FANDANGO

FANDANGO DELIVERABLE

Deliverable No.:	D3.4
Deliverable Title:	Ground truth development for FANDANGO system assessment
Project Acronym:	Fandango
Project Full Title:	FAke News discovery and propagation from big Data and artificial intelliGence Operations
Grant Agreement No.:	780355
Work Package No.:	WP3
Work Package Name:	Data gathering and preprocessing for Data Lake population(UPM)
Responsible Author(s):	David Martín Gutiérrez (UPM), Maria Camila Garcia(LVT), Daniele Vannella (LVT)
Date:	31.12.2019
Status:	v1.0 - Final Version
Deliverable type:	REPORT
Distribution:	PUBLIC

REVISION HISTORY

VERSION	DATE	MODIFIED BY	COMMENTS
V0.1	19.09.2019	David Martín Gutiérrez, Camila Garcia	First draft
V0.2	01.12.2019	Panagiotis Stalidis	Contributions, 2
V0.3	01.12.2019	DavidMartìn Gutierrez	Contributions, 3
V0.4	05.12.2019	Javier De Vega	Contribution, 1.1.2
V0.5	01.09.2019	Ilke Lemmelijn	Contribution, 1.1.2
V0.6	02.01.2020	Vittoriano Vancini	Contribution,1.1.2
V0.7	15.12.2019	Camila Garcia	Contribution 1.2.1
V0.9	03.01.2020	Daniele Vannella	Quality check
V1.0	08.01.2020	Panagiotis Stalidis	External review
V1.0	10.01.2020	Fulvio D'Antonio	Finalized document

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ABBREVIATION	DESCRIPTION
H2020	Horizon 2020
EC	European Commission
WP	Work Package
EU	European Union

EXECUTIVE SUMMARY

This document is a deliverable of the FANDANGO project funded by the European Union's Horizon 2020 (H2020) research and innovation programme under grant agreement No 780355.

It is a public report that describes the criteria that have been defined and used to collect data and build ground truth, the training sets, and the test sets that have been using to train and test all the machine learning models applied for the tasks described in WP4.

More specifically, Chapter 1 describe the criteria adopted to build the ground truth according to the needs of NLP based ML model. Chapter 2 describe the criteria adopted for media analysis. Finally, Chapter 3 describe the criteria adopted for source credibility analysis.

According to the iterative methodology in carrying out the activities of the project adopted by the Consortium, these criteria may be changed, if needed. All changes will be reported as annex in the following deliverables.

INTRODUCTION

Fake news denotes a type of yellow press which intentionally presents misinformation or hoaxes spreading through both traditional print news media and recent online social media. The amount of disseminated information and the rapidity of its diffusion make it practically impossible to assess reliability in a timely manner, highlighting the need for automatic online hoax detection systems. Many perspectives on who creates fake news, how and why it is created, how it propagates, and how it can be detected motivate the need for an in-depth analysis.

Fundamental human cognition and behavior theories developed across various discipline such as psychology, philosophy, social science, and economics provide invaluable insights for fake news analysis. Firstly, these theories introduce new opportunities for qualitative and quantitative studies of big fake news data, which to date, have been rarely available. Secondly, they facilitate building well-justified and explainable models for fake news detection and intervention, as well as introducing means to develop datasets that provide "ground truth" for fake news studies. It has been conducted a comprehensive literature survey across various disciplines and twenty well-known theories that can be potentially used to study fake news have been identified[1]:

		Term	Phenomenon
Style-based		*Undeutsch hypothesis* [Undeutsch 1967]	A statement based on a factual experience differs in content and quality from that of fantasy.
		Reality monitoring [Johnson and Raye 1981]	Actual events are characterized by higher levels of sensory- perceptual information.
		Four-factor theory [Zuckerman et al. 1981]	Lies are expressed differently in terms of arousal, behavior control, emotion, and thinking from truth.
Propagation-based		*Backfire effect* [Nyhan and Reifler 2010]	Given evidence against their beliefs, individuals can reject it even more strongly.
		Conservatism bias [Basu 1997]	The tendency to revise one's belief insufficiently when presented with new evidence.
		Semmelweis reflex [Bálint and Bálint 2009]	Individuals tend to reject new evidence because it contradicts with established norms and beliefs.
User-based (User's Engagement and Role)	Social influence	*Attentional bias* [MacLeod et al. 1986]	An individual's perception is affected by his or her recurring thoughts at the time.
		Validity effect [Boehm 1994]	Individuals tend to believe information is correct after repeated exposures.
		Bandwagon effect [Leibenstein 1950]	Individuals do something primarily because others are doing it.
		Echo chamber effect [Jamieson and Cappella 2008]	Beliefs are amplified or reinforced by communication and repetition within a closed system.
		Normative influence theory [Deutsch and Gerard 1955]	The influence of others leading us to conform to be liked and accepted by them.
		Social identity theory [Ashforth and Mael 1989]	An individual's self-concept derives from perceived membership in a relevant social group.
		Availability cascade [Kuran and Sunstein 1999]	Individuals tend to adopt insights expressed by others when such insights are gaining more popularity within their social circles
	Self-influence	*Confirmation bias* [Nickerson 1998]	Individuals tend to trust information that confirms their preexisting beliefs or hypotheses.
		Illusion of asymmetric insight [Pronin et al. 2001]	Individuals perceive their knowledge to surpass that of others.
		Naïve realism [Ward et al. 1997]	The senses provide us with direct awareness of objects as they really are.
		Overconfidence effect [Dunning et al. 1990]	A person's subjective confidence in his judgments is reliably greater than the objective ones.
	Benefit influence	*Prospect theory* [Kahneman and Tversky 2013]	People make decisions based on the value of losses and gains rather than the outcome.
		Valence effect [Frijda 1986]	People tend to overestimate the likelihood of good things happening rather than bad things.
		Contrast effect [Hovland et al. 1957]	The enhancement or diminishment of cognition due to successive or simultaneous exposure to a stimulus of lesser or greater value in the same dimension.

Figure 1: Different approaches to fake news detection

The aim of Fandango is to aggregate and verify different typologies of news data, media sources, social media, open data, so as to detect fake news and provide a more efficient and verified communication. Applying AI technologies allows to have promising solutions to tackle misinformation issues, one of the most important steps, when dealing with AI and machine learning models are the collection of data and all the criteria to gather and retrieve useful information to be used as datasets.

1. GROUND TRUTH FOR TEXT ANALYSIS

Fake news has become an important topic of research in a variety of disciplines including linguistics and computer Science. By definition, misinformation refers to the spread of false information (opposite to facts) in the context of news. In order to automatically detect instances of misinformation, one approach is to combine machine learning and computational linguistic techniques to build predictive models on labeled instances of news articles (fake/real or false/true items). Such models could then be applied to an unseen news article and score its veracity with respect to its linguistic characteristics.

An intuitive framing of the fake news problem in NLP and machine learning would like to understand how to classify news text into fake and legitimate instances. In the case of Fandango's project, it applies especially to the case of full text, as opposed to tweets or headlines distributed on social media – because text classification relies mainly on the linguistic characteristics of longer text. Deception detection in text has a broad literature in NLP, and fake news articles can be considered a category of deceptive text. Methods used for text classification vary from classic machine learning algorithms using a set of pre-defined linguistic features to modern neural network models which mainly rely on pre-trained word vectors and embedded representations resulting from processing large amounts of textual data.

The main challenge in this line of research is collecting quality data, i.e. , instances of fake and real news articles on a balanced distribution of topics.

