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

Giovanni Valer[1,*], Alan Ramponi[2] and Sara Tonelli[2]

[1] *University of Trento, Department of Information Engineering and Computer Science – Trento, Italy*
[2] *Fondazione Bruno Kessler (FBK), Digital Humanities Unit – Trento, Italy*

Abstract

English. Data for building fact-checking models for Italian is scarce, often contains ambiguous claims, and lacks textual diversity. This makes it hard to reliably apply such tools in the real world to support fact-checkers' work. In this paper, we propose a categorization of claim ambiguity and label the largest Italian test set based on it. Moreover, we create challenge sets across two axes of variation: *genres* and fact-checking *sources*. Our experiments using transformer-based semantic search show a large drop in performance under domain shift, and indicate the benefit of models' abstention in case of lacking evidence.

Italiano. *I dati per la creazione di modelli di fact-checking per l'italiano sono esigui, contengono spesso affermazioni ambigue e presentano una limitata diversità stilistica. Questo rende l'uso dei modelli risultanti da parte dei fact-checkers poco affidabile. In questo lavoro classifichiamo l'ambiguità delle affermazioni contenute nel più grande test set di fact-checking per l'italiano e creiamo dei nuovi challenge test set che riflettono stili e fonti differenti. I nostri esperimenti basati sulla ricerca semantica mostrano un notevole calo delle prestazioni in caso di cambio di dominio e indicano l'utilità dell'astensione da parte dei modelli in caso di limitate evidenze.*

Keywords

Automated fact-checking, claim ambiguity, domain shift, models' abstention, semantic search

1. Introduction

Countering the spread of misinformation is one of the major challenges of our society, but human fact-checkers struggle to cope with the increasing amount of content being published. On these bases, in recent years automated fact-checking has gained increasing attention in the NLP community, resulting in a significant body of works and initiatives, e.g., the Fact Extraction and VERification Workshop (FEVER), at its 7th edition in 2023 [1].

Research efforts in NLP for automated fact-checking span over a plurality of tasks, from claim detection to verdict prediction and justification production [2]. Nevertheless, languages other than English, one of them being Italian, are mostly overlooked in current NLP. Specifically, little work has been done to build annotated corpora for the Italian language, which is currently included in just a handful of multilingual datasets, i.e., X-Fact [3] and FakeCovid [4]. To exacerbate the problem, most datasets for automated fact-checking not only include underspecified claims for which verdicts are hard-to-impossible to be determined [5], but also typically lack domain diversity, making it difficult to ascertain the reliability of the resulting fact-checking systems on different genres (e.g., from news headlines to posts on social media).

In this paper, we aim to advance automated fact-checking in Italian by examining claim ambiguity in the largest test set to date, and providing means to measure and mitigate the impact of domain shift along *genres* and *sources* dimensions. Our study shows that automated fact-checking is still far from being reliably applied in the real-world, and indicates the benefit of models' abstention in case of lacking evidence for verification.

Contributions *i)* We propose a categorization of claim ambiguity, *ii)* annotate the Italian test portions of X-Fact according to it, and *iii)* create challenge test sets for studying automated fact-checking in Italian under domain shift. We further *iv)* assess performance shift using transformer-based semantic search, *v)* highlighting the benefit of abstention in the case of insufficient evidence.

2. Fact-checking Data

Among the fact-checking datasets comprising Italian, we select X-Fact [3] for our study since it represents a more diversified set of topics and comprises a larger amount of claims in the Italian language than FakeCovid [4].

