Tackling the Headline Incongruity Problem using Stance Detection

Simon Mariën , Pradeep Murukannaiah
TU Delft

Abstract

The increasing amount of untrusted content on the internet is a worrisome trend. The headline of an article can be adjusted to influence a potential readers attention and click-through rate. This clickbait or sensationalism can mislead the reader as the headline does not accurately represent the information in the corresponding article body.

This headline incongruity problem has received some recent attention from the research community with proposed datasets and approaches. However, there still lacks an overarching paper that tries to answer the current state of tackling the problem. This paper aims to fill this gap, look at recent proposed datasets and approaches, and compare them. These results will be discussed, reflected on, and formulated to answer the research question in conclusion. The main conclusions from this research are that the most suited dataset for comparison purposes is the Real-News based dataset and the Graph Neural Network by Yoon et al. performed the best.

1 Introduction

The headline incongruity problem occurs when the headline of an article does not accurately represent the information contained in the article [1]. An incongruent headline causes worrisome problems. First of all, it can mislead a reader who does not read the article body. Additionally, the stance and opinion of a reader can be misshaped with such a headline. Latsly, an incongruent headline can tempt the reader to reading an article that does not fully state what he is looking for.

The headline incongruity problem has some promising solutions involving stance detection [2][3][4]. The purpose of stance detection is to identify the stance of the text author towards a target (an entity, concept, event, idea, opinion, claim, and topic) either explicitly mentioned or implied within the text. The author can be in favour of, against, or neutral towards a proposition or target [5]. With stance detection being a rather new Natural Language Processing (NLP) task with diverse application areas [6], there is still a lot to discover and improve. This task lends itself well for the headline incongruity problem. The stance of the article body towards the headline is determined. In the simplest form, the article body with the headline can be classified as congruent or incongruent. Some literature uses related or unrelated [7][8].

In 2017, the Fake News Challenge [9] introduced a competition where multiple teams had to solve the headline incongruity problem using stance detection as a first step to combat fake news. Since then, multiple improvements and new approaches have been introduced, but a survey where recent possible solutions and approaches to this problem are analyzed and compared is missing in the literature. Based on this gap, we seek to answer the following research question:

How effective are stance detection methods in solving the headline incongruity problem?

Here, the effectiveness can mean a combination of factors including but not limited to accurate, efficient, and precise. In order to provide a comprehensive answer to this question, several subquestions are formulated.

1. What is the headline incongruity problem, and how can stance detection help to solve it?
2. What are the state-of-the-art stance detection models that solve the headline incongruity problem?
3. How do different models compare when trained and tested on the same dataset?
4. What are the pros and cons of each model?
5. What can be improved in future models to make them perform better compared to current models?
6. What issues can be solved using a stance detection based headline incongruity model?

Recent datasets and models in the headline incongruity area have been collected to answer the subquestions mentioned above. These datasets and models are looked at with a critical eye. Derived from these points of critique, the Real-News based dataset by Yoon et al. [4] is the most suited for comparison of models. It has a large sample size, is tested, and is released in a readable format. Comparison tests on the most promising models have been made when trained on this dataset. Due to process limitations, only three models have been tested in this research. The results for the other two models are copied from the paper that introduced the dataset. The Graph Neural Network from Yoon et al. [4] performed the best on the Real-News based dataset. From the

models tested in this paper, the model by team SOLAT in the SWEN had the highest performance. From these results, exciting pros and cons of the tested models have been discussed. Lastly, a conclusion on the current state with additional possible future improvements of tackling the headline incongruity problem has been made.

In Section 2, related work and background information on the research topic is discussed. In the following section, the method to answer the research question is explained. Next, in Section 4, the results and findings are discussed as well as the environment in which they were obtained. These results are compared to known numbers and placed in a broader context in Section 5. The interesting pros and cons of the tested models that become clear from the results are also discussed. The next section summarizes the answers to the research question and formulates recommendations for further research. After the conclusion, the ethical aspects and reproducibility of the research are analyzed.

2 Background and Related work

This section gives a background on the research area. Additionally, it presents a discussion on recent datasets and model approaches related to the research question.

