This project has received funding from the European Union's Horizon 2020 research and innovation programme under **grant agreement No 780355**

European Commission

FANDANGO

FANDANGO DELIVERABLE

Deliverable No.:	D4.2
Deliverable Title:	Multilingual text analytics for misleading messages detection prototypes
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.:	4
Work Package Name:	Fake News identifiers, machine learning and data analytics
Responsible Author(s):	Daniele Vannella, Fulvio D'Antonio, Camila Garcia
Date:	31.07.2019
Status:	v1 - Final
Deliverable type:	REPORT
Distribution:	PUBLIC

REVISION HISTORY

VERSION	DATE	MODIFIED BY	COMMENTS
V0.1	19.07.2019	Camila Garcia	First draft
V0.2	25.07.2019	Daniele Vannella	contributions
V0.3	29.07.2019	Daniele Vannella	contribution
V0.4	30.07.2019	Fulvio D'Antonio	First review
V1	8.08.2019	Daniele Vannella Saverio Gravina	Final review

TABLE OF CONTENTS

ABBREVIATIONS

ABBREVIATION	DESCRIPTION
H2020	Horizon 2020
EC	European Commission
WP	Work Package
EU	European Union
ML	Machine Learning
NLP	Natural Language Processing
NLI	Natural Language Inference
SVM	Support Vector Machine
LIWC	Language Inquiry Word Count

LIST OF FIGURES

List of Tables

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 how the task of multilingual text analytics which aims to detect misleading messages has been faced in this project.

The main focus of this deliverable is the description of a statistical learning model to classify news and evaluate how trustworthy they are. This process starts from a survey of the state of the art in fake detection, continues with the analysis of stylometric and syntactic features and finish with the report of the classification algorithms tested and their performance evaluations.

Introduction

Machine learning (ML) and Natural Language Processing (NLP) tools can be applied to prevent misinformation issues globally and reduce its impact.

Artificial intelligence technologies promise to significantly automate certain activities that fact-checkers and journalists conduct to determine whether an article reports accurate facts or is instrumentalized in spreading deceptive messages to manipulate people's conscience and damage the image of individuals, companies and public bodies.

However, assessing the veracity of a news story is a complex task. Even Artificial Intelligence implementations require a multi-method approach, and the process for such analysis can be divided into sub-tasks that are facilitated by standard techniques and open source libraries.

FANDANGO textual analysis deals with trustworthiness analysis using an agnostic approach. it aims to analyze the news without using the semantic of the words, taking advantage of supporting terms in multilingual and cross-domains tasks.

This document reports:

- Exploration of Fake news detection state-of-art
- Fandango text analysis approach
- Strategies for the creation of the Ground truth
- Experiments results
- Conclusion

1. Fake news detection

Fake news recognition has a particular growing interest due to the increment of online misinformation. Fandango will face the misinformation in four different ways:

1. considering textual content of an article and analysing its syntax and many linguistic features
2. building a graph to collect and connect authors and organizations in order to apply some graph analysis
3. analysing videos editing
4. analysing images editing

Fundamental theories of human cognitive and behavioural science provide invaluable insights for fake news analysis. Firstly, these theories introduce new opportunities for qualitative and quantitative studies of big fake news data which, today, has 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[1].

In the following sections the analysis of textual content is explained and reported, supported by several academic publications.

1.1. State-of-art:

Nowadays, there are four important lines of research among the automated classification of genuine and fake articles[1]:

- Knowledge-based, focusing on the false knowledge in fake news:

A knowledge-based perspective aims to analyze and/or detect fake news, using a process known as fact-checking. This process aims to assess news authenticity by comparing the knowledge extracted from to-be-verified news content with known facts (verified).

- Style/Content-based, concerned with how fake news is written and its content:

Style-based is attributable to the proposal of Argamon-Engelson[2] et al. (1998). Formally, fake news style can be defined a *set of quantifiable characteristics (e.g., machine learning features) that can well represent fake news and differentiate fake news from truth.* Studying fake news detection from a Style-based perspective is emphasized on investigating the news content and to assess its intention.

- Propagation-based, focused on how fake news spreads:

A propagation-based perspective studies fake news taking advantage of the information related to the dissemination of fake news, e.g., how it propagates and users spreading it.

- Credibility-based, investigating the credibility of its creators and spreaders:

The credibility-based perspective can overlap with a propagation based study of fake news, where relationships between news articles and components such publishers, authors or posts are considered. In other words, it studies news-related and social related information.

The state-of-art focuses on style/content-based perspective is explored in more details. Since in Fandango's project, this is the approach chosen.

1.1.1. State-of-art Style/Content-based

Deception studies in general do not only consider the inherent characteristics of the style of news articles. A linguistic approach, instead, attempts to identify text properties, such as writing style and content, that is seen as an interesting hint to discriminate real from fake news articles.

The assumption for this approach is that linguistic aspects of a text (punctuation usage, part-of-speech tags, word type choices and emotional valence of a text) are not under the total control of the writer, who often is guided by unconscious mind in ruling writing choices. This reveals important insights into the nature of a text. Moreover it gives to machine learning models the possibility to explore different kind of features, in order to get more information to implement in an algorithm.

While deception analysis and detection have long been an active area of research and have focused on the general style of deceptive (i.e., intentionally false) content across various types of information (e.g., statements, online messages, and reviews, and uniformly regards them as deception (i.e., disinformation)), the development of style-based fake news studies is still in an early stage, with only a limited number of such publication. In addition, most academic approaches use principally LIWC[1] (Linguistic Inquiry and Word Count).

- *LIWC is a text analysis program available for purchase. It calculates the degree to which various categories of words are used in a text, and can process texts ranging from e-mails to speeches, poems and transcribed natural language in either plain text or Word formats.*

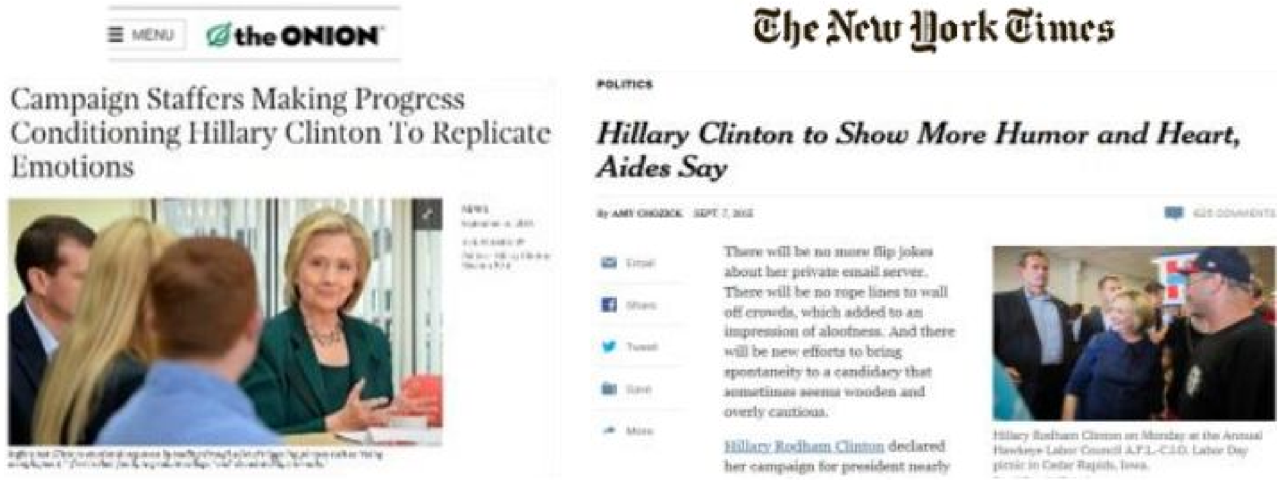
In Volkova et al.2017[3] is presented an analytic study on news language in the context of political fact-checking and fake news detection. The authors compare the language of real news with the satire, hoaxes, and propaganda ones to find linguistic characteristics of untrustworthy text. Experiments show that media fact-checking remains to be an open research question and stylistic clues can help to determine the truthfulness of a text.

A similar case of study has been reported by Victoria Rubin (Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News[4]). This paper provides a conceptual overview of satire and humor, elaborating and illustrating the unique features of satirical news, which mimics the format and style of journalistic reporting.

The authors collected a dataset of 360 news articles as a wide-ranging and diverse data sample, trying to mimic the scope of US and Canadian national newspapers, as much as possible (Figure 1):

[1] Site: http://www.liwc.net/

- Satirical news sites:
  - The Onion
  - The Beaverton
- Legitimate news sources:
  - The Toronto Star
  - The New York Times



Campaign Staffers Making Progress Conditioning Hillary Clinton To Replicate Emotions

Hillary Clinton to Show More Humor and Heart, Aides Say

By AMY CHOZICK SEPT. 7, 2015

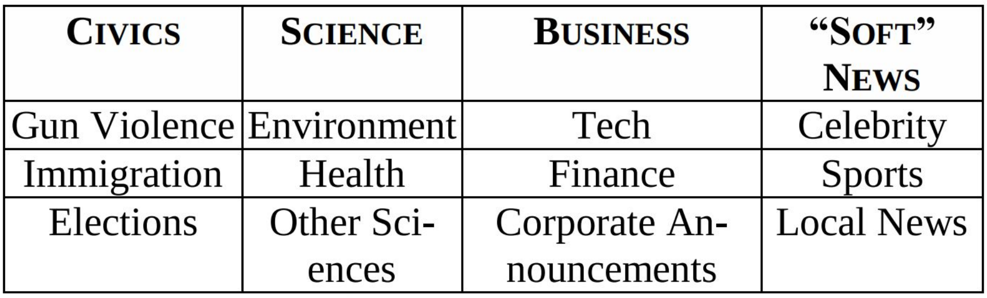
There will be no more flip jokes about her private email server. There will be no rope lines to wall off crowds, which added to an impression of aloofness. And there will be new efforts to bring spontaneity to a candidacy that sometimes seems wooden and overly cautious.

