Truth or Lie? Automatically Fact Checking News

Lucas Azevedo
Insight Centre for Data Analytics
Galway, Ireland
lucas.azevedo@insight-centre.org

ABSTRACT

In the actual scenario of ever-growing data consumption speed and quantity, factors like news source decentralization, citizen journalism and democratization of media, make the task of manually checking and correcting disinformation across the internet impractical or infeasible . Here, there is an imperative need for a fast and reliable way to account for the veracity of what is produced and spread as information: Automatic fact-checking.

In this work we present the problem of fact-checking in the era of big data and post-truth. Some existing approaches for this task are presented and their main features discussed and compared. Concluding, a new approach inspired on the best components of the existing ones is presented.

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*;

KEYWORDS

Natural Language Processing; Automatic Fact-checking; Deception Detection

ACM Reference Format:
Lucas Azevedo. 2018. Truth or Lie? Automatically Fact Checking News. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3184558.3186567

1 PROBLEM

The ongoing migration of the news business to the web, has as an immediate consequence a reduction of working places for journalists and with that, a reduction of the quality of its main product: information. For instance typically, there were five editions before a news source was published, now, publishers tend to have one or two. Sometimes, there is no edition. [28] Combined with the lack of reviews, the increasing precariousness of the journalist career led to a series of functions stacking for a single employee resulting in almost no filtering of the content produced.

Another significant impact on the media scenario is what specialists call citizen journalism[25], a process of democratization of the ways of producing media that is both cause and effect of the decrease of its costs.

In this new paradigm of media blogs, forums and social networking websites are not subject to traditional journalistic standards which affects the accuracy of information reported by these sources.[21] Thus, it became viable to produce a new kind of media business: Fake News.

Nevertheless, this new structure has many supporters that imagine this new model of information distribution as being the outcome of a wave of democratization in the media, rupturing the monopoly of big news companies, but also many adversaries that see in it "a world without editors, of unfettered spin, where the loudest or most agreeable voice wins and where truth is the first casualty"[23].

In this scenario, journalistic fact-checking arises as a measure to prevent false news, hoaxes and incomplete or neglected information to spread. Many press companies, websites and journalistic groups (listed in section 3) work on the hard tasks of: monitoring social media, identifying potential false claims and debunking or confirming them, always presenting arguments that support their verdict and these arguments' sources. But manual fact checking is an intellectually demanding and laborious process, and as Jonathan Swift once said in his classic essay "the Art of political lying": "Falsehood flies, and truth comes limping after it". [1]

To better understand the necessity of improvements in the automatic fact-checking field, add to the above described scenario, that when it comes to identifying a false claim, we, humans cannot perform a simple binary classification over deceptive statements with an accuracy much better than chance, In fact, "just 4% better, based on a meta-analysis of more than 200 experiments."[3] .Furthermore, humans typically find only one-third of text-based deceptions [11, 13]. This reflects the so-called 'truth bias' or the notion that people are more apt to judge communications as truthful [26].

At last, we should not forget that during the process of debunking fake news, fact-checkers can often over expose claims that are false, exaggerated or half-truths[14], potentially increasing the number of people that would interpret the bad piece of information as truth. This might happen more often than we expect due to the fact that humans have an automatic unconscious response when exposed to a counter-evidence to a false idea in that they believe is true. This exposure paradoxically not only fails to debunk the false idea but increases the confidence of it in their minds. This effect is known as the Backfire effect, and it has to be in the center of the discussion to any fact-checking project that aims to deliver its results with the objective of clarification of the population.

Outside the journalistic scope, fact-checking and deception detection techniques have been already used in areas as interpersonal psychology, law enforcement, credibility assessments, police work and homeland security, computer-mediated communication and NLP (or text analytics)[23] In this work we try to bring some of

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '18 Companion, April 23–27, 2018, Lyon, France
© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.
ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186567

the advancements in those areas to the field of Computational Fact-Checking.

