HOAXPEDIA: A Unified Wikipedia Hoax Articles Dataset

Hsuvas Borkakoty[1], Luis Espinosa-Anke[1,2]
[1]Cardiff NLP, School of Computer Science and Informatics, Cardiff University, UK
[2]AMPLYFI, UK
{borkakotyh,espinosaankel}@cardiff.ac.uk

Abstract

Hoaxes are a recognised form of disinformation created deliberately, with potential serious implications in the credibility of reference knowledge resources such as Wikipedia. What makes detecting Wikipedia hoaxes hard is that they are often written according to the official style guidelines and would pass as legitimate articles from a written quality standard. In this work, we first confirm the above assumption with a systematic analysis of similarities and discrepancies between legitimate and hoax Wikipedia articles, and introduce HOAXPEDIA, a collection of 311 hoax articles (from existing literature and official Wikipedia lists), together with semantically similar legitimate articles, which together form a binary text classification dataset aimed at fostering research in automated hoax detection. We report results of several models, hoax-to-legit ratios, and the amount of text classifiers are exposed to (full article vs the article's definition alone). Our results suggest that detecting deceitful content in Wikipedia based on content alone is feasible but very hard. We complement our analysis with a study on the distributions in edit histories and find that looking at this feature alone yields better classification results. [1]

1 Introduction

Wikipedia is, as Hovy et al. (2013) define it, the "largest and most popular collaborative and multilingual resource of world and linguistic knowledge", and it is acknowledged that its accuracy is on par with or superior to, e.g., the Encyclopedia Britannica (Giles, 2005). However, as with any other online platform, Wikipedia is also the target of online vandalism, and *hoaxes*, a more obscure, less obvious form of vandalism[2], constitute a significant threat to Wikipedia's overall integrity (Kumar et al., 2016; Wong et al., 2021; Wang and McKeown, 2010), among others, because of its "publish first, ask questions later" policy (Asthana and Halfaker, 2018). Although Wikipedia employs community based New Page Patrol systems to check the credibility of a newly created article, the process is always in backlog[3], making it overwhelming (Schneider et al., 2014).

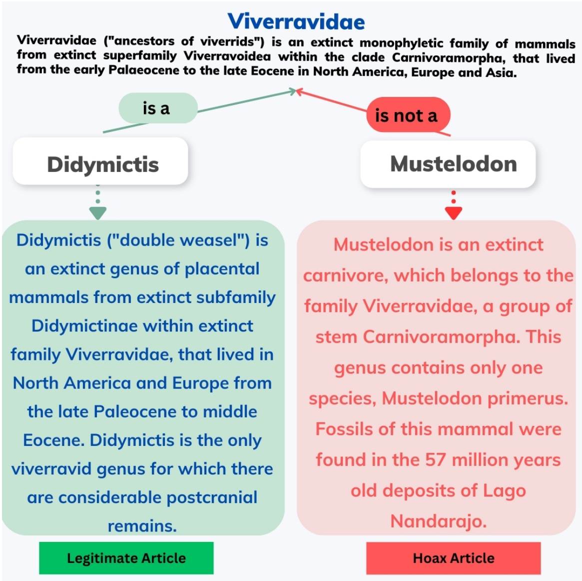


Figure 1: An example of the nature of the Hoaxpedia dataset. It contains hoax (red) articles as well as semantically similar legitimate articles (green), which pose a hard problem for a text-based classifier due to their textual similarities.

Hoax articles (as shown in Figure 1), are created to deliberately spread false information (Kumar et al., 2016), harm the credibility of Wikipedia as a knowledge resource and generate concerns

[1]The Dataset is available at: `https://huggingface.co/datasets/hsuvaskakoty/hoaxpedia` and associated codes are available at: `https://github.com/hsuvas/hoaxpedia_dataset.git`

[2]`https://en.wikipedia.org/wiki/Wikipedia:Do_not_create_hoaxes`.

[3]`https://en.wikipedia.org/wiki/Wikipedia:New_pages_patrol`.

Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia, pages 53–66
November 16, 2024 ©2024 Association for Computational Linguistics

among its users (Hu et al., 2007). Since manual inspection of quality is typically a lagging process (Dang and Ignat, 2016), the automatic detection of such articles is highly desirable. However, most works in the literature have centered their efforts on the metadata associated with hoax articles, e.g., user activity, appearance features or revision history (Zeng et al., 2006; Elebiary and Ciampaglia, 2023; Kumar et al., 2016; Wong et al., 2021; Hu et al., 2007; Susuri et al., 2017). For example, Adler et al. (2011) introduced a vandalism detection system using metadata, content and author reputation features, whereas Kumar et al. (2016) provide a comprehensive study of hoax articles and their timeline from discovery to deletion. In their work, the authors define the characteristics of a successful hoax, with a data-driven approach based on studying a dataset of 64 articles (both hoax and legitimate), on top of which they train statistical classifiers. Furthermore, other works have compared network traffic and features of hoax articles to those of other articles published the same day (Elebiary and Ciampaglia, 2023), and conclude that hoax articles attract more attention after creation than *cohort* (or legitimate) articles. Finally, Wong et al. (2021) study various Wikipedia vandalism types and introduce the Wiki-Reliability dataset, which comprises articles based on 41 author-compiled templates. This dataset contains 1,300 articles marked as hoax, which are legitimate articles with false information, a.k.a hoax facts (Kumar et al., 2016).

In this paper, we propose to study hoax detection only by looking at textual content. If successful, this would have obvious advantages in the transferrability of models to other platforms. To this end, we first construct a dataset (HOAXPEDIA) containing 311 hoax articles and around 30,000 *plausible negative examples*, i.e., legitimate Wikipedia articles that are semantically similar to hoax articles, so that the set of distractors *covers similar topics* (since similarity in style is assumed) to hoax articles (e.g., a newly discovered species). We also explore whether a Wikipedia definition (the first sentence of the article) can provide any kind of hints towards its veracity. Our results (reported at different ratios of hoax vs. legitimate articles) suggest that style and shallow features are certainly not the best predictors, but combining language models (LMs) with metadata features (e.g., an article's revision history) is a promising direction. Our contributions in this work can be summarised as follows.

