Assessing Italian News Reliability in the Health Domain through Text Analysis of Headlines

Luca Giordano
UNIOR NLP Research Group
University of Naples "L'Orientale"
giordanoluca.uni@gmail.com

Maria Pia di Buono
UNIOR NLP Research Group
University of Naples "L'Orientale"
mpdibuono@unior.it

Abstract

Fake news detection and fact checking represent challenging research areas in Natural Language Processing (NLP), especially in the health domain, which presents specific characteristics to be dealt with. On the one hand, online sources have become one of the main channels to retrieve health-related information. On the other hand, most of the time such online information suffers from lack of quality and requires domain-specific knowledge to be assessed. Therefore, the spread of untrustworthy health-related content urges to be mitigated since it may represent a threat for lives.

To this aim, we develop a domain-specific annotated dataset suitable for training automatic systems to assess Italian news reliability. Our proposal tries to overcome some of the limitations of the available datasets by applying an in-depth text analysis to obtain a more fine-grained reliability assessment in the health domain.

1 Introduction

Lately, the use of online sources for retrieving health information has become widespread, and thus an important source of medical advice (Dai et al., 2020). Particularly, social media platforms (SMPs) seem to be one of the most preferred channels to search and share information, especially in the health domain (Chen et al., 2018). As proved by several scholars (e.g., Finney Rutten et al. (2019); Basch et al. (2017)), the Internet and SMPs represent the main source of information for adults and also adolescents that are active users and searchers for online health information (Greškovičová et al., 2022).

Nevertheless, online health information is affected by several limitations with reference to its quality (Melchior and Oliveira, 2022). The lack of quality in information may generate two main types of untrustworthy content, namely disinformation and misinformation (Lazer et al., 2018). Nowadays, fighting the spread of untrustworthy and low-quality content through fake news detection and/or fact checking represents one of the main challenges to be faced. This is particularly true in the medical domain because such untrustworthy health-related content threaten lives (Anoop et al., 2020).

The Covid-19 pandemic has exacerbated the problem and brought out the need for gold standard datasets and predefined benchmarks for automated approaches, which have been neglected before that, as revealed by Viviani and Pasi (2017). In fact, the scarcity of comprehensive resources, mainly datasets, for fake health news detection slows down the development of novel approaches devoted to detect misinformation and disinformation within this domain (Dai et al., 2020).

Still, the development of resources suitable for assessing information and news in the health domain is far to be fully satisfied, mainly with reference to some domain-specific aspects and languages.

For this reason, in this paper we present a domain-specific annotated dataset suitable for training automatic systems to assess Italian news reliability. Our proposal tries to overcome some of the limitations of the available datasets and to propose a more fine-grained assessment of health-related news, achieved through an in-depth text analysis. Our main contributions are three: (i) proposing a set of stylometric, lexical, and sentiment features to assess news reliability; (ii) developing a domain-specific dataset for the Italian language[1]; (iii) providing a first baseline for the developed dataset.

The rest of the paper is organized as follows. In the next section, we present studies which are relevant to our analysis, referring mainly to the

[1]The dataset is publicly available at https://github.com/unior-nlp-research-group/TRADISAN.

development of datasets for fake news detection. In Section 3, we introduce our methodology, our dataset and the feature set. In Section 4 we explain the experimental setup and present the results. Finally in Section 5 conclusion and future work are discussed.

2 Related Work

The majority of studies published and resources made available focus on a binary classification of the veracity of English news at document-level (that is, an overall veracity rating either True or False for the whole news), although tested by means of different kinds of analysis (such as a range of linguistic features, e.g., Choudhary and Arora (2021); Kasseropoulos and Tjortjis (2021), sentiment analysis, e.g., Alonso et al. (2021) and others). As shown in D'Ulizia et al. (2021), out of the 27 datasets surveyed in the paper, 14 present a binary veracity classification (such as Shu et al. (2020); Tacchini et al. (2017)), while only 4 of them a three-way rating scale (such as Thorne et al. (2018)) and 6 a four-way one (such as Santia and Williams (2018)). Furthermore, 22 out of 27 are monolingual English datasets, only 2 are focused on the Health domain (Posadas-Durán et al., 2019; Jwa et al., 2019) and all of them are annotated at document-level.

Although in Bonet-Jover (2022) the classification proposed is still binary (Reliable/Unreliable), it is noteworthy that the author works on Spanish and that the annotation proposal is focused on the individual annotation of different structural and content elements of the news, therefore going beyond the document-level of analysis.

