"A Little Birdie Told Me ... " - Inductive Biases for Rumour Stance Detection on Social Media

Karthik Radhakrishnan* **Tushar Kanakagiri*** **Sharanya Chakravarthy***
Vidhisha Balachandran
Language Technologies Institute
Carnegie Mellon University
{kradhak2, tkanakag, sharanyc, vbalacha}@cs.cmu.edu

Abstract

The rise in the usage of social media has placed it in a central position for news dissemination and consumption. This greatly increases the potential for proliferation of rumours and misinformation. In an effort to mitigate the spread of rumours, we tackle the related task of identifying the stance (Support, Deny, Query, Comment) of a social media post. Unlike previous works (Fajcik et al., 2019; Yang et al., 2019), we impose inductive biases that capture platform specific user behavior. These biases, coupled with social media fine-tuning of BERT allow for better language understanding, thus yielding an F_1 score of 58.7 on the SemEval 2019 task on rumour stance detection.

1 Introduction

Social media has seen an exponential growth, replacing traditional news sources as the primary news source. The apparent value of interesting truth-like news and the ease of access to such news jointly make social media a hotbed for rumours, misinformation and fake news. In the absence of an authority to verify or debunk a rumour, social media users often share their own thoughts on its veracity, creating a collaborative inter-subjective sense-making to determine the veracity of the rumour. Hence, an important step in achieving the objective of veracity detection is tracking how other users opine on the accuracy of the rumourous story (Zubiaga et al., 2018).

The content on these social media platforms vary in their topics, style, sentiments, and structure (Manikonda et al.). Reddit, for example, is used for gathering a comprehensive view of opinions from users in a short period of time. On the other hand, an event on Twitter is alive for a longer duration and is used for following the development and evolution of an event (Priya et al., 2019).

It is also important to effectively utilize the context surrounding a particular tweet and model the exchanges in a conversation as they often contain crucial background information.

Given the extensive usage of sarcasm and rhetoric in expressing opinions (Carvalho et al.), understanding 'social media' style of text is also essential for effective stance identification.

In this work, we impose inductive biases accounting for the underlying social media platform, conversational context, and noisy social media style of text to improve on the task of rumour detection. To the best extent of our knowledge, this is the first work that applies inductive biases inspired from a deep analysis of communities and their usage patterns to the task of stance identification. We achieve a Macro F_1 score of 58.7 on the 2019 SemEval RumourEval task with novel techniques that surpass state of the art models (non-ensemble) by 2 F_1. The code for our approaches will be made available on GitHub[1].

2 Task Definition

To judge the veracity of a social media post, it is useful to analyze the surrounding discourse (comments/replies) by other users. The discourse is initiated by a SOURCE post and followed by tree-structured threads. Each post in a thread is made in response to a PARENT post that immediately precedes it. This problem was modeled as a SemEval shared task - RUMOUREVAL (Gorrell et al., 2019), consisting of two subtasks.

A. Stance Classification - Given a source post introducing a rumour and the ensuing conversation thread, classify the source and each post in the thread into one of 4 categories.

- **SUPPORT** : The author of the response supports the veracity of the rumour.

*Equal contribution

[1] https://github.com/sharanyarc96/SocialMediaRumorStanceDetection

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 244–248

- **DENY** : The author of the response denies the veracity of the rumour.
- **QUERY** : The author of the response asks for additional evidence in relation to the veracity of the rumour.
- **COMMENT** : The author of the response makes their own comment without a clear contribution to assessing the veracity of the rumour.

B. Veracity Prediction - Classify the rumour as TRUE, UNVERIFIED, or FALSE.

In this work, we focus on the Stance Classification (highlighted in Appendix A). The heavy class imbalance (highlighted in Appendix B) coupled with the low inter-annotator agreement ($\sim$63 %) (Derczynski et al., 2017) makes this a challenging task.

3 Related Work

This section highlights prior work on the RumourEval dataset. Table 1 provides a short summary of each of the models analysed below.

