information

Article

Combating Fake News in “Low-Resource” Languages: Amharic Fake News Detection Accompanied by Resource Crafting

Fantahun Gereme [1,*], William Zhu [1], Tewodros Ayall [2] and Dagmawi Alemu [2]

[1] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China; wfzhu@uestc.edu.cn

[2] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; 201714060124@std.uestc.edu.cn (T.A.); 201714060101@std.uestc.edu.cn (D.A.)

* Correspondence: fantishb@gmail.com or fantishb@std.uestc.edu.cn

Abstract: The need to fight the progressive negative impact of fake news is escalating, which is evident in the strive to do research and develop tools that could do this job. However, a lack of adequate datasets and good word embeddings have posed challenges to make detection methods sufficiently accurate. These resources are even totally missing for “low-resource” African languages, such as Amharic. Alleviating these critical problems should not be left for tomorrow. Deep learning methods and word embeddings contributed a lot in devising automatic fake news detection mechanisms. Several contributions are presented, including an Amharic fake news detection model, a general-purpose Amharic corpus (GPAC), a novel Amharic fake news detection dataset (ETH_FAKE), and Amharic fasttext word embedding (AMFTWE). Our Amharic fake news detection model, evaluated with the ETH_FAKE dataset and using the AMFTWE, performed very well.

Keywords: fake news; deep learning; Amharic corpus; dataset; word embedding

Citation: Gereme, F.; Zhu, W.; Ayall, T.; Alemu, D. Combating Fake News in “Low-Resource” Languages: Amharic Fake News Detection Accompanied by Resource Crafting. *Information* **2021**, *12*, 20. https://doi.org/10.3390/info12010020

Received: 15 December 2020
Accepted: 4 January 2021
Published: 7 January 2021

1. Introduction

Online media, specifically social media, is easily accessible, cheap, suitable for commenting and sharing, and more timely [1–3], which enables it to be favored by many, especially youngsters. However, it also has a dark side: the propagation of hate speech and inauthentic information, such as fake news. Fake news refers to news articles that are intentionally and verifiably false [4,5]. Fake news is increasingly becoming a threat to individuals, governments, freedom of speech, news systems, and society as a whole [3,6,7]. It disturbs the authenticity balance of the news system, creating real-life fears in the world’s societies. To express the spread and bad effect of fake news during the current pandemic, the WHO warned against fake news in the COVID-19 infodemic (https://www.who.int/dg/speeches/detail/director-general-s-remarks-at-the-media-briefing-on-2019-novel-coronavirus---8-february-2020). It said that while the virus spreads, misinformation makes the job of our heroic health workers even harder; it diverges the attention of our decision-makers and it causes confusion and spreads fear in the general public. The list of practical examples of the impacts of fake news is becoming extensive and the danger is already eminent.

To reduce the adverse effects of fake news, governments, the tech industry, and individual researchers have been trying to devise various mechanisms. Governments tried to enact legal proclamations that they believe will suppress fake news. For example, the government of Ethiopia has enacted the Hate Speech and Disinformation Prevention and Suppression Proclamation No.1185/2020 (https://www.accessnow.org/cms/assets/uploads/2020/05/Hate-Speech-and-Disinformation-Prevention-and-Suppression-Proclamation.pdf), though this looks less helpful as creators of fake news hide themselves, and this obscurity leaves no trace for the law. Facebook, Google, Twitter, and YouTube tried to take technological measures, using certain tools. In the development of fake news detection tools, linguistic resources play crucial roles. However, “low-resource” languages, mostly African languages, such as Amharic, lack such resources and tools.

Amharic (አማርኛ, Amarəñña), with the only African-origin script named Ethiopic/Fidel, is the second most spoken Semitic language in the World, next to Arabic, and it is the official working language of the Ethiopian government. As more Ethiopians are living outside their home, the Amharic language speakers in different countries of the world is also growing. In Washington DC, Amharic has gotten the status as one of the six non-English working languages [8]. Furthermore, Amharic is considered as a sacred language by Rastafarians across the world. Despite this, the Amharic language is one of the “low-resource” languages in the world, which lacks the tools and resources important for NLP (natural language processing) and other techno-linguistic solutions. To the best of our knowledge, for the Amharic language, there is no fake news detection dataset and we could not find work done to detect fake news written in Amharic. Moreover, there is a lack of quality Amharic word embedding. The available Amharic corpora are not sufficient and some of them are not freely open to the public. This is a total disadvantage, not being able to benefit from technology solutions.

