Generating Fact Checking Summaries for Web Claims

Rahul Mishra[1], **Dhruv Gupta**[2], **Markus Leippold**[3]

[1]University of Stavanger, Norway
[1]rahul.mishra@uis.no
[2]Max Planck Institute for Informatics, Germany
[2]dhgupta@mpi-inf.mpg.de
[3]University of Zurich, Switzerland
[3]markus.leippold@bf.uzh.ch

Abstract

We present SUMO, a neural attention-based approach that learns to establish the correctness of textual claims based on evidence in the form of text documents (e.g., news articles or Web documents). SUMO further generates an extractive summary by presenting a diversified set of sentences from the documents that explain its decision on the correctness of the textual claim. Prior approaches to address the problem of fact checking and evidence extraction have relied on simple concatenation of claim and document word embeddings as an input to claim driven attention weight computation. This is done so as to extract salient words and sentences from the documents that help establish the correctness of the claim. However, this design of claim-driven attention does not capture the contextual information in documents properly. We improve on the prior art by using improved claim and title guided hierarchical attention to model effective contextual cues. We show the efficacy of our approach on datasets concerning political, healthcare, and environmental issues.

1 Introduction

Most of the information consumed by the world is in the form of digital news, blogs, and social media posts available on the Web. However, most of this information is written in the absence of facts and evidences. Our ever-increasing reliance on information from the Web is becoming a severe problem as we base our personal decisions relating to politics, environment, and health on unverified information available online. For example, consider the following unverified claim on the Web:

"Smoking may protect against COVID-19."

A user attempting to verify the correctness of the above claim will often take the following steps: issue keyword queries to search engines for the claim; going through the top reliable news articles; and finally making an informed decision based on the gathered information. Clearly, this approach is laborious, takes time, and is error-prone. In this work, we present SUMO, a neural approach that assists the user in establishing the correctness of claims by automatically generating explainable summaries for fact checking. Example summaries generated by SUMO for couple of Web claims are given in Figure 1.

Prior approaches to automatic fact checking rely on predicting the credibility of facts (Popat et al., 2017), instance detection (Ma et al., 2018; Xu et al., 2018), and fact entailment in supporting documents (Parikh et al., 2016). The majority of these methods rely on linguistic features (Popat et al., 2017; Potthast et al., 2018; Qazvinian et al., 2011), social contexts, or user responses (Ma et al., 2015) and comments. However, these approaches do not help explain the decisions generated by the machine learning models. Recent works such as (Atanasova et al., 2019; Mishra and Setty, 2019; Popat et al., 2018) overcome the explainability gap by extracting snippets from text documents that support or refute the claim. (Mishra and Setty, 2019; Popat et al., 2018) apply claim-based and latent aspect-based attention to model the context of text documents. (Mishra and Setty, 2019) model latent aspects such as the speaker or author of the claim, topic of the claim, and domains of retrieved Web documents for the claim. We observe in our experiments that in prior works (Mishra and Setty, 2019; Popat et al., 2018), the design of claim guided attention in these methods is not effective and latent aspects such as the topic and speaker of claims are not always available. The snippets extracted by such models are not comprehensive or topically diverse. To overcome these limitations, we propose a novel design of claim and document

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 81–90
Online, Nov 19, 2020.

Claim:	***Smoking may protect against COVID-19***	**Label: False**	**Verdict: False**
Summary:	The current evidence suggests that the severity of COVID is higher among smokers, prevent the health risk linked to the excessive consumption or misuse" of nicotine products by people hoping to protect themselves from COVID-19. Evidence from China, where COVID-19 originated, shows that people who have cardiovascular and respiratory conditions caused by tobacco use, or otherwise, are at higher risk of developing severe COVID-19 symptoms. HO urges researchers, scientists and the media to be cautious about amplifying unproven claims that tobacco or nicotine could reduce the risk of COVID-19. Smoking is also associated with increased development of acute respiratory distress syndrome, a key complication for severe cases of COVID-19.		
Claim:	***Deforestation has made humans more vulnerable to pandemics***	**Label: True**	**Verdict: True**
Summary:	Deforestation can directly increase the likelihood that a pathogen will be transferred from wildlife species to humans through the creation of suitable habitats for vector species. Climate change, including deforestation which drives it, is a key driver of cross-species transmission which is where zoonotic emerging diseases come from . There is a correlation between deforestation and the rise in the spread of infectious diseases affecting humans. Deforestation forces various species into smaller, shared habitats and increases encounters between wildlife and humans. Habitat destruction and fragmentation due to deforestation can also increase the frequency of contact between humans, wildlife species, and the pathogens they carry . This can occur through direct transfer of pathogens from animals to humans or indirectly through cross-species transfer of pathogens from wildlife to domesticated species . Deforestation could be to blame for the rise of infectious diseases like the novel coronavirus.		