2.1 Headline incongruity problem

The headline incongruity problem describes headlines that do not accurately represent the information contained in the articles within which they occur [1]. This is a problem in the current society where everyone can create content using the internet, unlike in the past, where only certified news channels and newspapers published news.

Two specific issues can be solved with a model that tackles the headline incongruity problem. The first issue is clickbait. This is a headline type that is mainly used by tabloids and online-native digital media sites. It is designed to withhold information to entice the reader to read on, or in most cases, to click. Figure 1 is an example of an article that uses clickbait to grab a readers attention and withhold information for the reader. A second issue is sensationalism. This headline type has the goal to dramatise an otherwise non-dramatic story. It does not force the reader to click by withholding information. An example of a sensational headline is the following.

The first lady of swearing! How a ten-year-old Michelle Obama lost out on a 'best camper' award because she wouldn't stop cursing [10].

2.2 Stance detection

When a speaker describes an object to express attitudes and relationships towards the object, the speaker is forming a stance. A stance is considered a social act of sharing a speaker's point of view on an object with an audience, sometimes prompting listeners to adopt their stance as well [11].

Stance detection is the use of algorithms to detect the stance of a text towards a target. The author can be in favour, against or neutral towards a proposition or target [5].

2.3 Fake news challenge

The Fake News Challenge (2017) is a competition where stance detection and the headline incongruity problem come together. The Fake News Challenge aims to explore how artificial intelligence technologies, particularly machine learning and natural language processing, might be leveraged to combat the fake news problem. They believe that these AI technologies hold promise for significantly automating parts of the procedure human fact-checkers use today to determine if a story is real or a hoax [9]. The goal of the FNC-1 challenge is to determine the perspective (or stance) of a news article relative to a given headline. An article's stance can either agree or disagree with the headline, discuss the same topic, or it is entirely unrelated [8]. This article stance can form the first step in a pipeline to combat fake news without human intervention. In the competition participated 50 teams that all developed their model and ran test data to evaluate their performance and create a ranking. The best performing model had a relative score of 82.02 out of 100, which is not perfect, so there is room for improvement [9].

Figure 1: Incongruent headline example.

2.4 Data sets

To train and evaluate machine learning models, we need data. For the headline incongruity problem, a data entry should consist of a headline, article body, and label. The label for the problem in its simplest form is 0 (congruent) or 1 (incongruent). The three datasets released for this problem are discussed below.

2.4.1 FNC-1 dataset (2017)

The FNC-1 challenge released a train and test dataset with 49,972 annotated and 25,413 unannotated headline-body pairs [1]. The annotation classes of this dataset are not binary but can be {Agree, Disagree, Discuss, Unrelated}. This can have the advantage that a pair that is not significantly congruent or incongruent can be classified as Discuss or Unrelated.

In the original fake news challenge, the final score of a model is calculated with following formula:

$$\begin{aligned}\text{score} = 0.75 \times (\text{correct}(\text{agree}, \text{disagree}, \text{discuss})) \\ + 0.25 \times (\text{correct}(\text{unrelated}))\end{aligned} \quad (1)$$

Correct(agree, disagree, discuss) stands for the number of times the model predicts agree as agree, or disagree as dis-

[1] https://github.com/FakeNewsChallenge/fnc-1

agree, or discuss as discuss. Correct(unrelated) are the number of times the model predicts unrelated as unrelated. The formula thus gives a higher weight to agree, disagree, and discuss instances compared to unrelated instances. This score calculation takes the large number of unrelated instances into account, as can be seen in Table 1. Nevertheless, there is still a significant imbalance between the related classes. Classifying related and unrelated articles is not difficult (best-performing systems reach an F1 score of about 0.99 for unrelated classification [8]). When a model with this accuracy of relatedness classification classifies all agree, disagree, and discuss classes always to a discuss label, the model receives an FNC-1 score of 0.833. This is higher than the top-performing model submitted in the competition in 2017, as can be seen in Table 2.

Table 1: Label distribution FNC-1 dataset [8].

