Automatic and Manual Detection of Generated News: Case Study, Limitations and Challenges

Jérémie Bogaert
UCLouvain
Belgium
jeremie.bogaert@uclouvain.be

Antonin Descampe
UCLouvain
Belgium
antonin.descampe@uclouvain.be

Marie-Catherine de Marneffe
The Ohio State University
USA
demarneffe.1@osu.edu

François-Xavier Standaert
UCLouvain
Belgium
fstandae@uclouvain.be

Abstract

In this paper, we study the exploitation of language generation models for disinformation purposes from two viewpoints. Quantitatively, we argue that language models hardly deal with domain adaptation (i.e., the ability to generate text on topics that are not part of a training database, as typically required for news). For this purpose, we show that both simple machine learning models and manual detection can spot machine-generated news in this practically-relevant context. Qualitatively, we put forward the differences between these automatic and manual detection processes, and their potential for a constructive interaction in order to limit the impact of automatic disinformation campaigns. We also discuss the consequences of these findings for the constructive use of natural language generation to produce news items.

Keywords: Disinformation detection technologies.

ACM Reference Format:
Jérémie Bogaert, Antonin Descampe, Marie-Catherine de Marneffe, and François-Xavier Standaert. 2022. Automatic and Manual Detection of Generated News: Case Study, Limitations and Challenges. In *MAD '22: ACM International Workshop on Multimedia AI against Disinformation, June 27–30, 2022, Newark, N*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The spreading of misinformation, disinformation and fake news on social media platforms has been a topic of increasing interest over recent years [15]. Current advances in Natural Language Generation (NLG), based on neural machine learning, make it easy to generate massive amounts of seemingly consistent texts that can subsequently contribute to increase information disorder. There has therefore been work for automatically detecting misleading content under the term of "fake news detection". However, it is important to note that this general research direction covers quite different operational goals (see for example [4, 21, 27] for surveys and [11, 20, 22, 25] for a few technical case studies).

One possible goal is to gauge the veracity of a news item. But this goal faces the (non-technical) challenge of defining what a fake news is [14]. Besides, many different elements can be taken into account to detect whether news are genuine or fake: the source of the information and the sites, people or platforms relaying them, the author's reputation, a fact-checking process confronting the content with an external knowledge base, the style and the lexical field found in the text, etc. Among these elements, some are subjective (e.g., the author's or publishing sites' reputation) and most are based on contextual information not found in the news content itself. As a result, identifying labeled data sets for supervised machine learning in the context of fake news detection is inherently error-prone, which potentially implies label noise [9]. It also makes the reasons for which news are detected as fake using such data sets hard to interpret.

In order to circumvent these difficulties, fake news detection is sometimes restricted to the (better defined) problem of distinguishing machine-generated news from human-generated ones [25]. Such a step back is interesting since it provides a basis for the detection of neural potential fake news, or more generally for the detection of mass disinformation relying on automatic text generation. We follow this line of work and extend it from two perspectives.

First, we study quantitatively the news generated by a state-of-the-art language generation model called `Grover` [25]. While such an evaluation is generally provided in papers introducing new language models (including `Grover`), we extend it in two directions. On the one hand, we observe that

MAD '22, June 27–30, 2022, Newark, NJ
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

news in general, and fake news in particular, are usually targeting recent and fast evolving topics. We therefore build an experiment where classifiers are required to detect whether submitted news are human- vs. machine-generated, for text that is more and more remote from the training database. We consider different classifiers for this purpose (namely logistic regression, support vector machine, naive Bayes, and long short-term memory). On the other hand, we evaluate a manual detection experiment performed over four weeks, with 30 participants, each of them asked to evaluate 10 human- or machine-generated pieces of news per week. Unsurprisingly, our results confirm that domain adaptation is a challenge for language (as for other) models [24]: text generated by Grover for unseen topics is identified as machine-generated both by simple automatic tools and by manual detection. These results thus raise the question of how fast such language generation models can be adapted to newsfeeds, e.g., thanks to emerging plug-and-play language models [6].

Second, we investigate qualitatively the decisions made by the two detection processes (automatic and manual), and the reasons for which news are labeled as human- or machine-generated by the annotators in our manual-detection experiment. We first observe that news items that are incorrectly classified are not systematically the same for the two types of detection. We then leverage the reasons our annotators gave as open answers in our manual-detection experiment to explain their label choice (machine- or human-generated). Practically, we propose a systematization of the reasons in four types (syntax, semantic, background, other), each of them split into four subtypes (detailed in subsection 5.1). Discussing the possibility (or lack thereof) of exploiting these reasons in automatic detection tools, we highlight a potential complementarity between the automatic and manual detection of machine-generated news. We finally conclude the paper by motivating such a combined detection with relevant examples taken from our manual detection experiment.