Model	Description
BranchLSTM (Kochkina et al., 2017)	LSTM-based stance prediction using tweet branches
BLCU (Yang et al., 2019)	Inference-chain based GPT with word and tweet features
BUT-FIT (Fajcik et al., 2019)	BERT ensemble for stance classification, without hand-crafted features
EventAI (Li et al., 2019)	Ensemble of ML and Rule-based models with extensive feature engineering

Table 1: Related Work

3.1 Feature Engineering

EventAI, BranchLSTM and BLCU employed extensive feature engineering as described below.

- **Lexicon Based:** BranchLSTM utilized the count of negation words and swear words. BLCU made more extensive use of lexicons and looked for the presence of positive and negative words, swear words, query words and different classes of verbs.
- **Relation To Other Posts:** BranchLSTM and EventAI used cosine similarity between source and target embeddings. BLCU used the depth of the post in the thread.
- **Content Based:** BLCU checked for the presence of punctuations, hashtags, URLs and "RT". EventAI used similar features along with mentions of special accounts and hashtags (@cnn, #fakenews etc).
- **Tweet Role:** BranchLSTM and EventAI had features indicating whether the tweet was a source or a reply.
- **Tweet and User Metadata:** BLCU used tweet and user features such as favorite and retweet counts, follower and friend counts etc.

3.2 Pre-training

Since unsupervised pre-training for word representations has demonstrated success on a large variety of NLP tasks, the top performing models use pre-trained contextual word representations. BUT-FIT uses BERT (Devlin et al., 2019), BLCU uses GPT (Radford, 2018) and CLEARumor (Baris et al., 2019) uses ELMo (Peters et al., 2018). However, since none of these models are trained on Twitter/Reddit data, fine-tuning on social media data might help capture its idiosyncrasies such as usage of emoticons, opinion-centric text as opposed to fact-centric text, shorter sentences etc.

3.2.1 Intra-thread context

BranchLSTM and EventAI used similarity of the target post with other parts of the thread as features. Additionally, BranchLSTM treated a conversation thread as a set of linear branches. They defined a branch as a chain of tweets that included a leaf post and all its parents all the way to the source post.

BLCU utilized the entire conversation thread by concatenating it with the target post.

BUT-FIT made the assumption that the stance of the target post depends only on itself, the source post, and the previous post in the thread.

4 System Description

Our system (Figure 1) utilizes the content from the SOURCE and PARENT tweets as additional context following previous work (Fajcik et al.,2019;Yang et al.,2019) which noted that the above two tweets mostly contain sufficient information to classify a TARGET tweet correctly. In this work, we leverage various inductive biases and propose late fusion in §4.1, social media fine-tuning to better leverage BERT in §4.2, discrimination between social

media platforms in §4.3, domain-specific features over generic textual features in §4.4, and transition priors to better capture conversation dynamics in §4.5.

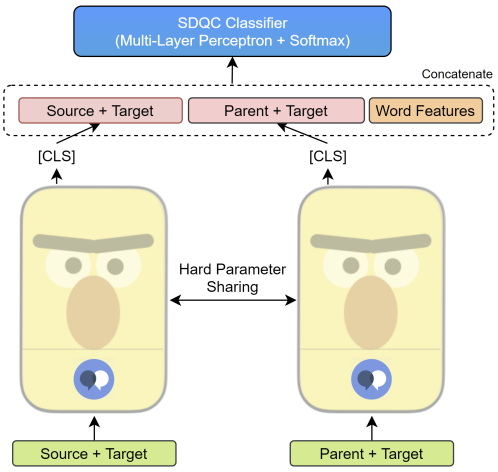

Figure 1: Model Architecture showcasing Late Fusion of SOURCE + TARGET and PARENT + TARGET representations along with additional features. BERT Input is of the form - [CLS] CONTEXT [SEP] TARGET. SOURCE is the post initiating the conversation. The TARGET post is made in response to the PARENT.

4.1 Late Fusion

We attend over the SOURCE and PARENT tweets separately followed by a late fusion of their [CLS] representations. This architecture allows for the TARGET tweets to independently attend over the SOURCE and the PARENT tweet, ensuring the capture of complementary information. This avoids the dilution of context that occurs through the combination of SOURCE and PARENT as context.

