

Learning Disentangled Latent Topics for Twitter Rumour Veracity Classification

John Dougrez-Lewis, Elena Kochkina, Maria Liakata,
Yulan He

What is a Rumour?

“An item of circulating information whose veracity status is yet to be verified at the time of posting” - Zubiaga et al. 2018

Why Bother?

Anyone can post rumours on social media, which can pose as news if propagated widely enough.

This is especially problematic when performed by well-connected establishments or individuals.

About two-thirds of Americans obtain news on social media¹.

Examples of dangerous rumours:

- Drinking bleach can cure Covid-19.
- Political disinformation.

[1] <https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/>

Real Consequences



Maryland Emergency Management Agency (MDMEMA) 

@MDMEMA



ALERT : We have received several calls regarding questions about disinfectant use and [#COVID19](#).

This is a reminder that under no circumstances should any disinfectant product be administered into the body through injection, ingestion or any other route.

5:24 PM · Apr 24, 2020



21.3K



See the latest COVID-19 information on Twitter

Research Aims

- Given a Twitter rumour from a *previously unseen* event*, accurately determine its veracity.

Initially Restricted to information from the Twitter thread and its responses.

Then Using background information from the internet.

*Event = real-world happening which causes many rumours to be posted online.

Challenge

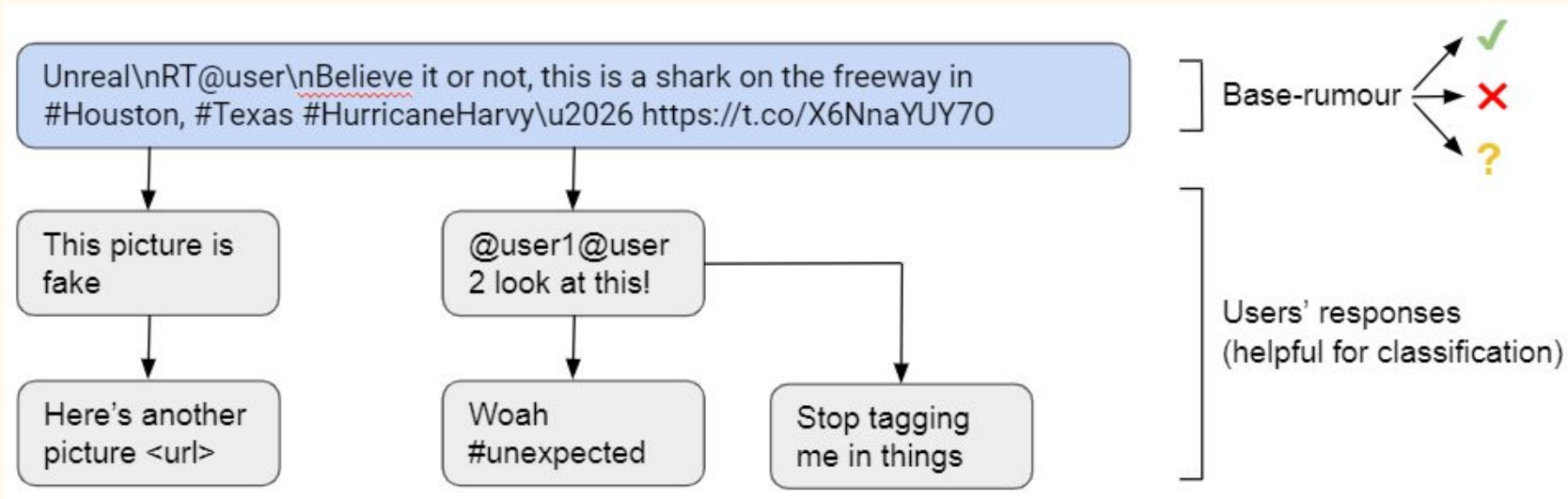
Much existing work focuses on rumours of previously seen events.

We focus on previously *unseen* events, for which there is less literature and results are lower.

Many important constructs of vocabulary and dialect are unique to the rumours of specific events, necessitating the learning of *more general* features for previously unseen events.

