Exploring the Usability of Persuasion Techniques for Downstream Misinformation-related Classification Tasks

Nikolaos Nikolaidis[1], Jakub Piskorski[2], Nicolas Stefanovitch[3]

[1]Athens University of Economics and Business, Athens, Greece

[2]Polish Academy of Sciences, Warsaw, Poland

[3]European Commission Joint Research Centre, Ispra, Italy

LREC-COLING 2024

Content

- Persuasion Techniques
- Motivation
- Datasets
- Experiments
- Conclusion

Persuasion techniques

- **Attack on reputation:** of a person/object, e.g., credibility, hypocrisy, "negative" experience and deeds, association to "negative" entity, doubt,
- **Justification:** appeal to value, fear, authority, popularity, etc.,
- **Simplification:** a statement is made that excessively simplifies a problem, usually regarding the cause, the consequences or the existence of choices,
- **Distraction:** changing the focus of the main topic/argument,
- **Call:** encouragement to act or think in a particular way,
- **Manipulative wording:** use of confusing, vague, emotional, non-neutral, unclear language, eggagerating.

Reference: slightly extended taxonomy presented in: (Piskorski et. al, 2023)

Persuasion techniques: Examples

'Fascist' Anti-Vax Riot Sparks COVID Outbreak in Australia.
Name Calling-Labelling

He talks like an EU official! Guilt by Association

Since the Pope said that this aspect of the doctrine is true we should add it to the creed. Appeal to Authority

Because everyone else goes away to college, it must be the right thing to do.
Appeal to Popularity

Lately, there has been a lot of criticism regarding the quality of our product. We've decided to have a new sale in response, so you can buy more at a lower cost!
Introducing Irrelevant Information

Today, women want the vote. Tomorrow, they'll want to be doctors and lawyers, and then combat soldiers. Consequential Oversimplification

How stupid and petty things have become in Washington" Loaded Language

Motivation: Persuasion is everywhere!

Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. Horum omnium fortissimi sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos animos pertinent important, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proeliis cum Germanis contendunt, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. Belgae ab extremis Galliae finibus oriuntur, pertinent

Manipulative Wording: Loaded Language

Attack on Reputation: Smears

Manipulative Wording: Exaggeration

Justification: Appeal to Values

Commissio
PopulusQue
Europaea

So, given a measure of **persuasiveness**:

- **Are Fake News more persuasive?**
- **Is more persuasive content more unreliable/deceiving?**
- **What can we say about a content of an article/source from its Persuasion Technique usage?**

We needed to develop capacity to answer such questions.

Motivation (cont.)

- **(RQ1):** Do Persuasion Techniques detected in texts **exhibit discriminatory power**?
  - and for what kind of tasks?
- **(RQ2):** Which of the **Persuasion Technique-derived features** yields the highest discriminatory power?
- **(RQ3):** What is the **contribution of each Persuasion Technique** to the overall discriminatory power of the specific features?

Datasets

Datasets: inclusion criteria

We set the following requirementes:

- Text should be **Long-form articles** (not tweets/microblogs)
- Text should be **available** (not only URLs)
- Text should **not be heavily distorted** (e.g. noise added for copyright reasons)
- Task should depend **on the text itself** (not metadata)
- Task **not based on an external oracle**/ground-truth

Additionally, we shought to include:

- Corpora with **gold labels** (per document annotation) over those with only silver labels (per domain/source)
- Corpora with **non-English** text

Datasets

Ended up with **8 corpora**:

Dataset	Task	Size	Balance	Language
LOCO	conspiracy theory detection	96,743	32.9%	EN
HND	hyperpartisan news detection	1,273 gold	63%	EN
		754,000 silver	50%	EN
COVID-DISINFO	conspiracy theory	23,509	21%	MULTI
Spanish FNC	fake news detection	971	49.4%	ES
Kaggle FN 2018	fake news detection	20,800	50%	EN
FakeNewsCorpus	fake news detection	96,000	50%	EN
QProp	propaganda detection	51,294	11.2%	EN
FANG-COVID	fake news detection	41,242	32%	DE

Table: The datasets characteristics. The value in the "balance" column indicates the proportion of target class in the dataset.

Experiments

Experiments: Methodology

We want to evaluate the discriminatory power of PT output on each **binary task.**

For each **document**:

1. We run the **Persuasion Technique classifier** (on sentence-level).
2. Based on the output, we compute several **PT features**.

Then, for each **feature:**

1. We calculate the **normalized frequency histograms** of each feature.
2. For each histogram, we compute the discrimination **metrics**.

Experiments: Features

Given a text with PT output, we wish to "pool" the label information in a single value.

