Mid Presentation

CSE 712
Symbolic Machine Learning - II

Group 3

Members:

Auninda Alam-21166050
Marjan Tahreen-21166049
Shohag Rana-21366015

INTRODUCTION

Detecting Attackable Sentences in Arguments

Yohan Jo, Seojin Bang, Emaad Manzoor, Eduard Hovy, Chris Reed

CONTRIBUTIONS

Introduced the problem of detecting attackable sentences in arguments

Analyzed driving reasons for attacks in arguments and the effects of sentence characteristics

The performance of machine learning models for detecting attackable sentences

LITERATURE REVIEW

Aristotle (2007) suggested three aspects of argument persuasiveness.

Wachsmuth et al. (2017b) summarized various aspects of argument quality studied in argumentation theory and NLP

Some research took empirical approaches and collected argument evaluation criteria from human evaluators (Habernal and Gurevych, 2016a; Wachsmuth et al., 2017a)

Some studies aimed to model the salience of individual sentences in attacked arguments (Jo et al., 2018; Ji et al., 2018)

Dataset

Dataset		Train	Val	Test
Attacked	#posts	25,839	8,763	8,558
	#sentences	420,545	133,090	134,375
	#attacked	119,254	40,163	40,354
Successful	#posts	3,785	1,235	1,064
	#sentences	66,628	20,240	17,129
	#successful	8,746	2,718	2,288

Table 1: Data statistics. "Attacked" contains posts with at least one attacked sentence. "Successful" contains posts with at least one successfully attacked sentence.

F1	Personal opinion (28%)
F2	Invalid hypothetical (26%)
F3	Invalid generalization (13%)
F4	No evidence (11%)
F5	Absolute statement (7%)
F6	Concession (5%)
F7	Restrictive qualifier (5%)
F8	Other (5%)

(b) Motivating factors for attacks.

Source

The Dataset was formed using the online discussions from the "Change My View (CMV)" subreddit.

Labelling

Each sentence in a post was labelled into three categories, i.e. successfully attacked, un-successfully attacked and unattacked.

Feature Extraction

- content
- external knowledge
- proposition types
- tone

Model

	Attacked			Successful		
	P@1	A@3	AUC	P@1	A@3	AUC
Random	35.9	66.0	50.1	18.9	45.0	50.1
Length	42.9	73.7	54.5	22.3	52.1	55.7
LR	47.1	76.2	61.7	24.2	54.5	59.3
(×) Content	45.2	74.4	58.1	24.0	52.6	57.0
(×) Knowledge	47.0	76.0	61.7	24.1	54.3	59.0
(×) Prop Type	46.7	75.9	61.5	24.4	53.6	59.0
(×) Tone	47.0	76.0	61.9	25.2	56.2	59.4
BERT	49.6	77.8	64.4	28.3	57.2	62.0
Humans†	51.7	80.1	–	27.8	54.2	–

