LUX (Linguistic aspects Under eXamination): Discourse Analysis for Automatic Fake News Classification

Anonymous ACL-IJCNLP submission

Abstract

The democratization/decentralization of both the production and consumption of information has resulted in a subjective and often misleading depiction of facts known as Fake News - a phenomenon that is effectively shaping the perception of reality for many individuals. Manual fact-checking is time-consuming and cannot scale and although automatic fact-checking, vis a vis machine learning holds promise, it is significantly hindered by a deficit of suitable training data. We present both a novel dataset, VERITAS(VERIfying Textual Aspects), a collection of fact-checked claims, containing their original documents and LUX(Language Under eXamination), a text classifier that makes use of an extensive linguistic analysis to infer the likelihood of the input being a piece of fake-news.

1 Introduction

Also defined as the intentional or unintentional spread of false information (K et al., 2019), Fake News has found fertile ground in the actual scenario of ever-growing data consumption and generation, where factors like news source decentralization, citizen journalism, democratization of media and astroturfing[1] (Lee, 2010) make the task of manually checking and correcting disinformation across the internet impractical if not infeasible, (Shao et al., 2016) despite the great amount of effort put by Fact-Checking Agencies, i.e. groups of journalists that manually identify and investigate rumours conveyed by Fake-news articles.

Consequently it is imperative that we develop an efficient and reliable way to account for the veracity of what is produced and spread as information: this process is known as automatic fact-checking. (Hassan et al., 2015)

[1]Astroturfing is the practice of masking the sponsors of a message or organization to make it appear as though it originates from and is supported by grassroots participants.

Although the there has been significant research effort to tackle the task of automatic fact-checking (Azevedo, 2018) the deficit of data collections containing organic news article - in their entirety - that were manually labeled with respect to their veracity is a common obstacle for the development of classifier models, especially the ones focused on supervised learning and/or document-level analysis. The absence of such data makes researchers rely on other approaches, e.g., stance determination (Popat et al., 2017), knowledge base matching (Wu et al., 2014), trust assessment of sources (Balakrishnan and Kambhampati, 2011), data structuring (Conroy et al., 2015), network pattern analysis (Shao et al., 2016), etc.

In this work we present the challenges faced in the process of developing a language model enriched by discourse features for fake-news detection, along with experimental results. The contributions of this work are mainly two: the Dataset Creation process, described in Section 2 and the introduction of the Text Classification Model, named LUX, in Section 3.

Section 4 brings a comprehensive evaluation of both VERITAS and LUX, while also featuring an ablation analysis of the latter.

2 Datasets for Fake News Classification

2.1 Available Corpora on Fake News

The deficit of suitable corpora for the intended approach is the main influence behind the creation of the VERITAS Dataset, and by consequence, the VERITAS Annotator. Below we present a list of datasets commonly used in related tasks. Note that, although those are valuable resources for many related tasks, none of them include all of the three most important characteristics required for a content based supervised classifier: significant volume of entries, gold standard labels and the whole fake

news articles (i.e., the origin).

Emergent16 a collection of 300 rumoured claims and 2,595 associated news articles - a counterpart to 'origin' in the VERITAS Dataset. Each claim's veracity is estimated by journalists after they have judged that enough evidence has been collected (Ferreira and Vlachos, 2016). Besides the claim labeling, each associated article is summarized into a headline and also labelled regarding its stance towards the claim. Because of this labelling of origins and the fixed structured of the website we could obtain some valid examples with a scraper.[2] Unfortunately they sum up less than 100 usable claim-origin pairs (discussed in subsection 2.3).

LIAR17 includes around 13K human-labeled short statements which are rated by the fact-checking website PolitiFact into: "pants on fire", "false", "barely true", "half true", "mostly true", or "true" (Wang, 2017). The domain-restricted data as well as the reduced length of text that can be retrieved from this corpus makes it unsuitable for linguistic fake news detection for generic domains.