Regarding the Italian language, to the best of our knowledge, there seems to exist only one publicly accessible dataset of Italian news annotated according to their veracity value, namely HoaxItaly (Pierri et al., 2020): it is a dataset composed of 1.2M tweets referring to 37k Italian news in total, divided into 3566 fact-checked true news and 32,686 fake news. However, the news domain is generic, the assessment is binary and at document-level.

With reference to the set of features typical of trustworthy and untrustworthy news respectively, several studies highlight different kinds of linguistic patterns.

In Biyani et al. (2016) the authors show that the degree of informality of a webpage, as measured by different metrics, is a strong indicator of it being a clickbait, that is an article with a misleading headline, exaggerating the content on the landing page. The amount of superlatives, quotes, exclamations, upper case letters, question marks and other indicators are used as features for a machine-learning model which achieves a 74.9% F1 score in predicting clickbaits.

Horne and Adali (2017) apply a set of linguistic features to three datasets in order to analyze the language of news articles in the political domain. They show that stylistic features such as the length of the article, the use of punctuation, the amount of personal pronouns, nouns and adverbs, the lexical redundancy of the text and others, applied both to the headline and to the body of the news, can help distinguish between real and fake news. Their findings are mostly confirmed by Shrestha and Spezzano (2021), who conduct a reproducibility study, and in addition show that also other factors, such as emotion and readability features are helpful in the fake news detection task.

In Rashkin et al. (2017) the authors show that features such as the amount of swear words, hedge words, sexual-related words, negations, superlatives and others appear to be typical of fake political news, while a frequent use of numbers, money-related words, assertive expressions and comparatives appear to be typical of true political news.

Greškovičová et al. (2022) show that seemingly minor editorial elements, such as poor grammar or boldface, in addition to the presence of superlatives, clickbaits and appeal to authority in health-related messages, which are all typical elements of untrustworthy news, influence and distort the perception of the credibility of news among secondary school students.

3 Methodology

Dai et al. (2020) identified several challenges that have to be addressed in fake health news detection, as they are specific of this domain. In fact, fake health news may require specialized knowledge to be recognized more than fake news in other domains.

Furthermore, health news are also easier to be manipulated, in that they can be easily transformed into misinformation or disinformation just by stating the association as causation or mix-

ing up the absolute risk and relative risk, which, as Dai et al. (2020) point out, require just minor modifications of the true information.

Thus, the proposed methodology tries to combine the identification of trustworthy sources together with the integration of linguistic and sentiment features selected by means of an in-depth analysis. To our aims, we adopt the criterion of reliability instead of veracity, to distinguish untrustworthy news from trustworthy ones and assume that stylometric, lexical and sentiment-based characteristics can be representative of the degree of news reliability.

As first step, we collect a list of news sources (i.e., online newspapers) which have been classified as trustable or untrustable by Newsguard[2], Media Bias/Fact Check[3], Bufale.net[4] and Butac[5], two international and two italian fact checking organizations which, among other activities, publish analyses and reports on news sources' trustworthiness. Furthermore, we take into account the data and analysis provided in the Digital News Report 2022 for Italy published by the Reuters Institute for the Study of Journalism[6]. Therefore, we create two lists of sources, respectively a *trustworthy* list and an *untrustworthy* list (Table 1).

We use these sources to extract a set of health-related news, using the classification by categories provided by the newspapers themselves together with a topic-label based extraction. This allows us to come up with a list of both trustworthy and untrustworthy news. Then, we perform a linguistic analysis to select a set of features that are representative of news reliability.

3.1 Data Collection

The list of trustworthy sources is made up of 12 Italian news outlets (e.g., Il Sole 24 Ore[7], la Repubblica[8], ANSA[9]), while the list of unstrustworthy sources is made up of 26 Italian news outlets (e.g., Voxnews[10], Dionidream[11], Byoblu[12]) for a total amount of 38 news sources (Table 1).

In order to collect the data from our sources, we write Python scripts tailored to each news outlet in order to scrape the news content. We exploit the Python libraries *pandas*[13], *requests*[14], *beautifulsoup*[15] and *newspaper3k*[16], which stem from machine-learning and data science. We aim at extracting the URLs of each article in the health-related categories of the news outlets and through those we extract the news content, that is the article's source, date of publication, headline, body of text and links to its images, if any (Table 2).