Problem Formulation

- P@1
- A@3
- AUC

ML Models

- Logistic Regression
- BERT

Baseline Models

- Random
- Length

Results

Based on the Computational Model run 10 times

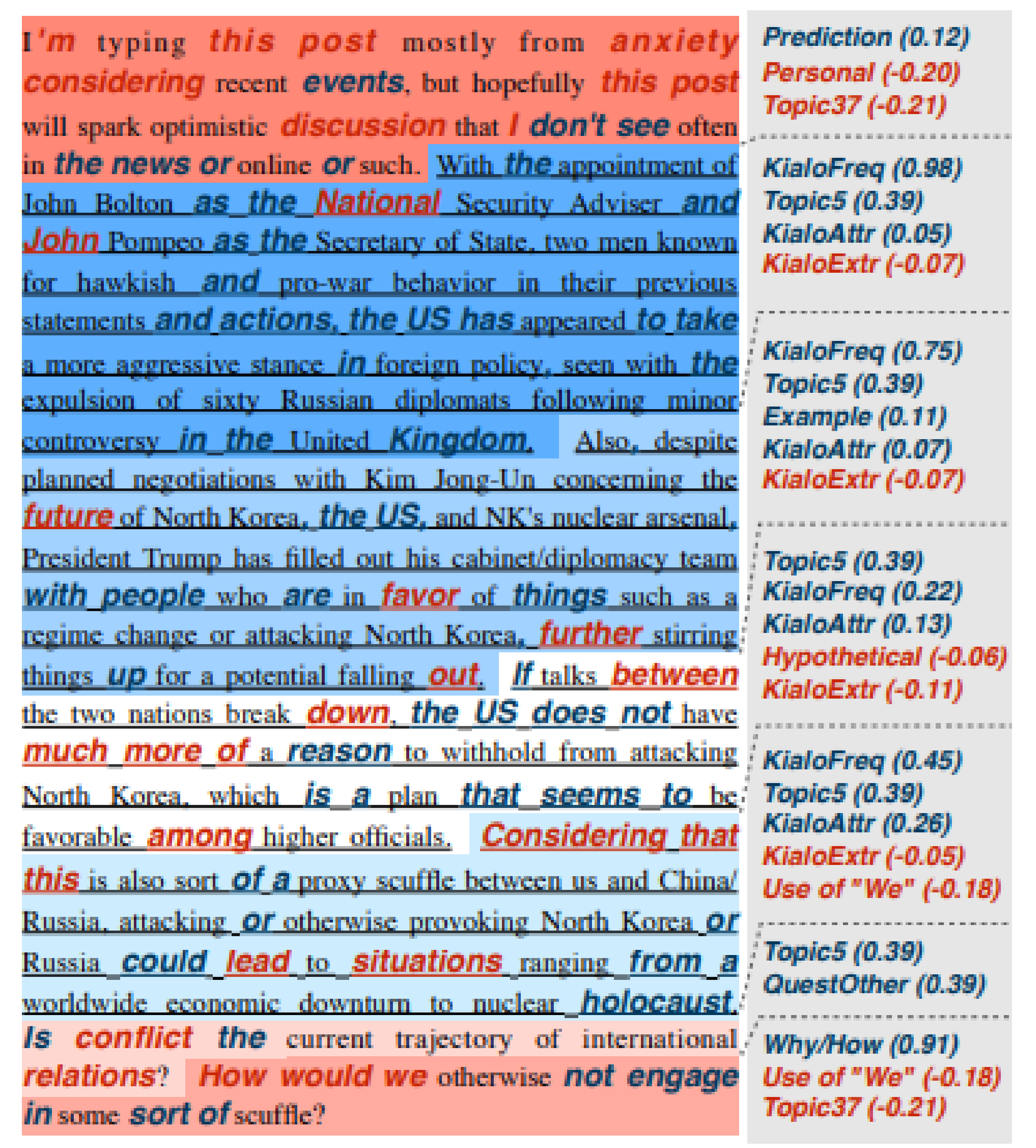

Figure 2: Prediction visualization. Background color indicates predicted attackability (blue: high, red: low). Successfully attacked sentences are underlined. Features with high/low weights are indicated with blue/red.

LR and BERT Outperform

Both the LR and BERT models significantly outperform the baselines, while the BERT model performs best.

	Attacked			Successful		
	P@1	A@3	AUC	P@1	A@3	AUC
Random	35.9	66.0	50.1	18.9	45.0	50.1
Length	42.9	73.7	54.5	22.3	52.1	55.7
LR	47.1	76.2	61.7	24.2	54.5	59.3
(×) Content	45.2	74.4	58.1	24.0	52.6	57.0
(×) Knowledge	47.0	76.0	61.7	24.1	54.3	59.0
(×) Prop Type	46.7	75.9	61.5	24.4	53.6	59.0
(×) Tone	47.0	76.0	61.9	25.2	56.2	59.4
BERT	49.6	77.8	64.4	28.3	57.2	62.0
Humans†	51.7	80.1	–	27.8	54.2	–

Conclusion