Specifically, words can take event-specific meanings in the context of an event due to the author assuming the reader's knowledge about it.

Problem Diagram

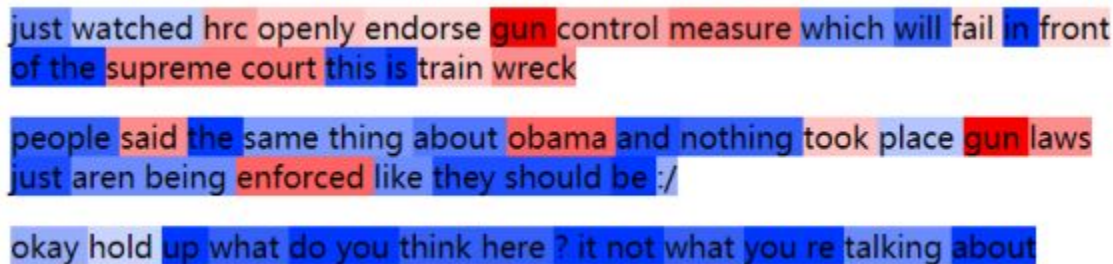


Approach

Disentanglement - separating *what* is being said from *how* it is being said.

Hypothesis - the dialect (how) will be better for prediction than the factual content (what) for previously unseen events.

Example of *what* (red) and *how* (blue) from the original paper on the model¹.



just watched hrc openly endorse gun control measure which will fail in front
of the supreme court this is train wreck

people said the same thing about obama and nothing took place gun laws
just aren being enforced like they should be :/

okay hold up what do you think here ? it not what you re talking about

[1] Zeng et al. 2019: What You Say and How You Say it: Joint Modeling of Topics and Discourse in Microblog Conversations