FakeNewsNet18 is a data repository containing a collection of around 22K real and fake news obtained from Politifact and GossipCop[3] fact-checking websites. Each row contains an ID, URL, title, and a list of tweets that shared the URL. It also includes linguistic, visual, social, and spatiotemporal context regarding the articles. This repository could still be used for supervised learning models if it were not for the fact that it does not provide sufficiently long texts to be used by a classifier based on linguistic aspects. For the same reason, CREDBANK (Mitra and Gilbert, 2015) and PHEME (Derczynski and Bontcheva, 2014) are also unsuitable for the authors' use case. Those three datasets focus on network indicators (e.g. number of retweets, sharing patterns, etc.) of fake news, instead of its contents. CREDBANK is a crowd sourced corpus of "more than 60 million tweets grouped into 1,049 real-world events, each annotated by 30 human annotators", while PHEME includes 4,842 tweets, in the form of 330 threads, related to 9 events.

FEVER18 The FEVER corpus (Thorne et al., 2018), created a set of more than 185K claims by modifying sentences from a collection of 50K Wikipedia articles. Annotators were then given the task to annotate other sentences from the same article in respect to their stance towards the modified sentence. The corpus is the largest to our knowledge, but since it is synthetically created and focused on a sentence-level stance classification approach, it is unlikely to perform efficiently on heterogeneous web documents as a fake news classifier.

Snopes19 (Hanselowski et al., 2019) provides a large collection of more than 16 thousand manually annotated text snippets extracted from 6,422 snopes.com articles. Unfortunately, less than half of those snippets present a stance (agreeing or disagreeing) towards the fact-checked claim. Also, the annotated snippets are, by definition, only a portion of the original article. Nevertheless, an origin identification process could generate a significant amount of valid examples from this data.

Due the restricted length of this submission a more detailed description of the following dataset is not provided, although is important to include them in this list: **BuzzFeed16** (Potthast et al., 2018), **Kaggle**[4] and **NELA17** (Horne et al., 2018).

2.2 The VERITAS Dataset

The VERITAS Dataset is, to our knowledge, the most complete data collection of manually annotated claims in regards to their veracity by being the only one to contain not only the mentioned veracity labels, but also the document that originated the checked claim in its entirety. Its creation can be better understood as a two step process: **1)** Fact-Checking articles scraping and **2)** Claim Origin Identification.

Step 1: Scraping FCAs As the cost for manually checking a large number of disputed claims is extensive, both in time and money, we have started the dataset creation process by scraping articles from fact-checking agencies and consequently

[2] While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler.

[3] https://www.gossipcop.com

[4] https://www.kaggle.com/mrisdal/fake-news/data

trusting the work made by their journalists that undertake the processes of: 1)selecting controversial claims, 2)leveraging web documents that either support or deny those statements to 3) finally come to a veracity verdict. In simple terms, a Fact-Checking Article (FCA) is a narrative of this investigative process.

For each scraped FCA, we create an entry in the dataset and extract a number of attributes, most importantly: the claim, the veracity label, and the list of hyperlinks to the mentioned web documents, which we call *Origin Candidates*, since they will be the subject of the Origin Identification process. The code used to scrap the pages is openly available[5].

Step 2: Claim Origin Identification One of the most important steps of the dataset creation pipeline was a task we defined as "origin identification". In short, after three automatic ways of identifying the article in which a fact-checked claim originated were carried out and yielded non-satisfactory results, it was decided that a manual annotation process would be used to select the correct entries from the totality of the dataset. An annotation tool[6] was developed in order to make the task easier and faster. This annotation process not only provided a large and complete version of the dataset, but also leaves a possibility for an automation of the origin identification process as a future improvement of the project.

