A Appendix

A.1 Evidence extraction classification problem: baselines and agreement bound

model	recall m	precis. m	F1 m
agreement bound	0.769	0.725	0.746
random baseline	0.500	0.500	0.500
majority vote	0.343	0.500	0.407

Table 10: Evidence extraction classification problem: baselines and agreement bound (m = macro)

A.2 Error analysis

A.2.1 Stance detection

Below we give an instance of a misclassified ETS. Even though the ETS supports the claim, the lexical overlap is relatively low. Most likely, for this reason, the model predicts *refute*.

Example: —————————————

Claim: *The Reuters news agency has proscribed the use of the word 'terrorists' to describe those who pulled off the September 11 terrorist attacks on America.*

ETS: *Reuters' approach doesn't sit well with some journalists, who say it amounts to self-censorship. "Journalism should be about telling the truth. And when you don't call this a terrorist attack, you're not telling the truth," says Rich Noyes, director of media analysis at the conservative Media Research Center. ...*

—————————————

model \ gold	support	refute	no stance
support	**472**	**86**	175
refute	41	80	51
no stance	141	74	**531**

Table 11: Stance detection confusion matrix (`AtheneMLP`)

A.2.2 Evidence extraction

The model wrongly predicts sentences when the topic of the sentences is similar to the topic of the claim, but the sentence is not relevant for the validation of the claim:

Example: —————————————

Claim: *The Department of Homeland Security uncovered a terrorist plot to attack Black Friday shoppers in several locations.*

FGE: *Bhakkar Fatwa is a small, relatively unknown group of Islamic militants and fanatics that originated in Bhakkar Pakistan as the central leadership of Al Qaeda disintegrated under the pressures of U.S. military operations in Afghanistan and drone strikes conducted around the world.* —————————————

A.2.3 Claim validation

The FGE are contradicting and the classifier predicts *refuted* instead of *supported*.

Example: —————————————

Gold standard: *supported*; Prediction: *refuted*

Claim: *As a teenager, U.S. Secretary of State Colin Powell learned to speak Yiddish while working in a Jewish-owned baby equipment store.*

FGE: *As a boy whose friends and employers at the furniture store were Jewish, Powell picked up a smattering of Yiddish. He kept working at Sickser's through his teens, ... picking up a smattering of Yiddish ... A spokesman for Mr. Powell said he hadn't heard about the spoof ...*

—————————————

model \ gold	supported	refuted	NEI
supported	36	26	13
refuted	**38**	**203**	**53**
NEI	18	42	27

Table 12: Confusion matrix for claim validation `BertEmb` (NEI: not enough information)