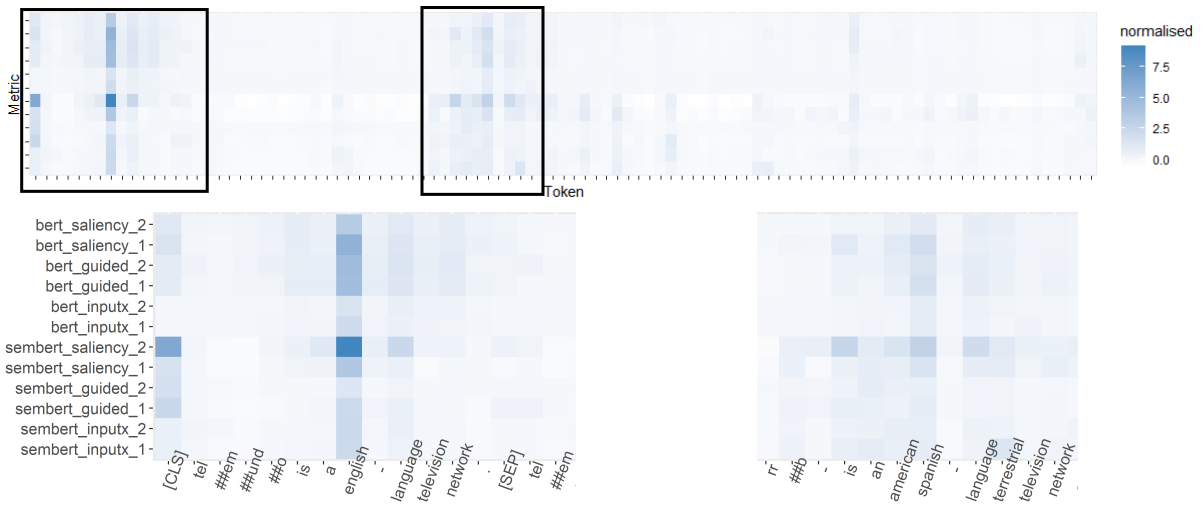

Figure 6: Saliency Scores of the *Telemundo* example with BERT and SemBERT. The above plot shows the entire claim and evidence input, and the plots under it zoom into the relevant parts, delimited with black frames above.

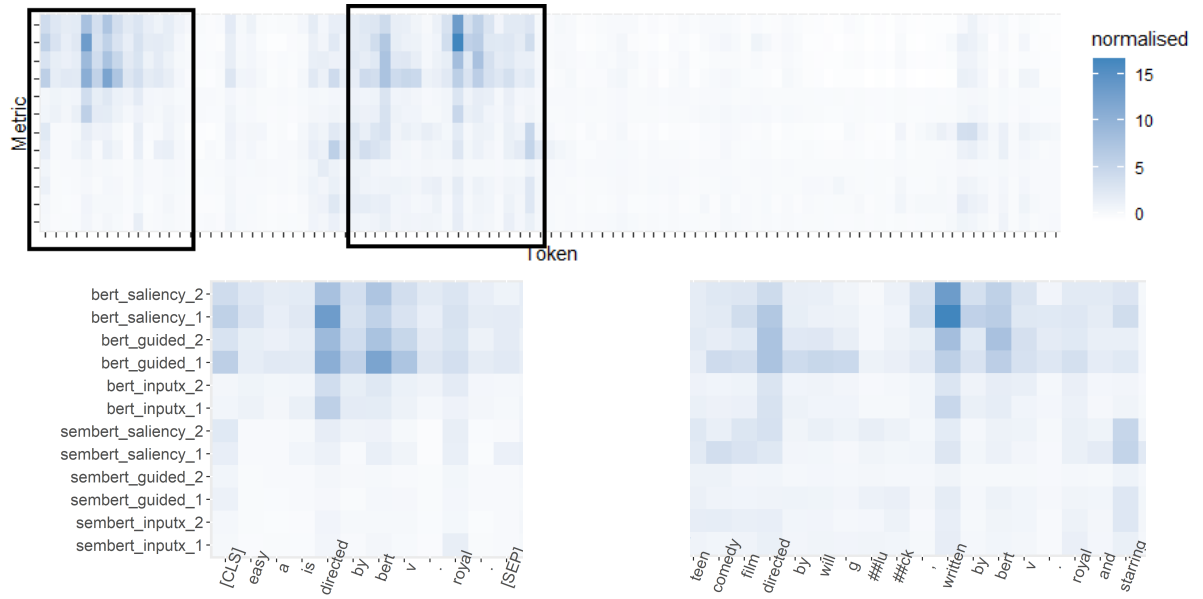

Figure 7: Saliency Scores of the *Easy A* example with BERT and SemBERT. The above plot shows the entire claim and evidence input, and the plots under it zoom into the relevant parts, delimited with black frames above.