Quality data is linked to the concept of Ground Truth , which is defined as factual data that have been observed or measured, and can be analyzed objectively. Ground Truth has not been inferred. If the data are based on an assumption, subject to opinion, or up for discussion, then, by definition, those are not Ground Truth data. In other words, Ground truth is a term used in statistics and machine learning that means checking the results of machine learning for accuracy against the real world.

In fake news detection context, it is hard to get stick to that definition since there has to be an assumption to declare an article as fake or not. Someone has to take a decision based on a criteria or a strong assumption; in fact, the first question we need to answer in addressing fake news detection through text classification is what to consider as a representative instance of fake news. In other domains related to deceptive text, such as fake product review detection, objective criteria can be designed when labelling the fake instances: a review written by someone who has not bought or used the product, or someone recruited by the product seller for the specific duty of writing a review would be considered fake. Fake news can also be defined as news articles written by amateurs (rather than journalists) recruited with the express purpose of generating content in favour or against an entity or policy, to promote a specific idea, or for financial gain such as attracting clicks for ads. This is one of the main reasons why datasets are often affected by bias or, rather, any trend or deviation from the truth.

It is usually difficult to know exactly what biases are present in a dataset, so the next best option is to think carefully about the process that generated the data. The two most common problems of collecting data for a ground truth are:

- ❖ Unbalanced data: determining the balance of ground truth is generally hard because in most cases the only information about ground truth comes from datasets which may contain biases.

- ❖ Simple Bias vs. label bias in data : Label bias occurs when the data-generating process systematically assigns labels differently for different groups.

Accuracy is often a component of fairness definitions, however optimizing for accuracy when data labels are biased can perpetuate biased decision making.

Sample bias occurs when the data-generating process samples from different groups in different ways. For example, an analysis of New York City's stop-and-frisk policy found that Black and Hispanic people were stopped by police at rates disproportionate to their share of the population (while controlling for geographic variation and estimated levels of crime participation).

A dataset describing these stops would contain sample bias because the process by which data points are sampled is different for people of different races. Sample bias compromises the utility of accuracy as well as ratio-based comparisons, both of which are frequently used in definitions of algorithmic fairness.

1.1. TRAINING SET IN FAKE NEWS DETECTION

Automatic fake news detection is a challenging problem in deception detection, and it has tremendous real-world political and social impacts. However, statistical approaches to combating fake news have been dramatically limited by the lack of labeled benchmark datasets. The problem of fake news detection is more challenging than detecting deceptive reviews, since the political language on TV interviews, posts on Facebook and Twitters are mostly short statements. However, the lack of manually labeled fake news dataset is still a bottleneck for advancing computational-intensive, broad-coverage models in this direction.

Although many people assume that instances of fake news are everywhere, in this project it's been challenging to compile a large enough dataset of reliably labelled fake news articles. In the following sections, there is a brief description of how data collection for fake news detection have been approached and carried out in some research work.

1.1.1 TRAINING SET USED IN ACADEMIC FIELD

Vlachos and Riedel (2014)[2] are the first to release a public fake news detection and fact-checking dataset, but it only includes 221 statements, which does not permit machine learning based assessments.

An interesting approach to build a training set is the one proposed in Automatic detection of fake news[3], where they started collecting a dataset of legitimate news belonging to six different domains (sports, business, entertainment, politics, technology, and education). The news was obtained from a variety of mainstream news websites (predominantly in the US) such as the ABCNews, CNN, USAToday, NewYorkTimes, FoxNews, Bloomberg, and CNET among others. To ensure the veracity of the news, they conducted manual fact-checking on the news content, which included verifying the news source and cross-referencing information among several sources. Using this approach, we collected 40 news in each of the six domains, for a total of 240 legitimate news.

Moreover, they collected fake news using crowdsourcing. So in order to generate fake versions of the news in the legitimate news dataset, they used crowdsourcing via Amazon Mechanical Turk, which has been successfully used in the past for collecting deception data on several domains, including opinion reviews and controversial topics such as abortion and death penalty.

However, collecting deceptive data via AMT showed additional challenges on the news domain. Firstly, the reporting language used by journalists might differ from AMT workers language (e.g., journalistic vs. informal

style). Secondly, journalistic articles are usually lengthier than consumer reviews and opinions, thus increasing the difficulty of the task for AMT workers as they would be required to read the full news article and create a fake version from it.

To address the former, the workers had to the extent possible to emulate a journalistic style in their writing. This decision was motivated by the 5th point of the fake news corpus guidelines described below in paragraph 1.1.2, which suggests to obtain news with homogeneous writing style.

An AMT task has been set up and workers had to generate a fake version of provided news. Each hit included the legitimate news headline and its corresponding body. Workers have been instructed to produce both a fake headline and a fake news body within the same topic and length as the original news.

Workers were also requested to avoid unrealistic content and to keep the names mentioned in the news. The fake news were produced by unique authors, as we allowed only a single submission per worker. It took approximately five days to collect 240 fake news. Each hit was manually checked for spam and to make sure workers followed the provided guidelines.

The final set of fake news consists of 31,990 words. Each fake news has on average 132 words and approximately 5 sentences. Table 1 shows a sample fake news, along with its legitimate version, in the technology domain.

LEGITIMATE	FAKE
Nintendo Switch game console to launch in March for $299 The Nintendo Switch video game console will sell for about $260 in Japan, starting March 3, the same date as its global rollout in the U.S. and Europe. The Japanese company promises the device will be packed with fun features of all its past machines and more. Nintendo is promising a more immersive, interactive experience with the Switch, including online playing and using the remote controller in games that don't require players to be constantly staring at a display. Nintendo officials demonstrated features such as using the detachable remote controllers, called "Joy-Con," to play a gun-duel game. Motion sensors enable players to feel virtual water being poured into a virtual cup.	**New Nintendo Switch game console to launch in March for $99** Nintendo plans a promotional roll out of it's new Nintendo switch game console. For a limited time, the console will roll out for an introductory price of $99. Nintendo promises to pack the new console with fun features not present in past machines. The new console contains new features such as motion detectors and immerse and interactive gaming. The new introductory price will be available for two months to show the public the new advances in gaming. However, initial quantities will be limited to 250,000 units available at the sales price. So rush out and get yours today while the promotional offer is running.

Figure 2: Sample legitimate and crowdsourced fake news

A new dataset called LIAR[4] has been introduced in May 2017 , which includes 12,836 short statements labeled for truthfulness, subject, context/venue, speaker, state, party, and prior history. Additionally, in contrast to crowdsourced datasets, the instances in LIAR are collected in grounded, more natural contexts, such as political debates, TV ads, Facebook posts, tweets, interviews, news releases, etc. In each case, the labeler provides a lengthy analysis report to ground each judgment, and the links to all supporting documents are also provided.

LIAR dataset includes 12.8k human labeled short statements from POLITIFACT's API, and each statement is evaluated by a POLITIFACT.COM editor for its truthfulness. After initial analysis, we found duplicate labels, and merged the full-flop, half-flip, no-flip labels into false, half-true, true labels respectively. We consider six fine-grained labels for the truthfulness ratings: *pants-fire, false, barely-true, half-true, mostly-true, and true*. The distribution of labels in the LIAR dataset is relatively well-balanced and compared to prior datasets, LIAR is an order of magnitude larger, which enables the development of statistical and computational approaches to fake news detection.