The rest of the paper is structured as follows. In Sections refsec:design and 3, we describe the two main steps of our experiment, namely the selection of human-generated news and the generation of news with a language model, on topics that are increasingly remote from the training database. In section 4, we detail the quantitative part of our results and confirm the challenge of domain adaptation for language models in front of both automatic and manual detectors. In section 5, we detail the qualitative part of our results and exhibit differences between the automatic and manual detection of machine-generated and human-generated news. We conclude by putting forward the possibility of combined detection as an interesting direction for further research.

2 News selection & generation

In this paper, we aim to evaluate the language generation capacity of the Grover language model for news that are increasingly remote from the training database. In this section, we describe our methodology for generating such news. For this purpose, we first identify some of the main topics covered by the RealNews database on which Grover was trained [25]. We then explain the (more and more challenging) categories of news that we used in our experiment.

2.1 News topics

Grover has been released in 2019 and was trained on the RealNews database. This database is composed of 37 million news which are a subset of CC-NEWS, a larger dataset available via common crawl, updated daily and containing news collected from information websites all over the world.[1] Each piece of news contains multiple fields such as the publication date, the author(s)' name(s), the title and the main text. Grover is able to generate using only some of these fields, like the title and the date. It is available in three versions (base, large and mega). We next use the base version. The motivations for selecting Grover are twofold: first, it is a state-of-the-art language generation model available in open source; second it allows comparison with the NeurIPS 2019 paper of Zellers et al. Yet, other language models could be considered, which we leave as an open problem.

In order to gain some understanding of the RealNews' topics, we applied Latent Dirichlet Allocation (LDA) on a fraction of its articles. LDA is a statistical model that allows clustering by assuming each data point to be a mixture of some unobserved groups [1]. For text data, the components of the mixture are the topics and they are identified thanks to the presence of representative words in each document. We used the pyLDAvis package for this purpose.[2] Figure 1 illustrates the results that we obtained for exemplary clusters and topics. For each cluster (on the left of the figure), LDA outputs a list of associated words (on the right of it).

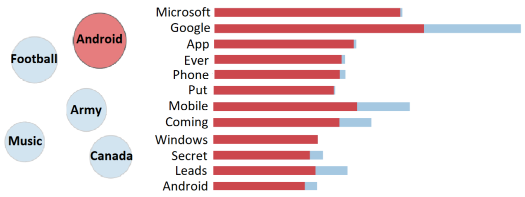

Figure 1. LDA illustration. Exemplary clusters are on the left. Associated words are on the right, with the ratio between their number of appearances in the "Android" cluster (in red) and their number of appearances in any clusters (in blue).

Concretely, we used LDA to identify 25 clusters from which we manually isolated 15 ones such that their associated words were easily and naturally connected with a well identified topic. Namely: *Black Friday, Facebook, Wall*

[1] https://commoncrawl.org.

[2] https://github.com/bmabey/pyLDAvis.

street, Football, American Senate, Air Crash, Olympic Games, Android, Healthcare, Army, Music, Immigration, Christmas, Education and *Canada*. Note that we primarily selected these (timeless) topics because they are quite unrelated to each other. For completeness, we additionally selected 5 event-triggered topics: *Metoo Movement, Notre-Dame Fire, Trump's Election, Ebola* and *Brexit*. As will be discussed in section 4, the timeless or event-triggered nature of the topics does not have a significant impact on our main conclusions.

Besides these 15+5 topics, we also chose topics outside the `RealNews` database. For this purpose, we adopted the even simpler approach of selecting articles of `CC-NEWS` that are subsequent to the ones in `RealNews`. As illustrated in Figure 2, the `RealNews` articles' were all published before 2019 (more precisely April 2019 which is `Grover`'s release date). Therefore, we manually selected 5 other topics that took place after this date: *Joe Biden's Election, Covid19, Beiruth Explosion, Georges Floyd's Death* and *Capitol Invasion*.

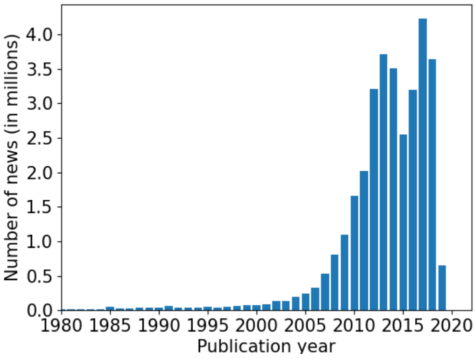

Figure 2. Publication years of `RealNews` articles.