Then, we remove broken links and articles with missing information, as well as duplicate articles from the same source. We also remark a potentially interesting phenomenon: 28 articles among the ones extracted from the trustworthy sources and 17 among the ones from the untrustworthy sources present an identical headline, despite being published by different sources. This might suggest plagiarism among news outlets. We keep these articles in our dataset since they might be significant, although we are aware that the presence of duplicates might affect the training data. Nevertheless, they represent a small part within the total amount of data. From the trustworthy list we keep a total of 9.973 news, which amount to 156.372 sentences and 4.925.379 tokens (we adopt the default AntConc token definition "Character Classes"[17]); from the untrustworthy list we keep a total of 22.128 news, which amount to 611.433 sentences and 17.648.641 tokens. Therefore, the corpus is made up of a total of 32.101 news published between November 1999 and February 2023, and it amounts to 767.805 sentences and 22.574.020 tokens (Table 3). To the aim of the present analysis we consider just news headlines, which amount to a total of 351.104 tokens.

3.2 Linguistic Analysis

In order to select the features suitable for our news assessment, we perform an initial analysis of our corpus to identify a first set of linguis-

[2] https://www.newsguardtech.com/it/
[3] https://mediabiasfactcheck.com/
[4] https://www.bufale.net/
[5] https://www.butac.it/
[6] https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/italy
[7] https://www.ilsole24ore.com/
[8] https://www.repubblica.it/
[9] https://www.ansa.it/
[10] https://voxnews.info/
[11] https://dionidream.com/
[12] https://www.byoblu.com/
[13] https://pandas.pydata.org/
[14] https://requests.readthedocs.io/en/latest/
[15] https://beautiful-soup-4.readthedocs.io/en/latest/
[16] https://newspaper.readthedocs.io/en/latest/
[17] https://laurenceanthony.net/software/antconc/releases/AntConc4011/help.pdf, p.13

Trustworthy	Untrustworthy	
Salute.gov	Eticamente.net	Vacciniinforma
ISS	Raffaele Palermo News	eVenti Avversi
la Repubblica	Nexus Edizioni	ByoBlu
il Post	Scienza e Conoscenza	Come Don Chisciotte
Vaccinarsi.org	COMILVA	VoxNews
il Fatto Quotidiano	Vivo in Salute	Mag24
TPI	The Living Spirits	Dagospia
AGI	Il Paragone	Filosofia e Scienza
ANSA	Database Italia	controinformazione.info
Focus	Ingannati	Disquisendo
Il Sole 24 Ore	CheSuccede	Essere Informati
Corriere della Sera	SocialBuzz!	Eurosalus
	Silenzi e Falsità	Dionidream

Table 1: Data Sources

ID	Source	Date	Headline	Text	Image	URL
3762	la Repubblica	2019/04/12	Fagioli e spinaci tengono lontano il tumore della vescica	...	Image1.jpg	...
9626	Il Sole 24 Ore	2023/01/12	Più contagi, non casi più gravi e lo scudo dei vaccini: ecco perché la variante Kraken non deve fare paura	...	Image1.jpg	...
15526	ByoBlu	2022/09/21	"BILL GATES HA GESTITO IL COVID PER ARRICCHIRSI": ORA SE NE ACCORGE ANCHE IL MAINSTREAM	...	Image1.jpg	...
18104	VoxNews	2021/04/09	RECORD DI MORTI SPALMATI: 718 IN 24 ORE, 9 APRILE SCORSO ANNO ERANO STATI 612	...	Image1.jpg	...

Table 2: Examples of Trustworthy (IDs 3762 and 9626) and Untrustworthy (IDs 15526 and 18104) Entries from our Corpus

.

List	# News	# Sentences	# Tokens
Trust.	9.973	156.372	4.925.379
Untrust.	22.128	611.433	17.648.641
TOTAL	31.101	767.805	22.574.020

Table 3: Corpus Description

tic aspects denoting (un)reliability. We adopt a method which includes a top-down approach, namely applying features already used by other scholars for other languages and domains (see Section 2), and a bottom-up approach, that is we analyse the dataset and collect features that arise from our set of news.
We obtain a total number of 31 features (Table 4) accounting for three different levels of analysis, namely stylometry, lexicon, and sentiment.