4.2 Understanding conversational constructs

One pattern of errors exhibited by the previous models is due to their lack of ability to understand the 'social media' style of text. Common conversational constructs like sarcasm / rhetoric (usually intended to attack / refute someone as opposed to being a genuine question seeking more information) were wrongly labelled as QUERY due to the text containing symbols like "?" or interrogative words.

To combat this, we use Conversational BERT [2] which is trained on social media and dialogue data. We further fine-tune this model on tweets from the RumourEval dataset and the larger PHEME dataset to incorporate additional background knowledge about rumourous tweets.

4.3 Domain Separation

Our rumours originate from two different social media domains - Twitter and Reddit. Though all prior work has trained models on a combination of data from both sources, we argue for domain separation owing to the fundamental differences in the type of content and interactions on these platforms.

Twitter rumours are concerned with breaking news (Charlie Hebdo shooting, Ferguson unrest) while Reddit rumours are around long-standing conspiracy theories (Flat earth, benefits of Nicotine etc). Reddit discussion threads are shorter and converge sooner i.e. it takes fewer replies to collect the required information. But in case of Twitter, obtaining information that resolves a rumour is a more continuous and a longer process (Priya et al., 2019).

Hence, we train separate models on the Twitter and Reddit data and later aggregate the results.

4.4 Additional Features

Prior work has experimented with inclusion of lexical, sentiment, and emotional features but report little to no improvement (likely because BERT already captures these features). We instead run a TF-IDF vectorizer to extract most discriminatory features for each class and use a subset along with BERT. It also worth noting that these features varied between Twitter and Reddit, further corroborating our hypothesis for domain separation.

4.5 Incorporating a prior

Upon deeper analysis of our training data, we observed that social media conversations tend to follow certain patterns - a QUERY stance is less likely to follow another QUERY stance (questions are usually followed by answers) while SUPPORT stance is highly likely to follow another SUPPORT stance (users espousing the same opinion). We incorporate this inductive bias via a post-processing module where we linearly interpolate the confidence scores from our model and the prior.

5 Experimental Setup

We use the HuggingFace Transformers library [3] to fine-tune BERT_{base} on the Sequence Classification task. We use the Adam optimizer (Kingma and Ba,

[2] docs.deeppavlov.ai/en/master/features/models/bert.html

[3] https://github.com/huggingface/transformers

ID	Example	Prediction	Comments
1	So multiple doctors don't count but only Hillary's do? Do you even understand her conditions?	SUPPORT	Conversational Pre-training helps differentiate between genuine queries and rhetorical questions
2	Necessary precaution? Isn't it better to close shop for some hours than to risk lives?	COMMENT	It isn't obvious that this reply is rhetorical, but since its parent was tagged as QUERY and queries don't normally follow each other, prior guides the model (Before priors, scores - 0.35 QUERY & 0.33 COMMENT)
3	**Tweet 1** - "WERE YOU THERE THOUGH" **Tweet 2** - "Your mind just can't fathom that can it?"	COMMENT	The gold label is QUERY though these questions are rhetorical
4	**Source** -"At least 10 killed in shooting" **Tweet 1** - "11 Killed now" *in reply to* **Source** **Tweet 2** - "11 Killed" *in reply to* **Source**	-	Gold labels are different {Tweet 1: SUPPORT, Tweet 2: DENY} though the texts have the same meaning

Table 2: Qualitative examples from our model

2014), with a learning rate of 1.5e-6 and batch size of 32 and train on an NVIDIA Tesla T4 GPU.

6 Results and Error Analysis

Our approach achieves an F_1 of 58.7, outperforming non-ensemble approaches by 2 F_1 as shown in Table 3. Though ensembles from BUT-FIT and BLCU achieve a higher F_1 score, we do not ensemble our model owing to high computational cost for training and inference (for ex. BUT-FIT ensembles over 100 BERT$_{large}$ models). We report our best and average (over 5 random seeds) on the RumourEval 2019 dev dataset. Table 2 shows some qualitative examples from our model.

Model	Macro-F1
BUT-FIT BERT$_{base}$ (Average)	51.4
BranchLSTM	49.3
BUT-FIT BERT$_{large}$ (Average)	56.2
BLCU (Best Reported)	56.6
Ours (Average)	56.7
Ours (Best)	**58.7**

Table 3: Comparison with state of the art models Ensembles from BUT-FIT and BLCU produce scores that are higher than those presented here. We show results of comparable non-ensemble versions of state of the art models.