2.2 News categories

The topics' selection described above provided us with the human-generated news for our experiment. We now detail how we produced their machine-generated counterpart. As already mentioned, `Grover` can generate news based on their title and date only, but can additionally exploit other fields. We thus defined our news categories as follows:

- **In RealNews-Full.** Human-generated news from the `RealNews` database. Machine-generated news created using all the fields of the human-generated news. This category focuses on the topics: *Black Friday, Facebook, Wall Street, Football & American Senate.*
- **In RealNews-Title/Date.** Human-generated news from the `RealNews` database. Machine-generated news created using the date/title fields of the human-generated news. This category focuses on the topics: *Air crash, Olympic Games, Android, Healthcare & Army.*
- **Topic in RealNews.** Human-generated news outside the `RealNews` database on topics seen in it. Machine-generated news created using the date/title fields of the human-generated news. Selected topics: *Music, Immigration, Christmas, Education & Canada.*
- **Topic out of RealNews:** Human-generated news outside the `RealNews` database, on topics that are not seen in it. Machine-generated news created using the date/title fields of the human-generated news. Selected topics: *Joe Biden's Election, Covid19, Beiruth Explosion, Georges Floyd's Death & Capitol Invasion.*

The first two categories are aimed to determine whether the addition of fields beyond the title/date of machine-generated news has a significant impact on their quality. The last two ones are increasingly remote from `Grover`'s training database and are aimed to answer our research question: how well can a language model deal with domain adaptation? Coming from `CC-NEWS`, the articles in these categories have the same format as those in the two first categories.

3 Machine-generated news detection

We now move to the description of the automatic and manual detections that we aim to evaluate in our experiment.

3.1 Automatic detection

We trained four machine learning models in order to classify human-generated and machine-generated news:

- **Logistic Regression** (LR) is a staple for classification tasks in general, and it is often used for text classification [18]. It models the relation between a response variable and one or more explanatory variables as a logistic function. Logistic regression is flexible, easy to use and usually leads to interpretable results [13].
- **Support Vector Machines** (SVM) are another widely used algorithm for classification problems [3]. Intuitively, they implement the idea that the inputs to classify can be non-linearly mapped to a high-dimension feature space where a linear decision surface is constructed in order to discriminate pairs of classes [5].
- **Naive Bayes** (NB) is a probabilistic classifier that assumes independence between the features to apply Bayes' Theorem. It is one of the simplest models, but can often achieve surprisingly good results [10].
- **Long Short-Term Memory** (LSTM) is a more complex model that became popular over the last years [16, 26]. In contrast with the previous tools, it is able to deal efficiently with sequentially dependent data (which is usually the case of news) [12]. LSTM is often considered as a baseline deep learning approach.

As usual for text classification, all the news we used for training and testing models were pre-processed using standard techniques. First, we tokenized news into words and assigned part-of-speech to each word. We then removed all the not fully alphabetical words (e.g., numbers) and we lemmatized the remaining ones using the WordNetLemmatizer from the nltk.stem package.[3] We finally vectorized these lemma into numbers. For the LR, SVM and NB models, we used the simple Term Frequency-Inverse Document Frequency (TF-IDF) statistic to produce vectors of words' weights [19]. For the LSTM model, we used the words embedding layer available in the Keras package.[4] The rationale behind this choice is that the LSTM can take advantage of the words' order.

3.2 Manual detection

In order to evaluate the previous automatic detection in front of a manual detection and put forward their potential differences, we set up a manual detection experiment with 30 participants during 4 weeks. At a high level, we asked participants to decide whether pieces of news they read were human-generated or machine-generated and to give a confidence score for their decision. Additionally, an open field allowed them to explain their choices. We built the experiment according to the following partition rules:

- Each participant reviews 10 news items per week.
- Each of these news item is on a different topic.
- The 10 news are balanced: they contain 5 human-generated and 5 machine-generated news.
- Participants never receive both a human-generated news and its machine-generated counterpart.

Participants were not informed of these partition rules.

As usual in statistical evaluations, we designed our experiment to be able controlling different plausible sources of variance that could affect our conclusions. That is, while we are primarily interested in determining whether the (automatic and manual) detection becomes harder with the more and more challenging categories of news defined in the previous section, we also investigated other explanatory variables. Namely, we first analyzed the impact of the type of topics (i.e., timeless or event-triggered) and the weeks (to see whether the participants' detection ability evolves over time). We next analyzed the impact of the title. For this purpose, participants were asked to give a first decision without seeing the news' title and to confirm (or change) it after seeing the title. Finally, we split the participants into groups according to their background (junior engineering, senior engineering, humanities and life sciences). Each group reviewed the same news items, meaning that each item is annotated 3 times (once per group). Participants were not rewarded.

Concretely, this experiment was run on a website that we designed rather than on platforms such as Amazon Mechanical Turk.[5] On the one hand, it allows having a better control of the experiment plan. On the other hand, it limits the amount of answers since the selection of participants and what was requested of them was more constrained.