Stylometric Features The stylometric features we take into account refer to sentence and word length (by characters), the use of uppercase style, the frequency of consecutive question and exclamation marks, frequency of quotes, double quotes and single quotes, ellipses and direct discourse. We also compute the amount of typos through a customized Contextual Spell Checker[18], a deep-learning based Noisy Channel Model Spell Algorithm trained on the PAISÀ Corpus[19], one of the largest publicly available corpora of Italian Web texts, licensed under Creative Commons.
The number of words written in uppercase, the number of long words (understood as being longer than 6 characters) and the number of typos are all weighted values accounting for the length of the sentence.

Lexical Features The lexical features we compute are the number of adverbs, comparatives, superlatives, currency-related words (such as *dollar*), negative adverbs, nouns, proper nouns, adjectives, possessive adjectives other than the 1st and 2nd singular and digits. Additionally, we exploit the Revised HurtLex (Tontodimamma et al.,

[18] Towards Data Science - Training a Contextual Spell Checker for Italian Language
[19] https://www.corpusitaliano.it/

2022), a lexicon of offensive, aggressive, and hateful words divided into 17 categories in over 50 languages in order to compute the number of occurrences of such words in the corpus. In the revised version, every Italian headword is annotated with an offensiveness level score, derived by applying an Item Response Theory model to the ratings provided by a large number of annotators (Tontodimamma et al., 2022). Therefore, we also compute the total offensiveness score of the sentence based on the scores of the words contained in it.
Furthermore, we also count the occurrences of domain-specific *buzzwords*, understood by the definition provided by the Cambridge Dictionary: "a word or expression from a particular subject area that has become fashionable by being used a lot, especially on television and in the newspapers"[20]. For this purpose, we compile a gazetteer of 73 words and phrases extracted from the top 300 keywords in the corpus sorted by likelihood and from the top ranking bigrams and trigrams sorted by frequency. Some examples of buzzwords in our gazetteer are *vaccino* (vaccine), *covid*, *coronavirus*, *sintomi* (symptoms), *immunità di gregge* (herd immunity), *lockdown*, *AIDS*, *green pass*, *vaiolo delle scimmie* (monkeypox) and *no vax*. We assume that Covid-19 global impact, urgency, and relevance as a major health crisis have led to a significant concentration of Covid-19-related keywords in the corpus, despite the pandemic started only in 2020, while the corpus contains news up to 1999. This might be evidence of the impact of the pandemic on news production in Italy. Therefore, we choose to keep this statistical bias in our buzzwords gazetteer as well. All lexical features, except for the offensiveness score, are weighted values accounting for the length of the sentence.

Sentiment Features Additionally, we exploit the adoption of sentiment-related features. This comes from the fact that several scholars (Alonso et al., 2021; Bhutani et al., 2019; Ajao et al., 2019) have recognized that the polarity and strength of sentiments expressed in text can improve the results in fake news and rumor detection tasks.
Thus, we apply NRC Emotion Intensity Lexicon (Mohammad and Turney, 2013) to detect and evaluate the presence of emotions-related words within the texts, such as *anger*, *joy*, and *trust*. In fact, we notice that news from the untrustworthy sources are characterized by a more frequent use of words associated with negative emotions, such as anger, e.g., Example (1), while trustworthy news tend to express more positive emotions, such as joy or trust, e.g., Example (2).

Source: Disquisendo - ID: 26847
Il governo italiano ha dichiarato GUERRA agli italiani. OBBLIGO VACCINALE che passa da 4 a 12 e fino a 16 anni!! SVEGLIAAAAA!! (The Italian government has declared WAR on the Italians. COMPULSORY VACCINATION goes from 4 to 12 and up to 16 years!! WAKE UUUUUP!!) (1)

Source: La Repubblica - ID: 4148
Lo smartwatch? Può salvare la vita (letteralmente) (The smartwatch? It can (literally) save lives) (2)

Furthermore, through SentITA (Nicola, 2018) we also consider the sentiment polarity of the headlines, as untrustworthy news tend to present a mostly negative polarity while trustworthy news a mostly positive one (Shrestha and Spezzano, 2021).