The final structure of each entry contains the following fields: Fact-Checking Article URL, Checked Claim, Claim Label, Tags, FCA date, Origin URL, Origin Domain, Origin Body, Origin Title, Origin Summary, Origin Keywords, Origin Date and Origin Author. Given the limited space, a more in-depth description of each field is not provided but can be found within the supplementary material (appendix 1) and also along an extensive description of the origin annotation process in `removed for anonimity purposes`.The past versions of the dataset are also openly available[7]

2.3 Consolidation of VERITAS Dataset

A consolidation of the VERITAS dataset followed the large annotation process over the scraped FCA pages that augmented both the quantity of annotated origins (1032 consolidated origins from more than 10k annotations) and the quality of the annotations, measured by Krippendorff's Alpha[8], reaching a substantial score of 0.6014. This consolidation generated the fourth version of the dataset, here addressed as V4.

Given the constant structure of `Emergent.info` articles, we have also incorporated its few valid claims, i.e., the ones with "true" or "false" verdict, and their respective sources.

Although the majority of origins obtained from Emergent were linked to "true" claims, when aggregated to the consolidated origins from VERITAS v4.0, the data collection showed a false/true class imbalance ratio of ≈ 1.44. Therefore, in order to quickly obtain "true" labeled news articles to balance the scraped Dataset, reporting articles were scraped and automatically labeled as "true" and composed a **separated** dataset where their headlines are used for the claim field. The sources of those articles were selected according to studies determining the least biased[9] and/or most trusted[10] news outlets in the U.S..

The authors are aware that the label assumption of those articles is far from ideal. Notwithstanding, it offers another **option** of palliative solution for the label unbalance issue and yielded positive results in similar works (Horne and Adali, 2017; Ireland, 2018). It should, however, be tested with caution and compared with other - also sub-optimal - methods, i.e., discarding "false" entries and/or implementing class weights on the model training. Both the collection of reporting articles and the emergent articles are provided **separately** so they can be optionally disregarded and eventually substituted by gold-standard data. Table 1 gives a bit of information about each subset.

Since the improvement of incorporating the entries from emergent was still to be evaluated by the proposed classifier, two different sample sets from the trusted sources were created, to balance both the v4.0 dataset by itself (V4+T1), as well as the concatenation of VERITAS and emergent (V4+EM+T2). The evaluation results will be presented at Section 4, as they are also the evaluation

[5] https://anonymous.4open.science/r/df8101bf-cb3f-40ef-be9f-03b6d0ff401c/

[6] https://anonymous.4open.science/r/d77fe72f-8921-472b-af78-f7bd14f202ca/

[7] https://anonymous.4open.science/r/2ffd7e58-8717-43e7-8c6f-feacc4a00d47/

[8] https://en.wikipedia.org/wiki/Krippendorff%27s_alpha

[9] https://www.businessinsider.com/most-biased-news-outlets-in-america-cnn-fox-nytimes-2018-8

[10] businessinsider.com/most-and-least-trusted-news-outlets-in-america-cnn-fox-news-new-york-times-2019-4

Table 1: VERITAS Subsets

	#E	#T	#F	#U
VERITAS v4.0 (V4)	1032	276	664	92
Emergent (EM)	865	308	179	378
Trusted1 (T1)	388	388	-	-
Trusted2 (T2)	259	259	-	-
V4+T1	1420	664	664	92
V4+EM+T2	2156	843	843	470

*Columns represent **#E**: total entries, **#T**:true entries, **#F**: false entries, **#U**:unverified entries*

for the linguistic model. By comparing the two balanced sets we can have a better understanding of the quality of the data obtained from emergent, keeping in mind that the difference in volume of entries would still affect the performance.

3 LUX - Language Under eXamination

The core contribution of this work is the investigation of the usage of linguistic aspects as discriminative features in a text classification model that should determine whether the given article is fake or not. We call this classifier LUX, short for Language Under eXamination.