4 Quantitative analysis

In this section, we present the quantitative results of our experiment. We start by detailing how we selected the parameters of our machine learning models and give one model's learning curves for illustration. We next describe our main result, namely the impact of the news categories on the accuracy of the automatic and manual detections. We note that we use the (easiest to interpret) accuracy metric because of our balanced experiment plan. We finally discuss the impact of the other explanatory variables of our experiment.

Model parameters. In order to avoid overfitting, we built all our models based on the following process. For each news category, we first extracted a validation set of 400 news out of the 2,000 ones available. Within the 1,600 remaining items, we then selected the model parameters by using a 4-fold cross validation (in which all the sets are balanced and contain the same number of human- and machine-generated news). For all models, the best parameters were obtained by performing a grid search (i.e., testing parameter combinations). For the LSTM, we additionally selected the number of epochs such that the training and test losses don't diverge for more than 5 epochs. Having found the best parameters, we finally retrained the model on the 1,600 (training and test) news and evaluated their accuracy on the validation set.

The training curves of the LR model for the different news categories, where we artificially reduce the size of the training set, are in Figure 3. We observe a good convergence and similar results were obtained for the other models. For the last point of the curves, we additionally report the following (Gaussian) confidence interval:

$$\left[\overline{\alpha} - z * \sqrt{\frac{\overline{\alpha} * (1-\overline{\alpha})}{n}}, \overline{\alpha} + z * \sqrt{\frac{\overline{\alpha} * (1-\overline{\alpha})}{n}}\right],$$

where $\overline{\alpha}$ is the estimated accuracy, z is set to 1.96 for a 95% confidence interval and $n = 400$ is the size of the validation set. Similar results were obtained with bootstrap confidence intervals (i.e., without the Gaussian assumption).

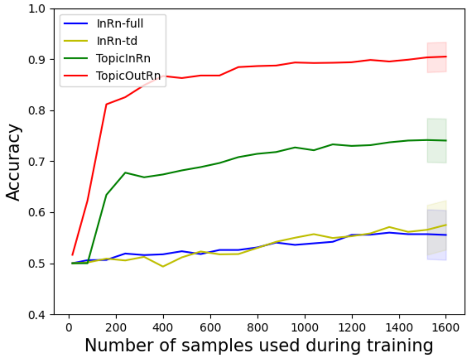

Figure 3. Learning curves of the LR detection tool.

[3] https://www.nltk.org/api/nltk.stem.html.
[4] https://keras.io/api/.
[5] https://www.mturk.com/.

Results. Figure 4 summarizes the results we obtained for our four machine learning models (i.e., the last point of their training curves) and for the manual detections of our experiment reported with the same confidence interval.

As far as the automatic detection is concerned, we observe that all the models lead to similar results and the three main categories of news (i.e., "in RealNews", "Topic in RealNews" and "topics out of RealNews") indeed lead to increasingly easy detections, reflected by a higher accuracy. In particular, the "topic out of RealNews" category leads to detections with an accuracy of approximately 90%. By contrast, the addition of all the fields (besides the title and the date that are always used) to the news generation does not have a significant impact on the detection (see for example "inRn-full" vs. "inRn-td" in Figure 4). The same observation holds for the type of topics, as illustrated in Appendix, Figure 8, where we additionally analyzed the "topic in RealNews" category with both timeless and event-triggered topics.

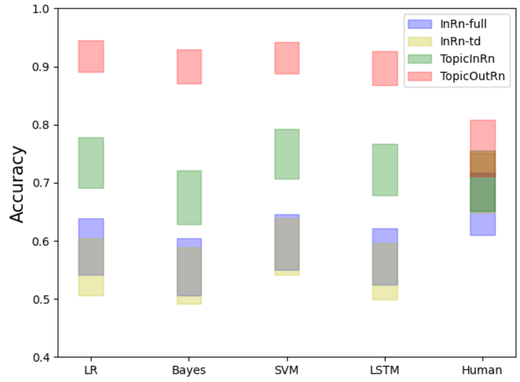

Figure 4. Accuracy of automatic and manual tools to detect human-generated from machine-generated news.

Comparing these accuracies with the ones reported in [25], where Grover is used in order to detect the news it generated, we observe that much simpler models allow detecting machine-generated news when domain adaptation is required (i.e., for the "topic out of RealNews" category). So these results support the claim that detecting machine-generated news on fast evolving topics, that typically require domain adaptation, is feasible even without access to complex (and expensive) models. By contrast, and unsurprisingly, the accuracies we reach for the "in RealNews" and "Topic in RealNews" categories are significantly lower than in [25], presumably due to the simpler models we use.