3.3 Reliability Assessment

We perform an analysis of news headlines from both trustworthy and untrustworthy sets, according to the aforementioned features and use these results to define a textual model. The textual model characterizes the set of untrustworthy news headlines and presents the following linguistic aspects:

- Longer headlines (by characters);
- Frequent use of uppercase style;
- Presence of consecutive question and exclamation marks;
- Higher frequency of ellipses, typos, double and single quotes (but less direct discourse);
- Higher frequency of adverbs, superlatives, first person singular pronouns and negative adverbs;
- Limited use of comparatives, currency-related words, nouns, adjectives, second person singular pronouns and digits;
- Higher frequency of words and phrases from

[20] https://dictionary.cambridge.org/it/dizionario/inglese/buzzword

Stylometric	Lexical	Sentiment
char_count	adverb_count_w	nrc_anger_w
uppercase_word_w	comp_w	nrc_trust_w
long_w_w	superl_count_w	nrc_joy_w
consecutive_question_count	currency_w	opos
consecutive_excla_count	rev_hurtlex_count_w	oneg
quotes_count	hurtlex_score	
dou_quotes_count	neg_adverbs_count_w	
single_quote_count	noun_count_w	
ellipses_count	prop_noun_count_w	
direct_discourse	adj_count_w	
typo_count_w	adj_poss_others_w	
	1st_pers_sing_w	
	2nd_pers_sing_w	
	digits_w	
	buzzwords_count_w	

Table 4: List of the 31 Reliability Features

the HurtLex lexicon and a higher offensiveness score;

- Higher frequency of proper nouns;
- Slightly higher frequency of buzzwords;
- Lower frequency of lexical items related to trust and joy.

Then, the textual model is employed to assess the headline reliability.

3.4 Dataset Creation

On the basis of such methodology, we create a dataset which contains the information related to the textual model for assessing reliability (Figure 1). The selected features are annotated according to their pertaining level, that is stylometric (`styl`), lexical (`lex`), and sentiment (`sent`).

In addition to this annotation at title-level, we also provide the dataset with additional annotations (Table 5), such as lemmatization (L), Part-of-speech tagging (PoS), Inside–Outside–Beginning chunk-tagging and (IOB) Named Entity Recognition tagging (NER). We test the annotated dataset performing an experiment to evaluate the results from some of the most common classifiers.

4 Experiment

We conduct a series of experiments to test our hypothesis, i.e. the assumption that stylometric, lexical and sentiment-based features can be suitable for assessing news reliability. Therefore, the main aim of these experiments is to test how fit our feature set is for an automatic assessment of news reliability. Although the final goal is a fine-grained (multi-class) automatic reliability annotation of the whole dataset, for the sake of these experiments and its contextual aim (i.e., testing the feature set and the generalizability of the results for the dataset annotation, rather than the classification granularity and performance *per se*), we assume that every article from the untrustworthy and trustworthy lists make up only two separate classes, therefore configuring it as a binary classification problem.
Since the dataset is imbalanced, we perform an undersampling process, i.e. we extract a sample of random unstrustworthy news equal to the (smaller) subset of trustworthy news (9973 samples). We end up with two equally sized subsets which amount to a total of 19946 samples. We justify the undersampling since the final number of samples is still a considerable amount. Finally, we do not stratify the sampling process neither on date of publication, nor source of provenance nor any other factor since we aim at a subset as randomized as possible. After the random undersampling, the subset of untrustworthy news contains 2 of the 17 duplicates, while the subset of trustworthy news keeps all its original 28 duplicates.

Environmental Setup All code was written and compiled in Python 3.10 on Linux Ubuntu 23.04 and several packages and libraries were exploited, such as *pandas*, *NumPy*[21], *SpaCy*[22], *NLTK*[23], *Transformers*[24], *scikit-learn*[25], *fastText*[26] and *PyTorch*[27].

[21] https://numpy.org/
[22] https://spacy.io/
[23] https://www.nltk.org/index.html
[24] https://huggingface.co/docs/transformers/index
[25] https://scikit-learn.org/stable/
[26] https://fasttext.cc/
[27] https://pytorch.org/

```
{
 "id": 16733,
 "source": "Come Don Chisciotte",
 "date": 2020-10-13 17:03:25+00:00,
 "url": "https://comedonchisciotte.org/la-farsa-dei-tamponi-e-degli-asintomatici/",
 "headline": "La FARSA dei Tamponi e degli Asintomatici",
       "styl":[
           {"char_count": 35.0, "uppercase_word_w": 0.142857, "long_w_w": 0.285714,
           "consecutive_question_count": 0, "consecutive_excla_count": 0, ...}
       ],
       "lex":[
           {"adverb_count_w": 0.0, "comp_w": 0.0, "superl_count_w": 0.0,
           "currency_w": 0.0, "rev_hurtlex_count_w": 0.0, ...}
       ],
       "sent":[
           {"nrc_anger_w": 0.1428571428571428, "nrc_trust_w": 0.0, "nrc_joy_w": 0.0,
           "opos": 0.021332342, "oneg": 0.9845031}
}
```


