Gradual Argumentation Evaluation for Stance Aggregation in Automated Fake News Detection

Imperial College London

Neema Kotonya & Francesca Toni | Department of Computing, Imperial College London

{n.kotonya18, f.toni}@imperial.ac.uk

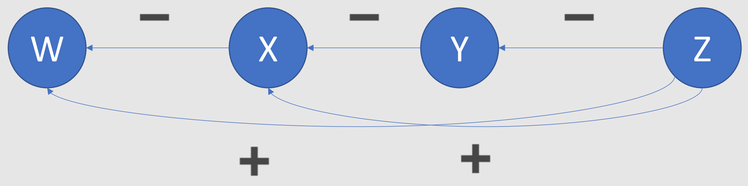
INTRODUCTION

Fake news has existed for as long as news has been in circulation. However, the automated detection of fake news has recently become a hot-button topic. To our knowledge, this work is the first attempt to apply argumentation to the problem of fake news detection.

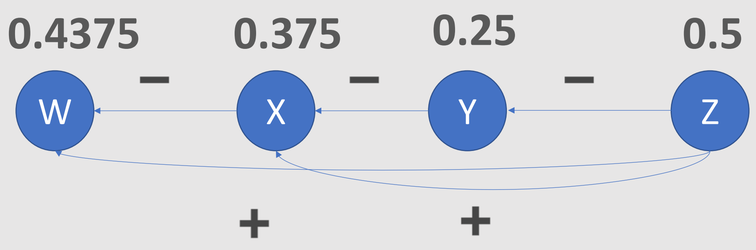
BACKGROUND

An abstract argumentation framework is a pair $\langle \mathcal{A}, \mathcal{R} \rangle$ consisting of arguments $\mathcal{A}$ and binary attack relation $\mathcal{R}$ over these arguments.

Bipolar Argumentation Frameworks (BAFs) incorporate support relations as well as attack relations.



Discontinuity-Free Quantitative Argumentation Debates (DF-QuAD) is a method for evaluating the strength of an argument according to the dialectical strength of its attackers and supporters.



METHOD

u1/source tweet: *Up to 20 held hostage in Sydney Lindt Cafe siege ⟨URL⟩ ⟨URL⟩* [SUPPORT]

—**u2/reply 1**: *"@u1: Up to 20 held hostage in Sydney Lindt Cafe siege ⟨URL⟩ ⟨URL⟩."* [SUPPORT]

—**u3/reply 2**: *Sick. "@u1: Up to 20 held hostage in Sydney Lindt Cafe siege ⟨URL⟩ ⟨URL⟩"* [SUPPORT]

—**u4/reply 3**: **@u1 @u10** *oh god !!!!* [COMMENT]

—**u5/reply 4**: **@u1** *at least they've got good chocolate* [COMMENT]

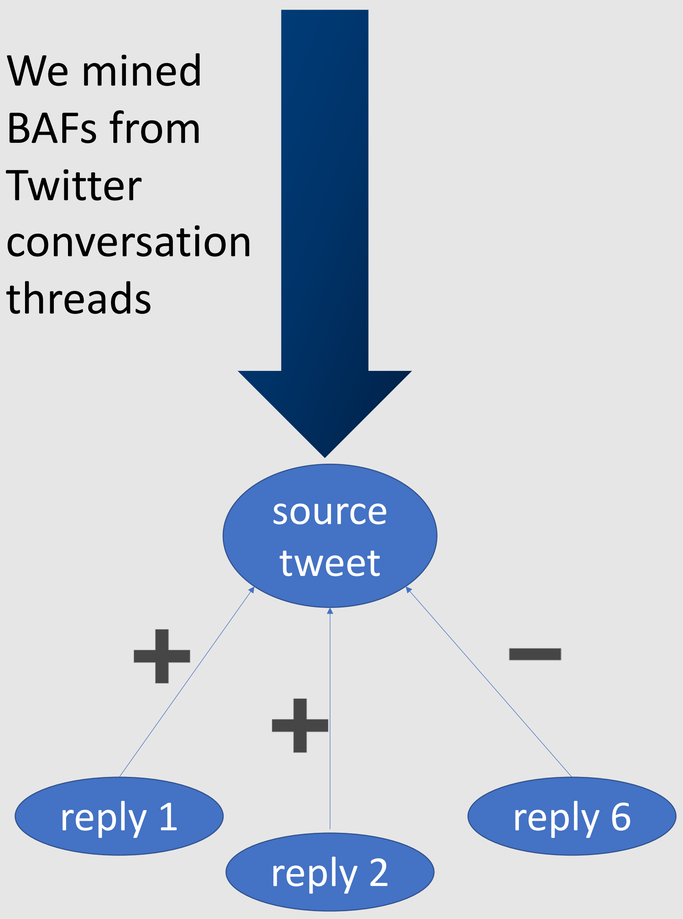
—**u6/reply 5**: **@u5** *you are an insensitive idiot!* [COMMENT]