Statement: *"The last quarter, it was just announced, our gross domestic product was below zero. Who ever heard of this? Its never below zero."*
Speaker: Donald Trump
Context: presidential announcement speech
Label: Pants on Fire
Justification: According to Bureau of Economic Analysis and National Bureau of Economic Research, the growth in the gross domestic product has been below zero 42 times over 68 years. Thats a lot more than "never." We rate his claim Pants on Fire!

Figure 3: Random Sample from the LIAR dataset.

In May 2019, Torabi Asr and Taboada tuble published an academic paper[6], where they described how MisInfoText were compiled. It's a repository built with a focus on quality data collection and based on the continuous effort of fact-checking websites in finding and labelling instances of fake news. The full text of the news articles is available, together with veracity labels previously assigned based on manual assessment of the articles' truth content. They have conducted experiments on the data that we have so far collected to show the gaps and sources of data imbalance in the topics covered by fact-checkers section.

They considered the data collection strategy for building a fake news detection system based on the definition one adopts for the task. In the majority of previous work described, instances of fake news were collected from a list of suspicious websites. In this way, the resulting dataset is balanced across classes, and split into training, validation and test sets. However, the noisy strategy to label all articles of a publisher based on its reputation would bias a classifier trained on this data, limiting its power to distinguish individual truthful news articles from misinformation instances. In other words, data that were collected in this fashion would not be suitable for learning linguistic patterns of deception; it would rather help distinguish general writing style of a group of news websites (the rumour or clickbait style).

In order to build a text classification system to detect false from true content based on linguistic cues, they needed news articles assessed individually and labelled with respect to their level of veracity. This type of data collection is labour-intensive, as it involves fact-checking for each and every news article. A variety of fact-checking websites perform this analysis on real news. Therefore, one way to collect data on rumours and false news is to take advantage of these websites and to try to automatically scrape information such as the true vs. false headlines and hopefully their sources.

In order to address the lack of data with reliable labels, they started a repository of news articles texts that have been labelled by fact-checking websites. MinInfoText repository consists of three major sections:

- List of datasets collected by scraping fact-checking websites, whose basic function is to find and tag suspicious news.

- List of other datasets that have been published in previous NLP papers and are useful for building misinformation detection models.
- List of potential fact-checking websites that we have not yet used in our data collection effort but could be employed in future work.

Dataset	Size and type	Labelling system	Notes
Allcott and Gentzkow (2017)	156 news articles	5-Way (false to true)	Collected from Snopes, Politifact and Buzzfeed fact-checking pages, focused on 2016 US election
Ferreira and Vlachos (2016)	1612 news articles	2-Way (false/true)	Unbalanced, originally developed for stance-detection
Rubin et al. (2016)	360 news articles	2-Way (satirical/legitimate)	Balanced by topic and label. A variety of topics.
Zhang et al. (2018)	40 news articles	Multiple (credibility indicators)	Continuous effort with the future goal of annotating 10,000 articles
Pérez-Rosas et al. (2017)	480 news articles	2-Way (fake/legitimate)	Balanced by topic and label. Fake items were artificially generated by Turkers.
Pérez-Rosas et al. (2017)	200 news articles	2-Way (fake/legitimate)	Balanced by topic and label. Focused on celebrity stories.
Wang (2017)	12.8K short statements	6-Way (false to true)	Collected using the Politifact API
Thorne et al. (2018)	185K short statements and supporting/refuting Wikipedia documents	2-Way (original/mutated)	Originally developed for stance-detection. Mutated claims were artificially generated.

Figure 3.1: Currently available datasets with texts individually labelled for veracity.

1.1.2 GUIDELINES FOR FAKE NEWS DETECTION BASED ON TEXT

Some guidelines for a Fake News Corpus, have been proposed, in building a fake news dataset. Nine mine requirements of a fake news corpus proposed by [5](Rubin et al., 2016):

- include both fake and real news items
- contain text-only news items
- have a verifiable ground-truth
- be homogeneous in length
- be homogeneous writing style
- contain news from a predefined time frame
- be delivered in the same manner and for the same purpose (e.g. humor, breaking news) for fake and real cases
- be made publicly available
- should take language and cultural differences into account.

The ground-truth remains challenging since verification with absolute certainty is very effort and time consuming, whether all the content of real news items is in fact true.

In Fandango in the approach taken, there is no reason why to use a homogeneous writing style corpus, since the model would like to catch that information. In order to work around the problem for which a corpus can't contain news from a predefined times, a retrain of the model has been scheduled each time a slot of time or a slot of data stored is exceeded. For example after 10.000 new articles ingested in the system or after one month of collecting new articles.

1.2 FANDANGO'S GROUND TRUTH

Taking into consideration all of the above mentioned requirements, three different methods are chosen to annotate documents and create the so-called training dataset. In particular, all the articles in a training set have to belong to a category, in this case the labels are "legitimate" and "fake".

In the next 3 paragraphs the three strategies are explained in detail.

1.2.1 DATASETS AVAILABLE ONLINE

The first approach is to collect existing datasets applied in academic research, in order to have a comparison with state-of-art solutions. The first two datasets are available from a paper called Exploiting Tri-Relationship for Fake News Detection[5], including online articles whose veracities have been identified by experts. The other datasets come from kaggle.

In particular, The FakeNewsDatabase dataset contains news in six different domains: technology, education, business, sports, politics, and entertainment. The legitimate news included in the dataset were collected from a variety of mainstream news websites predominantly in the US such as the ABCNews, CNN, USAToday, NewYorkTimes, FoxNews, Bloomberg, and CNET among others. The fake news included in this dataset consists of fake versions of the legitimate news in the dataset, written using Mechanical Turk as described before.

The Celebrity dataset contains news about celebrities (actors, singers, socialites, and politicians). The legitimate news in the dataset was obtained from entertainment, fashion and style news sections in mainstream news websites and from entertainment magazines websites. The fake news was obtained from gossip websites such as Entertainment Weekly, People Magazine, RadarOnline, and other tabloid and entertainment-oriented publications. The news articles were collected in pairs, with one article being legitimate and the other fake (rumors and false reports). The articles were manually verified using gossip-checking sites such as "GossipCop.com", and also cross-referenced with information from other entertainment news sources on the web.

This collection of datasets have been aggregated together to compose a single one. This choice was dictated by the need to have a comparison with the state-of- art and all the new innovative approaches proposed by researchers.

Here the link of the datasets taken in this approach available only in english:

- www.buzzfeed.com
- www.kaggle.com/mrisdal/fake-news/data
- www.kaggle.com/jruvika/fake-news-detection
- github.com/GeorgeMcIntire/fake_real_news_dataset
- www.kaggle.com/sumanthvrao/fakenewsdataset

text	title	type
Print They should pay all the back all the mon...	Muslims BUSTED: They Stole Millions In Gov't B...	bias
Why Did Attorney General Loretta Lynch Plead T...	Re: Why Did Attorney General Loretta Lynch Ple...	bias
Red State : \nFox News Sunday reported this mo...	BREAKING: Weiner Cooperating With FBI On Hilla...	bias
Email Kayla Mueller was a prisoner and torture...	PIN DROP SPEECH BY FATHER OF DAUGHTER Kidnappe...	bias
Email HEALTHCARE REFORM TO MAKE AMERICA GREAT ...	FANTASTIC! TRUMP'S 7 POINT PLAN To Reform Heal...	bias
Print Hillary goes absolutely berserk! She exp...	Hillary Goes Absolutely Berserk On Protester A...	bias
BREAKING! NYPD Ready To Make Arrests In Weiner...	BREAKING! NYPD Ready To Make Arrests In Weiner...	bias
BREAKING! NYPD Ready To Make Arrests In Weiner...	WOW! WHISTLEBLOWER TELLS CHILLING STORY Of Mas...	bias
\nLimbaugh said that the revelations in the Wi...	BREAKING: CLINTON CLEARED...Was This A Coordin...	bias