Figure 1: Example summaries generated by SUMO for unverified claims on the Web.

title driven attention, which better captures the contextual cues in relation to the claim. In addition to this, we propose an approach for generating summaries for fact-checking that are non-redundant and topically diverse.

Contributions. Contributions made in this work are as follows. First, we introduce SUMO, a method that improves upon the previously used claim guided attention to model effective contextual representation. Second, we propose a novel attention on top of attention (Atop) method to improve the overall attention effectiveness. Third, we present an approach to generate topically diverse multi-document summaries, which help in explaining the decision SUMO makes for establishing the correctness of claims. Fourth, we provide a novel testbed for the task of fact checking in the domain of climate change and health care.

Outline. The outline for the rest of the article is as follows. In Section 2, we describe prior work in relation to our problem setting. In Section 3, we formalize the problem definition and describe our approach, SUMO, to generate explainable summaries for fact checking of textual claims. In Sections 4 and 5, we describe the experimental setup that includes a description of the novel datasets that we make available to the research community and an analysis of the results we have obtained. In Section 6, we present the concluding remarks of our study.

2 Related work

We now describe prior work related to our problem setting. First, we describe works that rely only on features derived from documents that support the input textual claim. Second, we describe works that additionally include features derived from social media posts in connection to the claim. Third and finally, we describe works that rely on extracting textual snippets from text documents to explain a model's decision on the claim's correctness.

2.1 Content Based Approaches

Prior approaches for fact checking vary from simple machine learning methods such as SVM and decision trees to highly sophisticated deep learning methods. These works largely utilize features that model the linguistic and stylistic content of the facts to learn a classifier (Castillo et al., 2011; Ma et al., 2016; Qazvinian et al., 2011; Rashkin et al., 2017). The key shortcomings of these approaches are as follows. First, classifiers trained on linguistic and stylistic features perform poorly as they can be misguided by the writing style of the false claims, which are deliberately made to look similar to true claims but are factually false. Second, these methods lack in terms of user response and social context pertaining to the claims, which is very helpful in establishing the correctness of facts.

2.2 Social Media Based Approaches

Works such as (Qian et al., 2018; Shu et al., 2019; Yang et al., 2019) overcome the issue of user feedback by using a combination of content-based and context-based features derived from related social media posts. Specifically, the features derived from social media include propagation patterns of claim related posts on social media and user responses in the form of replies, likes, sentiments, and shares. These methods outperform content-based methods significantly. In (Yang et al., 2019), the authors propose a probabilistic graphical model for causal mappings among the post's credibility, user's opinions, and user's credibility. In (Qian et al., 2018), the authors introduce a user response generator based on a deep neural network that leverages the user's past actions such as comments, replies, and posts to generate a synthetic response for new social media posts.

2.3 Model Explainability

Explaining a machine learning model's decision is becoming an important problem. This is because modern neural network based methods are increasingly being used as black-boxes. There exist few machine learning models for fact checking that explain this decision via summaries. Related works (Mishra and Setty, 2019; Popat et al., 2018) achieve significant improvement in establishing the credibility of textual claims by using external evidences from the Web. They additionally extract snippets from evidences that explain their model's decision. However, we find that the claim-driven attention design used in these methods is inadequate, and does not capture sufficient context of the documents in relation to the input claim. The snippets extracted by these methods are often redundant and lack topical diversity offered by Web evidences. In contrast, our method enhances the claim-driven attention mechanism and generates a topically diverse, coherent multi-document summary for explaining the correctness of claims.

3 SUMO

We now formally describe the task of fact checking and explain SUMO in detail. SUMO works in two stages. In the first stage, it predicts the correctness of the claim. In the second stage, it generates a topically diverse summary for the claims. As input, we are provided with a Web claim $c \in C$, where C is a collection of Web claims and a pseudo-relevant set of documents $D = \{d_1, d_2, \ldots, d_m\}$, where m is the number of results retrieved for claim c. The documents $d \in D$ are retrieved from the Web as potential evidences, using claim c as a query. Each retrieved document d is accompanied by its title t and text body bd, i.e. ($d = \langle t, bd \rangle$). We define the representation of each document's body as a collection of k sentences as $bd = \{s_1, s_2, ..., s_k\}$ and each sentence as the collection of l words as $\{w_1, w_2, ..., w_l\} \in \mathbb{W}$, where $\mathbb{W}$ is the overall word vocabulary of the corpus. By k and l, we denote the maximum numbers of sentences in a document and the maximum number of words in a sentence, respectively. We use both WORD2VEC and pre-trained GloVe embeddings to obtain the vector representations for each claim, title, and document body. The objective is to classify the claim as either true or false and automatically generate a topically diverse summary pieced together from D for establishing the correctness of the claim.

