Stanceosaurus 2.0 : Classifying Stance Towards Multicultural Misinformation

Code, Data, Paper(s)

Jonathan Zheng*, Anton Lavrouk*, Tarek Naous, Ashutosh Baheti, Ian Ligon, Alan Ritter, Wei Xu

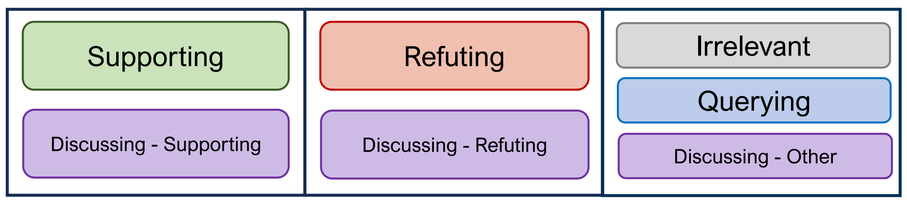
College of Computing – **Georgia Institute of Technology**

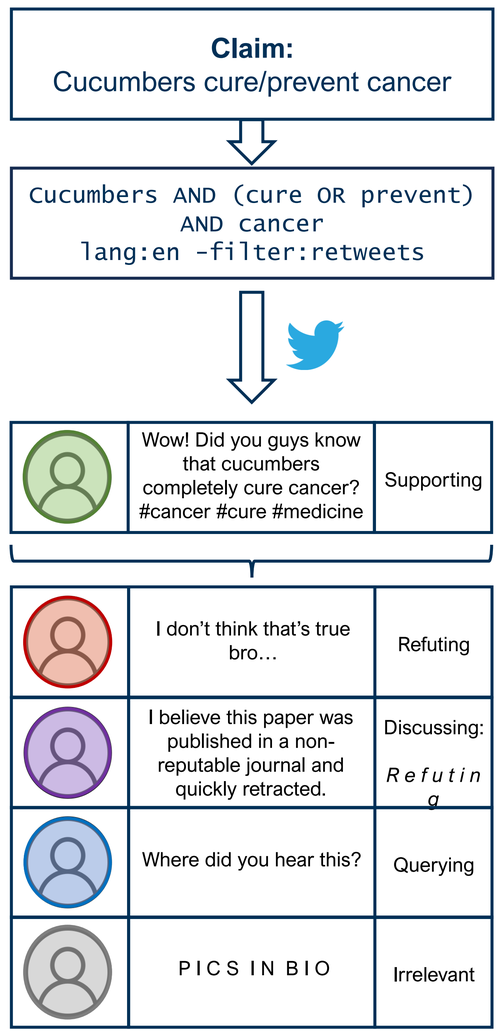
Motivation

- Previous misinformation corpora neglect non-English languages.
- It is important for misinformation datasets to be adaptable to other datasets, so that we may continue expanding our knowledge base.
- The class imbalance issue is highly prevalent in misinformation corpora.

Data Collection

- **Step One: Collecting Misinformation Claims**
  We randomly sample multiple fact checking sites for a diverse set of misinformation claims.
- **Step Two: Writing/Running Queries**
  We write Twitter/X API queries, such that we maximize relevant tweets, while at the same time trying to balance class distribution.
- **Step Three: Annotation**
  We annotate up to 50 tweets and 100 context tweets for each claim. Fine grained annotation using both stance and leaning allows for an interchange between 3 and 5 class (see below).





Corpus Statistics

English	Arabic	Hindi	Spanish	Russian
20,707 tweets	4,009 tweets	3,317 tweets	1,966 tweets	1,907 tweets

Discussing	Irrelevant	Supporting	Refuting	Querying
20,707 tweets	4,009 tweets	3,317 tweets	2,044 tweets	511 tweets