Figure 1: Annotation example for the headline ID: 16733 *La FARSA dei Tamponi e degli Asintomatici* (The FRAUD of Swabs and Asymptomatic patients), source: Come Don Chisciotte

H	ID	Token	L	PoS	IOB	NER
18192	1	AstraZeneca	AstraZeneca	PROPN	B	ORG
18192	2	vietato	vietare	VERB	O	–
18192	3	in	in	ADP	O	–
18192	4	...	...	...	...	...

Table 5: Annotation example extracted from the headline ID: 18192 *AstraZeneca vietato in Germania sotto ai 60 anni, Merkel: "Impossibile nascondere l'insicurezza"* (AstraZeneca banned in Germany under 60 years old, Merkel: "Impossible to hide uncertainty"), source: VoxNews

The Neural Network runs on an NVIDIA GeForce RTX™ 3060 Laptop GPU with CUDA v12.0.

Feature Selection In order to reduce computational cost, avoid overfitting, increase generalizability, and contribute to the explainability of the models, we apply statistical-based feature selection techniques, aiming at reducing the number of input variables to only those that have the strongest relationship with the target variable (Butcher and Smith, 2020). We adopt a filter-based univariate feature selection method. In detail, since we are dealing with numerical input variables and categorical output variables, we perform an analysis of variance (ANOVA) to compute the ANOVA correlation coefficient (F-value). ANOVA test is used to compare the means of different groups on a dependent variable and to determine whether the difference in group means is due to random variation or if they represent true population differences. Its assumptions are independence, homogeneity of variances of the residuals and a normal distribution (Butcher and Smith, 2020). We assume that each feature is independent from the other, and, since we conduct the analysis on two equally big subsets built *ad-hoc*, we can also assume feature homogeneity (Sawyer, 2009).

Regarding normality of distribution, several scholars, e.g., Lumley et al. (2002); Ghasemi and Zahediasl (2012), show that with large sample sizes the distribution of data can be ignored, as the potential violation of the normality assumption does not cause problems. Moreover, the adoption of the ANOVA test is justified due to its robustness under conditions of non-normally distributed data, as proved by Schmider et al. (2010) and Blanca Mena et al. (2017). Since ANOVA test can be suitable for both normal and non-normal distributions, especially with large sample sizes and our sample size amounts to 19946 samples, we choose not to test normality and to perform directly the ANOVA test.

#	Top Features	F-value
1	prop_noun_count_w	1068.08
2	uppercase_word_w	832.12
3	char_count	630.26
4	dou_quotes_count	393.81
5	ellipses_count	387.03
6	single_quote_count	367.05
7	quotes_count	338.06
8	direct_discourse	223.11
9	noun_count_w	193.67
10	typo_count_w	172.22
11	long_w_w	170.29
12	oneg	159.22
13	hurtlex_score	128.98

Table 6: Top features calculated using ANOVA

Features are then sorted in descending order by the F-value computed with the ANOVA test to determine the importance. We choose to consider the topK features that have an F-value of more than 100. We therefore keep the top 13 features (Table 6).

