Never guess what I heard... Rumor Detection in Finnish News: a Dataset and a Baseline

Hämäläinen, Mika

The Association for Computational Linguistics
2021-05

Hämäläinen , M , Alnajjar , K , Partanen , N & Rueter , J 2021 , Never guess what I heard... Rumor Detection in Finnish News: a Dataset and a Baseline . in A Feldman , G Da San Martino , C Leberknight & P Nakov (eds) , Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda . The Association for þÿComputational Linguistics , pp. 39 44 , Workshop on NLP for Internet

http://hdl.handle.net/10138/330266

Never guess what I heard... Rumor Detection in Finnish News: a Dataset and a Baseline

Mika Hämäläinen, Khalid Alnajjar, Niko Partanen and Jack Rueter
Department of Digital Humanities
University of Helsinki
firstname.lastname@helsinki.fi

Abstract

This study presents a new dataset on rumor detection in Finnish language news headlines. We have evaluated two different LSTM based models and two different BERT models, and have found very significant differences in the results. A fine-tuned FinBERT reaches the best overall accuracy of 94.3% and rumor label accuracy of 96.0% of the time. However, a model fine-tuned on Multilingual BERT reaches the best factual label accuracy of 97.2%. Our results suggest that the performance difference is due to a difference in the original training data. Furthermore, we find that a regular LSTM model works better than one trained with a pretrained word2vec model. These findings suggest that more work needs to be done for pretrained models in Finnish language as they have been trained on small and biased corpora.

1 Introduction

Contemporary online news media contains information from more and less reputable sources, and in many cases the reliability of individual news articles can be very questionable. This has far reaching impact on society and can even influence decision making, as everyone continuously encounters such material online. This is a real issue as identified by Paskin (2018). In their study, they found out that participating people could only, on the average, distinguish fake news from real news half of the time, and none of the participants was able to identify all samples of fake and real news correctly.

Ever since the 2016 US elections fake news and misinformation have become a hot topic in the English speaking world (Allcott and Gentzkow, 2017), however other countries are no less immune to the spread of such information. In this study we present for the first time a method for rumor detection in Finnish news headlines. Finnish language has not yet received any research interest in this topic, and for this reason we also propose a new dataset[1] for the task. We treat this task as a classification problem where each news headline is categorized as being either rumorous or factual.

Rumor detection is a very challenging task, and we believe that truly satisfactory results need to leverage other methods than only natural language processing. Whether a given text is a rumor or not is very strongly connected to real world knowledge and continuously changing world events that we don't believe that this can be solved within the analysis of individual strings without a larger context. However, if we can perform even a rough classification at this relatively simple level, this could be used as one step in more robust and complex implementations. Therefore, our initial approach should be seen as a baseline for future implementations, while it can be seen as an important advancement in this work, and in this case it is the starting point, as related work in matters of this topic for Finnish remains nonexistent.

2 Related Work

Rumor detection has in recent years become an active topic of investigation, especially due to the complex influence it has on modern societies through social media. There has been other work on rumor detection for languages other than English as well. Alzanin and Azmi (2019) studied rumor detection in Arabic tweets and Chernyaev et al. (2020) in Russian tweets. Recently, Ke et al. (2020) has also presented a method for rumor detection in Cantonese. A closely related topic, stance detection, has been studied in a comparable corpus of French Tweets (Evrard et al., 2020). In this section, we describe some of the related work in more detail.

Rubin et al. (2016) harnessed satire in the task of fake news detection, in their study, this figure of

[1] Our dataset is freely available for download on Zenodo https://zenodo.org/record/4697529

Proceedings of the 4th NLP4IF Workshop on NLP for Internet Freedom, pages 39–44
June 6, 2021. ©2021 Association for Computational Linguistics

language, that has also sparked research interest in detection (Li et al., 2020) and generation (Alnajjar and Hämäläinen, 2018) on its own, was useful in detecting fake news. They proposed an SVM (support vector machines) approach capturing five features: *Absurdity*, *Humor*, *Grammar*, *Negative Affect* and *Punctuation*. The idea of satire in fake news detection was also studied later on by Levi et al. (2019).

Tree LSTMs have been used recently in rumor detection (Kumar and Carley, 2019). They train the models on social media text which contains interactions as people reply to statements either providing supporting or contradicting statements. Their model is capable of taking these replies into account when doing predictions.

Sujana et al. (2020) detect rumors by using multiloss hierarchical BiLSTM models. The authors claim the hierarchical structure makes it possible to extract deep information form text. Their results show that their model outperforms a regular BiLSTM model.