3.1 Predicting Claim Correctness by Neural Attention

We now describe SUMO's neural architecture (see Figure 2) that helps in predicting the correctness of the input claim along with its pseudo-relevant set of documents. The model additionally learns the weights to words and sentences in the document's body that help ascertain the claim's correctness. First, we need to encode the pseudo-relevant documents that support a claim. To this end, as a **sequence encoder**, we use a Gated Recurrent Unit (GRU) to encode the document's body content. Claim and document's title are not encoded using sequence encoder; we explain the method to represent them in detail in upcoming sections.

Claim-driven Hierarchical Attention., aims to attend salient words that are significant and have relevance to the content of the claim. Similarly, we aim to attend the salient sentences at the sentence level attention. Recent works have used claim guided attention to model the contextual representation of the retrieved documents from the Web. These approaches provide claim-guided attention by first concatenating the claim word embeddings with document word embeddings and then applying a dense softmax layer to learn the attention weights as follows:

$$r_i = c_i \parallel d_i \;\; \& \;\; a_i = \tanh(W_a r_i + b_a) \qquad \alpha = \text{softmax}(a_i), \tag{1}$$

where c_i and d_i are the i^{th} claim and document

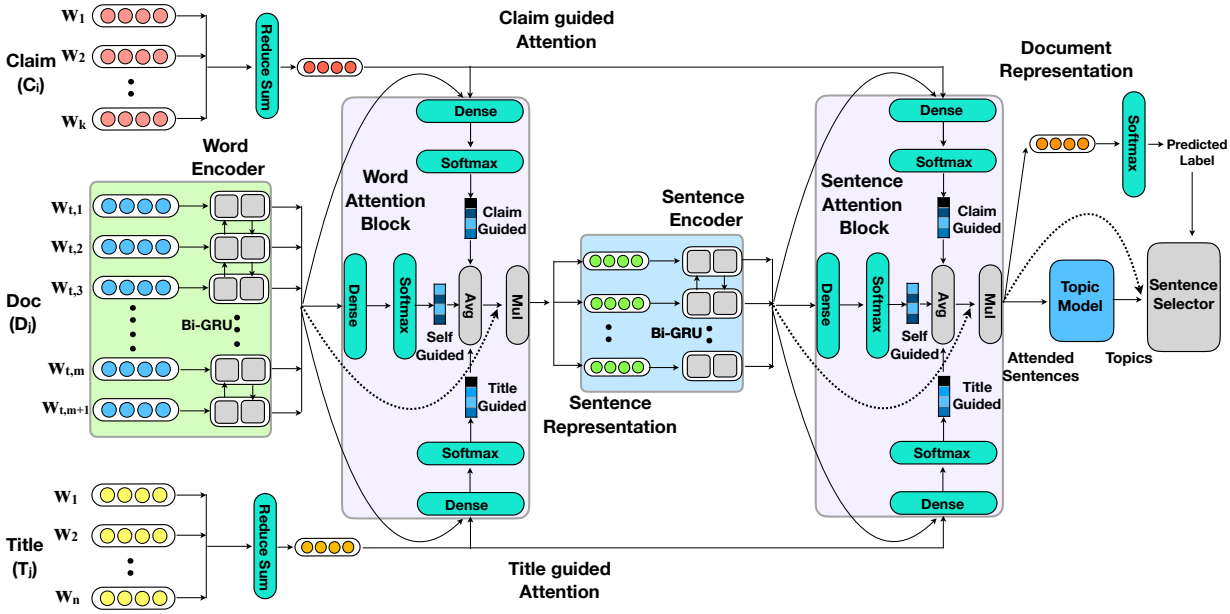

Figure 2: SUMO's neural network architecture for establishing the correctness of Web claims.

embeddings. W_a and b_a are the weight matrix and bias and α is the learned attention weight. However, during experiments, we observe that applying claim-based attention provides an inferior overall document representation. Therefore, we do not concatenate the claim and document embeddings before attention weight computation.

Each claim c_i is consists of l maximum number of words as $\{w_1, w_2, ..., w_l\}$. We represent each claim c_i as the summation of embeddings of all the words contained in it as: $Cl_i = \sum_{j=1}^{l} f(w_j)$, where $f(w_j)$ is the word embedding of the j_{th} word of claim c_i. Claim representation Cl_i and hidden states h_j from the GRU are used to **compute word-level claim-driven attention weights** as:

$$u_{j,i} = \tanh(W_{j,i} h_j + b_{j,i})$$
$$\alpha^C_{j,i} = \text{softmax}(u^\top_{j,i} Cl_i), \tag{2}$$

where $W_{j,i}$ and $b_{j,i}$ are the weight matrix and bias, $\alpha^C_{j,i}$ is the word level claim driven attention weight vector, and $h_j = (h_{j,1}, h_{j,2}, ..., h_{j,l})^\top$ represents the tuple of all the hidden states of the words contained in the j^{th} sentence. To **compute sentence level claim-driven attention weights**, we use claim representation Cl_i and hidden states h^S_j from the sentence level GRU units as concatenations of both forward and backward hidden states $h^S_j = \overrightarrow{h^S_j} \parallel \overleftarrow{h^S_j}$ as follows:

$$u_j = \tanh(W_j h^S + b_{j,i})$$
$$\alpha^C_j = \text{softmax}(u^\top_j Cl_i), \tag{3}$$

where W_j and b_j are the weight matrix and bias, $h^S = (h^S_1, h^S_2, ..., h^S_l)^\top$ is the combination of all hidden states from sentences, and $\alpha^C_j = (\alpha_{j,1}, \alpha_{j,2}, ..., \alpha_{j,k})^\top$ is the sentence level claim-driven attention weight vector for the j^{th} document.

