From Clickbait to Fake News Detection:
An Approach based on Detecting the Stance of Headlines to Articles

Peter Bourgonje, Julian Moreno Schneider, Georg Rehm
DFKI GmbH, Language Technology Lab
Alt-Moabit 91c, 10559 Berlin, Germany
{peter.bourgonje,julian.moreno_schneider,georg.rehm}@dfki.de

Abstract

We present a system for the detection of the stance of headlines with regard to their corresponding article bodies. The approach can be applied in fake news, especially clickbait detection scenarios. The component is part of a larger platform for the curation of digital content; we consider veracity and relevancy an increasingly important part of curating online information. We want to contribute to the debate on how to deal with fake news and related online phenomena with technological means, by providing means to separate related from unrelated headlines and further classifying the related headlines. On a publicly available data set annotated for the stance of headlines with regard to their corresponding article bodies, we achieve a (weighted) accuracy score of 89.59.

1 Introduction

With the advent of social media and its increasingly important role as a provider and amplifier of news, basically anyone, anywhere, can produce and help circulate content for other people to read. Traditional barriers to publishing content (like a press to print newspapers or broadcasting time for radio or television) have disappeared, and with this, at least part of traditional quality control procedures have disappeared as well. Basic journalistic principles like source verification, fact checking and accountability can be easily bypassed or simply ignored by individuals or organisations publishing content on Twitter, Facebook or other social networks. The impact of this situation is illustrated by the predominance of terms like "trolls", "fake news", "post-truth media" and "alternative facts". There is evidence that these developments and their effects are not harmless but can have a significant impact on real-world events, which is illustrated by a description of the role of social media in the 2016 US presidential election by (Allcott and Gentzkow, 2017), and by a study on the effectiveness and debunking strategies of rumours surrounding the Affordable Care Act by (Berinsky, 2017).

While the cause of this situation may have its roots in many different aspects of modern society, and hence needs to be approached from several different angles, we aim to make a contribution from the angle of Language Technology and Natural Language Processing. We consider fully-automated procedures for fact-checking, clickbait detection or fake news classification not feasible at this point (Rehm, 2017), but aim to support the community by providing means of detecting articles or pieces of news that need to be approached with caution, where a human has to make final decisions (on credibility, legitimacy etc.), but is aided by a set of tools. The approach described in this paper can serve as the back-end of such a smart set of tooling around fact-checking and can augment news coming from both traditional and non-traditional (social media) sources. We envision the resulting set of tools as a collection of expert tools for specific job profiles (like a journalist or a news editor), or in the shape of a simple browser plug-in, flagging unverified or dubious content to the end user.

The work presented in this paper was carried out under the umbrella of a two-year research and technology transfer project, in which a research centre collaborates with four SME partners that face the challenge of having to process, analyse and make sense of large amounts of digital content. The companies cover four different use cases and sectors (Rehm and Sasaki, 2015) including journalism. For these partners we develop a plat-

Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism, pages 84–89
Copenhagen, Denmark, September 7, 2017. ©2017 Association for Computational Linguistics

form that provides access to language and knowledge technologies (Bourgonje et al., 2016a,b). The services are integrated by the SME partners into their own in-house systems or those of clients.