Figure 4: Fake news sample from kaggle dataset

Body	Headline	Label
Image copyright Getty Images\nOn Sunday mornin...	Four ways Bob Corker skewered Donald Trump	1
LONDON (Reuters) - "Last Flag Flying", a comed...	Linklater's war veteran comedy speaks to moder...	1
The feud broke into public view last week when...	Trump's Fight With Corker Jeopardizes His Legi...	1
MEXICO CITY (Reuters) - Egypt's Cheiron Holdin...	Egypt's Cheiron wins tie-up with Pemex for Mex...	1
Country singer Jason Aldean, who was performin...	Jason Aldean opens 'SNL' with Vegas tribute	1
JetNation FanDuel League; Week 4\n% of readers...	JetNation FanDuel League; Week 4	0
In 2012, Kansas lawmakers, led by Gov. Sam Bro...	Kansas Tried a Tax Plan Similar to Trump's. It...	1

Figure 5: Kaggle dataset with fake and legitimate news

1.2.2 SEMI-AUTOMATIC ANNOTATION SYSTEM

Semi-automatic annotation method consists in making a list, verified by expert journalists, supporting some sources labeled as legitimate and some others as fake.

This list is composed by 10 source domains for each language(available in the following link) ,equally divided into the two categories requested. Below are reported all the domains for each language.

GOOD DOMAINS	BAD DOMAINS
www.ansa.it	www.il-giornale.info
www.repubblica.it	www.tg24-ore.com
www.corriere.it	tg-news24.com
www.ilsole24ore.com	il-quotidiano.info
www.rainews.it	www.lavocedelpatriota.it

Table 1: Source domains italian

GOOD DOMAINS	BAD DOMAINS
elpais.com	www.mediterraneodigital.com
elmundo.es	www.elespiadigital.com
elconfidencial.com	www.alertadigital.com
www.rtve.es/noticias/	latribunadecartagena.com
cadenaser.com	casoaislado.com

Table 2: Source domains in spanish

GOOD DOMAINS	BAD DOMAINS
www.vrt.be/vrtnws/nl	www.ninefornews.nl
www.standaard.be	www.dagelijksestandaard.nl
www.tijd.be	jdreport.com
www.nrc.nl	revolutionaironline.com
nos.nl/nieuws	www.wanttoknow.nl

Table 3: Source domains in dutch

GOOD DOMAINS	BAD DOMAINS
www.apnews.com	www.infowars.com
www.washingtonpost.com	sputniknews.com
www.wsj.com	russia-insider.com/en/
www.theguardian.com	en.news-front.info
www.reuters.com	www.express.co.uk/news/politics

Table 4: Source domains in english

As mentioned before, all the domains contained in the list have been chosen by journalist partner involved in the project. The justification of their choice is provided below, VRT collected dutch sources, ANSA collected italian sources and CIVIO collected spanish sources, all together collected english sources.

PARTNER	JUSTIFICATION
ANSA	The “good” news websites indicated are known for their accuracy and adherence to the facts. For this reason, they are among the most red in Italy. They also publish news distributed by ANSA that as a News Agency is set to high standards of quality check. The websites labeled “bad” according to an internal Fandango convention for ease of use, are among those indicated by Italian Fact-Checkers as the most prone to bias, disinformation and fake news.
VRT	VRT is a public broadcaster that has a social role to inform Flemish readers/viewers/listeners correctly. All topics provided on VRT channels are fact checked thoroughly. The other websites indicated as “good domains” are websites that are trusted by our journalists. The websites indicated as “bad domains” are websites that claim to publish news but are written in a very subjective, satirical or politically tinted way. Therefore, these are sources that our journalists don’t trust.
CIVIO	The sites labeled as "good domains" are long-established legacy media with standards of journalistic ethics. They apply the basic standards i.e disclosing the sources of their information, separating editorial content from news content. The sites labeled as "bad domains" include sites with an appearance of news publishers but have direct ties with political parties or extreme-right organizations. One of them is directly run by a extreme-right party currently concurring in electoral processes. All of them spread propaganda - including racist and homophobic messages- and promote their own political or activist platforms. Some of them are being or have been sued for defamation, hate speech and other issues. One of these sites recently closed its activities with its director leaving the country with a pending judiciary charge.

Table 5: Journalist justification of the choice of reliable and unreliable websites

Since datasets available online are very clean/well-preprocessed, this method was needed in order to start crawling real-dirty data. In big data field one of the most important needs is about the strategies to manage large volume of structured, unstructured and semi-structured data and almost all big data sets are dirty i.e. the set may contain inaccuracies, missing data, miscoding and other issues that influence the strength of big data analytics.

One of the biggest challenges in big data analytics is to discover and repair dirty data; failure to do this can lead to inaccurate analytics and unpredictable conclusions. Data cleaning is an essential part of managing and analyzing data. In order to evaluate this aspect in data, data quality needs to be mentioned. Data quality is generally described as the capability of data to satisfy stated and implied needs when used under specified conditions.

When dealing with real-world data, dirty data is the norm rather than the exception. It's continuously needed to predict correct values, impute missing ones, and find links between various data artefacts such as schemas and records. Data cleaning can't be considered as a piecemeal exercise (resolving different types of errors in isolation), and instead leverage all signals and resources to accurately predict corrective actions.

Moreover, data quality is a combination of data content and form. Where data content must contain accurate information and data form essentially be collected and visualized in an approach that creates data functioning. Content and form are significant considerations to reduce data mistakes, as they shed light on the fact that the task of repairing dirty data needs beyond simply providing correct data.

Inaccurate data refers to any field containing wrong values. A right value of data will bring accurate and signified arrangement of consistency and unambiguousness.

Incomplete data from missing data is produced by data sets basically missing values. These types of data are considered concealed when the amount of values is identified in a set, but the values themselves are unidentified, and it is also known to be condensed when there are values in a set that are eliminated. Inconsistent data arise as a consequence of data redundancy; i.e. same data value is stored in different files which may be in different formats. Duplicate data, are identical data entries that have been added by a system multiple times.

The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making. While it has been the focus of many researchers for several years, individual problems have been addressed separately. These include missing value imputation, outliers detection, transformations, integrity constraints violations detection and repair, consistent query answering, deduplication, and many other related problems such as profiling and constraints mining. In general, cleaning data may be time-consuming, but lots of tools have cropped up to make this crucial task more bearable. The Python community offers a host of libraries for making data orderly and legible—from styling DataFrames. So within this project the libraries that are used for data cleaning are in Python, one of them is Dora, which is designed for exploratory analysis, specifically, automating the most painful parts of it, like data cleaning. Cleansing functions include:

- Reading data with missing and poorly scaled values
- Imputing missing values
- Scaling values of input variables

1.2.3 MANUAL ANNOTATION SYSTEM:

Manual annotation system is a strategy decided in Fandango's project to avoid bias in the articles to consider as a training set. The bias can be well-explained as follows:

- Sample Bias : Sampling bias in training data has to be considered as result of human input. Since Machine learning models are predictive engines that train on a large mass of data based on the past. They are made to predict based on what they have been trained to predict. These predictions are only as reliable as the human collecting and analyzing the data. The decision makers have to remember that if humans are involved in any part of the process, there is a greater chance of bias in the model.