- We systematically contrast a set of proven Wikipedia hoax articles with legitimate articles.
- We propose HOAXPEDIA, a novel Wikipedia Hoax article dataset with 311 hoax articles and 30,000 semantically similar legitimate articles collected from Wikipedia.
- We conduct binary classification experiments on HoaxPedia, using a range of language models (including LLMs), features, and hoax-to-legitimate ratio.

2 Related work

In what follows, we give a brief overview of disinformation detection, the datasets available for the community and the role of Wikipedia in disinformation detection, as our work falls in the intersection between disinformation detection and Wikipedia research.

Disinformation detection and datasets: Disinformation and misinformation are two types of false information, they differ in that misinformation is inaccurate information created or propagated unknowingly, whereas disinformation is inaccurate information deliberately created to mislead the intended consumer (Hernon, 1995; Fallis, 2014; Kumar et al., 2016; Ireton and Posetti, 2018). Nonetheless, both are harmful to information quality and reliability, thus posing risks toward different aspects of society (Su et al., 2020). Alam et al. (2021) survey disinformation detection from a multi-modal perspective (specifically, text, images, audio, and video), with text being the most common. Datasets used for disinformation detection can be divided based on the length of input or claim: short sentences (such as tweets or Reddit posts) vs articles (common type being news articles), where most of the datasets follow claim-evidence based format (Su et al., 2020). The short sentences or claim based datasets are mostly sourced from social media, such as X (formerly Twitter) (Castillo et al., 2011; Derczynski et al., 2017; Zubiaga et al., 2018; López and Madhyastha, 2021), Reddit (Gorrell et al., 2018; Qu et al., 2022), or fact checking websites like Politifact[4] (Wang, 2017), Snopes[5] (Vo and Lee, 2020), or a combination of different fact checking websites (Augenstein

[4] https://www.politifact.com/
[5] https://www.snopes.com/

et al., 2019). These datasets usually contain claims, verification labels and evidences to back the label. Article level datasets, on the other hand, are varied, and focus on state-backed propaganda (Heppell et al., 2023), German multi-label disinformation (the GerDISDETECT dataset) (Schütz et al., 2024), or narratives at conflict dataset containing news articles (Sinelnik and Hovy, 2024), which mostly focuses on news article or propaganda based disinformation spreading. The datasets mentioned above are specialized towards topic/trend based or news based disinformation, with no specialization on Wikipedia.

Wikipedia in disinformation detection: Wikipedia, as described by McDowell and Vetter (2020), serves as a source of information validation as backed by its large set of articles contributed by community. This is seen in action for fact verification task datasets such as FEVER (Thorne et al., 2018b), TabFactA (Chen et al., 2019), or the FNC-1 (Fake News Challenge-1) dataset (Pomerleau and Rao, 2017). Here, evidences for claims are collected from Wikipedia articles (eg. FEVER, FNC-1) and tables (eg. TabFactA). However, being a product of community effort, Wikipedia is also prone to vandalism and inaccurate contents (McDowell and Vetter, 2020), and the community outlines different policies to combat these issues[6]. We also find efforts to automatize the process of detecting vandalism contents from Natural Language Processing perspective. Previously, feature based approaches extracted from metadata and editor behaviour were used to detect vandalism (Wu et al., 2010; Javanmardi et al., 2011; Heindorf et al., 2016). Implementation of early warning systems based on metadata and editor behavior is found in the work of Kumar et al. (2015), where they propose a dataset of page metadata and a set of autoencoder-based classifiers. Yuan et al. (2017) propose an edit history based approach, where they use behaviour of users over time as feature to create the embedding space for multi-source LSTM networks (Hochreiter and Schmidhuber, 1997). Additionally, real-time machine learning based Wikipedia edit scoring system named ORES (Halfaker and Geiger, 2020), and multilingual vandalism detection system (Trokhymovych et al., 2023) contributes to a high-end edit based vandalism detection systems that are deployed in Wikipedia. However, these approaches do not consider article text as a marker to detect vandalism.

While Wikipedia marks hoax articles as form of vandalism (Thorne et al., 2018a), we argue that the vandalism and hoax detection fields have not yet met - although there are notable exceptions (Kumar et al., 2016; Wong et al., 2021), and thus our work aims to establish a stronger tie between them with a single dataset unifying existing work in addition to gathering any available proven hoax article from additional sources.

Data Source	Data points
Kumar et al. (2016)	64
Elebiary and Ciampaglia (2023)	95
Wikipedia List of Hoaxes	
Collected from Wikipedia	87
Collected from Internet archive	65
Total	311

Table 1: Data sources used to construct HOAXPEDIA and their corresponding number of data points from each source.

3 HOAXPEDIA Construction

HOAXPEDIA is constructed by unifying five different resources that contain known hoaxes, e.g., from Kumar et al. (2016); Elebiary and Ciampaglia (2023), as well as from the URLs available in the official Wikipedia hoaxes list[7] and the Internet Archive. Articles extracted from the Internet Archive are the ones that are deleted from Wikipedia but are redirected from the list of Hoaxes as 'Archived version' to the Internet Archive[8]. The statistics of the articles collected from different sources are given in Table 1. We manually verify each of the articles we collect from Wikipedia and Internet Archive as a hoax using their accompanied deletion discussion and reasons for citing them as a hoax.