X-Fact contains 31,189 non-English textual claims from 25 languages, among which are 1,513 Italian claims based on *Pagella Politica* (PP)[1] and *Agenzia Giornalistica Italia* (AGI)[2] fact-checks. The original veracity labels for

Code: https://github.com/jo-valer/fact-checking-ita-abstention

CLiC-it 2023: Ninth Italian Conference on Computational Linguistics, November 30th, 2023 – December 2nd, 2023, Venice, Italy

* Corresponding author.
giovanni.valer@studenti.unitn.it (G. Valer); alramponi@fbk.eu (A. Ramponi); satonelli@fbk.eu (S. Tonelli)

[1] Pagella Politica website: https://pagellapolitica.it/
[2] Agenzia Giornalistica Italia website: https://www.agi.it/

the claims derive from different sources in multiple languages, and therefore have been homogenized by Gupta and Srikumar (2021) [3] to a fixed label set, i.e., *true*, *mostly-true*, *partly-true*, *mostly-false*, *false*, as well as *complicated* for cases whose original labels have been found hard to be mapped to the proposed label set.[3]

The data is structured into training, development, and test splits. Both the training and development sets have been extracted from PP and include 943 and 125 claims, respectively. The test set instead comprises an *in-domain* portion (190 claims from PP) and an *out-of-domain* one (255 claims from AGI). We remove instances marked as *complicated* by Gupta and Srikumar (2021) [3] from the test set since they do not provide any information about claim veracity. As a result, while the *in-domain* test portion remains the same (i.e., 190 claims), the size of the *out-of-domain* test portion decreases to 160 claims due to the filtering of 95 claims (i.e., 37.3%).

3. Annotation and Challenge Sets

3.1. Claim Categorization

Textual claims that undergo fact-checking may be hard or even impossible to verify due to ambiguity and underspecified language. When it comes to datasets for automated fact-checking, the additional context that has been used by fact-checkers is typically not included, making labels for claims in such decontextualized conditions to change from concrete verdicts (e.g., *true*, *false*) to being unverifiable [5]. Moreover, claims and associated verdict labels that are derived from fact-checking websites and included in most datasets may cause further ambiguity. Indeed, the claim often corresponds to the headline of the article describing the statement that has been verified, but the verdict label typically refers to the latter, and thus the claim-label pair may not match the original statement-label association (see the "Discordant label" ambiguity class described further on).

In this section, we provide a categorization of the reasons why a claim may be ambiguous[4] and annotate the Italian *in-domain* and *out-of-domain* test sets of X-Fact accordingly, expanding the observed causes of ambiguity beyond e.g., underspecification due to ill-defined terms [6] and pronouns [7].

Reasons for claim ambiguity The reasons why a claim may be ambiguous are identified based on a preliminary assessment of the test portions of X-Fact and of past literature [6, 7]. In the following, we provide ambiguity classes ordered by decreasing severity and accompanied by definitions and examples:

1. **Missing information**: the claim does not contain information that calls for verification:

   > "Di Battista e la guerra in Afghanistan." [En: *"Di Battista and the war in Afghanistan."*]: <u>mostly-true</u>

2. **Lack of context**: the claim does not provide enough context (e.g., *who*, *when*, and *where*) or contains ill-defined terms and pronouns, and thus can not be unambiguously verified:

   > "Siamo al nono mese consecutivo di riduzione degli sbarchi." [En: *"We are in the ninth consecutive month of reduced arrivals by sea."*]: <u>true</u>

3. **Discordant label**: the fact-checked statement has been rewritten in a negated form or as its opposite, but the label reflects the veracity of the original statement:

   > "No, la Banca d'Italia non è controllata dalle banche private." [En: *"No, the Bank of Italy is not controlled by private banks."*]: <u>partly-true</u>

4. **Claim as question**: the fact-checked statement has been rewritten as a question. Although it may preserve the information necessary for fact-checking purposes, this alteration does not represent an actual claim:

   > "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?"*]: <u>partly-true</u>

5. **No ambiguity**: the claim is unambiguous and therefore presents sufficient information for automated fact-checking purposes:[5]

   > "In Italia ci sono 18 milioni di persone a rischio povertà." [En: *"In Italy there are 18 million people at risk of poverty."*]: <u>true</u>

Annotating claim ambiguity We manually annotate claims for ambiguity on both Italian test sets of X-Fact following the proposed categories. We focus our efforts on test portions since these represent data that should be used to reliably assess automated fact-checking systems. This results in 350 annotated claims (i.e., 190 *in-domain* and 160 *out-of-domain*, cf. Section 2). If more than one ambiguity class is applicable for a given claim, the more

[3] We leave out the label *other* from our discussion since it is present only in some non-Italian subsets which are not part of this study.