We define the task of Computational Fact Checking as a "four state process, whether they are done - separately or not - by humans and/or machines" [2]. The stages are (i) monitor social and news media, (ii) identify check-worthy claims, (iii) check claims regarding their veracity and (iv) create and publish a verdict. In this work, we are mainly interested in automating the **third step,** which represents perhaps the most complex of the four, we will be deliberately focusing on written language, since it represents the type of data most frequently encountered on the Web [20], although acoustic and other non-linguistic features were also found to be useful for this task [15] and might be later explored in future work.

2 STATE OF THE ART

The state of the art in fact-checking comprises many different approaches that can be represented in a spectrum that varies from generic domain but theoretical to practical but domain specific approaches.[2] In the central section of this spectrum are the machine learning approaches, where we are going to focus.

The problem with the process of automating the conventional approach is that it is extremely hard to translate the operations made by the journalist to a computer, mainly because these operations greatly vary from case to case, hence the focus of many fact-checking works into machine learning-based approaches.

This is again a broad field and encompasses many techniques with many differences between them. Below we present a better classification of these techniques based on the type of data used as input: Unstructured content-based, Structured content-based and Context-based techniques.

2.1 Unstructured content-based techniques

"In spite of the attempt to control what they (liars) are saying, language "leakage" occurs with certain verbal aspects that are hard to monitor such as frequencies and patterns of pronoun, conjunction, and negative emotion word usage." [6]. The aspects cited by Conroy et al. are not the only ones to become more prominent in deceptive texts; by identifying and measuring them on unstructured text, a neural network can be trained to classify a text into deceptive or not. These aspects are also called Linguistic Based Cues (LBCs). Amongst many others, some examples of LBCs are: polarity, objectivity, occurrence of hedge words, count of words by POS tag, pausality, redundancy, emotiveness. [10][30]

Many of these LBCs have been used by different protocols for text - and multimodal - analysis on different domains, even outside the investigative journalism area. Examples of these protocols are: Criteria-Based Content Analysis (CBCA), Reality Monitoring (RM), Scientific Content Analysis (SCAN), Verbal Immediacy (VI), Interpersonal Deception Theory (IDT) Statement Validity Analysis (SVA) and Behavioral Analysis Interview (BAI). Each one of the cited protocols is defined by which set of LBCs are to be accounted when trying to classify a discourse into deceptive or not. [10][30]

For different automatic linguistic approaches, different sets of LBCs are selected and have their occurrence measured in the piece of text that is having its veracity checked, then, these measurements are used as input features in a classification machine learning model.

Some of these methods also rely on using search engines to gather documents - in form of blog texts, news articles, social media posts, etc. - related to the fact candidate and using LBCs to determine how does that document contributes to supporting or oppose it. In this case, the problem of source quality becomes extremely important but a positive factor also comes up: the easy presentation of documents that support the classification output.

On the negative side with respect to linguistic methods based on features that mainly rely on word frequency, "any resolution of ambiguous word sense remains non-existent." [6, 17]

Nevertheless, methods based on retrieving non-structured data using search engines have shown able to "detect emerging false or true claims with a macro-averaged accuracy of 80% within five days of it's publication on the web, with as low as six reporting articles per-claim" [22].

2.2 Structured content-based techniques

Another way to use machine learning techniques into verifying the veracity of a statement is to first try to convert the natural language sentence into a structured form, which can be performed by an Information Extraction (IE) process. [21]

This approach might seem simpler and more efficient than the others as the answers derived from knowledge bases (KBs) are more deterministic, but it has many drawbacks. The first is the IE process and the inability to come up with an accurate classification if it fails. Also, finding domain-specific structured data is not always possible, creating the appropriate query is another difficulty, and besides that, there is a chance of losing potential important information that was present in the original text during the process of translating into structured form. Is important to say that, when compared to unstructured content-based techniques, these methods have a weaker or non-existent reasoning that support the verdict over a truth/false claim.

The main problems and objectives of projects based in this approach can be summarized in some questions presented by as "How to find datasets relevant to given claims?", "How to automate the translation from text to structured claim?", followed by "How to formulate queries to "check these claims. There are some other open research questions for less simple cases, sucg as "how to check claims that cannot be readily derived from structured data?", "can we automatically generate counterarguments for claims classified as false?" and "can we automatically "reverse-engineer" vague claims to recover the omitted details?". Some of these questions are answered in [14, 29].