In terms of negative examples, while we could have randomly sampled Wikipedia pages, this could have introduced a number of biases in the dataset, e.g., hoax articles contain historical events, personalities or artifacts, and thus we are interested in capturing a similar breadth of topics, entities and

[6] https://en.wikipedia.org/wiki/Wikipedia:Vandalism

[7] https://en.wikipedia.org/wiki/Wikipedia:List_of_hoaxes_on_Wikipedia

[8] Example archived article: https://web.archive.org/web/20230608103922/https://en.wikipedia.org/wiki/Rainbow_fish_%28mythology%29

sectors in the negative examples so that a classifier cannot use "shortcuts" for effective classification. These negative examples correspond to authentic content. This is achieved by verifying they do not carry the Db-hoax flag, which Wikipedia's New Page Patrol policy uses to mark potential hoaxes. Within this set, we extract negative examples as follows. Let H be the set of hoax articles, and W the set of candidate *legitimate* Wikipedia pages, with $T_H = \{t_{H^1}, \ldots, t_{H^p}\}$ and $T_W = \{t_{W^1}, \ldots, t_{W^q}\}$ their corresponding vector representations, and p and q the number of hoax and candidate Wikipedia articles, respectively. Then, for each SBERT (all-MiniLM-L6-v2) (Reimers and Gurevych, 2019) title embedding $t_{H^i} \in T_H$, we retrieve its top k nearest neighbors (NN) from T_W via cosine similarity $\cos$. We experiment with different values for k, specifically $k \in \{2, 10, 100\}$:

$$\text{NN}\,(t_{H^i}) = \{t_{W^j} : j \in J_k(t_{H^i})\}$$

where $J_k(t_{H^i})$ contains the top k cosine similarities in T_W for a given t_{H^i}, and

$$\cos\,(t_{H^i}, t_{W^j}) = \frac{t_{H^i} \cdot t_{W^j}}{||t_{H^i}||||t_{W^j}||}$$

The result of this process is a set of positive (hoax) articles and a set of negative examples, which we argue is similar in both style and topic, effectively removing topic bias from the dataset.

4 Text Based Analysis on HOAXPEDIA

For a better understanding of article structure, and leverage the text and its features to distinguish between hoax and legitimate articles, we run different analysis in surface level and designing classifiers to identify hoax articles. We do not consider metadata that comes along with the Wikipedia articles, as metadata are platform-specific, which we argue can have a negative impact on transferability.

4.1 Hoax vs. Legitimate, a Surface-Level Comparison

To maintain longevity and avoid detection, hoax articles follow Wikipedia guidelines and article structure. This raises the following question: *"how (dis)similar are hoaxes with respect to a hypothetical legitimate counterpart?"*. Upon inspection, we found comments in the deletion discussions such as *"I wouldn't have questioned it had I come across it organically"* (for the hoax article *The Heat is On* [9]), or *"The story may have a "credible feel" to it, but it lacks any sources"*, a comment on article *Chu Chi Zui*[10]. Comments like these highlight that hoaxes are generally well written (following Wikipedia's guidelines), and so we proceed to quantify their stylistic differences in a comparative analysis that looks at: (1) article text length; (2) sentence and word length; and (3) a readability metrics.

Article Text length distribution: Following the works of Kumar et al. (2016), we conduct a text length distribution analysis with hoax and legitimate articles, and verify they show a similar pattern (as shown in Figure 2), with similar medians for hoax and legitimate articles, specifically 1,057 and 1,777 words, respectively.

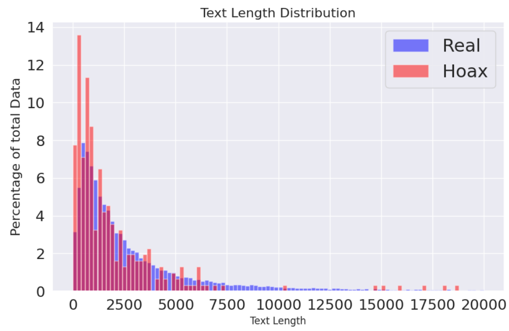


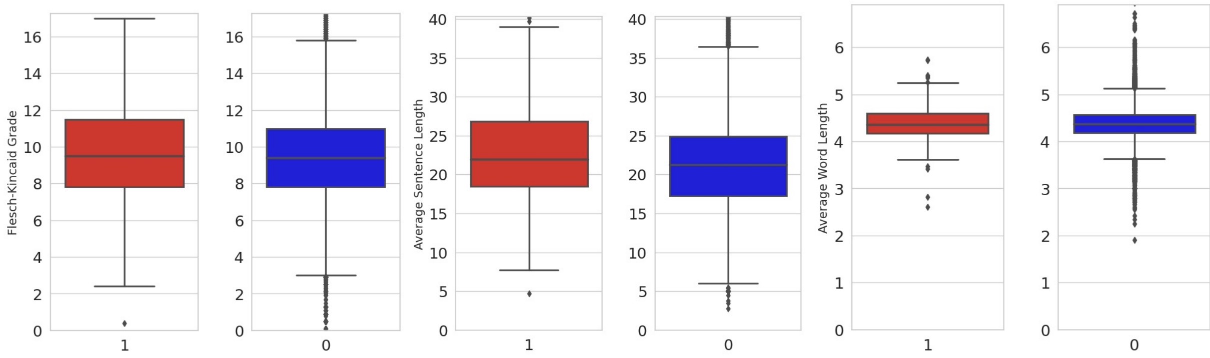
Figure 2: Text length distribution for hoax and legitimate articles (with percentage of data points shown in y-axis).

Average sentence and word length: Calculating average sentence and word length for hoaxes and legitimate articles separately can be a valuable proxy for identifying any obvious stylistic or linguistic (e.g., syntactic complexity) patterns. We visualize these in a series of box plots in Figure 3. They clearly show a similar style, with sentence and word length medians at 21.23 and 22.0, and 4.36 and 4.35 for legitimate and hoax articles respectively.