[4] In the remainder of this paper, we use "ambiguity" as a broad term that also includes underspecified language.

[5] Note that real-world facts and the subsequent claims are often time-, space-, and culture-dependent. We relax the "perfect unambiguity" requirement in the context of this work.

Table 1
Distribution of claims across ambiguity classes in the *in-domain* (PP) and *out-of-domain* (AGI) test sets. Claims that present sufficient information for fact-checking are in **bold**. Note that claims with labels providing no information about veracity (i.e., those originally marked as *complicated* in X-Fact) are not included in these counts (cf. Section 2).

Ambiguity class	PP test *in-domain*		AGI test *out-of-domain*	
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%

severe one is chosen. For instance, if a claim falls under both "claim as question" and "lack of context" categories, then the latter is applied. Annotation is carried out by a native speaker of Italian. Since the ambiguity classes are rather straightforward, no double annotation was performed. The distribution of annotated claims among classes and test set portions is presented in Table 1.

3.2. Creation of Challenge Test Sets

A typical assumption in most machine learning algorithms is that training and test data follow the same underlying distribution [8]. This is reflected by datasets in which diversity in textual types is rather limited, which makes it hard to assess the performance of automated fact-checking *into the wild*, such as under genre shift (i.e., from article headlines to user-generated content on social media). Although X-Fact includes *in-domain* and *out-of-domain* sets, these mainly reflect different fact-checking sources rather than textual genres.

To provide the research community with means to investigate and mitigate the impact of genre shift on automated fact-checking in Italian, we extend X-Fact with new challenge test sets. We rewrite the subset of claims from the *in-domain* and *out-of-domain* Italian test sets which exhibit sufficient information for fact-checking purposes (i.e., those in bold in Table 1, namely "claim as question" and "no ambiguity", totalling 117 claims for the *in-domain* test set and 137 claims for the *out-of-domain* one) in two different versions. The first one, which we call *news-like* (nl), resembles the language style of a newspaper headline. It is close in style to the original claim and thus meant to probe minimal shift. The second one, *social-like* (sl), is written trying to imitate social media jargon, using e.g., hashtags and abbreviations, and introducing typos. This is meant for assessing performance in scenarios in which automated fact-checking has to be applied to social media posts. Such process has also taken into account claim veracity, to ensure label consistency between the original text and the rewritten one.

For instance, given the claim: "Di Maio ha ragione: il M5S è una delle principali forze politiche in Europa." [English: "*Di Maio is right: the M5S is one of the main political forces in Europe.*"], the two additional claim versions that we create are the following:

- **News-like**: "Il M5S si conferma una delle principali forze politiche in Europa, secondo Di Maio." [English: "*The M5S is confirmed as one of the main political forces in Europe, according to Di Maio.*"]
- **Social-like**: "Il #M5S è traa i partiti maggiori d'Europa!!!" [English: "*The #M5S is amongg the major parties in Europe!!!*"]

As a result, two *in-domain* (nl and sl, 117 claims each) and two *out-of-domain* (nl and sl, 137 claims each) test sets are created as two variants of the original test sets. Detailed statistics for all the subsets are in Table 2.

4. Experiments

4.1. Experimental Setup

We conduct experiments on automated fact-checking along two axes of variation: source (*in-domain* vs *out-of-domain*) and genre (*news-like* vs *social-like*), using the data splits presented in Table 2.

Method All experiments employ a semantic search method for evidence retrieval based on Sentence-BERT [9], followed by majority-driven veracity classification. Compared to standard sentence classification e.g., using BERT [10], this makes automated fact-checking more transparent, since instances used to determine the veracity of input claims can be inspected and shown to the end user.

