LESA: Linguistic Encapsulation and Semantic Amalgamation Based Generalised Claim Detection Approach (Appendix for EACL-2021 publication)

Shreya Gupta[†*], **Parantak Singh**[‡*], **Megha Sundriyal**[†],
Md Shad Akhtar[†], **Tanmoy Chakraborty**[†]

[†] *IIIT-Delhi, India.* [‡] *Birla Institute of Technology and Science, Pilani, Goa, India.*
{shreyag, meghas, shad.akhtar, tanmoy}@iiitd.ac.in,
f20170109@goa.bits-pilani.ac.in

A Datasets

There exist a few datasets (Peldszus and Stede, 2015; Stab and Gurevych, 2017) for claim detection in online text; however, most of them are formal and structured texts. Despite the abundance of tweets, literature does not suggest any significant effort for claim detection in Twitter, and arguably, the prime reason is the lack of large-scale dataset.

Therefore, we attempted to develop a new and relatively larger dataset for claim detection in OSM platforms. We collected $\sim 10,000$ tweets from various sources (Carlson, 2020; Smith, 2020; Celin, 2020; Chen et al., 2020; Qazi et al., 2020) and manually annotated them. We additionally included claim tweets of Alam et al. (2020) and CLEF-2020 (Barrón-Cedeño et al., 2020).

To the best of our knowledge there exists only one relevant study (by (Alam et al., 2020)) that proposed guidelines for annotating claims in tweets. Some problems with the guidelines are as follows:

- The authors classified tweets as factually verifiable claims and non-factually verifiable claims, which is only a subset of the claims that exist.

- They do not consider personal opinions as claims. Therefore the tweet *"im actually starting to feel like europe is trying to make sure the virus spreads in africa"* has been labelled non-claim. This is problematic because personal opinions with societal implications have the potential to result in conspiracy theories and cause public unrest.

- They label a tweet as a claim only if the entire sentence is a claim; claims existing in a sub-sentence or sub-clause is not considered as a claim. However from this example *"note to with covid19 you cant declare bankruptcy settle out of court pay it to keep quiet hope it disappears in the spring ignore deaths claim its a hoax covid19"*, which has been labelled non-claim, it is clear that claims can be made in sub-clauses (the claim here being covid-19 is a hoax) and are equally important to detect and verify.

- They do not consider indirect claims. For example the following tweet is labelled non-claim: *"folks when you say the corona virus isnt a big deal it only kills the disabled elderly chornicallyill and immunocompromised the implication is that those people are expendable please be more careful"* This tweet indirectly implies that corona only kills the disabled, elderly, chronically ill and immunocompromised persons which is clearly a claim.

- They do not consider claims made in sarcasm or humour. For example, *"crap this virus is turning all the people into pigeons coronavirus"* is labelled as a non-claim under their guidelines.

The aforementioned reasons motivated us to form our own set of guidelines that are more exclusive, nuanced and applicable to our understanding of a claim.

In this section, we present how we annotated the tweets, extrapolating the set of extensive guidelines we formed for the process and the pre-processing methods we adopted.

A.1 Data preprocessing

Before annotating the dataset, we perform the preliminary task of data cleaning. Our pre-processing stage involves removing hashtags, URLs, user handles and non-ASCII characters. All tweets with

[*] First two authors have equal contributions. The work was done when Parantak was an intern at LCS2 Lab, IIIT-Delhi.

character count less than 20 and word count less than 4 are also removed, owing to the lack of context for their interpretation. Finally, we spell-check the words using symspellpy[1]. The final dataset contains 9,894 tweets which were split into 70 : 15 : 15 for training, validation and testing.

A.2 Data Annotation Guidelines

Our official definition adopted for claims is to *state or assert that something is the case, with or without providing evidence or proof*. Following are our guidelines for what qualifies as a claim. Anything that does not qualify as a claim, is labelled as 'non-claim'. However, certain clarifying guidelines are also given for non-claims in Section A.4.

Note that some guidelines for claims also contain situations (with examples) where the guidelines do not generalise and input is labelled non-claim; vice versa also holds. Such exceptions are preceded by an asterisk (*).

Labels: The tweets are labelled as 1 for claim, 0 for non-claim and x in obscure situations. This is considered only when the language of the tweet is incomprehensible or if no guideline can be referred to annotate it.

A.3 Guidelines for 'claims'

- Tweets mentioning statistics, dates or numbers.
  Example: *["just 1 case of corona virus in india and people are crazy for masks daily 400 people die in road crashes still no craze for helmetsthinking face safetysaves be it virus or road crashes"]*

- Tweets mentioning a personal experience.
  Example: *["i live in seattle i have all symptoms of covid19 and have a history of chronic bronchitis since i work in a physical therapy clinic with many 65 patients and those with chronic illnesses i decided to be responsible and go to get tested this is how that went"]*

- Tweets 'reporting' something to be true or an instance to have happened or will happen.
  Example: *["breaking boris johnson says he visited kettering hospital shook hands with corona patients but the hospital doesnt have corona cases shaking hands would be dangerous not sure how ill gloss over the fact the pm is a liar a complete fucking idiot but ill find a way x"]*