As for the manual detection, the results obtained are with lower statistical confidence due to the limited amount of samples we could evaluate (100 per category vs. 2,000 per category for automatic detection). Yet, we observe that the accuracy of the "in RealNews" category is confidently lower than the one of the "topics out of RealNews" category. Interestingly, we also observe in Figure 5 that the accuracy of the manual detectors depends on the (real) label of the news to classify, and is significantly higher when the items are human-generated, which is in contrast with the automatic detection, where human-generated and machine-generated news are classified with similar accuracies. So this figure provides a first indication that automatic and manual detection may not exploit the same features of the news, which we will further discuss in the following section.

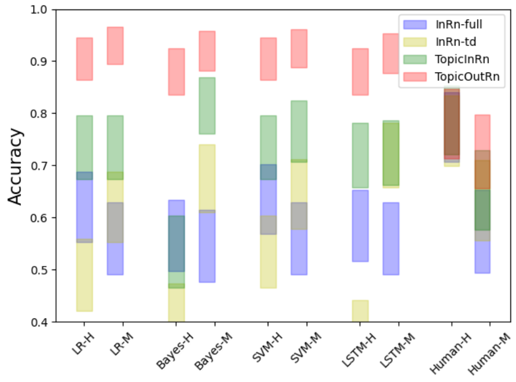

Figure 5. Accuracy of automatic and manual detection tools evaluated on machine- and human-generated news.

Other variables. Besides the dependency of the detection on the main categories, we evaluated the potential dependency between the labels predicted and other potential explanatory variables (i.e., the types of topics, the week, the titles and the groups). None of these variables was found to have a statistically significant impact on the predicted labels with the amount of samples collected in our experiment.

5 Qualitative analysis

In this last section, we finally analyze justifications that participants gave during the manual detection experiment and give some illustrative cases where automatic detection and manual detection lead to different outcomes.

5.1 Justification types and subtypes

We categorized the reasons given by the participants into three main types: *Syntax*, *Semantics* and *Background*, ranging from more local & internal to more global & external.

For the *Syntax* type, we considered 4 sub-types: *Punctuation & Connectors* (PC) for the kind and amount of punctuation & connectors; *Words Repetitions* (WR) for the repetitions of certain words; *Patterns' Repetitions* (PR) for the repetition of syntactic patterns (e.g., sentence constructions); *Grammar & Typos* (GT) for spelling or grammatical mistakes.

For the *Semantics* type, we considered 4 sub-types: *Sentence Meaning* (SM) for meaningless sentences; *Transitions*

(T) for logical transitions between sentences; *Goal & Structure* (GS) for the definition of a main point in the news items and a text structure that is coherent with respect to that main point; *Internal Coherence* (IC) for inconsistencies between pieces of information appearing in the item.

For the *Background* type, we considered 4 sub-types: *External Coherence* (EC) for inconsistencies with respect to the reader's knowledge; *Writing Style* (WS) for the subjective evaluation of the writing skills; *Vocabulary & Expressions* (VE) for the use of particular vocabulary words or typical expressions; *References* (R) for the presence and validity of links, citations, . . . , referred to within the news item.[6]

To these three main types, we add an *Others* type, either for reasons that relate to specific features of our models or when several of the previous types are present in the justification It contains 4 sub-types as well: *Numbers* (N) for the presence of numbers that trigger either an internal (e.g., 3+2 = 6) or external (e.g., 107% of the people) inconsistency; *Subjectivity & Opinion* (SO) for the expression of a viewpoint or opinion in the news; *Believable* (B) for news that in general do not contradict the reader's prior knowledge, whether related to the *Syntax*, *Semantic* or *Background* types; *Comparisons* (C) for justifications made by comparing different news.

Note that these types and sub-types of justifications can be used positively (resp., negatively) to argue in favor of (resp., against) the fact that a news item was machine-generated.

Based on this taxonomy, we represent the distribution of the reasons' types given by the participants during the experiment in Figure 6. While, as in the previous section, we lack samples to turn these histograms into quantitative conclusions, we can nevertheless extract two useful observations. First, this figure shows that the types of reasons used by the participants are not the same for machine-generated and human-generated news. This complements the indication given by Figure 5 that the accuracy of manual detection differs for machine-generated and human-generated news. Second, it shows that the detection of machine-generated news relies more on local types of reasons (i.e., *Syntax* and *Semantics*) while the detection of human-generated news relies more on global ones (i.e., *Background* and *Others*).

The more detailed Figure 7 leads to similar observations while also questioning the possibility to improve automatic detection tools with more advanced machine learning models and the possibility to improve manual detections by combining them with an automated detection assistant.