Formally, given an input claim $t_i \in T$, where T is a test set among test-(nl|sl)$_{(id|ood)}$ (cf. Table 2, *bottom*), the goal is to find the most relevant claim(s) $\{e_1, ..., e_n\} \in E$, $n \leq |E|$, where E is the evidence set (i.e., union of train, dev, and test$_{(id|ood)}$;[6] cf. Table 2, *top*), and n represents the maximum number of most similar claims to retrieve from it. In order to retrieve such evidence claims, $t_1, ..., t_{|T|}$ and $e_1, ..., e_{|E|}$ are all assigned an embedding v^{t_i} and v^{e_i}, respectively, using a pre-trained multilingual SentenceTransformers model[7] with default hyperparameters [11]. Then, the cosine similarity between

[6] This makes sure that veracity information for input claims is actually available, thus allowing us to study the impact of sources and genres in a *controlled* setting.

[7] `paraphrase-multilingual-MiniLM-L12-v2`. We use a multilingual model since preliminary experiments with Italian ones

Table 2
Distribution of claims across original veracity labels (T: *true*, MT: *mostly-true*, PT: *partly-true*, MF: *mostly-false*, F: *false*), mapped labels (T: *true*, F: *false*), sources (PP: *Pagella Politica*, AGI: *Agenzia Giornalistica Italia*), and genres (NL$_{orig}$: *news-like* as in its original form, NL: rewritten as *news-like*, SL: rewritten as *social-like*). TRAIN and DEV sets follow the original distribution as in X-FACT. All test sets contain the subset of instances that exhibit sufficient information for fact-checking (cf. Table 1).

				Original labels					*Mapped labels*		
Id	**Subset description**	**Source**	**Genre**	**T**	**MT**	**PT**	**MF**	**F**	**T**	**F**	**Tot**
TRAIN	Training set	PP	NL$_{orig}$	215	234	266	0	228	449	494	943
DEV	Development set	PP	NL$_{orig}$	27	33	30	0	35	60	65	125
TEST$_{id}$	Test set (*in-domain*)	PP	NL$_{orig}$	24	28	39	0	26	52	65	117
TEST$_{ood}$	Test set (*out-of-domain*)	AGI	NL$_{orig}$	94	0	0	0	43	94	43	137
TEST-NL$_{id}$	TEST$_{id}$ as *news-like*	PP	NL	24	28	39	0	26	52	65	117
TEST-NL$_{ood}$	TEST$_{ood}$ as *news-like*	AGI	NL	94	0	0	0	43	94	43	137
TEST-SL$_{id}$	TEST$_{id}$ as *social-like*	PP	SL	24	28	39	0	26	52	65	117
TEST-SL$_{ood}$	TEST$_{ood}$ as *social-like*	AGI	SL	94	0	0	0	43	94	43	137

the input claim embedding v^{t_i} and those of all evidence claims $v^{e_1}, ..., v^{e_{|E|}}$ is computed, i.e., $sim(v^{t_i}, v^{e_k}), k = 1, ..., |E|$. If $sim(v^{t_i}, v^{e_k}) > \tau$, where τ is a similarity threshold in $[0, 1]$, the claim e_k is a candidate for determining the veracity label of t_i. All candidate claims are sorted by similarity score and the most recurring label among the top n evidence claims is finally assigned to the input claim t_i. If no evidence claim is found or there is a tie among label counts from retrieved claims, then the model abstains. We believe that the possibility to abstain, rather than forcibly assigning a label, is highly desirable in real-world scenarios, since it is not always possible to assess the veracity of a claim.

Settings In order to isolate the impact on performance of genres and sources from the actual availability of relevant evidence (i.e., verified claims about the input claim's topic), we mainly focus on experiments in a *controlled* setting. This ensures that information for verification of each input claim is available in the evidence set E. Nevertheless, we also present results in a *non-controlled* setup for reference. Specifically, the latter does not include TEST$_{(id|ood)}$ as part of the evidence set E.