- Tweets containing verified facts also account for a claim, a veracious claim that is.
  (*Note*: a fact known by one, may not necessarily be known by another)
  Example: *["The Chinese CDC has started research and development of a vaccine for the #coronavirus."]* - known fact

- Tweets that negate a claim are also accounted as claims.
  Example: *["disinfectants are not a cure for coronavirus"]*

- Tweets that indirectly (subtly) imply that something is true.
  Example 1: *["b52questions 1 why is rudy not under arrest 2 why is harvey not in rikers 3 why is cuccinelli still working 4 has barr quit yet 5 when is flynn being sentenced 6 who trusts pence and mrs miller with messaging about coronavirus 7 are rs happy wtheir guy"]* - indirectly implies rudy is not under arrest and harvey is not in rikers
  Example 2: *["do rich people know theyll get the virus if poor people cant be tested and diagnoseddo rich people know theyll get the virus if poor people cant be tested and diagnosed"]* - indirectly implies rich people will get the virus if poor people can't be tested

- Claims made in sarcasm or humour.
  Example 1: *["@TheDailyShow Newsflash! #trumpfact If you paint your face #orange you will be #immune to #coronavirus"]*
  Example 2: *["RT @_saraellen: If you've ever used messers bathroom you're immune to corona virus."]*
  Example 3: *["corona virus minding its business by avoiding africa and going to other continents"]*
  Example 1 and 2 are both examples of claims, even if evidently sarcastic; Example 3 is a humour oriented claim.

- All opinions are not claims. Opinions that have societal implications are considered as claims.
  Example 1: *["@derekgilbert I think the Chinese stole a bio weapon https://t.co/RcF6XUJv4b, sent it to Wuhan*

[1] https://pypi.org/project/symspellpy/

China, it got out somehow and they cover it up with a story about it originating at a nearby market. They know how bad the virus is and quarantine entire cities. https://t.co/ZwMqsiWGau"]

Example 2: *["I think Burger King fries are better than Mc'D's"]*
Example 1 is an opinion that claims something to have happened, whose veracity will affect a certain section of the society. Hence it will be marked as a claim. Example 2, on the other hand, is a personal belief that majorly impacts only the person making the tweet and is hence marked non-claim.

- Tweet that says something is true and provides an attachment as evidence or to support the statement.
  Example 1: *["RT @Jawn42: If you ate here growing up, you're immune to the Coronavirus. https://t.co/b9a0hm171b"]*
  Example 2: *["RT @TerminalLance: This kills Coronavirus in the system https://t.co/iOFNlkSrUj"]*

  * However, if a person says something is provided in the attachment, that will not be a claim.
  Example: *["im stunned by the depth of coronavirus information being released in singapore on this website you can see every known infection case where the person lives and works which hospital they got admitted to and the network topology of carriers all laid out on a timeseries link"]*

- A claim can be a sub-part of a question.
  Example: *["Does the pneumonia shot help protect from developing pneumonia caused by #covid19"]* - the claim here being that pneumonia is caused by COVID-19.

A.4 Guidelines for 'non-claims'

- Hoping that something happens or feeling something is true is not claiming it.
  Example 1: *["World doesn't end if u don't give your opinions about corona virus. I'm drinking #nilavembu boiled in hot water and hope it prevents. #COVID-19 #coronapocolypse"]*
  Example 2: *["I feel like I'm immune to coronavirus."]*

- Inclusion of words that project doubt over the said statement.
  Example: *["politicalelle Political correctness infecting the #Coronavirus . Let's change words describing the virus maybe that will cure it."]*

  * Tweets containing doubt-casting words can still, however, contain claims.
  Example: *["Coronavirus may have originated in lab linked to China's biowarfare program #coronavirus https://t.co/2NSWidMkoa"]*

- Urging one to not claim something or to spread misinformation is not a claim.
  Example: *["#Covid19 - Dear all: Stop telling the public that plaquenil/Azithromycine is a cure!!!!! Plz some leadership is needed regarding this matter!"]*

- Questioning a possible claim is not a claim.
  Example: *["Do disinfectants really cure Corona?"]*

  * However, a tweet containing a question can still comprise a claim.
  Example: *["Would you like to promote a cure that can kill 90% #COVID-19 virus in the body in 3-15 min?"]*
  The above tweet claims that there exists a cure that can kill 90% of COVID-19 virus in the body

- Warning someone against a claim is not a claim.
  Example: *["If you think drinking disinfectants will cure #Covid_19 , you deserve death #trump"]*
  The above tweet may be hate speech but it does not say something is true or false. Hence it does not fall under the jurisdiction of claims.

B Experimental Details

In this section, we report a few additional experimental results and other supplementary details.