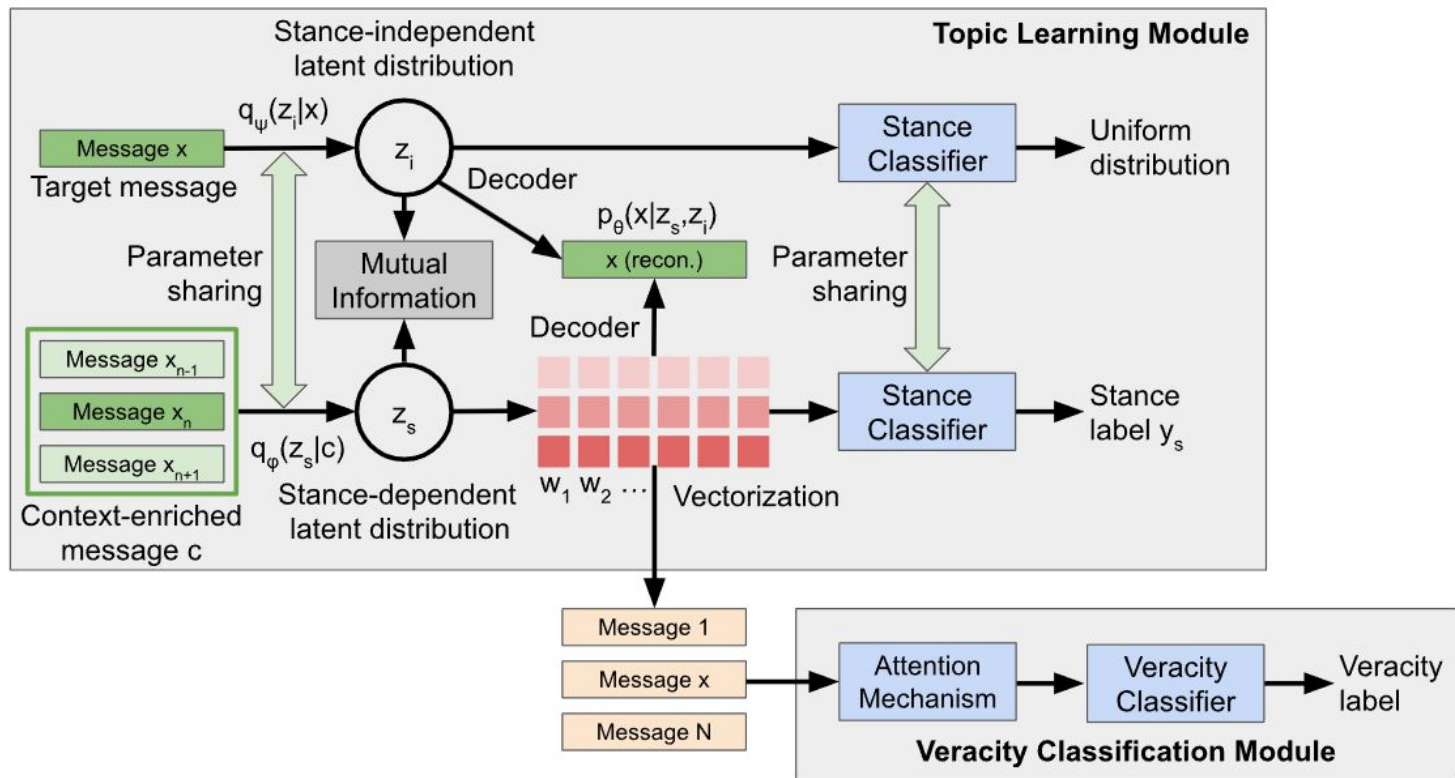
Disentanglement Details

The text is encoded into two latent representations {message, context} used internally by the model with a variational autoencoder.

The model aims to optimize the following:

- Minimize the similarities between *what* and *how*.
- Maximize the reconstruction quality of the original text from both factors together.
- Maximise the similarity of the latent representations of message and context to those of others

Model Diagram



Disentanglement Extension

Since the stances of responses to a rumour are predictive of its veracity¹, a model predicting stances correctly should also perform better on veracity (due to overlap between predictors of the two).

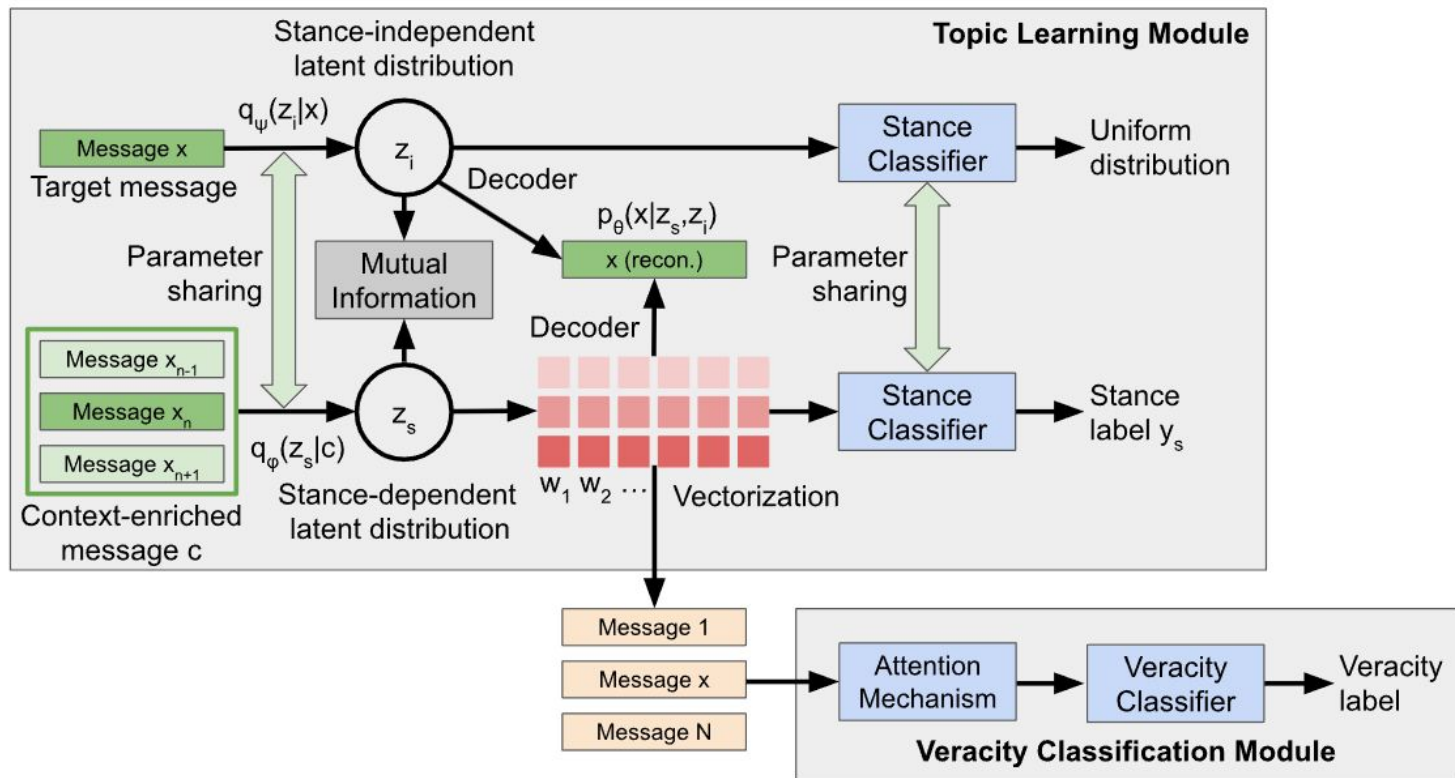
Thus we use the previously generated latent representations of message and context to predict stance.