Title-driven Hierarchical Attention. The objective of using the document title is to guide the attention in capturing sections in the document that are more critical and relevant for the title. Articles convey multiple perspectives, often reflected in their titles. By title-driven attention, we attend to those words and sentences that are not covered in claim-driven attention. Title-driven attention at both word and sentence level can be computed in a similar fashion as claim-driven attention. Each title t_i is comprised of l maximum number of words as $\{w_1, w_2, \ldots, w_l\}$. We represent each claim t_i as the summation of embeddings of all the words contained in it as: $T_i = \sum_{j=1}^{l} f(w_j)$. Title-driven attention weights for both words and sentence level

can be computed as follows:

$$\begin{aligned} u_{j,i} &= \tanh(W_{j,i}h_j + b_{j,i}) \\ \alpha^T_{j,i} &= \text{softmax}(u^\top_{j,i}T_i) \\ u_j &= \tanh(W_j h^S + b_{j,i}) \\ \alpha^T_j &= \text{softmax}(u^\top_j T_i). \end{aligned} \tag{4}$$

Hierarchical Self-Attention. Self-attention is a simplistic form of attention. It tries to attend salient words in a sequence of words and salient sentences in a collection of sentences based on the self context of a sequence of words or a collection of sentences. In addition to claim-driven and title-driven attention, we apply self-attention to capture the unattended words and sentences which are not related to claim or title directly but are very useful for classification and summarization. Self-attention weights for both words and sentence level can be computed as follows:

$$\begin{aligned} u_{j,i} &= \tanh(W_{j,i}h_j + b_{j,i}) \\ \alpha^{Sl}_{j,i} &= \text{softmax}(u^\top_{j,i}) \\ u_j &= \tanh(W_j h^S + b_{j,i}) \\ \alpha^{Sl}_j &= \text{softmax}(u^\top_j), \end{aligned} \tag{5}$$

where $\alpha^{Sl}_{j,i}$ and α^{Sl}_j are the self-attention weight vectors at word and sentence levels respectively.

Fusion of Attention Weights. We combine the attention weights from the three kinds of attention mechanisms: claim-driven, title-driven, and self-attention at both the word and sentence levels. At the word level, we set:

$$\alpha_j = (\alpha^C_{j,i} + \alpha^T_{j,i} + \alpha^{Sl}_{j,i})/3 \tag{6}$$

$$S_j = \alpha^\top_j h_j, \tag{7}$$

where $\alpha^C_{j,i}$, $\alpha^T_{j,i}$, and $\alpha^{Sl}_{j,i}$ are the attention weight vectors from claim, title and self-attention at the word level. S_j is the formed sentence representation after overall attention for the j^{th} sentence. At the sentence level, we set:

$$\alpha^S_j = (\alpha^C_j + \alpha^T_j + \alpha^{Sl}_j)/3 \tag{8}$$

$$doc = \alpha^\top_j h^S, \tag{9}$$

where α^C_j, α^T_j, and α^{Sl}_j are the attention weight vectors from claim, title, and self-attention at the sentence level, and doc is the formed document representation after overall attention.

Attention on top of Attention (Atop). Although the fusion of the three kinds of attention weights as an average of them works well, we realize that we lose some context by averaging. To deal with this issue, we use a novel attention on top of attention (Atop) method. We concatenate all three kinds of attentions α_{con} and α^S_{con} at both the word and sentence levels correspondingly. We apply a $\tanh$ activation based dense layer as a scoring function and subsequently, a softmax layer to compute attention weights for each of three kinds of attention:

$$\begin{aligned} \text{At word level:} \quad \alpha_{con} &= (\alpha^C_{j,i} \parallel \alpha^T_{j,i} \parallel \alpha^{Sl}_{j,i}) \\ u_{wa} &= \tanh(W_{wa}\alpha_{con} + b_{wa}) \\ \beta^w &= \text{softmax}(u_{wa}) \\ S_j &= \beta^w_1 \alpha^C_{j,i} + \beta^w_2 \alpha^T_{j,i} + \beta^w_3 \alpha^{Sl}_{j,i} \\ \text{At sentence level:} \quad \alpha^S_{con} &= (\alpha^C_j \parallel \alpha^T_j \parallel \alpha^{Sl}_j) \\ u_{sa} &= \tanh(W_{sa}\alpha^S_{con} + b_{sa}) \\ \beta^s &= \text{softmax}(u_{sa}) \\ doc &= \beta^s_1 \alpha^C_j + \beta^s_2 \alpha^T_j + \beta^s_3 \alpha^{Sl}_j, \end{aligned} \tag{10}$$

where β^w and β^s are the learned attention weight vectors for three kinds of attentions at the word and sentence levels, and doc is the formed document representation after Atop attention.