2.3 Context-based techniques

The third and last kind of approach listed under the ML section does not take the textual content into account and is more suitable for short texts such as micro-blogs. Instead, using (social) network information as spreading patterns, time stamp data, user profile, engagement score, reach and repost occurrences is the main characteristic of this approach.

Other textual classifiers that make use of natural language processing (NLP) features and ML can achieve a high performance by ignoring semantic aspects and focusing only at the grammatical and/or structural information of the text, as "useful fact-checking

can be done without understanding anything about the claim itself" [2].

Future work will show us what can ultimately be reached within the field. Questions about the capabilities of an automated fact-checking system are many ("can we automatically link claims to structured datasets of the related topic?", "can the queries that would answer our question be formulated without human intervention?", "can we anticipate what claims may be made soon?" [14], etc..) but the improvements made put us in an optimistic position.

Once again restricting in the scope of research, we will be focusing in how to make use of the listed machine learning approaches, combining linguistic, structural and contextual information. First we describe the variety of different processes and features that can be included in the task of performing fact-checking and at the end of this section we present Table1 where we categorize the many articles cited based on the approaches they take.

Before presenting Table1, we have enumerated and explained the distinct categories on which the methods will be grouped. The above described type of data is also one of the aspects analyzed in the studied methods.

Granularity of analysis

This aspect defines the semantic level from where the features are extracted. It goes from the higher granularity level of a whole document (in the case of LBC's), passing through the sentence level (where sentiment analysis can be applied), to arrive at the word level (BOW, keywords).

Machine Learning Features

The selection and engineering of features is by far the most unique aspect of each project. The main difficulty in this process is to be concise, i.e. use a set of features that include the most amount of information without being redundant.

LBC's are a great example of good features for this task, as they have been frequently used in psychology [20]. In [8], De Paulo evaluated more than 100 different cues regarding their entailment to deception presence in texts. Zhou, in [30], does the same with 27 other LBC's.

Analyzing the style of writing of a source can be of great value, as stated by [21] and confirmed by [22]. After proving the correlation between the objectivity of a text and its veracity, i.e., the more objective a text is, the higher the chance that it is carrying true facts. Expectations rise for classifier accuracy improvement if similar aspects to objectivity are measured in the input text and taken into account in the form of features. [21]

Entailments between writing style and veracity of a document are explored by other projects [22, 27]. Many fact-checking methods also rely on features often used by sentiment analysis classifier[18]

On the other hand, there are also works that suggest that lower-level features such as word quantity, verb quantity, and sensory ratio should be more often used due to their importance and generality across different models. [19]. Between these methods, some advocate the use of large feature sets, using "message content, user profiles and holistic statistics on [network] diffusion patterns" [4]. Other work observed an increase accuracy when reducing the size of their model's input vector by only taking into account the most frequent bigrams instead of all uni and/or bigrams[21].

Contextual information is highly acclaimed as being an important resource for feature engineering that is often neglected.[16, 31] "User information can be a strong clue in the initial broadcast, content features are important throughout entire propagation periods, and structural and temporal patterns help for longitudinal diffusion" [19]

When relying mainly on contextual features as the propagation tree of an input instance, the task of fact-checking can be modeled as a similarity problem, avoiding the painstakingly process of feature engineering by using kernel-based methods. [7, 19]

Type of Classifier

In Table 1 presented below, we also list the kind of machine (deep) learning model which are used to classify the input, as well as the number of classes determined by the authors of each project.

Source Dependency

A big improvement on the output of a model can be achieved by assessing the quality of sources used to measure the veracity of a claim, especially for projects that rely on multiple sources and see variety of sources as a positive aspect. In this cases, the problem of source dependency becomes extremely important, since it can be intuitive to think that having different sources with similar opinions towards a claim is positive, and in fact is, but only in the case where there is no source duplication, and this has to be evaluated[9]. This task is harder than many would think, as the original sources (in the sense of being a non-duplicate source) are more likely to make similar comments about aspects of the claim subject as other original sources.[6]

Keeping a record for each source is the way to be able to evaluate its quality. By comparing what that specific source has stated about a claim which veracity is known, the source reliance can be updated. In order to find out which sources are dependent to others, a good approach is to analyze in which frequency both suspect sources come up with a false value for a known claim veracity.[9]

Accountability/Reasoning/Justification

Having the capability of presenting supporting evidence besides accurately classifying a document or claim as being true or false, is one of the pillars of the "Holy Grail", term coined by [14] to refer to the ultimate objective of systems that tackle the challenge of Computational Fact Checking.