English Stance Detection

Model	Stanceosaurus (unseen claims) Precision	Recall	F1
BERT$_{base}$+ CE	51.1±1.1	50.5±2.0	50.4±1.6
+ weighted CE	50.5±1.9	52.7±1.1	51.3±1.3
+ CB$_{\text{foc}}$	50.6±1.3	55.7±2.1	52.5±1.0
BERT$_{large}$+ CE	54.3±0.8	53.0±0.6	53.6±0.6
+ weighted CE	53.8±1.3	53.8±1.2	53.6±1.0
+ CB$_{\text{foc}}$	53.9±1.2	53.7±1.1	53.6±0.5
BERTweet$_{base}$+ CE	53.1±1.2	52.2±1.6	52.3±1.0
+ weighted CE	51.8±1.0	55.2±1.4	53.1±0.7
+ CB$_{\text{foc}}$	51.3±0.6	56.8±0.6	53.5±0.3
BERTweet$_{large}$+ CE	60.6±2.0	60.2±1.0	60.2±1.1
+ weighted CE	**60.8**±1.6	60.2±1.0	60.2±0.5
+ CB$_{\text{foc}}$	59.8±1.3	**62.8**±1.5	**61.0**±0.8

Evaluation

Claims are split into train, dev and test sets. Models are evaluated on unseen claims.

Class-Imbalance Issue

Class-balanced focal loss achieves better performance than weighted CE.

$$\text{CB}_{\text{foc}}(\hat{s}, y) = -\underbrace{\frac{1-\beta}{1-\beta^{n_y}}}_{\text{reweighting}} \underbrace{\sum_{m \in C} (1-p_m)^{\gamma} \log(p_m)}_{\text{focal loss}}.$$

Russian

Loss	Precision	Recall	F1
CE	53.55±0.8	35.33±0.7	36.15±1.3
Weighted CE	44.38±0.2	42.84±0.5	42.09±0.1
CBFL	45.60±1.5	46.98±2.0	43.94±0.2

Spanish

Loss	Precision	Recall	F1
CE	50.26±1.9	40.86±0.7	41.81±1.0
Weighted CE	54.12±0.4	42.65±0.5	43.75±0.4
CBFL	51.26±2.2	44.15±0.9	43.83±1.0

Hindi

Loss	Precision	Recall	F1
CE	52.1±2.9	39.4±2.0	40.8±2.5
Weighted CE	55.0±4.2	42.4±1.4	44.3±1.8
CBFL	53.0±3.4	44.1±1.7	45.3±1.5

Arabic

Loss	Precision	Recall	F1
CE	44.8±4.0	40.1±2.5	40.0±2.0
Weighted CE	44.1±3.3	40.7±1.6	39.7±1.7
CBFL	46.1±2.6	44.7±1.1	43.1±0.2

Cross-Lingual Transfer

Experiment

We train mBERT on English Stanceosaurus, and evaluate on Russian, Spanish, Hindi, and Arabic.

Results

Each language show's comparable performance to the other, showing potential merits of the data.

Dataset Adaptability

We show that an **EasyAdapt** combined version of **Stanceosaurus** and **RumourEval** scores a higher F1 on both datasets test sets.

Train \ Test	Stanceosaurus Prec.	Rec.	F1	RumourEval Prec.	Rec.	F1
Stanceosaurus	66.9	**67.1**	66.8	44.8	43.8	41.2
RumourEval	39.8	43.6	40.6	**79.6**	59.7	65.7
Combined	66.6	66.0	66.2	61.1	63.4	60.6
EasyAdapt	**68.3**	67.0	**67.4**	74.4	**62.6**	**65.8**

English Performance On Unseen Claims

Fact Check Source	#test	Precision	Recall	F1
AAP Fact Check	452	50.1	39.4	40.6
AFP Fact Check CAN	824	71.5	54.7	58.7
AFP Fact Check NZ	224	64.2	63.7	63.5
Blackdotresearch	516	65.7	62.0	60.4
Factly	447	59.4	68.2	62.5
FullFact	474	57.0	55.4	55.8
Poynter	118	73.2	61.3	63.0
Politifact	614	57.7	53.7	51.8
Snopes	1402	61.4	52.0	54.4
All	5071	62.9	54.3	57.1

Set-up

We test BERTweet's ability to generalize toward regional claims by training on international claims. We create a new train/test/dev set based on claim location.

Results

Results vary wildly between sources. Two sources with the most international data have the highest F1 scores.