Prediction and Optimization. We use the overall document representation doc in a softmax layer for the classification. To train the model, we use standard softmax cross-entropy with logits as a loss function, we compute $\hat{y}$, the predicted label as:

$$\hat{y} = \text{softmax}(W_{cl} doc + b_{cl}). \tag{11}$$

3.2 Generating Explainable Summary

Recent works retrieve documents from the Web as external evidence to support or refute the claims and thereafter extract snippets as explanations to model's decision (Mishra and Setty, 2019; Popat et al., 2018). However, the extracted snippets from these methods are often redundant and lack topical diversity. The objective of our summarization algorithm is to provide ranked list of sentences that are: novel, non-redundant, and diverse across the topics identified from the text of the documents. In this section, we outline the method we utilize for achieving this objective.

Multi-topic Sentence Model: Each sentence in the document that is retrieved against the claim is modeled as a collection of topics: $s =$

$\langle a^{(1)}, a^{(2)}, \ldots a^{(k)} \rangle$. Let $\mathcal{A}$ be the set of topics $a_i \in \mathcal{A}$ across all candidate sentences from all the pseudo relevant set of documents D for the claim.

Objective. We formulate the summarization task as a diversification objective. Given a set of relevant sentences $\mathcal{R}$ which are attended by Atop attention in SUMO while establishing the claim's correctness. We have to find the *smallest* subset of sentences $\mathcal{S} \subseteq \mathcal{R}$ such that *all* topics $a_i \in \mathcal{A}$ are covered. This is a variation of the Set Cover problem (Agrawal et al., 2009; Korte and Vygen, 2002; Vazirani, 2001; Williamson and Shmoys, 2011; Johnson, 1974; Lovász, 1975; Chvátal, 1979). However, unlike IA-Select (Agrawal et al., 2009) we do not choose to utilize the Max Coverage variation of the Set Cover problem. Instead, we formulate it as Set Cover itself (Korte and Vygen, 2002; Vazirani, 2001). That is, given a set of topics $\mathcal{A}$, find a minimal set of sentences $\mathcal{S} \subseteq \mathcal{R}$ that cover those topics (Vazirani, 2001). Additionally, the inclusion of each sentence in the subset $\mathcal{S}$ has a *cost* associated with it, given by:

$$cost(s) = (Score)^{-1}$$
$$Score = (\lambda\theta_s + (1-\lambda)(W_{wa} + W_{sa})), \quad (12)$$

where θ_s is the topic distribution score for sentence s computed using a topic model (e.g., Latent Dirichlet Allocation (Blei et al., 2003)), $W_{wa} = \sum_{i=1}^{l} W_{wa}(i)$ is the average of attention weights of the words contained in sentence s, W_{sa} is the attention weight of the sentence s, and λ is a parameter to be tuned. We briefly describe our adaptation of the Greedy algorithm, which provides an approximate solution to the Set Cover problem, based on the discussion in (Korte and Vygen, 2002; Vazirani, 2001; Williamson and Shmoys, 2011; Johnson, 1974; Lovász, 1975; Chvátal, 1979).

Algorithm 1: *Adaption of the approximate Greedy algorithm for Set Cover problem from (Korte and Vygen, 2002; Vazirani, 2001; Williamson and Shmoys, 2011; Johnson, 1974; Lovász, 1975; Chvátal, 1979) to our topical diversification problem setting. At each iteration, a sentence is chosen that covers the most number of topics reflected by topic distribution score and has the highest attention weights. As an output, we are assured a non-redundant, novel, and a diversified set of sentences.*

Input: $\mathcal{A}$: Set of topics learned from the topic model for diversification.
$\mathcal{R}$: Set of sentences, attended by Atop.
Output: $\mathcal{S} \subseteq \mathcal{R}$: Diversified set of sentences over $\mathcal{A}$
$\mathcal{S} \leftarrow \phi$ // $\mathcal{S}$ contains diversified sentences
$\mathcal{A}' \leftarrow \phi$ // $\mathcal{A}'$ contains topics covered by $\mathcal{S}$
while $\mathcal{A}' \neq \mathcal{A}$ **do**
  /* identify the sentence that covers the most topics and is highly relevant for fact-checking */
  $s^* \leftarrow \arg\min_{s \in \mathcal{R} \setminus \mathcal{S}} \frac{cost(s)}{|\mathcal{A} - \mathcal{T}'|}$
  $\mathcal{A}' \leftarrow \mathcal{A}' \cup \{a_{s^*}\}$ // a_{s^*} is the dominant topic of sentence s^*
  $\mathcal{S} \leftarrow \mathcal{S} \cup s^*$
end