Label mapping Since TEST$_{ood}$ and its challenge sets have only *true* and *false* classes, the dataset labels have been mapped as: {*true*, *mostly-true*} $\rightarrow$ *true*, {*partly-true*,[8] *mostly-false*, *false*} $\rightarrow$ *false*.

Metrics We use macro F_1 score for evaluation to account for the unbalanced class distribution in the test sets (cf. *ood* ones, Table 2). When computing the F_1 score, abstention is counted as a wrong prediction, because on a controlled setup the model has access to relevant evidence and thus should not abstain. We also measure the correct (COR), error (ERR), and abstention (ABS) rates, by respectively counting the cases in which the model correctly or wrongly predicts a veracity label, or abstains.

4.2. Model Selection

Our method depends on two hyperparameters: the maximum number n of evidence claims and the threshold τ. We tested values of n in the search space $\{1, 2, 3, 4, 5\}$ across all axes of variation (i.e., data splits in Table 2, *bottom*) and thresholds $\tau \in [0.30, 0.85]$ (with step 0.05),[9] finding that $n = 1$ gives on average the best macro F_1 across all configurations (cf. Figure 1).[10] As a result, we use $n = 1$ in the rest of this paper, and present all τ values in the aforementioned range for discussing the trade-off between errors and abstention.

4.3. Results and Discussion

We present the results across sources, genres, and setups in Figure 2, highlighting the trade-off between abstention, and correct/wrong predictions.

Genre shift has a large impact on performance By comparing results on TEST-NL$_{id}$ with those of TEST-SL$_{id}$ (Figure 2a and 2b) and results on TEST-NL$_{ood}$ with that of TEST-SL$_{ood}$ (Figure 2c and 2d), we see a substantial drop

resulted in worse performance. We hypothesize this is due to pretraining data used by the latter.

[8] Indeed, *partly-true* is used in PP for claims that are wrong but based on a grain of truth.

[9] The range is motivated by preliminary experiments: we found that $\tau < 0.30$ and $\tau > 0.85$ are not informative, since the method retrieves almost all or no claims, respectively.

[10] Interestingly, we observe that $n = 2$ and $n = 4$ values result to low F_1 scores. This is because retrieving an even number of evidence claims leads to a higher probability of abstention, as *true* and *false* evidence claims may be in equal number, and abstention is considered as an error when calculating the F_1 score.

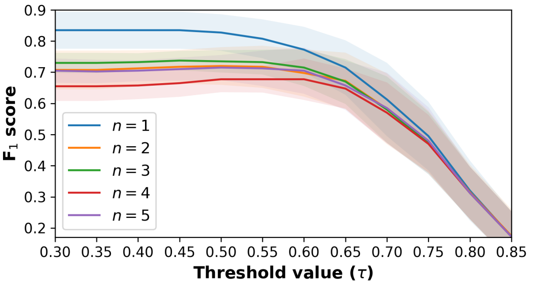


Figure 1: Impact of n and τ hyperparameter values on macro F_1 across test sets. Lines and shading indicate average scores and standard deviation across test sets, respectively.

Table 3
Detailed results across metrics and test sets in the *controlled* setting with hyperparameter values $n = 1$ and $\tau = 0.6$.

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

in COR and an increase in ERR on *social-like* test sets. We present selected results for $\tau = 0.6$ in Table 3, i.e., the threshold for which, on average, the ABS ratio still has a higher impact on ERR than COR before mainly impacting COR (cf. Figure 2). The F_1 scores largely drop from 0.86 to 0.74 and from 0.82 to 0.67 when testing the model on data derived from the same or a different fact-checking source, respectively. Such findings attest the impact of genres on the performance in case of available evidence for veracity prediction. This is confirmed in the *non-controlled* setup (results not shown for brevity), albeit the F_1 score exhibits a smaller drop due to confounding reasons such as the lack of relevant evidence.