Previous work investigated the use of such linguistic aspects as features for similar tasks such as deception detection (Reichel and Lendvai, 2016; Zhou et al., 2004), document clustering (Yu and Hatzivassiloglou, 2003a), text classification (Louis and Nenkova, 2011; Biyani et al., 2016) among others. Related works make use of few (mainly one) of those aspects and the majority of them report an improvement of their results by doing so.

Here we present a set of linguistic aspects that were shown to be correlated to deception. For each of these aspects, we present their contextual definition, along with a short literature review and a description of the methods we use to evaluate its presence or absence in a given piece of text. The objective is to build LUX (Language Under eXamination), a Fake News Classifier, effectively using these linguistic aspects to estimate the likelihood of an article containing fake news. Here, we present the results obtained with two baseline language models (BERT[11] (Devlin et al., 2018) and Word2Vec (W2V) (Mikolov et al., 2013)) towards building this classifier.

We are aware of an imbued redundancy that our features might present, since the aspects analyzed by the different approaches, in some cases, overlap with each other, but expect that the eventual bias this redundancy might add to the model can be overcome with the implementations of techniques such as LDA (Linear Discriminant Analysis) or PCA (principal component analysis).

3.1 Linguistic Aspects

Subjectivity Louis and Nenkova (Louis and Nenkova, 2011) observed that general sentences tend to be more subjective. Some of the shallow features that are correlated to the subjectivity level of a sentence are also used in their model, for example, punctuation marks, average number of characters and average number of words.

Pattern[12], a python library for text analysis, states in its section about subjectivity: "Written texts can be broadly categorized into two types: facts and opinions." Based on a lexicon of adjectives produced for product review analysis, pattern.en provides a function that maps the subjectivity score of a sentence to a range between 0 and 1 depending on the number of adjectives it contains. It also provides implementations of measuring functions for mood and polarity.

Riloff et Wiebe (Riloff et al., 2003) presents a methodology for the creation of the MPQA Subjectivity Lexicon. In summary, the authors: 1) use an automatic subjectivity classifier to label data while also 2) identifying patterns present in the sentences labeled as subjective and 3) use the learned patterns to improve the classification model(1) and iterate between the three steps, making bootstrapping possible. The MPQA Lexicon is also used for us to measure the subjectivity of a given text. Based on the lexicon, (Wilson et al., 2005) also created OpinionFinder, a Subjectivity Classifier.

Another interesting method was presented by (Yu and Hatzivassiloglou, 2003a), where a Naive Bayes classifier is trained over a Wall Street Journal dataset containing two classes:

[11] Bidirectional Encoder Representations from Transformers

[12] https://pypi.org/project/Pattern/

Subjective (every article with type Editorial or Letter to Editor) and Objective (Business or News). By analysing low level features on the texts, the NB classifier achieved a 0.91 recall and 0.86 precision on the binary classification task.

Specificity Zhou et al. (Yu and Hatzivassiloglou, 2003b) uses specificity and measures it by words depicting the following aspects: perceptual information (sounds, smells, physical sensations and visual details) and spatio-temporal. (Fuller et al., 2009) measure bi-logarithmic type-token ratio (LogTTR) for evaluating specificity.

(Li and Nenkova, 2015) introduced Speciteller, a python framework for fast and accurate prediction of sentence specificity, which was enhanced and presented by (Ko et al., 2019). It introduces a new algorithm that adjust its weights to the training set, making it applicable to any domain, out-of-the-box. This is the implementation we are going to use.

Complexity (Biyani et al., 2016) focused on the detection of click-baits (that can be seen as a subcategory of fake news) and reported that features used to measure the formality of a text were the most correlated to click bait articles. Using a slang lexicon and a list of bad words, as well as several readability scores, they obtained a reasonable F-1 score of 74,9.

A 1999 paper by (Heylighen and Dewaele, 1999) presents a famous metric for Formality evaluation, named the F-measure (not to be confused with the F1 score). (Pavlick and Tetreault, 2016) present a statistical model for predicting formality, but do not provide access to the model's code.