Dataset	Agree	Disagree	Discuss	Unrelated
FNC-1	7.4%	2.0%	17.7%	72.8%

Hanselowski et al. proposed a new metric (F_1m) that is not affected by this class imbalance. The class-wise F_1 scores are the harmonic means of the precisions and recalls of the four classes [8]. This metric can also indicate which classes are not yet accurately predictable. In the underlying table, the original and new metrics are displayed calculated by Hanselowski et al.. Using the F_1m metric, the Athene team developed the best performing model and is the new winner of the FNC-1 challenge.

Table 2: Top performing models FNC-1 challenge [9][8].

Model	FNC-1 score	F_1m
SOLAT in the SWEN	0.8202	0.582
Athene (UKP Lab)	0.8197	0.604
UCL Machine Reading	0.8172	0.583

2.4.2 NELA-17 based dataset (2019)

Yoon et al. released a dataset to train their model to tackle the headline incongruity problem, which will be discussed in this section. In comparison with the FNC-1 dataset, this dataset is much larger. To increase the dataset size, they released a version where an article body text is split into their paragraphs and make sub-pairs of headline-paragraph. This method also reduces the length of text that a model should process. The data with full-body text is named 'whole', the one with split paragraphs is called 'paragraph'.

The dataset used for their testing is created using South Korean news articles published in 2016 and 2017. To decrease the language barrier, they processed the word tokens to integers. Next to the Korean one, an English dataset in the same integer format is released, but there are no evaluation results reported for this set. The English dataset is generated from articles originating from the NELA-2017 dataset. Horne et al. [12] introduced this dataset to help researchers study complex and diverse news related challenges. The final generated dataset by Yoon et al. uses binary labels {congruent, incongruent}.

2.4.3 Real-News based dataset (2021)

Yoon et al. [4] described a new and improved model, with a corresponding dataset, in 2021. The used news articles originate from Real News [13] which crawled the web for news articles from 2016 to 2019. From these 32 million articles, 7 million originating from the most trustworthy news sources were selected. A small sample was manually checked to confirm the congruity of trustworthy news sources.

Two types of datasets are generated, a random and a similar dataset. The incongruent headline-body pairs are generated by switching a randomly chosen amount of paragraphs from the original article body with paragraphs from other articles. The random dataset uses paragraphs from random articles from the Real News corpus. For the similar dataset, headline similarity is measured by the Euclidean distance of the fastText embeddings pre-trained on the WikiNews corpus [14]. To avoid the same news story, articles that were published in a small-time period are not used. The similar incongruent data is then created by using paragraphs from these similar news stories.

The random or similar incongruent pairs are then appended to an equal amount of congruent pairs to complete the dataset. This dataset uses binary labels {congruent, incongruent}.

Table 3: Overview of size and label number of discussed datasets [6][4][2].

	Train	Dev	Test	Labels
FNC-1	49,972	-	25,413	4
NELA-17 whole	1.7M	100,000	100,000	2
NELA-17 paragraph	14.20M	834,064	100,000	2
Real-News similar	1,347,097	9,493	9,435	2
Real-News random	1,360,095	9,478	9,395	2

2.5 Approaches

After the FNC-1 challenge, a couple of new models were proposed to tackle the headline incongruity problem. They tried new approaches, and some claimed to have better accuracy than all previous models. In this research, only models with a public paper and code base for reproduction purposes are considered. In this subsection, the approach of the three best performing models of the FNC-1 challenge is analyzed. Next, the Deep Hierarchical Encoder model of Yoon et al. is discussed. Finally, the most promising model of the same team using a Graph Neural Network is inspected.

2.5.1 Team SOLAT in the SWEN (2017)

Team SOLAT in the SWEN won the FNC-1 challenge. After testing several different models, they concluded that an ensemble of multiple models had the best result. Their final submission was an ensemble based on a 50/50 weighted average between gradient-boosted decision trees and a deep CNN. The output of the CNN is then converted to the needed 4-class output by an MLP.