We can also set one of the two latent representations {message, context} to be *not* predictive of stance, instead aiming to predict a uniform distribution.

This technique can also be used with veracity in place of stance for the context.

[1] Dungs et al. 2018: Can rumour stance alone predict veracity?

Model Diagram



Results and Conclusion

Components Used	MacroF1
Stance-independent	0.434
Stance-dependent	0.375
Both together	0.395

Contrary to the initial hypothesis, the *what* (factual content) outperforms the *how* (mannerisms in Twitter thread). This is suggestive of more factual overlap between independent events than expected.

We achieve SOTA results for Accuracy and both True and Unverified classes.

Model	False	True	Unverified	Accuracy	MacroF1
Kochkina et al. (2018)	0.212	0.647	0.330	0.492	0.396
Cheng et al. (2020)	0.504	0.480	0.465	0.521	0.484
Li et al. (2019b)	-	-	-	0.483	0.418
Simple Baseline	0.201	0.413	0.407	0.395	0.339
BERT Baseline	0.113	0.592	0.326	0.405	0.345
Dual Independent	0.161	0.578	0.352	0.445	0.361
Disentanglement	0.164	0.642	0.531	0.528	0.434

External Evidence - Google Search Queries

Preprocessed

The base tweet, with a few tweaks.

Shortened with StanfordNLP

The tweet is parsed, and some desired structural components are kept.

Shortened with ClausIE

The tweet is broken into subject-predicate-object triples, kept in-place.

Search Strategies

Original Rumour:	<i>MORE: Massacre suspects believed to have taken hostage and holed up in small industrial town northeast of Paris: <url> #CharlieHebdo</i>
Query Strategy	Query Text
Preprocessed	before:2015-01-09 MORE : Massacre suspects believed to have taken hostage and holed up in small industrial town northeast of Paris :
StanfordNLP	before:2015-01-09 (Charlie Hebdo) Massacre suspects small industrial town northeast
ClausIE	before:2015-01-09 (Charlie Hebdo) Massacre suspects believed to have taken hostage holed up in small industrial town northeast of Paris

Preprocessed > ClausIE > StanfordNLP

Key takeaway: Search works best nowadays if stopwords and grammatical constructs are kept