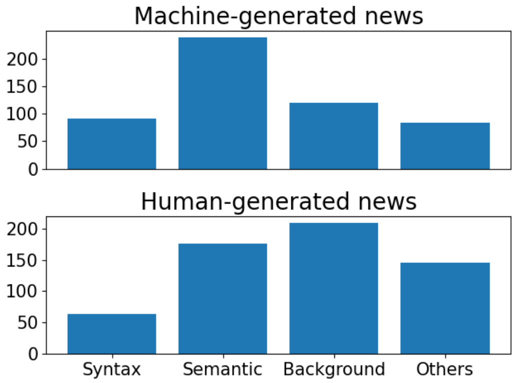

Figure 6. Histogram of the participants' justification types for machine-generated and human-generated news.

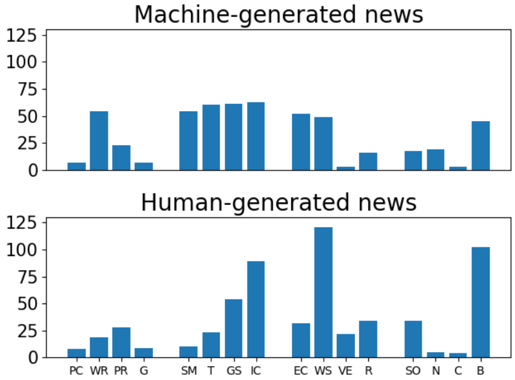

Figure 7. Histogram of the participants' justification subtypes for machine-generated and human-generated news.

Regarding the first possibility, it is worth noticing that among the reasons currently listed by the participants in our manual detection experiment, only a few are exploitable by the simple models we used as automatic detectors. For example, our pre-processing step removes the punctuation, the connectors, all the non fully alphabetical words (e.g., URLs, numbers, . . .) and the inflection forms (cancelling potential grammar mistakes). Furthermore, our three first models (i.e., LR, SVM and NB) rely on a TF-IDF vectorizer, implying that they leverage words' frequencies that are only reflected in some justification sub-types (e.g., *Comparisons*, *Words Repetitions*, *Vocabulary & Expressions*, and to some extent *Writing Style* and *Subjectivity & Opinion* if related to the words frequencies). By being able to capture sequential text features, the LSTM can in theory exploit more justifications sub-types like *Pattern Repetitions*, *Sentence Meaning*, *Transitions* and to some extent *Goal & Structure* and *Internal Coherence*, but would require more training data for this purpose.

More positively, the simplicity of our automatic detection tools emphasizes their potential complementarity with manual detection. That is, none of the participants in our experiment mentioned the analysis of words' frequencies (or other statistical measures that can be extracted by machine learning algorithms) in their justifications. This suggests that (even simple) automatic detectors could be an interesting

[6] WS and VE sub-types are background as they relate to the participants' a priori expectation of how humans and computers generate news.

asset to help manual detection (e.g., by generating warning signals for news that would be worth special attention).

5.2 Manual vs. automatic detection

As an illustration of the potential complementarity between the manual and automatic detection of machine-generated vs. human-generated news, we finally discuss a few examples where these two types of detection lead to contradictory outcomes. In order to identify them, we first listed these cases exhaustively (e.g., in Table 1 and Table 2 where we can see that a higher detection accuracy leads to a higher similarity between manual and automatic detection).[7]

Models	Output =		Output ≠		
	Hg/Hg	Mg/Mg	Hg/Mg	Mg/Hg	Total
LR	28	21	11	40	100
Bayes	30	20	12	38	100
SVM	30	18	14	38	100
LSTM	42	14	18	26	100

Table 1. Agreement between manual and automatic detections for the "In `RealNews`-Full" Category. Hg/Hg: both predict human-generated, Mg/Mg: both predict machine-generated, Hg/Mg: automatic detection predicts human-generated and manual detection predicts machine-generated, Mg/Hg: automatic detection predicts machine-generated and manual detection predicts human-generated.

Models	Output =		Output ≠		
	Hg/Hg	Mg/Mg	Hg/Mg	Mg/Hg	Total
LR	46	35	10	9	100
Bayes	46	32	13	9	100
SVM	25	34	11	30	100
LSTM	38	36	9	17	100

Table 2. Agreement between manual and automatic detections for the "Topic out of `RealNews`" Category.

A first example is given in Appendix B, subsection B.1. This machine-generated news item is correctly classified by all the models but incorrectly classified by all the participants of our experiment. The main reason given for this (incorrect) decision is that the writer's viewpoint was given in some sentences (highlighted in red), which participants considered as typically human. A second example is given in Appendix B, subsection B.2. This machine-generated news is also incorrectly classified by all the participants. Here the reason was that the names of the authors (highlighted in red) appear at the beginning and at the end of the items, which participants considered as a proof of consistency. Interestingly, both examples suggest that the a-priori understanding of the participants about what language models can achieve (or lack thereof) can play a role in their decisions, although this was not detected quantitatively in our experiments. Hence, repeating such experiments and considering clearly defined groups of participants with a limited understanding of language models and groups that are trained to use them, with more samples, would be an interesting follow-up work.