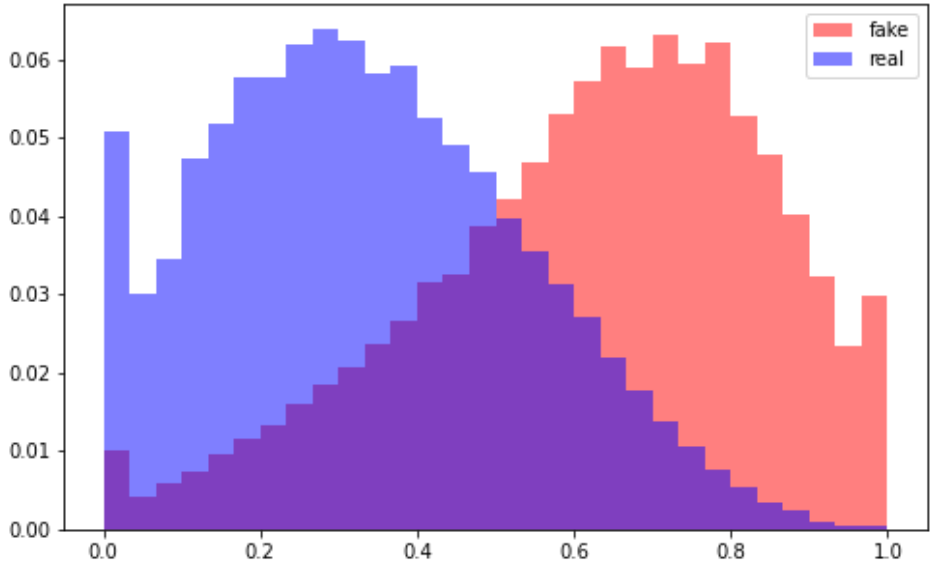
$$dens_p(D) = \min\left(\frac{\text{density}(D)}{|D|}, 1\right) \qquad spr_p(D) = \frac{\text{sent}(D)}{|D|}$$

$$div_p(D) = \min\left(\frac{\text{div}(D)}{q}, 1\right) \qquad comp_p(D) = \frac{\text{densityCo}(D)}{\text{density}(D)}$$

$$pos_p(D) = \frac{\text{Median}(\text{sentPos}(D))}{|D|} \qquad rarity_p(D) = \sum_{t \in PT} \text{freq}_t(D) \times \text{IDF}(t)$$

$$dens_c(D) = \min\left(\frac{\text{density}(D)}{|D|_c}, 1\right) \qquad spr_c(D) = \frac{c(D)}{|D|_c}$$

Experiments: Measures



Given two histograms (one per-class), we wish to quantify the divergence between the two.

- Absolute distance
- Jensen-Shannon divergence

Experiments: Models

Mode	Gran. Train	Gran. Eval	Focus	P	R	micro F_1	macro F_1
binary	binary	binary	paragraph	0.895	0.691	0.780	-
binary	binary	binary	sentence	0.753	0.531	0.623	-
binary	binary	binary	token	0.614	0.266	0.371	-
multi	fine	binary	paragraph	0.890	0.773	0.827	-
multi	fine	binary	sentence	0.757	0.599	0.669	-
multi	fine	binary	token	0.664	0.499	0.570	-
multi	fine	coarse	paragraph	0.664	0.536	0.593	0.489
multi	fine	coarse	sentence	0.532	0.387	0.448	0.345
multi	fine	coarse	token	0.405	0.265	0.320	0.261
multi	fine	fine	paragraph	0.537	0.297	0.470	0.332
multi	fine	fine	sentence	0.423	0.300	0.351	0.258
multi	fine	fine	token	0.316	0.206	0.249	0.202

Table: The performance of XLM-RoBERTa in different settings.

- Base model: based on **XLM-RoBerta** large
- Trained on SemEval2023Task3 Persuasion Technique dataset

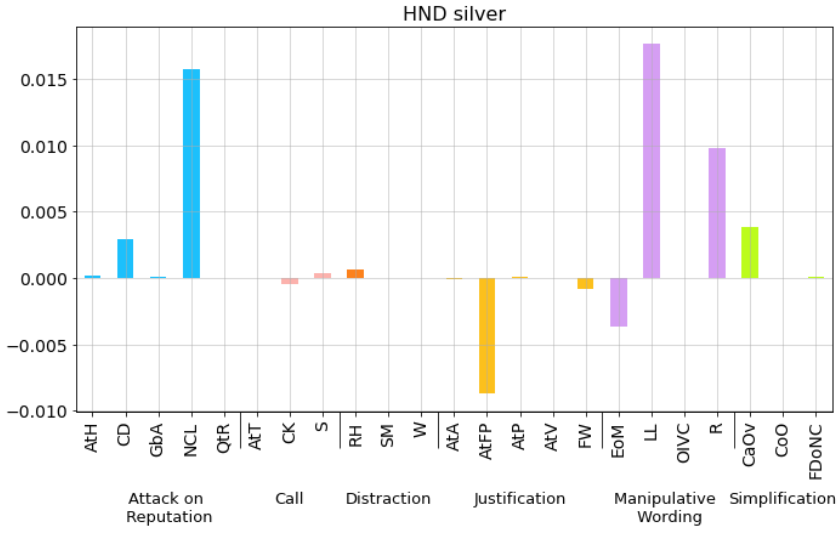
What works and what does not?