Previous work on Finnish news materials include a study by (Ruokolainen et al., 2019), where the articles were annotated for named entities. In addition, other researchers have targeted Finnish news materials, especially historical newspapers that are openly available. Furthermore, (Mela et al., 2019) has studied NER (named entity recognition) in the context of historical Finnish newspapers, and (Marjanen et al., 2019; Hengchen et al., 2019) have tested methods for analyzing broader changes in a historical newspaper corpus. Our work departs from this, as we focus on modern newspaper headlines.

Additionally, to our knowledge there has not been any previous work on rumor detection for Finnish, which makes our work particularly novel and needed.

3 Data

We collect data from a Finnish news aggregation website[2], in particular, we crawl the news headlines in the rumor category to form samples of rumor data. In addition, we crawl the headlines in the category news from Finland to compile a list of headlines that do not contain rumors but actual fact-based news stories. This way we have gathered 2385 factual and 1511 rumorous headlines totaling to 3896 samples. As a preprocessing step,

[2]https://www.ampparit.com/

	rumor	factual
train	1057	1669
test	227	358
validation	227	358

Table 1: The size of the data splits on a headline level

we tokenize the headlines with NLTK (Loper and Bird, 2002) word tokenizer.

We shuffle the data and split it 70% for training, 15% for validation and 15% for testing. The actual sizes can be seen in Table 1. We use the same splits for all the models we train in this paper. An example of the data can be seen in Table 2. The dataset has been published with an open license on Zenodo with a permanent DOI[3]. The splits used in this paper can be found in the dataset_splits.zip file.

4 Detecting Rumors

In this section, we describe the different methods we tried out for rumor detection. We compare LSTM based models with transfer learning on two different BERT models.

We train our first model using a bi-directional long short-term memory (LSTM) based model (Hochreiter and Schmidhuber, 1997) using OpenNMT (Klein et al., 2017) with the default settings except for the encoder where we use a BRNN (bi-directional recurrent neural network) (Schuster and Paliwal, 1997) instead of the default RNN (recurrent neural network). We use the default of two layers for both the encoder and the decoder and the default attention model, which is the general global attention presented by Luong et al. (2015). The model is trained for the default 100,000 steps. The model is trained with tokenized headlines as its source and the rumor/factual label as its target.

We train an additional LSTM model with the same configuration and same random seed (3435) with the only difference being that we use pretrained word2vec embeddings for the encoder. We use the Finnish embeddings provided by (Kutuzov et al., 2017)[4]. The vector size is 100 and the model has been trained with a window size 10 using skipgrams on the Finnish CoNLL17 corpus.

In addition, we train two different BERT based sequence classification models based on the Finnish BERT model FinBERT (Virtanen et al.,

[3]https://zenodo.org/record/4697529

[4]http://vectors.nlpl.eu/repository/20/42.zip

Headline	Rumor
Tutkimus: Silmälaseja käyttävillä ehkä pienempi riski koronatartuntaan *Study: People wearing eyeglasses may have a smaller risk of getting corona*	true
Koronaviruksella yllättävä sivuoire - aiheutti tuntikausien erektion *Coronavirus has a surprising symptom - caused an erection that lasted for hours*	true
Korona romahdutti alkoholin matkustajatuonnin *Corona caused a collapse in traveler import of alcohol*	false
Bussimatka aiheutti 64 koronatartuntaa *A bus trip caused 64 corona cases*	false

Table 2: Examples of rumorous and factual headlines related to COVID-19 from the corpus

	Overall	Factual	Rumor
LSTM	84.9%	93.2%	71.8%
LSTM + word2vec	71.6%	94.4%	35.6%
FinBERT	**94.3%**	93.2%	**96.0%**
Multilingual BERT	91.8%	**97.2%**	83.3%

Table 3: Overall and label level accuracies for each model

2019) and Multilingual BERT (Devlin et al., 2019), which has been trained on multiple languages, Finnish being one of them. We use the transformers package (Wolf et al., 2020) to conduct the fine tuning. As hyperparameters for the fine tuning, we use 3 epochs with 500 warm-up steps for the learning rate scheduler and 0.01 as the strength of the weight decay.

This setup takes into account the current state of the art at the field, and uses recently created resources such as Finnish BERT model, with our own custom made dataset. Everything is set up in an easily replicable manner, which ensures that our experiments and results can be used in further work on this important topic.

5 Results

In this section, we present the results of the models, in addition, we explain why these result were obtained by contrasting the task into the training data of each pretrained model. The accuracies of the models can be seen in Table 3.