Taking the example of *Easy A*, we start by checking that the REFUTES label of SemBERT is not random by changing the claim to *Easy A is written by Bert V. Royal*. SemBERT passes this test and outputs SUPPORTS. Following the tests for semantic structure in Ribeiro et al. (2020)'s CheckList, we modify the evidence by changing the order of the propositions, creating symmetric relations and swapping them to active form. The new versions of the evidence are:

1. **Order change**: Easy A, stylized as easy A, is a 2010 American teen comedy film *written by Bert V. Royal, directed by Will Gluck*, and starring Emma Stone, [...]. ← **Refutes**
2. **Order change**: Easy A, stylized as easy A, is a 2010 American teen comedy film written by Bert V. Royal, starring Emma Stone, [...], *and directed by Will Gluck*. ← **Refutes**
3. **Symmetric relation**: Easy A, stylized as easy A, is a 2010 American teen comedy film *directed by Will Gluck and Bert V. Royal* and starring Emma Stone, [...]. ← **Supports**
4. **Remove the *written by* proposition**: Easy A, stylized as easy A, is a 2010 American teen comedy film directed by Will Gluck, and starring Emma Stone, [..]. ← **Refutes**
5. **Active form**: Easy A, stylized as easy A, is a

2010 American teen comedy film. *Will Gluck directed the film , and Bert V. Royal wrote it.* ← **Refutes**

With all the variations of the evidence presented above, SemBERT always outputs the right label, while BERT just outputs the right label in the last piece of evidence, which contains the same information but in active form. These tests suggest that SemBERT does have capabilities regarding semantic structure that are missing in BERT. However, more systematic tests should be performed in this direction.

6 Conclusion and Future Work

In this work we have investigated if semantic information could help to improve the reasoning process when inferring the truth label of a claim given some pieces of evidence. To this goal, we have used two different semantic parsers and the architecture of the pre-trained model SemBERT (Zhang et al., 2020a). For our experiments, we have used the FEVER dataset (Thorne et al., 2018), which requires building a model that, given some pieces of evidence, can output if a claim is supported, refuted, or the evidence does not give enough information.

We have performed several experiments on top of the SemBERT architecture, such as training models with different kinds of semantic information, different sets of semantic tags, and with an additional attention mechanism to represent the semantic information. In terms of label accuracy, all our experiments have outperformed the baseline, which was a BERT model with no additional semantic information. Our best model uses Semantic Role Labels and a set of 10 different tags, with no additional attention mechanism. This model achieves a label accuracy of 75.37 on the development set and 72.18 on the test set, outperforming the baseline by 1.5 and 1.3 percentage points respectively. Future work could include testing the impact of these semantic structures in models such as RoBERTa (Liu et al., 2019) or XLNet (Yang et al., 2019).

To better understand the contribution of the semantic information, we have performed some explainability tests with our best model. These have shown that the SRL knowledge might be contributing to guiding the model in semantically complex sentences that include several propositions or passive forms.

To keep moving towards systems that can contribute to the work of fact-checkers, future research on claim verification should take two directions. On the one hand, there is a need to develop large datasets that are more similar to naturally-occurring claims. On the other hand, NLI models for claim verification should output more explanatory justifications to their conclusions, which would make these systems more trust-worthy.

In this work, we have not dealt with the task of evidence retrieval. In FEVER, this task is limited by the static Wikipedia database that comes with the dataset. However, in real-world scenarios defining the boundaries of what is trust-worthy information is a challenge that goes beyond research in NLP and reaches the fields of journalism, politics and even philosophy. The non-static nature of what is a true fact is an additional challenge to evidence retrieval.