Readability Analysis: Readability analysis gives a quantifiable measure of the complexities in text, revealing distinguishable patterns for disguising disinformation through hoaxes or convey clear, factual content. For readability analysis, we use the Flesch-Kincaid (FK) Grading system (Flesch,

[9] https://en.wikipedia.org/wiki/Wikipedia:Articles_for_deletion/The_Heat_Is_On_(TV_series)

[10] https://en.wikipedia.org/wiki/Wikipedia:Articles_for_deletion/Chu_Chi_Zui



(a) Flesch-Kincaid Grade (b) Average Sentence Length (c) Average Word Length

Figure 3: Results of different stylistic analyses on Hoax (red) and legitimate (blue) articles.

2007), a metric that indicates comprehension difficulty when reading a passage in the context of contemporary academic English. After obtaining an average for both hoax and legitimate articles, we visualize these averages again in Figure 3, we find a median of 9.4 for legitimate articles and 9.5 for hoax articles, which again highlights the similarities between these articles.

4.2 Classification Experiments

We cast the problem of identifying hoax vs. legitimate articles as a binary classification problem. Our experiments are aimed to explore the impact of data imbalance and content length, and we evaluate a suite of pre-trained LMs as well as a set of open sourced LLMs. We split the dataset into non overlapping train and test (with 80:20 ratio for positive instances for definition and fulltext settings), due to the smaller number of positive instances (311), as well as for the fact that we want to test the models for their abilities on unseen test data. The experimental settings and results are discussed below.

4.2.1 Pre-trained Language Models

We evaluated the BERT family of models (BERT base and large (Devlin et al., 2019), RoBERTa-base and large (Liu et al., 2019), Albert-base and large (Lan et al., 2019)), as well as T5 (Base and Large) (Raffel et al., 2020) and Longformer (Base) (Beltagy et al., 2020) with the same training configuration (as mentioned in Appendix B) and generation objective as *Binary classification* for T5 models. In terms of data size, we consider the three different scenarios outlined in Section 3 (2x, 10x and 100x negative examples). This approach naturally increases the challenge for the classifiers. The details about the data used in different settings are given in Appendix A.

In addition to the three different settings for positive vs. negative ratios, we also explore *how much text is actually needed to catch a hoax*, or, in other words, *are definition sentences in hoax articles giving something away*? This is explored by running our experiments on the full Wikipedia articles, on one hand, and on the definition (first sentence alone), on the other. This latter setting is interesting from a lexicographic perspective because it helps us understand if the Wikipedia definitions show any pattern that a model could exploit. Moreover, from the practical point of view of building a classifier that could dynamically *"patrol"* Wikipedia and flag content automatically, a definition-only model would be more interpretable (with reduced ambiguity and focusing on core meaning/properties of the entity) and could have less parameters (handling smaller vocabularies, and compressed knowledge), which would have practical retraining/deployment implications in cost and turnaround.

We compare several classifiers and analyze whether model size (in number of parameters) is correlated with performance of data imbalance and content length scenarios, reporting the results in F1 on positive class (hoax). In definition only setting, we find that models evaluated on datasets that are relatively balanced (2 real articles for every hoax) show a stable performance, but they degrade drastically as the imbalance increases. RoBERTa proves to be most consistent, with an F1 of around 0.6 for all three settings, whereas Albert models perform poorly (with some interesting behavior discussed later). For the full text setting, we find that Longformer models performs well, with an F1 of 0.8. Surprisingly, the largest model we evaluated (T5-large) is not the best performing model, although this could point to underfitting (dataset being small for model this size). Another interesting behavior of T5-large is that in the 1H2R data split, performance on definition and full text setting are the same. On the other side, we find that Albert models are the ones showing the highest improvement when going from definition to full text. This is interesting, as it shows a small model may miss nuances in definitions but can still compete with, or even outperform, larger models.

A perhaps not too surprising observation is that all models improve after being exposed to more text, as seen in Table 2, increasing their F1 by about 20% on average and sometimes even up to 30%. This confirms that definitions alone are not a sufficiently strong signal for detecting hoax articles, although there are notable exceptions. Moreover, in terms of absolute performance, the RoBERTa models perform decently, although significantly below their full-text settings. It is interesting to note that the Longformer base yields much better results in the 1H100R split when exposed only to definitions. This is indeed a surprising and counterintuitive result that deserves future investigation.

Effect of Definitions on Classifying Hoaxes We also test the importance of definition sentences in the full text setting though removing the definition sentence from each row and running classification on RoBERTa-Large, the most consistent model in our experiments. The results shown in Table 3, suggest that F1 decreases about 2% for the positive class when the definition sentence is missing. This shows that definitions show critical information about entities and events in Wikipedia, but often are not the place where hoax features would emerge, and therefore removing them from the full text does not change much of the story.

4.2.2 Large Language Models

We explore the capabilities of open-source Large Language Models (LLM) to detect hoax articles through our proposed dataset. We select Llama2-7B and 13B (Touvron et al., 2023), Llama3-8B (Dubey et al., 2024), and Mistral-7B models (Jiang et al., 2023) for the experiments, and the prompts used are given in the Appendix C. We consider prompt-based tuning and supervised fine-tuning (Touvron et al., 2023) as our experiment settings.

Prompting: For prompting, we consider zero-shot and few-shot prompts, as given in Appendix C, and the input setting are for both definition and fulltext. We report the results for F1-scores on positive class in Table 4. The results show that Llama2-13B models perform the best for both settings (definition and fulltext). Notably, performance difference between the definition and fulltext setting is marginal, as opposed to fine-tuned LMs in Table 2.

Fine-tuning: We fine-tune the LLMs with HOAXPEDIA in supervised fine-tuning (Touvron et al., 2023) paradigm. The results of fine-tuning as F1-scores for both definition and fulltext setting are shown in Table 5, with significant improvement across all the settings for all the models. Llama3 shows most consistency and is the best model across the scenarios, with a performance improvement of more than 25%.

