

When You Doubt, Abstain: A Study of Automated Fact-checking in Italian Under Domain Shift

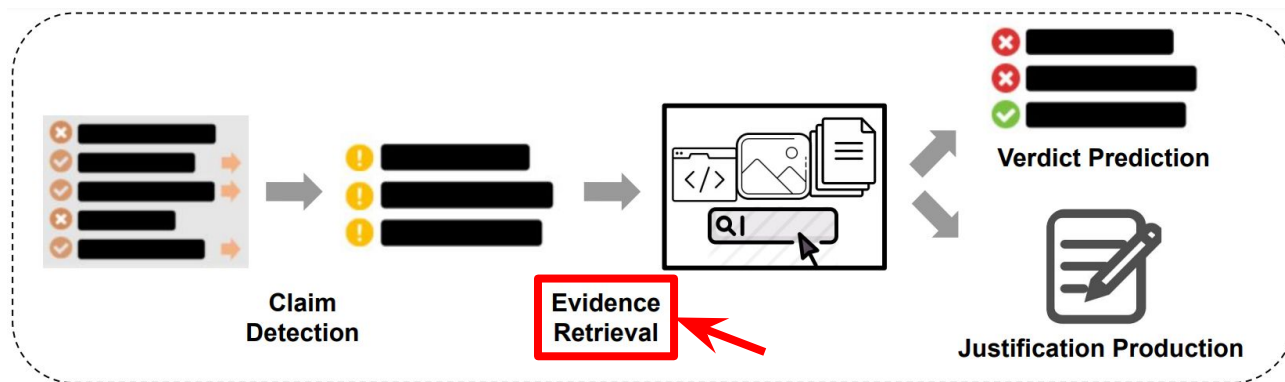
Giovanni Valer¹, Alan Ramponi², Sara Tonelli²

¹ University of Trento

² Fondazione Bruno Kessler (FBK)

Introduction

- Countering the **spread of misinformation**
- Automated tools



Guo et al., 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*.

- **Ambiguity of claims** rarely tackled \Rightarrow **Abstention?**
- **Italian** language overlooked

Fact-checking Data

Italian portion of **X-Fact** (*Gupta and Srikumar, 2021*)

Sources:

- Pagella Politica (PP)
- Agenzia Giornalistica Italia (AGI)

Splits:

- Training set (943) – PP
- Development set (125) – PP
- Test set
 - *In-domain* (190) – PP
 - *Out-of-domain* (160) – AGI

Claim Ambiguity

Based on a preliminary assessment
of the test portions of X-Fact

Missing information

The claim **does not contain information** that
calls for verification:

Example from X-Fact

it. Di Battista e la guerra in Afghanistan.

en. *Di Battista and the war in Afghanistan.*

mostly-true

Lack of context

The claim does not provide enough context (e.g., *who*, *when*, and *where*) or contains underspecified language, ill-defined terms and pronouns:

Example from X-Fact

it. Siamo al nono mese consecutivo di riduzione degli sbarchi.

en. *We are in the ninth consecutive month of reduced arrivals by sea.*

true

Discordant label

The statement has been **reported in a negated form** or as its opposite, but the label reflects the veracity of the original statement:

Example from X-Fact

it. No, la Banca d'Italia non è controllata dalle banche private.

en. *No, the Bank of Italy is not controlled by private banks.*

partly-true

Claim as question

The fact-checked statement has been reported as a question:

Example from X-Fact

it. Davvero la triplice sede del Parlamento europeo costa oltre 200 milioni di euro l'anno?

en. *Does the triple seat of the European Parliament really cost over 200 million euros per year?*

partly-true

Claim Ambiguity

Ambiguity class	PP test		AGI test	
	<i>in-domain</i>		<i>out-of-domain</i>	
Missing information	47	24.7%	6	3.8%
Lack of context	13	6.8%	17	10.6%
Discordant label	13	6.8%	0	0.0%
Claim as question	31	16.3%	0	0.0%
No ambiguity	86	45.3%	137	85.6%
Total	190	100.0%	160	100.0%

We focus on the highlighted subsets

Creation of Challenge Test Sets

Goal: **assess the performance** of automated fact-checking **under genre shift**

Rewritten test sets in two different styles (for both we have 117 *in-domain* claims and 137 *out-of-domain* claims):

News-like
(newspaper headline)

“Il M5S si conferma una delle principali forze politiche in Europa, secondo Di Maio.”