  The sample data used for training has to be as close a representation of the real scenario as possible. There are many factors that can bias a sample from the beginning and those reasons differ in each domain(i.e. business, security, medical, education).

- Prejudice Bias: it is a cause of human input. Prejudice occurs as a result of cultural stereotypes in the people involved in the process. Social class, race, nationality, gender can creep into a model that can completely and unjustly skew the results of your model. Involving some of these factors in statistical modelling for research purposes or to understand a situation at a point in time is completely different to predicting who should get a loan when the training data is skewed against people of a certain race, gender and/or nationality.

- Confirmation bias: It is a well-known bias that has been studied in the field of psychology and directly applicable to how it can affect a machine learning process. If the people of intended use have a pre-existing hypothesis that they would like to confirm with machine learning the people involved in the modelling process might be inclined to intentionally manipulate the process towards finding that answer.

- Group attribution bias: This type of bias results from a model trained with data that contains an asymmetric view of a certain group. For example, in a certain sample dataset if the majority of a certain gender would be more successful than the other or if the majority of a certain race makes more than another, the model will be inclined to believe these falsehoods. There is label bias in these cases. In actuality, these sorts of labels should not make it into a model in the first place. The sample used to understand and analyse the current situation cannot just be used as training data without the appropriate pre-processing to account for any potential unjust bias.

In Fandango's project it's been implemented a dedicated user interface, where journalist can annotate data and consequently try to avoid all the issues with bias on data, mentioned above. Basically a manual annotation is provided by users, through this interface linked to Fandango's data silos. All the articles proposed to the users to get annotated,come from a list of domains. This list contains websites that have a considerable rate of published fake news. 1000 articles per language have to be read, written and annotated by final users that are professional journalists.

To complete this task, journalists have 2 different options how to annotate articles, facilitating as much as possible this procedure. The first annotation option allows a user to retrieve an article from Fandango's database, read it, assign it a label and save it. In the second one instead, users can copy and paste a url, assign a label to the article and save it in the database. Each article has to be annotated by 3 different annotators, in order to validate the fairness of the data annotated.

The most frequent result for each news is taken as label. Articles returned to the final user to get annotated, come from a list made by journalists themselves. This list contains websites to be under control because they are not declared fake but have a considerable rate of suspicious articles.

Language	Domain to annotate
IT	www.ilgiornale.it
IT	www.liberoquotidiano.it
IT	www.ilfattoquotidiano.it
ES	www.larazon.es
ES	www.lavanguardia.com
ES	www.abc.es
ES	www.eldiario.es
ES	www.elespanol.com
EN	www.thesun.co.uk
EN	www.dailymail.co.uk
EN	www.breitbart.com
EN	cen.at
EN	www.defendevropa.com/2017/news/electoral-commission-arron-banks
EN	voiceofeurope.com
EN	gellerreport.com
EN	www.zerohedge.com
EN	v4report.com
NL	hln.be
NL	www.nieuwsblad.be
NL	newsmonkey.be
NL	www.gva.be
NL	www.nd.nl
NL	www.hbvl.be

NL	www.rtlnieuws.nl
NL	www.nu.nl
NL	www.dewereldmorgen.be
NL	nl.metrotime.be/news
NL	www.mo.be
NL	nieuws.vtm.be
NL	nl.express.live
NL	www.apache.be
NL	doorbraak.be
NL	www.krapuul.nl
ES	okdiario.com
ES	www.huffingtonpost.es
ES	publico.es
ES	libertaddigital.com
ES	periodistadigital.com
ES	vozpopuli.com
ES	elindependiente.com
ES	elplural.com
ES	lamarea.com
ES	www.elsaltodiario.com
ES	europapress.es
ES	efe.com
ES	www.cope.es
ES	lavozdegalicia.es
ES	elperiodico.com
ES	elcorreo.com

Table 6: Web-site where articles to annotate will be drawn

There are two main points to check the fairness of the data annotated. The first one is the insertion of articles, that are validated and verified to be true, these articles are considered as flags, they should be annotated by users as true. The second one is the calculation of Cohen's kappa coefficient. It's a statistic that is used to measure inter-rater reliability (and also Intra-rater reliability) for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as Coheńs kappa coefficient takes into account the possibility of the agreement occurring by chance. There is controversy surrounding Cohen's kappa due to the difficulty in interpreting indices of agreement. Some researchers have suggested that it is conceptually simpler to evaluate disagreement between items. So it measures the agreement between two raters who each classify N items into C mutually exclusive categories.

$$k \equiv \frac{p_o - p_\epsilon}{1 - P_\epsilon} = 1 - \frac{1 - P_o}{1 - P_\epsilon}$$

where p_o is the relative observed agreement among raters (identical to accuracy), and $p_\epsilon$is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. If the raters are in complete agreement then κ = 1. If there is no agreement among the raters other than what would be expected by chance(as given by p e),κ = 0. It is possible for the statistic to be negative, which implies that there is no effective agreement between the two raters or the agreement is worse than random.

Since Fandango has 3 reference contexts, all the articles proposed to the final users are subjected to a filter to discriminate off topics news, using techniques that take inspiration from the famous words representation of documents, word-to-vect and sense-embedding.

This filter is built by initially calculating a meaning for each news article, converting the article in a list of vectors. Each word in the article is represented in the vector by its correlated words. Then, the centroid of the vectors of words that appear in the article is calculated in order to generate a principal vector representing the article. Finally, the distance between the article representation vector and the topics representation vectors (filter) is measured. A detailed description of what is required and the necessary steps is:

- Websites list
- Bag-of-words used as topic filter

Steps:

1. Words transformation into vector(300 feature) from a pre-trained vocabulary.
2. Bag-of-words creation to use as topic filter.

   Let W_t be a topic, where $t = 1,2,3.$Considering 3 topics
3. Convert words into vectors correspondents in the vocabulary.

   $F(W) \rightarrow V$ where $V = (v_1, v_2, v_3, v_4, \ldots.., v_{300})$
4. Convert each word in a document as described before

   Let D be a document so we can represent it as $D = (V_1, V_2, V_3, \ldots, V_l)$

   and l represents the total number of words in D

   let replace each $V_i, i = 1,2,3, \ldots, l$

The result is a matrix M with this shape $300\ x\ l$ like the one below

$$\begin{bmatrix} v_1^1 & v_1^2 & . & . & . & v_1^l \\ v_2^1 & v_1^2 & . & . & . & v_2^l \\ . & . & . & . & . & . \\ . & . & & & & \\ v_{300}^1 & v_{300}^2 & . & . & & v_{300}^l \end{bmatrix}$$

5. Calculate a centroid for the document, taking as result the mean for each row

$$c_D = \frac{1}{l}\sum_{j=1}^{l} \quad v_j$$

At the end, a centroid vector will be obtained

$$C_D \ = \ (c_1, c_2, c_3, \ldots, c_{300})$$

6. Compare vector C_D with the bag-of-words for each topic, using a cosine similarity:

$$cosine\ similarity(C_D\ * W_t)$$

Set a threshold for the result of the measure above, and take all the documents that overcome that established values, for example if the threshold is 0.7 and result of the cosine similarity is 0.5 that document will be rejected. This procedure, at the moment, doesn't consider the present of oov, but we're aware of this problem.