Hillary Rodham Clinton declared her campaign for president nearly

Hillary Rodham Clinton on Monday at the Annual Hawkeye Labor Council A.F.L.-C.I.O. Labor Day picnic in Cedar Rapids, Iowa.

Figure 1: Two news articles about Hillary Clinton: from the Onion and The New York Times.

. The dataset covers 4 domains (Figure 2)



CIVICS	SCIENCE	BUSINESS	"SOFT" NEWS
Gun Violence	Environment	Tech	Celebrity
Immigration	Health	Finance	Sports
Elections	Other Sciences	Corporate Announcements	Local News

Figure 2 - Sample News Topicality. 5 Canadian and 5 American satirical and legitimate article pairs were collected on 12 topics across 4 domains.

The authors proposed an SVM-based algorithm, enriched with five predictive set of features:

1. **Absurdity** - It was defined by the unexpected introduction of new named entities (people, places, locations) within the final sentence of satirical news. (Tools used Natural Language Toolkit6 - NLTK[2] , Part of Speech tagger[3] and Named Entity Recognizer).
2. **Humor** - it was based on the premises of opposing scripts and maximizing semantic distance between two statements as method of punchline identification.

[2] https://www.nltk.org/

[3] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

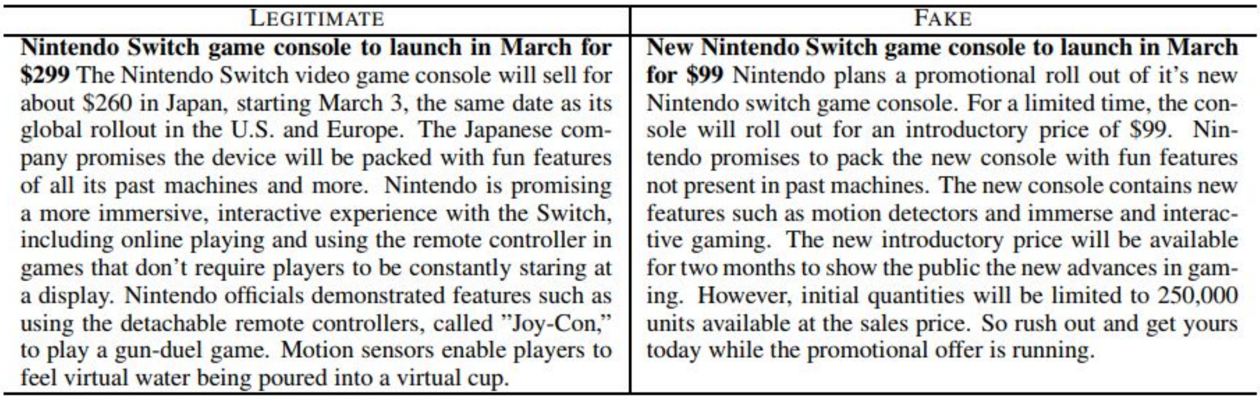
3. **Grammar** - it was the set of normalized term frequencies matched against the Linguistic Inquiry and Word Count (LIWC) 2015 dictionaries.
4. **Negative Affect** and **Punctuation**[4] - they were assigned as feature weights representing normalized frequencies based on term-for-term comparisons with LIWC 2015 dictionaries.

Their best predicting features combination (Absurdity, Grammar and Punctuation) detects satirical news with a 90% precision and 84% recall (F-score=87%).

In contrast to previous works primarily focused on satire detections, Veronica Perez-Rosas et al. (Automatic Detection of Fake News []) focused on fake news detections.

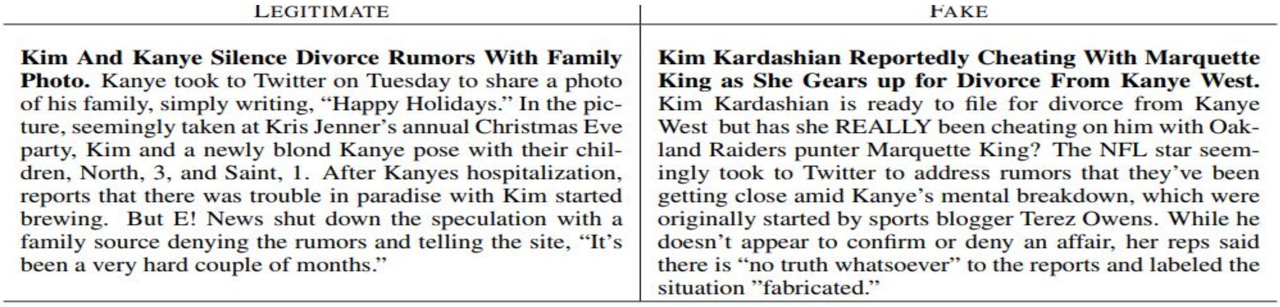
It provides two novel datasets for the task of fake news detection, covering six different news domains.

- Crowdsourced Fake News Dataset (FakeNewsAMT) - (Figure 1) To generate fake versions of the legitimate news items, the authors made use of crowdsourcing via Amazon Mechanical Turk (AMT). This Dataset covers six news domains (sports, business, entertainment, politics, technology, and education)
- Web Dataset Celebrity - (Figure 2) The Dataset is a collect of news of public figures as they are frequently targeted by rumors, hoaxes, and fake reports.



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 3 - Sample legitimate and crowdsourced fake news



LEGITIMATE	FAKE
Kim And Kanye Silence Divorce Rumors With Family Photo. Kanye took to Twitter on Tuesday to share a photo of his family, simply writing, "Happy Holidays." In the picture, seemingly taken at Kris Jenner's annual Christmas Eve party, Kim and a newly blond Kanye pose with their children, North, 3, and Saint, 1. After Kanyes hospitalization, reports that there was trouble in paradise with Kim started brewing. But E! News shut down the speculation with a family source denying the rumors and telling the site, "It's been a very hard couple of months."	**Kim Kardashian Reportedly Cheating With Marquette King as She Gears up for Divorce From Kanye West.** Kim Kardashian is ready to file for divorce from Kanye West but has she REALLY been cheating on him with Oakland Raiders punter Marquette King? The NFL star seemingly took to Twitter to address rumors that they've been getting close amid Kanye's mental breakdown, which were originally started by sports blogger Terez Owens. While he doesn't appear to confirm or deny an affair, her reps said there is "no truth whatsoever" to the reports and labeled the situation ”fabricated."

Figure 4 - Sample legitimate and web fake news in the Celebrity domain

[4] Best predicting feature combination

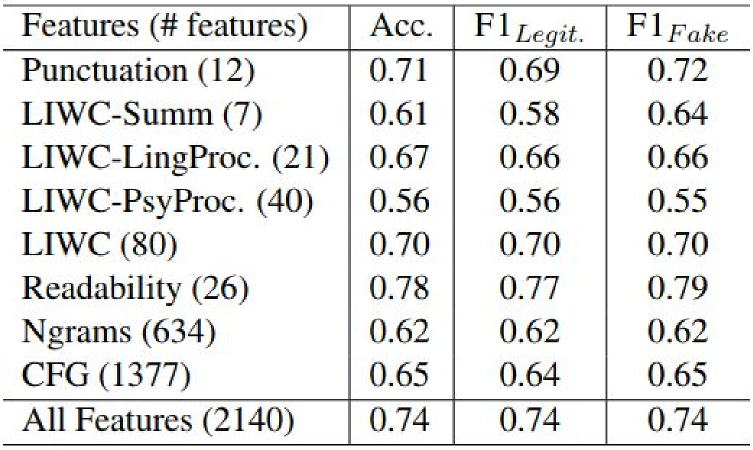
Moreover, the authors describe the collection, annotation, and validation process in detail and show several exploratory analysis on the identification of linguistic differences in fake and legitimate news content, using an ML model.

To build the fake news detection models, the authors starts by extracting several sets of linguistic features:

- **Ngrams** - unigrams and bigrams
- **Punctuation** - This includes punctuation characters such as periods, commas, dashes, question marks and exclamation marks
- **Psycholinguistic features**. They use the LIWC lexicon to extract the proportions of words that fall into psycholinguistic categories.
- **Readability -** it indicate text understandability
- **Syntax -** set of features derived from production rules based on context free grammars (CFG) trees using the Stanford Parser [5]

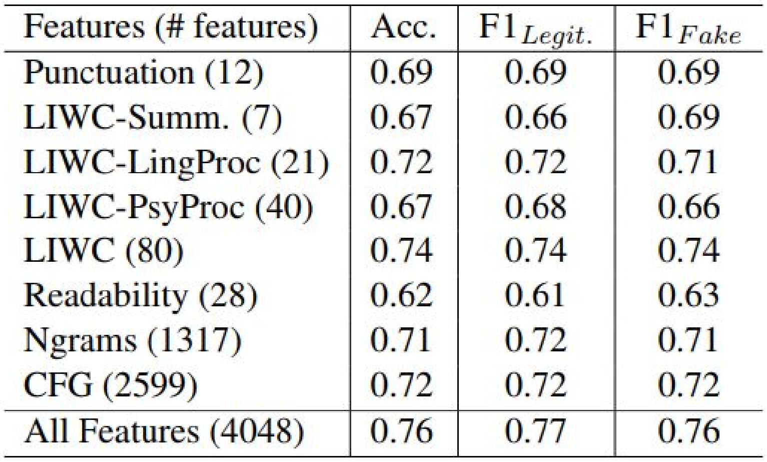
The experiments have been done with an SVM classifier and using five-fold cross validation, with accuracy, precision, recall, and F-score as performance metrics.

The performers reported in figure 5 and 6 suggest that an integrated use of linguistic, syntactic and semantic features is useful to discriminate between real and fake news content (accuracy 0.74 and 0.76).