4.2.3 Perplexity Experiments with LLMs

We consider perplexity as an indicator for LLMs to predict the distribution of Hoax and legitimate articles, with the hypothesis that LLMs will have difficulty predicting the contents of hoax articles, resulting in higher perplexity. We test the LLMs in both definition and fulltext settings. The average perplexity results for both settings are shown in Figure 4, revealing that there is a significant difference between the perplexity of hoax and legitimate articles in both settings. This suggests that LLMs struggle to predict the distribution of Hoax articles.

5 Comparing Revision Activities of Hoax and Legitimate Articles

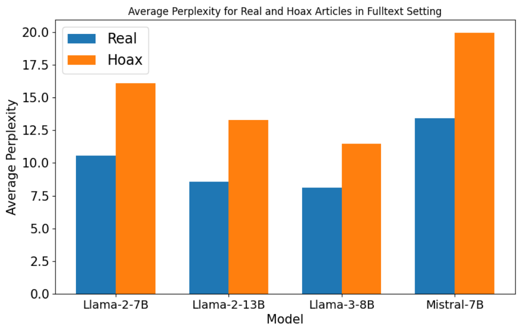
Analysing the revision timelines of hoaxes and legitimate articles can reveal valuable insights into activity patterns on those articles from the Wikipedia community. We investigate the revision activity patterns by collecting timelines of hoax and legitimate articles (in all three hoax-to-legitimate ratios mentioned above) and add these timelines to HOAXPEDIA. However, since some of the hoax articles were deleted from Wikipedia at the time of this experiment, we were only able to obtain 164 hoax articles out of 311 in our dataset. We explore the revision history timelines of legitimate and hoax articles through changepoints and dense regions in timelines and experiment with the binary classification problem of identifying hoax articles through their timelines.

5.1 Exploratory Analysis

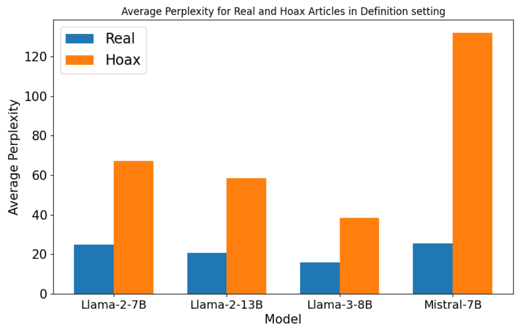
We analyze timeline patterns through the use of a dense region identification algorithm, namely Bayesian Online Changepoint Detection (BOCPD) (Adams and MacKay, 2007), followed by Kernel

Model	Model Size	Definition			Fulltext		
		1H2R	**1H10R**	**1H100R**	**1H2R**	**1H10R**	**1H100R**
Albert-base-v2	12M	0.23	0.17	0.06	0.67	0.47	0.11
Albert-large-v2	18M	0.28	0.30	0.15	0.72	0.63	0.30
BERT-base	110M	0.42	0.30	0.14	0.55	0.57	0.32
RoBERTa Base	123M	0.57	0.59	0.53	0.82	0.75	0.63
Longformer-base	149M	0.43	0.35	0.54	0.80	0.78	0.67
T5-Base	220M	0.48	0.25	0.14	0.51	0.27	0.23
BERT-large	340M	0.43	0.36	0.17	0.61	0.64	0.33
RoBERTa-large	354M	0.58	0.63	0.62	0.84	0.81	0.79
T5-large	770M	0.54	0.32	0.13	0.54	0.43	0.37

Table 2: F1 on the positive class - *hoax* at different degrees of data imbalance for definition-only and fulltext setup (H: Hoax, R: Real).



(a) Average perplexity scores for LLMs in the fulltext setup.



(b) Average perplexity scores for LLMs in the definition setup.

Figure 4: Average perplexity scores in fulltext and definition only setups for legitimate (real) and hoax articles.

Model	Setting	Precision	Recall	F1
RoBERTaL	1H2R	0.83	0.80	0.82
RoBERTaL	1H10R	0.82	0.71	0.76
RoBERTaL	1H100R	0.67	0.51	0.58

Table 3: Performance of RoBERTa-Large on binary classification without definition sentences in articles (with hoax to real ratio for fulltext setup in Settings column) on positive class - *hoax* (H: Hoax, R: Real).

Model Name	Zero-shot		Few-shot	
	Definition	**Fulltext**	**Definition**	**Fulltext**
Llama2-7B	0.48	0.50	0.51	0.52
Llama2-13B	0.57	0.58	0.59	0.59
Llama3-8B	0.33	0.40	0.35	0.40
Mistral-7B	0.53	0.56	0.54	0.58

Table 4: F1 score on positive class - *hoax* for prompting experiment in zero and few shot setting for definition-only and fulltext setup.

Model	Definition			Fulltext		
	1H2R	**1H10R**	**1H100R**	**1H2R**	**1H10R**	**1H100R**
Llama2-7B	0.76	0.47	0.49	0.66	0.48	0.47
Llama2-13B	0.80	0.48	0.50	0.60	0.63	0.50
Llama3-8B	0.80	0.48	0.50	0.83	0.67	0.50
Mistral-7B	0.71	0.55	0.49	0.68	0.53	0.49

Table 5: F1 score for LLM fine-tuning in degrees of data imbalance for definition-only and fulltext setup (H: Hoax, R: Real).

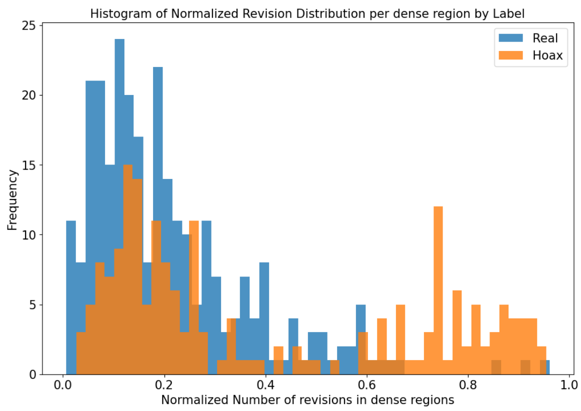


Figure 5: Histogram of normalized distribution for number of revisions in dense regions for hoax and legitimate (real) article.