Team SOLAT in the Swens model received an FNC-1 score of 0.8202 but based on the balanced metric F_1m they scored 0.582 which made them the worst performing team as can be seen in Table 2. This teams paper and source code[2] on GitHub are publicly available [15]. In their GitHub repository are instructions to reproduce their original submission.

An interesting fact and point of critique that was pointed out by Hanselowski et al. [8] is that the CNN model underperforms dramatically on the small FNC-1 dataset. The CNN solely received an FNC-1 score of 0.502. The gradient-boosted decision trees received an FNC-1 score of 0.830 independently which is surprisingly higher than the combined submission with a score of 0.8202.

2.5.2 Team Athene (UKP Lab) (2017)

The next team used an ensemble of 5 MLP configuration, each with seven hidden layers. Separate bag-of-words unigram features feed them with a TF-IDF configuration combined with baseline features. The MLPs are randomly initialized, and predictions are made with a hard vote by these [16].

Team Athens model received an FNC-1 score of 0.8197 but based on the balanced metric F_1m they scored 0.604 which made them the top-performing team as can be seen in Table 2. This teams paper and source code[3] on GitHub are publicly available [7]. In their GitHub repository are instructions to reproduce their original submission.

2.5.3 Team UCL Machine Reading (2017)

The third best performing model in the FNC-1 challenge is what they self claim as 'a simple but though-to-beat baseline for the FNC-1 stance detection task'[16]. They use lexical and similarity features passed through an MLP with one hidden layer. Text inputs are represented as TF and TF-IDF. For the hidden layer, a ReLU activation function is used.

This model received an FNC-1 score of 0.8172 and an F_1m of 0.583 as can be seen in Table 2. This teams paper and source code[4] on GitHub is publicly available [16]. In their GitHub repository are instructions to reproduce their original submission.

2.5.4 Deep Hierarchical Encoder (2019)

Yoon et al. got inspired by an approach that models textual similarity among question-answer pairs using a hierarchical architecture [17]. They proposed two methods, but we will only discuss the best performing method here. The Attentive Hierarchical Dual Encoder (AHDE) encodes the entire text input from the word to paragraph-level via employing a two-level hierarchy of the RNN architecture. Bi-directional RNNs are added in paragraph-level RNN to exploit information from the past and the future. In Figure 2 a diagram of the ADHE model is shown. For each paragraph, the word-level RNN encodes the word sequences to hidden states h_{1-t}. Next, the hidden states are fed into the next hierarchy RNN which models a sequence of paragraphs while preserving the order. The hidden word states h_{1-t} get encoded to paragraph level hidden states u_p. Then every u_p of the article is aggregated with its headline to a_i. From this, the incongruence score is computed. For a more detailed explanation with all formulas, I recommend consulting the research paper [2].

Next to this RNN, they came with a method where the headline gets compared with each body text paragraph. This information gets then aggregated to form a more accurate decision. They call it the Independent Paragraph (IP) method.

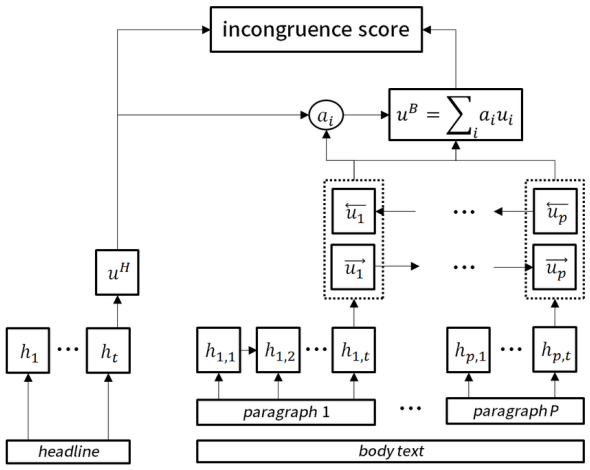

Figure 2: AHDE diagram [5]

On their own proposed Korean dataset (Subsection 2.4.2), they reported accuracy of 0.904 and an AUROC (Area Under Receiver Operating Characteristic) of 0.959 for whole headline-body pairs. For headline-paragraph (IP method) pairs, the accuracy is 0.895 and AUROC 0.977. Different from the three previous approaches is that this is a binary classification (congruent/incongruent). In their research, they also tested this model on the four label FNC-1 dataset. They could do this by transforming the dataset to the binary labels "unrelated" and "other". The ADHE model got an accuracy of 0.844. They claim that the lack of accuracy can be blamed on the small size of the FNC-1 dataset, which is a disadvantage for the proposed hierarchical neural network. In additional tests with a larger dataset, they claim to achieve the best accuracy of any other approaches solely [2].