Bag of words for each topic and for each language are reported from table 6 to 9:

Language	Word	Topic
NL	klimaat	Climate
NL	klimaatopwarming	Climate
NL	koolstofdioxide	Climate
NL	klimaatspijbelaar	Climate
NL	groene stroom	Climate
NL	klimaatstaking	Climate
NL	hernieuwbare energie	Climate
NL	CO2	Climate
NL	broeikasgas	Climate
NL	fossiele brandstof	Climate
NL	migratie	Immigration
NL	asiel	Immigration
NL	bootvluchteling	Immigration
NL	illegaal	Immigration

NL	migrant	Immigration
NL	allochtoon	Immigration
NL	migratieachtergrond	Immigration
NL	grens	Immigration
NL	vluchteling	Immigration
NL	verblijfsvergunning	Immigration
NL	parlement	EU
NL	lobbyen	EU
NL	Europarlementslid	EU
NL	commissie	EU
NL	belastingen	EU
NL	verkiezingen	EU
NL	liberaal	EU
NL	nationalist	EU
NL	partij	EU
NL	populisme	EU
NL	Europese Unie	EU
NL	Europese Commissie	EU

Table 7: words to create a filter of topics in dutch

Language	Word	Topic
ES	clima	Climate
ES	calentamiento global	Climate
ES	dióxido de carbono	Climate
ES	fracking	Climate
ES	cambio climático	Climate
ES	modelo energético	Climate
ES	energías renovables	Climate
ES	CO2	Climate
ES	efecto invernadero	Climate
ES	combustibles fósiles	Climate
ES	migración	Immigration
ES	asilo	Immigration

ES	ruta migratoria	Immigration
ES	trata	Immigration
ES	emigrante	Immigration
ES	inmigrante	Immigration
ES	Frontex	Immigration
ES	frontera	Immigration
ES	refugiado	Immigration
ES	permiso de residencia	Immigration
ES	Parlamento Europeo	EU
ES	lobby	EU
ES	Europarlamentario	EU
ES	Comisión Europea	EU
ES	impuestos	EU
ES	elecciones europeas	EU
ES	euroescéptico	EU
ES	mercado único	EU
ES	moneda única	EU
ES	Eurozona	EU
ES	Partido Popular Europeo	EU
ES	ALDE	EU
ES	Unión Europea	EU
ES	UE	EU
ES	visado	Immigration
ES	migrante	Immigration
ES	Open Arms	Immigration
ES	éxodo	Immigration
ES	migrantes	Immigration
ES	refugiados	Immigration
ES	emigrantes	Immigration
ES	inmigrantes	Immigration
ES	Samos	Immigration
ES	Lesbos	Immigration

Table 8: Words to create a filter of topics in spanish

Language	Word	Topic
IT	UE	EU
IT	Unione Europea	EU
IT	Consiglio europeo	EU
IT	Commissione europea	EU
IT	Europarlamento	EU
IT	Parlamento europeo	EU
IT	Europarlamentare	EU
IT	Filoeuropeista	EU
IT	Euroscettico	EU
IT	PPE	EU
IT	S&D	EU
IT	Bilancio UE	EU
IT	Eurozona	EU
IT	Euro	EU
IT	Moneta unica	EU
IT	Mercato unico	EU
IT	Crisi UE	EU
IT	Euroburocrati	EU
IT	Euroburocrazia	EU
IT	Brexit	EU
IT	Parlamentare europeo	EU
IT	Filoeuropeismo	EU
IT	Euroscetticismo	EU
IT	TFEU	EU
IT	TEU	EU
IT	Trattato di Roma	EU
IT	Banca Centrale Europea	EU
IT	BCE	EU
IT	Banca d'investimento europea (EIB)	EU
IT	Alto rappresentante dell'Unione per gli affari esteri e la politica di sicurezza	EU
IT	Profughi	EU
IT	Rifugiati	EU

IT	Asilo politico	EU
IT	Immigrati illegali	EU
IT	Attentati terroristici	EU
IT	Terroristi	EU
IT	Terrorismo	EU
IT	Estremisti	EU
IT	Islamizzazione UE	Immigration
IT	Paradiso Rifugiati	Immigration
IT	Attentati terroristici	Immigration
IT	Terrorismo	Immigration
IT	Estremismo anti UE	Immigration
IT	Suprematisti bianchi	Immigration
IT	Trattato di Dublino	Immigration
IT	Trattato di Schengen	Immigration
IT	Trattato di Maastricht	Immigration
IT	Mercato emissioni CO2 ETS	Climate
IT	Anidride carbonica	Climate
IT	Energia verde (Green)	Climate
IT	Fonti sostenibili	Climate
IT	Protocollo di Kyoto	Climate
IT	Accordo di Parigi	Climate
IT	Surriscaldamento	Climate
IT	Inquinamento	Climate
IT	Polveri sottili	Climate
IT	Energie rinnovabili	Climate
IT	Fotovoltaico	Climate
IT	Eolico	Climate
IT	Fracking	Climate

Table 9: Words to create a filter of topics in italian

Language	Word	Topic
EN	EU	EU
EN	European Union	EU
EN	European Council	EU
EN	European Parliament	EU
EN	European Commission	EU
EN	Member of the European Parliament	EU
EN	MEP	EU
EN	European Commissioner(s)	EU
EN	President of the European Commission	EU
EN	European People's Party (EPP)	EU
EN	Progressive Alliance of Socialists and Democrats (S&D)	EU
EN	EU Budget	EU
EN	Eurozone	EU
EN	Euro crisis	EU
EN	Single currency	EU
EN	Single market	EU
EN	Free trade	EU
EN	Bureaucracy	EU
EN	Bureaucratic monster	EU
EN	EU Cohesion	EU
EN	Brexit	EU
EN	Eurodeputy	EU
EN	Euroscepticism	EU
EN	pro-Europeanism	EU
EN	Eurosceptics	EU
EN	High Representative of the Union for Foreign Affairs and Security Policy	EU
EN	President of the European Council	EU
EN	TFEU	EU
EN	Treaty on the Functioning of the European Union	EU
EN	TEU	EU
EN	Treaty on European Union	EU
EN	Treaty of Rome	EU
EN	European Central Bank (ECB)	EU

EN	European Investment Bank (EIB)	EU
EN	Maastricht Treaty	Immigration
EN	Schenghen ConventionTrearty	Immigration
EN	Schenghen Agreement	Immigration
EN	Islamisation	Immigration
EN	Refugees Paradise	Immigration
EN	Refugee Crisis	Immigration
EN	Illegal entries	Immigration
EN	Terrorist attacks	Immigration
EN	Terrorism	Immigration
EN	Extremists	Immigration
EN	White Supremacists	Immigration
EN	Dublin Regulation Treaty	Immigration
EN	Kyoto Protocol	Climate
EN	Paris Agreement	Climate
EN	European cabon emission trading system (ETS)	Climate
EN	EU Air Pollution Control	Climate
EN	Sustainable Renowable Sources	Climate
EN	Particulate emissions control	Climate
EN	Carbon footprint	Climate
EN	Carbon dioxide	Climate
EN	Photovoltaic plant	Climate
EN	CO2 emissions	Climate
EN	Blue energy	Climate
EN	Solar plant	Climate
EN	Wind farm	Climate
EN	Fracking	Climate

Table 10: Words to create a filter of topics in english

USER INTERFACE FOR DATA ANNOTATION

As mentioned above, to facilitate the data annotation made by the final users, a dedicated interface has been implemented. The annotation interface has the same layout of the first Fandango's Platform version (more information in deliverable 5.2).