B.1 POS embeddings - Trigram

We compute POS embeddings by learning word2vec skip-gram model (Mikolov et al., 2013)

Models	**Noisy**		**Semi-Noisy**				**Non-Noisy**								**Wt Avg**	
	Twitter		OC		WTP		MT		PE		VG		WD			
	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$		
POS-only [2-gram]	0.41	0.59	**0.52**	0.14	**0.54**	0.23	0.42	0.00	0.43	0.06	**0.55**	0.30	**0.50**	0.09	0.47	0.47
POS-only [3-gram]	**0.49**	**0.74**	0.51	**0.17**	0.49	**0.24**	0.54	0.32	**0.51**	**0.32**	**0.55**	**0.39**	0.43	**0.10**	**0.50**	**0.62**
POS-only [4-gram]	0.23	0.23	0.50	0.05	0.53	0.14	**0.61**	**0.35**	0.43	0.06	0.53	0.18	0.47	0.00	0.41	0.19

Table 1: Macro F1 and claim-F1 for POS n-gram experiments.

Models	**Noisy**		**Semi-Noisy**				**Non-Noisy**								**Wt Avg**	
	Twitter		OC		WTP		MT		PE		VG		WD			
	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$		
POS-only [Bi-LSTM]	**0.49**	**0.74**	**0.51**	**0.17**	**0.49**	**0.24**	0.54	0.32	0.51	0.32	0.55	**0.39**	**0.43**	0.10	**0.50**	**0.62**
POS-only [transformer]	0.48	0.72	0.43	**0.18**	0.42	0.20	**0.56**	**0.37**	**0.57**	**0.36**	**0.58**	0.35	**0.43**	**0.20**	0.48	0.61

Table 2: Macro F1 and claim-F1 for POS architecture experiments

on the tri-gram POS sequence. The choice of tri-gram sequence is empirical. Table 1 shows results of our comparative experiments between bi-gram, tri-gram and four-gram sequences. As can be observed, c-$F1$ score for tri-grams is highest for six out of seven datasets with a weighted average increase of $\geq$ 30% over bi-grams and a three-fold increase over four-grams. Weighted Average m-$F1$ for tri-grams is also the highest by a margin of 6.38% over the next best performing bi-grams.

B.2 POS embeddings - Bi-LSTM

In Table 2, we report the experimental results to support the choice of using Bi-LSTM, as oppose to Transformers, for extracting the POS features. As evident from Table 2, both c-$F1$ and m-$F1$ scores for Bi-LSTM are better for 4 out of 7 datasets. Additionally, the weighted average for both metrics is higher in case of Bi-LSTM.

B.3 Hyperparameter

Detail about hyperparameters is given in Table 3. For the skip-gram model, we set *context window* = 6, *embedding dimension* = 20, and discard the POS sequence with *frequency* $\leq$ 2.

The DEP Embedding for each text distribution is prepared using the transformer architecture, wherein for our specific prototype, we compute dependency embeddings with *dimension* = 20 using 5 attention heads and a feed-forward *dimension* = 128. The attained representation is then pooled using *GlobalAveragePooling* and then passed through two linear layers with 64 and 32 hidden units respectively.

The BERT Embedding is trained as a downstream task wherein we use ”bert-base-uncased” provided as pre-trained Language Model by HuggingFace. We use an Adam Optimizer with a *learning rate* = $2e^{-5}$ and a *batch size* = 16 for the same.

Hyper-parameter	**Config**
BERT Fine-tuning	
Epochs	3
Optimizer	Adam
Learning Rate	$2e^{-5}$
Batch size	16
Dependency Encoder - Pretraining	
Epochs	10
Embedding dimension	20
Attention heads	5
Feed-Forward units	128
Dropout	0.3
Optimizer	Adam
Activation	ReLU & SoftMax
Loss function	Cross Entropy
Part-of-Speech - Pretraining	
Embedding dimension	20
Window span	6
Minimum count	2
Amalgamated Model	
Epochs	25
Batch size	256
Bi-LSTM units	128
Hidden units	256, 32, 16, 8
Dropout	0.3
Optimizer	Adam
Activation	ReLU & SoftMax
Loss function	Cross Entropy

Table 3: Hyper-parameters of our `LESA` model.

The BERT-layer from after being fine-tuned on our corpus is kept frozen in the final model, while the pooled output from the same is passed through two dense layers with 768 (default hidden size of the BERT configuration made available through Hugging Face) and 32 hidden units respectively to obtain a representation for further use in the model.

The information state obtained from the concatenation of the prior three representations is then processed by an attention layer followed by a dense layer of 16 units and ReLU activation function and 30% dropout for regularisation. Finally, two dense layers of 8 units and 2 units respectively culminate into a softmax for classification. The layers use ReLU and Softmax activation functions respectively. We use the Adam Optimizer and sparse categorical cross-entropy as the loss function for the main output as well as for the auxiliary outputs.

References

Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, and Preslav Nakov. 2020. Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. *arXiv preprint arXiv:2005.00033*.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *Advances in Information Retrieval*, pages 499–507, Cham. Springer International Publishing.

Carlson. 2020. Coronavirus tweets.

Sven Celin. 2020. Covid-19 tweets afternoon 31.03.2020.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health Surveill*, 6(2):e19273.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv: 1301.3781*.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. *SIGSPATIAL Special*, 12(1):6–15.

Shane Smith. 2020. Coronavirus (covid19) tweets - early april.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.