Classification We conduct a series of experiments, testing five different machine-learning classifiers (namely, Logistic Regression, Decision Tree, Multinomial Naive-Bayes, Random Forest and LinearSVC) and, for BERT, a Multi-Layer Perceptron (MLP) with different input combinations and different word embedding techniques (namely, GloVe, fastText, and pre-trained BERT Base). We split the data in 90:10 training and testing ratio and make sure that all the duplicates are always only in the training set, since, as stated in Section 3.1, they might have been generated through a process we want to take into account. We then perform a cross-validation on the training set, i.e. we split it into 10 train/validation subsets, while the test set remains unaltered. In each iteration, the training set is used for training while the validation set for validation. The performance measure reported is then the average of the values computed in the loop. For the MLP, the cross-validation is performed directly in the training loop, while, for the ML classifiers, through a GridSearchCV[28] technique implemented via scikit-learn, which also allows us to perform a hyperparameter optimization for every ML classifier. The cross-validation parameter is set to 10 folds. We then use the best estimator obtained for the classification task on the test set.
First, we try a classification taking only the whole 31 numerical features as input, without any word vector representation.
Then we ignore the features and classify the data only with three different word embedding techniques; we first try GloVe, then fastText and finally an italian XXL Bert Base transformer cased model pre-trained on the whole italian Wikipedia, OPUS corpus and the italian subset of OSCAR corpus, for a total amount of 13,138,379,147 tokens[29]. Being a Base model, it is made up of 12 layers of transformers block with a hidden size of 768 and 12 self-attention heads and has around 110M trainable parameters.
Then, we combine the different word embeddings with all 31 features.
Finally, we classify the data with a combination of the different word embeddings and only the top 13 features we obtained from the feature selection process. We implement the MLP with PyTorch: the pooled output of the BERT encoder is used as input, the dropout rate is set at 0.5, the activation function is ReLu, the optimizer Adam, the loss function CrossEntropy, and we found that the optimal number of epochs is 6. When combining BERT with the features, the linear layer takes as input a tensor of length equal to the pooled output of BERT + the number of features.

Results The results (Table 7) show that, as expected, state-of-the-art BERT is the best model, achieving an F1 score of 0.855 alone, 0.884 when combined with the top 13 features. A classification based exclusively on our entire feature set achieves an F1 score of 0.70, while with only the top 13 it decreases to 0.68. Although the score is slightly lower, it is noteworthy that less than half of the original feature set were used. This emphasizes the importance of the feature selection process, and this must be taken into account for the dataset annotation, for example by assigning different weights to different features. The use of our features in all settings (alone, in combination with word embeddings and with BERT) improves the results, although slightly. The improvement is more considerable for fastText word embeddings than GloVe.
These results show that this feature set can be a

[28] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[29] https://huggingface.co/dbmdz/bert-base-italian-cased

Model	Classifier	$P_{MacroAVG}$	$R_{MacroAVG}$	F1
All Features	RandomForest	0.70	0.70	0.70
Top13 Features	RandomForest	0.68	0.68	0.68
fastText	LinearSVC	0.76	0.76	0.76
fastText + All Features	LinearSVC	0.78	0.78	0.78
fastText + Top13 Features	LinearSVC	0.79	0.79	0.79
GloVe	LogisticRegression	0.78	0.78	0.78
GloVe + All Features	LogisticRegression	0.79	0.79	0.79
GloVe + Top13 Features	LogisticRegression	0.79	0.78	0.79
$BERT_{BASE}$	Multi-Layer Perceptron	0.855	0.855	0.855
$BERT_{BASE}$ *+ All Features*	Multi-Layer Perceptron	0.871	0.871	0.871
$BERT_{BASE}$ + Top13 Features	**Multi-Layer Perceptron**	**0.887**	**0.884**	**0.884**

Table 7: Experiment Results

starting point for assessing Italian news reliability in the health domain.

5 Conclusion and Future Work

In this paper, we present our preliminary work on the automatic reliability assessment of Italian news in the health domain. Our methodology is based on the use of trustworthy and untrustworthy sources and the definition and selection of a set of stylometric, lexical and sentiment features suitable for detecting misinformation and disinformation within health-related content. We believe that our approach can help improving the explainability of classification models thanks to our in-depth linguistic analysis. In addition, we also believe that the research community will be able to further exploit our annotated dataset to build upon this resource.

As future work, we intend to investigate further the linguistic features as well as the integration of information from external knowledge bases in order to check content manipulation. We also plan to extend our analysis to the whole news content and assign different weights to the features on the basis of their relevance and other linguistic and stylistic considerations related to this specific domain. Finally, we will investigate the integration of social media-related aspects, such as news network propagation, reach and engagement.

Acknowledgements

Luca Giordano has been supported by Borsa di Studio GARR "Orio Carlini" 2022/23 - Consortium GARR, the National Research and Education Network.
Maria Pia di Buono has been supported by Fondo FSE/REACT-EU - Progetti DM 1062 del 10/08/2021 "Ricercatori a Tempo Determinato di tipo A) (RTDA)". Azione IV.4 - Dottorati e contratti di ricerca su tematiche dell'innovazione/Azione IV.6 - Contratti di ricerca su tematiche Green.
The authors would like to thank Raffaele Manna for his support.