In this work, we tried to narrow down those gaps. We present several contributions that include the following:

- We collected and organized a huge Amharic general purpose corpus.
- We created Amharic fasttext word embedding.
- We prepared a novel fake news detection dataset for the Amharic language.
- We introduced a deep learning-based model for Amharic fake news detection.
- We performed a series of experiments to evaluate the word embedding and fake news detection model.

The rest of this document is organized as follows. In Section 2, we present the general-purpose Amharic corpus (GPAC). Section 3 explains the Amharic fasttext word embedding (AMFTWE). The Amharic fake news detection dataset (ETH_FAKE) is explained in Section 4. Section 5 is dedicated to the experiments, results, and discussion, while Section 6 concludes our work.

2. GPAC: General-Purpose Amharic Corpus

One of the challenges of content-based fake news detection is the absence of sufficient corpora to train word embeddings, which are used in a multiplicity of NLP applications, including fake news detection [1,9–15], either to represent the features for traditional classifiers, or to initialize the deep neural network embedding layers. Similarly, the shortage of an appropriate dataset to train fake news detection models is the other bottleneck. Especially African languages labeled “under-resourced”, such as Amharic, suffer from a shortage of such resources.

Amharic is a highly influential language with its own ancient script. Not to mention its early existence and applications, it has been the working language of courts, the military, language of trade, and everyday communications since the late 12th century, and remains the official language of the Ethiopian government today [16,17]. Most of the Ethiopian Jewish communities in Ethiopia and Israel speak Amharic. In Washington DC, Amharic became one of the six non-English languages in the Language Access Act of 2004, which allows government services and education in Amharic [8]. Furthermore, Amharic is considered as a sacred language by Rastafarians. Despite Amharic being highly powerful, it is still one of the “low-resource” languages in the world. A lack of sufficient corpora and linguistic tools to help use technology make the language disadvantaged in this regard. However, there have been a few works done to prepare the Amharic corpus and linguistic tools.

The Walta Information Center (WIC) corpus is a small-sized corpus with 210,000 tokens collected from 1065 Amharic news documents [18]. The corpus is manually annotated for POS tags. It is, however, too small for deep learning applications. The HaBit project corpus is another web corpus, which was developed by crawling the web [19]. The corpus is cleaned and tagged for POS using a TreeTagger trained on WIC.

The Crúbadán corpus was developed under the project called corpus building for a large number of under-resourced languages [20]. The Amharic corpus consists of 16,970,855 words crawled from 9999 documents. This corpus is just a list of words with their frequencies, which is inconvenient for word embedding and other deep learning applications.

The Contemporary Amharic Corpus (CACO) [21] is another corpus crawled from different sources. We checked and got about 21 million tokens from 25,000 documents in this corpus. As we can see in Table 1, the WIC is too small and the Crúbadán is just a list of words and thus inconvenient to train quality Amharic word embeddings; the remaining two are not sufficient for the data-hungry word embedding training. Of course, the POS-tagged corpora are not directly usable for this purpose. Thinking to fill these gaps, we created our own general-purpose Amharic corpus (GPAC (https://github.com/Fanpoliti/GPAC)) collected from a variety of sources. This version of GPAC includes about 121 million documents and more than 40 million tokens.

Table 1. The general-purpose Amharic Corpus (GPAC) compared with other Amharic corpora.

Corpus	Documents	Tokens	Remark
The Walta Information Center (WIC)	1065	210,000	
The HaBit project (am131516)	75,509	25,975,846	
The Crúbadán	9999	16,970,855	Just a list of words and frequencies
Contemporary Amharic Corpus (CACO)	25,199	21,863,015	
General-Purpose Amharic Corpus (GPAC)	121,071	40,601,139	

2.1. Data Collection

We collected data from diversified sources and prepared a general-purpose Amharic corpus (GPAC). There are two objectives for preparing this corpus. First, it will be used as a general resource for future NLP research and tool development projects for the "low-resource" language Amharic. Second, added to the other corpora, it will be used to create a good-quality Amharic word embedding, which itself has two objectives. As part of this fake news detection work, it is the backbone of the embedding layer. Secondly, it is a vital resource in many NLP applications and others.