Features (# features)	Acc.	$F1_{Legit.}$	$F1_{Fake}$
Punctuation (12)	0.71	0.69	0.72
LIWC-Summ (7)	0.61	0.58	0.64
LIWC-LingProc. (21)	0.67	0.66	0.66
LIWC-PsyProc. (40)	0.56	0.56	0.55
LIWC (80)	0.70	0.70	0.70
Readability (26)	0.78	0.77	0.79
Ngrams (634)	0.62	0.62	0.62
CFG (1377)	0.65	0.64	0.65
All Features (2140)	0.74	0.74	0.74

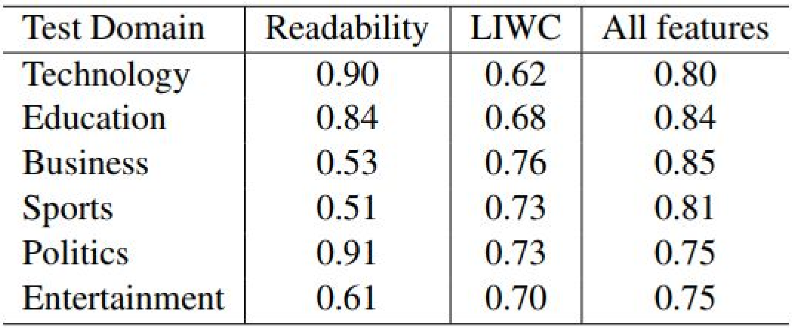
Figure 5 - Classification results for the FakeNewsAMT dataset collected via crowdsourcing.



Features (# features)	Acc.	$F1_{Legit.}$	$F1_{Fake}$
Punctuation (12)	0.69	0.69	0.69
LIWC-Summ. (7)	0.67	0.66	0.69
LIWC-LingProc (21)	0.72	0.72	0.71
LIWC-PsyProc (40)	0.67	0.68	0.66
LIWC (80)	0.74	0.74	0.74
Readability (28)	0.62	0.61	0.63
Ngrams (1317)	0.71	0.72	0.71
CFG (2599)	0.72	0.72	0.72
All Features (4048)	0.76	0.77	0.76

Figure 6 - Classification results for the Celebrity news dataset

As shown, the Readability has a drop from 0.78 to 0.62 inversely similar to LIWC-PsyProc. These find a hint respecting the important role of the domains in the fake news detection. For this reason, in a further experiment, the authors assess the cross-domain classification performance for the six news domains in the FakeNewsAMT dataset. Therefore, they trained a model using only five of the six domains in the dataset, and tested it on the remaining one. The results for the Cross Domain are shown in Figure 7.



Test Domain	Readability	LIWC	All features
Technology	0.90	0.62	0.80
Education	0.84	0.68	0.84
Business	0.53	0.76	0.85
Sports	0.51	0.73	0.81
Politics	0.91	0.73	0.75
Entertainment	0.61	0.70	0.75

Figure 7 - Cross-domain classification accuracy for the complete LIWC and readability feature sets. Training data consists of all but the test domains in the FakeNewsAMT dataset.

The results shows that fake and legitimate news in politics, education, and technology domains might be structurally similar to the fake and legitimate content in the other five domains.

The authors report 2 main conclusions:

- Computational linguistics can support the fake news identification process with an automated manner. The proposed linguistics-driven approach suggests that to differentiate between fake and genuine content it is worthwhile to look at the lexical, syntactic and semantic level of a news item in question (accuracy up to 76%). Therefore the linguistics features seem promising but should not be limited to these and should also include meta features (e.g. comments on the article), features from different modalities (e.g. the visual makeup of a website) etc.
- It is possible to build resources for the fake news detection task by combining manual and crowdsourced annotation approaches. Infact one dataset is collected via crowdsourcing

(FakeNewsAMT). The second dataset (Celebrity) is obtained directly from the web and covers celebrity news.

Another important contribution on style-base approach is reported in “A Stylometric Inquiry into Hyperpartisan and Fake News; Potthast et al.2017 [6]. The authors focused on demonstrating a connection between writing style and informative content of a news article. In particular they focused on two series of experiments that investigate style differences and similarities between:

- (left/right) hyperpartisan and mainstream news
- fake, legitimate, and satire news.

In order to explore the weight of style in the text, the used features are:

- non domain-specific features:
  a. n-grams, n in[1,3], of:
     i. characters
     ii. stop words
     iii. parts-of-speech.
  b. 10 readability scores[5]
  c. dictionary features ,each indicating the frequency of words from a tailor-made dictionary in a document.

- domain-specific features
  a. ratios of quoted words
  b. ration of external links
  c. number of paragraphs
  d. paragraphs average length.

The experiments have been done with a WEKA’s[6] random forest implementation with default settings and using 3-fold cross validation, with accuracy, precision, recall, and F-score as performance metrics.

The authors report 2 main results:

1. demonstrating that the writing style of the hyperpartisan left and right are more similar compared with a mainstream source*(F1=0.78)*.

2. studying how the writing style can be used to discriminate real from fake news. They conclude that stylometry is not the silver bullet as a style-based fake news detection *(F1=0.46).*

Moreover, in this paper is reported a Taxonomy of paradigms for fake news detection (Figure 2). This taxonomy is to find new categories. In particular, the authors split the style/context-based approach in two different categories:

[5] Automated Readability Index, Coleman Liau Index, Flesh Kincaid Grade Level and Reading Ease, Gunning Fog Index, LIX,McAlpine EFLAW Score, RIX, SMOG Grade, Strain Index

[6] Weka 3: Machine Learning Software in Java

- Context-based - Fake news items are identified via meta information and spread patterns, like:
    - Long et al. (2017)[7] shows that the writer's information can be a useful feature for fake news detection
    - Derczynski et al.(2017)[8] attempts to determine the veracity of a claim, based on the conversation collected on Twitter as one of the RumourEval tasks.

- Style-based - Divided into two sub-categories:
    - Deception detection - it originates from forensic linguistics and built on the Undeutsch[9] hypothesis. It focuses on single statements
        - Rubin et al. (2015) uses rhetorical structure theory as a measure of story coherence and as an indicator for fake news.
        - Pisarevskaya et al. [11] applied a framework called RST (see before)
    - Text categorization - It characterize a given document in terms of some very large class of stylistic features.
        - Rashkin et al. (2017) [12] performs statistical analysis of the stylistic differences between real, satire, hoax, and propaganda news.
        - Horne and Adali (2017) [13] use style features for fake news detection. Their final classifier uses only 4 features (number of nouns, type-token ratio, word count, number of quotes), which can be easily manipulated;

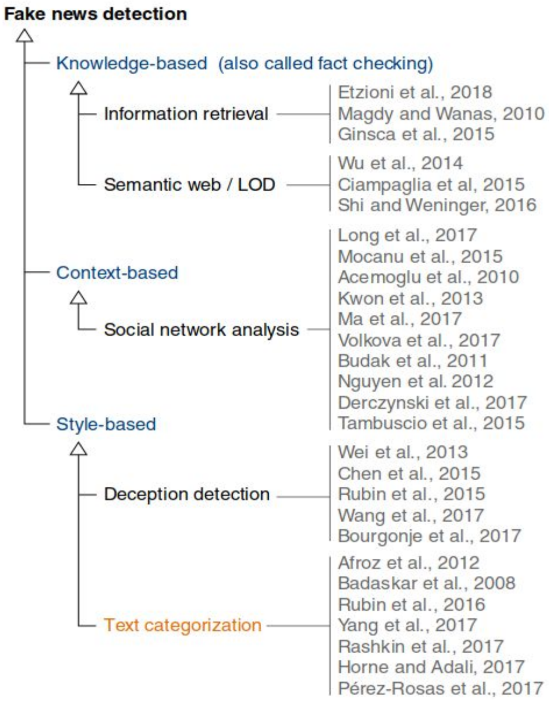


Figure 8 - Taxonomy of paradigms for fake news detection alongside a selection of related work.

Other important works in style-based perspective are: Bond et al.2017[14] Pisarevskaya's [11] and Hamid Karimi.

In Bond et al.2017[14] is analyzed the language of Donald Trump and Hillary Clinton from the debates as well as the tweets of millions of people during the presidential campaign. The authors used a reality monitoring[7] (RM) framework with LIWC to code candidates' language. The focus of the research is to determine, using a computerized algorithm grounded in memory theory, if veracity of candidates' public statements in the 2016 US presidential race can be successfully differentiated. The results showed that the model has classified 87% of fact-checked truth statements but only 28% of fact-checked lie statements (63% classification overall).

In Pisarevskaya's[11]publication there is a clear comparison between manual and automatic accuracy of fake news detection. Actually, the reported precision of human deception and detection ability in Russian language is 0.55. In order to carry their research, the authors collected a corpus of 134 news reports divided into truthful and fake news. They applied a framework called RST(Rhetorical Structure Theory). This framework (Mann and Thompson, 1988[15]) is addressed to the discourse level of text, where a text is represented as a hierarchical tree. Some parts are more essential (nucleus) than others (satellite). Afterwards, they segmented the texts manually and applied RST relations tagging to them. As to the experiments, three dataset models for machine learning were based on features from the Rhetorical Structure Theory and used the model based on features from the sentiment lexicon as a baseline. Finally, they chose Support vector machines (SVMs) and Random Forest to classify the news reports into 2 classes: truthful/deceptive. The predictive power of the model is the highest one, based simply on frequencies Rhetorical Structure Theory as a Feature for Deception Detection in News Reports of RST relation types in texts. The classification task was solved better by SVMs (linear kernel) for the chosen dataset (0.65 accuracy score).

A similar result connected to RST framework has been recently obtained in Learning Hierarchical Discourse-level Structure for Fake News Detection[16]. It is based on the incorporation of hierarchical discourse-level structure of fake and legitimate. This ML model learns and builds a discourse-level structure for legitimate/fake news in an automated and data-driven way. This method outperforms baselines like N-grams, LIWC BiG RNN-CNN, LSTM of more than 1.5% of accuracy, reaching exactly 82.19.

The most recent work[17], dated 17th July 2019, deals with fake news detection as natural language inference (NLI). Research declared to this method to be generally effective for the aforementioned task, in particular their method ensembles and combines NLI models train with a fine-tuned BERT model and a decomposable attention model with predictions made by observing transitivity relations.