In this paper, we aim to contribute to a first step in battling fake news, often referred to as stance detection, where the challenge is to detect the stance of a claim with regard to another piece of content. Our experiments are based on the setup of the first Fake News Challenge (FNC1).[1]. In FNC1, the claim comes in the form of a headline, and the other piece of content is an article body. This step may seem, and, in fact, is, a long way from automatically checking the veracity of a piece of content with regard to some kind of ground truth. But the problem lies exactly in the definition of the truth, and the fact that it is sensitive to bias. Additionally, and partly because of this, annotated corpora, allowing training and experimental evaluation, are hard to come by and also often (in the case of fact checker archives) not freely available. We argue that detecting whether a piece of content is related or not related to another piece of content (e. g., headline vs. article body) is an important first step, which would perhaps best be described as clickbait detection (i. e., a headline not related to the actual article is more likely to be clickbait). Following the FNC1 setup, the further classification of related pieces of content into more fine-grained classes provides valuable information once the "truth" (in the form of a collection of facts) has been established, so that particular pieces of content can be classified as "fake" or, rather, "false". Since this definitive, resolving collection of facts is usually hard to come by, the challenge of stance detection can be put to use combining the outcome with credibility or reputation scores of news outlets, where several high-credibility outlets disagreeing with a particular piece of content point towards a false claim. Stance detection can also prove relevant for detecting political bias: if authors on the same end of the political spectrum are more likely to agree with each other, the (political) preference of one author can be induced once the preference of the other author is known. Additionally, the stances of utterances towards a specific piece of content can provide hints on its veracity. (Mendoza et al., 2010) show that the propagation of tweets regarding crisis situations (like natural disasters) differs based on their content: tweets spreading news are affirmed by related tweets, whereas tweets spreading rumours are mostly questioned or denied. In this paper we propose a solution that involves the human-in-the-loop. We think that our approach can be a valuable part of solving the problem described above. The rest of this paper is divided into five sections. Section 2 reviews related work, Section 3 describes the data set used, Section 4 explains our approach in detail and Section 5 provides an evaluation. Our conclusions are presented in Section 6.

2 Related Work

The suggestion of using Language Technologies (NLP, NLU etc.) to design solutions for modern online media phenomena such as "fake news", "hate speech", "abusive language", etc. is receiving rapidly growing interest in the form of shared tasks, workshops and conferences. The awareness that LT can contribute to solutions related to these topics is present. Yet, at the same time, it is being acknowledged that the problem is much more complex than anything that can be solved by exploiting current state of the art techniques alone. The effect known as "belief perseverance" or "continued influence effect" (Wilkes and Leatherbarrow, 1988) and its influence on modern media and politics is described by (Nyhan and Reifler, 2015), who state that reasoning based on facts that have shown to be false, remains in place until an alternative line of reasoning has been offered. The credibility of a politician stepping down due to bribery accusations is not restored after only rejecting this explanation (by a letter from the prosecutors). In addition, an alternative explanation (like being named president of a university, but not being able to disclose this until the predecessor has stepped down) has to be provided. Another socio-psychological contribution on the topic of "fake news" and its consumption is presented by (Marchi, 2012) who report on a survey among teenagers and their news consumption habits. Although they have a slightly different definition of "fake news" than the one we use in this paper, the study presents a relevant overview of the consumption of news and the important aspects with different social groups. The authors claim that "authenticity" is highly valued among teenagers consuming news, hence their explained preference for blogs, satirical shows,

[1] http://www.fakenewschallenge.org

or basically anything other than traditional media outlets, which they consider “identical”, lacking contextual information and any authenticity. The acknowledgment that teenagers increasingly rely on news coming from non-traditional news sources underlines the need for new ways of dealing with challenges related to these alternative sources. (Conroy et al., 2015) present a useful overview of recent approaches towards “fake news” detection using NLP and network analyses. The authors include several state-of-the-art figures and acknowledge the fact that these numbers are domain-dependent, which is why it is difficult to arrive at a state-of-the-art figure independent of a specific use case and data set. From an NLP perspective, the challenge of dealing with this problem is further exemplified by the fact that annotated data is hard to find, and, if present, exhibits rather low inter-annotator agreement. Approaching the “abusive language” and “hate speech” problem from an NLP angle (Bourgonje et al., 2017), (Ross et al., 2016) introduce a German corpus of tweets and annotate it for hate speech, resulting in figures for Krippendorff’s α between 0.18 and 0.29, (Waseem, 2016) compare amateur (CrowdFlower) annotations and expert annotations on an English corpus of Tweets and report figures for Cohen’s Kappa of 0.14, (Van Hee et al., 2015) use a Dutch corpus annotated for cyberbullying and report Kappa scores between 0.19 and 0.69, and (Kwok and Wang, 2013) investigate English racist tweets and report an overall inter-annotator agreement of only 33%.

An approach similar to ours is described by (Ferreira and Vlachos, 2016), who introduce a data set and three-class classification (“for”, “against”, “observing”). In addition to a logistic regression classifier, the authors exploit dependency parse graphs, a paraphrase database (Pavlick et al., 2015) and several other features, to arrive at an accuracy of 73%. Another related approach is described by (Augenstein et al., 2016), who apply stance detection methods on the SemEval 2016 Task 6 data set. Their focus is on learning stances towards a topic in an unsupervised and weakly supervised way using a neural network architecture. (Babakar and Moy, 2016) present a useful and recent overview of fact checking approaches.