Finally, a third example is given in Appendix B, subsection B.3. This machine-generated news is correctly identified as such by all the participants but incorrectly classified by all the models. This time, participants mostly underlined a lack of goal and structure in the news, giving the impression that the sentences are weirdly connected. As mentioned in subsection 5.1, such semantic reasons are among those that are unlikely to be captured by our simple models, leaving the investigation of more complex models with more training data as another interesting follow-up question.

6 Conclusions

In this paper, we investigated different challenges raised in the news field by the recent improvements of automatic natural language generation based on language models.

First of all, in relation to the quality of machine-generated news and their likelihood to be perceived as human-generated, our experiments confirm the difficulty for language models to deal with domain shifting. Generating a seemingly consistent text on a topic outside those processed during training remains inherently challenging and requires further investigation in domain adaptation techniques like plug-and-play language models [6]. The implications of this observation in the news field are two-fold. On the one hand, the amount of time required for training or domain adaptation makes it less easy to spread harmful disinformation on recent event-triggered topics. On the other hand, for professional newsrooms, it also makes the perspective of using these techniques in a constructive manner quite faraway at this stage. Automatic content production has a great potential in journalism, whether it be for "speeding up production, increasing breadth of coverage, enhancing accuracy or enabling new types of personalization" [8], but without further advances, it will likely remain based on a combination of templates and structured data rather than on language models.

Second, in relation to the detection of machine-generated news, our results indicate a complementarity between the automatic and manual detection methods. Humans and machines do not focus on the same aspects, and our results confirm that computing power should be used for what it does best: scale and speed. Counting words in a huge amount of texts and detecting hidden patterns in the way sentences are built is definitely something computers do better and faster than humans. Therefore, enhancing manual detection based on semantic or background knowledge elements with

[7] Since every news was reviewed by 3 participants (one per group), the manual detection decision in the table is taken as the majority vote.

statistical syntactic indications that are automatically computed appears to be the most efficient way to detect such generated news. In newsrooms, besides finding and spreading news items, debunking and verifying information has always been part of the normative role journalists assign themselves [23]. With the advent of technologies enabling massive amount of credible machine-generated text, this role needs to evolve and take a computational turn. New skills need to enter the newsroom and this is yet another example of how journalism practices are not replaced but displaced by the recent improvements of automation in the production workflow. Whether machines will eventually be able to exploit semantic and background elements just as humans naturally do remains an open question. However, even a positive answer will not empty the role of journalists of what is the very essence of the profession: being accountable for the information disseminated and the editorial decisions that are taken. This hybridation of professional practices with accountability remaining on the human side is in line with the "human still in the loop" perspective recently put forward by Milosavljević and Vobič [17]. It emphasizes a pressing need to embed innovations into established human–machine relations in the news production process, which is amplified by the technical challenges of making algorithms accountable [7]. It also follows Bucher's general observation that algorithms "do not eliminate the need for human judgment and know-how in news work; they displace, redistribute, and shape new ways of being a news worker" [2].

Acknowledgments. François-Xavier Standaert is a senior associate researcher of the Belgian Fund for Scientific Research (FNRS-F.R.S.). This work has been funded in parts by the Walloon Region's ARIAC/TRAIL research project.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.

[2] Taina Bucher. 2018. *If...Then: Algorithmic Power and Politics.* Oxford University Press.

[3] Fabrice Colas and Pavel Brazdil. 2006. Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. In *Artificial Intelligence in Theory and Practice, IFIP 19th World Computer Congress, TC 12: IFIP AI 2006 Stream, August 21-24, 2006, Santiago, Chile (IFIP)*, Max Bramer (Ed.), Vol. 217. Springer, 169–178. https://doi.org/10.1007/978-0-387-34747-9_18

[4] Nadia K. 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.

[5] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Mach. Learn.* 20, 3 (1995), 273–297. https://doi.org/10.1007/BF00994018

[6] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *ICLR*. OpenReview.net.

[7] Antonin Descampe, Clément Massart, Simon Poelman, François-Xavier Standaert, and Olivier Standaert. 2022. Automated news recommendation in front of adversarial examples and the technical limits of transparency in algorithmic accountability. *AI Soc.* 37, 1 (2022), 67–80.

[8] Nicholas Diakopoulos. 2019. *Automating the News: How Algorithms Are Rewriting the Media* (harvard university press ed.).

[9] Benoît Frénay and Michel Verleysen. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Trans. Neural Networks Learn. Syst.* 25, 5 (2014), 845–869.

[10] David J Hand and Keming Yu. 2001. Idiot's Bayes—not so stupid after all? *International statistical review* 69, 3 (2001), 385–398.