[7] The process people use in deciding whether information initially had an external or an internal source[]

1.1.2. Fandango's approach

Fandango's approach starts from intuitions and experiments carried out in the publication "A Stylometric Inquiry into Hyperpartisan and Fake News[18]" and Veronica Perez-Rosas el al[25].

To differ from previous work, Fandango's approach features are only context agnostic. Therefore, all features based on text as n-gram, frequency of word, tf/idf LIWC and etc. have been removed. In other words a "context agnostic" strategy implies that each word of the text is not considered as a predictor variable.

This approach can be catalogued as exclusively style-based Fake News Analysis, addressing how contents and writing style of fake news can be different from the same one of true news.

The Fandango approach aims to overcome the main problems of context-based approaches, that are:

- Domain specificity
- Language dependent

In particular, to support language dependency the main research direction is to identify functions to correlate and match different languages between them. A specific example is the publication "Word translation without parallel data[19]". In this paper is described how a bilingual dictionary between two languages can be built without using any parallel corpora. This can be done by aligning monolingual word embedding spaces in an unsupervised way, without using any character information. Moreover, this model outperforms the existing supervised methods on cross-lingual tasks for some language pairs. Briefly description of this approach can be done step by step, as shown in Figure 4.

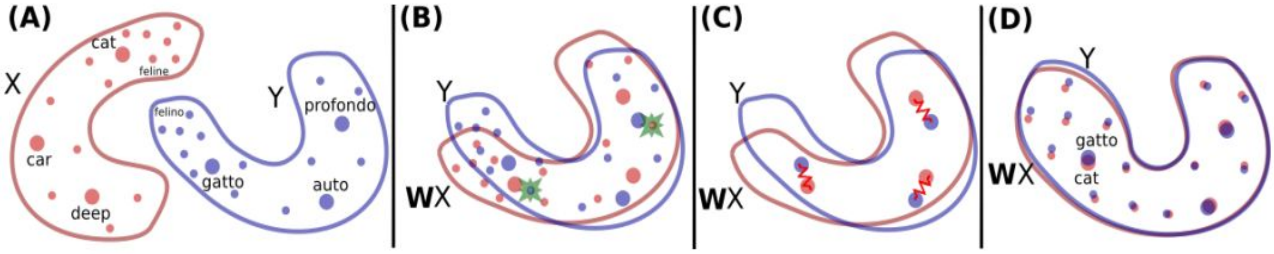


Figure 9 - word translation without parallel data description stey by step

(A) Let's consider two distributions of word embeddings, English words in red represented by X and Italian words in blue denoted by Y , to translate. Each dot on the image represents a word in that space.

(B) The frequency of the words is proportional to their sizes in the training corpus of that language. It's possible to learn a rotation matrix W which roughly aligns the two distributions.

(C) On the image, the green stars are randomly selected words that are fed to the discriminator to determine whether the two word embeddings come from the same distribution. This method uses frequent words aligned by the previous step as anchor points, and minimizes an energy function that corresponds to a spring system between anchor points. The refined mapping is then used to map all words in the dictionary.

(D) Finally, the translation process continues by using the mapping W and a distance metric, that expands the space where there is a high density of points (like the area around the word “cat”), so that “hubs” (like the word “cat”) become less close to other word vectors than they would otherwise.

This paper is the guidelines to support the experimentations of a cross-language models. In fact, one of future research is to use a single model to analyze articles in different languages. This can be availed the intuition from which, “agnostic” features behaviour of a language are similar to those of another language.

Fandango’s Tools and Output

The entire procedure for creating the machine learning classifiers, able to detect fake news, had to be repeated using sources and articles in italian, dutch and spanish. In some languages the features of the model cannot be implemented at all, due to their structure or complexity, to be transformed as functions.

Machine learning classifiers have been created in Python programming language, using libraries as DS4Biz Predictor, Scikit-learn, Spacy, nltk, polyglot, TreeTagger, tensorflow, keras and PySpark. It provides a score that corresponds to the probability of an article to be classified as reliable or unreliable.

The ML classifiers is interrogable via RESTful api (deliverable 3.1).

An example of the output provided by the analyzer is represented in JSON format with the following keys below:

a) **Identifier**: the unique identifier of the article. This value is provided during the preprocessing step, where the “texthash” is used for this purpose.

b) **TextRating**: this field is one of the scores calculated, gives a news prediction analysing article body and title.

2. Ground truth building

2.1. Description

Ground truth is a term used in various fields to basically refer to any kind of information provided by direct observation. In machine learning it is used to define a dataset with which a model is trained or to compare the model’s performance.

The different methodologies to build a ground truth are exposed In the state-of-art section . In particular, in Victoria Rubin et al.2016[20] they considered four different domains, two from satirical articles sources and two from legitimate news sources.

Another approach is reported in Veronica Perez-Rosas et al. in which the authors used the crowdsourcing via Amazon Mechanical Turk (AMT) to build a ground truth.

2.2. Strategy chosen to build FANDANGO's ground truth

Ground truth contains all the samples to train a model and all the articles have to belong to a category(legitimate/fake). In order to approach this task, three different methods are chosen to annotate documents and create the so-called training dataset.

In many of the academic material about fake news, there have been established some guidelines[6], in order to create a ground truth as informative and correct as possible.

Summarizing the points:

A. Include fake and legitimate
B. Contain text-only news items
C. Be homogeneous in length and writing style
D. Contains news from a predefined time frame
E. Be delivered in the same manner and for the same purpose
F. Be made publicly available

In order to satisfy this points, Fandango's GT is built using 3 different strategies.

1. Online Datasets Collection - This approach satisfies mainly the points A, B, C, F
2. Semi-Automatic annotation system - This approach satisfies mainly the points A, B, D
3. Manual annotation system - This approach satisfies mainly the points A, B, D , E

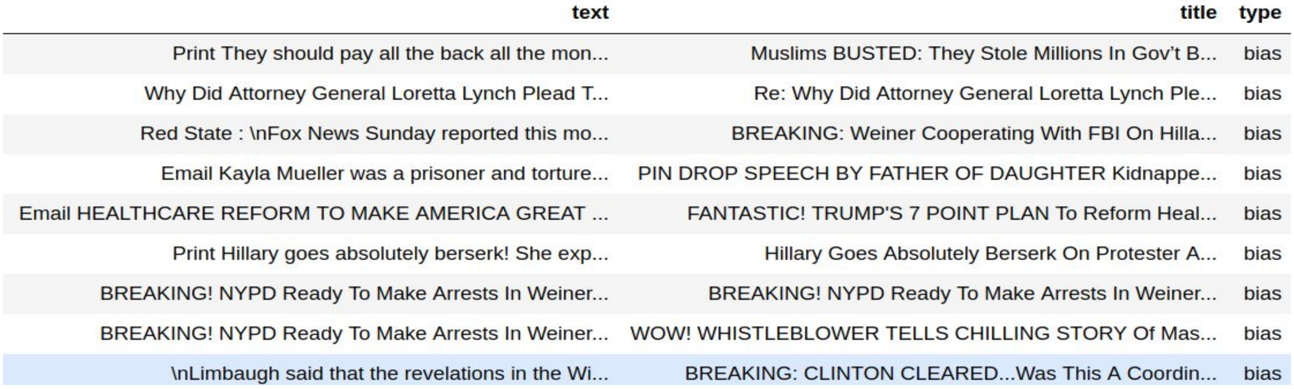
Online datasets collection:

Firstly, using existing dataset applied in academic research, allows 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, including online articles whose veracities have been identified by experts. The other two datasets come from kaggle. All these 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 approach proposed by researchers.

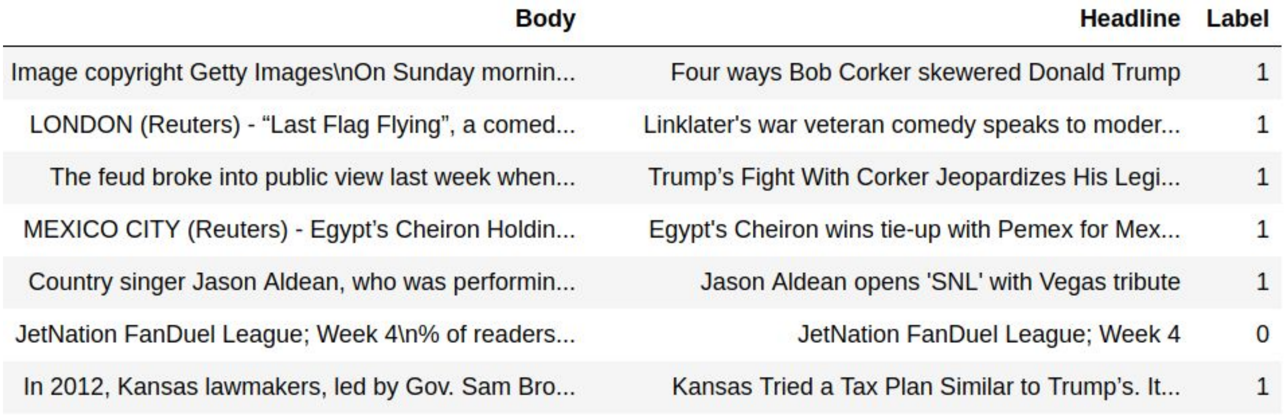
Here the link of the datasets taken in this approach available only in english:

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



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 10 - Fake news sample from kaggle



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 11 - kaggle dataset with legitimate and fake news

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 fake. This list is composed by 10 source domains for each language equally divided into the two categories requested, showed in tables from 1 to 4

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.comlavocedelpatriota
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.tijd.be	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