A lot is expected from mathematical models concerning to the reasoning over their outputs. Having an accurate model is good, but providing a justification alongside the verdict can allow a higher assurance to the user, deliver more information about the claim being analyzed and also help to understand and improve the model, in the development stages.

Evaluation Metrics

Most of the projects studied here rely on the harmonic mean between precision and recall well established measure for machine learning and, more specifically, NLP scientific works, also known as F1-score.

Table 1: Machine Learning-based Approaches for Automatic Fact-checking

Article	Input	Classifier	Features	Granularity
[10]	Unst.	Binary Perceptron (1L 3N)	Word Level	Constructed Set of LBCs
[20]	Unst.	Binary (Naive Bayes)	Word Level	LIWC, a set of LBCs
[21]	Struct.	Binary (Logistic Regression)	Word Level	Subj.+Sent. Lexicons
[19]	Unst./Struct.	Binary (Propagation Tree Kernel)	Word/Context	User prof./Timestamp
[22]	Unst.	Binary (CRF)	Word/Context	Trend/Content/S.Rank
[5]	Struct.	Binary (Weighted Knowledge Graph)	N/A	N/A

Article	Source Dep.	Accountability	Evaluation
[10]	N/A	By LBC	74% Ov. Acc
[20]	N/A	By Word Class	59.8% Ov. Acc
[21]	YES	Not mentioned but possible	70-90% Ov. Acc
[19]	N/A	NO	73-75% Ov. Acc
[22]	YES	By web documents	80% Ov. Acc.
[5]	N/A	NO	61-95% Ov. Acc

3 CONTRIBUTION AND METHODOLOGY

The results reached so far consist mainly of a careful study of the problem and its current solutions as well as a well defined direction be followed and steps to be taken, as described in the section below.

Based on the needs identified on the field and the directions pointed by the methods studied, the first step of our proposed approach consists of the construction of a dataset of manually fact-checked claims by different fact-checking websites as Channel 4 FC, Snopes.com, FactCheck.org, Politifact.com, FullFact.org and OpenEurope.org.uk. Some effort will have to be made to map the data from these different sources into a common structure. There is no reason not to share this resource, once it is built.

After obtaining the data, a further preprocessing step for is to identify candidate claims which claims are worth checking, for this, initially, we will be using ClaimBuster[14], a third-party framework that has presented good results.

Then the core step of the contribution: the development of a neural network for automatically binary classification of claims into true or false, to be trained on the sentences extracted from the dataset above described.

From an initially larger feature set, containing: frequent bi-grams, usage of hedges, assertive and factive verbs, and other high semantic level LBC's as objectivity, pausality, etc. A smaller optimal set of features will be selected to start the model training. This selection process will performed by sensitivity analysis over the initial and larger feature set. In order to to represent important aspects of a text, as certainties, speculations and doubts, higher-order linguistic features will be experimented as well.

Finally, after having the classifier trained on the described dataset, we plan to evaluate our model with data coming not only from the cited websites, but also from different sources and domains, including social media, blog posts, news articles, etc using the F-1 score in order to measure its accuracy.

4 CONCLUSIONS AND FUTURE WORK

The main conclusion of this work is that the process of feature engineering has great impact in the results of a classifier and higher level textual information can lead to higher accuracy. That said, multimodal approaches bring high expectations for improvements in automatic fact-checking's state-of-the-art, as more and more high-level correlations can be discovered by analyzing extra textual information.

As expected, different domains have their own particularities, but for linguistic approaches, some sets of common linguistic based cues are orthodox and most likely will be related to a higher overall accuracy if considered in the model, regardless of the genre of data.