3 Data Set

Our experiments are conducted on the dataset released by the organisers of the first Fake News Challenge (FNC1) on stance detection. The data set is based on the work of (Ferreira and Vlachos, 2016) and can be downloaded from the corresponding GitHub page, along with a baseline implementation for this task, achieving a score of 79.53.[2] The data consists of a set of headlines and articles that are combined with each other (multiple times, in different combinations) and annotated for one of four classes: “unrelated”, “agree”, “disagree”, “discuss”, indicating the stance of the headline towards the content of the article (see Table 1).

Unique headlines	1.648	
Unique articles	1.668	
Annotated pairs	49.972	100%
Class: unrelated	36.545	73%
Class: discuss	8.909	18%
Class: agree	3.678	7%
Class: disagree	840	2%

Table 1: Key figures of the FNC-1 data set

The FNC1 scoring method consists of first verifying whether a particular combination of headline and article has been correctly classified as “unrelated” (the corresponding class) or “related” (one of the classes “agree”, “disagree” or “discuss”). Getting this binary classification correct amounts up to 25% of the final, weighted score. The remaining 75% of the score consists of correctly classifying headline article pairs in the three remaining classes. The setup of our system adheres to this scoring method, and hence applies several classifiers sequentially, as explained in Section 4.

4 Approach and Methods

In line with the scoring system of the challenge, we first apply a procedure to decide whether a particular headline/article combination is related or unrelated. This is done based on n-gram matching of the lemmatised input (headline or article), using the CoreNLP Lemmatiser (Manning et al., 2014). The number of matching n-grams (where $n = 1..6$) in the headline and article is multiplied by length and IDF value of the matching n-gram

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

(n-grams containing only stop words or punctuation are not considered), then divided by the total number of n-grams. If the resulting score is above some threshold (we established 0.0096 as the optimal value), the pair is taken to be related.

A formal definition is provided in Equation 1: considering a headline and an article represented by two arrays (H and A) of all possible lemmatised n-grams when $n \in [1,6]$, $h(i)$ and $a(i)$ being the i^{th} element of arrays H and A, $len(\cdot)$ being a function that computes the length in tokens of a string (n-gram), TF_T^k being the frequency of appearance of term k in array T and IDF^k being the inverse document frequency of term k in all the articles.

$$sc = \frac{\sum_{i=1}^{len(H)} TF^{h(i)} * IDF^{h(i)}}{len(H) + len(A)} \quad (1)$$

where

$$TF^{h(i)} = \{(TF_H^{h(i)} + TF_A^{H(i)}) * len(h(i))\} \quad (2)$$

As shown in Table 1, the majority of "related" instances are of the class "discuss" and simply assigning this class to all "related" instances leads to an accuracy of 61.51 already (for this portion of the data set), as shown in the "Majority vote" column. To improve upon this baseline and to further classify the related pairs into "agree", "disagree" or "discuss", we use Mallet's Logistic Regression classifier implementation (McCallum, 2002) trained on headlines only (without lemmatisation or stop word removal), using the three classes. This resulted in a weighted score of 79.82 (column "3-class classifier"). In subsequent experiments, we introduced a (relative) confidence threshold: if the distance between the best scoring class and the second-best scoring class is above some threshold (we established 0.7 as the optimal value), the best-scoring class is assigned to the pair. If the difference was below the threshold, we used three binary classifiers to decide between the best scoring class and the second-best scoring class (i. e., one binary classifier for "agree"-"disagree", one for "agree"-"discuss" and one for "discuss"-"disagree"). These classifiers are trained on both the headline and the article (joined together, without lemmatisation or stop word removal). The results are shown in the column "Combined classifiers" in Table 2.

This setup leads to the best results on the data set. In other experiments we used more linguistically motivated features, some of them inspired by the work of (Ferreira and Vlachos, 2016). From rather basic ones (like a question mark at the end of a headline to detect "disagree" instances) to more sophisticated ones (like extracting a dependency graph, looking for negation-type typed dependencies and calculate their normalised distance to the root node of the graph, and compare this value for headline and article), but these did not improve upon the final weighted score reported in Table 2.