The results vary greatly, with tens of percentages between different approaches. It is important to note that while FinBERT gets the best overall accuracy and the best accuracy in predicting rumorous headlines correctly, it does not get the best accuracy in predicting factual headlines correctly, as it is actually Multilingual BERT that gets the best accuracy for factual headlines. This makes us wonder why this might be so. When we look at the training data for these models, we can see that Multilingual BERT was trained on Wikipedia[5], whereas FinBERT was mainly trained on data from an internet forum, Suomi24[6], that is notorious for misinformation, (33% of the data) and Common Crawl[7] (60% of the data). Only 7% of the training data comes from a news corpus. When we put the results into perspective with the training data, it is not at all the case, as the authors of FinBERT claim in their paper: *"The results indicate that the multilingual models fail to deliver on the promises of deep transfer learning for lower-resourced languages"* (Virtanen et al., 2019). Instead, based on our results, it is only evident that Multilingual BERT outperforms FinBERT on factual headlines as its training data was based on an encyclopedia, and that FinBERT is better at detecting rumors as its training data had a large proportion of potentially rumorous text from Suomi24 forum.

In the same fashion, we can explain the results of the LSTM models. A great many papers (Qi et al., 2018; Panchendrarajan and Amaresan, 2018; Alnajjar, 2021) have found that pretrained embeddings improve prediction results when used with an LSTM model, however, in our case, we were better off without them. While the data description (Zeman et al., 2017) was not clear on what the data of the pretrained word2vec model consists of (apart from it being from Common Crawl), we can still inspect the overlap in the vocabulary of the training data and the pretrained model. Our training and validation datasets contain 17,729 unique tokens, out of which 5,937 were not present in the pretrained model. This means that approximately 33% of the

[5] https://github.com/google-research/bert/blob/master/multilingual.md
[6] https://keskustelu.suomi24.fi/
[7] https://commoncrawl.org/

tokens in our dataset were simply not present in the pretrained model.

This is partially due to the English driven tradition of not lemmatizing pretrained models, however, for a language such as Finnish this means that a simple overlap in vocabulary is not enough, instead one would even need to have overlap in the syntactic positions where the words have appeared in the data of a pretrained model and in the training data of the model that would utilize the pretrained embeddings. It is important to note that the pretrained word2vec model does not have a small vocabulary either (2,433,286 tokens).

In order to study whether the issue arises from the fact that the word2vec model is not lemmatized or from the fact that its training data was from a different domain, we conduct a small experiment. We lemmatize the words in our training and validation dataset and the words in the vocabulary of the word2vec model by using the Finnish morphological FST (finite-state transducer) Omorfi (Pirinen, 2015) through UralicNLP[8] (Hämäläinen, 2019). After lemmatization, our corpus contains 10,807 unique lemmas, 2,342 out of which are still not in the lemmatized vocabulary of the word2vec model. This means that even on the lemma level, 21.7% of the words are not covered by the word2vec model. The lemmatized vocabulary size of the pretrained model is 576,535 lemmas. It is clear that a model leveraging from sub-word units could not alleviate the situation either, as such models are mainly useful to cope with inflectional forms, but not with completely new words that merely look familiar on the surface.

6 Conclusions

Our study shows that with the tested settings it is possible to differentiate the rumor and non-rumor categories with a very high accuracy. As the experiment setup was relatively simple, yet elegant, we believe that similar results can also be repeated for other languages for which rumor detection systems have not yet been created. The experiments reported here are just one part in creating such a system for Finnish language. We believe that the path towards more thorough solutions lies in larger manually annotated datasets that contain even more variation than the materials we have now used. Although, some of these datasets could be automatically generated by using Finnish semantic databases (Hämäläinen, 2018) and syntax realization (Hämäläinen and Rueter, 2018) in conjunction with existing Finnish news headline generation methods (Alnajjar et al., 2019).

Possibly the most relevant finding of our study lies, however, in the results we detected with different BERT models and were able to connect into the differences in training data. These findings are important much beyond just rumor detection, which is only one domain where these models are being continuously used. As the question of training data seemed to be an important one also for the word2vec model in the LSTM experiment, we can only conclude that the level of the existing pretrained models for Finnish is not good enough for them to work in many different domains. This is not a question of Finnish being "low resourced" (see Hämäläinen 2021), as huge amounts of text exist in Finnish online, but more of a question of not enough academic interest in producing high-quality models. This is something we will look into in the future.