2.5.5 Graph Neural Network (2021)

A team consisting of mostly the same members proposed an extension on the ADHE model using graphs. They analyzed that most of the earlier models have trouble when the article body gets too long. Their solution is dividing the article into their paragraphs and processing the article in these parts. The GHDE (graph-based hierarchical dual encoder) first computes a node representation for each headline and paragraph using a hierarchical RNN structure. These nodes are paired to calculate a matching score that functions as an edge weight for those nodes. When the graph is complete, the graph neural network propagates information between nodes to examine the article's incongruity. In the last step, all the

[2] https://github.com/Cisco-Talos/fnc-1

[3] https://github.com/hanselowski/athene_system

[4] https://github.com/uclnlp/fakenewschallenge

[5] https://github.com/david-yoon/detecting-incongruity

nodes information is combined to get a final prediction result [4].

This model received an accuracy of 0.852 and an AUROC of 0.928 on their own proposed similar dataset. On the random dataset, they received an accuracy of 0.959 and an AUROC of 0.989. From their comparison results, the newly proposed model performed the best. The next best performing model is the previously discussed ADHE model. Yoon et al. claim to outperform previous state-of-the-art models by a substantial margin (5.3%) on the AUROC curve [4].

3 Methodology

This research aims to determine the current state of solutions for the headline incongruity problem. I will tackle this problem by taking the most promising models in this domain and running comparison tests. This way, I can see where we are today and what still needs to be improved on.

Based on the advantages and disadvantages described below, I choose to use the Real-News based dataset for the comparison tests. The FNC-1 dataset is well tested and used, but the unbalanced distribution and relatively small sample size makes it not an appropriate alternative (Subsection 2.4.1). Another alternative would be to choose the NELA-2017 based dataset. It has a large sample size, and the companion paper shows a balanced label distribution. A disadvantage with this dataset is that it is only available in an integer format without a proper usability explanation. There exists a repository [6] that contains the code that is used to generate the English version in readable CSV format. Unfortunately, the code in its current form is not executable as it throws errors that are not easy to fix. I could not get it to work without debugging the codebase. The Real-News based dataset has a large sample size, is tested, and is released in a readable format. There is also the option to use the similar or random dataset. A possible disadvantage is the binary label distribution which can have an impact on the performance of models that were designed for four-labelled datasets.

All five approaches described in Section 2.5 were planned to be trained and tested to have a good view of all their characteristics. Due to time constraints, which is described in Section 4.3, only the models from the FNC are trained and tested. The most recent models (ADHE and GNN) are already tested on the to be used dataset. To use the binary dataset, the other models will have their four labels grouped as unrelated, which stands for incongruent, or congruent, which stands for agree, disagree, and discuss.

The data preprocessing, training, and testing was conducted on the TU Delft HPC cluster. The models from the FNC are not optimized for such a large dataset, so it will not be feasible to use a personal computer. This is because the FNC-1 dataset is small compared to the newer datasets, as can be seen in Table 3. These models use algorithms and data structures such as a 2D array that use many resources when ran on a large dataset. On the other hand, the cluster can queue multiple jobs and simultaneously use much computational power.

[6] https://github.com/sugoiii/detecting-incongruity-dataset-gen

4 Experimental Setup and Results

In this section, the experimental setup, comparison metrics and final results are discussed.

4.1 Environment

All tested models have different dependencies and environment requirements. For every model, a new python virtual environment has to be created with multiple libraries. Also, some models required an additional process to be able to train their model, e.g., the Athene team required a java Stanford corenlp parser to run simultaneously. As most of the models were implemented in 2017, they required rather old python versions and libraries. The TU Delft HPC cluster does not natively support many of these dependencies, so a workaround using Anaconda2 and installing packages from the source was required. The setup of environments and running these relatively old models took more time than first anticipated.