5 Evaluation

The first step of deciding whether a headline/article pair is related or not is done based on n-gram matching (of lemmatised n-grams). This procedure is rule-based and only relies on finding an optimal value for the threshold, based on the data. To arrive at an optimal value, we used all data and did not separate it into training and test sets. Since the subsequent classification methods are based on machine learning, the following evaluation figures are the result of 50-fold cross-validation, with a 90-10 division of training and test data, respectively.

Considering that the combination of headlines and article bodies has been performed randomly with many obviously unrelated combinations, the relatedness score of 93.27 can be considered relatively low.[3] Upon manual investigation of the cases classified as "unrelated" (but that were in fact of the "agree", "disagree" or "discuss" class), we found that the vast majority had headlines with different wordings that were not matching after lemmatisation. One concrete example with the headline "Small Meteorite Hits Managua" in its article body mentions "the Nicaraguan capital" but not "Managua" and "a chunk of an Earth-passing asteroid" instead of "small meteorite". To improve the approach for cases such as this one, we propose to include more sophisticated techniques to capture word relatedness in a knowledge-rich way as an important part of future work. The other way round, cases classified as related that were in fact annotated as "unrelated" contained words in the headline that were frequently mentioned in the article body. One example with the headline "SHOCK CLAIM: PGA Golfer Says Tiger Woods Is Suspended For Failed Drug Test" was combined

[3] The variation for this row in Table 2 is due to different runs (on different, random splits of the data).

	Majority vote	3-class classifier	Combined classifiers
Relatedness score	93.27	93.26	93.29
Three-class score	61.51	75.34	88.36
Weighted score	69.45	79.82	**89.59**

Table 2: Results of 50-fold cross-validation

with an article body about the divorce of Tiger Woods and Elin Nordegren. Here, we suggest, as part of future work, to include event detection, to move away from entity-based representations and put more focus on the event actually reported.

After deciding on relatedness, we are left with (on average) 1,320 instances. For the three-class classification of this set, we obtained (on average) 686 cases that scored above the scoring difference threshold and were assigned their class by this three-class Logistic Regression classifier. Of these, 642 were correct, resulting in an accuracy of 93.64 for this portion of the data set (i. e., "related"). The average number of cases where the scoring difference was below the threshold (634) were classified using the three binary classifiers. This resulted in 544 correctly classified instances, and a score of 85.83 for this section of the data set. Putting these scores together, the weighted score and the individual components are shown in Table 2, i. e., the relatedness score for the binary decision "related" or "unrelated" (25% of the weighted score) and the three-class score for the classification of "related" instances into "agree", "disagree" or "discuss" (75% of the weighted score). To get an idea of the effect of the first stage's error rate on the second stage of processing, we re-ran the experiments taking the "related" vs. "unrelated" information from the annotations directly. This resulted in a three-class score of 89.82, i. e., a 1.46 drop in accuracy due to classification errors in the first stage.

While these numbers look promising for initial steps towards tackling the challenge that fake news poses globally, we acknowledge that at least the 25% of the score (the relatedness score of 93.27) is not directly applicable in a real world scenario, since the data set was artificially boosted by randomly combining headlines and article bodies – a headline such as "Isis claims to behead US journalist" is combined with an article on who is going to be the main actor in a biopic on Steve Jobs. Although this headline/article pair was (obviously) tagged as "unrelated", this is not something that is usually encountered in a real-world scenario. For the more fine-grained classification of articles that have been classified as "related", the three-way classification is a relevant first step, but other classes may need to be added to the set, or a more detailed division may need to be made in order to take the next steps in tackling the fake news challenge. Additionally, we see the integration of known facts and general discourse knowledge (possibly through Linked Data), and the incorporation of source credibility information as important and promising suggestions for future research.