It could not go without saying that structural information is a resource that can be used for improvement of results, especially for social media, as linguistic information is less representative of the data due either the shortage of characters used, in the case of short texts or microblogs, or by the frequent usage of neologisms, acronyms and nonlinguistic data.

We also noted that different kinds of fake-news differ in the type of media they occur more often and that by using different types of media we can increase or decrease the detection of each one of them. For example: tweets are especially suited for hoax detection.[12]

After having the initial steps implemented and a running classifier prototype, many other techniques can be implemented and evaluated. That is the case of auxiliary contextual information, as shown in [19] and multi-modality as suggested by [15]. Some of the most promising features can be better measured if not only textual data is available as is the case of pausality, arousal and many others. [10, 24, 30]

There are still some room for improvement in the pre-classifying steps, where the semantic information of a document or claim is not represented clearly or not even in textual format, examples of these cases are metaphors, multi-sentence claims, complex paraphrases, emojis, acronyms, neologisms, etc.

5 ACKNOWLEDGEMENTS

This work is funded by the SFI Grant agreement No 12/RC/2289.


Co-financed by the European Union
Connecting Europe Facility

REFERENCES

[1] John Arbuthnot and Jonathan Swift. 1874. *The Art of Political Lying*. K. Tompkins.
[2] Mevan Babakar and Will Moy. 2016. The State of Automated Factchecking. *Full Fact* (2016).
[3] Charles F Bond Jr and Bella M DePaulo. 2006. Accuracy of deception judgments. *Personality and social psychology Review* 10, 3 (2006), 214–234.
[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.
[5] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015), e0128193.
[6] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.
[7] Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 423.
[8] Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin* 129, 1 (2003), 74.
[9] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment* 2, 1 (2009), 550–561.
[10] Christie M Fuller, David P Biros, and Rick L Wilson. 2009. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems* 46, 3 (2009), 695–703.
[11] JF George and BT Keane. 2006. Deception detection by third party observers. In *deception detection symposium, 39th annual Hawaii international conference on system sciences*.
[12] Cale Guthrie Weissman. 2015. "Computers can now tell the difference between real breaking news and internet trolls trying to dupe us". https://goo.gl/x2D2Zi. (2015).
[13] Jeffrey T Hancock, Jennifer Thom-Santelli, and Thompson Ritchie. 2004. Deception and design: The impact of communication technology on lying behavior. In *Proceedings of the SIGCHI*. ACM, 129–134.
[14] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. *world* (2015).
[15] Julia Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, et al. 2005. Distinguishing deceptive from non-deceptive speech. In *Ninth European Conference on Speech Communication and Technology*.
[16] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 1103–1108.
[17] David F Larcker and Anastasia A Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50, 2 (2012), 495–540.
[18] Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*. ACM, 342–351.
[19] Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 708–717.
[20] Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, 309–312.
[21] Ndapandula Nakashole and Tom M Mitchell. 2014. Language-Aware Truth Assessment of Fact Candidates.. In *ACL (1)*. 1009–1019.
[22] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1003–1012.
[23] Victoria L Rubin, Yimin Chen, and Niall J Conroy. 2015. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.
[24] Victoria L Rubin and Niall Conroy. 2012. Discerning truth from deception: Human judgments and automation efforts. *First Monday* 17, 3 (2012).
[25] Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014*. 18–22.
[26] Aldert Vrij. 2000. *Detecting lies and deceit: The psychology of lying and implications for professional practice*. Wiley.
[27] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2006. Learning subjective language. *Learning* 30, 3 (2006).
[28] Jim Wright. 2017. "Fake News, Facts, and Alternative Facts" [How News Has Changed]. https://www.coursera.org/learn/fake-news-facts-alternative-facts-michiganx-teachout-2x/lecture/sxW8D/veteran-journalist-on-how-news-has-changed. (2017). Accessed 10/04/17.
[29] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward computational fact-checking. *Proceedings of the VLDB Endowment* 7, 7 (2014), 589–600.
[30] Lina Zhou, Judee K Burgoon, Jay F Nunamaker, and Doug Twitchell. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation* 13, 1 (2004), 81–106.
[31] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11, 3 (2016), e0150989.