Fact-checking sources do matter, too By looking at the results on TEST-NL$_{ood}$ (Figure 2c) and TEST-SL$_{ood}$ (Figure 2d), we can observe that not only the COR percentages drop earlier compared to the *in-domain* source counterparts (i.e., Figure 2a and 2b), but also that errors accumulate in the presence of multiple dimensions of variation, i.e., source and genre (cf. Figure 2d vs Figure 2b). The F_1 score drops further from 0.86 to 0.82 and from 0.74 to 0.67 on NL and SL genres, respectively (Table 3). This is again confirmed by results in the *non-controlled* setting (Figure 2, *bottom*).

Table 4
Analysis of non-COR predictions across test sets in the *controlled* setting ($n = 1$, $\tau = 0.6$). *w/o evidence*: the correct evidence is not retrieved; *w/ evidence*: it is retrieved but replaced by a wrong claim with higher similarity to the input.

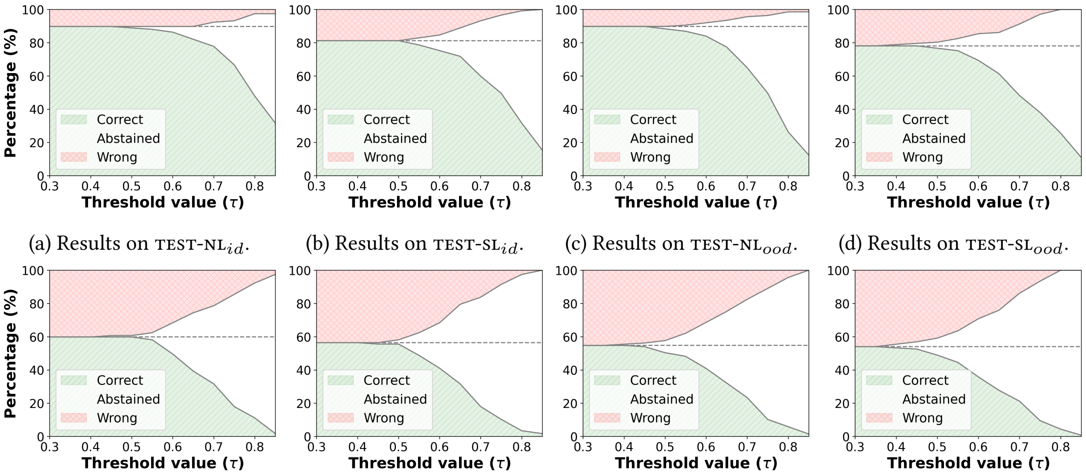
	ABS	ERR	
		w/o evidence	*w/ evidence*
TEST-NL$_{id}$	4 (25.0%)	4 (25.0%)	8 (50.0%)
TEST-SL$_{id}$	11 (37.9%)	7 (24.1%)	11 (37.9%)
TEST-NL$_{ood}$	11 (50.0%)	3 (13.6%)	8 (36.4%)
TEST-SL$_{ood}$	22 (52.4%)	11 (26.2%)	9 (21.4%)
Total	48 (44.0%)	25 (22.9%)	36 (33.0%)

Abstention helps in reducing errors When the model abstains (considering $n = 1$), there are no instances in E that are "similar enough" (i.e., τ) to the input claim. Intuitively, this reduces the impact of erroneous predictions "when in doubt". Figure 2 (*dashed lines*) provides insights into the impact of abstention on formerly COR and ERR percentages across all configurations, as τ varies. We can see that up to $\tau \cong 0.6$, abstention has the great advantage of reducing ERR while negligibly impacting COR. The trade-off between reducing ERR and preserving COR becomes evident with $\tau \gtrsim 0.7$, for which abstention comes mainly at the expense of COR. Even in the more challenging test set (i.e., TEST-SL$_{ood}$), COR predictions are more than double (i.e., 69.3%) than the incorrect ones (i.e., ERR+ABS), but this is not true for $\tau \geq 0.65$. In the *non-controlled* setting (Figure 2, *bottom*), on the other hand, this aspect is hard to assess due to spurious factors. By looking at Table 3, we can also see that error rates on OOD sets compared to ID ones do not increase, and actually moderately decrease on the NL genre (i.e., from 0.10 to 0.08).