[11] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *KDD*. ACM, 1803–1812.

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[13] David W. Hosmer and Stanley Lemeshow. 2000. *Applied Logistic Regression, Second Edition.* Wiley. https://doi.org/10.1002/0471722146

[14] Edson C. Tandoc Jr., Zheng Wei Lim, and Richard Ling. 2018. Defining "Fake News". *Digital Journalism* 6, 2 (2018), 137–153.

[15] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The Science of Fake News. *Science* 359, 6380 (2018), 1094–1096.

[16] Gang Liu and Jiabao Guo. 2019. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* 337 (2019), 325–338. https://doi.org/10.1016/j.neucom.2019.01.078

[17] Marko Milosavljević and Igor Vobič. 2019. Human Still in the Loop. *Digital Journalism* 7, 8 (Sept. 2019), 1098–1116. https://doi.org/10.1080/21670811.2019.1601576

[18] Tomas Pranckevicius and Virginijus Marcinkevicius. 2017. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Balt. J. Mod. Comput.* 5, 2 (2017). https://doi.org/10.22364/bjmc.2017.5.2.05

[19] Juan Ramos. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the First Instructional Conference on Machine Learning*. 29–48.

[20] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *CIKM*. ACM, 797–806.

[21] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.* 10, 3 (2019), 21:1–21:42.

[22] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor.* 19, 1 (2017), 22–36.

[23] Stephen J. A. Ward. 2018. Epistemologies of Journalism. In *Journalism*, Tim P. Vos (Ed.). De Gruyter, 63–82. https://doi.org/10.1515/9781501500084-004

[24] Wen Xu, Jing He, and Yanfeng Shu. 2020. Transfer Learning and Deep Domain Adaptation. In *Advances and Applications in Deep Learning*, Marco Antonio Aceves-Fernandez (Ed.). IntechOpen, Chapter 3.

[25] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *NeurIPS*. 9051–9062.

[26] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A C-LSTM Neural Network for Text Classification. *CoRR* abs/1511.08630 (2015). arXiv:1511.08630 http://arxiv.org/abs/1511.08630

[27] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* 53, 5 (2020), 109:1–109:40.

A Additional figures

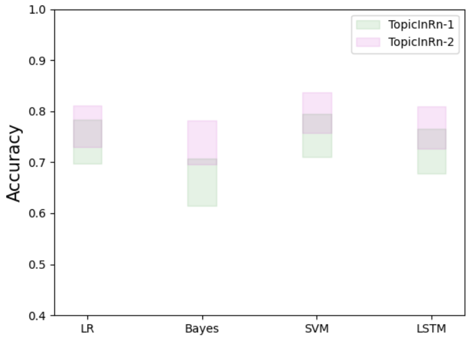

Figure 8. Accuracy of automatic and manual detection tools for timeless and event-triggered types of topics.

B Exemplary news

B.1 First example

Candid Obama admits: healthcare website glitches are a gift to critics

This story was updated at 2:12 p.m.
WASHINGTON - Barack Obama is expected to invoke [...]

If Obama starts off speaking about "infrastructure problems," he may as well be announcing that a healthcare act was the foundation of the federal government's economic success. Because of the website's weakness and the software shutdown, millions of people have been unable to use the healthcare exchanges or their federal employer plans for weeks and months.

As we move deeper into the election, Obama is [...]

That would be a good idea.

B.2 Second example

Beirut explosions create a dilemma for the world

By Marja Sokneski and Olga Rosendahl

MUZAFFARABAD, August 9 (Reuters) - As the Iranian nuclear talks in Rome wind down, Iranian leaders are under increasing pressure to commit more to resolving the Middle East's biggest nuclear crisis. [...]
"Iran was in Syria mainly to ease the crisis. What we need is a durable settlement that all sides can live with." He did not give more details.

(Reporting by Marja Sokneski and Olga Rosendahl; Editing by Catherine Evans)

B.3 Third example

16 feared dead after hot air balloon crashes in Texas

The International Space Station air ambulance has landed at the base of the Texas Air National Guard base, where the disaster happened Tuesday afternoon (July 30).

National Guardsmen [...]

A 14-foot air ambulance was taking part in a training exercise at the base that day that saw a hot air balloon shoot a hole in the grass before crashing into the base's hilly terrain.

Crews on site watched the shooter [...]

a 100-ton oxygen tank stuck above the ground, said Lt. Scott Anderson, spokesman for the station.

"I don't think we lost a single lifter or crew member," Anderson said. [...]

With the air ambulance's parachute, a safe launch position, and the power of the air rescue truck, the rocket typically takes about 10 minutes to complete.
KXAN requested an interview from the air force, as well as a telephone interview with the agency.