2.2. Data Processing

The preprocessing of the documents involves spelling correction, normalization of punctuation marks, and sentence extraction from documents for the purpose of randomizing the documents. Extracting each statement from individual documents and randomizing them helps make the corpus publicly available for researchers and tool developers without affecting the copyrights, if any. Different styles of punctuation marks have been used in the documents or articles. For quotation marks, different representations such as " ", " ", ‹‹ ››, ' ', ' ', or « » have been used. We normalized all types of double quotes by " ", and all single quotes by ' '. Other punctuation marks were normalized as follows: full stops (like:: and ።) by ።, hyphens (like:-, and ፡—) by ፦, and commas (like፥ and ÷) by ፣. Table 2 summarizes the various multi-domain data sources used to build the corpus.

Table 2. Data sources for the general-purpose Amharic corpus (GPAC).

Sources Types	Instances
News Papers	AddisAdmass, Reporter, Goolgule
Facebook	Government communication offices pages (Federal, Regional, Zonal, Woreda), university pages, think tank groups pages, media pages, etc.
Portals	EthiopiaNege, EthiopiaZare, Satenaw, ECADF
Forums	Ethiopian Review
Media	ESAT, Walta Information Center, BBC News Amharic
Books	Academic, fiction, historical, etc.
Religious Books	The Holy Bible, newspapers, others

3. Amharic Fasttext Word Embedding (AmFtWe)

Word embeddings have been used to represent the features for traditional classifiers, or as initializations in deep neural networks. Word embeddings are real-valued representations for text by embedding both the semantic and syntactic meanings obtained from an unlabeled large corpus and are perhaps one of the key advances for the remarkable performance of deep learning methods in challenging NLP (natural language processing) problems, such as content-based fake news detection [6,14,22]. They are widely used in NLP tasks, such as sentiment analysis [23], dependency parsing [24], machine translation [25], and fake news detection [1,9–15].

Considering the difficulty of the fake news detection problem, fake news detection methods using deep learning can benefit from good-quality word embeddings. Publicly available models, which are pre-trained on large amounts of data, have become a standard tool for many NLP applications, but are mostly available for the English language. Word embeddings for "low-resource" languages are absent or are very limited. For the Amharic language, a fasttext-based word embedding (cc_am_300) was trained by [26], using 300 dimensions. However, the number of word vectors are limited and also it contains uncleaned English tokens.

The distributional hypothesis used in [27–29] utilizes the idea that the meaning of a word is captured by the contexts in which it appears, to learn word embeddings. Thus, the quality of the word vectors directly depends on the amount and quality of data they were trained on. Based on this fact, in this work, we introduce a high-quality Amharic fasttext word embedding (AmFtWe (https://github.com/Fanpoliti/AMFTWE)) trained on a huge corpus (GPAC_CACO_WIC_am131516) obtained by merging and deduplicating four corpora (discussed in Section 2), namely, GPAC, am131516, WIC, and CACO, using a fasttext model with sub-word information [30]. Table 3 illustrates the architecture of the word embedding. As the quality of word embeddings directly depends on the amount and quality of data used, the AmFtWe is of high quality. This is manifested in the superior performance of our fake news detection model when it uses AmFtWe compared with cc_am_300 [26].

Table 3. The Amharic fasttext word embedding (AmFtWe) architecture.

Embedding Name	Corpus	Dimension	Size of the .vec File	Size of the .bin File
amftwe_300.bin/.vec	GPAC_CACO_WIC_am131516	300	1.91 gb	4.2 gb
amftwe_200.bin/.vec		200	1.27 gb	2.2 gb
amftwe_100.bin/vec		100	620 mb	1.2 gb
amftwe_50.bin/vec		50	312 mb	620 mb

The very reason we chose fasttext is because Amharic is one of the morphologically rich languages and it is possible to improve the vector representations for morphologically rich languages by using character-level information [30].

We evaluated AmFtWe using an extrinsic evaluation. In an extrinsic word embedding evaluation, we use word embeddings as the input features to a downstream task, in our case fake news detection, and measure the changes in performance metrics specific to that task [31–34]. For comparison purposes, we use the only available Amharic word embedding presented in [26]. This fake news detection, task-oriented evaluation, as presented in Section 5, shows that AmFtWe is a quality word embedding. The objective of preparing an AmFtWe is not only for the consumption of this paper; it is intended to be a valuable resource for future computational linguistic researches. For this reason, we will make it publicly available through our GitHub link (https://github.com/Fanpoliti/AMFTWE), with various dimensions and file formats (as shown in Table 3).