4.2 Dataset preprocessing

The FNC-1 models are not compatible with the chosen Real-News based dataset. The single TSV train and test files should be split into the needed multiple CSV files to use this dataset. Integer labels also need to be converted to a string format. The companion repository includes the python script used to preprocess the data using the Pandas library.

4.3 Process limitations

The setup of the multiple test environments through the command line took way longer than first anticipated. Next to this, the request for sufficient resource allocation to run the models on the HPC cluster was not incorporated in the initial research planning. As a result, the planned research was not feasible in the 10-week hard time constraint. With my supervisor's permission, I chose to copy the test results for both the ADHE and Graph Neural Network presented in the paper [4] accompanying the used dataset instead of reproducing them myself. Because they are not tested in the same environment, there could be some deviations compared to reproduced results. In the discussion section, the lack of reproduction will be taken into consideration to draw conclusions. Reproducing these results is still an open issue in future work.

4.4 Comparison metrics

The metrics accuracy and AUROC (Area Under Receiver Operating Characteristic) curve are used to compare the different models. This is mainly because the ADHE and graph neural network model are not reproduced, as can be read in the previous section. In the paper where the results are straight copied from, only accuracy and AUROC curve are mentioned. Accuracy is the fraction of predictions the model got right. The AUROC curve is a metric only usable in binary classification that tells how much the model can distinguish between classes. The higher the AUROC, the better the model predicts 0 classes as 0 and 1 classes as 1. AUROC measures how true positive rate (recall) and false-positive rate trade-off. AUROC is a good metric for unbalanced classes because it punishes a model with a bias to the most present class.

4.5 Results

In Tables 4 and 5, the accuracy and AUROC results from testing and the paper by Yoon et al. [4] can be found. Table 4 represents the results from the similar dataset. Table 5 represents the results from the random dataset.

Table 4: Test results on similar dataset [4].

Model	Accuracy	AUROC
SOLAT in the SWEN	0.7254	0.7665
Athene (UKP Lab)	0.6989	0.7323
UCL Machine Reading	0.6761	0.6769
ADHE	0.797	0.879
Graph Neural Network	**0.852**	**0.928**

Table 5: Test results on random dataset [4].

Model	Accuracy	AUROC
SOLAT in the SWEN	0.7543	0.7693
Athene (UKP Lab)	0.7140	0.7117
UCL Machine Reading	0.6993	0.6347
ADHE	0.922	0.971
Graph Neural Network	**0.959**	**0.989**

5 Discussion

In this section, the obtained results are compared to results in earlier research work. Further, a reflection and conclusion are formulated from these results.

From Tables 4 and 5, it can be seen that the Graph Neural Network performs the best with a wide margin. This shows that a headline with corresponding semantically close article paragraphs lends itself well for a graph structure. This model extends the ADHE model that came with a novel approach based on splitting articles into their paragraphs and words using a hierarchy of the RNN architecture. This approach performed significantly better than the three best performing models in the FNC-1 challenge. From the tested models in this research, the model by SOLAT in the SWEN performed the best. The configuration of five randomly initialized MLPs by team Athene that predicted using a hard vote performed persistently better than the single MLP by UCL Machine Reading.

In Table 6 and 7, the results reported by Yoon et al. [4] can be found. XGB stand for XGBoost and is a model that implements gradient boosted decision trees. This model uses the same technology as the XGB part of the model ensemble by Talos in the Swen. Bert is a transformer model that was pre-trained for a masked language model and with a next-sentence prediction objective [18]. The BERT approach stands for measuring the next-sentence prediction performance of BERT. The BDE model was trained using Bert as a backbone but with the weights of the pre-trained BERT network frozen because of the lack of computational resources.

Table 6: Results by Yoon et al. on the similar dataset [4].

Model	Accuracy	AUROC
XGB	0.700	0.776
BDE	0.654	0.712
BERT	0.510	0.487

Table 7: Results by Yoon et al. on the random dataset [4].