Density Estimation (KDE) (Węglarczyk, 2018), with which we obtain dense regions, which are significantly active periods in a page's revision period in comparison with the overall distribution. Figure 6 shows a comparison of two timelines with highlighted dense regions. We can see that the number of revisions are generally low for hoax articles, and that their dense regions are mostly around the beginning and end of the article's timeline. This can be attributed to New Page Patrol (NPP) for spike in the beginning and detection with deletion discussion for the end. To quantify this evidence, we divide the revision timelines of hoax and legitimate articles into quartiles and compute a normalized count of dense regions. The result for each quartile is given in Table 6, and clearly shows that the proportion of dense regions happening at the beginning and at the end are higher (especially close to the end of the article's life) for hoax articles than for legitimate ones. We also show in a histogram the normalized distribution of hoax and legitimate (real) revisions in Figure 5, which provides a full-picture summary of these edits. The distribution shown here is the density of revisions for hoax and legitimate articles with respect to the frequency of articles in that density. Based on this analysis, we further find that legitimate articles have 5.40x more revisions on average (81.70 for legitimate vs. 15.11 for hoax), but if we look at the relative density of each revision, hoax articles undergo more activity per region (0.21 for legitimate articles vs. 0.39 of hoax articles), which suggests that for the hoax articles, there is a "disproportionate hyperfocus" of the community at very concrete points in the lifespan of the article.

Quartile	Hoax	Real
Q1	0.69	0.75
Q2	0.02	0.17
Q3	0.04	0.22
Q4	0.75	0.42

Table 6: Average distribution of dense regions per quartile (timeline divided into four parts) for hoax and legitimate (real) articles.

5.2 Revision History based Classification

We formulate the detection of hoaxes as a binary classification problem with features collected from article revision histories (each containing a series of timestamps) for hoax and legitimate articles. To create the feature vector, we group those timestamps by month and year (MM-YYYY) to create the vocabulary[11] for our model. We use this vocabulary to obtain the TF-IDF features (Sparck Jones, 1972). We train a Support Vector Machine (SVM) (Vapnik, 2013) classifier with the TF-IDF features. We report F1 scores for the positive class in Table 7, with good performance (0.88 for the 1H2R setting) of the SVM classifier, although the performance decreases due to the data imbalance. This further proves that the revision history can be an important feature in the detection of hoaxes. However, we also argue that timeline alone may not be enough, as it is a statistical feature prone to outliers. Moreover, hoaxes are defined based on it's contents, thus we encourage the importance of content as the important feature for hoax article detection.

Data Split	Precision	Recall	F1
1H2R	0.86	0.91	0.88
1H10R	0.89	0.78	0.83
1H100R	0.97	0.69	0.80

Table 7: Results of SVM timeline classifier for label 1 (Hoax) for all data splits.

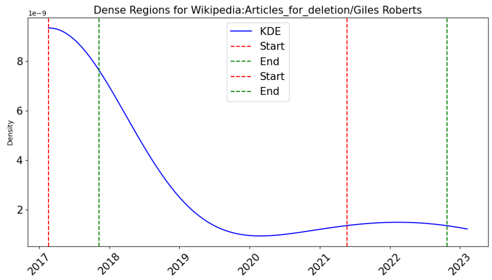
6 Conclusion and Future Work

We have introduced HOAXPEDIA, a dataset containing hoax articles extracted from Wikipedia, from a number of sources, from official lists of hoaxes, existing datasets, and the Web Archive. We paired these hoax articles with similar legitimate articles, and after analyzing their main properties (concluding they are written with very similar style and content), we report the results of a number of binary classification experiments, where we explore the impact of (1) positive to negative ratio; and (2) going from the whole article to only the definition. This is different from previous work in that we have exclusively looked at the content of these hoax articles, rather than metadata such as traffic or longevity. For the future, we would like to explore the approaches (Arora et al., 2024; Field et al., 2022) to reduce spurious artifacts that might appear during the creation of the dataset to strengthen the dataset. Additionally, utilizing approaches for building Wikipedia corpus controlling for topic or readability (Johnson et al., 2021; Trokhymovych et al., 2024) can improve the overall quality of the dataset. We would also like to further refine what the criteria are used by Wikipedia editors to detect hoax articles, turn those insights into a ML model, and explore other types of non-obvious online vandalism.

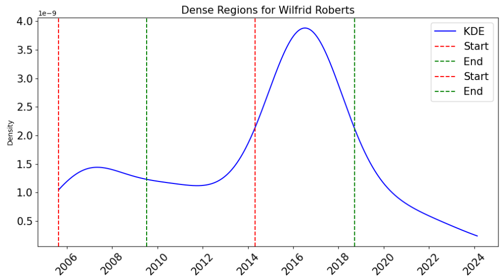
7 Limitations

We present a new dataset named HOAXPEDIA and associated baselines from a wide variety of language models / large language models. Our study shows that these types of dataset can be helpful

[11]Appendix D explains the process of creating a vocabulary from the revision history



(a) Revision history Plot for an example Hoax article.



(b) Revision history plot for an example legitimate article.