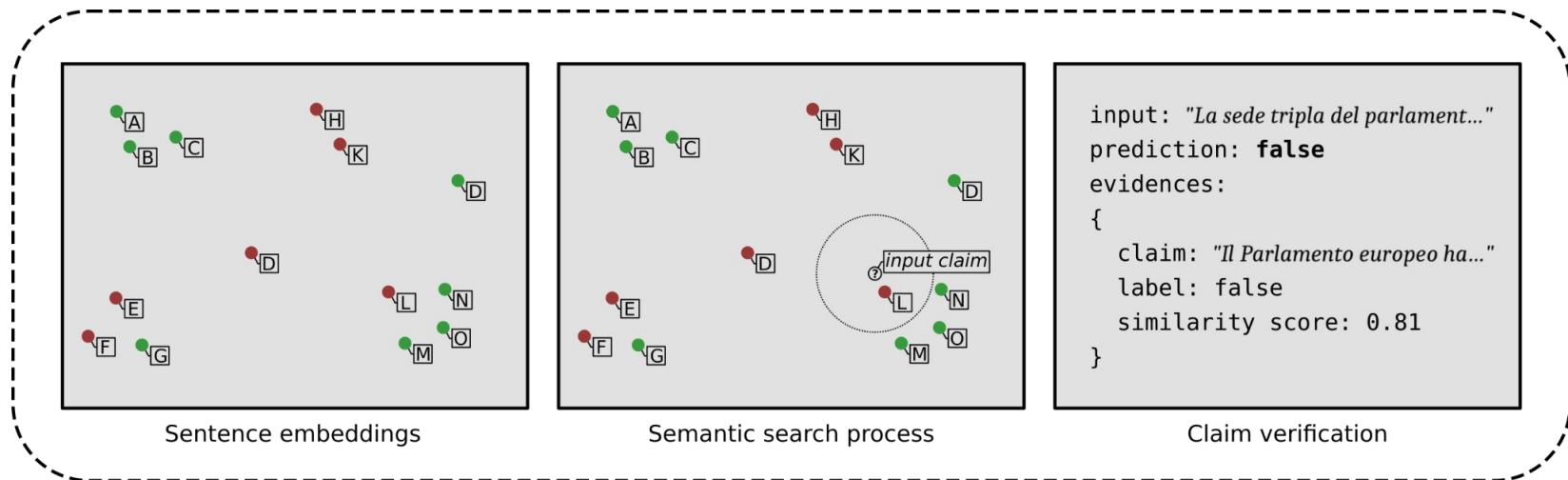
The M5S is confirmed as one of the main political forces in Europe, according to Di Maio.

Social-like
(social media post)

“Il #M5S è tra i partiti maggiori d’Europa!!!”

The #M5S is among the major parties in Europe!!!

Method



Semantic search method for evidence retrieval based on SentenceTransformers¹

Hyperparameter τ (cosine similarity threshold for abstention)

¹ paraphrase-multilingual-MiniLM-L12-v2.

Experimental Setup

Two axes of variation:

- **Source:** in-domain (ID) vs out-of-domain (OOD)
- **Genre:** news-like (NL) vs social-like (SL)

Two setups:

- **Controlled** (relevant information present in the evidence set)
- **Non-controlled** (relevant information may not be present in the evidence set)