4. ETH_FAKE: A Novel Amharic Fake News Dataset

The fake news problem is a recent phenomenon and already a research issue, although still relatively less explored [35]. Even though there are research studies done, the scarcity of standard datasets was a common issue raised by many researchers. Deep learning-based fake news detection has been shown to be effective [6,22]. However, the data-hungry nature of this approach and the absence of sufficient datasets have made the research outcomes limited. Even the trials made in preparation of fake news detection datasets focused on the English language. Fake news detection in "under-resourced" languages, such as Amharic, is difficult to do because of the absence of a dataset. Though Amharic is a widely used language, as we have discussed in Section 2, and the impact of fake news in the regions using the language is a big concern—both to the government and society—there has not been any fake news detection research done for the Amharic language and there is no Amharic fake news detection dataset.

This critical problem motivated us to do fake news detection research and prepare an Amharic fake news detection dataset. We created the first Amharic fake news detection dataset with fine-grained labels and named it ETH_FAKE (https://github.com/Fanpoliti/ETH_FAKE). ETH_FAKE consists of 3417 real news articles and 3417 fake news articles gathered from Amharic Facebook pages and online newspapers that accounts to a total of 6834 articles. Table 4 summarizes the architecture of the ETH_FAKE dataset. We discussed the data collection and preprocessing of the data under Sections 4.1 and 4.2.

Table 4. Composition of ETH_FAKE: a novel Amharic fake news detection dataset.

News Group	Number of Articles	News Sources	News Domain
Real News	3417	Facebook, Addis Admass newspaper, Reporter newspaper	Sport, politics, art, social, religion, education, economics, history
Fake News	3417	Facebook, Addis Admass newspaper, Reporter newspaper	Sport, politics, art, social, religion, education, economics, history

4.1. Data Collection

As there was no existing Amharic fake news detection dataset, it was compulsory to collect data from the scratch. Obtaining well balanced real and fake pieces of Amharic news is not an easy task; especially getting the fake articles was tiresome. Both real news and fake news articles were obtained from Facebook and two well-known Ethiopian private newspapers, Reporter and Addis Admass. Reporter (Amharic: ሪፖርተር) and Addis Admass (Amharic: አዲስ አድማስ) are private newspapers published in Addis Ababa, Ethiopia. Archives of the online Amharic versions of these newspapers were scrapped using a python script. Even though these newspapers are presumed to broadcast real news, we have done fact checking for the news pieces by employing four senior journalists and a linguist. Articles collected from Facebook passed through the same procedure to check whether they contain factual news.

Getting the fake news group was the utmost demanding task. As we could not get Amharic fake news sources, we were required to collect them piece-by-piece from scratch. We collected the fake news articles from Facebook and the aforementioned online newspapers. After identifying check-worthy Amharic Facebook pages, groups, and profiles, we scrapped the pages and fact-checked them manually using a group of senior journalists and a linguist. The ridiculously false statements like bombardment of an existing facility, bridge, or dam; contradictions to general truth; etc. are easily picked as FAKE, where as other contents were analyzed thoroughly. A FAKE label is attached to an article if the content is reporting completely false information or most of the content is verifiably false. Both the real news and the fake news come from multiple domains like sport, politics, arts, education, religion, economics, history, etc. Even though most of the fake news comes from Facebook, we tried to balance the domains of the real and fake sources.

4.2. Preprocessing

The absence of well-developed tools, like English and other languages, makes Amharic text preprocessing difficult. Its customary to find multiple languages intermingled in web-scraped texts. We cleaned out mixed texts of English from the news articles.

5. Automatic Fake News Detection: Model, Experiments, and Evaluation

5.1. Evaluation Metrics

We evaluated both the word embedding and the fake news detection model. For the fake news detection model, we used the accuracy, precision, recall, and F1 score as the evaluation matrices. For the evaluation of the word embeddings, we used the extrinsic evaluation technique, which is an evaluation of word embeddings using a specific task with metrics [31–34]. Our word embedding was evaluated against an existing pre-trained word embedding [26] using the same task—fake news detection.