Table 1: Dataset Statistics

PUBLIC DATASETS		
STATISTICS	POLITIFACT	SNOPES
#CLAIMS	3568	4341
#DOCUMENTS	29556	29242
#DOMAINS	3028	3267
NEW DATASETS		
STATISTICS	CLIMATE	HEALTH
#CLAIMS	104	100
#DOCUMENTS	1050	978
#DOMAINS	97	83

4 Evaluation

Datasets. We use two publicly available datasets, namely PolitiFact political claims dataset and Snopes political claims dataset (Popat et al., 2018) for evaluating SUMO's capability for fact checking. Dataset statistics for both the datasets are shown in Table 1. In the case of Politifact, claims have one of the following labels, namely: 'true', 'mostly true', 'half true', 'mostly false', 'false', and 'pants-on-fire,'. We convert 'true', 'mostly true', and 'half true' labels to the 'true' and the rest of them to 'false' label. For the Snopes dataset, each claim has either 'true' or 'false' as a label.

We evaluate SUMO for the task of summarization on PolitiFact, Snopes, Climate, and Health datasets. The two new datasets, Climate and Health, are about climate change and health care respectively. We test SUMO only on the PolitiFact and Snopes dataset for the task of fact checking as they are magnitudes larger than the new datasets that we release. The climate change dataset contains claims broadly related to climate change and global warming from `climatefeedback.org`. We use each claim as a query using Google API to search the Web and retrieve external evidences in the form of search results. Similarly, we create a dataset related to health

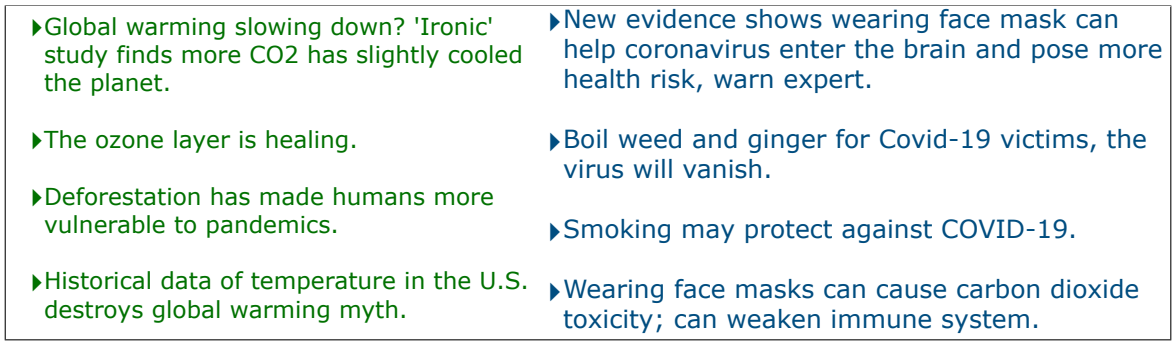

Figure 3: Examples from climate change and health care dataset

care that additionally contains claims pertaining to the current global COVID-19 pandemic from `healthfeedback.org`. Examples of claims from these two datasets are shown in Figure 3. We make the new datasets, publicly available to the research community at the following URL: `https://github.com/rahulOmishra/SUMO/`.

SUMO Implementation. We use TensorFlow to implement SUMO. We use per class accuracy and macro F_1 scores as performance metrics for evaluation. We use bi-directional Gated Recurrent Unit (GRU) with a hidden size of 200, word2vec (Mikolov et al., 2013), and GloVe (Pennington et al., 2014) embeddings with embedding size of 200 and softmax cross-entropy with logits as the loss function. We keep the learning rate as 0.001, batch size as 64, and gradient clipping as 5. All the parameters are tuned using a grid search. We use 50 epochs for each model and apply early stopping if validation loss does not change for more than 5 epochs. We keep maximum sentence length as 45 and maximum number of sentences in a document as 35. For the task of summarization, we use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) as a topic model to compute topic distribution scores and the dominant topic for each candidate sentence.

5 Results

5.1 Setup for the Task of Claim Correctness

We experiment with five variants of our proposed SUMO model and compare with six state-of-the-art methods. The six state-of-the-art methods are as follows. First, we have the basic Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997)) unit which is used with claim and document contents for classification. Second, we have a convolutional neural network (CNN) (Kim, 2014) for document classification. Third, we compare against the model proposed in (Tang et al., 2015) that uses a hierarchical representation of the documents using hierarchical LSTM units (Hi-LSTM). Fourth, we compare against the model proposed in (Yang et al., 2016) that uses a hierarchical neural attention on top of hierarchical LSTMs (HAN) to learn better representations of documents for classification. Fifth, we compare against the model proposed in (Popat et al., 2018) that uses a claim guided attention method (DeClarE) for correctness prediction of claims in the presence of external evidences. Sixth and finally, we compare against the recent work (Mishra and Setty, 2019) that improves on DeClarE method by using latent aspects (speaker, topic, or domain) based attention.