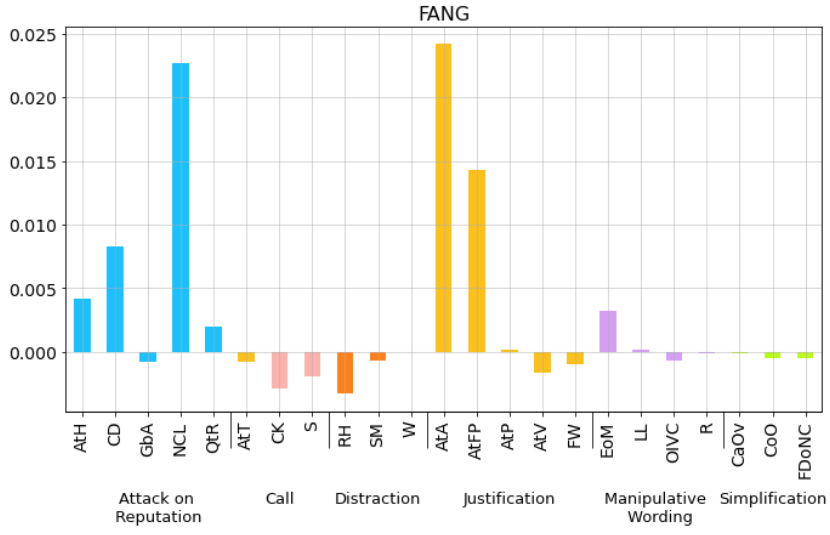
By measuring the discriminatory power on each corpus we were able to rank the performance on each feature.
We used the trivial *#sent* (no. of sentences) as a trivial control variable.

COVID DISINFO		FANG COVID		FakeNewsCoprus		HND gold		HND silver		Kaggle FN		LOCO		QProp		spanish FNC	
feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score
$dens_p$	0.503	$dens_c$	0.524	*#sent*	0.286	div_p	0.492	spr_c	0.292	$dens_p$	0.363	$dens_c$	0.436	div_p	0.496	$dens_c$	0.314
$dens_c$	0.503	spr_c	0.512	div_p	0.109	P	0.484	$dens_p$	0.281	$dens_c$	0.361	div_p	0.419	P	0.462	$dens_p$	0.308
spr_c	0.502	$dens_p$	0.506	$dens_p$	0.106	spr_p	0.441	spr_p	0.28	spr_p	0.359	$dens_p$	0.418	$dens_c$	0.461	spr_c	0.302
spr_p	0.501	spr_p	0.487	$dens_c$	0.103	$comp_p$	0.438	$dens_c$	0.276	spr_c	0.35	spr_c	0.414	$dens_p$	0.451	spr_p	0.296
div_p	0.501	P	0.451	spr_p	0.103	$dens_p$	0.42	div_p	0.253	*#sent*	0.31	P	0.407	$comp_p$	0.445	*#sent*	0.273
pos_p	0.475	*#sent*	0.244	spr_c	0.103	$dens_c$	0.395	P	0.227	P	0.204	spr_p	0.396	spr_c	0.439	pos_p	0.201
$comp_p$	0.4	div_p	0.188	pos_p	0.087	spr_c	0.372	$comp_p$	0.213	$comp_p$	0.154	$comp_p$	0.362	spr_p	0.431	$comp_p$	0.198
#sent	0.372	$comp_p$	0.179	$comp_p$	0.031	*#sent*	0.367	pos_p	0.166	div_p	0.14	pos_p	0.237	*#sent*	0.324	div_p	0.192
P	0.259	pos_p	0.161	P	0.014	pos_p	0.26	*#sent*	0.09	pos_p	0.129	*#sent*	0.123	pos_p	0.282	P	0.145

Table: Ranking of the features with respect to their discriminatory power for the various tasks using *absDistance* metric.

Contribution of Individual Techniques





We experimentally assessed the contribution of each individual PT, by performing a **leave-one-out experiment**.

- The higher the bar, the lower the discrimination metrics upon exclusion.

Observations and Insights

- Conspiracy, Hyperpartisanship and Propaganda are **efficiently distinguished**.
- Results are **mixed to negative** in Fake News related tasks (espacially FakeNewsCorpus).
- **Simpler features** (div_p, $dens_p$, spr_p) seem to be more robust.
- When narrowing down the results to **NER-only sentences**, results were sub-par.
- The sensitivity of individual PTs depends on the **individual task**.

Summary of results

Regarding our 3 **Research Questions**:

- **(RQ1)** PT features exhibit noticeable discriminatory power in:
  - Conspiracy theory detection
  - Hyperpartisan news detection
  - Propaganda detection
- **(RQ2)** Diversity (div_p), Density ($dens_p$), and Spread (spr_p) are the most reliable attributes across corpora
- **(RQ3)** Loaded Language and Name Calling or Labeling contribute noticeably to the discriminatory power.
  - specific histogram profile depends on the given task/corpus