Acknowledgements

This work has been partially funded by the projects DeepText (KK-2020-00088, SPRI, Basque Government) and DeepReading (RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE). Rodrigo Agerri acknowledges the funding of the UPV/EHU Colab 19/19 project "Tools for the analysis of parliamentary discourses: polarization, subjectivity and affectivity in the post-truth era", the RYC-2017–23647 fellowship and from the ANTIDOTE - EU CHIST-ERA project (PCI2020-120717-2) of the Agencia Estatal de Investigación through the INT-Acciones de Programación Conjunta Internacional (MINECO) 2020 call.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A Diagnostic Study of Explainability Techniques for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. 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)*, pages 4685–4697.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association*

for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, ACL '98/COLING '98, pages 86–90, USA. Association for Computational Linguistics.

Claire Bonial, Susan Brown, W. Corvey, Martha Palmer, Volha Petukhova, and Harry Bunt. 2011. An Exploratory Comparison of Thematic Roles in VerbNet and LIRICS.

Samuel 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*, pages 632–642.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017a. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Soheil Danesh, Tamara Sumner, and James H. Martin. 2015. SGRank: Combining Statistical and Graphical Methods to Improve the State of the Art in Unsupervised Keyphrase Extraction. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 117–126, Denver, Colorado. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel Weld. 2008. Open Information Extraction from the Web. *Commun. ACM*, 51:68–74.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. *EMNLP 2018*, page 103.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-Answer Driven Semantic Role Labeling: Using Natural Language to Annotate Natural Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Alfred Hermida. 2010. Twittering the news. *Journalism Practice*, 4:297–308.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv e-prints*, 1611:arXiv:1611.07270.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. AAAI Press.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2021. Proofver: Natural logic theorem proving for fact verification.

Nir Kshetri and Jeffrey Voas. 2017. The Economics of "Fake News". *IT Professional*, 19:8–12.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ACL*, abs/1910.13461.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *Proceedings of*

the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 523–534, Jeju Island, Korea. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866. Issue: 01.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage Re-ranking with BERT. *arXiv:1901.04085 [cs]*. ArXiv: 1901.04085.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A Survey on Natural Language Processing for Fake News Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv:1904.05255 [cs]*. ArXiv: 1904.05255.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer.

Amir Soleimani, Christof Monz, and Marcel Worring. 2019. *BERT for Evidence Retrieval and Claim Verification*. GroundAI.

J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR (workshop track)*.

Gabriel Stanovsky, Ido Dagan, and Mausam. 2015. Open IE as an Intermediate Structure for Semantic Tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 303–308, Beijing, China. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised Open Information Extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2019. Adversarial attacks against Fact Extraction and VERification. *arXiv:1903.05543 [cs]*. ArXiv: 1903.05543.

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, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science (New York, N.Y.)*, 359(6380):1146–1151.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Shuohang Wang and Jing Jiang. 2016. Learning Natural Language Inference with LSTM. *arXiv:1512.08849 [cs]*. ArXiv: 1512.08849.

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans,

Louisiana. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing Transformer Architectures with Encoders of Explicit Syntactic Interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online. Association for Computational Linguistics.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020a. Semantics-Aware BERT for Language Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9628–9635. Number: 05.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020b. SG-Net: Syntax-guided machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9636–9643. Issue: 05.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning Over Semantic-Level Graph for Fact Checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5):109:1–109:40.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36. Publisher: ACM New York, NY, USA.

A Appendix: Tags mapping

All Tags	Tags1 Tags	DREAM Tags
O	O	O
B-V	V	verb
I-V	V	verb
B-ARG0	ARG0	argument
I-ARG0	ARG0	argument
B-ARG1	ARG1	argument
I-ARG1	ARG1	argument
B-ARG2	ARG2	argument
I-ARG2	ARG2	argument
B-ARG4	ARG4	argument
I-ARG4	ARG4	argument
B-ARGM-TMP	TMP	temporal
I-ARGM-TMP	TMP	temporal
B-ARGM-LOC	LOC	location
I-ARGM-LOC	LOC	location
B-ARGM-CAU	ARGM	argument
I-ARGM-CAU	ARGM	argument
B-ARGM-PRP	ARGM	argument
I-ARGM-PRP	ARGM	argument

Table 4: Mapping between sets of SRL tags