The first functionality of the annotation platform allows users to retrieve an article from Fandango's silos. The articles proposed to be annotated come from the websites defined as "to monitor" by the end-users themselves.

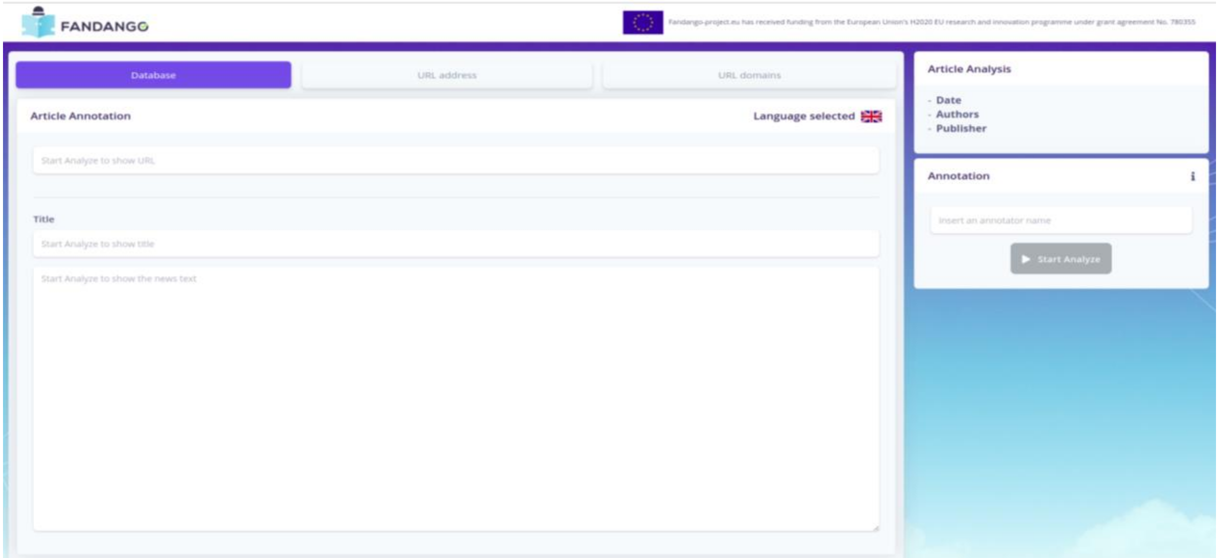

Figure 6 : Annotation User Interface

Once a user inserts her name and clicks on "start analyze", the system proposes an article to get annotated. The information available from the article are its headline/title, its body, its publisher and authors (when available). There are two options when the article appears. The user can read and annotate the presented article or skip it.Alternately, the user can copy and paste directly a link connected to a news item of her choice and assign a label to it. An important detail about the annotation rules is that an article is considered annotated when three annotators assigned to it a label, so it could happen that the system proposes again articles that a user skipped before.

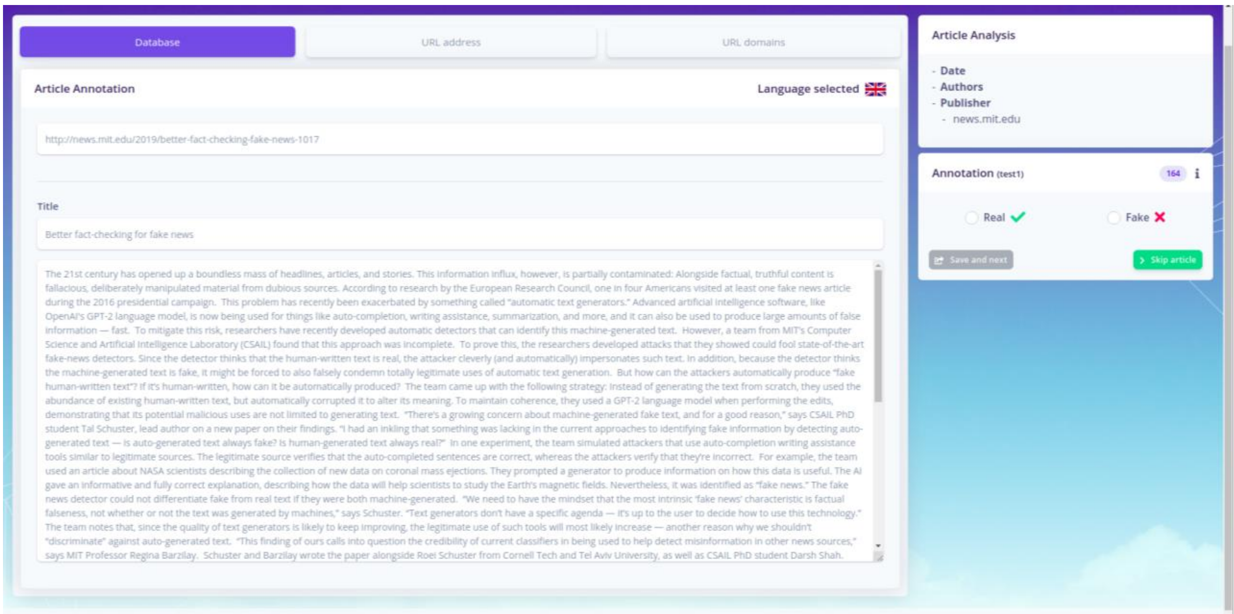

Figure 7 : Annotation User Interface in use

As shown in figure, the system is provided with a counter, which keeps track of the total number of annotated articles.