Results

Controlled setups

in-domain

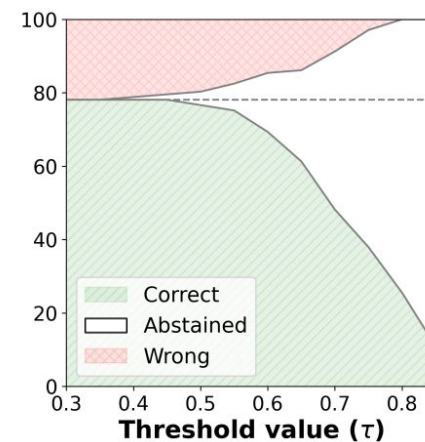
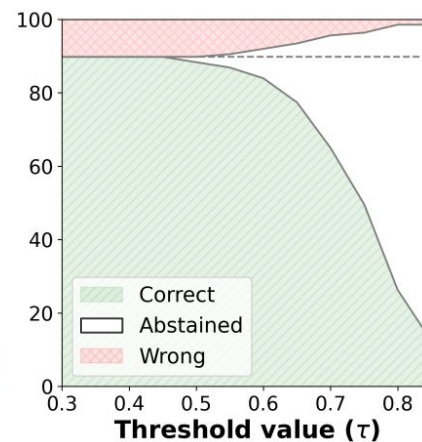
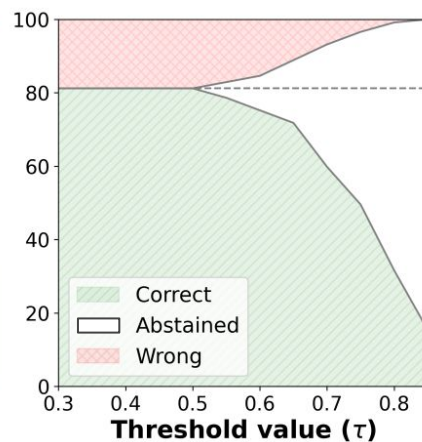
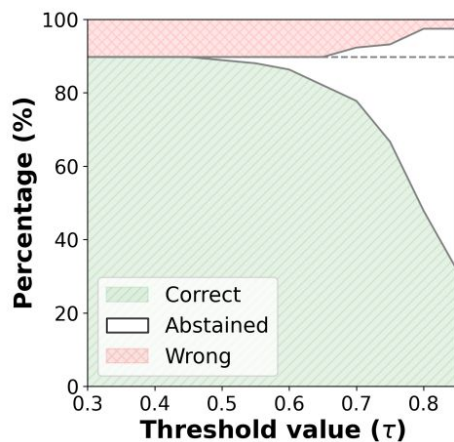
out-of-domain

news-like

social-like

news-like

social-like



Results

GENRE SHIFT

Genre shift has a **large impact** on performance

in-domain: macro F₁ score 0.86 → 0.74 (-0.12)
out-of-domain: macro F₁ score 0.82 → 0.67 (-0.15)

	macro F ₁ score	
	ID	OOD
NL	0.86	0.82
SL	0.74	0.67

	correct (cor)	
	ID	OOD
NL	0.86	0.84
SL	0.75	0.69

	abstention (abs)	
	ID	OOD
NL	0.03	0.08
SL	0.09	0.16

	error (err)	
	ID	OOD
NL	0.10	0.08
SL	0.15	0.15

Results

SOURCES

Fact-checking **sources do matter**,
too

news-like:

macro F₁ score
0.86 ↘ 0.82 (-0.04)

social-like:

0.74 ↘ 0.67 (-0.07)

macro F₁ score

ID → OOD

NL	0.86	0.82
SL	0.74	0.67

correct (cor)

	ID	OOD
NL	0.86	0.84
SL	0.75	0.69

abstention (abs)

	ID	OOD
NL	0.03	0.08
SL	0.09	0.16

error (err)

	ID	OOD
NL	0.10	0.08
SL	0.15	0.15


Results

ABSTENTION

Abstention **helps in reducing errors**

error rate

news-like:

0.10  0.08 (-0.02)

social-like:

0.15  0.15

macro F_1 score

	ID	OOD
NL	0.86	0.82
SL	0.74	0.67

correct (cor)

	ID	OOD
NL	0.86	0.84
SL	0.75	0.69

abstention (abs)

	ID	OOD
NL	0.03	0.08
SL	0.09	0.16

 error (err)

	ID	OOD
NL	0.10	0.08
SL	0.15	0.15

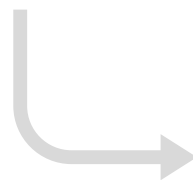
Error Analysis

Causes of errors:

33.0% **correct evidence discarded** (wrong evidence with higher similarity)

22.9% correct evidence not found → **abstention**

44.0% correct evidence not found → **wrong label**



22.2%
claim ambiguity

Conclusion

- Genres and sources have a **large impact on performance**
- **Abstention** to cope with **lack of sufficient evidence**
- **Semantic similarity** for transparent fact-checking

Future work:

- **Automating** claim ambiguity categorization
- **Assessing system efficacy with intended users**

Thank you!

✉ *giovanni.valer@studenti.unitn.it*

🐙 *<https://github.com/jo-valer/fact-checking-ita-abstention>*