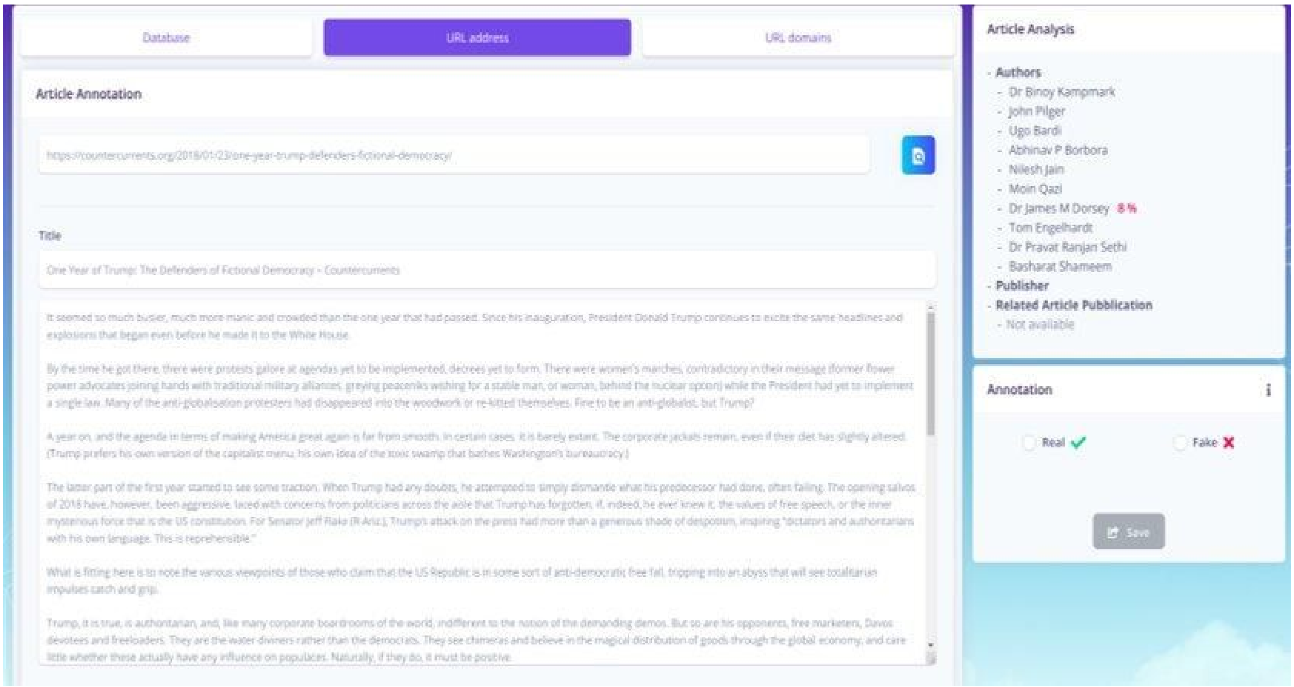
Table 4 - source domains in english

MANUAL ANNOTATION SYSTEM:

The goal is to annotate manually 1000 articles per language. The annotators are professional journalists (end-users).

A UI has been implemented to support the annotators in this task. It provided two ways to add a label to the articles:

1. A user can annotate an article from Fandango’s data silos.
2. A user can annotate an external article adding its URL.



Article Annotation

https://countercurrents.org/2018/01/23/one-year-trump-defenders-fictional-democracy/

Title

One Year of Trump: The Defenders of Fictional Democracy - Countercurrents

It seemed so much busier, much more manic and crowded than the one year that had passed. Since his inauguration, President Donald Trump continues to excite the same headlines and explosions that began even before he made it to the White House.

By the time he got there, there were protests galore at agendas yet to be implemented, decrees yet to form. There were women's marches, contradictory in their message (former flower power advocates joining hands with traditional military alliances, greying peaceniks wishing for a stable man, or woman, behind the nuclear option) while the President had yet to implement a single law. Many of the anti-globalisation protesters had disappeared into the woodwork or re-kitted themselves. Fine to be an anti-globalist, but Trump?

A year on, and the agenda in terms of making America great again is far from smooth. In certain cases, it is barely extant. The corporate jackals remain, even if their diet has slightly altered. (Trump prefers his own version of the capitalist menu, his own idea of the toxic swamp that bathes Washington's bureaucracy.)

The latter part of the first year started to see some traction. When Trump had any doubts, he attempted to simply dismantle what his predecessor had done, often failing. The opening salvos of 2018 have, however, been aggressive, laced with concerns from politicians across the aisle that Trump has forgotten, if, indeed, he ever knew it, the values of free speech, or the inner mysterious force that is the US constitution. For Senator Jeff Flake (R-Ariz.), Trump's attack on the press had more than a generous shade of despotism, inspiring "dictators and authoritarians with his own language. This is reprehensible."

What is fitting here is to note the various viewpoints of those who claim that the US Republic is in some sort of anti-democratic free fall, tripping into an abyss that will see totalitarian impulses catch and grip.

Trump, it is true, is authoritarian, and, like many corporate boardrooms of the world, indifferent to the notion of the demanding demos. But so are his opponents, free marketers, Davos devotees and freeloaders. They are the water diviners rather than the democrats. They see chimeras and believe in the magical distribution of goods through the global economy, and care little whether these actually have any influence on populaces. Naturally, if they do, it must be positive.

Article Analysis

- Authors
  - Dr Binoy Kampmark
  - John Pilger
  - Ugo Bardi
  - Abhinav P Borbora
  - Nilesh Jain
  - Moin Qazi
  - Dr James M Dorsey 8 %
  - Tom Engelhardt
  - Dr Pravat Ranjan Sethi
  - Basharat Shameem
- Publisher
- Related Article Pubblication
  - Not available

Annotation

○ Real ✓ ○ Fake ✗

Figure 12 - Annotation system UI

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 (Table 5).

Language	Domains
IT	http://www.ilgiornale.it
IT	https://www.liberoquotidiano.it
IT	https://www.ilfattoquotidiano.it
ES	https://www.larazon.es
ES	https://www.lavanguardia.com
ES	https://www.abc.es
ES	https://www.eldiario.es
ES	https://www.elespanol.com
EN	https://www.thesun.co.uk
EN	https://www.dailymail.co.uk
EN	http://www.breitbart.com

EN	http://cen.at
EN	https://www.defendevropa.com/2017/news/electoral-commission-arron-banks
EN	https://voiceofeurope.com
EN	https://gellerreport.com
EN	https://www.zerohedge.com
EN	http://v4report.com
NL	https://hln.be
NL	https://www.nieuwsblad.be
NL	http://newsmonkey.be
NL	https://www.gva.be
NL	https://www.hbvl.be
NL	https://www.nd.nl
NL	https://www.rtlnieuws.nl
NL	https://www.nu.nl
NL	http://www.dewereldmorgen.be
NL	https://nl.metrotime.be/news
NL	https://www.mo.be
NL	https://nieuws.vtm.be
NL	https://nl.express.live
NL	https://www.apache.be
NL	https://www.ad.nl/nieuws
NL	https://doorbraak.be
NL	https://www.krapuul.nl
ES	http://okdiario.com
ES	https://www.huffingtonpost.es
ES	http://publico.es
ES	http://libertaddigital.com
ES	http://periodistadigital.com
ES	http://vozpopuli.com
ES	http://elindependiente.com
ES	http://elplural.com
ES	http://lamarea.com
ES	https://www.elsaltodiario.com
ES	http://europapress.es

ES	http://efe.com
ES	https://www.cope.es
ES	http://lavozdegalicia.es
ES	http://elperiodico.com
ES	http://elcorreo.com
ES	https://www.lne.es
ES	http://farodevigo.es
ES	http://heraldo.es
ES	http://levante-emv.com
ES	http://diariovasco.com
ES	http://diariodenavarra.es
ES	https://www.eldiariomontanes.es
ES	http://elnortedecastilla.es
ES	http://laverdad.es

Table 5 - source domains

2.3. Filter of the news

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 [21]and word-embedding[22].

This filter is built, retrieving main meaning for each news, converting an article in a list of vectors, where each word is represented by its correlated words, and then calculating a centroid to generate a principal vector to measure how close the result is to the vectors(filter) representing topics. The detailed description what is described before is divided into required actions and some steps as follows:

Required Action:

- 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 $V = (v_1, v_2, v_3, v_4, ..., v_{300})$

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

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

   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 6 - 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 7 - 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 8 - 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

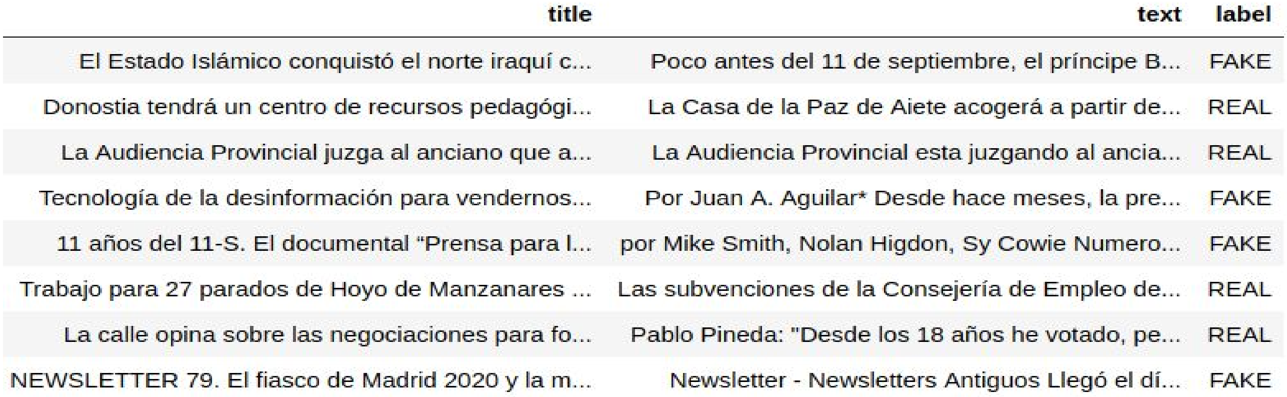
EN	Eurosceptics	EU
EN	High Representative of the Union for Foreign Affairs and Security Policy	EU
EN	President of the European Council	EU
EN	TFEU	EU
EN	Treaty on the Functioning of the European Union	EU
EN	TEU	EU
EN	Treaty on European Union	EU
EN	Treaty of Rome	EU
EN	European Central Bank (ECB)	EU
EN	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 Suprematists	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 9 - words to create a filter of topics in english