Figure 6: Revision history based dense region plots for hoax and legitimate articles with dense regions marked with dotted lines.

in the area of free text disinformation detection. However, there are some limitations to our work that we aim to address here. The sets proposed here are small, with only 311 positive examples (hoaxes), which can be attributed to the fact that we only collect the examples that are explicitly labeled as hoaxes, rather than articles under discussion for hoaxes. Additionally, in our experiments, we do not conduct further investigation for model behaviors such as performance improvement of Longformer models in the hardest setting. We leave these analysis in future work, as the scope of this work is to introduce this dataset and establish the baseline results with pre-trained LMs and LLMs. Finally, we do not compare the results with existing work, mainly with (Kumar et al., 2016), since the approaches mentioned in existing work are metadata dependent with different sets of features/approaches in consideration, and our approach is based on article text, we argue that the results may not be comparable. We also acknowledge that Wikipedia is a multilingual effort, and our dataset only contains data from Wikipedia in the English language, which can be a major limitation in multilingual landscape. We keep the multilingual extension of the hoax dataset as one of the future work.

8 Ethics Statement

This paper is in the area of online vandalism and disinformation detection, hence a sensitive topic. All data and code will be made publicly available to contribute to the advancement of the field. However, we acknowledge that deceitful content can be also used with malicious intents, and we will make it clear in any associated documentation that any dataset or model released as a result of this paper should be used for ensuring a more transparent and trustworthy Internet.

References

Ryan Prescott Adams and David JC MacKay. 2007. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

B. Thomas Adler, Luca de Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Computational Linguistics and Intelligent Text Processing*, pages 277–288, Berlin, Heidelberg. Springer Berlin Heidelberg.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021. A survey on multimodal disinformation detection. *arXiv preprint arXiv:2103.12541*.

Akhil Arora, Robert West, and Martin Gerlach. 2024. Orphan articles: The dark matter of wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 100–112.

Sumit Asthana and Aaron Halfaker. 2018. With few eyes, all hoaxes are deep. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 675–684, New York, NY, USA. Association for Computing Machinery.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Quang Vinh Dang and Claudia-Lavinia Ignat. 2016. Quality assessment of wikipedia articles without feature engineering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, page 27–30, New York, NY, USA. Association for Computing Machinery.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Anis Elebiary and Giovanni Luca Ciampaglia. 2023. The role of online attention in the supply of disinformation in wikipedia. *arXiv preprint arXiv:2302.08576*.

Don Fallis. 2014. A functional analysis of disinformation. *IConference 2014 Proceedings*.

Anjalie Field, Chan Young Park, Kevin Z Lin, and Yulia Tsvetkov. 2022. Controlled analyses of social biases in wikipedia bios. In *Proceedings of the ACM Web Conference 2022*, pages 2624–2635.

Rudolf Flesch. 2007. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.

Jim Giles. 2005. Special report internet encyclopaedias go head to head. *nature*, 438(15):900–901.

Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. Rumoureval 2019: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1809.06683*.

Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–37.

Stefan Heindorf, Martin Potthast, Benno Stein, and Gregor Engels. 2016. Vandalism detection in wikidata. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 327–336, New York, NY, USA. Association for Computing Machinery.

Freddy Heppell, Kalina Bontcheva, and Carolina Scarton. 2023. Analysing state-backed propaganda websites: a new dataset and linguistic study. *arXiv preprint arXiv:2310.14032*.

Peter Hernon. 1995. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government Information Quarterly*, 12(2):133–139.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

M. Hu, E. Lim, A. Sun, H. W. Lauw, and B. Vuong. 2007. Measuring article quality in wikipedia. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*.

Cherilyn Ireton and Julie Posetti. 2018. *Journalism, fake news & disinformation: handbook for journalism education and training*. Unesco Publishing.

Sara Javanmardi, David W. McDonald, and Cristina V. Lopes. 2011. Vandalism detection in wikipedia: a high-performing, feature-rich model and its reduction through lasso. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, page 82–90, New York, NY, USA. Association for Computing Machinery.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Isaac Johnson, Martin Gerlach, and Diego Sáez-Trumper. 2021. Language-agnostic topic classification for wikipedia. In *Companion Proceedings of the Web Conference 2021*, pages 594–601.

Srijan Kumar, Francesca Spezzano, and VS Subrahmanian. 2015. Vews: A wikipedia vandal early warning system. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 607–616.

Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International World Wide Web Conference*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Julio Amador Díaz López and Pranava Madhyastha. 2021. A focused analysis of twitter-based disinformation from foreign influence operations. In *Proceedings of the 1st International Workshop on Knowledge Graphs for Online Discourse Analysis (KnOD 2021) co-located with the 30th The Web Conference (WWW 2021)*, volume 2877. CEUR Workshop Proceedings.

Zachary J. McDowell and Matthew A. Vetter. 2020. It takes a village to combat a fake news army: Wikipedia's community and policies for information literacy. *Social Media + Society*, 6(3):2056305120937309.

Dean Pomerleau and Delip Rao. 2017. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. *Fake news challenge*.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *arXiv preprint arXiv:2205.12617*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jodi Schneider, Bluma S. Gelley, and Aaron Halfaker. 2014. Accept, decline, postpone: How newcomer productivity is reduced in english wikipedia by pre-publication review. In *Proceedings of The International Symposium on Open Collaboration*, OpenSym '14, page 1–10, New York, NY, USA. Association for Computing Machinery.

Mina Schütz, Daniela Pisoiu, Daria Liakhovets, Alexander Schindler, and Melanie Siegel. 2024. GerDIS-DETECT: A German multilabel dataset for disinformation detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7683–7695, Torino, Italia. ELRA and ICCL.

Antonina Sinelnik and Dirk Hovy. 2024. Narratives at conflict: Computational analysis of news framing in multilingual disinformation campaigns. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 225–237, Bangkok, Thailand. Association for Computational Linguistics.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Qi Su, Mingyu Wan, Xiaoqian Liu, Chu-Ren Huang, et al. 2020. Motivations, methods and metrics of misinformation detection: an nlp perspective. *Natural Language Processing Research*, 1(1-2):1–13.