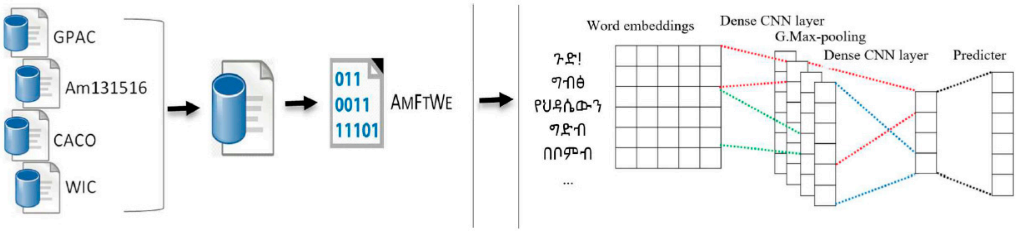

Figure 1. General diagram for the Amharic fake news detection model.

5.2. Experimental Setup

We wrote both the main project code and the data-scraping script in Python 3, using the TensorFlow r1.10, NumPy, and Keras library. Figure 1 depicts the Amharic fake news detection model based on Convolutional Neural Networks (CNNs). Since CNNs have been proven to show superior performance in text classification tasks, specifically in content-based fake news detection [6,14,22], we used them as model building methods. We present specific details as follows. The fake news detection dataset was preprocessed and split for training and validation in an 80/20 ratio (80% of it for training set and the remainder 20% for validation set). We used an embedding dimension of various sizes with 10,000 unique tokens and a 5000 sequence-length post padded with zeros. The output of the embedding layer was fed into a dense network of 128 neurons with the ReLU (Rectified Linear Unit) activation function. Then this output was passed into a one-dimensional GlobalMaxPooling layer. The output of the GlobalMaxPooling layer was again fed into a dense network of 128 neurons with the ReLU activation function, whose output was finally passed into a one-dimensional dense network with a sigmoid activation function. Rmsprop was used as the optimization technique and binary-crossentropy as the loss function.

We used our word embedding (AMFTWE) with the dimensions 50, 100, 200, and 300 to record the performance of the fake news detection model in different dimensions. For the comparison of the two word embeddings, AMFTWE and cc_am_300, we set up a separate experiment with 300 dimensions, because the pre-trained word embedding (cc_am_300 [26]) is available in 300 dimensions only.

5.3. Results and Discussion

As the experimental results depicted in Table 5 show, the Amharic fake news detection model performed very well. The model scored validation accuracy above 99% while using the 300- and 200-dimension embeddings. This good performance might be attributed to both the fake news detection model and the quality of the word embedding. We could not find a content-based fake news detection work for the Amharic language for comparison.

Table 5. Experimental result of the model performance using the cc_am_300 [10] and AMFTW*e* embeddings.

Word Embedding	Dataset	Embedding Dim	Model Performance			
			Acc	**Pre**	**Rec**	**F1**
cc_am_300	ETH_FAKE	300	.9883	.9850	.9882	.9866
AMFTWE	ETH_FAKE	50	.9715	.9631	.9713	.9672
AMFTWE	ETH_FAKE	100	.9890	.9898	.9847	.9872
AMFTWE	ETH_FAKE	200	.9921	.9910	.9930	.9920
AMFTWE	ETH_FAKE	300	.9936	.9930	.9941	.9935

We recorded a higher performance of our model when it uses the Amharic fasttext word embedding AMFTWE than using the existing Amharic word embedding cc_am_300 presented in [26]. Since the evaluation of cc_am_300 and AMFTWE was made in the same experimental setup, we can say that the higher score of the model using AMFTWE is due to the relatively higher quality of AMFTWE. Hence, AMFTWE will be a valuable resource for future related research. Regarding the choice of dimensions, higher dimensions of AMFTWE, obviously, made the model perform better than the lower dimensions. As the results are proximate, we may opt to use either the 200-dimensional or 300-dimensional pre-trained word embedding, based on our memory and storage availability.