Another famous work on the formality area is Coh-Metrix (Graesser et al., 2014), but the only access to its implementation is through a simple HTML portal, so we have discarded this option.

Fortunately, a python library[13] provides several readability measuring tools, including the implementations of the F-measure and CLScore, LIX and RIX, which were also used by (Biyani et al., 2016).

[13] https://pypi.org/project/readability/

Another python library[14], initially developed for the AFEL project, provides more measuring tools for semantic complexity analyzer. We make use of both libraries to implement the highest amount of unique metrics for Complexity, Formality and Readability.

Uncertainty (Szarvas et al., 2012) defines: "Uncertainty can be interpreted as lack of information: The receiver of the information cannot be certain about some pieces of information".

Victoria Rubin wrote her thesis (Rubin et al., 2006) on Certainty Identification. Following her work, Veronika Vincze also wrote a thesis (Vincze, 2015) on the same subject and, along with her group achieved great results (Vincze et al., 2008) on the CoNLL Shared Task 10, that aimed for the classification of uncertain texts from the BioScope corpus. The approach described in Vincze's thesis was implemented very conveniently as a python library for Uncertainty detection, that is used by us for uncertainty measurement.

(Reichel and Lendvai, 2016) tried to identify hoax-resolving tweets by using the ratio between four data augmented lexicons (knowledge, report, belief, and doubt) as features, along with low-level syntactic features, not achieving good results.

Loughran and McDonald Sentiment Word Lists and MPQA are Uncertainty Lexicons that are also used by us.

Affect (Pang and Lee, 2008) is an extensive review of the literature on sentiment analysis and opinion mining that encompasses the field of linguistic aspect evaluation, which this work is focused on.

(Whissell, 2004) provides the Dictionary of Affect in Language, which includes people's mean ratings for the Pleasantness, Activation, and Imagery of close to 9,000 words. The dictionary is a lexicon with ratings representing the two main dimensions of emotional space, valence and arousal, along with another rating for people's assessment of imageability, i.e., how easily it is to form a mental picture of a word.

[14] https://github.com/afel-project/pySemanticComplexity/blob/master/pysemcom.py

A better definition of Affect in the context of deception detection is necessary in order to decide which resource is more appropriate for the aspect evaluation, for now we are going to let the experiments evaluations indicate what is the most appropriate way of measuring affect for our task.

(Li and Nenkova, 2015) mention the MRC Psycholinguistic Database has words annotated w.r.t imageability among other aspects, while VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert, 2014) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to social media. Thus, it seems to be quite appropriate for us.

Verbal Immediacy (Mehrabian and Wiener, 1966) first defined Immediacy as a linguistic property that refers to the degree to which a source associates himself/herself with the topics of a message; that is, "immediacy is the degree to which a source approaches or avoids a topic". Based on that definition, (Zhou et al., 2004) measured it by analysing spatial and temporal terms, passive voice ratio, self reference manner and group reference manner, among others. Different works relate the non-immediacy to the presence of deception in text since these try to disassociate oneself from one's communication.

Negative affect and passive voice are some indicators of non-immediacy. Since the first is already addressed by us, we will be using a ratio between passive sentences over the total number of sentences to determine how passive is the text. In this context, a sentence is deemed passive, if it contains a "BE" verb followed by some other, non-BE verb, except for a gerund.

Diversity / Quantity / Pausality Those are syntactic features and some of the previous defined ones already make use of one or more ways of measuring them. For example, the diversity measurement is used to evaluate a sentence's Complexity. Still, there are many different ways to measure diversity and since we intend to remove the redundancy of the features anyways, we will measure it with many different formulas.

In a 2013 article, (Jarvis, 2013) proposed that the six properties of lexical diversity should be measured by Variability, Volume, Evenness, Rarity, Dispersion and Disparity. Using a python library[15], we measure some of those metrics.