4.4. Error Analysis

We collect ABS and ERR predictions across test sets in the *controlled* setting (with $n = 1$, $\tau = 0.6$) and perform a manual analysis. As shown in Table 4, 33.0% has the correct evidence retrieved at first ($sim(v^{t_i}, v^{e_k}) > \tau$), but this is later discarded because wrong evidence has higher similarity to the input. In the remaining cases, the method fails to retrieve the correct evidence, and thus either wrongly predicts the label (22.9%) or abstains (44.0%). Among the 55.9% (61) ERR only, 13.1% (8) is actually based on a correctly-retrieved relevant claim, but because of claim ambiguity in TRAIN and DEV sets (e.g., discordant labels), the prediction is wrong. In particular, such case accounts for 22.2% (8 out of 36) of errors with evidence. This gives a concrete measure of the impact of ambiguity on the fact-checking process.

As regards the performance shift on *social-like* sets



(a) Results on TEST-NL$_{id}$. (b) Results on TEST-SL$_{id}$. (c) Results on TEST-NL$_{ood}$. (d) Results on TEST-SL$_{ood}$.

(e) Results on TEST-NL$_{id}$. (f) Results on TEST-SL$_{id}$. (g) Results on TEST-NL$_{ood}$. (h) Results on TEST-SL$_{ood}$.

Figure 2: Percentage of correct, abstained, and wrong predictions across sources (*id*: PP, *ood*: AGI) and genres (NL: *news-like*, SL: *social-like*). Top: *controlled* setup; bottom: *non-controlled* setup. The dashed line splits abstention into formerly correct (*below* the line) and wrong (*above* the line) predictions.

compared to *news-like* ones, we observe that this is mainly due to the presence of tags (e.g., hashtags, user mentions) in input claims, which account for an average drop of 0.17 F$_1$. *Social-like* claims without tags instead show a more modest drop in performance, i.e., 0.03 F$_1$. Indeed, entities in the form of tags can greatly differ from their plain text counterparts (e.g., "*Autostrade per l'Italia*" uses "*@MyWayAspi*" as username), making semantic matching between the two not trivial.

5. Related Work

Research on automated fact-checking for Italian is very limited. Besides X-Fact, datasets comprising Italian are FakeCovid [4], a multilingual dataset with just 111 articles for Italian, and IRMA [12], a collection of unverified articles from websites classified as "untrustworthy" by fact-checkers.

As regards automated fact-checking on social media, it usually foresees three steps: claim detection, evidence retrieval and veracity prediction [2]. Our contribution is addressing both the second and the third step, in that we focus on detecting or leveraging already fact-checked claims using a semantic similarity approach. Similar to our method, Shaar *et al.* (2020) [13] used cosine similarity between the input and an already-verified claim. However, they do not address ambiguous instances. Hardalov *et al.* (2022) [14] use social media claims for which users have responded with a link to a fact-checking article, but again they do not consider ambiguity and abstention. Broadly, evidence sufficiency prediction has been recently proposed by Atanasova *et al.* (2022) [15] as a task for identifying if evidence is available for reliable fact-checking.

6. Conclusion

In this work, we show that domains do have a large impact on performance of automated fact-checking for Italian, and the faculty of abstention may be considered to cope with lack of sufficient evidence. Moreover, we contribute to the community by classifying claim ambiguity in the largest Italian test set to date and distributing Italian challenge test sets reflecting diversified domains. Future work includes complementing challenge sets with further versions by multiple annotators as well as automating claim ambiguity assessment. Moreover, the confidence level of the classifier could be investigated and measured with tailored metrics to improve automated fact-checking reliability in handling uncertain cases.