6. Conclusions

In this paper, we have studied Amharic fake news detection using deep learning and news content accompanied with the preparation of several computational linguistic resources for this “low-resource” African language. The lack of Amharic fake news detection research, especially due to the lack of both a fake news dataset and a good Amharic word embedding, as well as limitations in the existing Amharic corpora, have motivated us to contribute our share to fill these gaps. Together with the Amharic fake news detection model, we contributed several resources of paramount importance for this work and future research. We created ETH_FAKE: a novel Amharic fake news detection dataset with fine grained labels, which was collected from various multi-domain sources. Considering the lack of quality Amharic word embedding, we prepared AMFTWE: Amharic fasttext word embedding with sub-word information. GPAC, a general-purpose Amharic corpus, was the other contribution of this work. We used GPAC merged with other publicly available corpora to train AMFTWE, which, in turn, was used to initialize the embedding layer of our fake news detection model. Our fake news detection model performed very well using both word embeddings, cc_am_300 and AMFTWE. However, it exhibited a higher performance record when AMFTWE was used compared to cc_am_300, which could be attributed to the quality of our word embedding, which was trained on a relatively huge corpus. As deep learning methods require more data, this work may further be improved by increasing the size of ETH_FAKE and GPAC. On the other hand, using other word embedding algorithms, such as BERT (Bidirectional Encoder Representations from Transformers), could help train a word embedding possibly better than AMFTWE, provided the data-hungry nature of BERT is satisfied. However, crafting an Amharic fake news dataset and obtaining a large number of Amharic corpora will be challenging.

Author Contributions: Conceptualization, F.G. and W.Z.; methodology, F.G.; software, F.G. and D.A.; validation, F.G., W.Z., T.A. and D.A.; formal analysis, F.G.; investigation, F.G.; resources, T.A.;

data curation, T.A.; writing—original draft preparation, F.G.; writing—review and editing, F.G.; visualization, D.A.; supervision, W.Z.; project administration, F.G. and W.Z.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available corpora were used in this study, papers associated with these resources are properly cited.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bajaj, S. *The Pope Has a New Baby! Fake News Detection Using Deep Learning*; Stanford University: Stanford, CA, USA, 2017.
2. Ferreira, W.; Vlachos, A. Emergent: A novel data-set for stance classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; ACM: New York, NY, USA, 2016; pp. 1163–1168.
3. Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B.J.; Wong, K.; Cha, M. Detecting rumors from microblogs with recurrent neural networks. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), New York, NY, USA, 9–15 July 2016; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 2016; pp. 3818–3824.
4. Shu, K.; Slivaz, A.; Wangy, S.; Tang, J.; Liuy, H. Fake news detection on social media: A data mining perspective. *SIGKDD* **2017**, *19*, 22–36. [CrossRef]
5. Zhou, X.; Zafarani, R. 2018 A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.* **2020**, *53*. [CrossRef]
6. Gereme, F.B.; William, Z. Fighting fake news using deep learning: Pre-trained word embeddings and the embedding layer investigated. In Proceedings of the 3rd International Conference on Computational Intelligence and Intelligent Systems (CIIS 2020), Tokyo, Japan, 13–15 November 2020; ACM: New York, NY, USA, 2020. [CrossRef]
7. Mohammad, S.M.; Sobhani, P.; Kiritchenko, S. Stance and sentiment in tweets. *ACM Trans. Internet Technol.* **2017**, *17*, 1–23. [CrossRef]
8. District of Columbia Language Access Act Fact Sheet 2004. Available online: https://ohr.dc.gov/publication/know-your-rights-cards-amharic (accessed on 6 January 2021).
9. Karimi, H.; Tanh, J. Learning hierarchical discourse-level structure for fake news detection. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3432–3442.
10. Khan, J.Y.; Khondaker, T.I.; Iqbal, A.; Afroz, S. A benchmark study on machine learning methods for fake news detection. *arXiv* **2019**, arXiv:1905.04749.
11. Singhania, S.; Fernandez, N.; Rao, S. 3HAN: A deep neural network for fake news detection. In Proceedings of the International Conference on Neural Information Processing, Guangzhou, China, 14–18 November 2017; Springer: Cham, Switzerland, 2017.
12. Tagami, T.; Ouchi, H.; Asano, H.; Hanawa, K.; Uchiyama, K.; Suzuki, K.; Inui, K.; Komiya, A.; Fujimura, A.; Yanai, H.; et al. Suspicious news detection using micro blog text. In Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, Hong Kong, China, 1–3 December 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018.
13. Thota, A.; Tilak, P.; Ahluwalia, S.; Lohia, N. Fake news detection: A deep learning approach. *SMU Data Sci. Rev.* **2018**, *1*, 10.
14. Wang, W.Y. Liar-Liar pants on fire: A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, BC, Canada, 6 February 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 422–426.
15. Yang, Y.; Zheng, L.; Zhang, J.; Cui, Q.; Li, Z.; Yu, P. TI-CNN: Convolutional neural networks for fake news detection. *arXiv* **2018**, arXiv:1806.00749.
16. Ronny, M. Amharic as lingua franca in Ethiopia. *Lissan J. Afr. Lang. Linguist.* **2006**, *20*, 117–131.
17. Anbessa, T. Amharic: Political and social effects on English loan words. In *Globally Speaking: Motives for Adopting English Vocabulary in Other Languages*; Rosenhouse, J., Kowner, R., Eds.; Multilingual Matters: Bristol, UK, 2008; p. 165.
18. Demeke, G.A.; Getachew, M. Manual annotation of Amharic news items with part-of-speech tags and its challenges. In *Ethiopian Languages Research Center Working Papers*; Ethiopian Languages Research Center: Addis Ababa, Ethiopia, 2016; Volume 2, pp. 1–16.
19. Rychlý, P.; Suchomel, V. Annotated Amharic Corpora. In Proceedings of the International Conference on Text, Speech, and Dialogue(TSD2016), Brno, Czech Republic, 12–16 September 2016; Springer: Cham, Switzerland, 2016; pp. 295–302.
20. Scannell, K.P. The Crúbadán Project: Corpus building for under-resourced languages. *Cah. Cental* **2007**, *5*, 1.
21. Gezmu, A.M.; Seyoum, B.E.; Gasser, M.; Nürnberger, A. Contemporary Amharic Corpus: Automatically Morpho-Syntactically Tagged Amharic Corpus. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, Santa Fe, NM, USA, 2018*; Association for Computational Linguistics: Stroudsburg, PA, USA, August 2018; pp. 65–70.