Model	Accuracy	AUROC
XGB	0.687	0.756
BDE	0.720	0.799
BERT	0.512	0.561

When a comparison is made between the results of the ensemble model of SOLAT in the SWEN and the XGB model, it can be seen that the ensemble has an impact. In Section 2.5, there was a point of critique on the fact that the CNN severely underperformed compared to the XGB and even made the prediction worse than an individual XGB. The deep CNN model seemed to have finally improved the gradient-boosted decisions trees when trained on a substantially larger dataset than the FNC-1 one. A conclusion that can be drawn from this is that a model needs to be designed with the size of available data in mind. The Athene and UCL Machine Reading models performed somewhat the same in the FNC-1 challenge. However, using this training data size, a relatively small but clear performance gap can be seen. From the ADHE and Graph Neural Network results, it can be concluded that the processing of article bodies as an ensemble of paragraphs is a promising strategy. The BDE and BERT approaches could perform potentially better when the performance is not limited like in the results from Table 6 and 7. Pre-trained BERT models are trained with huge datasets that could be leveraged and tweaked for the headline incongruity problem.

6 Conclusions and Future Work

This paper focuses on the current state of solutions to tackle the headline incongruity problem. The Real-News based dataset from Yoon et al. is the most suitable for comparing models in this research area. From the models compared in this research, the Graph Neural Network by Yoon et al.[4] performed the best. From the tested models, the ensemble of XGB and CNN by team SOLAT in the SWEN [15] had the highest performance.

Some conclusions can be drawn from the results. Processing paragraphs separately has a positive result in terms of performance compared to processing whole article bodies. A Graph structure lends itself well for semantically close headline and article-paragraphs pairs. A Neural Network performs substantially better when trained on a large dataset. A model has to be designed with the size of the dataset in mind. In future research, a BERT approach with enough computational resources could be a good alternative.

There have been many improvements made to solve the problem in recent years. The problem of clickbait and fake news will keep growing, and solutions have to be found. I hope this research gives a good view of the current state of

tackling the headline incongruity problem using stance detection and contributes to the construction of more credible online environments for news consumption.

6.1 Future work

In future research, several open issues and recommendations can be formulated. Firstly, the ADHE and Graph Neural Network results can be reproduced in the same environment as the rest of the models. By doing this, the results can be compared appropriately. Another recommendation is the training of a BERT model with the proper computational resources. This can potentially be a good strategy to solve the headline incongruity problem and is worth testing.

7 Responsible Research

In the following section, ethical aspects and the reproducibility of the research methods are discussed.

7.1 Societal impact

Solving the headline incongruity problem has a positive societal impact. With an effective solution, all worrisome problems that an incongruent article creates could be solved. It could give the reader a congruity score or indicate if the article is a form of clickbait or sensationalism. The reader will not be misled by the headline when he does not read the article body. The stance and opinion of the reader will not be that misshaped if he knows if a headline is structured to grab a readers attention or dramatize. A reader will know beforehand if the article is worth reading based on the headline and the solution metrics. Lastly, a reader can not be tempted much to read an article when a clickbait or sensational headline is used.

7.2 Reproducibility

Next to this paper, a public repository is published that helps reproduce all the calculated results for this paper. The used Real-News dataset can be downloaded by following the process described on their GitHub page [7]. All used models can be found on their repositories which are referenced in their respective section. The public repository for this paper[8] contains all additional python scripts for preprocessing and calculation. Lastly, all publications and repositories used can be found in the references to ensure full reproducibility.

8 Acknowledgements

I want to express my deep and sincere gratitude to my research supervisor and responsible professor Asst. Prof. Pradeep Murukannaiah. His domain insight and helpful guiding helped me a lot. I would also want to thank Wout Haakman, Abel van Steenweghen, Jacob Roeters van Lennep, and Kristóf Vassall as my group members to together make our research proposal for the CSE3000 Research Project. The regular feedback, teamwork and support helped my research for the better. Lastly I want to thank my home university TU Delft for the available computing resources on the HPC cluster. Without this computing power, I would not be able to run all tests.