Arsim Susuri, Mentor Hamiti, and Agni Dika. 2017. Detection of vandalism in wikipedia using metadata features – implementation in simple english and albanian sections. *Advances in Science, Technology and Engineering Systems Journal*, 2:1–7.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and verification (fever) shared task. *arXiv preprint arXiv:1811.10971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Mykola Trokhymovych, Muniza Aslam, Ai-Jou Chou, Ricardo Baeza-Yates, and Diego Saez-Trumper. 2023. Fair multilingual vandalism detection system for wikipedia. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4981–4990.

Mykola Trokhymovych, Indira Sen, and Martin Gerlach. 2024. An open multilingual system for scoring readability of wikipedia. *arXiv preprint arXiv:2406.01835*.

Vladimir Vapnik. 2013. *The nature of statistical learning theory*. Springer science & business media.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. *arXiv preprint arXiv:2010.03159*.

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

William Yang Wang and Kathleen McKeown. 2010. "got you!": Automatic vandalism detection in Wikipedia with web-based shallow syntactic-semantic modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1146–1154, Beijing, China. Coling 2010 Organizing Committee.

Stanisław Węglarczyk. 2018. Kernel density estimation and its application. In *ITM web of conferences*, volume 23, page 00037. EDP Sciences.

KayYen Wong, Miriam Redi, and Diego Saez-Trumper. 2021. Wiki-reliability: A large scale dataset for content reliability on wikipedia. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2437–2442, New York, NY, USA. Association for Computing Machinery.

Qinyi Wu, Danesh Irani, Calton Pu, and Lakshmish Ramaswamy. 2010. Elusive vandalism detection in wikipedia: a text stability-based approach. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, page 1797–1800, New York, NY, USA. Association for Computing Machinery.

Shuhan Yuan, Panpan Zheng, Xintao Wu, and Yang Xiang. 2017. Wikipedia vandal early detection: from user behavior to user embedding. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 17*, pages 832–846. Springer.

Honglei Zeng, Maher A. Alhossaini, Li Ding, Richard Fikes, and Deborah L. McGuinness. 2006. Computing trust from revision history. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services*, PST '06, New York, NY, USA. Association for Computing Machinery.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2).

A Dataset Details

We release our dataset in 3 settings as mentioned in Section 4.2. The settings with data splits and their corresponding sizes are mentioned in Table 8.

			Number of Instances		
Dataset Setting	**Dataset Type**	**Split**	**Non-hoax**	**Hoax**	**Total**
1Hoax2legitimate	Definition	Train	426	206	632
		Test	179	93	272
1Hoax2legitimate	Full Text	Train	456	232	688
		Test	200	96	296
1Hoax10legitimate	Definition	Train	2,225	203	2,428
		Test	940	104	1,044
1Hoax10legitimate	Full Text	Train	2,306	218	2,524
		Test	973	110	1,083
1Hoax100legitimate	Definition	Train	20,419	217	20,636
		Test	8,761	82	8,843
1Hoax100legitimate	Full Text	Train	22,274	222	22,496
		Test	9,534	106	9,640

Table 8: Dataset details in definition-only and fulltext settings with number of hoax and legitimate article splits.

B Language Model Training Details

We train our Language Models with the configuration given below. We use one NVIDIA RTX4090 to train the LMs, one NVIDIA V100 and one NVIDIA A100 GPU to train the LLMs.

- Learning rate: 2e-06
- Batch size: 4 (for Fulltext experiments) and 8 (For Definition experiments)
- Epochs: 30
- Loss Function: Weighted Cross Entropy Loss
- Gradient Accumulation Steps: 4
- Warm-up steps: 100

C Prompt for LLM in-context learning

The instruction prompt used for LLMs in their in-context learning with examples for few shot experiment are given below.

You are a helpful knowledge management expert and excel at identifying whether an input Wikipedia article is a hoax or not. Wikipedia defines a hoax as 'a deliberately fabricated falsehood made to masquerade as truth'. You take an Wikipedia article as input and return with the label citing hoax(Label 1) or real(Label 0) based only on the text of the article. Given an article from Wikipedia, your task is to analyze the article text to identify if the article is hoax or real. The Hoax and real articles are defined as follows:

- Hoax: An article that is deliberately fabricated falsehood made to masquerade as truth.
- Real: An article which contains information about an existing entity and are not fabricated.

Your output should be a JSON dictionary with label that you found. Here are the possible labels with what they mean:

- 0 : The article is real article.
- 1 : The article is a hoax article.

Your input will be in the following format:
INPUT: { Text: <Article text> }
OUTPUT: { Label: <One of the label from the possible labels: 0 and 1, where 0 is real article and 1 is hoax article.> }
Please respond with only the JSON dictionary containing label. You are instructed strictly to return output only in the format given above, nothing else. No yapping.

Here are the examples used in few-shot experiments.

Example 1:
INPUT:
{ Text: Albion Dauti (born May 31, 1995 in Caracas, Venezuela) is a Venezuelan telenovela actor and presenter. Albion Dauti was born in Caracas, Venezuela. He studied acting at the Faculty of Arts in Caracas. In 2010 he started working as a presenter in Venevision. }
OUTPUT:
{ Label: 1 }
Example 2:
INPUT:
{ Text: Michelle Madhok (born May 26, 1971) is the Founder and CEO of White Cat Media Inc. - DBA SheFinds Media, parent company of online shopping publication SheFinds. com and MomFinds. com. She writes a weekly style column for New York's Metro newspaper and appears regularly on Fox News Channel, The Today Show, and The Tyra Banks Show. }
OUTPUT:
{ Label: 0 }

D Vocabulary creation for revision history classification

We generate the vocabulary for timeline via the following process.

1. We extract the revision history of each article and convert the all the timestamps to standardized date-time format.

2. Group the timestamps by month and year (MM-YYYY). We call this Binning.

3. Count the number of revisions for each bin.

4. Return a dictionary of month-year bins and their corresponding counts.