Other simple aspects are also taken into account, as the overall quantity of words in absolute number and by P.O.S.-tag as well as the pausality, measured by the ratio between punctuation marks and number of sentences.

4 Evaluation and results

In simple terms LUX is a binary model for classifying general text into fake news / real news and it was originally proposed as a way to evaluate the efficiency of the above mentioned linguistic features. Aiming for generality, this model takes a text document (that could be a long article or a simple headline) as sole input and outputs the probabilities of it being fake or not, based on its psycho-linguistic profile and contextual representation. For the latter, different types of text encodings were tested and it became clear that the usage of fixed-size BERT document embeddings outperformed Word2Vec, which was tested on RNN, LSTM (Hochreiter and Schmidhuber, 1997) and Bi-LSTM (Schuster and Paliwal, 1997), with the latter having the best results, but still inferior to BERT.

After performing a grid search with different optimizers, activation functions, learning rates, training epochs and fully connected layer(FLC) dimensions, the initial model was decided to be composed of a simple ReLu[16] activated 64-dimensional FLC with a dropout of 30% attached to the final layer, of dimensionality 2 where a softmax filter would represent the false and true labels probabilities. Adam (Kingma and Ba, 2014) was the best performing optimizer and a combination of $\alpha = 0.001$ over 100 training epochs generally yielded the best results. Figure 1 brings an outline of the model.

All the reported values in Table 2 for Accuracy and F1 score come from a 9-fold training over the data. The results for the two best baseline models are also included, namely the same model using only the BERT document embeddings and only the w2v embeddings over a simple Bi-LSTM with 128 dimensions on the recurrent layer.

[15] https://github.com/kristopherkyle/lexical_diversity

[16] https://deepai.org/machine-learning-glossary-and-terms/relu

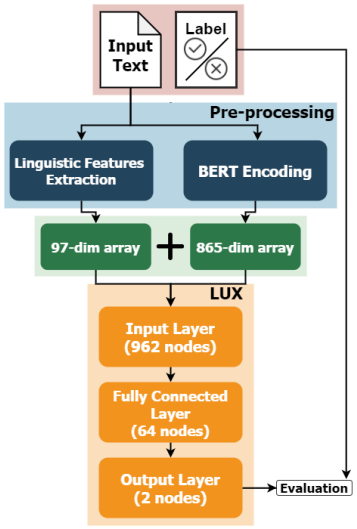

Figure 1: Outline of LUX classifier

Since the data from FEVER18 (Thorne et al., 2018) and Snopes19 (Hanselowski et al., 2019) is composed of short statements a comparative analysis is also presented alongside a V4+EM+T2 run using only the claims as input text, instead of the larger body texts.

The final input for each article is a an ensemble of a document embedding generated by BERT trained on BERT-Large uncased corpus[17] and the 97 linguistic features described in the previous section. An anonymous version of the code repository is available at https://anonymous.4open.science/r/56600808-abac-4999-a367-a279fdf08b69/.

Given the initial results, the robustness added from the a different source, i.e. emergent, with the benefit from balancing classes using the trusted news (T2) yielded the best results. Consequently, it was decided this was the selected subset for the linguistic features ablation analysis.

4.1 Ablation

Table 3 Brings the three most impactful positive and negative features, i.e. features that, when removed, most decrease or most increase the accuracy of the model, respectively. Those are all results using as base the best model run, i.e., LUX model over the V4+EM+T2 data, depicted in Table 2. A longer table containing the results for the full ablation analysis can be found within the supplementary material (appendix 2).