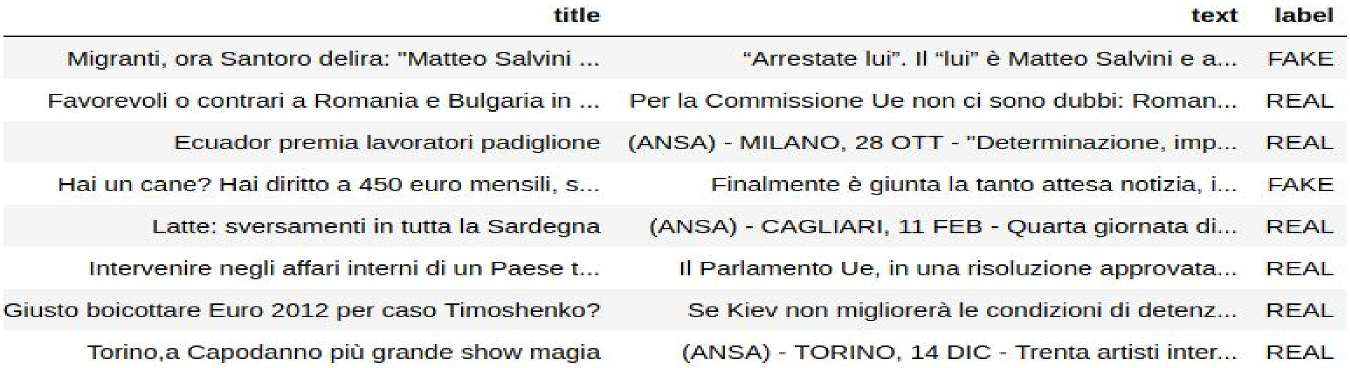
2.4 Example of ground truth in different language

In this section are shown portions di Ground truth for each language (Figure 13 - 15)



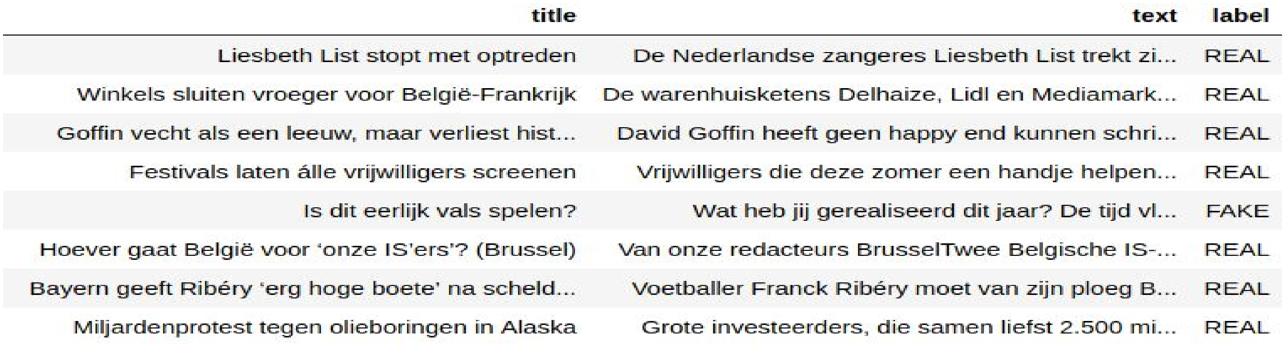
title	text	label
El Estado Islámico conquistó el norte iraquí c...	Poco antes del 11 de septiembre, el príncipe B...	FAKE
Donostia tendrá un centro de recursos pedagógi...	La Casa de la Paz de Aiete acogerá a partir de...	REAL
La Audiencia Provincial juzga al anciano que a...	La Audiencia Provincial esta juzgando al ancia...	REAL
Tecnología de la desinformación para vendernos...	Por Juan A. Aguilar* Desde hace meses, la pre...	FAKE
11 años del 11-S. El documental “Prensa para l...	por Mike Smith, Nolan Higdon, Sy Cowie Numero...	FAKE
Trabajo para 27 parados de Hoyo de Manzanares ...	Las subvenciones de la Consejería de Empleo de...	REAL
La calle opina sobre las negociaciones para fo...	Pablo Pineda: "Desde los 18 años he votado, pe...	REAL
NEWSLETTER 79. El fiasco de Madrid 2020 y la m...	Newsletter - Newsletters Antiguos Llegó el dí...	FAKE

Figure 13 - Ground truth in spanish



title	text	label
Migranti, ora Santoro delira: "Matteo Salvini ...	“Arrestate lui”. Il “lui” è Matteo Salvini e a...	FAKE
Favorevoli o contrari a Romania e Bulgaria in ...	Per la Commissione Ue non ci sono dubbi: Roman...	REAL
Ecuador premia lavoratori padiglione	(ANSA) - MILANO, 28 OTT - "Determinazione, imp...	REAL
Hai un cane? Hai diritto a 450 euro mensili, s...	Finalmente è giunta la tanto attesa notizia, i...	FAKE
Latte: sversamenti in tutta la Sardegna	(ANSA) - CAGLIARI, 11 FEB - Quarta giornata di...	REAL
Intervenire negli affari interni di un Paese t...	Il Parlamento Ue, in una risoluzione approvata...	REAL
Giusto boicottare Euro 2012 per caso Timoshenko?	Se Kiev non migliorerà le condizioni di detenz...	REAL
Torino,a Capodanno più grande show magia	(ANSA) - TORINO, 14 DIC - Trenta artisti inter...	REAL

Figure 14 - Ground truth in italian



title	text	label
Liesbeth List stopt met optreden	De Nederlandse zangeres Liesbeth List trekt zi...	REAL
Winkels sluiten vroeger voor België-Frankrijk	De warenhuisketens Delhaize, Lidl en Mediamark...	REAL
Goffin vecht als een leeuw, maar verliest hist...	David Goffin heeft geen happy end kunnen schri...	REAL
Festivals laten álle vrijwilligers screenen	Vrijwilligers die deze zomer een handje helpen...	REAL
Is dit eerlijk vals spelen?	Wat heb jij gerealiseerd dit jaar? De tijd vl...	FAKE
Hoever gaat België voor 'onze IS'ers'? (Brussel)	Van onze redacteurs BrusselTwee Belgische IS-...	REAL
Bayern geeft Ribéry 'erg hoge boete' na scheld...	Voetballer Franck Ribéry moet van zijn ploeg B...	REAL
Miljardenprotest tegen olieboringen in Alaska	Grote investeerders, die samen liefst 2.500 mi...	REAL

Figure 15 - Ground truth in dutch

3. Preprocessing

Data preprocessing is a set of operation to apply to a sample set in order to clean and take more useful information from it as well as adequate format to the collected information. In an initial phase, data crawled from sources domains aforementioned were in a raw status, so it has been applied some functions to avoid repetitions or record with important features missing, for example : all articles whose text or title fields are empty have been discarded because they are not articles indeed but other sorts of sources that the crawlers collected like cookies confirmation or advertisements. Publisher and language fields has to be checked since they report mistakes most of the time.

3.1. Preprocessing applied to text

Normally, there are many operations to apply to text, in order to clean or get it homogeneous. They are needed for making text from human language to machine-readable format for further processing.

In particular, different tasks go for normalize data for example: converting all letters to lower/upper case, removing punctuations and stop-words (very common words that have little meaning, such as 'the', 'and', etc.), accent marks and other diacritics, removing white spaces (tokenization), expanding abbreviations and more.

Since the aim of fandango's text classifier is to be 'content agnostic', all the operations described before are not going to be applied to the text. This decision comes from the syntax, the structures and the part of speech of an article and their distributions.

3.1.1. Example text raw and preprocessed text

RAW RECORD:

identifier: 8fd0d91d9a2d47f97ea1c97174f0baffa834840c286325590cc7fabec59f2a09

title: Formula One

text: Motor racing: Hamilton would rather fight Ferrari than battle with Bottas Five times world champion Lewis Hamilton says Mercedes should not be blamed for their dominance and he would far rather be in a battle with Formula One rivals /n Ferrari and Red Bull than team mate Valtteri Bottas /n/n.

source_domain: https://www.reuters.com/news/sports

PREPROCESSED RECORD:

headlinE: Formula One

articleBody: Motor racing: Hamilton would rather fight Ferrari than battle with Bottas Five times world champion Lewis Hamilton says Mercedes should not be blamed for their dominance and he would far rather be in a battle with Formula One rivals Ferrari and Red Bull than team mate Valtteri Bottas.

sourceDomain: www.reuters.com

language: en

4. Features engineering

Feature engineering, in this case, can be considered as a set of decisions to be taken during an article's analysis. These features are used to identify if a news is fake or legitimate in the best and most useful way as possible.

The features engineering can be composed in the following tasks: feature transformation, feature generation and extraction, feature selection. The quality of the results of a predictive machine learning model largely depends on the quality of the available features. Therefore, it's one of the most important procedures in the classifier creation pipeline.

4.1. Description

In order to classify an article as reliable or not, some specific features that can be used for machine learning model training. Those features are taken from a text and a body of an article and can be divided into the following macro-categories:

1) Simple frequency features: with these features, apparently simple, it would like to catch some articles specifics in terms of structure(normalized with respect to the length)

    a) stopword counter,
    b) characters counter,
    c) punctuation counter

2) Part of The Speech(POS) features: part of speech distribution can be a good indicator to understand the fake nature of an article since it's used in stylographic studies, based on the observation of how authors tend to write in relatively consistent, recognizable and unique ways.

    a) count adjective
    b) count adverbs
    c) count conjunctions
    d) count verbs
    e) count personal nouns

3) Advance frequency features:

    a) word average per paragraph
    b) lexical density
    c) mean segment type-token ratio (MSTTR)
    d) moving average type-token ratio (MATTR)

4) Readability indices[23]: readability describes the ease with which a document can be read. There exists plenty of different tests to calculate this index and all of them are considered to be predictions of reading ease.

    a) Flesch Index of readability
    b) FKG Read Level
    c) Automated Readability Index(ARI)

Features to implement in future:

There are plenty of features that can be added to our model. They have to be tested in order to understand if they bring additional and useful information to the machine learning classifier.