Effectiveness of Evidence

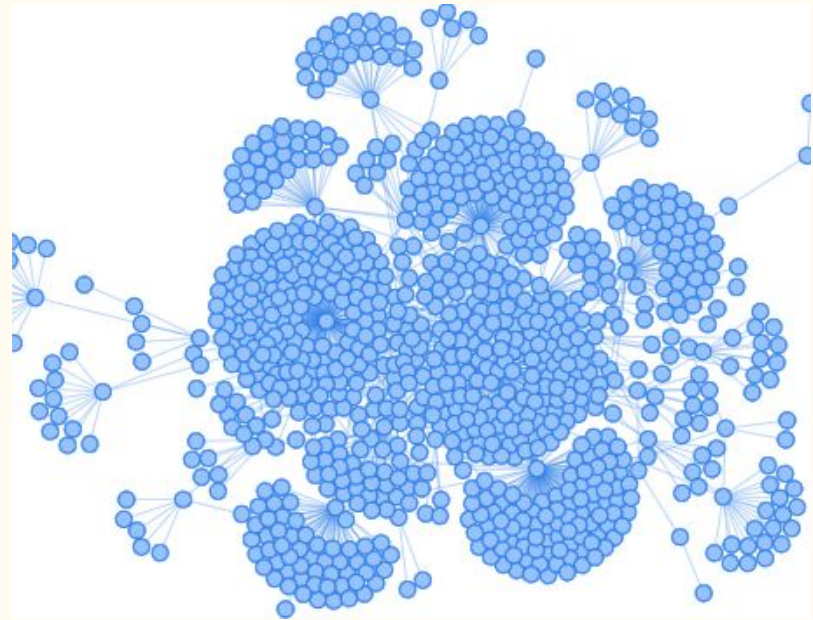
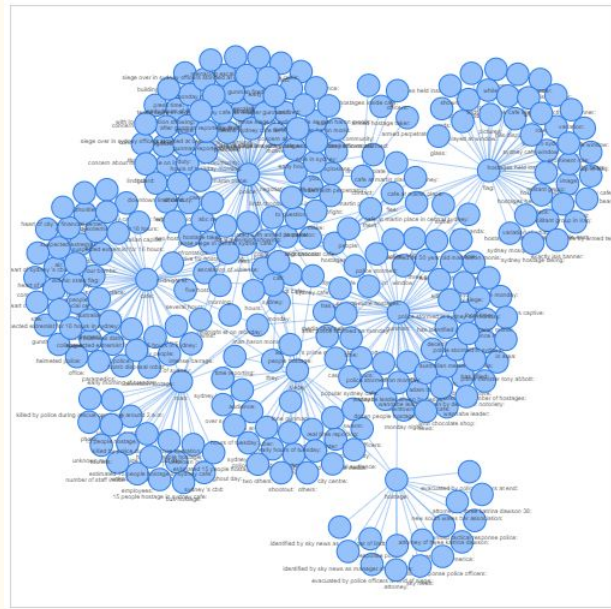
Rumour + Evidence consistently and substantially outperforms either Rumour or Evidence alone.

BERT	Ch	Fe	Ge	Ot	Sy	False	True	Unv	MacroF1
Rumour + Ev.	0.317	0.174	0.213	0.406	0.318	0.221	0.549	0.265	0.345
Rumour	0.306	0.134	0.315	0.345	0.320	0.209	0.562	0.242	0.338
Evidence	0.268	0.045	0.264	0.370	0.307	0.140	0.645	0.099	0.295
RoBERTa									
Rumour + Ev.	0.306	0.183	0.383	0.368	0.347	0.384	0.600	0.279	0.421
Rumour	0.290	0.113	0.260	0.420	0.309	0.211	0.549	0.232	0.331
Evidence	0.288	0.028	0.252	0.335	0.327	0.145	0.611	0.144	0.301
NLI-SAN									
Rumour + Ev.	0.354	0.256	0.365	0.591	0.458	0.186	0.480	0.250	0.405

The evidence we have retrieved is indeed highly useful!

(It tends to come from highly reputable sources, provided by Google)

Knowledge Graphs



Questions