The proposed five variants of our method SUMO are as follows. First, we have the SUMO-AW2V variant that corresponds to the basic SUMO model with word2vec embeddings. Second, we have SUMO-AtopW2V variant consists of the SUMO model with WORD2VEC embeddings. Furthermore, in SUMO-AtopW2V we use Atop method of attention fusion rather than a simple average. Third, we have the SUMO-AGlove variant, which is the basic SUMO model that uses GloVe embeddings. Fourth, we have the SUMO-AtopGlove variant, that consists of the SUMO model with GloVe embeddings. Moreover, in SUMO-AtopGlove, we use Atop method of attention fusion rather than a simple average. Fifth and finally, we have the SUMO-AtopGlove+source-Emb variant that is similar to SUMO-AtopGlove however with additional source embeddings (domains of retrieved documents).

5.2 Claim Correctness Task Results

The results for establishing claim correctness are shown in Table 2. We observe that the basic LSTM based model achieves 57.89% and 69.89% in terms of macro F_1 accuracy in prediction of claim correctness for POLITIFACT and SNOPES, respec-

Table 2: Comparison of the proposed models with various state of the art baseline models for two publicly available datasets.

POLITIFACT				SNOPES			
Model	**True Accuracy**	**False Accuracy**	**Macro F_1**	**Model**	**True Accuracy**	**False Accuracy**	**Macro F_1**
LSTM	53.51	56.32	57.89	LSTM	69.23	70.67	69.89
CNN	55.92	57.33	59.39	CNN	72.05	74.29	72.63
HAN	60.13	65.78	63.44	HAN	72.89	76.25	73.84
DeClarE (full)	68.18	66.01	67.10	DeClarE (full)	60.16	80.78	70.47
SADHAN-agg	68.37	78.23	75.69	SADHAN-agg	79.47	84.26	80.09
SUMO-AW2V	67.30	69.22	70.74	SUMO-AW2V	77.32	80.67	75.56
SUMO-AtopW2V	67.81	70.09	71.15	SUMO-AtopW2V	78.02	81.66	76.86
SUMO-AGlove	68.03	72.57	72.39	SUMO-AGlove	78.74	82.03	77.22
SUMO-AtopGlove	68.93	73.43	72.79	SUMO-AtopGlove	78.89	82.46	78.45
SUMO-AtopGlove+source-Emb	**69.33**	**80.08**	**77.69**	SUMO-AtopGlove+source-Emb	**81.29**	**86.82**	**82.93**

tively. The CNN model performs slightly better than LSTM as it captures the local contextual features better. The hierarchical attention network outperforms CNN with macro F_1 accuracy of 63.4% and 73.84%. The reason for this improvement is hierarchical representation using word and sentence level attention. The state of the art DeClarE model provides significant improvements on baseline methods with macro F_1 accuracy of 67.10% and 70.47%. This gain can be attributed to claim guided attention and source embeddings. However, we observe that this design of claim based attention is not very effective. The more recent work, SADHAN improves on DeClarE, which uses a similar design for claim-oriented attention and incorporates a more comprehensive structure by using several latent aspects to guide attention.

SADHAN outperforms DeClarE with macro F_1 accuracy of 75.69% and 80.09%, respectively. Interestingly, we observe that the basic SUMO model with word2vec embeddings performs better than DeClarE with source embeddings. This observation is a clear indication of the superiority of our claim- and title-driven attention design. The SUMO with Atop attention fusion is more effective than a simple average fusion of attention weights, which becomes apparent from the gain in macro F_1 accuracy in both the datasets. SUMO with pertained GloVe embeddings outperforms the word2vec versions of SUMO as the GloVe embeddings are trained on a large corpus and therefore captures better context for the words. SUMO-AtopGlove+source-Emb outperforms all the other models and it is statistically significant with a p-value of 2.79×10^{-3} for POLITIFACT and 3.09×10^{-4} for SNOPES. The statistical significance values were computed using a two sample Student's t-test. We notice that SUMO could not outperform SADHAN without source embeddings, as SADHAN uses the very complex structure, having three parallel models with hierarchical latent aspects guide attention. However, SADHAN has many drawbacks. First, it is challenging to train and requires more hardware resources and time. Second, the latent aspects are not available for all the Web claims. Therefore, it is not generalizable. Third, it fails to accommodate new values of latent variables at the test time.

5.3 Setup for the Task of Summarization

For the evaluation of the summarization capability of SUMO, we create gold reference summaries for claims. For creating the gold reference summaries, we include all the facts related to the claim, which are important for the claim correctness prediction, non-redundant, and topically diverse. We find that the descriptions provided for a claim on fact-checking websites such as `snopes.com` and `politifact.com` are suitable for this purpose. We use cosine similarity score of 0.4 between claims and sentences of description to filter out irrelevant or noisy sentences. As evaluation metrics, we use ROUGE-1, ROUGE-2, and ROUGE-L scores. The ROUGE-1 score represents the overlap of unigrams, while the ROUGE-2 score represents the overlap of bigrams between the summaries generated by the SUMO system and gold reference summaries. The ROUGE-L score measures the longest matching sequence of words using Longest Common Sub-sequence algorithm.