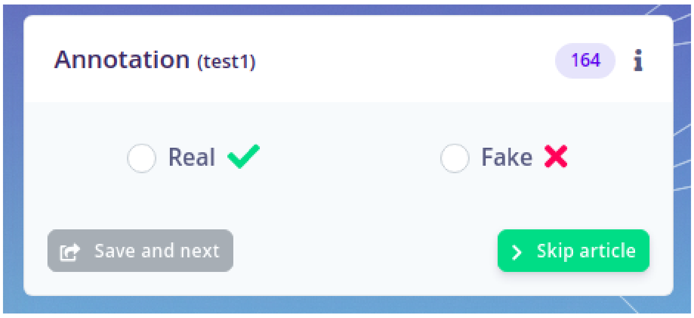

Figure 8 : Counter of the Annotation User Interface

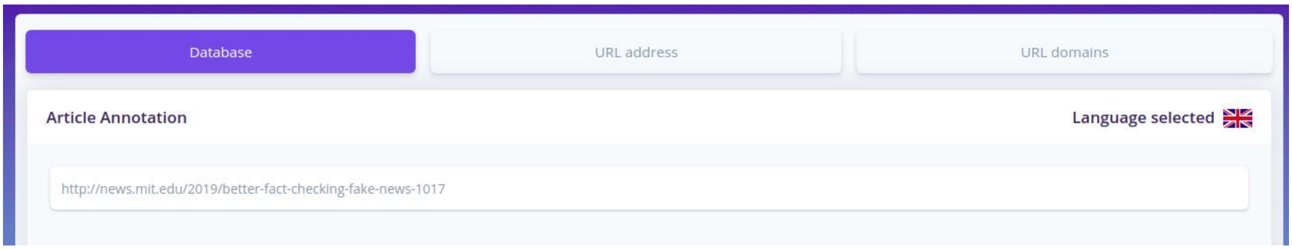

Figure 8.1 : language selector of the Annotation User Interface

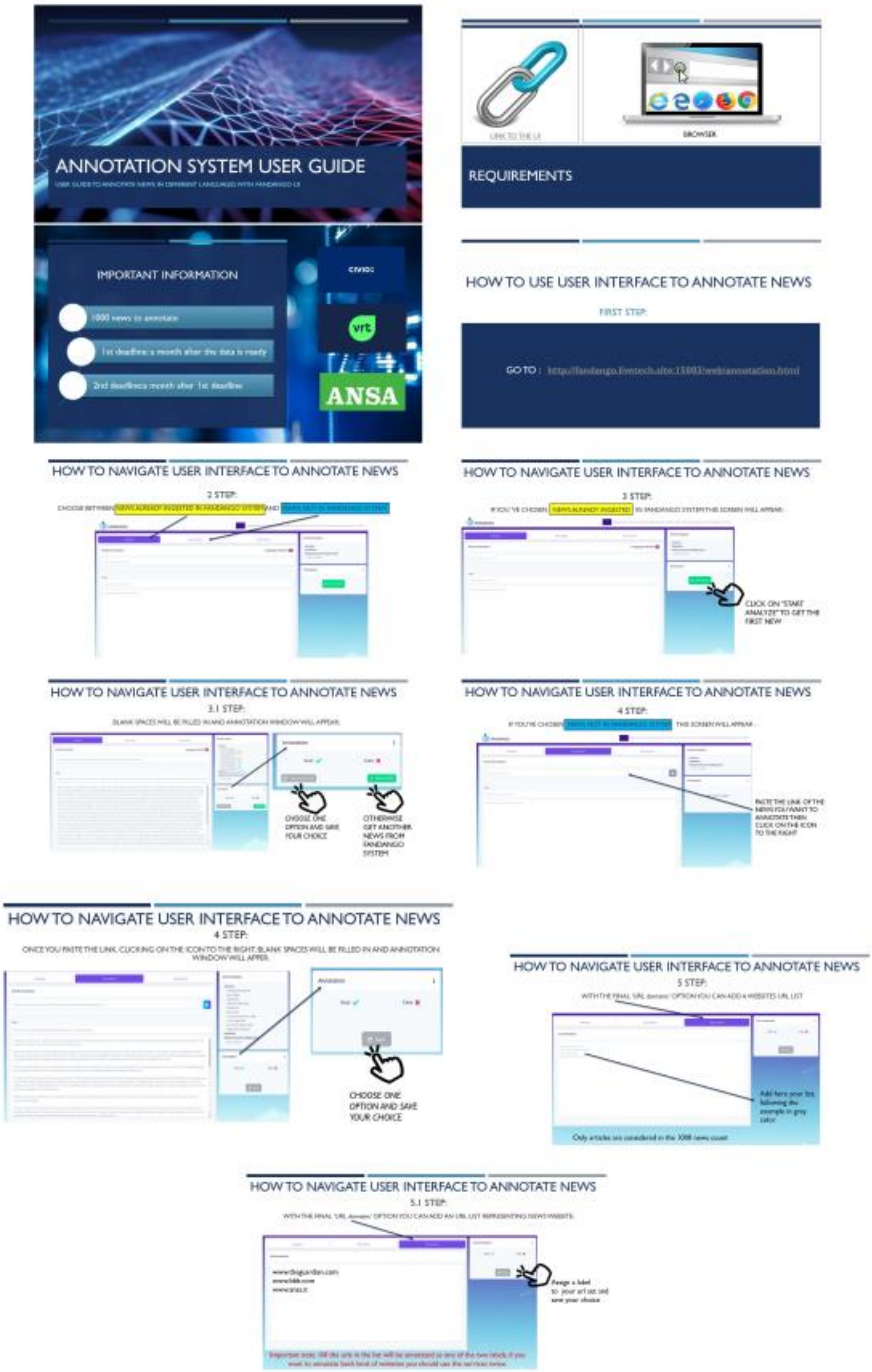

Figure 9: User Manual of Annotation User Interface

2. GROUND TRUTH FOR MEDIA ANALYSIS

One of the reasons to characterize a news article as fake or misleading is to include manipulated media. The media analysis tools that are implemented for the Fandango platform provide to the end users indications on the authenticity of media that are included in the examined articles. To build reliable and robust forgery detection models, training and testing data are very important. A good dataset has to include a large number of examples that cover all possible types of manipulation that are covered by the model as well as negative examples, i.e. media that are authentic and have not been manipulated. Additionally, subjectivity and bias should be absent. The major categories of media analyses that are provided for the Fandango project are: object detection, media manipulation and image localization.

During the last decade the extensive use of media has led to an increase in the volume of media forgeries which in turn has kindled scientific interest in the field of media forensics. Due to this interest, a number of media forensics challenges have sprouted targeting the detection of various types of attacks like image splicing and copy-move and video frame interpolation. Each competition provides training and testing datasets as well as a set of metrics for comparing methodologies. In the Fandango project the evaluation of each media manipulation analysis is performed by using an available public dataset provided by such a challenge.

Object detection and segmentation are research problems in the field of visual computing for over 20 years. Therefore, huge datasets and standard evaluation metrics exist and are used by Fandango. On the other hand, localization of images has gained a lot of traction recently with the publication of the second version of Google landmarks dataset and the execution of the Google Landmarks Challenge.

A detailed description of the datasets used for training and evaluating the machine learning models is provided in the description of each of the models in deliverables *D4.1 Spatio-temporal analytics and out of context fakeness markers prototypes* and *D4.3 Copy-move detection on audio-visual content prototypes*.

3. GROUND TRUTH FOR SOURCE CREDIBILITY (GRAPH ANALYSIS)

Analysing the set of publishers and authors that post and spread the news in the distinct networks considered in the project is crucial to detect whether the information is suspicious or reliable. Consequently, the groundtruth must take into consideration this aspect when it is built. To do so, a set of publishers must be labelled using the proper definitions that have been both described and discussed along this deliverable in order to be consistent to the rest of the analyzers. Using this groundtruth, the validation of the proposed models for measuring the source credibility for a certain publisher will be calculated using a set of error metrics that were properly described in *Deliverable 4.4: Source Credibility, Profiling and Social Graph Analytics Prototypes*.

On the other hand, the groundtruth building arises several problems for authors since there is no way to manually labelled a large number of authors because of the excessive time consuming and public datasets are only labelled considering the domain of the article but not the authors involved in it.

To mitigate this problem, authors associated to the same publisher will be validated using the groundtruth value of such publisher. Following this approach, there is no need to manually annotate a large set of authors and the committed error will be lower since it is assumed that every author that writes in a certain source is either as reliable or suspicious as the media source where he/she has published.

Consequently, the groundtruth for measuring the quality of the models regarding the credibility of a certain source or publisher will be built using both manual and semi-automatic annotation systems that are aforementioned described with more details. For more details regarding the procedures for obtaining the credibility of a source see *Deliverable 4.4: Source Credibility, Profiling and Social Graph Analytics Prototypes*.

BIBLIOGRAPHY

[1] Fake News: A Survey of Research, Detection Methods, and Opportunities - XINYI ZHOU and REZA ZAFARANI, Syracuse University, USA -December 2018.

[2] Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. Proceedings of the ACL 2014 Workshop on Language Technology and Computational Social Science.

[3] Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea. 2017. Automatic Detection of Fake News.

[4] William Yang Wang. 2017 “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection.

[5] Rubin, Victoria & Conroy, Nadia & Chen, Yimin & Cornwell, Sarah. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. 10.18653/v1/W16-0802.

[6] Torabi Asr, F., & Taboada, M. (2019). Big Data and quality data for fake news and misinformation detection. Big Data & Society.