[7] https://github.com/minwhoo/detecting-incongruity-gnn

[8] https://github.com/simonmarien/headline-incongruity-problem

References

[1] S. Chesney, M. Liakata, M. Poesio, and M. Purver, *Incongruent headlines: Yet another way to mislead your readers*, Jan. 2017. DOI: 10.18653/v1/W17-4210.

[2] S. Yoon, K. Park, J. Shin, H. Lim, S. Won, M. Cha, and K. Jung, "Detecting incongruity between news headline and body text via a deep hierarchical encoder", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 791–800.

[3] R. Mishra, P. Yadav, R. Calizzano, and M. Leippold, "Musem: Detecting incongruent news headlines using mutual attentive semantic matching", *CoRR*, vol. abs/2010.03617, 2020. arXiv: 2010.03617. [Online]. Available: https://arxiv.org/abs/2010.03617.

[4] S. Yoon, K. Park, M. Lee, T. Kim, M. Cha, and K. Jung, "Learning to detect incongruence in news headline and body text via a graph neural network", *IEEE Access*, vol. PP, pp. 1–1, Feb. 2021. DOI: 10.1109/ACCESS.2021.3062029.

[5] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, *SemEval-2016 task 6: Detecting stance in tweets*, San Diego, California, Jun. 2016. DOI: 10.18653/v1/S16-1003. [Online]. Available: https://www.aclweb.org/anthology/S16-1003.

[6] D. Küçük and F. Can, "Stance detection: A survey", *ACM Comput. Surv.*, vol. 53, no. 1, Feb. 2020, ISSN: 0360-0300. DOI: 10.1145/3369026. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3369026.

[7] H. Andreas, P. Avinesh, S. Benjamin, C. Felix, C. Debanjan, M. C. M., and G. Iryna, "Description of the system developed by team athene in the fnc-1", Jun. 2017. [Online]. Available: https://github.com/hanselowski/athene_system/blob/master/system_description_athene.pdf.

[8] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych, "A retrospective analysis of the fake news challenge stance-detection task", in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1859–1874. [Online]. Available: https://www.aclweb.org/anthology/C18-1158.

[9] *Fake news challenge stage 1 (fnc-i): Stance detection*. [Online]. Available: http://www.fakenewschallenge.org/.

[10] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading online content: Recognizing clickbait as "false news"", in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, ser. WMDD '15, Seattle, Washington, USA: Association for Computing Machinery, 2015, pp. 15–19, ISBN: 9781450339872. DOI:

10.1145/2823465.2823467. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/2823465.2823467.

[11] J. Du Bois, "The stance triangle", in. Jan. 2007, pp. 139–141. DOI: 10.1075/pbns.164.07du.

[12] B. D. Horne, W. Dron, S. Khedr, and S. Adali, "Sampling the news producers: A large news and feature data set for the study of the complex media landscape", *CoRR*, vol. abs/1803.10124, 2018. arXiv: 1803.10124. [Online]. Available: http://arxiv.org/abs/1803.10124.

[13] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news", *CoRR*, vol. abs/1905.12616, 2019. arXiv: 1905.12616. [Online]. Available: http://arxiv.org/abs/1905.12616.

[14] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations", *CoRR*, vol. abs/1712.09405, 2017. arXiv: 1712.09405. [Online]. Available: http://arxiv.org/abs/1712.09405.

[15] S. Baird, D. Sibley, and Y. Pan, Jun. 2017. [Online]. Available: https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html.

[16] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, "A simple but tough-to-beat baseline for the fake news challenge stance detection task", *CoRR*, vol. abs/1707.03264, 2017. arXiv: 1707.03264. [Online]. Available: http://arxiv.org/abs/1707.03264.

[17] S. Yoon, J. Shin, and K. Jung, "Learning to rank question-answer pairs using hierarchical recurrent encoder with latent topic clustering", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1575–1584. DOI: 10.18653/v1/N18-1142. [Online]. Available: https://www.aclweb.org/anthology/N18-1142.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding", *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.