6.1 Ablation Study

In this section, we analyze individual components of our contribution and report incremental improvements in Table 4.

Model	Macro-F1
Base Model	51.2
+ Conversational Pre-Training	53.7 **(+2.5)**
+ TF-IDF features	55.2 **(+1.5)**
+ Domain Sep. and Late Fusion	56.4 **(+1.2)**
+ Transition Priors	58.7 **(+2.3)**

Table 4: Effect of each of our inductive biases

Conversational Pre-Training allows the model to correctly interpret social media constructs like sarcasm, rhetoric (Table 2, Ex. 1) and yields a boost of 2.5 F_1.

TF-IDF features improve the score by 1.5 F_1 by biasing the model based on frequently used words/phrases for each stance.

Domain Separation and Late Fusion provide further gains, increasing the F_1 by 1.2. In addition to improving the score, domain separation is also essential for using TF-IDF features and prior as they are platform dependant.

Transition Priors increase the performance by 2.3 F_1 by guiding the prediction based on stance transition priors in cases where the model makes uncertain predictions (Table 2, Ex. 2).

7 Unsolvable Examples

The RumourEval dataset contains examples with noisy annotations (Table 2, Ex. 3) where the ground truth is mislabeled, thus penalizing our model for correct predictions. Additionally, few examples which have the same hierarchy and similar text (Table 2, Ex. 4) are assigned different labels (Possibly due to different interpretations among annotators) resulting in noisy training examples.

Another class of unsolvable examples stemmed from deleted tweets. If a particular tweet was deleted, the dataset attaches its children to their GRANDPARENT tweet. This presents issues as the children express opinions towards a deleted tweet. A potential solution would be to remove tweets where the '@' mention is towards an unseen author but we would risk further reducing the small number of training examples in our dataset.

8 Conclusion and Future Work

In this work, we showcased the efficacy of inductive biases to the task of stance classification and achieved a score of 58.7 F_1, surpassing existing approaches. We hope to utilize this model in other downstream tasks like veracity detection (Task B) and expand our inductive biases to other social media tasks such as fact verification and conversation derailment detection.

References

Ipek Baris, Lukas Schmelzeisen, and Steffen Staab. 2019. CLEARumor at SemEval-2019 Task 7: ConvoLving ELMo Against Rumors. *arXiv preprint arXiv:1904.03084*.

Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. Clues for Detecting Irony in User-Generated Contents: Oh...!! It's "so Easy";-), year = 2009, isbn = 9781605588056, publisher = Association for Computing Machinery, address = New York, NY, USA, url = https://doi.org/10.1145/1651461.1651471,.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. 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 *NAACL-HLT*.

Martin Fajcik, Lukáš Burget, and Pavel Smrz. 2019. BUT-FIT at SemEval-2019 Task 7: Determining the Rumour Stance with Pre-Trained Deep Bidirectional Transformers. *arXiv preprint arXiv:1902.10126*.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. *arXiv preprint arXiv:1704.07221*.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. eventAI at SemEval-2019 Task 7: Rumor Detection on Social Media by Exploiting Content, User Credibility and Propagation Information. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 855–859.

Lydia Manikonda, Ghazaleh Beigi, Huan Liu, and Subbarao Kambhampati. Twitter for Sparking a Movement, Reddit for Sharing the Moment: #metoo.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *ArXiv*, abs/1802.05365.

Shalini Priya, Ryan Sequeira, Joydeep Chandra, and Sourav Kumar Dandapat. 2019. Where should one get news updates: Twitter or reddit. *Online Social Networks and Media*, 9:17–29.

Alec Radford. 2018. Improving Language Understanding by Generative Pre-Training.

Ruoyao Yang, Wanying Xie, Chunhua Liu, and Dong Yu. 2019. BLCU_NLP at SemEval-2019 Task 7: An Inference Chain-based GPT Model for Rumour Evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1090–1096.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Comput. Surv.*, 51:32:1–32:36.