Standard summarization techniques are not useful in such a scenario as the objective of summarization with standard techniques is usually not fact-checking. Hence, we compare the SUMO results with an information retrieval (BM25) and a natural language processing based method (QuerySum). BM25 is a ranking function, which uses a probabilistic retrieval framework and ranks the documents based on their relevance to a given search query. We use Web claims as a query and apply BM25 to get the most relevant sentences from all

Table 3: Results for the Task of Summarization.

Model	ROUGE-1	ROUGE-2	ROUGE-L
BM25	26.08	14.78	29.98
QuerySum	29.78	16.49	30.16
SUMO	**33.89**	**19.21**	**35.92**

the documents retrieved for the claim. We also compare the results with the query-driven attention based abstractive summarization method QuerySum (Nema et al., 2017), which also uses a diversity objective to create a diverse summary. We use ROUGE metrics with a gold reference summary to evaluate the generated summaries.

5.4 Comparison of Summarization Results

Results for the task of summarization are shown in Table 3, the QuerySum method performs significantly better than BM25 with a ROUGE-L score of 30.16 as it uses query-driven attention and diversity objective, which results in a diverse and query oriented summary. The proposed model SUMO outperforms QuerySum with a ROUGE-L score of 35.92. We attribute this gain to the use of word and sentence level weights, which are trained using back-propagation with correctness label. We also notice that in QuerySum some sentences are related to the claim but are not useful for fact checking. Therefore, they are absent in the gold reference summary. The results for SUMO are statistically significant (p-value $= 1.39 \times 10^{-4}$) using a pairwise Student's t-test.

6 Conclusion

We presented SUMO, a neural network based approach to generate explainable and topically diverse summaries for verifying Web claims. SUMO uses an improved version of hierarchical claim-driven attention along with title-driven and self-attention to learn an effective representation of the external evidences retrieved from the Web. Learning this effective representation in turn assists us in establishing the correctness of textual claims. Using the overall attention weights from the novel Atop attention method and topical distributions of the sentences, we generate extractive summaries for the claims. In addition to this, we release two important datasets pertaining to climate change and healthcare claims.

In future, we plan to investigate the BERT (Devlin et al., 2019) and other Transformer (Vaswani et al., 2017) architecture based embedding methods in place of GloVe (Pennington et al., 2014) embeddings for better contextual representation of words.

References

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, page 5–14, New York, NY, USA. Association for Computing Machinery.

Pepa Atanasova, Jakob Grue, Simonsen Christina, and Lioma Isabelle. 2019. Generating Fact Checking Explanations.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *WWW*.

Vasek Chvátal. 1979. A greedy heuristic for the set-covering problem. *Math. Oper. Res.*, 4(3):233–235.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sepp Hochreiter and Schmidhuber. 1997. Long short-term memory. volume 9, page 1735–1780, Cambridge, MA, USA. MIT Press.

David S. Johnson. 1974. Approximation algorithms for combinatorial problems. *J. Comput. Syst. Sci.*, 9(3):256–278.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Bernhard Korte and Jens Vygen. 2002. Approximation algorithms. In *Combinatorial Optimization: Theory and Algorithms*, pages 361–396, Berlin, Heidelberg. Springer Berlin Heidelberg.

László Lovász. 1975. On the ratio of optimal integral and fractional covers. *Discret. Math.*, 13(4):383–390.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3818–3824. AAAI Press.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. CIKM '15, page 1751–1754.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. WWW '18, page 585–593, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Rahul Mishra and Vinay Setty. 2019. Sadhan: Hierarchical attention networks to learn latent aspect embeddings for fake news detection. ICTIR '19, page 197–204.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072, Vancouver, Canada. Association for Computational Linguistics.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

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 *WWW*, pages 1003–1012.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *EMNLP*, pages 22–32.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *ACL*, volume 1, pages 231–240.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. EMNLP '11.

Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural user response generator: Fake news detection with collective user intelligence. IJCAI '18, pages 3834–3840.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 312–320, New York, NY, USA. Association for Computing Machinery.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, Lisbon, Portugal. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. abs/1706.03762.

Vijay V. Vazirani. 2001. Approximation algorithms. New York, NY, USA. Springer-Verlag New York, Inc.

David P. Williamson and David B. Shmoys. 2011. The design of approximation algorithms. New York, NY, USA. Cambridge University Press.

Brian Xu, Mitra Mohtarami, and James Glass. 2018. Adversarial Domain Adaptation for Stance Detection. (Nips):1–6.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: A generative approach.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL: HLT*, pages 1480–1489.