—**u7/reply 6**: **@u1** *all reports say 13* [DENY]

—**u8/reply 7**: *"@u1: Up to 20 held hostage in Sydney Lindt Cafe siege ⟨URL⟩ ⟨URL⟩" - wonder if they'll get paid overtime* [COMMENT]

—**u9/reply 8**: *"@u1: Up to 20 held hostage in Sydney Lindt Cafe siege ⟨URL⟩ ⟨URL⟩" - Oh. My. God. I am SICK!* [COMMENT]

We mined BAFs from Twitter conversation threads



We employed DF-QuAD to evaluate the strength of source tweets.

We used the computed strength value predict the veracity of source tweets.

$$v(C) = \begin{cases} true\ if\ \sigma(C) > 0.5 \\ false\ if \sigma(C) \leq 0.5 \end{cases}$$

RESULTS

We leveraged a DF-QuAD-based evaluation method to predict the veracity (true/false) claims presented in the RumourEval tweets.

Dataset	Model	Agree			Disagree			Discuss		
		P	R	F1	P	R	F1	P	R	F1
FNC-1	GB	**.831**	.736	.781	.570	.322	.412	.926	**.972**	.934
	GRU	.645	.685	.665	.402	.244	.304	.876	.887	.882
	LSTM	.817	**.878**	**.846**	.652	**.493**	.562	**.964**	.955	**.960**
	BiLSTM	.829	.840	.835	**.676**	**.493**	**.570**	.949	.965	.957
Rumour Eval	LSTM	.166	**.490**	.248	**.160**	.0119	.0222	.753	.513	.610
	BiLSTM	**.178**	.430	**.252**	.105	**.0448**	**.0628**	**.759**	**.576**	**.655**

Precision (P), recall (R), and F1-score (F1) of stance detection classifiers on FNC1 test set and RumourEval.

	Stance aggregation method	Veracity Assessment (RumourEval Task B)					
		False			True		
		P	R	F1	P	R	F1
Gold standard labels (RumourEval Task A)	Credibility-weighted average	.581	.383	.462	.743	**.866**	.800
	DF-QuAD (DR)	**.625**	**.532**	**.575**	**.789**	.845	**.816**
	DF-QuAD (DR + NR)	.615	.511	.558	.781	.845	.811
LSTM stance detection labels	Credibility-weighted average	**.750**	.079	.143	.746	**.990**	**.851**
	DF-QuAD (DR)	.667	**.105**	**.182**	**.750**	.981	.850
	DF-QuAD (DR + NR)	.667	**.105**	**.182**	**.750**	.981	.850
Bidirectional LSTM stance detection labels	Credibility-weighted average	.400	.050	.089	.719	**.970**	.826
	DF-QuAD (DR)	**.500**	**.075**	**.130**	**.724**	**.970**	**.829**
	DF-QuAD (DR + NR)	**.500**	**.075**	**.130**	**.724**	**.970**	**.829**

Precision (P), recall (R), and F1-score (F1) of stance aggregation methods when applied to gold standard labels and the stance labels predicted by LSTM and BiLSTM trained stance classifiers

REFERENCES

Neema Kotonya and Francesca Toni. *Gradual Argumentation Evaluation for Stance Aggregation in Automated Fake News Detection.* Proceedings of the 6th Workshop on Argument Mining. 2019.