In general, as suggested by Schlichtkrull *et al.* (2022) [16], it would important to assess the system efficacy with its intended users, in order to evaluate any unforeseen harm possibly caused by actual applications of the technology. In the future, we therefore plan to test our system by including it in the workflow adopted by professional fact-checkers to verify possible cases of mis/disinformation.

Acknowledgments

This work has received financial support from the European Union's Horizon Europe research and innovation program under grant agreement No. 101070190 (AI4Trust).

References

[1] M. Akhtar, R. Aly, C. Christodoulopoulos, O. Cocarascu, Z. Guo, A. Mittal, M. Schlichtkrull, J. Thorne, A. Vlachos (Eds.), Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER), Association for Computational Linguistics, Dubrovnik, Croatia, 2023. URL: https://aclanthology.org/2023.fever-1.0.

[2] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206. URL: https://aclanthology.org/2022.tacl-1.11. doi:10.1162/tacl_a_00454.

[3] A. Gupta, V. Srikumar, X-fact: A new benchmark dataset for multilingual fact checking, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 675–682. URL: https://aclanthology.org/2021.acl-short.86. doi:10.18653/v1/2021.acl-short.86.

[4] G. K. Shahi, D. Nandini, Fakecovid - A multilingual cross-domain fact check news dataset for COVID-19, in: Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media, 2020. URL: https://workshop-proceedings.icwsm.org/abstract.php?id=2020_14. doi:10.36190/2020.14.

[5] P. Singh, A. Das, J. J. Li, M. Lease, The case for claim difficulty assessment in automatic fact checking, arXiv preprint arXiv:2109.09689 (2022). URL: https://arxiv.org/abs/2109.09689.

[6] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 18–22. URL: https://aclanthology.org/W14-2508. doi:10.3115/v1/W14-2508.

[7] V. Kocijan, T. Lukasiewicz, E. Davis, G. Marcus, L. Morgenstern, A review of Winograd schema challenge datasets and approaches, arXiv preprint arXiv:2004.13831 (2020). URL: https://arxiv.org/abs/2004.13831.

[8] A. Ramponi, B. Plank, Neural unsupervised domain adaptation in NLP—A survey, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 6838–6855. URL: https://aclanthology.org/2020.coling-main.603. doi:10.18653/v1/2020.coling-main.603.

[9] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[11] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4512–4525. URL: https://aclanthology.org/2020.emnlp-main.365. doi:10.18653/v1/2020.emnlp-main.365.

[12] F. Carrella, A. Miani, S. Lewandowsky, IRMA: the 335-million-word Italian coRpus for studying MisinformAtion, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2339–2349. URL: https://aclanthology.org/2023.eacl-main.171. doi:10.18653/v1/2023.eacl-main.171.

[13] S. Shaar, N. Babulkov, G. Da San Martino, P. Nakov, That is a known lie: Detecting previously fact-checked claims, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3607–3618. URL: https://aclanthology.org/2020.acl-main.332. doi:10.18653/v1/2020.acl-main.332.

[14] M. Hardalov, A. Chernyavskiy, I. Koychev, D. Ilvovsky, P. Nakov, CrowdChecked: Detecting previously fact-checked claims in social media, in: Proceedings of the 2nd Conference of the Asia-

Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 266–285. URL: https://aclanthology.org/2022.aacl-main.22.

[15] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, Fact checking with insufficient evidence, Transactions of the Association for Computational Linguistics 10 (2022) 746–763. URL: https://aclanthology.org/2022.tacl-1.43. doi:10.1162/tacl_a_00486.

[16] M. Schlichtkrull, N. D. Ousidhoum, A. Vlachos, The intended uses of automated fact-checking artefacts: Why, how and who, ArXiv abs/2304.14238 (2023). URL: https://api.semanticscholar.org/CorpusID:258352573.