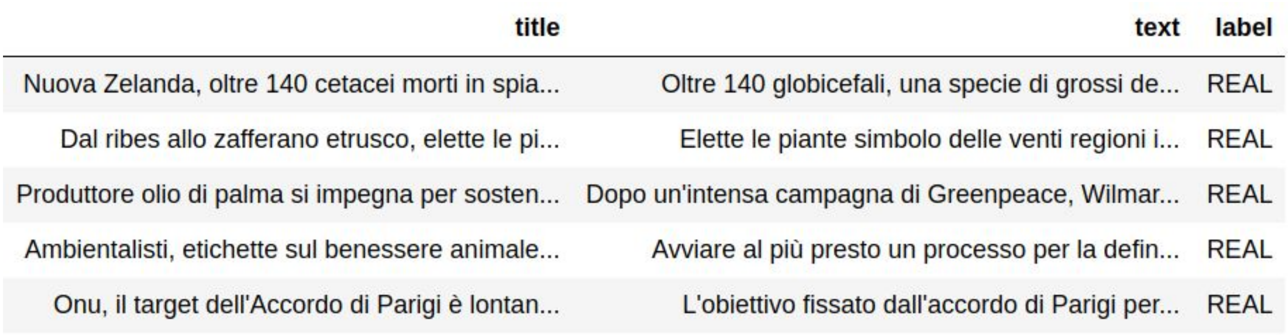
Different readability indices have to be implemented for western european languages, because of the different languages structure and syllable counts.

- Rate Index(RIX): it can be used on documents of any Western European language and its output is a score between 0(very easy) and 55+ (very difficult). It’s calculated as long word over the sentences in the document, where long words are those ones with more than 6 characters.
- Lesbarhets Index: its output is an index whose indicates a grade level. An index below 0.1 is grade 1 while 7.2 and above is college grade.

  The calculation is made with the following formula:

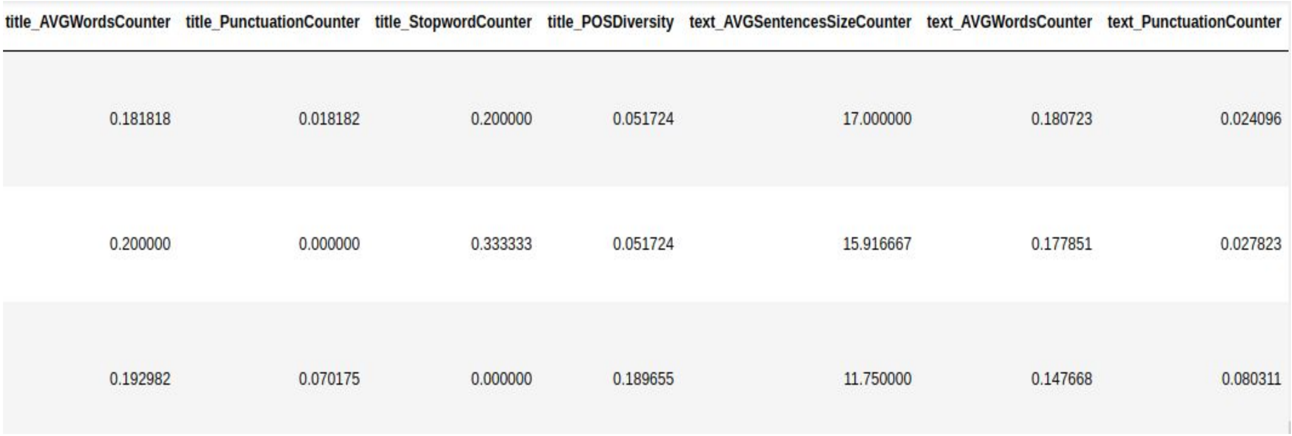
$$LIX = \left(\frac{TOTAL\ WORDS}{TOTAL\ SENTENCES} \right) + \left(\frac{LONG\ WORDS}{TOTAL\ WORDS} * 100 \right)$$

As it has been mentioned above, the machine learning text classifier has been trained considering just features that doesn’t concern with the semantic part of the article. An example of the transformation applied to the training set is shown below (Figure 16 and 17).



title	text	label
Nuova Zelanda, oltre 140 cetacei morti in spia...	Oltre 140 globicefali, una specie di grossi de...	REAL
Dal ribes allo zafferano etrusco, elette le pi...	Elette le piante simbolo delle venti regioni i...	REAL
Produttore olio di palma si impegna per sosten...	Dopo un'intensa campagna di Greenpeace, Wilmar...	REAL
Ambientalisti, etichette sul benessere animale...	Avviare al più presto un processo per la defin...	REAL
Onu, il target dell'Accordo di Parigi è lontan...	L'obiettivo fissato dall'accordo di Parigi per...	REAL

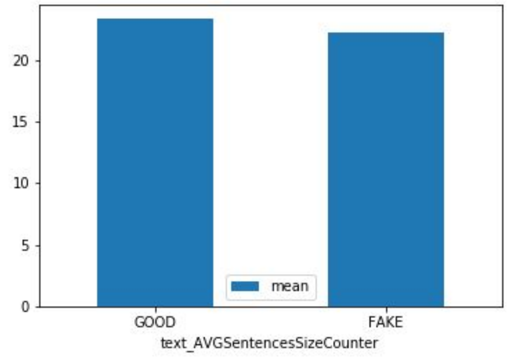
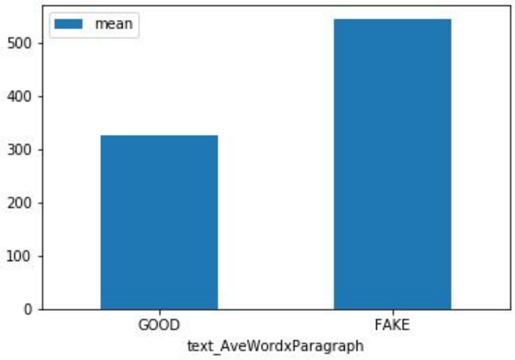
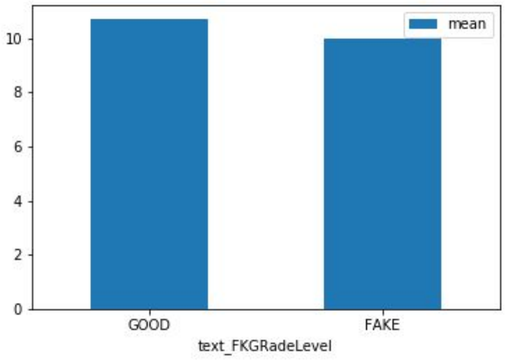
Figure 16 - training set before feature engineering

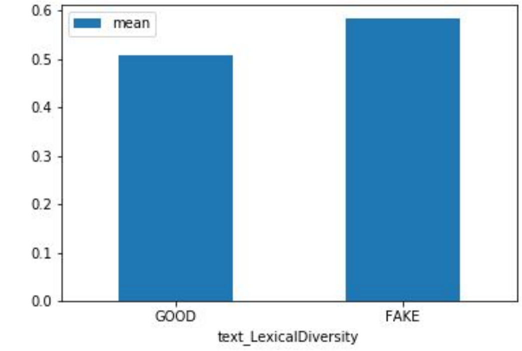
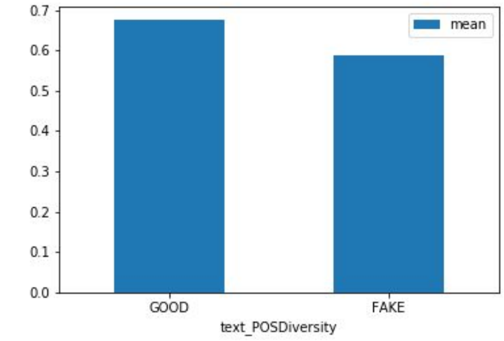
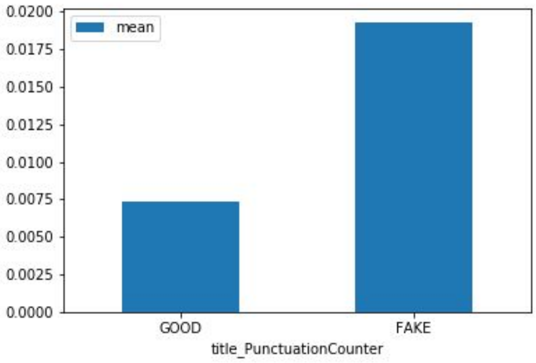


title_AVGWordsCounter	title_PunctuationCounter	title_StopwordCounter	title_POSDiversity	text_AVGSentencesSizeCounter	text_AVGWordsCounter	text_PunctuationCounter
0.181818	0.018182	0.200000	0.051724	17.000000	0.180723	0.024096
0.200000	0.000000	0.333333	0.051724	15.916667	0.177851	0.027823
0.192982	0.070175	0.000000	0.189655	11.750000	0.147668	0.080311

Figure 17 - training set after feature engineering

4.2.1 Example of features chosen for the machine learning classifier

Comparison between a good structure and fake structure	Description and Expectations
	**Text Average Sentence Size** It calculates the average of words in sentences of an article body legitimate articles tend to have more articulated sentences, so usually contain more words to motivate facts and provides insights.
	**Text Average Paragraph Size** It calculates the average of words contained in the paragraphs of a body article. Legitimate articles tend to have fewer but longer paragraphs,because journalists tend to write long articles and divide them into short paragraphs.
	**FKG Read Level** The Kincaid readability index indicates how difficult is the understanding of a text. Higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read. Ex: Legitimae News: FKGR=14.1261839465 Fake News: FKGR=11.1213519553 Legitimate articles are written to be easily understood, instead fake news tend to have a lack of clarity, so they will have a low grade of

	readability.
	**Lexical Diversity** It calculates the ratio of the number of distinct words used over the number words in total. fake articles tend not to use the same words several times because they have longer and less articulated sentences. Instead, good articles try to emphasize the concept several times to make it clear to the reader.
	**PoS Diversity** It calculates the ratio of the number of distinct PoS used over the total number of possible PoS[8]. Journalists tend to use more grammatical forms, so good articles will have more distinct PoS than fake news.
	**Punctuation Ratio** Calculates the ratio of the number of punctuation in the title to the title length . In real articles titles do not tend to be longer than one sentence, so the ratio between punctuation and length is low, unlike fake articles which tend to put a lot of emphasis in the titles.