Positive Features(PF): When individually removed, each of the 97 features of the model, 50 have report a decreased accuracy of the model by an average of 0.056%, where 21 'contribute' with more than the average of all the positive features and only 10 features decrease more than 1% accuracy when absent. All three top PF fall into the Quantity group, as P.O.S.-tag counts, while most of the most sophisticated, i.e. higher semantic level, make to the top 10. Besides the ones featured(pun intended) in the table, the top 10 also comprises, unordered: Pausability, Coleman-Liau informality score, specificity, measure of lexical textual diversity(MTLD), and three features from the semantic complexity evaluator (Venant and d'Aquin, 2019): assortativity, average number of in-links, and the density. In short, those features are metrics from a graph generated from entities identified in the text, when matched against DBpedia knowledge graph. They refer to, respectively, the similarity of connections with respect to the vertice the number of edges a vertex has to other vertices; the number of links that go from entities of the global DBPedia to the identified entities; and the density of a graph stresses how much nodes are connected to each other.

Negative Features(NF): As expected, the negative features account for the other 47 features. On average, each negative feature increases the accu-

Table 2: First Evaluation

Model	Dataset	Avg. Acc	Avg. F1
BERT*	V4	0.7365	0.734
W2V*	V4	0.6000	0.598
LUX	V4	0.7896	0.768
LUX	V4+T1	0.7603	0.757
LUX	EM	0.7911	0.778
LUX	V4+EM	0.7928	0.767
LUX	**V4+EM+T2**	**0.8050**	**0.804**
LUX	FEVER18	0.6942	0.691
LUX	Snopes19	0.7405	0.517
LUX	V4+EM+T2[#]	0.7723	0.708

**:Only the embeddings were used as input, these results serve as baselines to analyse the improvement added by LUX's linguistic features*

[#]:A version of V4+EM+T2 using the claim (and not the origin body) as input for comparison with other datasets focused on small texts.

[17] github.com/google-research/bert/blob/master/README.md

Table 3: Ablation Results

Feat.Idx	Feature	Avg. Acc	≈Removal Impact
	Most Positive Features		
55	'CD'	0.7864	-1.8%
74	'RBR'	0.7871	-1.7%
71	'PRP'	0.7904	-1.4%
	Most Negative Features		
17	diameter	0.8215	+1.6%
23	nbTypesStd	0.8208	+1.5%
81	'VBD'	0.8201	+1.5%

racy of the model by 0.6% when removed individually. From those, 17 have a better-than-average impact. Avoiding the risk of removing important features from the model and given the high number of negative features, we mention the 9 features that, when **not** considered, improved LUX's accuracy by more than 1%, but focus the discussion on the top 3. Our results point to the number of VBD (verbs, in the past tense form) in the input text as being the third least important feature of the model, while the top two NF are metrics from the same complexity evaluation approach mentioned above. They are nbTypesStd and diameter of nodes, meaning respectively: the standard deviation on the number of different link types per node and the "spreadness"(sic.) of concepts, i.e., the more unrelated and specific concepts we have, the higher the diameter will be. The other six NF improved the accuracy of the model in more than 1% when removed are: the number of words P.O.S.-tagged as PDT (predeterminer), two readability metrics (Dale–Chall and Flesch Reading Ease) and three other features from PySemCom: number of entities, entities density in the text, standard deviation over the number of in-links.

5 Conclusion and Future Work

In this work, the developments to the field were performed in two lines: the consolidation of the VERITAS Dataset, unique due the provision of organic origin for the labeled claims manully assigned by FCAs. Given the completeness of the released data, it can be an useful resource for a number of related tasks, namely: Document Retrieval, Stance Detection and Claim Validation. As a second contribution, we have confirmed the hypothesis that the inclusion of linguistic metrics as model features allows for a better text classification performance, at least in the target task of identifying fake-news.

After having set up an initial version of the classifier, named LUX, we could demonstrate an improvement from its first evaluation by increasing the quality and quantity of the training data, as well removing the most negative features from the model. The final LUX version performs better than both tested baselines. When used to evaluate the quality of datasets, LUX yields better scores when trained with VERITAS, than when compared with two other fake-news datasets, FEVER18 and Snopes19.