References

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2507–2511. IEEE.

Miguel A Alonso, David Vilares, Carlos Gómez-Rodríguez, and Jesús Vilares. 2021. Sentiment analysis for fake news detection. *Electronics*, 10(11):1348.

K Anoop, P Deepak, and VL Lajish. 2020. Emotion cognizance improves health fake news identification. In *IDEAS*, volume 2020, page 24th.

CH Basch, Patricia Zybert, Rachel Reeves, and CE Basch. 2017. What do popular youtubetm videos say about vaccines? *Child: care, health and development*, 43(4):499–503.

Bhavika Bhutani, Neha Rastogi, Priyanshu Sehgal, and Archana Purwar. 2019. Fake news detection using sentiment analysis. In *2019 twelfth international conference on contemporary computing (IC3)*, pages 1–5. IEEE.

Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. 2016. "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

M José Blanca Mena, Rafael Alarcón Postigo, Jaume Arnau Gras, Roser Bono Cabré, and Rebecca Bendayan. 2017. Non-normal data: Is anova still a valid option? *Psicothema, 2017, vol. 29, num. 4, p. 552-557*.

Alba Bonet-Jover. 2022. Veracity vs. reliability: Changing the approach of our annotation guideline.

Brandon Butcher and Brian J Smith. 2020. Feature engineering and selection: A practical approach for predictive models: by max kuhn and kjell johnson. boca raton, fl: Chapman & hall/crc press, 2019, xv+ 297 pp., $79.95 (h), isbn: 978-1-13-807922-9.

Liang Chen, Xiaohui Wang, and Tai-Quan Peng. 2018. Nature and diffusion of gynecologic cancer–related misinformation on social media: analysis of tweets. *Journal of Medical Internet Research*, 20(10):e11515.

Anshika Choudhary and Anuja Arora. 2021. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, 169:114171.

Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862.

Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2021. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518.

Lila J Finney Rutten, Kelly D Blake, Alexandra J Greenberg-Worisek, Summer V Allen, Richard P Moser, and Bradford W Hesse. 2019. Online health information seeking among us adults: measuring progress toward a healthy people 2020 objective. *Public Health Reports*, 134(6):617–625.

Asghar Ghasemi and Saleh Zahediasl. 2012. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2):486.

Katarína Greškovičová, Radomír Masaryk, Nikola Synak, and Vladimíra Čavojová. 2022. Superlatives, clickbaits, appeals to authority, poor grammar, or boldface: Is editorial style related to the credibility of online health messages? *Frontiers in Psychology*, page 5056.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. 2019. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*, 9(19):4062.

Dimitrios Panagiotis Kasseropoulos and Christos Tjortjis. 2021. An approach utilizing linguistic features for fake news detection. In *Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 International Conference, AIAI 2021, Hersonissos, Crete, Greece, June 25–27, 2021, Proceedings 17*, pages 646–658. Springer.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen. 2002. The importance of the normality assumption in large public health data sets. *Annual review of public health*, 23(1):151–169.

Cristiane Melchior and Mirian Oliveira. 2022. Health-related fake news on social media platforms: A systematic literature review. *new media & society*, 24(6):1500–1522.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Giancarlo Nicola. 2018. Bidirectional attentional lstm for aspect based sentiment analysis on italian. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12(108).

Francesco Pierri, Alessandro Artoni, and Stefano Ceri. 2020. Hoaxitaly: a collection of italian disinformation and fact-checking stories shared on twitter in 2019. *arXiv preprint arXiv:2001.10926*.

Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. 2019. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Giovanni Santia and Jake Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the international AAAI conference on web and social media*, volume 12, pages 531–540.

Steven F Sawyer. 2009. Analysis of variance: the fundamental concepts. *Journal of Manual & Manipulative Therapy*, 17(2):27E–38E.

Emanuel Schmider, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Bühner. 2010. Is it really robust? *Methodology*.

Anu Shrestha and Francesca Spezzano. 2021. Textual characteristics of news title and body to detect fake news: a reproducibility study. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 120–133. Springer.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca De Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Alice Tontodimamma, Lara Fontanella, Stefano Anzani, and Valerio Basile. 2022. An italian lexical resource for incivility detection in online discourses. *Quality & Quantity*, pages 1–19.

Marco Viviani and Gabriella Pasi. 2017. Credibility in social media: opinions, news, and health information—a survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 7(5):e1209.