6 Conclusions

We present a system for stance detection of headlines with regard to their corresponding article bodies. Our system is based on simple, lemmatisation-based *n*-gram matching for the binary classification of "related" vs. "unrelated" headline/article pairs. The best results were obtained using a setup where the more fine-grained classification of the "related" pairs (into "agree", "disagree", "discuss") is carried out using a Logistic Regression classifier at first, then three binary classifiers with slightly different training procedures for the cases where the first classifier lacked confidence (i. e., the difference between the best and second-best scoring class was below a threshold). We improve on the accuracy base line set by the organisers of the FNC1 by over 10 points and scored 9th place (out of 50 participants) in the actual challenge. As described in Section 1, the approach explained in this article can be part of the set of services needed by a fact-checking tool (Rehm, 2017). The first, binary classification of "related" vs. "unrelated" can be exploited for clickbait detection. The more fine-grained classification of related headlines can specifically support in the detection of political bias and rumour veracity (Srivastava et al., 2017).

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful feedback. The project "Digitale Kuratierungstechnologien" is supported by the German Federal Ministry of Education and Research, Wachstumskern-Potenzial (no. 03WKP45). http://www.digitale-kuratierung.de.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Working Paper 23089, National Bureau of Economic Research.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *CoRR*, abs/1606.05464.

Mevan Babakar and Will Moy. 2016. The State of Automated Factchecking.

Adam J. Berinsky. 2017. Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, 47(2):241–262.

Peter Bourgonje, Julian Moreno-Schneider, Jan Nehring, Georg Rehm, Felix Sasaki, and Ankit Srivastava. 2016a. Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer. In *The Semantic Web*, number 9989 in LNCS, pages 65–68. Springer.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. Automatic Classification of Abusive Language and Personal Attacks in Various Forms of Online Communication. In *Language Technologies for the Challenges of the Digital Age: Proc. of GSCL 2017*, Berlin.

Peter Bourgonje, Julian Moreno Schneider, Georg Rehm, and Felix Sasaki. 2016b. Processing Document Collections to Automatically Extract Linked Data: Semantic Storytelling Technologies for Smart Curation Workflows. In *Proc. of the 2nd Int. Workshop on NLG and the Semantic Web (WebNLG 2016)*, pages 13–16, Edinburgh, UK.

Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. In *Proc. of the 78th ASIS&T Annual Meeting*, ASIST '15, pages 82–82.

W. Ferreira and A. Vlachos. 2016. Emergent: a novel data-set for stance classification. The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proc. of the 27th AAAI Conf. on AI*, pages 1621–1622.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Toolkit. In *ACL*, pages 55–60.

R. Marchi. 2012. With facebook, blogs, and fake news, teens reject journalistic objectivity. *Journal of Communication Inquiry*, 36(3):246–262.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter Under Crisis: Can We Trust What We RT? In *Proc. of the First Workshop on Social Media Analytics*, pages 71–79.

Brendan Nyhan and Jason Reifler. 2015. Displacing misinformation about events: An experimental test of causal corrections. *Journal of Exp. Political Science*, 2(01):81–93.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proc. of the 53rd Annual Meeting of the ACL, Beijing, China, Volume 2: Short Papers*, pages 425–430.

Georg Rehm. 2017. An Infrastructure for Empowering Internet Users to handle Fake News and other Online Media Phenomena. In *Language Technologies for the Challenges of the Digital Age: Proc. of GSCL 2017*, Berlin.

Georg Rehm and Felix Sasaki. 2015. Digitale Kuratierungstechnologien – Verfahren für die effiziente Verarbeitung, Erstellung und Verteilung qualitativ hochwertiger Medieninhalte. In *Proc. of GSCL 2015*, pages 138–139.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proc. of NLP4CMC III: 3rd Workshop on NLP for CMC*, pages 6–9.

Ankit Srivastava, Georg Rehm, and Julian Moreno Schneider. 2017. DFKI-DKT at SemEval-2017 Task 8: Rumour Detection and Classification Using Cascading Heuristics. In *Proc. of SemEval-2017*, pages 477–481, Vancouver, Canada. ACL.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proc. of the Int. Conf. Recent Advances in NLP*, pages 672–680.

Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proc. of the 1st Workshop on NLP and Computational Social Science*, pages 138–142.

A. L. Wilkes and M. Leatherbarrow. 1988. Editing episodic memory following the identification of error. *The Quarterly Journal of Exp. Psychology*, 40(2):361–387.