22. Gereme, F.B.; William, Z. Early detection of fake news, before it flies high. In Proceedings of the 2nd International Conference on Big Data Technologies (ICBDT2019), Jinan, China, 28–30 August 2019; ACM: New York, NY, USA, 2019; pp. 142–148. [CrossRef]
23. Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; Qin, B. Learning sentiment-specific word embedding for twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics 1, Baltimore, MD, USA, 23–25 June 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1555–1565.
24. Ouchi, H.; Duh, K.; Shindo, H.; Matsumoto, Y. Transition-Based dependency parsing exploiting supertags. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2059–2068. [CrossRef]
25. Chen, K.; Zhao, T.; Yang, M.; Liu, L.; Tamura, A.; Wang, R.; Utiyama, M.; Sumita, E. A neural approach to source dependence based context model for statistical machine translation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 266–280. [CrossRef]
26. Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning word vectors for 157 languages. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; European Language Resources Association (ELRA): Paris, France, 2018.
27. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. In Proceedings of the International Conference on Learning Representations (ICLR 2013), Scottsdale, AZ, USA, 2–4 May 2013.
28. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 2, Lake Tahoe, SN, USA, 5–10 December 2013; Curran Associates Inc.: Red Hook, NY, USA, 2013; pp. 3111–3119.
29. Pennington, J.; Socher, R.; Manning, C.D. GloVe-Global Vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1532–1543.
30. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *TACL* **2017**, *5*, 135–146. [CrossRef]
31. Bairong, Z.; Wenbo, W.; Zhiyu, L.; Chonghui, Z.; Shinozaki, T. Comparative analysis of word embedding methods for DSTC6 end-to-end Conversation Modeling Track. In Proceedings of the 6th Dialog System Technology Challenges (DSTC6) Workshop, Long Beach, CA, USA, 10 December 2017.
32. Li, H.; Li, X.; Caragea, D.; Caragea, C. Comparison of word embeddings and sentence encodings as generalized representations for crisis tweet classification tasks. In Proceedings of the ISCRAM Asian Pacific Conference, Wellington, New Zealand, 4–7 November 2018.
33. Wang, B.; Wang, A.; Chen, F.; Wang, Y.; Kuo, C.J. Evaluating word embedding models: Methods and experimental results. *APSIPA Trans. Signal Inf. Process.* **2019**, *8*. [CrossRef]
34. Schnabel, T.; Labutov, I.; Mimno, D.; Joachims, T. Evaluation methods for unsupervised word embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 298–307.
35. Allcott, H.; Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **2017**, *31*, 211–236. [CrossRef]