As lines of research for our future work, a development of an automatic origin identification step for the VERITAS dataset would allow for a much larger version of it, which in turn could further enhance the classification model (LUX). If this step is achieved, a bootstrapping loop for claim veracity checking with origin identification would be complete, and both the inclusion of new entries to the data collection as well as the further training of classification model could be fully automated, having as their only bottleneck, the permanent scraping of manually fact-checked claims, which is already an automatic process.

Another enhancement being added to this work is the output and analysis of BERT attention weights (Vaswani et al., 2017) for both explainability and interpretability of the model. (Yin et al., 2016; Rush et al., 2015)

Increasing the size of the VERITAS dataset could also be achieved by leveraging the work done by (Hanselowski et al., 2019) and identifying as the origins of a claim, the website containing the snippets annotated as 'supportive' of the claim. This task is currently ongoing.

References

Lucas Azevedo. 2018. Truth or lie: Automatically fact checking news. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 807–811. International World Wide Web Conferences Steering Committee.

Raju Balakrishnan and Subbarao Kambhampati. 2011. Sourcerank: Relevance and trust assessment for deep web sources based on inter-source agreement. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 227–236, New York, NY, USA. ACM.

Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. 2016. " 8 amazing secrets for getting

more clicks": Detecting clickbaits in news streams using article informality. In *Thirtieth AAAI Conference on Artificial Intelligence*.

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):1–4.

Leon Derczynski and Kalina Bontcheva. 2014. Pheme: Veracity in digital social networks. In *UMAP workshops*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

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):695–703.

Arthur C Graesser, Danielle S McNamara, Zhiqang Cai, Mark Conley, Haiying Li, and James Pennebaker. 2014. Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 115(2):210–229.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL2019)*.

Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. *world*.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Benjamin D Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Eleventh International AAAI Conference on Web and Social Media*.

Benjamin D Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Twelfth International AAAI Conference on Web and Social Media*.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Sonnet Ireland. 2018. Fake news alerts: Teaching news literacy skills in a meme world. *The Reference Librarian*, 59(3):122–128.

Scott Jarvis. 2013. Capturing the diversity in lexical diversity. *Language Learning*, 63:87–106.

Anoop K, Manjary Gangan, Deepak P, and Lajish V L. 2019. *Leveraging Heterogeneous Data for Fake News Detection*, pages 229–264.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6610–6617.

Caroline W Lee. 2010. The roots of astroturfing. *Contexts*, 9(1):73–75.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 605–613, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Albert Mehrabian and Morton Wiener. 1966. Nonimmediacy between communicator and object of communication in a verbal message: application to the inference of attitudes. *Journal of Consulting Psychology*, 30(5):420.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Tanushree Mitra and Eric Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *Ninth International AAAI Conference on Web and Social Media*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

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*, pages 1003–1012. International World Wide Web Conferences Steering Committee.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Uwe D Reichel and Piroska Lendvai. 2016. Veracity computing from lexical cues and perceived certainty trends. *arXiv preprint arXiv:1611.02590*.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 25–32.

Victoria L Rubin, Elizabeth D Liddy, and Noriko Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing attitude and affect in text: Theory and applications*, pages 61–76. Springer.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750.

György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Rémi Venant and Mathieu d'Aquin. 2019. Towards the prediction of semantic complexity based on concept graphs.

Veronika Vincze. 2015. *Uncertainty detection in natural language texts*. Ph.D. thesis, szte.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):1–9.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Cynthia Whissell. 2004. Using computer-scored measures of emotion and style to discriminate among disputed and undisputed pauline and non-pauline epistles. *Perceptual and motor skills*, 98(3_suppl):1117–1125.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 34–35.

You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward computational fact-checking. *Proceedings of the VLDB Endowment*, 7(7):589–600.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272.

Hong Yu and Vasileios Hatzivassiloglou. 2003a. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics.

Hong Yu and Vasileios Hatzivassiloglou. 2003b. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 129–136.

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):81–106.