[8] The total number of Parts Of the Speech depends on the language. for Instance In English we use TreeTagger[12] therefore the total number of possible PoS is 58

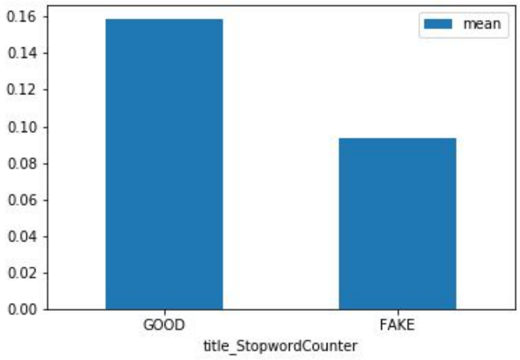
<table>
<tr>
<td>
</td>
<td><b>Stopword Ratio</b>
Calculates the ratio of the number of stopwords over the number of words in the title.
In legitimate articles titles tend to be shorter, so the ratio to stopwords is higher than fake titles.</td>
</tr>
</table>

Table 10 - Features explanation

5. Model selection

Text classification problems have been widely studied and used in many real applications over the last years. Especially with recent breakthroughs in Natural Language Processing and text mining, there is a huge interest in developing applications that leverage text classification methods.

Moreover there has been an exponential growth in the number of complex documents and texts that require a deeper understanding of machine learning methods to be able to accurately classify texts in many applications. In this case, different machine learning approaches were tested with positive results in natural language processing. The result in these learning algorithms relies on their capacity to understand complex models and non-linear relationships within data.

The algorithm tested are:

- Ridge Classifier
- Stochastic Gradient Descent Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- Passive Aggressive Classifier
- Deep Learning models
- LightGBM
- VotingClassifier
    - SGDClassifier
    - BernoulliNB
    - MultinomialNB
    - ComplementNB

5.1. Experimental results:

The approach described above has been compared to the academic paper “Learning Hierarchical Discourse-level Structure for Fake News Detection[16]”, since their approach takes as training sets most of the samples used to build Fandango’s classifier.

In that research has been highlighted the following two points:

- the fact that the structures of fake news documents, at the discourse-level, are substantially different from those of the legitimate ones.
- the identification of insightful and interpretable information from extracted structures can delineate intrinsic differences between fake and legitimate documents.

It means that the main objective is quite similar and it is to find some features in the structural shape of a news to discriminate them in a classifier.

Report Performance		
	Method	Accuracy (%)
Hamid Karimi et al. 2019[9]	BiGRNN-CNN	77.06
	LSTM[w+1]	80.54
	HDSF	82.19
Fandango[10] **The main results are shown below*	LGBM	**86.11**
	Deep Learning model (Mainly Layers: dense, dropout)	77.21

Table 11 - Report Performance

[9] Use 5 differents dataset, in Fandango’s training set 2 of this datasets has been ignored because they reported many different labels and most of the samples were claims, it could add some bias to the results: https://www.buzzfeed.com; http://www.politifact.com/

[10]Total Number of Items annotated Fake Article 6214 ; Real Article 5241. The performances have been calculated splitting the G.T. in training set (80%) and test set (20%)

The algorithms that best performs is Light GBM (Table 11), with the following parameters:

- limit the max depth for tree model : -1 as default
- type of algorithm used : traditional Gradient Boosting Decision tree
- number of leaves in full tree : 100
- the maximum depth of tree: -1(means no limit)
- the impact of each tree on the final outcome: 0.1
- number of boosting iterations: 200
- number of threads for LightGBM: -1 (no limit)

These are the preliminary results, actually using LIME[24] framework it's possible to obtain a local explanation of a prediction, without any kind of distinction of the model's complexity. This explanation doesn't allow to consider the model as black box. In other words, without any justification in the choice of the classifier.

Further experiments will be carried out on the data manually recorded by our journalists and on the labelled domains. Unfortunately, the models resulting from the training using these different G.T. will not be comparable with the state of the art algorithms. However, different authors might be involved in the performance evaluation, using their models on Fandango's datasets.

5.2. Future Agnostic experiments:

The use of different language in Fandango's project is a very remarked objective, so more experiments will be carried in order to understand how to approach this challenge.

Therefore, a wide range of expressions will be performed to study and analyze the performance of a model, trained in a language, and to calculate a performance in another different language.

In addition, a generic template will be formed using a dataset containing all the news in all languages, and tested on different test sets, one for each language of Fandango. Even if it seems an ambitious challenge, preliminaries experiments,not reported in this deliverable, indicate that this direction could lead to good results.

6. Conclusions:

The big spread of fake news and its impact on democracy, justice, and public trust has increased the demand for fake news analysis, detection and intervention.

Current academic research focuses on fake news from four perspectives: the false information it carries, its writing style/stylometric, its propagation patterns, and the credibility of creators and spreaders.

Fandango, with its agnostic approach, is trying to face the problem at a higher level of abstraction than classic experiments, addressing it from a knowledge-based perspective, studying fake news from a style-based perspective to emphasize the news content investigation.

However, knowledge-based studies aim to evaluate the authenticity of the given news, while style-based studies aim to assess news intention. These intuitions and fundamental theories have motivated and made possible style-based deception studies, whether for statements, online communications, online reviews, or news articles.

7. References:

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

[2]Argamon, Shlomo & Rachel Shimoni, Anat. (2003). Automatically Categorizing Written Texts by Author Gender. Lit Linguistic Computing. 17.

[3] Volkova, Svitlana & Shaffer, Kyle & Yea Jang, Jin & Hodas, Nathan. (2017). Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. 647-653. 10.18653/v1/P17-2102.

[4] Rubin, Victoria & Conroy, Nadia & Chen, Yimin & Cornwell, Sarah. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. June 2016.

[5] Dan Klein and Christopher D. Manning. 2003. A parsing: fast exact Viterbi parse selection. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 40-47. DOI: https://doi.org/10.3115/1073445.1073461

[6] A Stylometric Inquiry into Hyperpartisan and Fake News - Martin Potthast; Johannes Kiesel; Kevin Reinartz; Janek Bevendorff and Benno Stein - ACL 2018

[7] Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, andChu-Ren Huang. 2017. Fake news detection throughmulti-perspective speaker profiles. InProceedings ofthe Eighth International Joint Conference on NaturalLanguage Processing, IJCNLP 2017, Taipei, Taiwan,November 27 - December 1, 2017, Volume 2: ShortPapers, pages 252–256.

[8] SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours - Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, Arkaitz Zubiaga - April 2017.

[9] Udo Undeutsch. 1967. Beurteilung der glaubhaftigkeitvon aussagen.Handbuch der Psychologie, 11:26–181.

[10] Victoria Rubin, Niall Conroy, and Yimin Chen. 2015.Towards News Verification: Deception DetectionMethods for News Discourse. InProceedings of theHawaii International Conference on System Sciences(HICSS48) Symposium on Rapid ScreeningTechnologies, Deception Detection and CredibilityAssessment Symposium, Kauai, Hawaii, USA

[11] Rhetorical Structure Theory as a Feature for Deception Detection in News Reports in the Russian Language Pisarevskaya D. Institute for System Programming of the RAS, Moscow, Russia - June 2017.

[12] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, SvitlanaVolkova, and Yejin Choi. 2017. Truth of varyingshades: Analyzing language in fake news and politicalfact-checking. InProceedings of the 2017 Conference On Empirical Methods in Natural LanguageProcessing, EMNLP 2017, Copenhagen, Denmark,September 9-11, 2017, pages 2931–2937.

[13] Benjamin D. Horne and Sibel Adali. 2017. This justin: Fake news packs a lot in title, uses simpler,repetitive content in text body, more similar to satire than real news.CoRR, abs/1703.09398

[14] Bond, G. D., Holman, R. D., Eggert, J.-A. L., Speller, L. F., Garcia, O. N., Mejia, S. C., Mcinnes, K. W.,Ceniceros, E. C., and Rustige, R. (2017) ‘Lyin' Ted’, ‘Crooked Hillary’, and ‘Deceptive Donald’: Language of Lies in the 2016 US Presidential Debates. Appl. Cognit. Psychol., 31: 668– 677. doi: 10.1002/acp.3376.

[15] Rhetorical Structure Theory: Toward a functional theory of test organization - William C. MANN and SANDRA A. THOMPSON.

[16] Learning Hierarchical Discourse-level Structure for Fake News Detection - Hamid Karimi, Computer Science and Engineering. Michigan State University- Aprile 2019.

[17] Fake News Detection as Natural Language Inference - Kai-Chou Yang, Timothy Niven, Hung-Yu Kao - February 2019.

[18] A Stylometric Inquiry into Hyperpartisan and Fake News - Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, Benno Stein - July 2018

[19] Word translation without parallel data - Alexis Connea, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer , Hervé Jégou - January 2018.

[20] 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.

[21] T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean 2013 Distributed representations of words and phrases and their compositionality.

[22] Turian, Joseph, Lev Ratinov, and Yoshua Bengio. "Word representations: a simple and general method for semi-supervised learning." Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010.

[23] Project report for IKT407 in Autumn 2007 by Thomas Jakobsen and Thomas Skardal, November 2007

[24] "Why Should I Trust You?": Explaining the Predictions of Any Classifier - Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin - August 2016

[25] Automatic Detection of Fake News Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea - August 2017