Further work is needed from a qualitative perspective to see what exactly leads to a certain classification, and which kind of error types can be detected. Since the classification was done solely based on linguistic features of the text, represented by the strings, we must assume there are lexical and stylistic differences that are very systematic. Not unlike in the case of the existing methods for rumor detection, our models did not have access to any real world knowledge about the rumors or factual and non-factual information at the time when the headlines were written. It is obvious that a very well functioning system can only be built in connection to this kind of sources, as the fact that something is a rumor is ultimately connected to the content and real world knowledge, and not just the words in the string. However, we argue that our system could already be useful in a rough classificatory tasks where rumor containing news could be automatically selected for manual verification, or for verification with a more specialized neural network. Naturally further work also has to take into account more non-rumor text types and genres, so that certain degree of robustness can be reached.

[8] We use the dictionary forms model

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Khalid Alnajjar. 2021. When word embeddings become endangered. In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*. Rootroo Ltd.

Khalid Alnajjar and Mika Hämäläinen. 2018. A master-apprentice approach to automatic creation of culturally satirical movie titles. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 274–283.

Khalid Alnajjar, Leo Leppänen, and Hannu Toivonen. 2019. No time like the present: methods for generating colourful and factual multilingual news headlines. In *Proceedings of the 10th International Conference on Computational Creativity*. Association for Computational Creativity.

Samah M Alzanin and Aqil M Azmi. 2019. Rumor detection in Arabic tweets using semi-supervised and unsupervised expectation–maximization. *Knowledge-Based Systems*, 185:104945.

Aleksandr Chernyaev, Alexey Spryiskov, Alexander Ivashko, and Yuliya Bidulya. 2020. A rumor detection in Russian tweets. In *International Conference on Speech and Computer*, pages 108–118. Springer.

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.

Marc Evrard, Rémi Uro, Nicolas Hervé, and Béatrice Mazoyer. 2020. French tweet corpus for automatic stance detection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6317–6322.

Mika Hämäläinen. 2018. Extracting a semantic database with syntactic relations for finnish to boost resources for endangered uralic languages. *The Proceedings of Logic and Engineering of Natural Language Semantics 15 (LENLS15)*.

Mika Hämäläinen and Jack Rueter. 2018. Development of an open source natural language generation tool for finnish. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 51–58.

Simon Hengchen, Ruben Ros, and Jani Marjanen. 2019. A data-driven approach to the changing vocabulary of the 'nation'in english, dutch, swedish and finnish newspapers, 1750-1950. In *Digital Humanities 2019*.

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

Mika Hämäläinen. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*. Rootroo Ltd.

Liang Ke, Xinyu Chen, Zhipeng Lu, Hanjian Su, and Haizhou Wang. 2020. A novel approach for Cantonese rumor detection based on deep neural network. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1610–1615. IEEE.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proc. ACL*.

Sumeet Kumar and Kathleen Carley. 2019. Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5047–5058, Florence, Italy. Association for Computational Linguistics.

Andrei Kutuzov, Murhaf Fares, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 58th Conference on Simulation and Modelling*, pages 271–276. Linköping University Electronic Press.

Or Levi, Pedram Hosseini, Mona Diab, and David Broniatowski. 2019. Identifying nuances in fake news vs. satire: Using semantic and linguistic cues. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 31–35, Hong Kong, China. Association for Computational Linguistics.

Lily Li, Or Levi, Pedram Hosseini, and David Broniatowski. 2020. A multi-modal method for satire detection using textual and visual cues. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 33–38.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, page 63–70, USA. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Jani Marjanen, Ville Vaara, Antti Kanner, Hege Roivainen, Eetu Mäkelä, Leo Lahti, and Mikko Tolonen. 2019. A national public sphere? analysing the language, location and form of newspapers in finland, 1771–1917. *Journal of European Periodical Studies*, 4(1):54–77.

Matti La Mela, Minna Tamper, and Kimmo Tapio Kettunen. 2019. Finding nineteenth-century berry spots: Recognizing and linking place names in a historical newspaper berry-picking corpus. In *Digital Humanities in the Nordic Countries Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*. CEUR-WS. org.

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Danny Paskin. 2018. Real or fake news: who knows? *The Journal of Social Media in Society*, 7(2):252–273.

Tommi A Pirinen. 2015. Development and use of computational morphology of finnish in the open source and open science era: Notes on experiences with omorfi development. *SKY Journal of Linguistics*, 28:381–393.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2019. A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, pages 1–26.

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

Yudianto Sujana, Jiawen Li, and Hung-Yu Kao. 2020. Rumor detection on Twitter using multiloss hierarchical BiLSTM with an attenuation factor. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.